Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Infection, Genetics and Evolution 79 (2020) 104212

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event

D. Paraskevis[a,*], E.G. Kostaki[a], G. Magiorkinis[a], G. Panayiotakopoulos[b], G. Sourvinos[c], S. Tsiodras[d]

[a] Department of Hygiene Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, Athens, Greece
[b] National Public Health Organization (NPHO), Athens, Greece
[c] Laboratory of Clinical Virology, School of Medicine, University of Crete, Heraklion, Greece
[d] Medical School, National and Kapodistrian University of Athens, Athens, Greece

ARTICLE INFO

Keywords:
Novel coronavirus
Genomic sequence analysis
Phylogenetic analysis
Recombination
Origin
Molecular epidemiology

ABSTRACT

Background: A novel coronavirus (2019-nCoV) associated with human to human transmission and severe human infection has been recently reported from the city of Wuhan in China. Our objectives were to characterize the genetic relationships of the 2019-nCoV and to search for putative recombination within the subgenus of sarbecovirus.

Methods: Putative recombination was investigated by RDP4 and Simplot v3.5.1 and discordant phylogenetic clustering in individual genomic fragments was confirmed by phylogenetic analysis using maximum likelihood and Bayesian methods.

Results: Our analysis suggests that the 2019-nCoV although closely related to BatCoV RaTG13 sequence throughout the genome (sequence similarity 96.3%), shows discordant clustering with the Bat_SARS-like coronavirus sequences. Specifically, in the 5′-part spanning the first 11,498 nucleotides and the last 3′-part spanning 24,341–30,696 positions, 2019-nCoV and RaTG13 formed a single cluster with Bat_SARS-like coronavirus sequences, whereas in the middle region spanning the 3′-end of ORF1a, the ORF1b and almost half of the spike regions, 2019-nCoV and RaTG13 grouped in a separate distant lineage within the sarbecovirus branch.

Conclusions: The levels of genetic similarity between the 2019-nCoV and RaTG13 suggest that the latter does not provide the exact variant that caused the outbreak in humans, but the hypothesis that 2019-nCoV has originated from bats is very likely. We show evidence that the novel coronavirus (2019-nCov) is not-mosaic consisting in almost half of its genome of a distinct lineage within the betacoronavirus. These genomic features and their potential association with virus characteristics and virulence in humans need further attention.

The family Coronaviridae includes a large number of viruses that in nature are found in birds and mammals (Kahn and McIntosh, 2005; Fehr and Perlman, 2015). Human coronaviruses, first characterized in the 1960s, are associated with a large percentage of respiratory infections both in children and adults (Kahn and McIntosh, 2005; Paules et al., 2020).

Scientific interest in Coronaviruses exponentially increased after the emergence of SARS-Coronavirus (SARS-CoV) in Southern China (Drosten et al., 2003; Ksiazek et al., 2003; Peiris et al., 2003). Its rapid spread led to the global appearance of more than 8000 human cases and 774 deaths (Kahn and McIntosh, 2005). The virus was initially detected in Himalayan palm civets (Guan et al., 2003) that may have served as an amplification host; the civet virus contained a 29-nucleotide sequence not found in most human isolates that were related to the global epidemic (Guan et al., 2003). It has been speculated that the function of the affected open reading frame (ORF 10) might have played a role in the trans-species jump (Kahn and McIntosh, 2005). A similar virus was found later in horseshoe bats (Lau et al., 2005; Li et al., 2005a). A 29-bp insertion in ORF 8 of bat-SARS-CoV genome, not found in most human SARS-CoV genomes, was suggestive of a common ancestor with civet SARS-CoV (Lau et al., 2005). After the SARS epidemic, bats have been considered as a potential reservoir species that could be implicated in

* Corresponding author at: Department of Hygiene, Epidemiology and Medical Statistics, Medical School, University of Athens, 75 Mikras Asias Street, 115 27 Athens, Greece.
E-mail address: dparask@med.uoa.gr (D. Paraskevis).

https://doi.org/10.1016/j.meegid.2020.104212
Received 27 January 2020; Accepted 28 January 2020
Available online 29 January 2020

