


RESEARCH ARTICLE

Open Access



# Horizontal gene transfer and recombination analysis of SARS-CoV-2 genes helps discover its close relatives and shed light on its origin

Vladimir Makarenkov<sup>1\*</sup> , Bogdan Mazouze<sup>2†</sup>, Guillaume Rabusseau<sup>2,3</sup> and Pierre Legendre<sup>4</sup>

## Abstract

**Background:** The SARS-CoV-2 pandemic is one of the greatest global medical and social challenges that have emerged in recent history. Human coronavirus strains discovered during previous SARS outbreaks have been hypothesized to pass from bats to humans using intermediate hosts, e.g. civets for SARS-CoV and camels for MERS-CoV. The discovery of an intermediate host of SARS-CoV-2 and the identification of specific mechanism of its emergence in humans are topics of primary evolutionary importance. In this study we investigate the evolutionary patterns of 11 main genes of SARS-CoV-2. Previous studies suggested that the genome of SARS-CoV-2 is highly similar to the horse-shoe bat coronavirus RaTG13 for most of the genes and to some Malayan pangolin coronavirus (CoV) strains for the receptor binding (RB) domain of the spike protein.

**Results:** We provide a detailed list of statistically significant horizontal gene transfer and recombination events (both intergenic and intragenic) inferred for each of 11 main genes of the SARS-CoV-2 genome. Our analysis reveals that two continuous regions of genes S and N of SARS-CoV-2 may result from intragenic recombination between RaTG13 and Guangdong (GD) Pangolin CoVs. Statistically significant gene transfer-recombination events between RaTG13 and GD Pangolin CoV have been identified in region [1215–1425] of gene S and region [534–727] of gene N. Moreover, some statistically significant recombination events between the ancestors of SARS-CoV-2, RaTG13, GD Pangolin CoV and bat CoV ZC45-ZXC21 coronaviruses have been identified in genes ORF1ab, S, ORF3a, ORF7a, ORF8 and N. Furthermore, topology-based clustering of gene trees inferred for 25 CoV organisms revealed a three-way evolution of coronavirus genes, with gene phylogenies of ORF1ab, S and N forming the first cluster, gene phylogenies of ORF3a, E, M, ORF6, ORF7a, ORF7b and ORF8 forming the second cluster, and phylogeny of gene ORF10 forming the third cluster.

**Conclusions:** The results of our horizontal gene transfer and recombination analysis suggest that SARS-CoV-2 could not only be a chimera virus resulting from recombination of the bat RaTG13 and Guangdong pangolin coronaviruses but also a close relative of the bat CoV ZC45 and ZXC21 strains. They also indicate that a GD pangolin may be an intermediate host of this dangerous virus.

**Keywords:** Evolution of SARS-CoV-2, Gene evolution, Horizontal gene transfer, Recombination, Phylogenetic network, Consensus tree

## Background

The recent outbreak of a serious pneumonia disease caused by the SARS-CoV-2 (i.e. COVID-19) pathogen has highlighted the danger of coronavirus spread between different zoonotic sources. Some important transfers of genetic information across species have been

\*Correspondence: makarenkov.vladimir@uqam.ca

<sup>†</sup>Vladimir Makarenkov and Bogdan Mazouze have contributed equally to this manuscript

<sup>1</sup>Département d'informatique, Université du Québec à Montréal, Montreal, QC, Canada

Full list of author information is available at the end of the article



observed during the first SARS outbreak, involving species from various wet markets in China [62]. Several recent studies have suggested that the only close relative of SARS-CoV-2 is the RaTG13 CoV found in *Rhinolophus affinis* (horseshoe bats) [22, 68]. Thus, these bats could be considered as the main natural reservoir of the SARS-CoV and SARS-CoV-2 viruses. However, recent analyses of the SARS-CoV-2 genome conducted by Lu et al. [40] and Lam et al. [31] have indicated its high resemblance in certain regions with different coronavirus genomes of Malayan pangolins (*Manis javanica*). Zhang et al. [66, 67] have reported that the SARS-CoV-2 genome is 91.02% identical to that of a Guangdong (GD Pangolin CoV—virus found in dead Malayan pangolins in the Guangdong province of China [37]). While at the whole-genome level RaTG13 remains the closest to SARS-CoV-2 coronavirus organism overall (these CoVs share 96% of whole genome identity), the receptor binding (RB) domain of the spike (S) protein of SARS-CoV-2 is much more similar to the RB domain of the GD Pangolin CoV than to that of RaTG13 [31, 66, 67]. Five key amino acid residues taking part in the interaction with human angiotensin-converting enzyme 2 (ACE2) are completely identical in SARS-CoV-2 and GD Pangolin CoV. However, these amino acids are different in the RB domain of RaTG13 and the Guangxi (GX) Pangolin CoV (virus found in Malayan pangolins in the Guangxi province of China). Three possible evolutionary hypotheses could be advanced to explain this paradigm. According to the first of them, these mutations may have occurred as a consequence of the phenomenon of parallel evolution, when distinct CoV organisms have undergone similar mutations, and thus have developed similar traits, in response to common evolutionary pressure. For example, it is possible that SARS-CoV-2 had acquired the RB domain mutations during adaptation to passage in cell culture, as has been observed for SARS-CoV [1]. The second reasonable hypothesis is divergent evolution favoring amino acid substitutions in the RaTG13 lineage, independent of recombination. Thus, GD Pangolin CoV and SARS-CoV-2 similarity could be the consequence of shared ancestry. According to the third hypothesis, two or more close relatives of SARS-CoV-2 may have been affected by some recombination events within a host species. Such a recombination could result in gene exchange between the CoV genomes. During this recombination, some whole genes of the donor CoV genome could be incorporated into the recipient CoV genome either directly (when the orthologous genes are absent in the recipient) or by supplanting in it the existing orthologous genes. This constitutes the *complete gene transfer model* that accounts for the phenomenon of intergenomic recombination [6]. Moreover, the recombination process could lead to the

formation of *mosaic* genes through *intragenic recombination* of the orthologous genes of the donor and recipient CoVs. The term mosaic comes from the pattern of interspersed blocks of sequences with different evolutionary histories. This constitutes the *partial gene transfer model* that accounts for the phenomenon of intragenomic recombination [7]. These models of reticulate evolution have been widely studied in the literature since the beginning of this century [2, 3, 10, 12, 18, 18, 23, 24, 26, 27, 34, 29, 35, 44–46].

In this paper, we provide arguments supporting the third hypothesis of the SARS-CoV-2 origin, according to which the SARS-CoV-2 genome is a chimera of the RaTG13 and GD Pangolin coronaviruses. Such a conclusion is in agreement with the recent results of Xiao et al. [61] and Li et al. [36]. Some authors, however, present evidence that there was not recombination between ancestors of GD Pangolin CoV and RaTG13, suggesting that the similarity pattern between their genomes is rather the result of recombination into RaTG13 from some unknown CoV strains [9].

In their study investigating the origins of SARS, Stavrinides and Guttman [55] highlighted that the SARS-CoV genome is a mosaic of some mammalian and avian virus genomes. These authors also pointed out that recombination between the ancestor viruses of SARS may have happened in the host-determining gene S. However, the work of Stavrinides and Guttman mainly addresses the deep evolutionary origins of the entire SARS CoV clade and has been criticized by some expert in the field [59]. Hu et al. [25] have more recently conducted evolutionary analysis of 11 bat CoV (i.e. SARSr-CoV) genomes discovered in horseshoe bats in the Yunnan province of China and found that they share high sequence similarity to SARS-CoV in the hypervariable N-terminal domain and the RB domain of gene S, as well as in some regions of genes ORF3 and ORF8. Hu et al. also reported that their recombination analysis provided evidence of frequent recombination events within genes S and ORF8 between these bat CoVs, and suggested that the direct progenitor of SARS-CoV may have originated from multiple recombination events between the precursors of different bat CoVs.

Human CoV strains discovered during previous SARS outbreaks have been often hypothesized to pass from bats to humans using intermediate hosts (e.g. civets for SARS-CoV and camels for MERS-CoV) [20], suggesting that SARS-CoV-2 may have also been transmitted to humans this way. It is worth noting, however, that some other studies from scientists in the field, such as Ralph Baric, Zheng-Li Shi and Peter Dazsak, have showed the potential for direct infection of humans from bat strains. They include both serological studies in rural China

near the caves where these bat viruses circulate as well as in vitro/cell culture studies indicating the ability of bat isolates to infect human cells directly [68]. Nevertheless, the discovery of such an intermediate host of SARS-CoV-2, if it existed, is key, as it could shed light on the evolution of this dangerous virus.

At the same time, it is also crucial to retrace the evolution of all genes of the SARS-CoV-2 genome, doing it on a gene-by-gene basis. Evolutionary patterns of various genes of SARS-CoV-2 could be quite different as some of them could be affected by specific horizontal gene transfer and recombination events. These events could witness that the SARS-CoV-2 genome is a mosaic genome obtained via recombination of various virus strains. Our findings suggest that the SARS-CoV-2 genome may be in fact formed via recombination of genomes close to RaTG13 and GD Pangolin CoV genomes, and be a close relative of bat CoV ZC45 and ZXC21.

## Results

In this section, we present a detailed analysis of putative gene transfer and recombination events that were detected in each of 11 main genes of SARS-CoV-2 as well as in the RB domain of the spike protein.

### SimPlot similarity analysis

Our SimPlot analysis (Fig. 1a) conducted with 25 CoV genomes (see the 'Methods' section for a detailed data description) shows that the Wuhan SARS-CoV-2 and RaTG13 genomes share 96.14% of whole-genome identity, while the Wuhan SARS-CoV-2 and GD Pangolin genomes are 90.34% identical. The RaTG13 and GD Pangolin CoV genomes are by far the closest ones to the SARS-CoV-2 genome. For example, only 85.43% of whole-genome identity is shared between the Wuhan SARS-CoV-2 and GX Pangolin CoV genomes. Given such a close resemblance between SARS-CoV-2 and RaTG13, bat is a likely reservoir of origin for SARS-CoV-2, as was the case during previous CoV outbreaks.

Then, we performed a detailed similarity analysis at the level of individual genes to compare the Wuhan SARS-CoV-2 gene sequences with the RaTG13, GD Pangolin CoV group sequences and Bat CoVZ group sequences (Fig. 2). Our gene-by-gene analysis revealed some regions where SARS-CoV-2 was more similar to GD Pangolin CoV than to RaTG13. These regions have been found in genes S (at the RB domain level), ORF3a, M, ORF7a and N, suggesting that some recombination events between GD Pangolin CoV and RaTG13 may have occurred not only in gene S (as it has been reported in previous studies; see [31] and [66, 67]), but also in four other genes of these CoV genomes. Moreover, we found that in some continuous gene regions, specifically in genes ORF1ab,

ORF3a, M and N, the SARS-CoV-2 gene sequences are much more similar to those of the bat CoV ZC45 and ZXC21 viruses than to those of the GD Pangolin CoVs, and sometimes even of RaTG13 (Fig. 2).

