SARS-CoV-2 Origins and Evolution: Insights from *Coronaviruses* Recombination and Phylogenetic Analysis

Alltalents Tutsirayi Murahwa (✉ alltalents.murahwa@uct.ac.za)

University of Cape Town Department of Clinical Laboratory Sciences <https://orcid.org/0000-0003-3427-7852>

Harris Onywera

University of Cape Town Department of Clinical Laboratory Sciences

Fredrick Nindo

IHU Mediterranee Infection

Research

Keywords: Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), Recombination, Phylogenetic analysis

Posted Date: July 7th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-30068/v2>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Abstract

Background: It is imperative in the midst of a global epidemic to investigate the origins of the infectious agent especially when it has reached parts of the world with either ailing economies or pre-existing political turmoil consistent with non-functional health systems.

Methods: To explore the possibility of cross species infection, genomic recombination and the emergence of novel coronaviruses in the near future we carried out recombination and phylogenetic analysis to determine the spatio-temporal evolution and origins of the current SARS-CoV-2 virus.

Results: Our findings prove using two robust recombination tools, RDPv4.100 and SimPlot3.5.1 analysis that SARS-CoV-2 is a recombinant of pangolin and bat RaTG13 sequences as been previously shown elsewhere. We also report one novel recombination event between two SARS-CoV-2 sequences (*SARS-CoV-2 sequence, MT188341, SARS-CoV-2 sequence, MT293183*). Bearing in mind that the prerequisite for recombination is the occurrence two viral sequences in the same reservoir, biological niche or host at the same time we postulate either co-infection with the two viral sequences, or superinfection, both scenarios have not been reported elsewhere.

Conclusion: The possibility of recombination between the SARS-CoV-2 sequences poses the likelihood of the emergence of new and maybe more or less virulent “strains” of the virus. We believe that the future of science lies in our ability to be able to use computational based methods to predict the genetic sequences of infectious agents of the next epidemics. The addition of more SARS-CoV-2 sequences has a bearing on the understanding of the origin, evolution and clinical outcome prediction of given viral genomes. More SARS-CoV-2 sequences are needed to elucidate our understanding of this family of viruses.

Introduction

Coronaviruses (CoVs) are of the order *Nidovirales*, sub-order *Cornidovirineae*, family *Coronaviridae* and of the subfamily *Orthocoronavirinae* according to the recent ICTV taxonomic classification (<https://talk.ictvonline.org/taxonomy/>). The subfamily is divided into four genera namely *Alphacoronaviruses*, *Betacoronaviruses*, *Gammacoronaviruses* and *Deltacoronaviruses* (1). Alpha-CoVs and Beta CoVs have been mainly associated with mammals while Gamma-CoVs and Delta-CoVs have been associated with birds (2, 3). The CoV genome is a single stranded positive sense RNA virus with a genome ranging from 27kbp to 31kbp, making it the largest known RNA genome to date (4). All CoVs share similarities in their genomic organisation, which comprises 16 non-structural proteins from nsp1 to nsp16 encoded by the ORF 1a/b at the 5' end (5). The genome also encodes four structural proteins: the spike protein (S), envelope protein (E), membrane protein (M) and the nucleocapsid protein (N) (6, 7).

CoVs history dates back to the 1960s, where a variety of animal pathological conditions ranging from transmissible gastroenteritis to infectious bronchitis and to encephalitis in mice, were described; canine

respiratory CoVs (8, 9), mouse hepatitis virus (MHV) (10, 11), bovine CoV (12), feline CoV (9) and turkey CoV (13) were among the first animal and bird diseases to be documented.

Up until late 2019 only six CoVs were known to infect humans (14); these are HCoV229E, HCoV-OC43, SARS-CoV, HCoV-NL63, HCoV-HKU1 and MERS (15). The 229e and OC43 are the most widely studied HCoVs and they account for about 15-29% of all human common cold cases (16). SARS-CoV was the aetiological agent behind the 2003 outbreak in China (17), MERS-CoV was the pathogen behind the 2012 outbreak in the Middle East (18) and SARS-CoV-2 is responsible for the current and ongoing outbreak that started in Wuhan, China in November 2019 (4).

