Secuenciación de genomas de SARS-CoV-2: herramienta clave en esta pandemia

Ángel Zaballos[1], Sarai Varona[2], María Iglesias-Caballero[3], Sara Monzón[2], Francisco Pozo[3], Inmaculada Casas[3], Isabel Cuesta[2]

[1]Unidad de Genómica, Unidades Centrales Científico-Técnicas, Instituto de Salud Carlos III, Majadahonda, Madrid, España. [2]Unidad de Bioinformática, Unidades Centrales Científico-Técnicas, Instituto de Salud Carlos III, Majadahonda, Madrid, España. [3]Unidad de Virus Respiratorios y Gripe. Centro Nacional de Microbiología. Instituto de Salud Carlos III.

Palabras clave: genómica, bioinformática, análisis filogenético, epidemiología molecular.

La pandemia causada por el nuevo coronavirus SARS-CoV-2 no sólo está afectando de manera muy sensible al modo de vida de miles de millones de personas, sino que, a la vez de reto científico y médico, está suponiendo un nuevo paradigma en la detección, caracterización y seguimiento epidemiológico del agente causal. Y sin lugar a duda la secuenciación completa del genoma de SARS-CoV-2 a partir de muestras clínicas es el estandarte de este nuevo paradigma. Con más de 45 millones de casos de infección detectados hasta octubre de 2020 (https://www.worldometers.info/coronavirus/), son más de 170000 los genomas depositados en la base de datos de GISAID en las mismas fechas (https://www.gisaid.org/), casi un 0,4% de todos los casos detectados mundialmente, que en principio puede parecer una cifra minúscula, pero a su vez nunca vista para otro agente infeccioso. Las ventajas de conocer la secuencia completa del genoma del virus que está infectando a individuos concretos hablan por sí mismas. Con todos los elementos genéticos determinados para cada espécimen es posible trazar con precisión los orígenes de los brotes, las cadenas de transmisión, la dispersión del agente y su propia evolución espacio-temporal. Pero no sólo aporta valiosa información epidemiológica, sino que los cambios genéticos en el virus pueden acarrear a su vez cambios fenotípicos con consecuencias en su infectividad, agresividad de la enfermedad causada y en el diseño de fármacos o vacunas.

SECUENCIACIÓN MASIVA DE GENOMAS DE SARS-COV-2

Como en muchas otras áreas de la microbiología, o en la misma genética humana, la secuenciación completa de genomas va camino de instaurarse como técnica de referencia en el ámbito de la identificación y caracterización molecular, impulsada cómo no, por la enorme capacidad de secuenciación de las distintas plataformas de secuenciación masiva actuales y la reducción de los costes por nucleótido, que llega a equiparar el coste de una secuenciación clásica de Sanger, una kilobase, con una gigabase de secuencia obtenida por los nuevos métodos.

La secuenciación del genoma de SARS-CoV-2 no presenta mayor obstáculo debido a su tamaño (unas 30 Kb) sino por el tipo de muestra, exudado nasofaríngeo en la mayoría de los casos, en la que además del RNA viral, cuya proporción dependerá fundamentalmente de la carga viral del individuo afectado, encontraremos ácidos nucleicos del paciente, así como de otros organismos que puedan estar coinfectando o formando parte de la flora respiratoria, y además en proporciones muy variables de cada uno de ellos. De ahí que los protocolos difieran en la introducción o no de algún paso de enriquecimiento en secuencias específicas del virus. En el caso de utilizar el RNA de la muestra sin selección alguna, lo que se suele denominar como aproximación metagenómica, se obtendrán mayoritariamente secuencias de RNAs humanos, siendo las secuencias virales entre un 5 y menos del 0.001 por ciento del total. Esto requiere generar en muchos casos decenas de millones de secuencias por muestra para garantizar una cobertura completa del genoma y, en su caso, la identificación de variantes en baja proporción, lo cual exige equipos con un gran rendimiento de secuenciación, un importante coste en reactivos y también adecuada capacidad bioinformática. Eso sí, esta aproximación viene con bonus extra, ya que se puede caracterizar además el transcriptoma del paciente, parte de sus variantes alélicas, su microbioma respiratorio y la presencia de otros virus (Xiao *et al.*, 2020). Sin embargo, y por las razones de costes antes citadas, las aproximaciones más extendidas son las apriorísticas, en las que, ya sea mediante amplificación específica por PCR o por captura con sondas específicas, se reduce considerablemente el número de secuencias requeridas para obtener el genoma completo del virus. En el caso de la amplificación, las distintas estrategias se basan en el diseño de varios cientos de amplicones solapantes que cubren prácticamente todo el genoma del virus. Una de las primeras, y más exitosas por el volumen de genomas secuenciados, fue la desarrollada por la red ARTIC (https://artic.network/ncov-2019), puesta a disposición de la comunidad científica en los inicios de la pandemia y diseñada en principio para ser usada con los secuenciadores basados en nanoporos (Minlon), aunque pronto se extendió al resto de plataformas (Illumina, Thermofisher). Además del bajo coste, la rapidez en la determinación y en el ensamblado de los genomas, así como el elevado número de muestras que se pueden procesar simultáneamente la han colocado como la estrategia preferida por muchos investigadores. Como desventajas, la sensibilidad a posibles variantes del virus, baja fiabilidad para variantes de baja frecuencia (identificación de cuasiespecies) y la mayor probabilidad de contaminaciones entre muestras. Una situación intermedia entre la estrategia de amplicones y la metagenómica es el uso de sondas diseñadas para capturar específicamente las secuencias virales, que pueden ser para un único genoma o ampliarse a otros virus respiratorios o incluso a múltiples familias. No