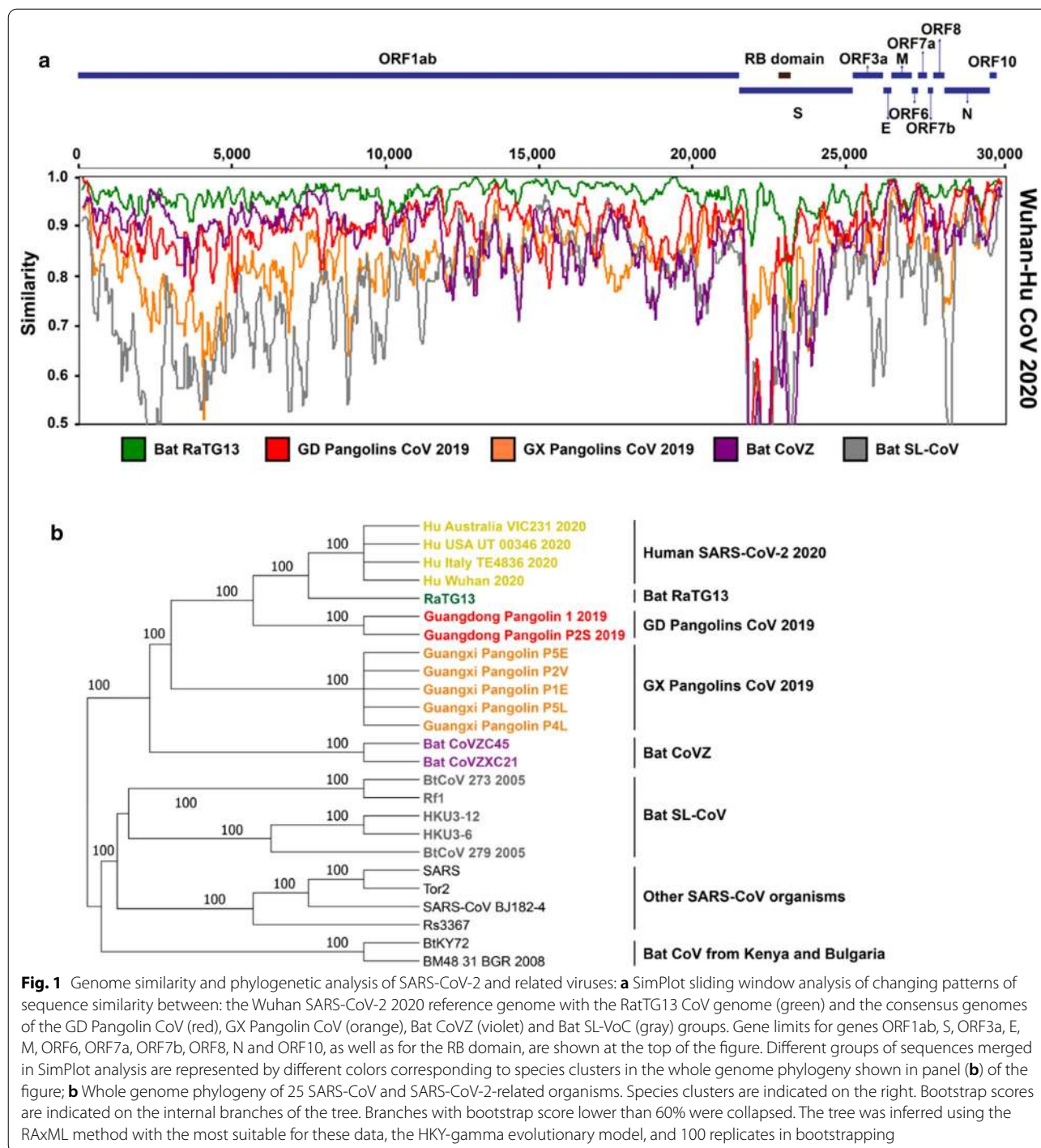
### $\Phi$ -test recombination analysis

We conducted gene-by-gene  $\Phi$ -test recombination analysis [10] to further investigate eventual recent recombination events which may have occurred between orthologous gene sequences of the Wuhan SARS-CoV-2, RaTG13, GD Pangolin 1 CoV, GD Pangolin P2S CoV, bat CoV ZC45 and bat CoV ZXC21 viruses. The  $\Phi$ -test was carried out with different sliding window sizes, varying from 50 to 400 (with a step of 50), and the window progress step of 1 as the window size can affect the test outcome. The results presented in Table 1 are reported for the window size corresponding to the smallest  $p$ -value found for a given gene. The  $p$ -values lower than or equal to the 0.05 threshold were considered as significant. They indicate the presence of recombination in the gene under study.

According to the  $\Phi$ -test (see Table 1), statistically significant recombination events involving these six coronaviruses have been detected in genes ORF1ab, S, ORF3a, ORF7a, ORF8, N and in the whole genome sequences. It is worth noting that recombination events in genes ORF1ab, S and ORF3a were detected with high confidence, with  $p$ -values of 0.0023, 0.0086 and 0.0000113, respectively.

### Horizontal gene transfer and recombination analyses

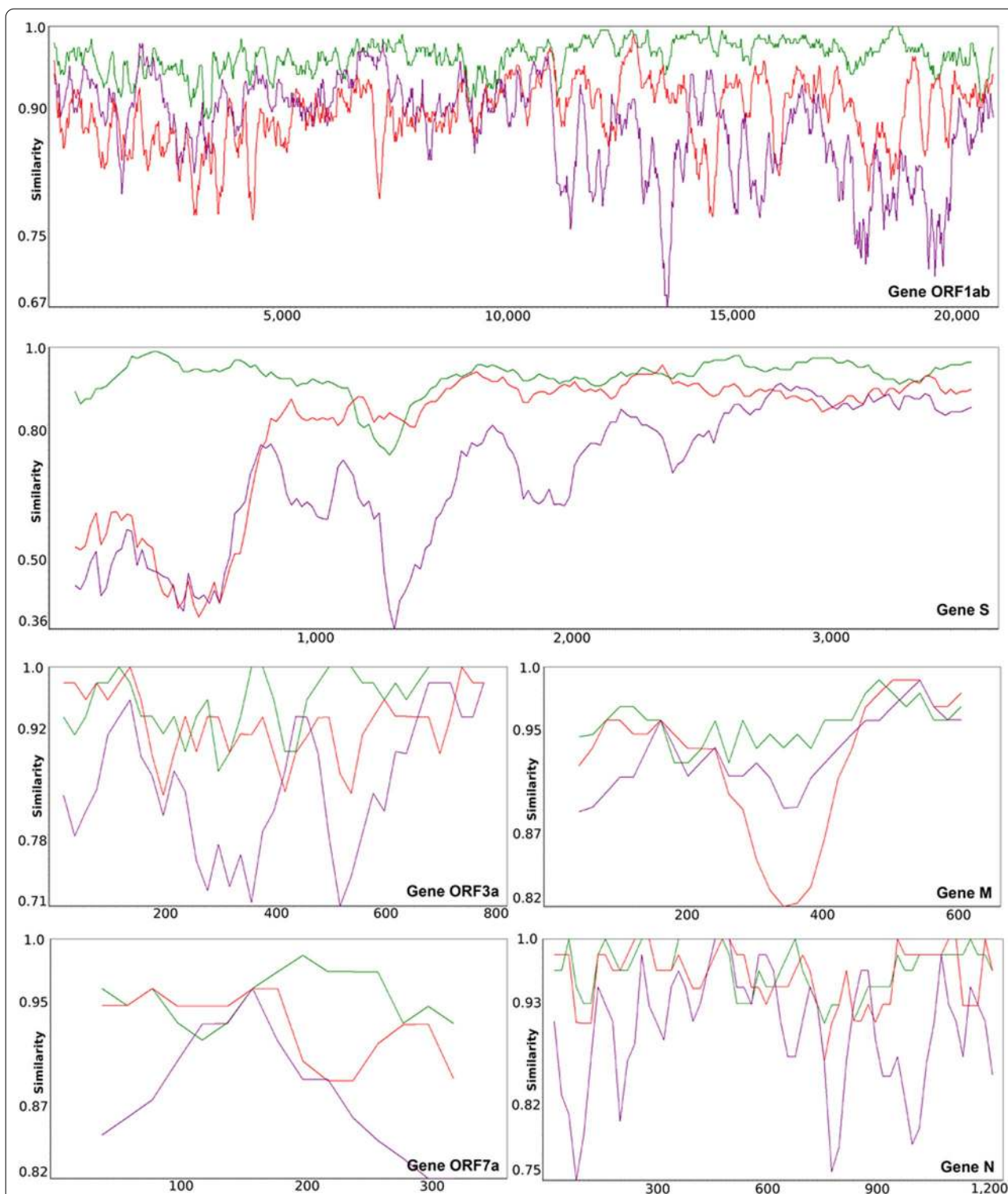
In addition to the SimPlot similarity and the  $\Phi$ -test recombination analyses, we also inferred the whole genome phylogeny (Fig. 1b) and all individual gene trees for 11 main genes of SARS-CoV-2 (Fig. 3a–k) as well as for the RB domain of the spike protein (Figs. 3l and 4a) using the RAxML method [54] with bootstrapping, and then conducted a detailed horizontal gene transfer-recombination analysis (Fig. 3) using the HGT-Detection program available on the T-Rex web server [8]. It is worth noting that the presentation of the whole genome tree could be a bit confusing in our context, as the central premise of this work is that the evolution of coronavirus organisms is driven by the gene transfer and recombination mechanisms, but it remains necessary and serves as a support tree topology to represent the available species groups (Fig. 1b) as well as the detected gene transfer and recombination events (Fig. 3). The HGT-Detection program allows one to infer all possible horizontal gene transfer events for a given group of species by reconciling the species tree (i.e. whole genome tree in our case) with different gene phylogenies built for whole individual genes or some of their regions [7, 14]. The bootstrap



support of the inferred gene transfers was also assessed by HGT-Detection. It is worth noting that the bootstrap support of horizontal gene transfer events (HGT) is usually lower than that of the related branches of the species and gene phylogenies. For example, in order to get an HGT bootstrap support of 100%, gene transfer from cluster C1 to cluster C2 in the species tree must be present

in all gene transfer scenarios inferred from all replicated multiple sequence alignments (MSAs) used in bootstrapping, and clusters C1 and C2 must be neighbor clusters in all gene trees inferred from replicated MSAs [6].

