PERSPECTIVAS FECHA DE PUBLICACIÓN: 26/2/2021

GISAID: INICIATIVA INTERNACIONAL PARA COMPARTIR DATOS GENÓMICOS DEL VIRUS DE LA GRIPE Y DEL SARS-CoV-2

Marta Hernández
Laboratorio de Microbiología y Biología Molecular. Instituto Tecnológico Agrario de Castilla y León. Valladolid. España.

Emilio García-Morán
Servicio de Cardiología. Hospital Clínico Universitario de Valladolid. Valladolid. España.

David Abad
Laboratorio de Microbiología y Biología Molecular. Instituto Tecnológico Agrario de Castilla y León. Valladolid. España.

José María Eiros
Servicio de Microbiología y Parasitología. Hospital Universitario del Río Hortega. Valladolid. España.

La aparición de amenazas de enfermedades infecciosas que comprometen la salud y la economía mundial, requiere de iniciativas globales que protejan el interés común. Con este objetivo aparecen las plataformas de almacenamiento genómico que permiten compartir datos para garantizar o mejorar la Salud de la población. Así surge GISAID con un primer objetivo de compartir secuencias genómicas de virus gripales que permiten seguir la evolución de los mismos y predecir el diseño de vacunas para anticiparse a la epidemia del año venidero. En este artículo se revisa el concepto de la iniciativa, y su aportación a la actual pandemia de SARS-CoV-2 con más de 230.000 genomas depositados en su repositorio (2 diciembre 2020).

GISAID: CONCEPTO

Una carta publicada en agosto de 2006 en la revista *Nature*[1] motivó la creación en 2008 de la iniciativa GISAID (*Global Initiative on Sharing All Influenza Data*) que permite la disponibilidad inmediata de genomas de los virus gripales (influenza virus)[2]. La información más compleja y específica de los virus es su secuencia genética, es decir las letras representativas de los nucleótidos que les permite su replicación y desarrollo. En el caso de los virus gripales y del SARS-CoV-2 se trata de ARN, ácido ribonucleico, en contraposición al ADN que es el material heredable en humanos y otros organismos complejos, así como en bacterias y otros virus. La secuencia completa del material genético de un virus se denomina genoma o borrador de genoma cuando este es incompleto. Esta información sirve para comprender la evolución de un brote epidémico. Los investigadores, incluida la OMS y la industria, pueden anticipar cada año el desarrollo de vacunas como la de la gripe. A 30 de mayo de 2020 GISAID albergaba 320.542 genomas de gripe mientras que a 2 de diciembre se dispone de 333.059 secuencias (incremento del 3,9%), siendo el registro más antiguo, un genoma del virus influenza del 7 de noviembre de 1985. Esta infraestructura de alojamiento de datos genéticos del virus de la gripe ha permitido una

respuesta rápida ante la crisis desatada por el SARS-CoV-2. Desde el pasado 10 de enero GISAID comenzó a ser también un repositorio genómico del virus SARS-CoV-2, que hasta el 30 de mayo alojaba 35.249 genomas y el 2 de diciembre existen 234.698 secuencias de SARS-CoV-2 depositadas, un incremento de más del 500% en 6 meses.

QUIÉN SON Y QUÉ DATOS OFRECE

La iniciativa GISAID recibe apoyo administrativo de *Freunde von GISAID e.V*. una asociación registrada sin fines de lucro en Munich (Alemania), organizada y operada exclusivamente con fines benéficos, científicos y educativos. La web está alojada por el Ministerio Federal de Agricultura y Alimentación, del gobierno federal alemán y recibe apoyo público-privado de los CDC norteamericanos (*Centers for Disease Control and Prevention*), el gobierno de Singapur a través de su *Agency for Science, Technology and Research* (*A*STAR*), la Fundación Sanofi Pasteur y la compañía SEQIRUS. Además, la OMS a través de sus Centros de Vigilancia GISRS, (*Global Influenza Surveillance and Response Systems*) localizados en 115 países y los laboratorios de referencia de la OIE/FAO (*World Organisation for Animal Health and Food and Agriculture Organization*) proporcionan supervisión científica en el desarrollo de esta plataforma. Para tener acceso a los datos es preciso registrarse y aceptar las condiciones de uso, que entre otros requerimientos obliga a agradecer el empleo de sus datos y limita el mostrarlos o compartirlos fuera de esta comunidad. Hay 21.533 usuarios registrados en GISAID en el momento actual (mayo 2020), que a título ilustrativo proceden 376 de España, 5.751 de Estados Unidos, 3.165 China, 1.096 el Reino Unido y 720 Alemania.