The origin of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has always been a question of many studies since November 2019 and even before the focus has been on the origins of SARS-CoV from the 2003 outbreak (1). It is imperative in the midst of a global epidemic to wonder why the origins of the infectious agent matter, especially when it has reached parts of the world with either ailing economies or political turmoil consistent with non-functional health systems. CoVs are normally hosted in, and their evolution moulded by bats (19, 20). It has been hypothesized that a large portion of the coronaviruses in humans are of bat origin (21). Obviously, a few research groups have as of late affirmed the hereditary comparability between SARS-CoV-2 and a bat betacoronavirus (22-27). The entire genome sequence of SARS-CoV-2 has 96.2% identity to a bat SARS-related coronavirus (SARSr-CoV; RaTG13) collected from Yunnan territory, China (28, 29), yet it is not fundamentally the same as the genomes of SARS-CoV (about 79%) or MERS-CoV (about half) (30, 31). It has additionally been affirmed that the SARS-CoV-2 uses a similar receptor, the angiotensin converting enzyme II (ACE2), as the SARS-CoV (25). In spite of the fact that the particular course of transmission from characteristic animal reservoirs to humans stays unclear (20, 27), a few studies have demonstrated that pangolins may have provided a partial spike gene to SARS-CoV-2; the basic functional sites in the spike protein of SAR-CoV-2 are about indistinguishable from one isolated in an infection from a pangolin (32-34). These findings indicate that SARS-CoV likely evolved in bats or pangolins over long periods of time. In general there are three theoretically plausible scenarios to explain the origins of SARS-CoV-2 (i) natural selection in animal host before zoonotic transfer (ii) natural selection in humans after zoonotic transfer and (iii) selection during laboratory passaging of experimental viruses (35). In any of the 3 scenarios the selection or change in the genome is due to either recombination or mutation (19).

Despite these recent discoveries, several aspects of the evolutionary tenets and driving forces behind the outbreak of SARS-CoV-2 remain unexplored (36). Here we further interrogated and investigated the likelihood of the most popular theory: of the origin of SARS-CoV-2 from bat and pangolin CoVs, by using recombination analysis of 60 reference CoV whole genomes and 413 SARS-CoV-2 sequences to explore the possibility of recombination as a driving force. This work provides new views into the factors driving the evolution of SARS-CoV-2 and its pattern of spread through the human population as shown in the phylogenetic analysis of 690 SARS-CoV-2 whole genome sequences from different geographical areas.

Methods

Design

This is an exploratory study to investigate the origin of CoVs as previously postulated in literature, CoVs can emerge from recombination events between human CoVs and other species CoVs (5). We first analysed the possibility of SARS-CoV-2 being a recombinant by aligning 60 CoV sequences with the SARS-CoV-2 reference sequence, then to detect the possibility of recombination within SARS-CoV-2 viruses a second recombination analysis was done using 413 SARS-CoV-2 sequences. In both recombination analyses tests were carried out to validate the recombination events. Phylogenetic analysis of SARS-CoV-2 sequences with previously known CoV reference and representative sequences was also done. Lastly we did a phylogenetic analysis of 690 SARS-CoV-2 sequences from the current outbreak to get a clearer picture of the velocity of the evolution of these sequences, from different geographical locations.

Source of sequence data

Four sequence data sets were used 1) 60 representative and reference CoV sequences 2) 413 SARS-CoV-2 complete genome sequences from the current outbreak 3) 473 CoV sequences, which were a combination of datasets 1 and 2 and 4) All the 690 SARS-CoV-2 complete genome sequences that were available from Genbank database at the time of the analyses. All the sequences including the reference SARS-CoV-2 sequence (**NC_045512.2**) were downloaded from Genbank (<https://www.ncbi.nlm.nih.gov/genbank/>) and the NCBI Virus resource (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/>), see also supplementary material 1a, b, c and d for the Fasta format of the sequences).

Alignments and phylogenetic analysis

All alignments as listed in Table 1 below were done using MAFFT version 7.45 (37), (<https://mafft.cbrc.jp/alignment/server/large.html>) (38). All phylogenetic trees were generated using the Neighbour-Joining method with 1000 bootstraps, and the Newick outputs were viewed and modified in iTOL (<https://itol.embl.de/tree/>). For the 60 CoV reference sequences (59 previously known CoVs and the reference SARS-CoV-2 sequence, **NC_045512.2**) the MAFFT alignment CLUSTAL output was used for the recombination analysis (R1), for the 473 Coronavirus Sequences (413 SARS-CoV-2 sequences from Genbank plus 60 Coronavirus reference sequences) the Newick format of the tree was uploaded, viewed and modified in iTOL as mentioned above, and the CLUSTAL alignment used for the recombination analysis (R2). And lastly for the 690 SARS-CoV-2 sequences the Newick format of the tree was also uploaded, viewed and modified in iTOL.

Recombination analysis

60 representative CoV sequences and the 473 sequence data sets as listed in Table 1 were included in the recombination analysis. We constructed alignments of these sequences using MAFFT and the CLUSTAL outputs were used for the recombination analyses. This alignment was analysed using RDP v4.100 (39)

(with default settings) which implements analysis of recombination using a suite of 9 recombination detection methods or algorithms: Recombination detection program (RDP) (39), BOOTSCAN (40), CHIMAERA (41), GENECONV (42), MAXIMUM X² (43), PhylPro (44), VisRD (45), LARD (46) and SISCAN (47). Only recombination events that were identified by at least four methods in the RDP v4.95 software were considered.

