## Insertions share similarity to HIV

The insertions were observed to be present in all the genomic sequences of 2019-nCoV virus available from the recent clinical isolates (Supplementary Figure 1). To know the source of these insertions in 2019-nCoV a local alignment was done with BLASTp using these insertions as query with all virus genome. Unexpectedly, all the insertions got aligned with Human immunodeficiency Virus-1 (HIV-1). Further analysis revealed that aligned sequences of HIV-1 with 2019-nCoV were derived from surface glycoprotein gp120 (amino acid sequence positions: 404-409, 462-467, 136-150) and from Gag protein (366-384 amino acid) (Table 1). Gag protein of HIV is involved in host membrane binding, packaging of the virus and for the formation of virus-like particles. Gp120 plays crucial role in recognizing the host cell by binding to the primary receptor CD4.This binding induces structural rearrangements in GP120, creating a high affinity binding site for a chemokine co-receptor like CXCR4 and/or CCR5.

## Discussion

The current outbreak of 2019-nCoV warrants a thorough investigation and understanding of its ability to infect human beings. Keeping in mind that there has been a clear change in the preference of host from previous coronaviruses to this virus, we studied the change in spike protein between 2019-nCoV and other viruses. We found four new insertions in the S protein of 2019-nCoV when compared to its nearest relative, SARS CoV. The genome sequence from the recent 28 clinical isolates showed that the sequence coding for these insertions are conserved amongst all these isolates. This indicates that these insertions have been preferably acquired by the 2019-nCoV, providing it with additional survival and infectivity advantage. Delving deeper we found that these insertions were similar to HIV-1. Our results highlight an astonishing relation between the gp120 and Gag protein of HIV, with 2019-nCoV spike glycoprotein. These proteins are critical for the viruses to identify and latch on to their host cells and for viral assembly (Beniac et al., 2006). Since surface proteins are responsible for host tropism, changes in these proteins imply a change in host specificity of the virus. According to reports from China, there has been a gain of host specificity in case 2019-nCoV as the virus was originally known to infect animals and not humans but after the mutations, it has gained tropism to humans as well.

Moving ahead, 3D modelling of the protein structure displayed that these insertions are present at the binding site of 2019-nCoV. Due to the presence of gp120 motifs in 2019-nCoV spike glycoprotein at its binding domain, we propose that these motif insertions could have provided an enhanced affinity towards host cell receptors. Further, this structural change might have also increased the range of host cells that 2019-nCoV can infect. To the best of our knowledge, the function of these motifs is still not clear in HIV and need to be explored. The exchange of genetic material among the viruses is well known and such critical exchange highlights the risk and the need to investigate the relations between seemingly unrelated virus families.

## Conclusions

Our analysis of the spike glycoprotein of 2019-nCoV revealed several interesting findings: First, we identified 4 unique inserts in the 2019-nCoV spike glycoprotein that are not present in any other coronavirus reported till date. To our surprise, all the 4 inserts in the 2019-nCoV mapped to

short segments of amino acids in the HIV-1 gp120 and Gag among all annotated virus proteins in the NCBI database. This uncanny similarity of novel inserts in the 2019- nCoV spike protein to HIV-1 gp120 and Gag is unlikely to be fortuitous. Further, 3D modelling suggests that atleast 3 of the unique inserts which are non-contiguous in the primary protein sequence of the 2019-nCoV spike glycoprotein converge to constitute the key components of the receptor binding site. Of note, all the 4 inserts have pI values of around 10 that may facilitate virus-host interactions. Taken together, our findings suggest unconventional evolution of 2019-nCoV that warrants further investigation. Our work highlights novel evolutionary aspects of the 2019-nCoV and has implications on the pathogenesis and diagnosis of this virus.