Los genomas y los datos se suben a esta web por investigadores de cualquier país del mundo. Para cada genoma se incluye además de la secuencia genética, otros datos clínicos y epidemiológicos de la muestra. GISAID provee el acceso abierto gratuito a los datos de este repositorio que incluyen el nombre científico del virus, el pase de cultivo en el que se secuenció (ya que pueden acumular mutaciones cuando se somete a crecimiento en cultivo celular), la fecha de la toma de la muestra, el lugar geográfico y físico de obtención de la muestra (elementos biológicos, entornos naturales como una cueva, etc.), la especie hospedadora (humana o animal), en el caso de la especie humana el género, edad, y estado clínico del individuo (hospitalizado, vivo, fallecido), el tipo de muestra de la que se aisló (frotis nasofaríngeo, esputo, lavado traqueobronquial, orina, heces, etc.) y datos del brote. Además, se incluyen datos técnicos sobre los modelos de los equipos utilizados en la secuenciación, la *depth of coverage*, es decir el número de veces que una determinada posición nucleotídica es secuenciada, y por tanto permite un ligero filtro por alta o baja cobertura.

Existen dos herramientas de búsqueda en la web de GISAID: Epiflu™ y EpiCoV™, que permiten la búsqueda multifactorial de virus de la gripe y SARS-CoV-2, respectivamente. Es posible rastrear y seleccionar genomas por fecha o lugar de aislamiento, país, subtipo, hospedador, laboratorio que deposita los datos, etc., e incluso una vez seleccionados se puede utilizar la herramienta *blast* (*Basic Local Alignment Search Tool)* y obtener un alineamiento múltiple de secuencias (*multiple sequence alignment,* msa).

GISAID es fundamentalmente una base de datos, pero convertir estos datos en información requiere un análisis bioinformático. Existe una íntima alineación de los datos

de GISAID con varios portales que también han desarrollado una rápida respuesta a la emergencia SARS-CoV-2. Un ejemplo es Nexstrain[4] que tiene como objetivo proporcionar una instantánea en tiempo real de la evolución de las poblaciones de patógenos y proporcionar visualizaciones interactivas. Permite observar conjuntos de datos actualizados frecuentemente, proporcionando una herramienta de vigilancia novedosa para las comunidades científicas y de salud pública. Nexstrain es un proyecto de código abierto escrito en Javascript y Python, que utiliza como repositorios principales datos genómicos de NCBI[4], GISAID[2] y ViPR[5].

PARA QUÉ SIRVE EL CONTENIDO DE GISAID

GISAID es una base de datos de considerable tamaño, dinámica y compleja, que permite a los usuarios registrados descargar los datos, analizarlos y elaborar literatura científica. Esto requiere conocimientos y herramientas de bioinformática, estadística y visualización de datos. Otra posibilidad interesante son los proyectos que hacen uso de la información de GISAID como son los portales Nexstrain y el Centro Nacional de Bioinformación de China[6]; ellos analizan y realizan una representación gráfica de la información genómica en árboles filogenéticos, dendrogramas interactivos elaborados a partir de los genomas anotados que permiten interpretar la evolución de las mutaciones. Nextstrain también integra información de otros virus como Zika, Dengue, Ébola, Gripe, Fiebre de Lassa, Virus del Nilo Occidental, Sarampión, Enterovirus y ahora Coronavirus[7]. Se indica por colores la región geográfica de procedencia o el laboratorio de envío del genoma, las mutaciones en epítopos o en los receptores de unión a la célula eucariota, la frecuencia de mutación y datos como la fecha de muestreo. Estos árboles con ramas se pueden ampliar o colorear según se desee consultar, por ejemplo, la procedencia, o un determinado genotipo que indique una mutación en un aminoácido. Resulta factible buscar mutaciones que definan un grupo de interés (clado), y luego se puede definir clados por cambios de aminoácidos o de nucleótidos, que ocurren con mayor frecuencia, y mediante su anclaje es posible redefinir el árbol con otra raíz. También se representa la variabilidad del alineamiento en función del tiempo y la frecuencia de la sustitución de un determinado aminoácido a lo largo del tiempo. Posibilita así mismo rastrear y reconstruir el árbol para una determinada mutación y mostrar esta información como un gráfico de barras de entropía en cada posición en el genoma. La selección de una posición en el genoma con entropía distinta de cero revela la distribución de la variante segregadora en la filogenia y en el mapa. Esto permite interrogar el cambio genético que puede ser adaptativo o subyacente a un cambio en la dinámica de la enfermedad. Para ello es importante el sesgo de muestreo, ya que la falta de datos puede oscurecer los enlaces de transmisión. Para muchos patógenos, la aparición y propagación de variantes de ganancia de función es una preocupación grave y resulta necesaria la monitorización continua de tales mutaciones putativamente adaptativas. Por ejemplo, las mutaciones identificadas por de Vries *et al* (2017) en el ámbito de los virus de la gripe aviar son fácilmente visibles en nextstrain.org/avian/h7n9.