Importantly, each detected horizontal gene transfer event can be interpreted in three ways: (1) It can represent a unique complete or partial HGT event involving



**Fig. 2** Gene-by-gene SimPlot similarity analysis performed to compare gene sequences of the Wuhan SARS-CoV-2 2020 reference genome with those of the RatTG13 genome (green), GD Pangolin CoV consensus genome (red) and Bat CoVZ consensus genome (violet). Similarity plots are presented for genes ORF1ab, S, ORF3a, M, ORF7a and N that encompass the most important overlaps between the RaTG13, GD Pangolin CoV and Bat CoVZ similarity curves

**Table 1** The results of the  $\Phi$  recombination test carried out for gene and whole genome sequences of Wuhan SARS-CoV-2, RaTG13, GD Pangolin 1 CoV, GD Pangolin P2S CoV, bat CoV ZC45 and bat CoV ZXC21

Region	$\Phi$ -test result (p-value)	Recombination detected (yes/no)	Window size
Gene ORF1ab	$2.23 \times 10^{-3}$	Yes	150
RB domain	Too short	–	
Gene S	$8.60 \times 10^{-3}$	Yes	200
Gene ORF3a	$1.13 \times 10^{-5}$	Yes	100
Gene E	0.56	No	200
Gene M	0.226	No	100
Gene ORF6	Too short	–	
Gene ORF7a	0.00155	Yes	200
Gene ORF7b	Too short	–	
Gene ORF8	0.0453	Yes	200
Gene N	0.024	Yes	50
Gene ORF10	Too short	–	
Whole genomes	0.0156	Yes	200

distant species, as discussed above; (2) it can represent the phenomenon of parallel evolution, which the involved species might have undergone; and (3) it can also represent the situation where a new species (i.e. a gene transfer recipient) was created by recombination of the donor species genome with the genome of a recipient neighbor in the species phylogeny (this was potentially the case of the gene transfers from GD Pangolin CoV to SARS-CoV-2, which is a neighbor of RaTG13 in the species phylogeny, found for genes S and N; see our gene-by-gene analysis below).

We will now discuss the evolution of 11 main genes of SARS-CoV-2 as well as that of the RB domain of the spike protein with emphasis on the specific horizontal gene transfer and recombination events detected for each of them.

### Evolution of gene ORF1ab

Gene ORF1ab occupies more than two thirds of coronavirus genomes. It encodes the replicase polyprotein, being translated from ORF1a (11826–13425 nt) and ORF1b (7983–8157 nt) [60], and plays an important role in virus pathogenesis [19].

The bootstrap support of the branches of the ORF1ab gene phylogeny is commonly very high, except for the branch connecting the clusters of GD and GX Pangolin CoVs (Fig. 3a). For this gene, we found a horizontal gene transfer-recombination event between the ancestors of the clusters of the Bat CoVZ organisms and the cluster including the RaTG13 and SARS-CoV-2 viruses. This

event affects a half of this long gene, occurring in gene region [1–11630] with bootstrap support of 61.5% (as this transfer does not affect the whole gene sequence, it is called a *partial HGT*). A transfer of the whole ORF1ab gene (i.e. a *complete HGT*) from the bat RF1 CoV to bat BtCoV 279 2005 has been detected with bootstrap score of 84.6%. The transfer between the cluster of the bat BTKY72 and BM48 31 BGR 2008 coronaviruses and the cluster involving SARS-CoV-2, RaTG13 and GD Pangolin CoVs has been detected on gene region [16326–17626] with bootstrap support of 61.5%. Finally, a deep phylogeny transfer from CoV organisms from the bottom part of the tree (i.e. SARS-like 2003–2013 viruses) to the cluster of GX Pangolins CoVs has been found in a relatively short region [1249–1421] with a high bootstrap score of 97.9%.

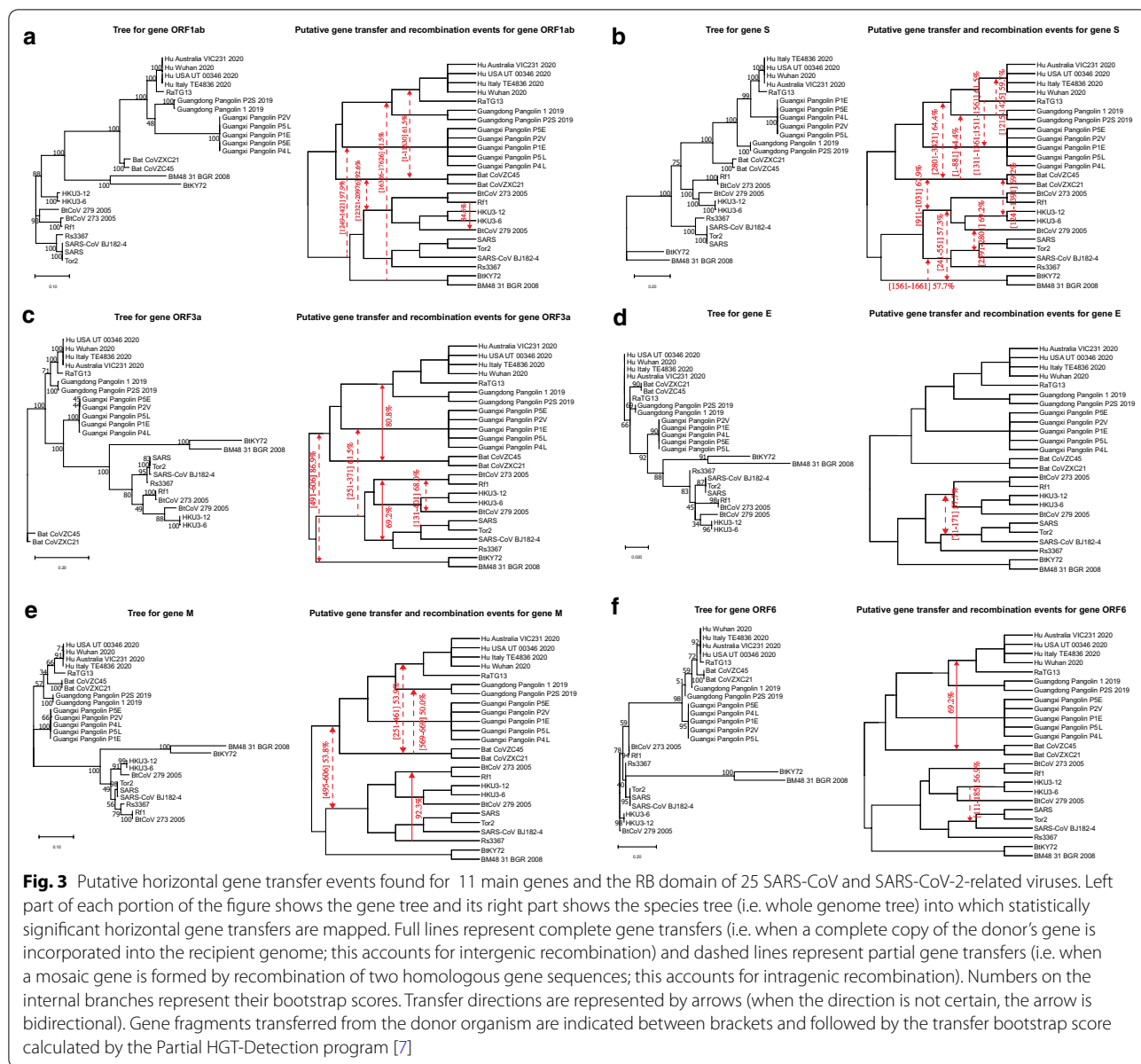
### Evolution of gene S

Gene S is among the most important coronavirus genes since it regulates the ability of the virus to overcome species barriers and allows interspecies transmission from animals to humans [65]. S proteins are responsible for the “spikes” present on the surface of coronaviruses, giving this virus family a specific crown-like appearance. These proteins are type I membrane glycoproteins with signal peptides used for receptor binding [60]. They play a crucial role in viral attachment, fusion and entry, being a target for development of antibodies, entry inhibitors and vaccines [58].

As shown on the SimPlot diagram (Figs. 1a and 2), this gene has the most variable sequences among all genes of coronavirus genomes. The bootstrap scores of internal branches of the phylogeny of gene S are all very high (Fig. 3b). For this gene, a partial gene transfer was found in region [1–881] between the Bat CoVZ group and the cluster of GD Pangolin CoVs with bootstrap support of 64.4%. Another partial transfer from GD Pangolin CoVs to SARS-CoV-2 was found on the interval [1215–1425] with bootstrap score of 59.7%. This transfer corresponds to the RB domain. Partial transfers were detected from RaTG13 to GX Pangolin CoVs in regions [1311–1361] and [1511–1561] with an average bootstrap score of 61.5%. Finally, another partial transfer was found between the bat CoVZ group and the cluster including BtCoV 273 2005, Rf1, HKU3-12, HKU3-6 and BtCoV 279 2005.

### Evolution of gene ORF3a

The protein 3a is unique to SARS-CoV and SARS-CoV-2. It is essential for disease pathogenesis. In SARS-related CoVs, it forms a transmembrane homotetramer complex with ion channel function and modulates virus release [39]. The sequences of ORF3a are as highly variable, almost as those of gene S. The bootstrap scores of



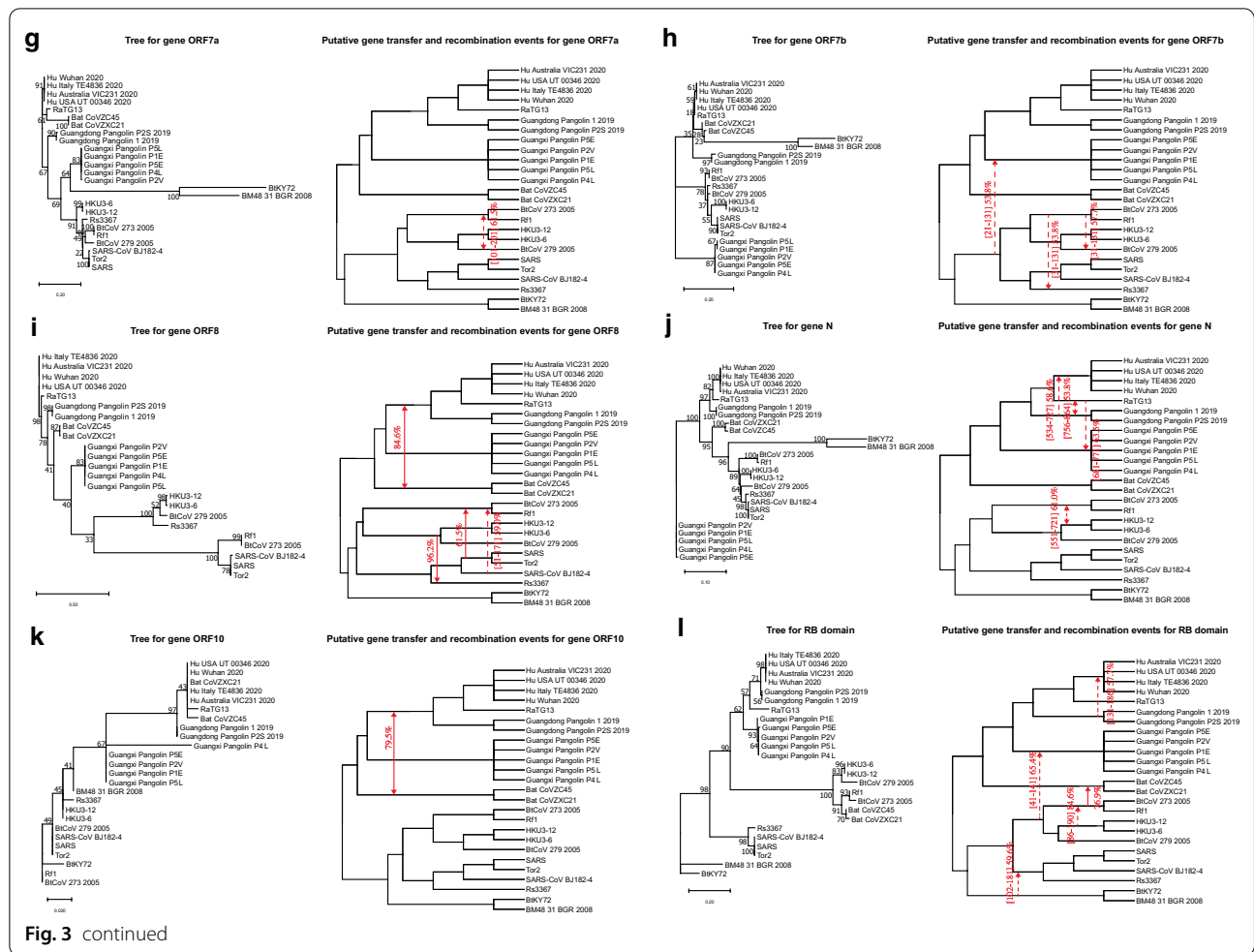
the ORF3a gene tree (Fig. 3c) are usually high, except for that of the branch connecting the cluster of GD Pangolin CoVs and the cluster of RatTG13 and SARS-CoV-2 (71%), and the branch connecting clusters of some old CoVs at the bottom part of the tree.