## References

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

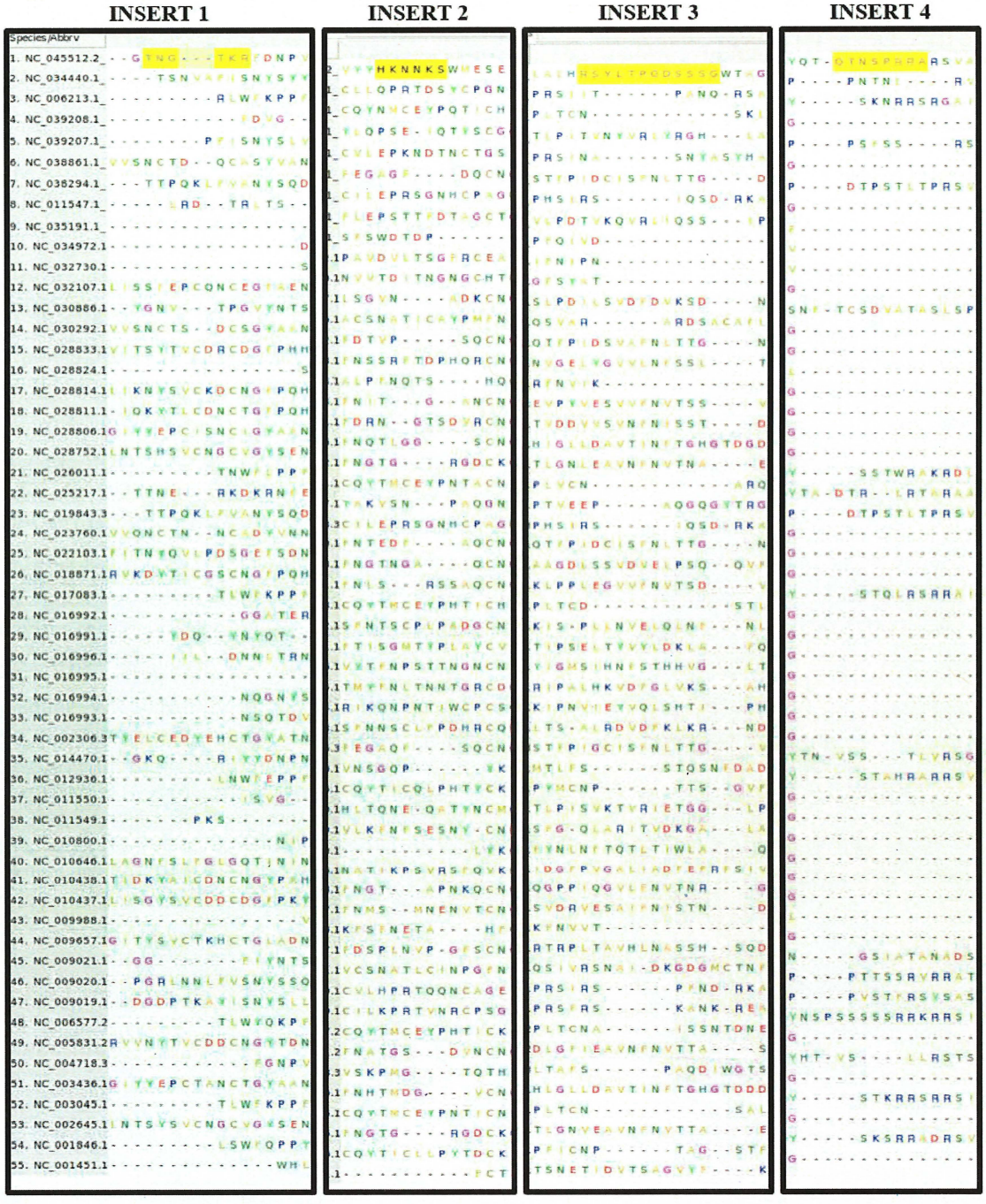



Fig.S1 Multiple sequence alignment of glycoprotein of *coronaviridae* family, representing all the four inserts.

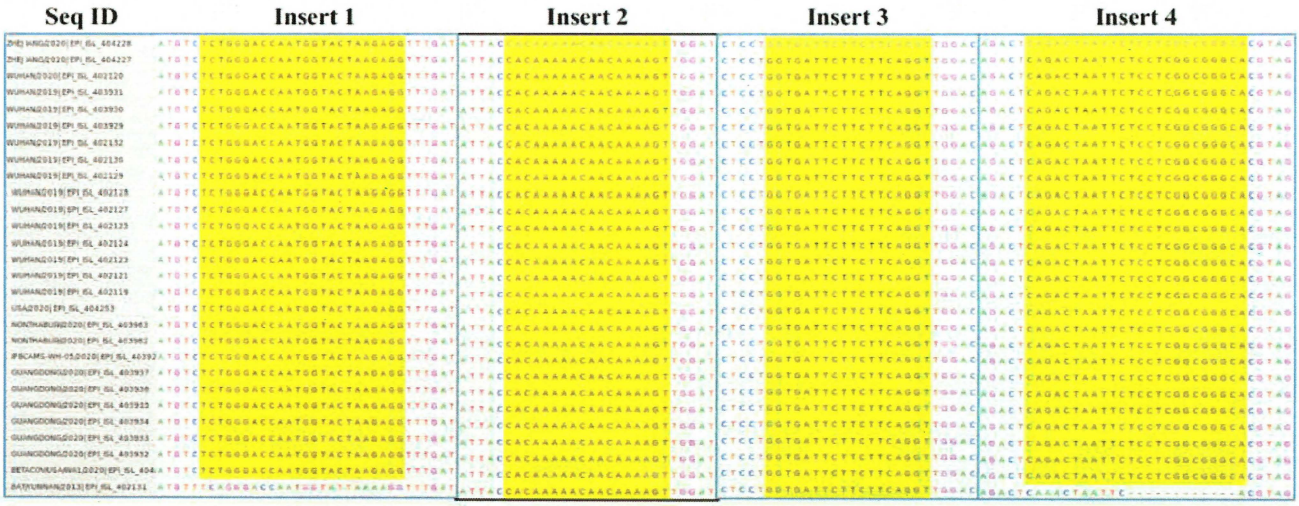



Fig.S2: All four inserts are present in the aligned 28 Wuhan 2019-nCoV virus genomes obtained from GISAID. The gap in the Bat-SARS Like CoV in the last row shows that insert 1 and 4 is very unique to Wuhan 2019-nCoV.