Un aspecto importante de la base de datos de GISAID que cabe mencionar, es que no prefiltra las secuencias depositadas de cualquier genoma y se basa en una acuerdo y compromiso del autor que envía los datos, siendo muy fácil realizar el envío del genoma. Incluso algunas plataformas comerciales

como las de la compañía Illumina permiten la secuenciación y ensamblado del genoma con un pipeline disponible en su web *BaseSpace* (*Illumina SARS-COV-2 NGS Data Toolkit*) que incluye el depósito directo del genoma mediante *GISAID Submission App*. Por ello cabe que se encontren genomas cuya calidad de secuencia es deficiente en un alto porcentaje de posiciones del genoma, y son indicadas por la letra N, como posición de contenido indeterminado. En nuestra experiencia de consulta de las 29.903 pares de bases nucleotídicas del SARS-CoV-2 nos hemos encontrado genomas con hasta un 5% de Ns en su secuencia. Si bien es verdad que cualquier dato bien interpretado puede ser útil, hay que prestar atención a los genomas con significativos segmentos indeterminados, estos genomas solo serían válidos para interpretar determinados segmentos concretos del genoma. La variabilidad que indican las mutaciones dentro de cada genoma puede estar distribuida de forma muy desigual a lo largo de la secuencia, aquellas posiciones (loci) que menos varían, es decir, más conservadas, pueden indicar segmentos implicados en funciones biológicas clave para el virus. Además, las mutaciones permiten establecer linajes de cepas que ilustran la cadena de transmisión o se correlacionan con aspectos clínicos, o permiten la identificación de brotes y reinfecciones.

EL GENOMA DE SARS-COV-2

El SARS-CoV-2, como cualquier otro virus, produce mutaciones cada vez que se multiplica dentro de una célula, pero presenta una estabilidad de secuencia superior a otros virus como el de la gripe debido a que dispone de un mecanismo intrínseco de corrección de errores durante la replicación del virus. Se trata de una proteína codificada en la orf1ab denominada nsp14 (ExoN) con actividad 3′–5′ exonucleasa que mantiene la estabilidad del genoma vírico y permite la escisión de errores, por lo que los coronavirus están acumulando mutaciones mucho más lentamente que otros virus ARN. Aun así, la variación genética referida a genomas completos, situada en el espacio y en el tiempo, permite analizar la cadena de transmisión de los aislados. Su genoma consta de 29.903 pares de bases, en cuyos extremos existen dos regiones UTR (*untranslated regions*), y en medio hay 10 ORF (*open reading frame*, segmentos codificantes) que se traducen en 25 proteínas. Dos tercios de todo el genoma lo ocupan las ORF1a y ORF1b que forman dos polipéptidos pp1a (nsp1-11) y pp1ab (nsp1-10, nsp12-16) procesados por dos cistein proteasas codificadas en nsp3 (papain-like proteasa; PLpro) y nsp5 (quimiotripsina-like protease) también conocida como 3CLpro. El otro tercio de genoma codifica las 4 proteínas estructurales, la ORF2 codifica la proteina espicular (S), la ORF4 codifica las proteínas de envoltura (E), la ORF5 la de membrana (M) y la ORF9 la nucleocápside (N). Los datos de GISAID permiten análisis filogenéticos periódicos de los genomas de modo que en un primer momento se establecieron se establecieron como dos cepas denominadas S y L, después en los primeros meses del año se definieron 3 grandes clados o variantes mayoritarios en SARS-CoV-2 en función de 3 mutaciones. Esto es una situación dinámica y el 26 de mayo, aparece un nuevo clado y dos subclados, se hicieron 6 agrupaciones filogenéticas, la L se dividió en V y G, y más tarde la G en GH y GR: L, 2.390 genomas; S, 2.367 genomas; G, 6.723 genomas; GR, 7.497 genomas; GH, 6.581 genomas; V, 2.374 genomas; y 1.372 genomas pertenecientes a otros clados. Hoy se han establecido 8 clados (S, L, V, G, GH, GR, GV y O) que incluyen una serie de mutaciones, hasta 31 recoge GISAID. La G incluye la variante D614G que es la

