Extraña similitud de las inserciones únicas en la proteína de punta 2019-nCoV con el VIH-1 gp120 y Gag

Prashant Pradhan[$1,2], Ashutosh Kumar Pandey[$1], Akhilesh Mishra[$1], Parul Gupta[1], Praveen Kumar Tripathi[1], Manoj Balakrishnan Menon[1], James Gomes[1], Perumal Vivekanandan*[1]and Bishwajit Kundu*[1]

[1]Kusuma School of biological sciences, Indian institute of technology, New Delhi-110016, India.
[2]Acharya Narendra Dev College, University of Delhi, New Delhi-110019, India
[$]Equal contribution
* Corresponding authors- email: bkundu@bioschool.iitd.ac.in
vperumal@bioschool.iitd.ac.in

Resumen:
Actualmente estamos siendo testigos de una gran epidemia causada por el nuevo coronavirus 2019 (2019- nCoV). La evolución del 2019- nCoV sigue siendo esquiva. Encontramos 4 inserciones en la glicoproteína de pico (S) que son únicas del 2019-nCoV y no están presentes en otros coronavirus. Es importante que los residuos de aminoácidos en las 4 inserciones tengan identidad o similitud con los del VIH-1 gp120 o VIH-1 Gag. Curiosamente, a pesar de que las plantillas son discontinuas en la secuencia de aminoácidos primarios, el modelo tridimensional del 2019-nCoV sugiere que convergen para constituir el sitio de unión del receptor. El hallazgo de 4 inserciones únicas en el 2019-nCoV, todas las cuales tienen identidad/similitud a los residuos de aminoácidos en las proteínas estructurales clave del VIH-1 es poco probable que sea de naturaleza fortuita. Este trabajo proporciona conocimientos aún desconocidos sobre el 2019-nCoV y arroja luz sobre la evolución y la patogenicidad de este virus con importantes implicaciones para el diagnóstico del mismo.

Introducción

Los Coronavirus (CoV) son virus de ARN de sentido positivo de una sola cadena que infectan a animales y humanos. Se clasifican en 4 géneros basados en la especificidad de su huésped: Alfacoronavirus, Betacoronavirus, Deltacoronavirus y Gammacoronavirus (Snijder et al., 2006). Hay siete tipos conocidos de CoV que incluyen el 229E y el NL63 (Genus Alphacoronavirus), el OC43, el HKU1, el MERS y el SARS (Genus Betacoronavirus). Mientras que la 229E, la NL63, la OC43 y la HKU1 infectan comúnmente a los seres humanos, el brote de SARS y MERS en 2002 y 2012 respectivamente se produjo cuando el virus pasó de los animales a los seres humanos causando una importante mortalidad (J. Chan et al., s.f.; J. F. W. Chan et al., 2015). En diciembre de 2019, se informó de otro brote de coronavirus en Wuhan (China) que también se transmitió de animales a humanos. La Organización Mundial de la Salud (OMS) ha denominado temporalmente a este nuevo virus como Coronavirus 2019-novel (2019-nCoV) (J. F.-W. Chan et al., 2020; Zhu et al., 2020). Aunque hay varias hipótesis sobre el origen de 2019-nCoV, la fuente de este brote en curso sigue siendo esquiva. Las pautas de transmisión de 2019-nCoV son similares a las pautas de transmisión documentadas en los brotes anteriores, incluso por contacto corporal o por aerosol con personas infectadas con el virus.

Se han recibido informes de casos de enfermedades leves a graves y de muertes por la infección en Wuhan. Este brote se ha extendido rápidamente a naciones distantes como Francia, Australia y Estados Unidos, entre otras. El número de casos dentro y fuera de China está aumentando considerablemente. Nuestra comprensión actual se limita a las secuencias del genoma del virus y a modestos datos epidemiológicos y clínicos. Un análisis exhaustivo de las secuencias de nCoV disponibles en 2019 puede proporcionar importantes pistas que pueden ayudar a avanzar en nuestro conocimiento actual para gestionar el brote en curso.

La glicoproteína de espiga (S) del cornonavirus se divide en dos subunidades (S1 y S2). La subunidad

S1 ayuda en la unión de los receptores y la subunidad S2 facilita la fusión de las membranas (Bosch et al., 2003; Li, 2016). Las glicoproteínas de pico de los coronovirus son determinantes importantes del tropismo tisular y del tipo de huésped. Además, las glicoproteínas de pico son objetivos críticos para el desarrollo de vacunas (Du et al., 2013). Por este motivo, las proteínas de pico representan el elemento más estudiado entre los coronavirus. Por lo tanto, hemos tratado de investigar la glicoproteína pico de la 2019-nCoV para comprender su evolución, la secuencia de características novedosas y las características estructurales utilizando herramientas computacionales.

Metodología

Recuperación y alineación de secuencias de ácido nucleico y proteínas

Recuperamos todas las secuencias de coronavirus disponibles (n=55) de la base de datos de genoma viral del NCBI (https://www.ncbi.nlm.nih.gov/) y utilizamos el GISAID (Elbe & Buckland-Merrett, 2017)[https://www.gisaid.org/] para recuperar todas las secuencias completas disponibles (n=28) de 2019-nCoV al 27 de enero de 2020. La alineación de secuencias múltiples de todos los genomas de coronavirus se realizó utilizando el software MUSCLE (Edgar, 2004) basado en el método de unión de de pares. De los 55 genomas de coronavirus, 32 genomas representativos de todas las categorías se utilizaron para el desarrollo de árboles filogenéticos mediante el programa informático MEGAX (Kumar et al., 2018). Se encontró que el pariente más cercano del SARS era el CoV. La región de la glicoproteína del CoV, del SARS y del 2019-nCoV se alinearon y visualizaron utilizando el software Multalin (Corpet, 1988). La secuencia de aminoácidos y nucleótidos identificada se alineó con la base de datos de todo el genoma viral utilizando BLASTp y BLASTn. La conservación de los patrones de los nucleótidos y aminoácidos en 28 variantes clínicas del genoma de 2019-nCoV se presentaron realizando una alineación de secuencias múltiples utilizando el software MEGAX. La estructura tridimensional de la glicoproteína 2019-nCoV se generó utilizando el servidor en línea SWISS-MODEL (Biasini et al., 2014) y la estructura se marcó y visualizó utilizando PyMol (DeLano, 2002).

