INFORME DEL GRUPO DE ANALISIS CIENTÍFICO DE CORONAVIRUS DEL ISCIII (GACC-ISCIII).

SECUENCIACIÓN GENÉTICA: ¿QUÉ ES Y PARA QUÉ SIRVE?

15 de abril de 2020

IMPORTANTE: *Este informe está realizado con la evidencia científica disponible en este momento y podrá ser actualizado si surgen nuevas evidencias.*

RESUMEN DIVULGATIVO

La secuenciación genética es una tecnología que permite conocer y descifrar el código genético que tienen todos los seres vivos. Se trata de 'leer' ese código, que contiene información imprescindible para su desarrollo y funcionamiento, como si de un libro de instrucciones genéticas se tratase. Estas señas de identidad, que definen las características y la 'firma genética' de los organismos biológicos, vienen 'inscritas' en moléculas llamadas ácidos nucleicos, formadas por nucleótidos.

En el caso de los virus hay un importante debate científico sobre si son realmente organismos vivos, ya que no son capaces de realizar algunas de las funciones biológicas primordiales. En todo caso, la secuenciación genómica del nuevo coronavirus ha sido desde su descubrimiento uno de los principales objetivos, ya que es la puerta de entrada para poder conocerlo y combatirlo.

En lo que va de año 2020 se han conseguido secuenciar miles de genomas completos del coronavirus, gracias al análisis de muestras de pacientes afectados por la enfermedad COVID-19. Lograr esta secuenciación es fundamental para conocer mejor el virus y definir sus características y comportamiento. De entrada, la secuenciación permitió clasificarlo, definirlo e incluirlo como un nuevo miembro de las familias de virus ya conocidas, bautizándolo como SARS-CoV-2. La secuenciación genómica del SARS-CoV-2 ha permitido averiguar su origen (ver informe sobre origen del coronavirus), saber cómo se transmite (ver informe sobre mecanismos de transmisión), investigar su capacidad de difusión y contagio, y lograr información necesaria para el futuro desarrollo de fármacos y vacunas.

En la actualidad la mayoría de centros de investigación son capaces de hacer secuenciación genética. Hay diferentes tecnologías para llevarla a cabo. La secuenciación de Sanger, una de las primeras en desarrollarse y clave para automatizar el proceso de secuenciación que se conoce hoy, sigue siendo una referencia. A lo largo de los años han ido surgiendo nuevas tecnologías que permiten obtener más información del organismo secuenciado de manera más rápida. Entre ellas destacan tecnologías como Illumina e IonTorrent, consideradas parte de la segunda generación de secuenciación genómica, y Pacific Bioscience y Oxford Nanopore, que ya forman parte de una tercera generación de esta tecnología.

La secuenciación genómica ha protagonizado uno de los grandes hitos científicos del siglo XXI, la presentación del Proyecto Genoma Humano, que desveló nuestro código genético y que ha revolucionado el estudio de nuestras características biológicas y la lucha contra las enfermedades. Entre las aplicaciones de la secuenciación están el mayor conocimiento de los

orígenes de las especies, la detección precoz de síndromes y de genes asociados a enfermedades y la identificación de personas en ciencia forense, entre otras.

INFORME COMPLETO

¿Qué es la secuenciación genética?

Todos los seres vivos están formados por diferentes tipos de macromoléculas, que incluyen los glúcidos, los lípidos, las proteínas y los ácidos nucleicos. Estos últimos, están formados por la repetición de unidades básicas llamadas nucleótidos que se unen formando largas cadenas. Los ácidos nucleicos tienen una importante función, ya que contienen el código genético necesario para el desarrollo y funcionamiento de todos los seres vivos, es decir para la vida. Existen dos tipos principales de ácidos nucleicos, el ácido desoxirribonucleico (ADN), formado por dos cadenas de nucleótidos entrelazadas que forman una estructura de doble hélice, y el ácido ribonucleico (ARN) que está formado por una única cadena. La estructura básica de los ácidos nucleicos, el nucleótido, está siempre formada por tres elementos, un glúcido, una base nitrogenada y un grupo fosfato. Las bases nitrogenadas son las que diferencian unos nucleótidos de otros y en el caso del ADN pueden ser: adenina o A, timina o T, citosina o C y guanina o G. La secuenciación genética es, por tanto, la determinación del orden de nucleótidos en una molécula de ácido nucleico mediante procesos físico-químicos.