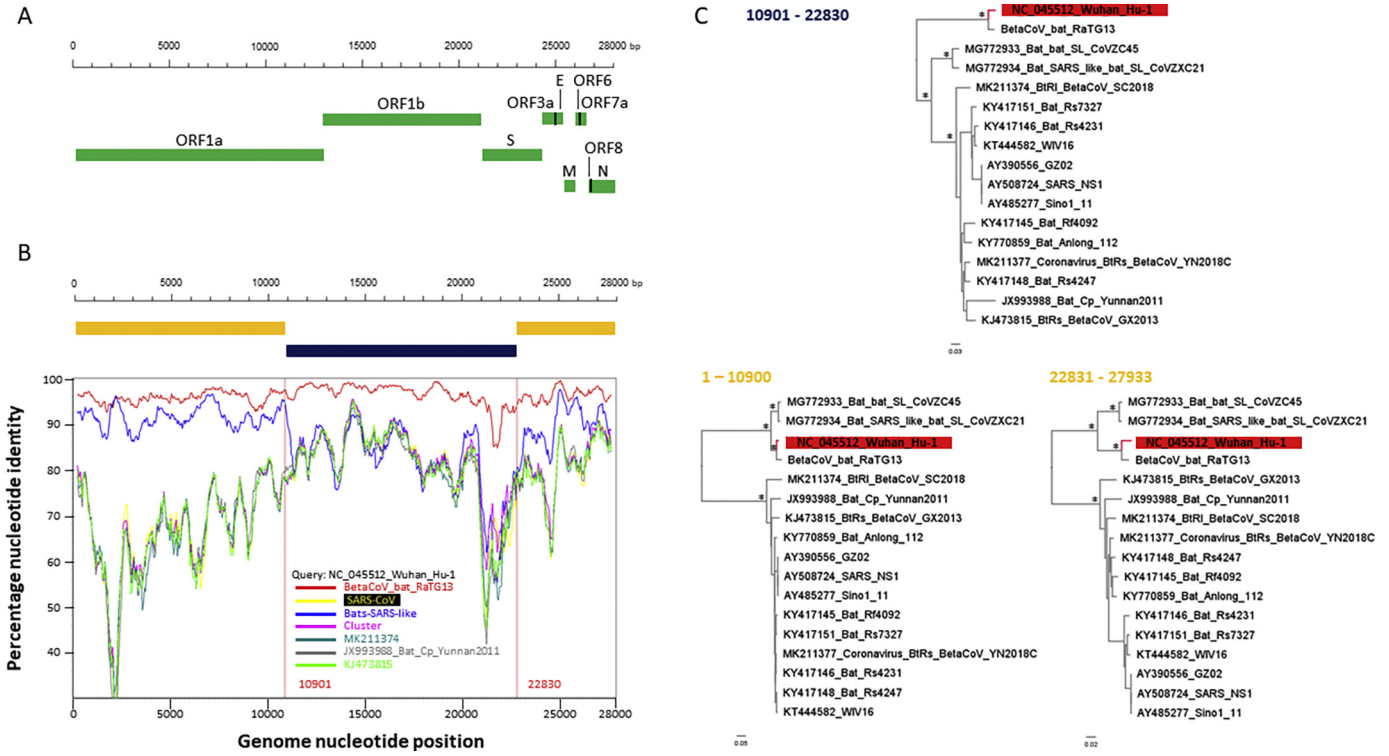

Fig. 1. A. Genomic organization of the novel coronavirus (2019-nCoV) according to the positions in the edited alignment. B. Simplot of 2019-nCoV (NC_045512_Wuhan_Hu-1) against sequences within the subgenus sarbecovirus. Different colours correspond to the nucleotide similarity between the 2019-nCoV and different groups. The regions with discordant phylogenetic clustering of the 2019-nCoV with Bats_SARS-like sequences are shown in different colours. C. Maximum likelihood (ML) phylogenetic trees inferred in different genomic regions as indicated by the Simplot analysis. The genomic regions are shown in numbers at the top or at the left of the trees. The 2019-nCoV sequence is shown in red and stars indicate important nodes received 100% bootstrap and 1 posterior probability support. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

future coronavirus-related human pandemics (Cui et al., 2019). During 2012 Middle East Respiratory coronavirus (MERS-CoV) emerged in Saudi Arabia (Zaki et al., 2012; Hajjar et al., 2013) and has since claimed the lives of 919 out of 2521 (35%) people affected (ECDC, 2020). A main role in the transmission of the virus to humans has been attributed to dromedary camels (Alagaili et al., 2014) and its origin has been again traced to bats (Ithete et al., 2013).

Ever since both SARS and MERS-CoV (due to their high case fatality rates) are prioritized together with "highly pathogenic coronaviral diseases other than MERS and SARS" under the Research and Development Blueprint published by the WHO (World Health Organization, 2018).

A novel coronavirus (2019-nCoV) associated with human to human transmission and severe human infection has been recently reported from the city of Wuhan in Hubei province in China (World Health Organization, 2020; Hui et al., 2020). A total of 1,320 confirmed and 1,965 suspect cases were reported up to 25 January 2020; of the confirmed cases 237 were severely ill and 41 had died (World Health Organization, 2020). Most of the original cases had close contact with a local fresh seafood and an animal market (Zhu et al., 2020; Perlman, 2020).