Resultados

Extraña similitud de las nuevas inserciones en la proteína de punta 2019-nCoV con el VIH-1 gp120 y Gag

Nuestro árbol filogenético de todos los coronavirus sugiere que 2019-nCoV está estrechamente relacionado con el CoV del SARS [Fig1]. Además, otros estudios recientes han vinculado el 2019-nCoV con el CoV del SARS. Por lo tanto, comparamos las secuencias de glicoproteínas de pico del 2019-nCoV con el CoV del SARS (número de adhesión del NCBI: AY390556.1). Tras un examen cuidadoso de la alineación de las secuencias encontramos que la glicoproteína de punta de 2019-nCoV contiene 4 inserciones [Fig.2]. Para investigar más a fondo si estas inserciones están presentes en cualquier otro coronavirus, realizamos una alineación de secuencias múltiples de las secuencias de aminoácidos de la glicoproteína de pico de todos los coronavirus disponibles (n=55) [véase la Tabla S.File1] en NCBI refseq (ncbi.nlm.nih.gov) esto incluye una secuencia de 2019-nCoV[Fig.S1]. Encontramos que estas 4 inserciones [inserciones 1, 2, 3 y 4] son únicas para 2019-nCoV y no están presentes en otros coronavirus analizados. Otro grupo de China había documentado tres inserciones que comparaban menos secuencias de glicoproteínas de pico de los coronavirus. Otro grupo de China había documentado tres inserciones que comparaban menos secuencias de glicoproteínas de pico de los coronavirus (Zhou et al., 2020).

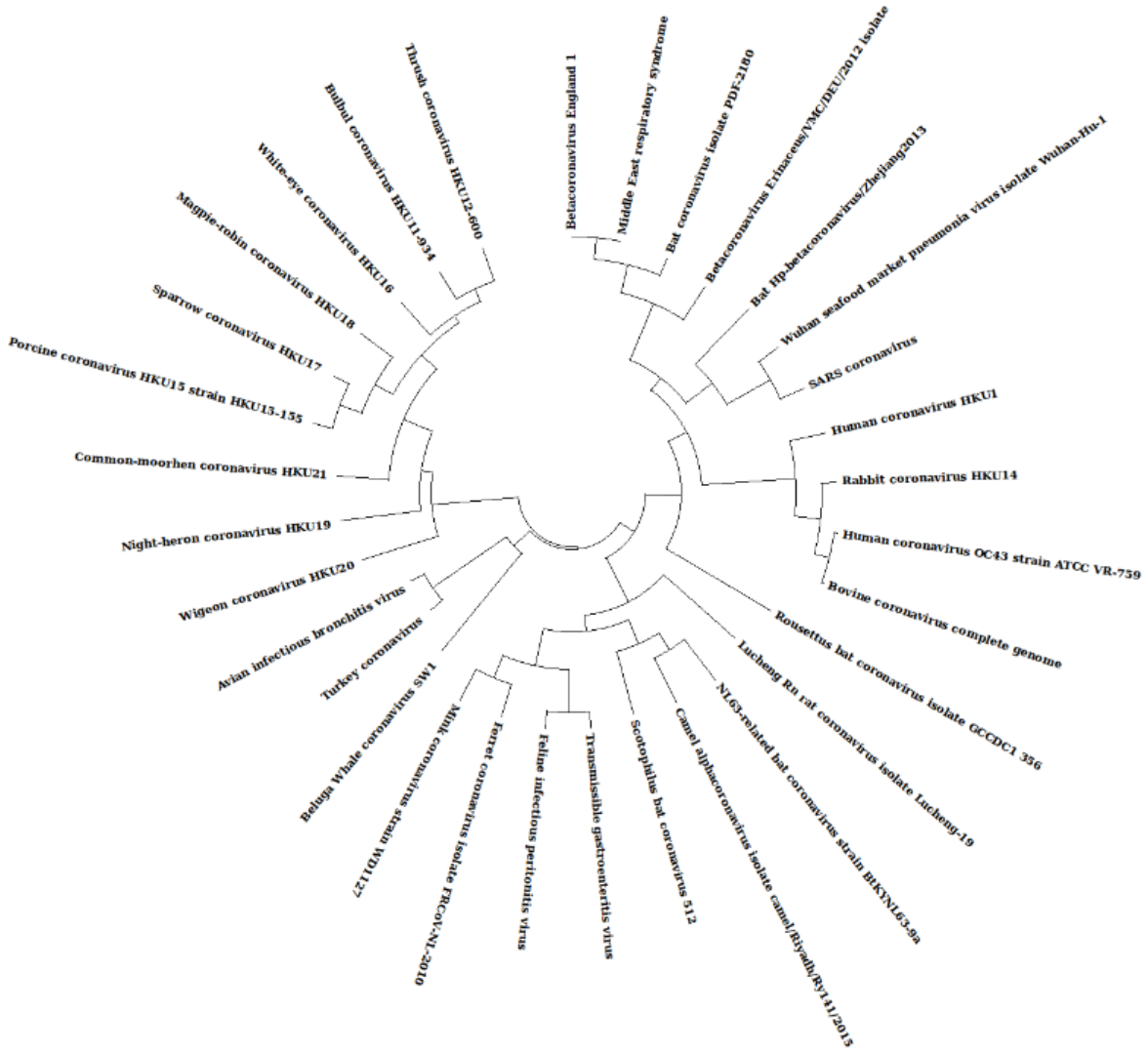


Figura 1: La genealogía de máxima verosimilitud muestra la evolución de 2019- nCoV: La historia evolutiva fue inferida usando el método de máxima verosimilitud y el modelo basado en la matriz de JTT. Se muestra el árbol con la mayor probabilidad logarítmica (12458.88). Los árboles iniciales para la búsqueda heurística se obtuvieron automáticamente aplicando los algoritmos Neighbor-Join y BioNJ a una matriz de distancias en pares estimadas usando un modelo JTT, y luego seleccionando la topología con valor de logaritmo de verosimilitud superior. Este análisis implicó 5 secuencias de aminoácidos. Había un total de 1387 posiciones en el conjunto de datos final. Los análisis evolutivos se realizaron en MEGA X.