Posiblemente el principal hito del siglo XXI en ciencia haya sido el Proyecto Genoma Humano y la publicación en el año 2001 del primer borrador de la secuencia del genoma humano. Es decir, se consiguió determinar la secuencia genética contenida en los 23 cromosomas humanos o dicho de otra manera el orden de los más de 4,500 millones de nucleótidos que conforman el genoma humano. Esta carrera por conseguir la secuencia del genoma humano tuvo importantes consecuencias en el desarrollo de nuevas tecnologías que permitieran producir las grandísimas cantidades de datos que exigía un proyecto de esta envergadura.

¿Cómo se lleva a cabo?

La secuenciación genética es una técnica que se desarrolla en prácticamente todos los laboratorios de investigación biológica y médica en la actualidad y existen diferentes tecnologías para conseguirlo. El método de referencia actualmente es la secuenciación de Sanger por electroforesis capilar, pero en los últimos 20 años se ha producido una gran expansión de nuevos métodos que se conocen en su conjunto como secuenciación de alto rendimiento y que incluye métodos de segunda y tercera generación. Todos estos métodos conviven actualmente debido a que tienen aplicaciones diferentes.

1.- Secuenciación de Sanger por electroforesis capilar. Este método es una modificación de la estrategia diseñada en 1975 por el científico Frederick Sanger en la que se utilizaban nucleótidos modificados químicamente para que al añadirse a una nueva cadena que se está formando, ésta no pudiera continuar [1]. Es decir, funcionan como nucleótidos de paro. El desarrollo de las técnicas fluorescentes, la mejora en las enzimas necesarias para llevar a cabo el proceso y la introducción de la electroforesis capilar permitieron automatizar el proceso y llegar hasta los equipos que tenemos actualmente [2].

2.- Métodos de alto rendimiento de segunda generación. La característica principal de estos nuevos métodos es su capacidad de llevar a cabo millones de reacciones de secuenciación de

forma simultánea. El desarrollo de estos métodos permitieron grandes avances en el Proyecto Genoma Humano. Dentro de este apartado, actualmente existen dos importantes tecnologías, Illumina e IonTorrent basadas en principios diferentes.

- **Illumina**. Actualmente, la referencia es la tecnología llevada a cabo por la compañía de San Diego (EEUU), que utiliza una amplificación previa en los fragmentos a secuenciar para formar millones de clusters de ADN. Utiliza también nucleótidos modificados asociados a moléculas fluorescentes que al unirse a la cadena que sirve como molde, emiten una señal que es captada por una cámara. El análisis de las imágenes permite determinar qué nucleótido es el que se ha unido en cada ciclo de la reacción, permitiendo determinar la secuencia genética.
- **IonTorrent**. Otra tecnología de este tipo utilizada actualmente es IonTorrent de la compañía ThermoFisher. En este caso, se utilizan micropocillos donde se coloca el ADN a secuenciar y se añaden nucleótidos. Cuando uno de estos nucleótidos se añaden a la nueva cadena que se está formando, libera ciertas moléculas cargadas, incluyendo iones de hidrógeno, que alteran el pH. Estas plataformas integran potentes sensores de pH que detectan la unión del nucleótido en cada ciclo.

3.- Métodos de alto rendimiento de tercera generación. Estos métodos van un paso más allá y son capaces de secuenciar moléculas de ADN sin amplificación previa, sin necesidad de seguir la estrategia de secuenciación por síntesis que siguen los métodos de segunda generación. Este tipo de métodos incluyen tecnologías como Pacific BioSciences u Oxford Nanopore.

- **Pacific Biosciences**. Esta tecnología está basada en guías de onda "modo cero" (*zero-mode waveguide*), que son estructuras donde se integra una polimerasa (enzima capaz de unir nucleótidos para formar cadenas de ácidos nucleicos) y que aprovechan el comportamiento de la luz para determinar qué nucleótidos marcados con moléculas fluorescentes son añadidas en cada ciclo de secuenciación.
- **Oxford Nanopore**. Esta tecnología consigue determinar la secuencia genética de un ácido nucleico al hacerlo pasar mediante campos eléctricos a través de un nanoporo en el que se detectan cambios en la densidad del flujo eléctrico

¿Para qué sirve en la investigación sobre SARS-CoV-2?