Full-genome sequence analysis of 2019-nCoV revealed that belongs to betacoronavirus, but it is divergent from SARS-CoV and MERS-CoV that caused epidemics in the past (Zhu et al., 2020). The 2019-nCoV along with the Bat_SARS-like coronavirus forms a distinct lineage within the subgenus of the sarbecovirus (Zhu et al., 2020).

Our objectives were to characterize the genetic relationships of the 2019-nCoV and to search for putative recombination within the subgenus of sarbecovirus.

Viral sequences were downloaded from NCBI nucleotide sequence database (http://www.ncbi.nlm.nih.gov). The BatCoV RaTG13 sequence was downloaded from the GISAID BetaCov 2019–2020 repository (http://www.GISAID.org). The sequence was reported in Zhou et al. (2020). Full-genomic sequence alignment was performed using MAFFT v7.4.2. (Katoh and Standley, 2013) and manually edited using MEGA v1.0 (Stecher et al., 2020) according to the encoded reading frame. Putative recombination was investigated by RDP4 (Martin, 2015) and Simplot v3.5.1 (Lole et al., 1999) and discordant phylogenetic clustering in individual genomic fragments was confirmed by phylogenetic analysis using maximum likelihood (ML) and Bayesian methods. ML trees were reconstructed using Neighbor-Joining (NJ) with ML distances or after heuristic ML search (TBR) with GTR + G as nucleotide substitution model as implemented in PAUP* 4.0 beta (Swofford, 2003). The GTR + G was used in Bayesian analysis as implemented in MrBayes v3.2.7 (Huelsenbeck and Ronquist, 2001). Phylogenetic trees were viewed using FigTree v1.4 (http://tree.bio.ed.ac.uk/software/figtree/).

A similarity plot was performed using a sliding window of 450 nts moving in steps of 50 nts, between the query sequence (2019-nCoV) and different sequences grouped according to their clustering pattern. The similarity plot (Fig. 1A,B) suggested that the RaTG13 was the most closely related sequence to the 2019-nCoV throughout the genome. The genetic similarity between the 2019-nCoV and RaTG13 was 96.3% (p-distance: 0.0369). On the other hand, a discordant relationship was detected between the query and the sequences of the Bat_SARS-like coronavirus (MG772934 and MG772933) (Fig. 1C). Specifically in in the 5′-part of the genome spanning the first 10,901 nts of the alignment that correspond to the 11,498 nucleotides of the prototype strain (NC_045512) and the last 3′-part spanning 22,831–27,933 positions (24,341–30,696 nucleotides in the NC_045512), 2019-nCoV and RaTG13 formed a single cluster with Bat_SARS-like coronavirus sequences (Fig. 1C). In the middle region spanning the 3′-end of ORF1a,

Fig. 2. Neighbor joining distance tree of the BLAST search results of the 2019-nCoV (NC_0455_Wuhan_Hu_1) sequence in the genomic region 10,901–22,830 of the alignment.

the ORF1b and almost half of the spike regions (10,901–22,830 nts in the alignment or 11,499–24,340 of the NC_045512), 2019-nCoV and RaTG13 grouped in a separate distant lineage within the sarbecovirus branch (Fig. 1B, C). In this region the 2019-nCoV and RaTG13 is distantly related to the Bat_SARS-like coronavirus sequences. Phylogenetic analyses using different methods confirmed these findings. A BLAST search of 2019-nCoV middle fragment revealed no considerable similarity with any of the previously characterized corona viruses (Fig. 2).

Our study suggests that the new corona virus (2019-nCoV) is not a mosaic and it is most closely related with the BatCoV RaTG13 detected in bats from Yunnan Province (Zhou et al., 2020). The levels of genetic similarity between the 2019-nCoV and RaTG13 suggest that the latter does not provide the exact variant that caused the outbreak in humans, but the hypothesis that 2019-nCoV has originated from bats is very likely. On the other hand, there is evidence for discordant phylogenetic relationships between 2019-nCoV and RaTG13 clade with their closest partners, the Bat_SARS-like coronavirus sequences. In accordance with previous analysis (http://virological.org/t/ncovs-relationship-to-bat-coronaviruses-recombination-signals-no-snakes/331), Bat_SARS-like coronavirus sequences cluster in different positions in the tree, suggesting that they are recombinants, and thus that the 2019-nCoV and RaTG13 are not (Ji et al., 2020; Magiorkinis et al., 2004). One previous study based on codon usage analyses suggested that the spike protein of 2019-nCoV might have originated from one yet-unknown unsampled coronavirus through recombination (Ji et al., 2020). Codon usage analyses can resolve the origin of proteins with deep ancestry and insufficient phylogenetic signal or invented de novo. The recently-published bat coronavirus sequence however provides strong phylogenetic information to resolve the origin of the Spike protein, as well as the rest of the genome, suggesting a uniform ancestry across the genome. We have previously shown that phylogenetic discordance in deep relationships of coronaviruses is common and can be explained either by ancient recombination event or altered evolutionary rates in different lineages, or a combination of both (Magiorkinis et al., 2004). Our study rejects the hypothesis of emergence as a result of a recent recombination event. Notably, the new coronavirus provides a new lineage for almost half of its genome, with no close genetic relationships to other viruses within the subgenus of sarbecovirus. This genomic part comprises half of the spike region encoding a multifunctional protein responsible also for virus entry into host cells (Babcock et al., 2004; Li et al., 2005b). The unique genetic features of 2019-nCoV and their potential association with virus characteristics and virulence in humans remain to be elucidated.