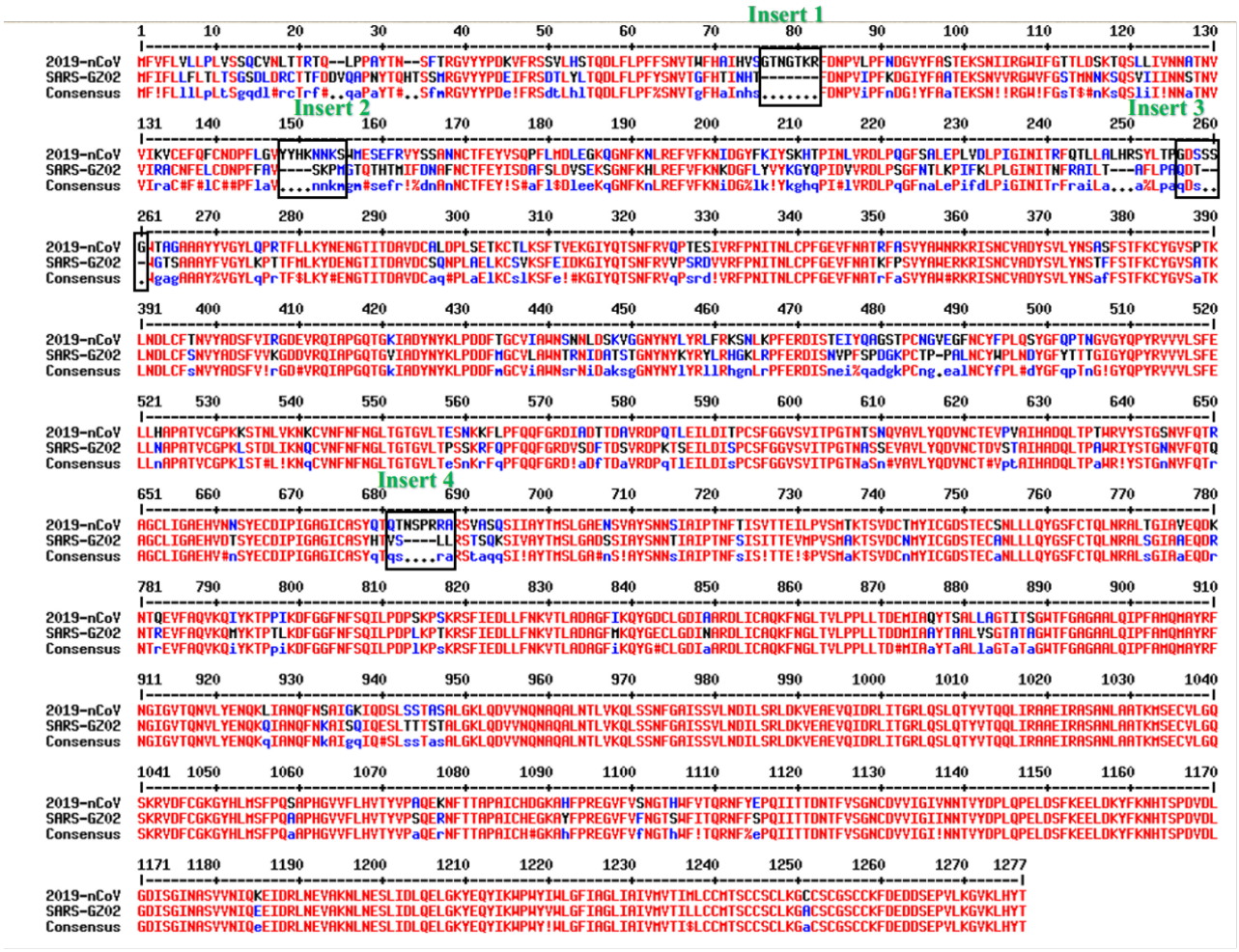


Figura 2: Alineación de secuencias múltiples entre las proteínas de punta de 2019-nCoV y el SARS. Las secuencias de las proteínas pico de 2019-nCoV (Wuhan-HU-1, Accession NC_045512) y del SARS CoV (GZ02, Accession AY390556) fueron alineadas usando el software MultiAlin. Los sitios de diferencia están resaltados en recuadros.

Luego analizamos todas las secuencias de longitud completa disponibles (n=28) de 2019-nCoV en GISAID (Elbe & Buckland-Merrett, 2017) como el de 27 de enero de 2020 para la presencia de estas inserciones. Como la mayoría de estas secuencias no están anotadas, comparamos las secuencias de nucleótidos de la glicoproteína de punta de todas las secuencias disponibles de 2019-nCoV usando BLASTp. Curiosamente, las 4 inserciones se conservaron absolutamente (100%) en todas las secuencias disponibles de 2019- nCoV analizadas [Fig.S2, Fig.S3].

Luego tradujimos el genoma alineado y encontramos que estas inserciones están presentes en todos los virus Wuhan 2019-nCoV excepto en el virus 2019-nCoV de murciélago como huésped [Fig.S4]. Intrigados por las 4 inserciones altamente conservadas únicas de 2019-nCoV quisimos entender su origen. Para ello, utilizamos la alineación local de 2019-nCoV con cada inserción como consulta contra todos los genomas de los virus y consideramos los aciertos con una cobertura de secuencia del 100%. Sorprendentemente, cada una de las cuatro inserciones se alineaban con segmentos cortos de las proteínas del Virus de Inmunodeficiencia Humana-1 (VIH-1). Las posiciones de los aminoácidos de las inserciones en 2019-nCoV y los residuos correspondientes en gp120 del VIH-1 y en gag del VIH-1 se muestran en el cuadro 1. Las primeras 3 inserciones (inserción 1, 2 y 3) alineados a segmentos cortos de residuos de aminoácidos en la gp120 del VIH-1. La inserción 4 alineada a la Gag del VIH-1. La inserción 1 (6 residuos de aminoácidos) y la inserción 2 (6 residuos de aminoácidos) en el pico de glicoproteína de 2019-nCoV son 100% idénticos a los residuos mapeados a la gp120 del VIH-1. La inserción 3 (12 residuos de aminoácidos) en 2019- nCoV se mapea a VIH-1 gp120 con lagunas [ver Tabla 1]. La inserción 4 (8 residuos de aminoácidos) se mapea a VIH-1 Gag con lagunas.