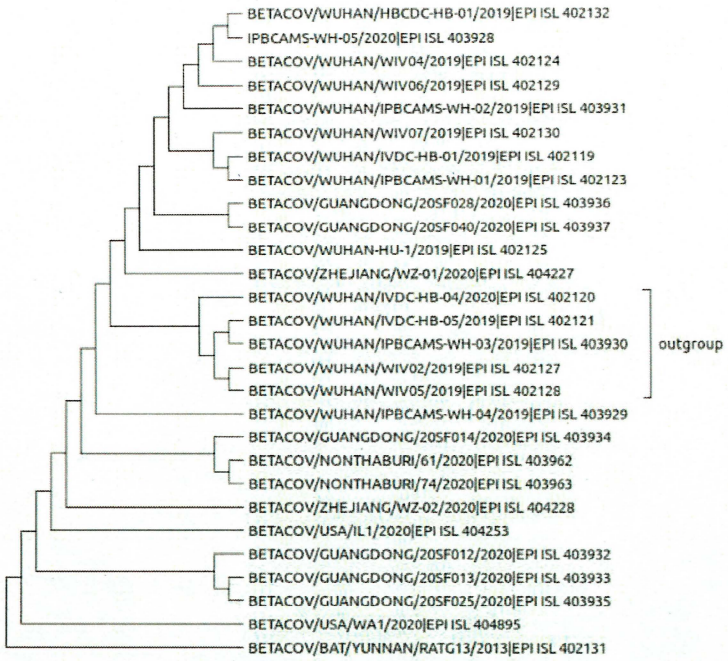
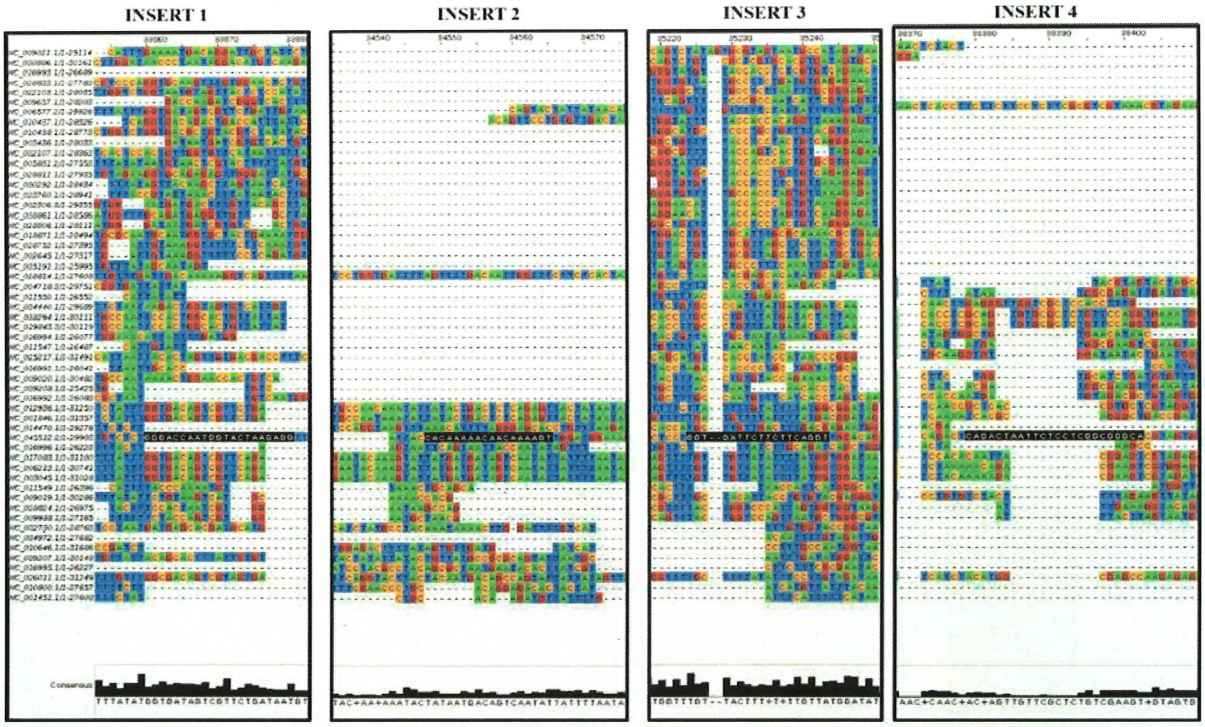



Fig.S3 Phylogenetic tree of 28 clinical isolates genome of 2019-nCoV including one from bat as a host.




Supplementary Fig 4. Genome alingment of Coronaviridae family. Highlighted black sequences are the inserts represented here.