requiere un gran número de secuencias para obtener el genoma final, pero es más costoso tanto en reactivos como en manipulación que la secuenciación de amplicones.

ANÁLISIS BIOINFORMÁTICO DE SECUENCIAS DE SARS-COV-2

La bioinformática es una disciplina clave en el análisis de grandes datos derivados de las tecnologías de alto rendimiento, como son las secuencias de genomas generadas por secuenciación masiva. Especialmente en esta pandemia, la bioinformática ha tenido una relevancia muy significativa, generando la metodología y proporcionando las herramientas necesarias para el análisis, procesamiento e interpretación de los datos obtenidos en la secuenciación del genoma viral. Desde que el SARS-CoV-2 fue anunciado por el gobierno de China el 31 de diciembre de 2019, se tardaron apenas 10 días en obtener la primera secuencia completa del genoma del virus, y ser depositada en bases de datos públicas como GeneBank o GISAID (Wu *et al.*, 2020) .

En el momento en que la OMS determinó el 20 de enero de 2020 que el nuevo coronavirus era una emergencia sanitaria pública de importancia mundial, la comunidad bioinformática asumió la responsabilidad de crear protocolos de análisis estandarizados y eficientes, adaptados a las características del genoma del virus SARS-CoV-2. Los primeros protocolos de análisis para el ensamblado del genoma de SARS-CoV-2 fueron los desarrollados por la red de ARTIC y por la comunidad GalaxyProject. La primera, que ya había desarrollado protocolos estandarizados para los virus de la gripe, Ébola y Zika, diseñó con gran rapidez otro capaz de analizar los datos generados en la secuenciación de amplicones de SARS-CoV-2 mediante la tecnología de nanoporos. Sin embargo, el número de muestras que es posible secuenciar con los dispositivos basados en nanoporos es muy inferior al que se puede obtener con los de Illumina, razón por la cual se adaptaron los métodos de análisis a los datos generados por estos secuenciadores, ampliándolo además a otros métodos de enriquecimiento de genoma viral, como es el uso de sondas de captura. Con el esfuerzo colaborativo de cientos de bioinformáticos (https://github.com/virtual-biohackathons/covid-19-bh20), han surgido una gran variedad de protocolos de análisis y de colaboraciones a nivel mundial, resultando herramientas como Viralrecon (https://github.com/nf-core/viralrecon) para reconstruir el genoma viral a partir de los datos de secuenciación masiva.

Dos son las aproximaciones en las que se basan los diferentes protocolos de obtención de la secuencia de los genomas virales, las basadas en genomas de referencia y las de ensamblado *de novo*. La primera consiste en el mapeado de las lecturas de las muestras sobre el genoma del virus SARS-CoV-2 de Wuhan, con la posterior determinación y filtrado de variantes entre ambas secuencias y generación de un genoma consenso que contiene las variantes propias de la muestra analizada (Grubaugh *et al.*, 2019) . En principio, esta aproximación tendría una clara desventaja y es que no permitiría identificar variantes estructurales en el genoma viral que no estuviesen en el genoma de referencia que se use. Sin embargo, el virus de SARS-CoV-2 no parece haber variado lo suficiente desde su aparición como para que esta estrategia no sea lo suficientemente eficaz, razón principal por la que es la más empleada mundialmente. La segunda aproximación, aunque menos empleada, sí que permitiría obtener estas variantes estructurales y consiste en el ensamblado *de novo*, de las lecturas obtenidas del secuenciador, sin usar un genoma de referencia. Aunque ya se disponía de diferentes programas para este tipo de ensamblado, algunos se han optimizado para la reconstrucción del genoma de SARS-CoV-2, como es el caso de CoronaSpades (Meleshko y Korobeynikov 2020). El ensamblado *de novo* sería la opción de análisis a elegir si se hace una aproximación de enriquecimiento mediante sondas y se desaconseja en el caso de amplicones porque las diferencias en profundidad de secuenciación entre los mismos, puede generar artefactos en el ensamblado.