Aunque las 4 inserciones representan tramos cortos y discontinuos de aminoácidos en la glicoproteína de pico de 2019-nCoV, el hecho de que los tres compartan la identidad o similitud de aminoácidos con la gp120 del VIH-1 y la Gag del VIH-1 (entre todas las proteínas virales anotadas) sugiere que no se trata de un hallazgo fortuito. En otras palabras, se puede esperar esporádicamente una coincidencia fortuita para un tramo de 6-12 residuos de aminoácidos contiguos en una proteína no relacionada. Sin embargo, es poco probable que las 4 inserciones en la glicoproteína de punta 2019-nCoV coincidan fortuitamente con 2 proteínas estructurales clave de un virus no relacionado (VIH-1).
Los residuos de aminoácidos de las inserciones 1, 2 y 3 de la glicoproteína de punta 2019-nCoV que se mapearon a VIH-1 fueron parte de los dominios V4, V5 y V1 respectivamente en gp120 [Tabla 1]. Desde las inserciones de 2019-nCoV mapeados a regiones variables del VIH-1, no eran ubicuos en la gp120 del VIH-1, sino que se limitaban a secuencias seleccionadas del VIH-1 [véase S.File1] principalmente de Asia y África.
La proteína Gag del VIH-1 permite la interacción del virus con la superficie del huésped cargada negativamente (Murakami, 2008) y una alta carga positiva en la proteína Gag es una característica clave para la interacción huésped-virus. Al analizar los valores pI de cada una de las 4 inserciones en 2019-nCoV y los correspondientes tramos de residuos de aminoácidos de las proteínas del VIH-1, encontramos que a) los valores pI eran muy similares para cada par analizado b) la mayoría de estos valores pI eran de 10±2 [Ver Tabla 1] . Cabe destacar que, a pesar de las lagunas en las inserciones 3 y 4, los valores pI eran comparables. Esta uniformidad en los valores pI para las 4 inserciones merece una investigación más profunda.
Como ninguna de estas 4 inserciones está presente en ningún otro coronavirus, la región genómica que codifica estas inserciones representa los candidatos ideales para diseñar cebadores que puedan distinguir 2019-nCoV de otros coronavirus.

Motivos	Virus Glycoprotein	Alineación del motivo	La proteína del VIH y Región variable	Genoma del VIH País fuente/ subtipo	Número de residuos polares	Carga total	Valor pI
Inserción 1	2019-nCoV (GP) **HIV1(GP 120)**	**71 76** **TNGTKR** **TNGTKR** **404 409**	gp120-V4	Thailand*/ CRF01_AE	**5** 5	2 2	11 11
Inserción2	2019-nCoV (GP) **HIV1(GP 120)**	**145 150** **HKNNKS** **HKNNKS** **462 467**	gp120-V5	Kenya*/G	**6** 6	2 2	10 10
Inserción 3	**019- nCoV (GP)** **HIV1(GP 120)**	**245 256** **RSYL- - - -TPGDSSSG** **RTYLFNETRGNSSSG** **136 150**	gp120-V1	India*/C	8 10	2 1	10.84 **8.75**
Inserción 4	2019-nCoV (Poly P) **HIV1(gag)**	**676 684** **QTNS----------PRRA** **QTNSSILMQRSNFKG PRRA** **366** 348	Gag	India*/C	6 12	2 4	12.00 **12.30**

Tabla 1: Secuencias alineadas de 2019-nCoV y la proteína gp120 del VIH-1 con sus posiciones en la secuencia primaria de la proteína. Todas las inserciones tienen una alta densidad de residuos con carga positiva. Los fragmentos eliminados en las inserciones 3 y 4 aumentan la proporción de carga positiva en el área de superficie. *Por favor ver la Tabla 1 de la Supp. para los números de adhesión

Las nuevas inserciones son parte del sitio de unión del receptor de 2019-nCoV

Para obtener conocimientos estructurales y comprender el papel de estas inserciones en la glicoproteína 2019-nCoV, modelizamos su estructura basándonos en la estructura disponible de la glicoproteína de punta del SARS (PDB: 6ACD.1.A). La comparación de la estructura modelada revela que, aunque las inserciones 1,2 y 3 se encuentran en lugares no contiguos en la secuencia primaria de la proteína, se pliegan para constituir la parte del sitio de unión de la glicoproteína que reconoce el receptor del huésped (Kirchdoerfer y otros, 2016) (Figura 4). La inserción 1 corresponde al NTD (dominio N-terminal) y las inserciones 2 y 3 corresponden al CTD (dominio C-terminal) de la subunidad S1 en la glicoproteína de punta 2019-nCoV. La inserción 4 está en la unión de la SD1 (sub-dominio 1) y la SD2 (sub-dominio 2) de la subunidad S1 (Ou et al., 2017). Especulamos que estas inserciones proporcionan flexibilidad adicional al sitio de unión de la glicoproteína al formar un bucle hidrófilo en la estructura de la proteína que puede facilitar o mejorar las interacciones entre el virus y el huésped.