La comunidad científica mundial se ha lanzado a secuenciar el genoma del virus causante de la pandemia de la COVID-19, consciente de la importancia que puede tener esta información para abordar el desafío que supone la enfermedad actualmente. Así, se han conseguido secuenciar 8275 genomas completos del virus aislado de pacientes de todo el mundo y en un tiempo récord, incluyendo 5131 de Europa, 1779 de América del Norte, 821 de Asia, 399 de Oceanía, 90 de África y 54 de América del Sur. España ha secuenciado ya 151 genomas completos. Gracias a esta excelente fuente de información, los científicos tienen cada día más herramientas para avanzar en diferentes aplicaciones.

- **Clasificación del virus**. El análisis del genoma del SARS-CoV-2 ha permitido ubicar este nuevo virus en el árbol de la diversidad de los virus conocidos hasta la fecha mediante lo que se conoce como análisis filogenéticos. Así, la secuenciación del genoma del virus aislado de uno de los primeros pacientes detectados en la ciudad china de Wuhan ha permitido clasificar al virus secuenciado como un nuevo miembro de la familia *Coronaviridae*, subfamilia *Orthocoronavirinae*, género *Betacoronavirus*, subgénero *Sarbecovirus*, especie *coronavirus relacionado con síndrome respiratorio severo agudo* (SARS) [3].

- **Origen del virus**. Gracias a la secuenciación de genomas completos del virus se han podido realizar diferentes hipótesis acerca del origen del virus. Esto es posible porque se ha podido comparar su secuencia con la de otros virus aislados en animales y cuyo genoma ha sido previamente secuenciado y publicado para que cualquier investigador pueda estudiarlo. De este modo, el genoma de SARS-CoV-2 a lo que más se parece actualmente es a un virus aislado en murciélagos con el que comparte el 96% de su genoma [4]. También se ha podido determinar que es muy parecido a otro coronavirus aislado en un pangolín y con el que comparte un 90% del genoma [5].

- **Transmisión del virus**. La comparación de los genomas que se van secuenciando en todo el mundo así como la información de cuándo y dónde se han obtenido esas muestras permite a los investigadores identificar el posible inicio de la epidemia, trazar posibles rutas de transmisión entre ciudades y países así como monitorizar su diseminación geográfica. De forma paralela, este tipo de análisis permiten conocer cuánto está mutando el virus.

- **Patogenicidad del virus**. Gracias a la secuenciación del genoma del virus y a su posterior análisis se ha podido identificar que su genoma está compuesto por una única cadena de ARN de polaridad positiva formado por aproximadamente 30.000 nucleótidos. Se conocen al menos 6 marcos de lectura abiertos (ORFs) que incluyen ORF1a y ORF1b que codifican dos poliproteínas que son procesadas por al menos tres proteasas víricas para acabar produciendo 16 proteínas no estructurales. Las otras ORFs codifican para las proteínas estructurales que incluyen proteínas de la espícula, membrana, envuelta y nucleocápside [6]. Así, podemos conocer la secuencia específica de las proteínas que conforman la envuelta del virus, crucial para el proceso de ensamblado y liberación del virus. O de la glicoproteína de la espícula, que sabemos que está compuesta de dos subunidades (S1 y S2) y que se presenta en la partícula vírica en forma de homotrímero y así se une al receptor celular ACE2. Además, mediante el análisis de su secuencia podemos saber que la subunidad S2 contiene un péptido de fusión, un dominio transmembrana y otro citoplasmático y que está altamente conservada por lo que podría ser una importante diana terapéutica. Conocemos también que la subunidad S1 contiene el dominio de unión al receptor y que está mucho menos conservado con respecto a otros coronavirus (identidad de aminoácidos del 40%) [6]. Se conoce también la existencia de otros genes que no presentan homología con ningún otro gen de coronavirus encontrado hasta ahora como la ORF3b o la presencia de otros genes como el de la ORF8 que codifican para una proteína estructuralmente diferente a otras encontradas en SARS-CoV. El estudio en profundidad de estos genes puede resultar clave para entender la patogenicidad del virus [6].

