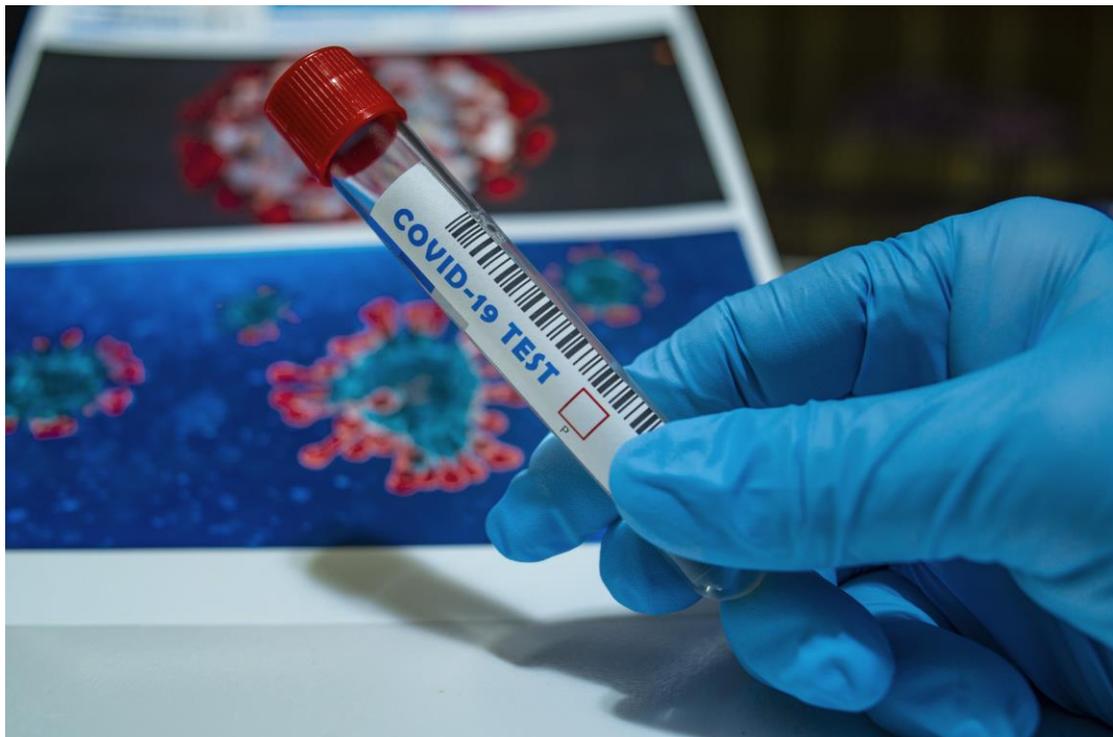


València, 2 de abril de 2025

Hallan errores en las secuencias genéticas del coronavirus incluidas en la mayor base de datos mundial

- Un estudio liderado por el CSIC descubre ‘artefactos’ en las secuencias con reparación de las deleciones del virus que provoca la COVID-19, lo que afecta a la infección y a la respuesta a vacunas
- Muchas de las secuencias con reparación de mutaciones en la proteína spike del virus, la llave para infectar las células humanas, se debieron a errores en el procesamiento de datos



Tras realizar un minado en la base de datos del SARS-CoV-2 más usada en la pandemia descubrieron errores./ Pixabay

Un equipo multidisciplinar liderado por el Instituto de Biología Integrativa de Sistemas (I2SysBio), centro mixto del Consejo Superior de Investigaciones Científicas (CSIC) y la Universitat de València (UV), acaba de publicar un estudio que descubre una nueva perspectiva sobre la capacidad del virus SARS-CoV-2 para mutar e infectar a los humanos. Mediante una revisión de la base de datos genéticos del virus más utilizada durante la pandemia, el equipo de investigación halló ‘falsos positivos’ en su capacidad

de reparación de las deleciones, un proceso que restaura secciones del genoma viral que afecta la capacidad del virus para replicarse o evadir el sistema inmunológico del huésped. En el trabajo, publicado en la revista *Virus Evolution*, participan investigadores del Instituto de Biomedicina de Valencia (IBV) del CSIC y del Instituto de Investigación Sanitaria La Fe (IIS-La Fe).