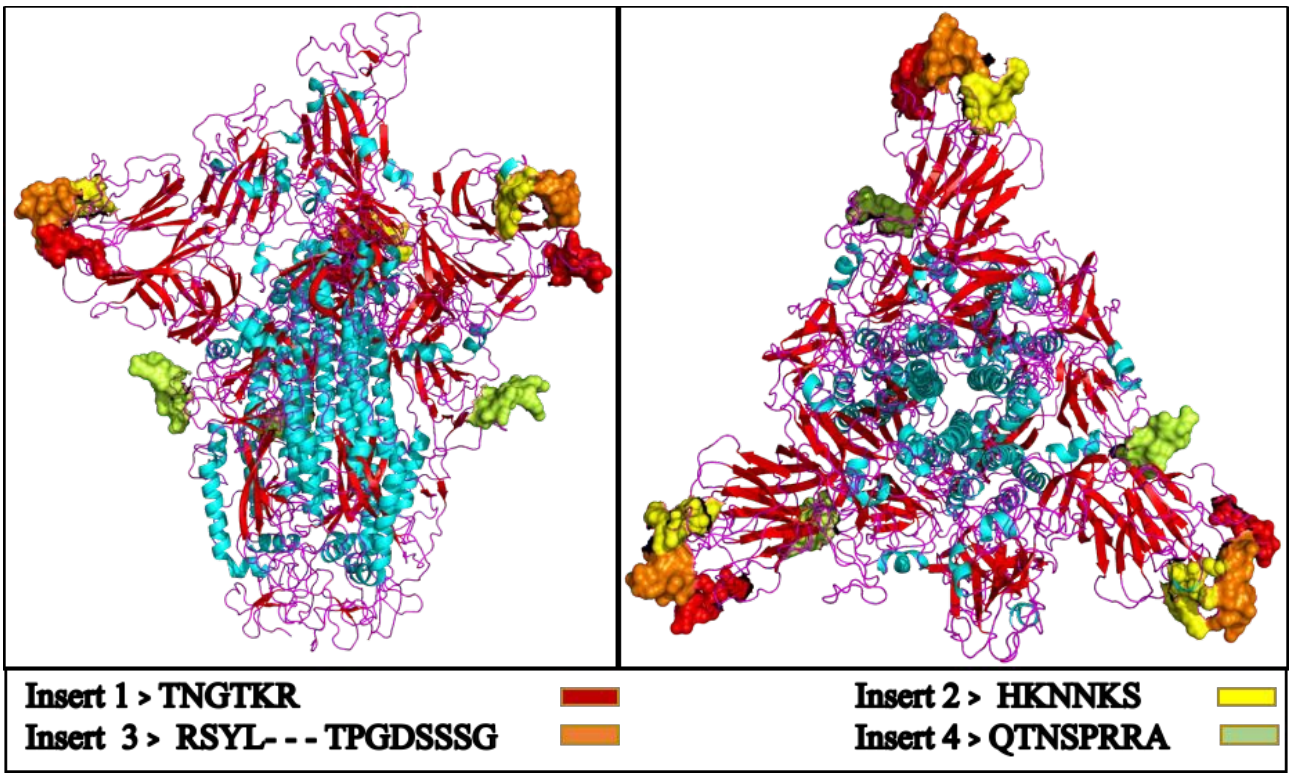


Figura 3. Glicoproteína de pico homotrimer modelada del virus 2019-nCoV. Las inserciones de la proteína de envoltura del VIH se muestran con cuentas de color, presentes en el sitio de unión de la proteína.

Análisis evolutivo de 2019-nCoV

Se ha especulado que 2019-nCoV es una variante del Coronavirus derivado de una fuente animal que se transmitió a los humanos. Considerando el cambio de especificidad para el huésped, decidimos estudiar las secuencias de la glicoproteína pico (proteína S) del virus. Las proteínas S son proteínas de superficie que ayudan al virus a reconocer y adherirse al huésped. Por lo tanto, un cambio en estas proteínas puede reflejarse como un cambio en la especificidad del huésped del virus. Para conocer las alteraciones en el gen de la proteína S de 2019-nCoV y sus consecuencias en los reordenamientos estructurales realizamos un análisis in-silico de 2019-nCoV con respecto a todos los demás virus. Una alineación de secuencias múltiples entre las secuencias de aminoácidos de la proteína S de 2019-nCoV, Bat-SARS-Like, SARS-GZ02 y MERS reveló que la proteína S ha evolucionado con una diversidad significativa más cercana a la del SARS-GZ02 (Figura 1).

Inserciones en la región de la proteína Spike de 2019-nCoV

Dado que la proteína S de 2019-nCoV comparte la ascendencia más cercana con el SARS GZ02, la codificación de la secuencia de las proteínas de punta de estos dos virus se comparó utilizando el software MultiAlin. Encontramos cuatro nuevas inserciones en la proteína de 2019-nCoV- "GTNGTKR" (IS1), "HKNNKS" (IS2), "GDSSSG" (IS3) y "QTNSPRRA" (IS4) (Figura 2). Para nuestra sorpresa, estas inserciones de secuencias no sólo estaban ausentes en la proteína S del SARS, sino que tampoco se observaron en ningún otro miembro de la familia Coronaviridae (Figura suplementaria). Esto es sorprendente, ya que es bastante improbable que un virus haya adquirido tales inserciones únicas de forma natural en un corto período de tiempo.

Las inserciones comparten similitudes con el VIH

Se observó que las inserciones estaban presentes en todas las secuencias genómicas del virus 2019-nCoV disponibles en los aislamientos clínicos recientes (Figura suplementaria 1). Para conocer la fuente

de estas inserciones en 2019-nCoV se hizo una alineación local con BLASTp usando estas inserciones como consulta con todo el genoma del virus. Inesperadamente, todas las inserciones se alinean con el Virus de Inmunodeficiencia Humana-1 (VIH-1). Análisis posteriores revelaron que las secuencias alineadas del VIH-1 con 2019-nCoV se derivaron de la glicoproteína gp120 de superficie (posiciones de la secuencia de aminoácidos: 404-409, 462-467, 136-150) y de la proteína Gag (366-384 aminoácido) (Tabla 1). La proteína mordaza del VIH participa en la unión de la membrana del huésped, en el empaquetamiento del virus y en la formación de partículas similares al virus. Gp120 juega un papel crucial en el reconocimiento de la célula huésped al unirse al receptor primario CD4. Esta unión induce reordenamientos estructurales en GP120, creando un sitio de unión de alta afinidad para un co-receptor de quimiocinas como CXCR4 y/o CCR5.

Discusión

El actual brote de 2019-nCoV justifica una investigación exhaustiva y la comprensión de su capacidad para infectar a los seres humanos. Teniendo en cuenta que ha habido un claro cambio en la preferencia del huésped de coronavirus anteriores a este virus, estudiamos el cambio en la proteína de punta entre 2019-nCoV y otros virus. Encontramos cuatro nuevas inserciones en la proteína S de 2019-nCoV cuando la comparamos con su pariente más cercano, el SARS CoV. La secuencia del genoma de los 28 aislamientos clínicos recientes mostró que la secuencia que codifica estas inserciones se conserva entre todos estos aislamientos. Esto indica que estas inserciones han sido adquiridas preferentemente por el 2019-nCoV, proporcionándole una ventaja adicional de supervivencia e infectividad. Profundizando más, encontramos que estas inserciones eran similares al VIH-1. Nuestros resultados destacan una relación asombrosa entre la gp120 y la proteína Gag del VIH, con la glicoproteína de punta 2019-nCoV. Estas proteínas son fundamentales para que los virus identifiquen y se aferren a sus células huéspedes y para el ensamblaje viral (Beniac et al., 2006). Dado que las proteínas de superficie son responsables del tropismo del huésped, los cambios en estas proteínas implican un cambio en la especificidad del huésped del virus. Según informes de China, se ha producido una ganancia de especificidad del huésped en el caso 2019-nCoV, ya que se sabía que el virus originalmente infectaba a los animales y no a los seres humanos, pero después de las mutaciones, ha ganado tropismo también para los seres humanos.