For this gene, a full gene transfer was found between the cluster of SARS-CoV-2, RaTG13 and GD Pangolin CoVs, and the group of Bat CoVZ viruses, with statistical significance of 80.8%. A partial gene transfer was detected between the lower part of the tree (i.e. SARS-like 2003–2013 viruses), excluding bat viruses from Kenya and Bulgaria (i.e. BtKY72 and BM48 31 BGR 2008), and the GX Pangolin CoVs in region [251–371] with bootstrap

support of 61.5%. The remaining identified transfers were a deep phylogeny transfer and two transfers between CoVs from the previous SARS outbreak.

**Evolution of gene E**

This gene is among the shortest in the SARS-CoV-2 genome. The protein E is a small transmembrane protein associated with the envelope of coronaviruses. It is well conserved among all CoVs (Fig. 3d). Gene E is usually not a good target for phylogenetic analysis because of its short sequence length [60]. The bootstrap scores of its gene phylogeny (Fig. 3d) are mostly mediocre. A single gene transfer-recombination event has been found for



this gene. It affects region [71–171] and involves CoVs from the previous SARS outbreak.

**Evolution of gene M**

This gene is particularly important since it is responsible for assembly of new virus particles [28]. Bootstrap scores of the gene tree are on average much lower than those of genes ORF1ab, S and ORF3a (Fig. 3e).

Here, we detected two partial gene transfers involving the viruses of the Bat CoVZ group and affecting: (1) the SARS-CoV-2 and RaTG13 CoVs in region [251–461] with bootstrap support of 53.9%, and (2) the GD Pangolin CoVs in region [569–669] with bootstrap support of 50%. Moreover, a complete gene transfer from Rs367 to the cluster of Rf1 and BtCoV 273 2005 was found for this gene with bootstrap score of 92.3%.

**Evolution of gene ORF6**

Gene ORF6 impacts the expression of transgenes [47]. The gene phylogeny of ORF6 (Fig. 3f) has multiple internal branches with low bootstrap support. For this gene,

we found a complete gene transfer between the Bat CoVZ group and the cluster including SARS-CoV-2 and RaTG13 with bootstrap support of 69.2%. Furthermore, a partial gene transfer between HKU3-12 CoV and the SARS-CoV cluster in region [111–185] with bootstrap score of 67.7% was also found for this gene.

**Evolution of genes ORF7a and ORF7b**

The proteins encoded by coronavirus genes ORF7a and 7b have been demonstrated to have proapoptotic activity when expressed from cDNA [52]. The phylogenies of these short genes are very unresolved (Fig. 3g, h). A gene transfer-recombination event found for ORF7b involves CoVs from the lower part of the tree, excluding CoVs of Kenyan and Bulgarian bats, and the GX Pangolin CoVs. It occurred in region [21–131] with bootstrap score of 53.8%. Interestingly, a transfer affecting the bat CoVs related to the previous SARS outbreak was found in both of these genes.



### Evolution of gene ORF8

It has been recently shown that the SARS-CoV-2 viral protein encoded from gene ORF8 shares the least homology with SARS-CoV among all the viral proteins, and that it can directly interact with MHC-I molecules, significantly downregulating their surface expression on various cell types [66, 67]. The gene phylogeny of ORF8 has several internal branches with low bootstrap scores (Fig. 3i). It has been established that ORF8 protein of SARS-CoV has been acquired through recombination from SARS-related coronaviruses from greater horseshoe bats [32]. A complete gene transfer was found for this gene between the cluster of SARS-CoV-2, RaTG13 and GX Pangolin CoVs and the bat CoVZ group with bootstrap support of 84.6%. The other detected transfers concerned the viruses of the SARS-CoV group.

### Evolution of gene N

The nucleocapsid protein (N) is one of the most important structural components of SARS-related coronaviruses. The primary function of this protein is to encapsulate the viral genome. It is involved in the formation of the ribonucleoprotein through interaction with the viral RNA [28]. Several studies report that the protein N interferes with different cellular pathways, thus being a crucial regulatory component of the virus as well [56]. The phylogeny of gene N (Fig. 3j) is usually well resolved, except for two clusters in the SARS-CoV part of the tree with bootstrap scores of 64% and 45%. Our SimPlot analysis showed that the SARS-CoV-2 gene sequence of gene N is almost as similar to the RaTG13 gene sequence as it is to the GD Pangolin gene sequence. Precisely, for gene N, the Wuhan SARS-CoV-2 and RaTG13 viruses share 96.9% of the whole-gene identity, while the Wuhan SARS-CoV-2 and GD Pangolin CoV gene sequences are 96.19% identical.

Three statistically significant gene transfer-recombination events have been detected for this gene. The most interesting of them is the partial gene transfer from the cluster of GD Pangolin CoVs towards the cluster of SARS-CoV-2 found in region [534–727] with bootstrap support of 58.6%. Another partial transfer, between RaTG13 and GD Pangolin CoVs, was detected in region [756–864] with bootstrap support of 53.8%. Finally, a partial transfer between the cluster containing the BtCoV 273 2005 and Rf1 viruses and the cluster of HKU CoVs was found in region [551–721] with bootstrap score of 61.0%.

### Evolution of gene ORF10

The protein ORF10 of SARS-CoV-2 includes 38-amino acids and its function is unknown [63]. The phylogeny of this short gene is not well resolved (Fig. 3k) with several

internal branches having bootstrap support under 50%. The only complete gene transfer-recombination event detected for this gene affects the Bat CoVZ virus group and the cluster of SARS-CoV-2, RaTG13 and GD Pangolin CoVs. Its bootstrap score is 79.5%.

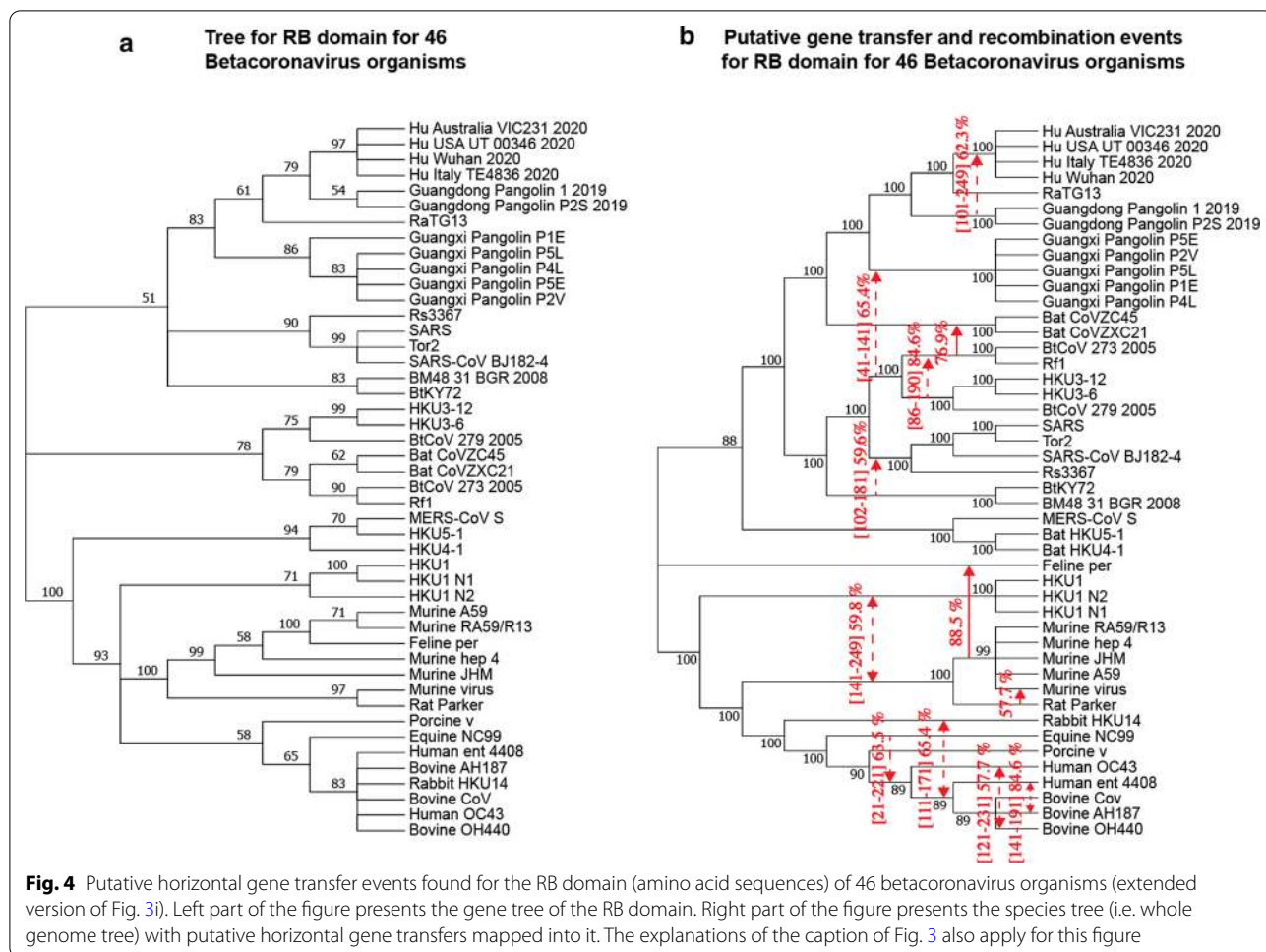
### Evolution of the RB domain

We carried out an independent analysis of the main evolutionary events characterizing the RB domain of the spike protein because of its major evolutionary importance. The S protein mediates the entry of the virus into the cells of the host species by binding to a host receptor through the RB domain located in its S1 subunit, and then merging the viral and host membranes in the S2 subunit. As SARS-CoV, SARS-CoV-2 also recognizes ACE2 as its host receptor binding to the S protein of the virus. Thus, the RB domain of the SARS-CoV-2 S protein is the most important target for the development of virus attachment inhibitors and vaccines [58].