El trabajo liderado por el grupo de Patogenómica del I2SysBio en colaboración con el grupo de Biología Viral del mismo instituto ofrece una perspectiva innovadora sobre ciertos cambios genéticos raros en la proteína *spike* del SARS-CoV-2, la 'llave' que utiliza el coronavirus para infectar nuestras células. La investigación se centró en los llamados eventos de reparación de deleciones en esta proteína, en los que el virus aparenta corregir su genoma.

Tras realizar un minado de datos masivo en la base de datos de genomas del virus SARS-CoV-2 más usada en la pandemia, llamada GISAID, descubrieron que varios de los hallazgos iniciales se debían probablemente a errores introducidos por el procesamiento de datos en grandes bases de datos genéticas. Los métodos informáticos empleados para analizar millones de secuencias virales pueden inducir equivocaciones, generando la impresión de que el virus repara sus deleciones con mayor regularidad. Al comparar estos datos ya procesados con la información obtenida directamente de la secuenciación de genomas (lecturas de secuenciación), el equipo ha logrado obtener una visión más realista de los cambios genéticos que sufre el virus.

Menos del 60% de eventos de reparación confirmados

“Utilizando el repositorio de secuencias genéticas GISAID estimamos una frecuencia muy alta de estos eventos de reparación de deleciones que se espera que sean raros”, explica **Mireia Coscollá Devís**, investigadora del CSIC que lidera el estudio. “Nos dimos cuenta de que las secuencias de la base de datos de GISAID están procesadas por cada laboratorio de forma diferente y contenía muchos falsos positivos para este tipo de marcadores. Así, aunque en determinados casos pudimos confirmar que se trataba de un fenómeno real, en la mayoría eran consecuencia del procesamiento de las secuencias”, revela.

Así, “vimos que menos del 60 por ciento de los eventos de reparación de deleciones se podía confirmar. Aunque no hemos podido cuantificarlo exactamente para todos, podemos comprar las proporciones del marcador en varias bases de datos, y vemos que la diferencia es de 5 a 51 veces menos frecuente de lo que aparecía en las bases de datos procesadas”, calcula la investigadora del CSIC.

Aunque estos eventos de reparación son poco comunes, el estudio evidencia que, cuando ocurren, pueden afectar de forma sutil al comportamiento del virus. “Por ejemplo, ciertas reparaciones pueden modificar la forma en que el virus ingresa a las células o influir en la respuesta a los anticuerpos generados por la vacunación”, asegura Coscollá, algo que el equipo de investigación demostró mediante experimentos *in vitro*.

Intercambio de datos genómicos de patógenos

Así, “nuestra investigación pone de relieve la importancia de examinar detenidamente los datos genéticos para evitar conclusiones erróneas”, remarca la investigadora del CSIC. La Organización Mundial de la Salud (OMS) recomienda una política de intercambio de datos genómicos de patógenos para proteger la salud pública. Sin embargo, en España no existe una recopilación central de datos de secuencias de patógenos humanos, animales y ambientales, como tampoco existe una política para el intercambio de datos anonimizados entre instituciones de salud y científicas. Esto dificulta el seguimiento y la respuesta a las enfermedades infecciosas incluyendo el seguimiento de la resistencia a los antimicrobianos, destacan los investigadores.

El trabajo ha sido financiado por el Ministerio de Ciencia, Innovación y Universidades y por la Unión Europea con fondos NextGenerationEU/PRTR a través de la PTI+ Salud Global del CSIC. Además, está apoyado por la Generalitat Valenciana y el Fondo Social Europeo a través de la ayuda CIACIF/2022/333. El trabajo computacional se realizó en Garnatxa, el clúster de computación de alto rendimiento (HPC) del Instituto de Biología Integrativa de Sistemas.

Referencia:

Miguel Álvarez-Herrera, Paula Ruiz-Rodríguez, Beatriz Navarro-Domínguez, Joao Zulaica, Brayan Grau, María Alma Bracho, Manuel Guerreiro, Cristóbal Aguilar-Gallardo, Fernando González-Candelas, Iñaki Comas, Ron Geller, Mireia Coscollá, ***Genome data artifacts and functional studies of deletion repair in the BA.1 SARS-CoV-2 spike protein***, *Virus Evolution*, 2025; <https://doi.org/10.1093/ve/veaf015>