Avanzando, el modelado en 3D de la estructura de la proteína mostró que estas inserciones están presentes en el sitio de unión de 2019-nCoV. Debido a la presencia de patrones de gp120 en la glicoproteína de punta de 2019-nCoV en su dominio de unión, proponemos que estas inserciones de motivos podrían haber proporcionado una mayor afinidad hacia los receptores de las células huésped. Además, este cambio estructural también podría haber aumentado el rango de células huésped que 2019-nCoV puede infectar. Hasta donde sabemos, la función de estos patrones aún no está clara en el VIH y debe ser explorada. El intercambio de material genético entre los virus es bien conocido y ese intercambio crítico pone de relieve el riesgo y la necesidad de investigar las relaciones entre familias de virus aparentemente no relacionadas.

Conclusiones

Nuestro análisis del pico de glicoproteína de 2019-nCoV reveló varios hallazgos interesantes: Primero, identificamos 4 inserciones únicos en la glicoproteína pico de 2019-nCoV que no están presentes en ningún otro coronavirus reportado hasta la fecha. Para nuestra sorpresa, las 4 inserciones en el 2019-nCoV se mapearon a segmentos cortos de aminoácidos en el VIH-1 gp120 y Gag entre todas las proteínas de virus anotadas en la base de datos del NCBI. Esta asombrosa similitud de las nuevas inserciones en la proteína de punta de 2019-nCoV con la gp120 y Gag del VIH-1 es poco probable que sea fortuita. Además, el modelado tridimensional sugiere que al menos 3 de las inserciones únicas que no son contiguos en la secuencia de la proteína primaria de la glicoproteína de punta 2019-nCoV convergen para constituir los componentes clave del sitio de unión del receptor. Cabe destacar que las 4 inserciones tienen valores de pI de alrededor de 10 que pueden facilitar las interacciones entre virus y huésped. En conjunto, nuestros hallazgos sugieren una evolución no convencional de 2019-nCoV que justifica una mayor investigación. Nuestro trabajo destaca los nuevos aspectos evolutivos del 2019-nCoV y tiene implicaciones en la patogénesis y el diagnóstico de este virus.

Referencias

Beniac, D. R., Andonov, A., Grudeski, E., & Booth, T. F. (2006). Architecture of the SARS coronavirus prefusion spike. *Nature Structural and Molecular Biology*, *13*(8), 751–752. https://doi.org/10.1038/nsmb1123

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., Bordoli, L., & Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku340

Bosch, B. J., van der Zee, R., de Haan, C. A. M., & Rottier, P. J. M. (2003). The Coronavirus Spike Protein Is a Class I Virus Fusion Protein: Structural and Functional Characterization of the Fusion Core Complex. *Journal of Virology*, *77*(16), 8801–8811. https://doi.org/10.1128/jvi.77.16.8801-8811.2003

Chan, J. F.-W., Kok, K.-H., Zhu, Z., Chu, H., To, K. K.-W., Yuan, S., & Yuen, K.-Y. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerging Microbes & Infections*, *9*(1), 221–236. https://doi.org/10.1080/22221751.2020.1719902

Chan, J. F. W., Lau, S. K. P., To, K. K. W., Cheng, V. C. C., Woo, P. C. Y., & Yuen, K.-Y. (2015). *Middle East Respiratory Syndrome Coronavirus: Another Zoonotic Betacoronavirus Causing SARS-Like Disease*. https://doi.org/10.1128/CMR.00102-14

Chan, J., To, K., Tse, H., Jin, D., microbiology, K. Y.-T. in, & 2013, undefined. (n.d.). Interspecies transmission and emergence of novel viruses: lessons from bats and birds. *Elsevier*.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research*. https://doi.org/10.1093/nar/16.22.10881

DeLano, W. L. (2002). The PyMOL Molecular Graphics System, Version 1.1. *Schr{ö}dinger LLC*. https://doi.org/10.1038/hr.2014.17

Du, L., Zhao, G., Kou, Z., Ma, C., Sun, S., Poon, V. K. M., Lu, L., Wang, L., Debnath, A. K., Zheng, B.-J., Zhou, Y., & Jiang, S. (2013). Identification of a Receptor-Binding Domain in the S Protein of the Novel Human Coronavirus Middle East Respiratory Syndrome Coronavirus as an Essential Target for Vaccine Development. *Journal of Virology*, *87*(17), 9939–9942. https://doi.org/10.1128/jvi.01048-13

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. https://doi.org/10.1093/nar/gkh340

Elbe, S., & Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges. https://doi.org/10.1002/gch2.1018

Kirchdoerfer, R. N., Cottrell, C. A., Wang, N., Pallesen, J., Yassine, H. M., Turner, H. L., Corbett, K. S., Graham, B. S., McLellan, J. S., & Ward, A. B. (2016). Pre-fusion structure of a human coronavirus spike protein. Nature. https://doi.org/10.1038/nature17200

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution. https://doi.org/10.1093/molbev/msy096

Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. Annual Review of Virology, 3(1), 237–261. https://doi.org/10.1146/annurev-virology-110615-042301

Murakami, T. (2008). Roles of the interactions between Env and Gag proteins in the HIV-1 replication cycle. Microbiology and Immunology, 52(5), 287–295. https://doi.org/10.1111/j.1348-0421.2008.00008.x

Ou, X., Guan, H., Qin, B., Mu, Z., Wojdyla, J. A., Wang, M., Dominguez, S. R., Qian, Z., & Cui, S. (2017). Crystal structure of the receptor binding domain of the spike glycoprotein of human betacoronavirus HKU1.