We first studied the evolution of the RB domain for the 25 original organisms that are strongly phylogenetically related to SARS-CoV and SARS-CoV-2. The RB domain amino acid phylogeny (Fig. 3l) commonly exhibits high bootstrap scores of its internal branches, except for the root branch of the cluster of GD Pangolin CoVs (56%), the branch separating RaTG13 from the cluster of SARS-CoV-2 and GD Pangolin CoVs (57%), and the branch separating the cluster of GX Pangolin CoVs from the cluster including the SARS-CoV-2, GD Pangolin CoV and RaTG13 viruses (62%). Thus, the location of both pangolin clusters and that of RaTG13 are the most uncertain in this tree.

Our similarity analysis showed that SARS-CoV-2 and RaTG13 share 89.47% of the whole-gene identity, while the SARS-CoV-2 and GD Pangolin CoV RB domain amino acid sequences are 97.36% identical. Consequently, some exchange of genetic material between the clusters of GD Pangolin CoVs and SARS-CoV-2 would be expected here. In fact, a gene transfer from GD Pangolin CoVs to SARS-CoV-2 was detected in region [131–186] with bootstrap support of 57.7%. Another transfer found for this tree was that from the cluster containing BtCoV 273 2005, Rf1, HKU3-12, HKU3-6 and BtCoV 279 2005 to GX Pangolin CoVs in region [41–141] with bootstrap support of 65.4%.

Moreover, we inferred an extended version of the RB domain tree, adding to it 21 coronavirus organisms (Fig. 4a) labeled as common cold CoV in the GISAID coronavirus tree [53] and other coronavirus organisms available in GenBank [49]. These additional CoVs include Human betacoronavirus 2c EMC/2012 (i.e. MERS-CoV S), Human coronavirus (i.e. HKU1 and its isolates N1 and N2), Human coronavirus OC43 and

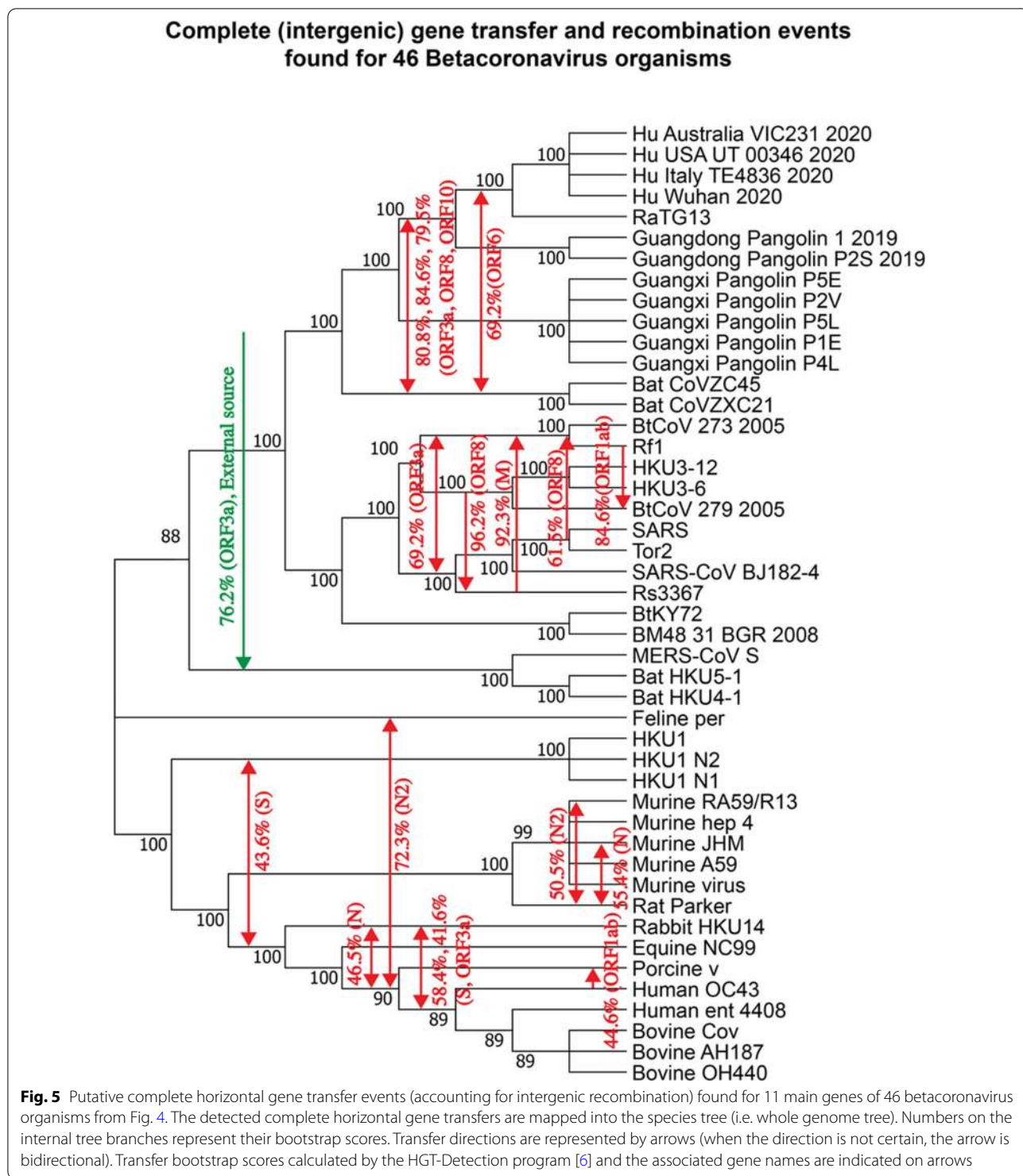


Human enteric coronavirus strain 4408. This extended analysis allowed us to discover some additional intergenic and intragenic gene transfer-recombination events involving human-related coronaviruses, including the exchange of genetic material between: (1) Murine CoVs and HKU1-related viruses, i.e. a complete gene transfer detected with high bootstrap score of 88.5%, and a partial transfer detected with bootstrap score of 59.8%, (2) Equine CoV and the cluster including Human OC43 and Enteric CoVs, i.e. a partial gene transfer with bootstrap score of 63.5%; (3) Rabbit HKU14 CoV and the cluster including the Enteric human CoV, and the three bovine CoVs, i.e. a partial gene transfer with bootstrap score of 65.4%; (4) Human OC43 CoV and Bovine OH440 CoV, i.e. a partial gene transfer with bootstrap score of 57.7%; and (5) Human Enteric CoV and Bovine AH187 CoV, i.e. a partial gene transfer with bootstrap score of 84.6%. Interestingly, no any gene transfer-recombination event between the viruses of the upper part of the species tree, containing SARS-related coronaviruses (Fig. 4b), and the lower

part of this tree, containing MERS-related, HKU1-related, OC43 CoV-related and Enteric CoV-related viruses, has been detected.

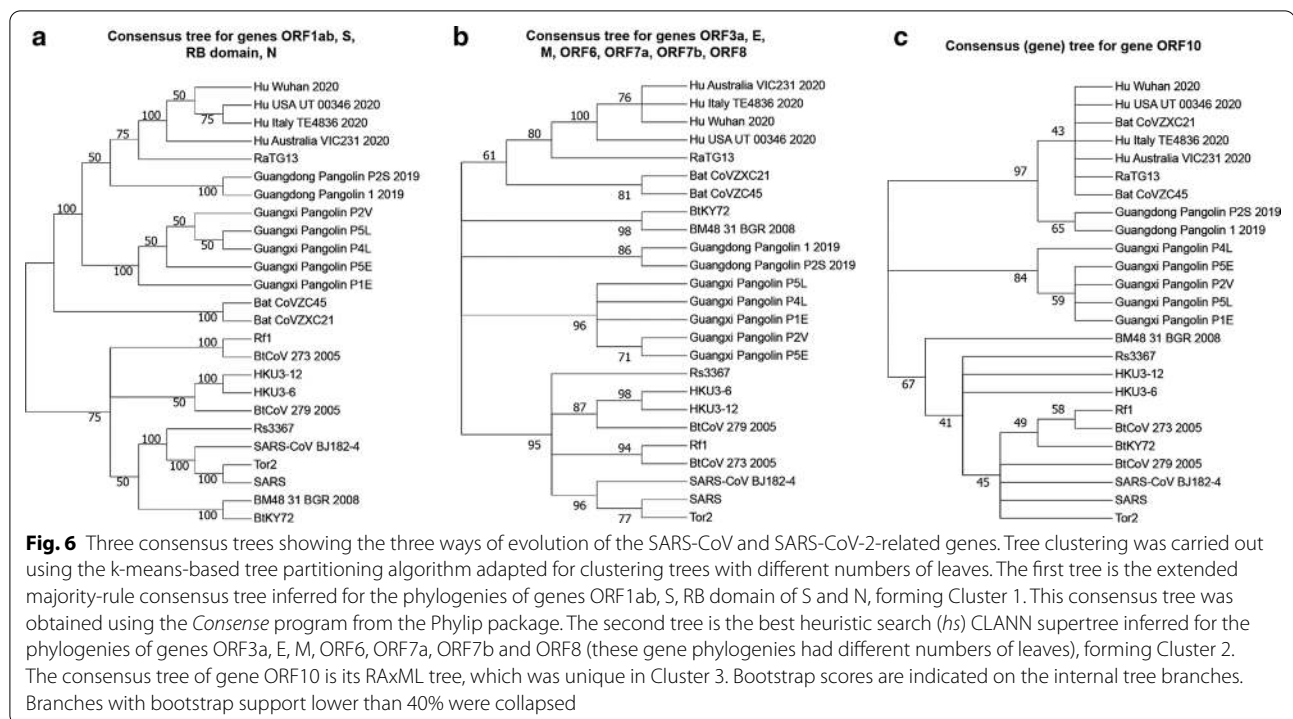
**Analysis of intergenic recombination in 46-species phylogeny**

We also carried out an analysis of intergenic (complete gene transfers) recombination events in the 46-species phylogeny discussed in the previous section. This extended analysis allowed us to discover some important gene transfer-recombination events that affected the lower part of this larger phylogeny (see Fig. 5), including the 25 SARS-CoV-2-related viruses from Fig. 1b as well as MERS-CoV, HKU1 coronavirus strains, Human coronavirus OC43, Human enteric coronavirus, and different Bat, Rat Parker, Murine, Feline, Equine, Bovine, Porcine and Rabbit coronavirus strains. The analysis was conducted using the HGT-Detection program [6]. The transfers of the higher part of the tree presented in Fig. 5 are the complete gene transfers from Fig. 3. The



most significant transfers found for the lower part of the tree include those between: (1) Murine JHM and Rat Parker coronaviruses found for gene N (bootstrap score of 55.4%); (2) Feline CoV and the ancestor of the cluster comprising Porcine, Human OC43, Human enteric

and three Bovine CoVs for gene N2 (bootstrap score of 72.3%); (3) Rabbit CoV and the ancestor of the cluster comprising Human OC43, Human enteric and three Bovine CoVs for gene S (bootstrap score of 58.4%); and, finally, (4) an interesting complete transfer affecting the



cluster of MERS-CoV S and two of its close relatives, i.e. bat coronaviruses HKU-4 and HKU-5, for gene ORF3a (bootstrap score of 76.4%). The gene transfer accounting for this recombination event most likely stems from an external source (a CoV organism which is absent in the tree). It is marked by a green arrow in Fig. 5. It is worth noting that no intergenic recombination events between coronaviruses from the higher and lower parts of the extended CoV species phylogeny tree were discovered.