Una vez determinados, los genomas consenso se pueden subir a los repositorios públicos existentes, como GISAID o ENA (https://www.ebi.ac.uk/ena/browser/about), para que estén al alcance de toda la comunidad científica. Ambos repositorios están haciendo esfuerzos para unificar criterios de calidad y análisis, para que la comparación de secuencias revele relaciones filogenéticas robustas que faciliten entender la evolución temporal del virus y determinar las cadenas de transmisión a nivel mundial (Alm *et al.,* 2020). Además, ayudará a conocer las variantes virales circulantes, información necesaria para, mantener las herramientas de diagnóstico viral basadas en PCR efectivas, realizar el seguimiento de una posible vacuna y conocer su eficacia, o identificar cuasiespecies virales con posible incidencia en el desarrollo futuro de la pandemia.

LA CLASIFICACIÓN DEL VIRUS SARS-COV-2 A TRAVÉS DE SU GENOMA

El análisis del genoma completo de patógenos se ha revelado como una importante herramienta en el estudio de la epidemiología molecular de las enfermedades infecciosas. La existencia de plataformas de publicación y de análisis de secuencias, como las ya citadas y otras como NextStrain, han permitido la visualización en tiempo real de las secuencias disponibles de los diferentes países, facilitando el estudio de la distribución del virus y la identificación de mutaciones que pudiesen derivar en posibles adaptaciones al hospedador humano o cambios en las características del virus (van Dorp *et al.,* 2020). Esto ha facilitado la definición de una nomenclatura para los diferentes clados o ramas con las variantes que han aparecido en su desarrollo y expansión. Las principales propuestas para clasificar filogenéticamente las secuencias de SARS-CoV-2 son las de las plataformas NextStrain y GISAID. La primera propone cinco grandes clados. Dos de ellos emergieron ya en 2019: el clado 19A, que se considera el grupo raíz, y el 19B, definido por los cambios C8782T y T28144C. Los tres clados restantes agrupan secuencias de virus circulantes en 2020: el 20A, que se distingue del 19A por las sustituciones C3037T, C14408T y A23403G, el 20B, caracterizado por tres cambios consecutivos G28881A, G28882A y G28883C y el 20C, que presenta las sustituciones C1059T y G25563T. En los dos primeros clados, 19A y 19B, se agrupan las secuencias que circularon durante los primeros meses en Asia, mientras que el clado 20A comprende las secuencias de Europa de comienzos del 2020. Los otros dos clados del 2020 comprenden a las secuencias mayoritarias en Europa (20B) y Norteamérica (20C). La clasificación propuesta por GISAID está basada en la combinación de nueve marcadores genéticos que permite que el 95 %

de las secuencias de SARS-CoV-2 puedan ser clasificadas en seis grupos filogenéticos bien definidos que van desde los dos grupos iniciales S y L, hasta la posterior evolución del clado L en los grupos V y G, quedando este último finalmente dividido en los clados GH y GR. Estos nombres refieren a mutaciones que sirven para describir el grupo. Por ejemplo, el cambio D614G en la espícula caracterizó al grupo descrito como clado G. La unificación de criterios a la hora de nombrar los diferentes clados es una tarea pendiente, estando todas estas propuestas aún en evaluación, ya que, por ejemplo, actualmente ninguna de estas nomenclaturas refleja alguna propiedad fenotípica del virus, como pudieran ser variantes antigénicas, si bien el virus es antigénicamente similar hasta el momento.

La distribución de los diferentes clados en los países que conforman la región Europea de la OMS es muy variada, destacando en España la alta proporción de clados 19B/S y 20A/G. Este hecho puede deberse a un efecto fundador o a sesgos muestrales (Díez-Fuertes *et al.,* 2020). pero también se ha podido condicionar por la duración de las restricciones de viaje y las diferentes medidas implementadas en cada país. En el caso de restricciones de viaje tempranas probablemente se vería reducida la incidencia temprana y por lo tanto los clados más prevalentes serían los 19A/L/V/O frente a los clados 20A/G que posteriormente se convirtieron en los clados dominantes (Alm E *et al.*, 2020).