- **Diseño de fármacos antivirales**. El estudio de la secuencia del virus, permite conocer qué posiciones del mismo son clave para infectar células humanas. En este sentido, gracias a la información generada en otros virus relacionados, como el SARS o el MERS, conocemos que los residuos L455, F486, Q493, S494, N501 and Y505 de la proteína de la espícula son clave para la unión al receptor humano ACE2 [7]. También que ciertas variantes raras observadas ya en virus SARS-CoV-2 como V483A, G476S, L455I, F456V y S494P, han sido previamente asociados a una ligera menor afinidad por el receptor, así como a una antigenicidad alterada en posiciones equivalentes en los virus MERS y SARS-CoV [8, 9, 10]. Así como la proteína de la espícula del SARS-CoV-2 es bastante diferente a la del SARS-CoV (identidad del 76%), existen otras proteínas mucho más conservadas y que constituyen otras posibles dianas del virus sobre las que diseñar fármacos antivirales. Estas proteínas son su proteasa y su polimerasa con las que comparte un 96% y un 97%, respectivamente. Por tanto, cualquier fármaco diseñado

frente a estas proteínas del SARS-CoV podrían tener una alta probabilidad de funcionar también frente a SARS-CoV-2.

- **Diseño de vacunas**. Actualmente existen diferentes estrategias para el desarrollo de vacunas frente a SARS-CoV-2. El prototipo chino consiste en integrar en el genoma de otro virus denominado adenovirus, el gen de la espícula del SARS-CoV-2. Con esto conseguiríamos que el sistema inmune de la persona vacunada reaccionara frente a dicha proteína del SARS-CoV-2 y que actuara en el momento que esta persona vacunada se infectara por el virus. Por su parte, el prototipo de EEUU es una molécula de ARN también basada en el gen de la espícula del SARS-CoV-2. Existen otras posibles estrategias que incluyen virus inactivados o atenuados donde el producto vacunal consistiría en la inoculación del virus completo de forma que no sea capaz de desencadenar la enfermedad pero sí una respuesta inmune potente frente al virus. Todas estas estrategias requieren conocer el genoma o partes del mismo para poder empezar con el diseño de la propia vacuna [11].

Madrid, 15 de abril de 2020

Informe realizado por Francisco Diez-Fuertes. Resumen divulgativo: José A. Plaza.

Grupo de Análisis Científico de Coronavirus del Instituto de Salud Carlos III.

Integran este grupo los Drs Mayte Coiras, Francisco Diez, Elena Primo, Cristina Bojo, Beatriz Pérez-Gómez, Francisco David Rodríguez, Esther García-Carpintero, Luis María Sánchez, José A. Plaza y Débora Alvarez. Está coordinado por el Dr José Alcamí.

REFERENCIAS

1.- Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975 Mayo 25;94(3):441–448.

2.- Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res. 1985; 13(7):2399-412.

3.- Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. A new coronavirus associated with human respiratory disease in China. Nature 2020; 579(7798):265-269.

4.- Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature Medicine 2020

5.- Tsan-Yuk Lam T, Ho-Hin Shum M, Zhu HC, Tong YG, Ni XB, Liao YS, Wei W, Cheung WYM, Li WJ, Li LF, Leung GM, Holmes EC, Hu YL, Guan Y. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. Nature 2020.

6.- 11.- Cascella M, Rajnik M, Cuomo A, et al. Features, Evaluation and Treatment Coronavirus (COVID-19) [Updated 2020 Mar 20]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554776/

7.- Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. Journal of Virology 2020;94(7)e00127-20.

8.- Kleine-Weber H, Elzayat MT, Wang L, Graham BS, Müller MA, Drosten C, Pöhlmann S, Hoffmann M. Mutations in the Spike Protein of Middle East Respiratory Syndrome Coronavirus Transmitted in Korea Increase Resistance to Antibody-Mediated Neutralization. Journal of Virology 2019, 93(2) e01381-18.

9.- Wu K, Peng G, Wilken M, Geraghty RJ, Li F. Mechanisms of Host Receptor Adaptation by Severe Acute Respiratory Syndrome Coronavirus. The Journal of Biological Chemistry 2012; 287, 8904-8911.

10.- Rockx B, Donaldson E, Frieman M, Sheahan T, Corti D, Lanzavecchia A, Baric RS. Escape from Human Monoclonal Antibody Neutralization Affects In Vitro and In Vivo Fitness of Severe Acute Respiratory Syndrome Coronavirus, The Journal of Infectious Diseases 2010;201(6):946–955.

11.- Amanat F, Krammer F. SARS-CoV-2 Vaccines: Status Report. Immunity 2020.

IMÁGENES COMPLEMENTARIAS