más prevalente en el mundo y el clado GV incluye la "nueva" variante europea A222V.

En conclusión, la disponibilidad de un número elevado de genomas de patógenos como el SARS-CoV-2 o la gripe, a través de una iniciativa como GISAID, permite ver fenómenos en la transmisión del virus, así como identificar elementos importantes en su secuencia genética. Esto último aporta información que repercutirá positivamente en el correcto enfoque de modelos vacunales (mutagénesis dirigida o ensamblaje de fragmentos genómicos) y la investigación de dianas terapéuticas, que puede hacerse incluso previamente a la investigación biológica, para cribado de múltiples candidatos por características estructurales codificadas (*molecular docking*).

La aportación de este tipo de iniciativas a la Salud Pública y a la Medicina Clínica es clave. Las nuevas aplicaciones de secuenciación masiva que generan gran cantidad de datos de forma económica, conllevan un cambio en el paradigma del análisis microbiológico pasando de un ámbito laboratorial a uno computacional[8]. Sin embargo, es imparable su aplicación en la rutina del diagnóstico puesto que ofrecen una información de carácter epidemiológico (tipado), clínico (resistencia a antimicrobianos, genes de virulencia), trazabilidad (mapas geográficos de origen de brotes), detección de reinfecciones, evolución de virulencia, etc. Esta web permite de forma gratuita de momento el depósito y acceso a las secuencias genómicas y no infiere un coste adicional al procesado, permite un retorno invaluable ya que dimensiona el estudio de poblaciones a un ámbito mundial y no meramente local en vistas a la evidente globalización que afecta a la Salud del conjunto de la humanidad. Además, iniciativas como esta genera costes que de momento están siendo asumidos por los gobiernos que los promueven para el mantenimiento de los servicios web y el personal adscrito es financiado por las instituciones científicas de las cuales dependen, pero aporta una gran ventaja para el resto de la comunidad científica y para los profesionales clínicos y del ámbito de la salud pública, y un ejemplo impecable de colaboración.

BIBLIOGRAFÍA

1. Boosting access to disease data. Nature 442, 957 (2006). https://doi.org/10.1038/442957a

2. Global Initiative on Sharing All Influenza Data (consultado el 30/05/2020, require registrarse). Disponible en: www.gisaid.org

3. Real-time tracking of pathogen evolution (consultado el 30/05/2020). Disponible en: https://nextstrain.org

4. National Center for Biotechnology Information (consultado el 30/05/2020). Disponible en: www.ncbi.nlm.nih.gov

5. Virus Pathogen Resource (consultado el 30/05/2020). Disponible en: www.viprbrc.org

6. China National Center for Bioinformation (consultado el 30/05/2020). Disponible en: https://bigd.big.ac.cn/ncov

7. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. Nextstrain: real-time tracking of pathogen evolution, Bioinformatics, 34, 23(2018), pp 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

8. Hernández M, Quijada NM, Rodríguez-Lázaro D, Eiros JM. Aplicación de la secuenciación masiva y la bioinformática al diagnóstico microbiológico clínico. Revista Argentina de Microbiología. https://doi.org/10.1016/j.ram.2019.06.003 https://www.sciencedirect.com/science/article/pii/S0325754119300811