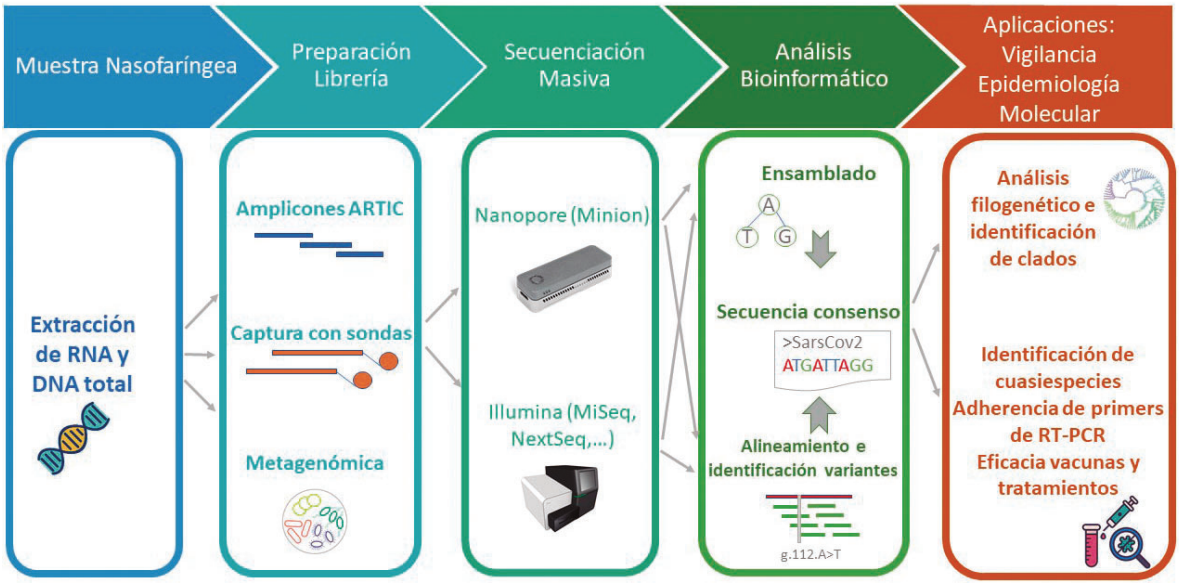

Figura 1. Esquema que resume el uso de la secuenciación del genoma completo en la vigilancia e investigación del SARS-CoV-2.

Esta pandemia ha puesto de manifiesto la importancia de la secuenciación genómica en dos aspectos clave en enfermedades infecciosas, la identificación rápida del patógeno que causa la infección, y su caracterización, seguimiento y evolución que faciliten el control de la infección. El estado de desarrollo de la secuenciación y de su metodología de análisis en las unidades de genómica y bioinformática de los centros de investigación y sanitarios ha facilitado enormemente su aplicación durante esta pandemia, aunque es necesario llegar a protocolos estandarizados y armonizados que permitan la comparación de datos de una forma fidedigna. La vigilancia epidemiológica molecular se ha consolidado como una herramienta necesaria y decisiva que debe de funcionar de forma unificada a nivel mundial para poder controlar las enfermedades infecciosas.

REFERENCIAS:

Alm E *et al.* (2020). Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Eurosurveillance 25: 32.doi: https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001410.

Díez-Fuertes F *et al.* (2020). A founder effect led early SARS-CoV-2 transmission in Spain. J. Virol. JVI.01583-20 doi: 10.1128/JVI.01583-20.

Grubaugh ND *et al.* (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol. 20: 8 doi: 10.1186/s13059-018-1618-7.

Meleshko D y Korobeynikov A (2020). CoronaSPAdes: from biosynthetic gene clusters to coronaviral assemblies," bioRxiv, p. 2020.07.28.224584, doi: 10.1101/2020.07.28.224584.

van Dorp L *et al.* (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. 83: 104351. doi: 10.1016/j.meegid.2020.104351.

Wu F *et al.* (2020). A new coronavirus associated with human respiratory disease in China. Nature 579: 265–269. doi: 10.1038/s41586-020-2008-3.

Xiao M *et al.* (2020). Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. Genome Med. 12: 57 doi: 10.1186/s13073-020-00751-4.