#### Cluster analysis of CoV gene phylogenies

Finally, we also carried out the CoV gene tree clustering to identify genes following similar evolutionary patterns (i.e. having similar gene tree topologies). This analysis was performed using the k-means-based tree clustering algorithm adapted for clustering phylogenies with different numbers of leaves [57] as some gene trees contained less than 25 species (see the ‘Methods’ section). Our results indicate that coronavirus genes followed three different patterns of evolution as the phylogenies of 11 CoV genes and that of the RB domain of the spike protein were partitioned into 3 disjoint clusters. Figure 6 presents the consensus trees of the detected tree clusters. The first of them (Fig. 6a) obtained using the *Consense* program from the Phylip package [17] is the extended majority rule consensus tree of the gene phylogenies of ORF1ab, S, RB domain of S and N. The supertree inferred by CLANN [13]

(in this case we could not use *Consense* since the species BM48\_31\_BGR\_2008 and BtKY72 were missing in the gene phylogeny of ORF8) for the gene phylogenies of ORF3a, E, M, ORF6, ORF7a, ORF7b and ORF8 is shown in Fig. 6b. The consensus tree of the third cluster, containing a unique representative—gene tree of ORF10, is its RAxML gene phylogeny (Fig. 6c).

#### Discussion

Recombination is a prevalent process contributing to diversity of most viruses, including SARS-CoV-2 and other betacoronavirus organisms. It allows viruses to overcome selective pressure and adapt to new hosts and environments [48]. In this work, we conducted thorough gene-by-gene horizontal gene transfer and recombination analysis of SARS-CoV-2-related viruses. Even though gene borders are not always a natural demarcation of regions where recombination might occur, e.g. two large portions of gene ORF1ab exhibit different evolutionary histories, the performed sliding window analysis allowed us to treat each considered genetic segment as an independent gene region having its own evolutionary history. We first performed a comparative analysis of four strains of SARS-CoV-2 with 21 close members of the SARS-CoV family, which revealed multiple horizontal gene transfer and recombination events among these coronavirus organisms. The most striking of them were statistically significant gene transfers from Guangdong Pangolin

CoV to SARS-CoV-2 found in gene S (i.e. this transfer most likely accounts for a putative recombination event between Guangdong Pangolin CoV and RatTG13 in the RB domain region of the spike protein) and gene N (i.e. this transfer most likely accounts for a putative recombination event between Guangdong Pangolin CoV and RatTG13 in region [534–727] of this gene). These findings are in support of the hypothesis that SARS-CoV-2 genome is a chimera resulting from recombination of the RatTG13 and Guangdong Pangolin CoV genomes. According to some recent studies, the discovery of SARS-CoV-like coronaviruses from pangolins with similar RB domain, provides the most parsimonious explanation of how SARS-CoV-2 could acquire it via recombination or mutation [1, 36, 61]. The fact that the highlighted gene transfer-recombination events were detected only between Guangdong Pangolin CoV and SARS-CoV-2, but not between Guangxi Pangolin CoV and SARS-CoV-2, is another argument in favor of a mosaic recombinant origin of the SARS-CoV-2 genome, and against the parallel evolution paradigm. The confirmed recombination events in genes S and N (Fig. 3b and j) as well as the SimPlot recombination analysis of genes S, ORF3a, M, ORF 7a and N (Fig. 2) suggest that Guangdong pangolin is a likely intermediate host of SARS-CoV-2, on its way of transmission from bats to humans. The bat RatTG13 virus strain could infect a Guangdong pangolin, which was already a bearer of its own CoV, probably similar to that found in Guangxi pangolins. The two CoV genomes could then recombine and the resulting recombinant evolve into a SARS-CoV-2 mosaic strain which was then transmitted to humans. Furthermore, we also discovered multiple gene exchanges between the cluster including the bat CoVZC45 and CoVZXC21 viruses, and the cluster including the SARS-CoV-2 strains and RatTG13. Some statistically significant gene transfer-recombination events between these CoV clusters were found in 6 of 11 coronavirus genes, namely in ORF1ab, ORF3a, M, ORF6, ORF8 and ORF10 (according to the gene transfer-recombination analysis conducted with HGT-Detection; see Fig. 3), and in genes ORF1ab, S, ORF3a, ORF7a, ORF8 and N (according to recombination analysis conducted with  $\Phi$ -test; see Table 1). These findings confirm that not only the RatTG13 and GD Pangolin CoVs have influenced the evolution of SARS-CoV-2, but also the bat CoVZC45 and CoVZXC21 coronaviruses or their common ancestor.

The intergenic recombination analysis of the extended 46-species coronavirus phylogeny allowed us to detect eight additional statistically significant gene transfer events affecting the lower part of the extended coronavirus phylogeny. Among them, we need to highlight a complete gene transfer affecting the cluster of MERS-CoV S and two bat coronaviruses (HKU-4 and HKU-5),

stemming from an external source, which was found for gene ORF3a (see Fig. 5). No recombination events between coronaviruses from the higher and lower parts of the extended CoV phylogeny were found (see Figs. 4b and 5). This finding suggests that the gene transfer-recombination history of coronaviruses from the SARS-CoV-2 cluster (common cold CoVs from the higher part of the extended 46-species phylogeny) and that of coronaviruses from the cluster including MERS-related, HKU1-related, OC43-related and Enteric-related CoVs (lower part of the extended 46-species phylogeny) can be studied separately.

It is worth noting that we have also tried to add to our data set some extra SARS-CoV-2 genomes (in addition to the four originally considered SARS-CoV-2 genomes from Wuhan in China, Italy, Australia and USA), but realized that it did not lead to discovery of any further gene transfer-recombination events in which these extra SARS-CoV-2 organisms could be involved. This happens because of a very high sequence similarity between the available SARS-CoV-2 genomes. For example, the root genomes of the two main SARS-CoV-2 lineages, A and B (according to a recent SARS-CoV-2 sequence classification proposed in [50]), share 99.89% of whole genome identity (we measured it between the following coronavirus organisms: Lineage\_A\_EPI\_ISL\_406801 and Lineage\_B\_MN908947.3; see [50] for more details), while the Hu\_Wuhan\_2020 and Hu\_Australia\_VIC231\_2020 SARS-CoV-2 genomes analyzed in our work share 99.98% of whole genome identity. The difference between the existing SARS-CoV-2 genomes can be mainly explained by the presence of particular sets of mutations, with respect to the root sequence. These mutations are few in numbers and are usually not contiguous. They can be detected by a simple genome comparison and do not necessitate the use of horizontal gene transfer and recombination detection methods. It should be mentioned that when extra SARS-CoV-2 genomes were added to the 46 betacoronavirus organisms analyzed in our work, they were always involved into the same gene transfer-recombination events as the four originally considered SARS-CoV-2 genomes.

We also conducted a cluster analysis of the coronavirus gene trees in order to identify genes with similar evolutionary histories. This analysis revealed the presence of three clusters of gene phylogenies: the first cluster includes the phylogenies of genes ORF1ab, S, RB domain and N, the second cluster includes the phylogenies of genes ORF3a, E, M, ORF6, ORF7a, ORF7b and ORF8, and the third cluster contains only the phylogeny of gene ORF10. For example, the phylogenies of genes S and N, whose evolution was affected by the highlighted

recombination events between Guangdong Pangolin CoV and RaTG13, were assigned to the same cluster.

## Conclusion

The main finding of our work is a detailed list of statistically significant horizontal gene transfer and recombination events inferred for 11 main genes and the RB domain of the spike protein of SARS-CoV-2 and related betacoronavirus genomes (see Figs. 3, 4 and 5). The main advantages of the conducted gene transfer and recombination analysis, compared to other recent works in the field [9, 31, 66, 67], is that it allowed us not only to identify genes and genomes that have been affected by recombination, but also to determine the donor and recipient organisms for each detected recombination event, to find out whether this event was intergenic or intragenic, and to assess its statistical significance via a bootstrap score. Our detailed horizontal gene transfer and recombination analysis was conducted for each of 11 main genes of the SARS-Cov-2 genome, involving 46 betacoronavirus organisms. The obtained results (see Figs. 1, 2, 3, 4 and 5) suggest that SARS-Cov-2 could not only be a chimera resulting from recombination of the bat RaTG13 and Guangdong pangolin coronaviruses but also a close relative of the bat CoV ZC45 and ZXC21 virus strains. They also indicate that a Guangdong pangolin may be an intermediate host of SARS-CoV-2 prior to its transmission to humans. Furthermore, our topology-based clustering analysis of coronavirus gene trees revealed a three-way evolution of SARS-CoV-2 genes (see Fig. 6).

It is worth mentioning that some incongruencies among horizontal gene transfer and recombination detection methods may exist when applied to the same data [5]. In this work, we used three different methods for detecting gene transfer and recombination events ( $\Phi$ -test recombination analysis of Bruen et al. [10], intergenic HGT analysis of Boc et al. [6], and both intergenic and intragenic HGT analysis of Boc and Makarenkov [7]) to study the evolution SARS-CoV-2 and related betacoronaviruses. While the method of horizontal gene transfer analysis of Boc et al. [6] is based on the comparison of the species and gene tree topologies, that of Boc and Makarenkov [7] conducts a sliding window analysis of aligned gene sequences. Both of them make part of a phylogenetic approach. In the future, it would be interesting to validate our horizontal gene transfer and recombination results using composition-based ("parametric") horizontal gene transfer detection methods [51].