Nature Communications. https://doi.org/10.1038/ncomms15216

Snijder, E. J., van der Meer, Y., Zevenhoven-Dobbe, J., Onderwater, J. J. M., van der Meulen, J., Koerten, H. K., & Mommaas, A. M. (2006). Ultrastructure and origin of membrane vesicles associated with the severe acute respiratory syndrome coronavirus replication complex. Journal of Virology, 80(12), 5927–5940. https://doi.org/10.1128/JVI.02501-05

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., … Shi, Z.-L. (2020). Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. BioRxiv. https://doi.org/10.1101/2020.01.22.914952

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine, NEJMoa2001017. https://doi.org/10.1056/NEJMoa2001017

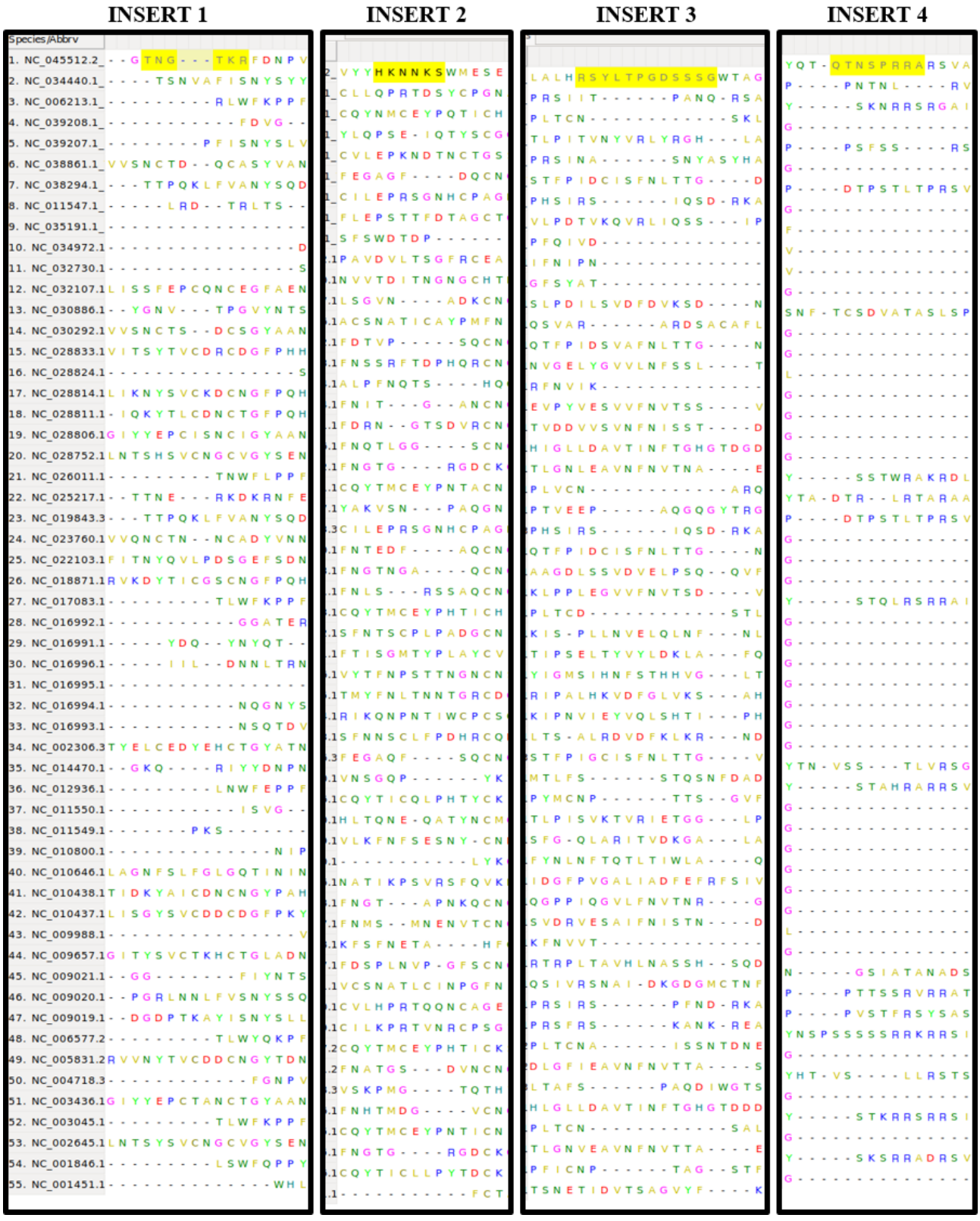


Fig.S1 Alineación de secuencia múltiple de glucoproteína de la familia coronaviridae, que representa los cuatro insertos.

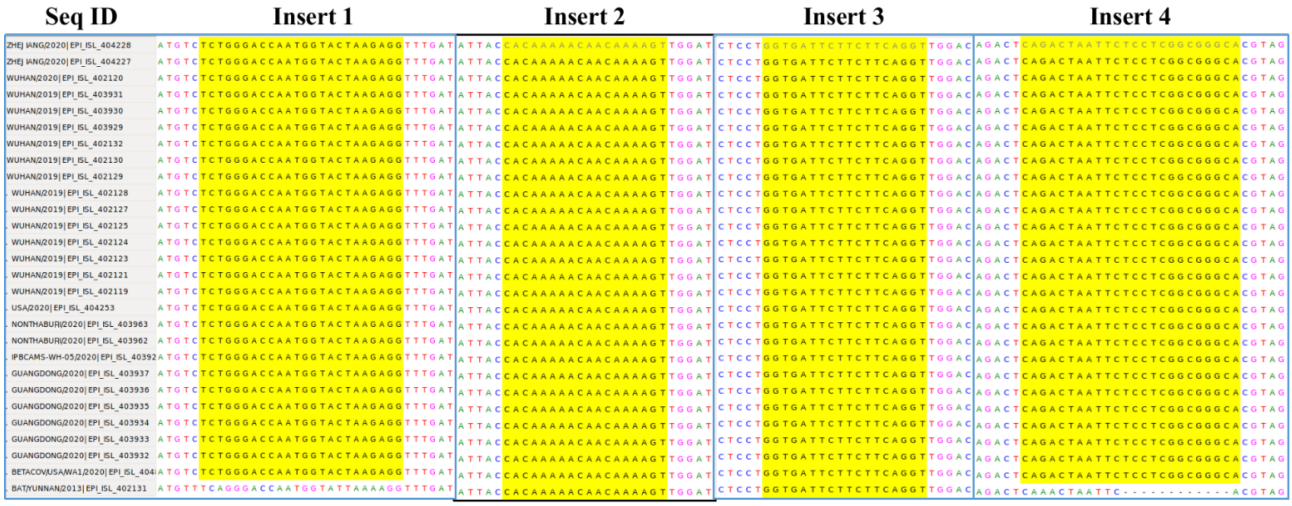


Fig.S2: Las cuatro inserciones están presentes en los genomas alineados del virus 28 Wuhan 2019-nCoV obtenidos del GISAID. El hueco en el Bat-SARS Like CoV en la última fila muestra que la inserción 1 y 4 es muy exclusivo de Wuhan 2019-nCoV.