Declaration of Competing Interest

All authors report no conflict of interest related to the submitted work.

References

Alagaili, A.N., et al., 2014. Middle East respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. mBio 5 (2) e00884–14.
World Health Organization, 2018. Annual review of diseases prioritized under the Research and Development Blueprint World Health Organization, Informal consultation, 6-7 February 2018, Geneva, Switzerland.
Babcock, G.J., et al., 2004. Amino acids 270 to 510 of the severe acute respiratory syndrome coronavirus spike protein are required for interaction with receptor. J. Virol. 78 (9), 4552–4560.
Cui, J., Li, F., Shi, Z.L., 2019. Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17 (3), 181–192.
Drosten, C., et al., 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N. Engl. J. Med. 348 (20), 1967–1976.
ECDC, January 2020. Risk assessment guidelines for infectious diseases transmitted on aircraft (RAGIDA) Middle East Respiratory Syndrome Coronavirus (MERS-CoV). In: ECDC Technical Report.
Fehr, A.R., Perlman, S., 2015. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol. Biol. 1282, 1–23.
Guan, Y., et al., 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302 (5643), 276–278.
Hajjar, S.A., Memish, Z.A., McIntosh, K., 2013. Middle East Respiratory Syndrome Coronavirus (MERS-CoV): a perpetual challenge. Ann. Saudi Med. 33 (5), 427–436.
Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17 (8), 754–755.
Hui, D.S., et al., 2020. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - the latest 2019 novel coronavirus outbreak in Wuhan, China. Int. J. Infect. Dis. 91, 264–266.
Ithete, N.L., et al., 2013. Close relative of human Middle East respiratory syndrome coronavirus in bat, South Africa. Emerg. Infect. Dis. 19 (10), 1697–1699.
Ji, W., et al., 2020. Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human. Int. J. Infect. Dis. 91, 264–266.
Kahn, J.S., McIntosh, K., 2005. History and recent advances in coronavirus discovery. Pediatr. Infect. Dis. J. 24 (11 Suppl) p. S223–7, discussion S226.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (4), 772–780.
Ksiazek, T.G., et al., 2003. A novel coronavirus associated with severe acute respiratory syndrome. N. Engl. J. Med. 348 (20), 1953–1966.
Lau, S.K., et al., 2005. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc. Natl. Acad. Sci. U. S. A. 102 (39), 14040–14045.
Li, W., et al., 2005a. Bats are natural reservoirs of SARS-like coronaviruses. Science 310 (5748), 676–679.
Li, W., et al., 2005b. Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. EMBO J. 24 (8), 1634–1643.
Lole, K.S., et al., 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. 73 (1), 152–160.
Magiorkinis, G., et al., 2004. Phylogenetic analysis of the full-length SARS-CoV sequences: evidence for phylogenetic discordance in three genomic regions. J. Med. Virol. 74 (3), 369–372.

Martin, D.P., et al., 2015. RDP4: detection and analysis of recombination patterns in virus genomes. Virus Evol. 1 (1) p. vev003.
Paules, C.I., Marston, H.D., Fauci, A.S., 2020. Coronavirus infections-more than just the common cold. JAMA. https://doi.org/10.1001/jama.2020.0757.
Peiris, J.S., et al., 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet 361 (9366), 1319–1325.
Perlman, S., 2020. Another decade, another coronavirus. N. Engl. J. Med. https://doi.org/10.1056/NEJMe2001126.
Stecher, G., Tamura, K., Kumar, S., 2020. Molecular evolutionary genetics analysis (MEGA) for macOS. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msz312. pii: msz312.
Swofford, D.L., 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
World Health Organization, 2020. Novel Coronavirus (2019-nCoV) Situation report- 5, 25 January 2020. Geneva, Switzerland.
Zaki, A.M., et al., 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. 367 (19), 1814–1820.
Zhou, P., et al., 2020. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. bioRxiv p. 2020.01.22.914952.
Zhu, N., et al., 2020. A novel coronavirus from patients with pneumonia in China, 2019. In: N Engl J Med.