## Methods

### Data description

We explored the evolution of 25 coronavirus organisms, including a cluster of four SARS-CoV-2 genomes (from

Wuhan in China, Italy, Australia and USA, taken from different clusters of the GISAID human coronavirus tree available at: <https://www.gisaid.org>; [53]), two GD Pangolin CoV genomes (obtained from dead Malayan pangolins during an anti-smuggling operation in the Guangdong province of China) and five Guangxi Pangolin CoV genomes (obtained from the Beijing Institute of Microbiology and Epidemiology). We also included in our analysis the RaTG13 bat CoV genome from *Rhinolophus affinis* from the Yunnan province of China, along with the cluster of two Bat CoVZ organisms, comprising the bat CoV ZC45 and ZXC21 viruses, collected in the Zhejiang province of China in 2018 and five bat CoV genomes sampled in bats across multiple provinces of China from 2006 to 2010 (denoted as BtCoV 273 2005, Rf1, HKU3-12, HKU3-6 and BtCoV 279 2005 in the whole genome CoV phylogeny, see Fig. 1b). Finally, we also considered the SARS-CoV-related genomes from the first SARS outbreak (i.e. human SARS, Tor2, SARS-CoV BJ182-4 CoVs and bat Rs3367 CoV found in *Rhinolophus sinicus*) and two CoV strains coming from bats found in Kenya and Bulgaria (BtKY72 and BM48 31 BGR 2008). Most of these CoV genomes have been originally considered by Lam et al. [31]. Moreover, for an extended analysis of putative gene transfer-recombination events affecting the RB domain of the spike protein and intergenic recombination events (complete gene transfers) affecting all the genes under study, we considered 21 additional coronavirus organisms, including viruses labeled as common cold CoVs in the GISAID coronavirus tree [53] and other CoV organisms studied by Prabakaran et al. [49], they are available in GenBank [4] at: <https://www.ncbi.nlm.nih.gov/Structure/cdd/PF09408>). Additional file 1: Table 1 in Additional Material reports full organism names, host species and GenBank or GISAID accession numbers for all CoV genomes analyzed in this study.

We first carried out the SimPlot [38] similarity analysis of coronaviruses most closely related to SARS-CoV-2, comparing the Wuhan SARS-CoV-2 reference genome to a consensus genomes of five CoV groups (GD Pangolin CoVs, GX Pangolin CoVs, RaTG13, Bat CoVZ and Bat SL-CoV, see Fig. 1a). The GD Pangolin group in our analysis consisted of two Guangdong Pangolin CoVs available in GISAID (GD Pangolin 1 and GD Pangolin P2S in Fig. 1b). This explains differences in the results of our SimPlot similarity analysis with the results of Lam et al. [31], who considered only the first of these GD Pangolin CoVs in their SimPlot analysis. To avoid possible inconsistency and program crashes during the SimPlot similarity analysis [38],  $\Phi$ -test recombination analysis [10] and horizontal gene transfer detection [7], we replaced missing nucleotides in the low-coverage regions of the GD Pangolin P2S CoV genome by the corresponding nucleotides

of the GD Pangolin 1 CoV genome. This allowed us to better highlight intersections between the GD Pangolin CoV group similarity curve and the RaTG13 similarity curve (see Figs. 1a and 2), which may indicate the presence of gene transfer-recombination events between the GD Pangolin and RaTG13 coronaviruses.

Multiple sequence alignments for all gene and genome CoV sequences (in the Fasta format) used in this study as well as all inferred phylogenetic trees (in the Newick format) are available at: [http://www.info2.uqam.ca/~makarenkov\\_v/Supplementary\\_Material.zip](http://www.info2.uqam.ca/~makarenkov_v/Supplementary_Material.zip).

### Methods details

The VGAS (Viral Genome Annotation System) tool [64], designed to identify automatically viral genes and perform gene function annotation, was used to validate all CoV genes extracted from GenBank and GISAID. Multiple sequence alignments for 11 CoV genes of the 25 original, and then 46 (for an extended analysis), betacoronavirus organisms (nucleotide sequences), and for the RB domain of the spike (S) protein (amino acids), were carried out using the MUSCLE algorithm [16] with default parameters of the MegaX package (version 10.1.7) [30]. These alignments were used to infer gene trees presented in Figs. 3 and 4 (left part of each portion of the figure). The whole genome CoV sequences for the original 25, and then 46 (for an extended analysis), betacoronavirus organisms were aligned in MegaX using the same version of MUSCLE. The whole genome alignments were used to infer species trees (see Figs. 1b and 4b). The alignment accuracy for all gene and genome alignments was verified manually base by base. The GBLOCKS tool (version 0.91b; [11]) from the Phylogeny.fr web server [15] was used to eliminate sites with large proportions of gaps. The less stringent correction option of GBLOCKS was used.

The maximum likelihood (ML) gene and genome phylogenies were inferred using the RAxML algorithm (version v0.9.0; [54]). Each tree was constructed under the best-fit DNA/amino acid substitution model found using MegaX, and available in RAxML, for the corresponding multiple sequence alignment. The best available substitution model for genes ORF1ab, S, N and the whole genomes was (GTR+G+I), for genes ORF3a, E, ORF6, ORF7a it was (HKY+G), for gene ORF7b it was (HKY+I), for gene ORF8 it was (HKY+G+I), for gene ORF10 it was (JC), and for the RB domain it was (WAG+G) (see Additional file 1: Table 2). In each case, the bootstrap scores of internal branches of all phylogenies were calculated using 100 bootstrap replicates. All gene and genome trees were originally drawn in MegaX.

The Partial HGT-Detection program [7] from the T-Rex web server [8] was used to infer directional horizontal gene transfer-recombination networks for 11 CoV

genes and the RB domain of the spike protein (see Figs. 3 and 4b). Rooted ML genome tree of 25, and then 46, CoV organisms, playing the role of the species tree, and multiple sequence alignments (for 11 CoV genes and the RB domain of the spike protein) provided by MUSCLE, were used as input parameters of the Partial HGT-Detection and HGT-Detection programs [6] (latter program was used for the extended analysis of intergenic recombination events). The PhyML algorithm [21] with 100 bootstrap replicates was carried out to infer trees from different gene regions for each position of the sliding window used in Partial HGT-Detection. This algorithm was carried out separately with sliding window sizes of 10, 25, 50 and 100 sites as well as with the whole sequence alignments (to infer gene transfers of whole genes). The sliding window advancement sizes (i.e. step sizes) of 1 (for short genes) and 10 (for long genes and whole genomes) sites was used in our analysis. Gene transfer-recombination events with bootstrap support of at least 50% identified by Partial HGT-Detection (see Figs. 3—right portion of each panel, and 4b), and at least 40% identified by HGT-Detection (see Fig. 5), were represented by mapping them into the species tree. Some of these transfers may in fact be explained by the paradigm of parallel evolution when species sharing similar environment undergo similar mutations and develop similar traits.

SimPlot v3.5. [38] was used to carry out a sliding window analysis and determine patterns of sequence similarity using as reference the Wuhan SARS-CoV-2 2020 genome. This genome was compared to the RaTG13 genome as well as to the consensus genomes of the Guangdong Pangolin CoV group, Guangxi Pangolin CoV group, Bat CoVZ group and Bat SL-VoC group (see Fig. 1a). These consensus genomes were the default consensus genomes generated by SimPlot v3.5 in order to represent a group of species. In addition, gene-by-gene SimPlot similarity analysis was performed to compare the genes of the Wuhan SARS-CoV-2 2020 reference genome with those of the RaTG13, Guangdong Pangolin CoV and Bat CoVZ group genomes (Fig. 2).

The  $\Phi$ -recombination test [10] was performed to detect recombination patterns among individual genes and whole genome sequences of the Wuhan SARS-CoV-2, RaTG13, GD Pangolin 1 CoV, GD Pangolin P2S CoV, CoV ZC45 and CoV ZXC21 viruses (see Table 1). The  $\Phi$ -test was conducted with sliding windows of sizes 50 to 400 (with a step of 50) and the window progress step of 1. The version of the  $\Phi$ -test used in our study was that provided by David Bryant at his web site: <https://www.maths.otago.ac.nz/~dbryant/software.html>.

Tree clustering (Fig. 6) was carried out using the k-means-based tree clustering algorithm adapted for clustering trees with different numbers of leaves [57]

because some gene trees contained less than 25 species. The latest version of the tree clustering program was used (it is available at: <https://github.com/TahiriNadi/KMeansSuperTreeClustering>). The program was run with the following options—Tree clustering method: k-means; cluster validation index: Calinski-Harabasz; penalization parameter  $\alpha = 0$ ; Tree distance: Robinson and Foulds topological distance (not squared; see [33, 41, 43, 57]). The only difference with the default parameters of the program was that we set the penalization parameter  $\alpha$  to 0 because 11 out of 12 trees contained a full set of 25 species.

For the first cluster of trees (i.e. trees of genes ORF1ab, S, RB domain of S, and N) inferred for the full list of 25 species, the *Consense* program of the Phylip package [17] was used to infer the extended majority-rule consensus tree. As the sequences of BM48\_31\_BGR\_2008 and BtKY72 were missing in the multiple sequence alignment of gene ORF8, we applied a super-tree reconstruction method to retrace consensus evolutionary patterns for the cluster of genes ORF3a, E, M, ORF6, ORF7a, ORF7b, and ORF8. The *CLANN* program [13] was used with the best heuristic search (*hs*) and *bootstrap* options with 100 replicates to infer a supertree for these genes. The consensus tree of the third cluster, containing the only tree of ORF10, is the ORF10 gene tree inferred with RAxML.

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s12862-020-01732-2>.

**Additional file 1: Table 1.** The full virus names, abbreviations, host species and GenBank/GISAID accession numbers for all genomes analysed in our study. **Table 2.** The most appropriate evolutionary model found by the MegaX program for each gene/genome region analysed in our study, with the corresponding optimal parameters Gamma (G) and Intensity (I), as well as the values of Bayesian Information Criterion (BIC).

## Abbreviations

ACE2: Angiotensin-converting enzyme 2; CoV: Coronavirus; COVID-19: Coronavirus disease 2019; GD pangolin: Guangdong pangolin; GX pangolin: Guangxi pangolin; HGT: Horizontal gene transfer; MERS-CoV: Middle East respiratory syndrome coronavirus; MSA: Multiple sequence alignment; ORF: Open reading frame; RB domain: Receptor binding domain; SARS: Severe acute respiratory syndrome; SARS-CoV: Severe acute respiratory syndrome coronavirus.

## Acknowledgements

We thank Compute Canada and Université du Québec à Montréal for providing us with necessary computational resources. We also thank Dr. Fernando Gonzalez-Candelas and two anonymous reviewers for their valuable comments on this manuscript.

## Authors' contributions

VM, GR and PL supervised and designed the study. BM and VM performed data processing, recombination and horizontal gene transfer analyses. All authors read and approved the final manuscript.

## Funding

We thank the Canadian Institute for Advanced Research (CIFAR Catalyst Project CF-0136), Canada CIFAR AI Chair, and the Natural Sciences and Engineering Research Council (NSERC Grant No. 249644) for funding this work. BM received support as a Graduate Student Fellow from CIFAR and NSERC. The funding bodies (CIFAR and NSERC) played no role in the design of the study, analysis and interpretation of the data, and the writing of the manuscript.

## Availability of data and materials

The datasets supporting the conclusions of this article are available in our data archive at: [http://www.info2.uqam.ca/~makarenkov\\_v/Supplementary\\_Material.zip](http://www.info2.uqam.ca/~makarenkov_v/Supplementary_Material.zip) and in Additional Material.

## Ethics approval and consent to participate

All the experiments carried out in this study are in accordance with Canadian legislation, and the research performed does not require any ethical permits in Canada.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup> Département d'informatique, Université du Québec à Montréal, Montreal, QC, Canada. <sup>2</sup> Montreal Institute for Learning Algorithms (Mila), Montreal, QC, Canada. <sup>3</sup> Département d'informatique et de Recherche Opérationnelle, Université de Montréal and Canada CIFAR AI Chair, Montreal, QC, Canada. <sup>4</sup> Département de Sciences Biologiques, Université de Montréal, C. P. 6128, Succursale Centre-Ville, Montreal, QC H3C 3J7, Canada.