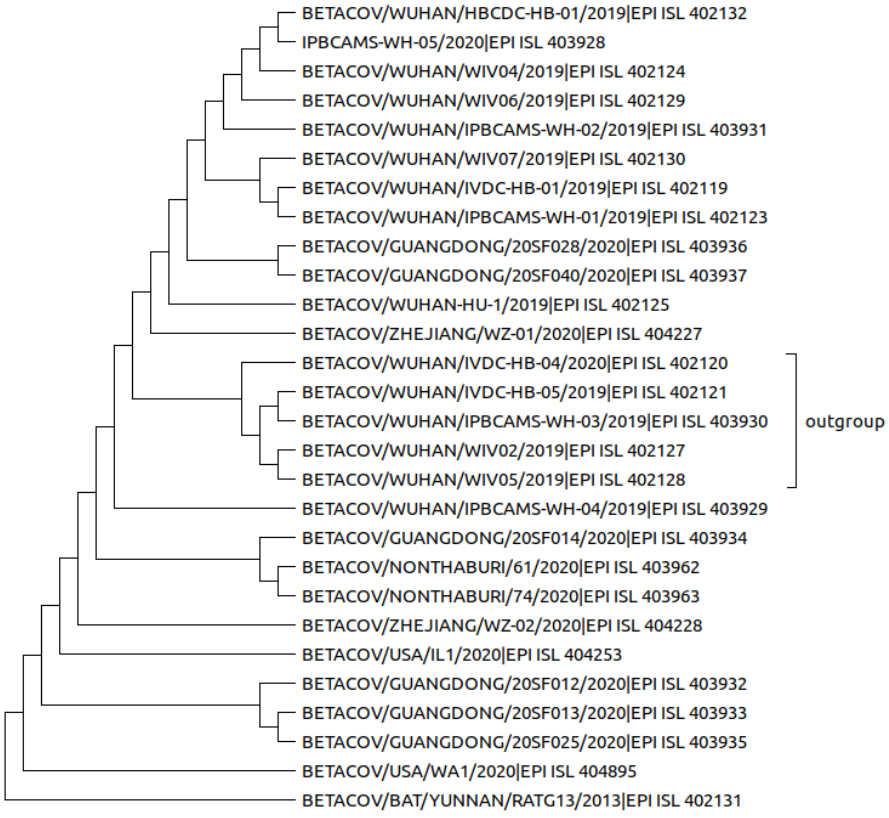
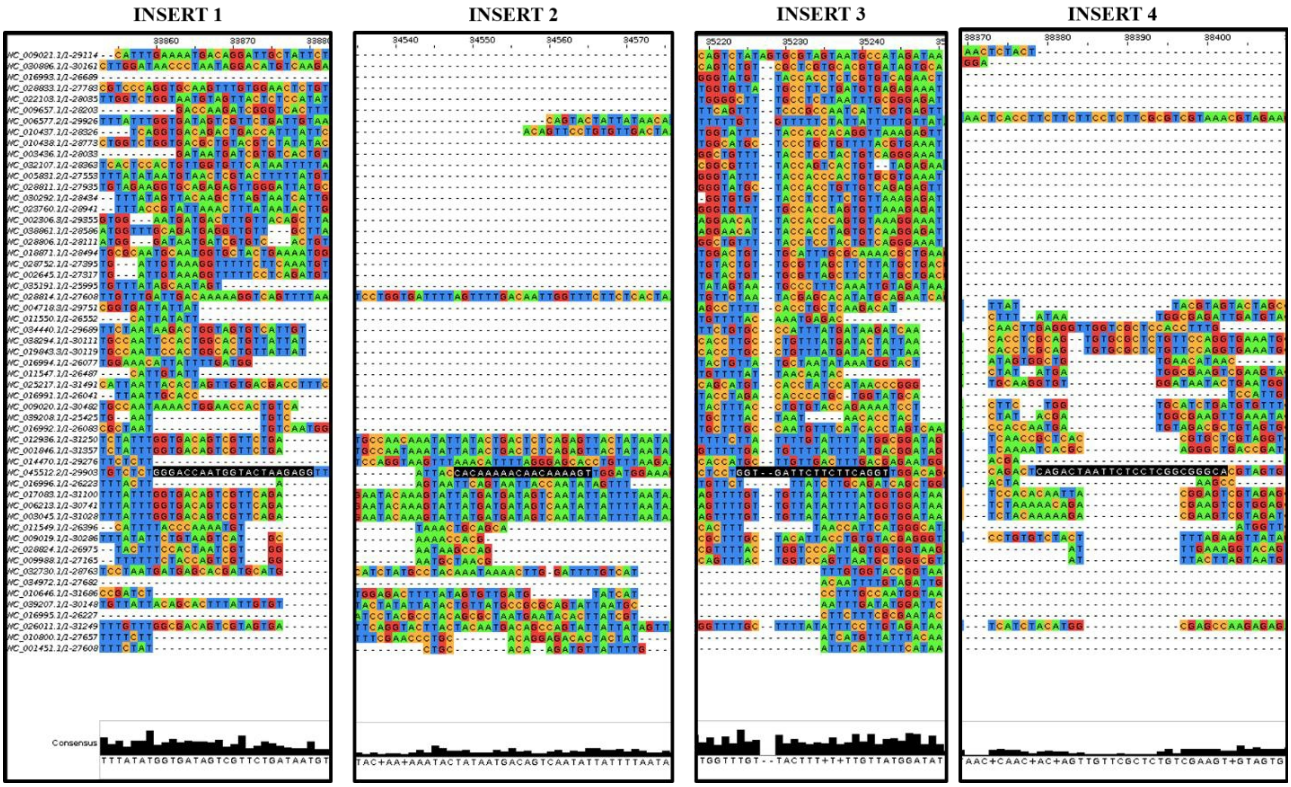


Fig.S3 Árbol filogenético de 28 aislamientos clínicos del genoma de 2019-nCoV incluyendo uno de murciélago como huésped.



Suplemento de la figura 4. Alineación del genoma de la familia Coronaviridae. Las secuencias negras resaltadas son las inserciones representadas aquí

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.30.927871