Received: 8 September 2020 Accepted: 8 December 2020

Published online: 21 January 2021

## References

- Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nat Med*. 2020;26:450–2.
- Arenas M. The importance and application of the ancestral recombination graph. *Front Genet*. 2013;4:206.
- Baptiste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L, Whitfield J. Networks: expanding evolutionary thinking. *Trends Genet*. 2013;29:439–41.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res*. 2007;36:D25–30.
- Becq J, Churlaud C, Deschavanne P. A benchmark of parametric methods for horizontal transfers detection. *PLoS ONE*. 2010;5:e9989.
- Boc A, Philippe H, Makarenkov V. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Syst Biol*. 2010;59:195–211.
- Boc A, Makarenkov V. Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic Acids Res*. 2011;39:e144–e144.
- Boc A, Diallo AB, Makarenkov V. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Res*. 2012;40:W573–9.
- Boni MF, Lemey P, Jiang X, Lam TTY, Perry B, Castoe T, Rambaut A, Robertson DL. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *bioRxiv*. 2020. <https://doi.org/10.1101/2020.03.30.015008>.
- Bruen T, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. *Genetics*. 2006;172:2665–81.
- Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 2000;17:540–52.
- Corel E, Lopez P, Méheust R, Baptiste E. Network-thinking: graphs to analyze microbial complexity and evolution. *Trends Microbiol*. 2016;24:224–37.
- Creevey CJ, McInerney JO. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics*. 2005;21:390–2.



14. Denamur E, Lecomte G, Darlu P, Tenaillon O, Acquaviva C, Sayada C, Sunjevaric I, Rothstein R, Elion J, Taddei F, Radman M, Matic I. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell*. 2000;103:71–21.
15. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M, Claverie J-M, Gascuel O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*. 2008;36:W465–9.
16. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform*. 2004;5:113.
17. Felsenstein J. 1993. PHYLIP (phylogeny inference package). Available from <https://evolution.genetics.washington.edu/phylip.html>.
18. Glazko G, Makarenkov V, Liu J, Mushegian A. Evolutionary history of bacteriophages with double-stranded DNA genomes. *Biol Direct*. 2007;2:36.
19. Graham RL, Sparks JS, Eckerle LD, Sims AC, Denison MR. SARS coronavirus replicase proteins in pathogenesis. *Virus Res*. 2008;133:88–100.
20. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *J Virol*. 2010;84:3134–46.
21. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res*. 2005;33:W557–9.
22. Guo YR, Cao QD, Hong ZS, Tan YY, Chen SD, Jin HJ, Tan KS, Wang DY, Yan Y. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak—an update on the status. *Mil Med Res*. 2020;7:1–10.
23. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2006;23:254–67.
24. Huson DH, Rupp R, Scornavacca C. Phylogenetic networks: concepts, algorithms and applications. Cambridge: Cambridge University Press; 2010.
25. Hu B, Zeng LP, Yang XL, Ge XY, Zhang W, Li B, Xie J-Z, Shen X-R, Zhang Y-Z, Wang N, Luo D-S, Zheng X-S, Wang M-N, Daszak P, Wang L-F, Cui J, Shi Z-L. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog*. 2017;13:e1006698.
26. Jin G, Nakhleh L, Snir S, Tuller T. Maximum likelihood of phylogenetic networks. *Bioinformatics*. 2006;22:2604–11.
27. Jin G, Nakhleh L, Snir S, Tuller T. Inferring phylogenetic networks by the maximum parsimony criterion: a case study. *Mol Biol Evol*. 2007;24:324–37.
28. Kandeel M, Ibrahim A, Fayed M, Al-Nazawi M. From SARS and MERS CoVs to SARS-CoV-2: moving toward more biased codon usage in viral structural and nonstructural genes. *J Med Virol*. 2020;92:660–6.
29. Koonin EV, Makarova KS, Aravind L. Horizontal gene transfer in prokaryotes: quantification and classification 1. *Ann Rev Microbiol*. 2001;55:709–42.
30. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35:1547–9.
31. Lam TT-Y, Jia N, Zhang Y-W, Shum MH-H, Jiang J-F, Zhu H-C, Tong Y-G, Shi Y-X, Ni X-B, Liao Y-S, Li W-J, Jiang B-G, Wei W, Yuan T-T, Zheng K, Cui X-M, Li J, Pei G-Q, Qiang X, Cheung WY-M, Li L-F, Sun F-F, Qin S, Huang J-C, Leung GM, Holmes EC, Hu Y-L, Guan Y, Cao W-C. Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins. *Nature*. 2020;583:282–5.
32. Lau SKP, Feng Y, Chen H, Luk HKH, Yang W-H, Li KSM, Zhang Y-Z, Huang Y, Song Z-Z, Chow W-N, Fan RYY, Ahmed SS, Yeung HC, Lam CSF, Cai J-P, Wong SSY, Chan JFW, Yuen K-Y, Zhang H-L, Woo PCY. Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. *J Virol*. 2015;89:10532–47.
33. Leclerc B, Makarenkov V. On some relations between 2-trees and tree metrics. *Discrete Math*. 1998;192(1–3):223–49.
34. Legendre P. Special section on reticulate evolution. *J Classif*. 2000;17:153–95.
35. Legendre P, Makarenkov V. Reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst Biol*. 2002;51:199–216.
36. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong XP, Chen Y, Gnanakaran S, Korber B, Gao F. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv*. 2020;6:eabb9153.
37. Liu P, Chen W, Chen JP. Viral metagenomics revealed Sendai virus and coronavirus infection of Malayan pangolins (*Manis javanica*). *Viruses*. 2019;11:979.
38. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol*. 1999;73:152–60.
39. Lu W, Zheng BJ, Xu K, Schwarz W, Du L, Wong CK, Chen J, Duan S, Deubel V, Sun B. Severe acute respiratory syndrome-associated coronavirus 3a protein forms an ion channel and modulates virus release. *Proc Natl Acad Sci USA*. 2006;103:12540–5.
40. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Yuhai B, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet*. 2020;395:565–74.
41. Makarenkov V, Leclerc B. Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees. In: *Mathematical hierarchies and biology*. Providence: American Mathematical Society; 1996. p. 183–208.
42. Makarenkov V, Legendre P. Improving the additive tree representation of a dissimilarity matrix using reticulations. In: *Data analysis, classification, and related methods*. Berlin: Springer; 2000. p. 35–40.
43. Makarenkov V, Leclerc B. Comparison of additive trees using circular orders. *J Comput Biol*. 2000;7:731–44.
44. Makarenkov V, Legendre P. From a phylogenetic tree to a reticulated network. *J Comput Biol*. 2004;11:195–212.
45. Makarenkov V, Legendre P, Desdevises Y. Modelling phylogenetic relationships using reticulated networks. *Zool Scr*. 2004;33:89–96.
46. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*. 2003;3:2.
47. Mortiboys H, Furnston R, Bronstad G, Aasly J, Elliott C, Bandmann O. UDCA exerts beneficial effect on mitochondrial dysfunction in LRRK2G2019S carriers and in vivo. *Neurology*. 2015;85:846–52.
48. Pérez-Losada M, Arenas M, Galan JC, Palero F, Gonzalez-Candelas F. Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infect Genet Evol*. 2015;30:296–307.
49. Prabakaran P, Gan J, Feng Y, Zhu Z, Choudhry V, Xiao X, Ji X, Dimitrov DS. Structure of severe acute respiratory syndrome coronavirus receptor-binding domain complexed with neutralizing antibody. *J Biol Chem*. 2006;281:15829–36.
50. Rambaut A, Holmes EC, Hill V, O'Toole A, McCrone J, Ruis C, du Plessis L, Pybus O. A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. *Nat Microbiol*. 2020;5:1403–7.
51. Ravenhall M, Škunca N, Lassalle F, Dessimoz C. Inferring horizontal gene transfer. *PLoS Comput Biol*. 2015;11:e1004095.
52. Schaecher SR, Touchette E, Schriewer J, Buller RM, Pekosz A. Severe acute respiratory syndrome coronavirus gene 7 products contribute to virus-induced apoptosis. *J Virol*. 2007;81:11054–68.
53. Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance*. 2017;22(13):30494.
54. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 2006;22:2688–90.
55. Stavrinos J, Guttman DS. Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J Virol*. 2004;78:76–82.
56. Surjit M, Lal SK. The SARS-CoV nucleocapsid protein: a protein with multifarious activities. *Infect Genet Evol*. 2008;8:397–405.
57. Tahiri N, Willems M, Makarenkov V. A new fast method for inferring multiple consensus trees using k-medoids. *BMC Evol Biol*. 2018;18:48.
58. Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S, Zhou Y, Du L. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell Mol Immunol*. 2020;17(6):613–20.
59. Weiss SR, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiol Mol Biol Rev*. 2005;69:635–64.

60. Woo PC, Huang Y, Lau SK, Yuen KY. Coronavirus genomics and bioinformatics analysis. *Viruses*. 2010;2:1804–20.
61. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, Li N, Guo Y, Li X, Shen X, Zhang Z, Shu F, Huang W, Li Y, Zhang Z, Chen R-A, Wu Y-J, Peng S-M, Huang M, Xie W-J, Cai Q-H, Hou F-H, Liu Y, Chen W, Xiao L, Shen Y. Isolation and characterization of 2019-nCoV-like coronavirus from Malayan pangolins. *BioRxiv*. 2020. <https://doi.org/10.1101/2020.02.17.951335>.
62. Xu R-H, He J-F, Evans MR, Peng G-W, Field HE, Yu D-W, Lee C-K, Luo H-M, Lin W-S, Lin P, Li L-H, Liang W-J, Lin J-Y, Schnur A. Epidemiologic clues to SARS origin in China. *Emerg Infect Dis*. 2004;10:1030.
63. Yoshimoto FK. The proteins of severe acute respiratory syndrome coronavirus-2 (SARS CoV-2 or n-COV19), the cause of COVID-19. *Protein J*. 2020;39:198–216.
64. Zhang KY, Gao YZ, Du MZ, Liu S, Dong C, Guo FB. Vgags: a viral genome annotation system. *Front Microbiol*. 2019;10:184.
65. Zhang CY, Wei JF, He SH. Adaptive evolution of the spike gene of SARS coronavirus: changes in positively selected sites in different epidemic groups. *BMC Microbiol*. 2006;6:88.
66. Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr Biol*. 2020. <https://doi.org/10.1016/j.cub.2020.03.022>.
67. Zhang Y, Zhang J, Chen Y, Luo B, Yuan Y, Huang F, Yang T, Yu F, Liu J, Liu B, Song Z, Chen J, Pan T, Zhang X, Li Y, Li R, Huang W, Xiao F, Zhang H. The ORF8 protein of SARS-CoV-2 mediates immune evasion through potentially downregulating MHC-I. *bioRxiv*. 2020. <https://doi.org/10.1101/2020.05.24.111823>.
68. 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, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579:270–3.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