Phylogenetic Incongruence testing

The Shimodaira-Hasegawa (SH) test (48) using W-IQ-TREE v1.6.10 (49) was used to interrogate and corroborate the RDPv4.100 results. CLUSTAL alignments (50-52) of all the 60 CoV whole-genome reference sequences and the 473 sequences dataset, were used to compute the log-likelihoods of phylogenetic trees in W-IQ-TREE (<http://iqtree.cibiv.univie.ac.at>) (49). The tool tests tree topology, estimates model parameters such as substitution rates and optimizes tree branch lengths to lessen computational usage. Default settings of the W-IQ-Tree were used, including best fit model (53) and ultra-fast bootstrap analysis (1000 alignments) (54) to run tree topology analysis including the Kishino-Hasegawa (KH) test (55), Shimodaira-Hasegawa (SH) test (48) and approximately unbiased (AU) test (56) to test if there is a difference in evolutionary patterns amongst trees generated after removing the recombinant regions from the original sequences.

The two sets of trees constructed were denoted A1 and A2, A1 representing the trees generated from the original CoV reference sequences alignment and A2 denoting the trees generated after removing the recombinant regions from the original reference sequences, after the recombination analyses. The alignments without recombinant regions were generated automatically from RDP v4.95 (39) after the recombination analysis.

Pairwise percentage identity and Similarity plots

Using SARS-CoV-2_NC_045512.2 as the query sequence, we performed a similarity plot analysis using Simplot implemented in DAMBE6 (57) to estimate the pairwise genetic relatedness of the newly isolated SARS-CoV-2 with other reference corona viruses isolated from different species. We used default sliding window size of 500 nucleotides and step size of 50 under the TN93 distance (58) model as parameter settings for this analysis. The graph that is generated is a set of lines (or optionally strings of points) that reflect the similarity (or distance) of each reference sequence to the query sequence. In order to generate this plot, a sliding window is passed across the alignment in small steps (the window size and step size are selectable).

Table 1. Summary of analysis performed and sequence dataset used

Type of analysis	Sequence data set used
Alignments	<p><i>Alignment 1.</i> 60 reference Coronavirus sequences.</p> <p><i>Alignment 2. 473 sequences:</i> 413 SARS-CoV-2 sequences from the current outbreak plus 60 previously known Coronavirus reference sequences, obtained from the NCBI Virus resource.</p> <p><i>Alignment 3.</i> Phylogenetic analysis of 690 SARS-CoV-2 complete genome sequences from the current outbreak obtained from the NCBI Virus resource.</p>
Recombination analysis	<p>R1: 60 Coronavirus sequences (59 previously known CoVs and the reference SARS-CoV-2 sequence, NC_045512.2, Supplementary material 2a</p> <p>R2: 473 sequences: (413 SARS-CoV-2 sequences plus 60 reference sequences obtained from the NCBI Virus resource, Supplementary material 2b</p>
Shimodaira-Hasegawa test for phylogenetic incongruence	We used FASTA alignments of the 60 Coronavirus sequences, with and without recombination regions, to corroborate findings from the recombination analysis.
Phylogenetic tree construction	<p><i>Tree 1:</i> Constructed from Alignment 1(Supplementary material 3a, Newick format of the tree).</p> <p><i>Tree 2:</i> Constructed from Alignment 2 (Supplementary material 3b, Newick format of the tree).</p> <p><i>Tree 3:</i> Constructed from Alignment 3 (Supplementary material 3c, Newick format of the tree)</p>
Pairwise identity and Similarity plots	SARS-CoV-2 reference sequence, SARS-CoV (Frankfurt) and BatCov-HKU8, BetaCoV-Erinaceus (Hedgehog), bat-CoV RaTG13, MERS-CoV and 3 pangolin CoV sequences.

Results

Recombination analysis

R1: This analysis was done to initially check for the possibility of SARS-CoV-2 being a recombinant. Out of a total of 508 events, 28 events satisfied the criteria of at least 4 methods having significant p values, and of the 28 only two were of interest. Event 26 showed SARS-CoV-2 as a recombinant with a pangolin CoV and SARS-CoV as major and minor parents, and event 125 showed SARS-CoV as a recombinant with a pangolin CoV and bat CoV RaTG13 as major and minor parents respectively, see Table 2. Figures 1 and 2 also show the Neighbour-Joining (NJ) trees of the two events, event 26 tree shows how SARS-CoV-2 sequence, NC_045512.2 is most closely related to bat coronavirus RaTG13 (96.03%), and to the pangolin sequences (85.20%) as also shown in Table 3 in the pairwise percentage table. The event 26 NJ tree also

shows some distant relationship with a lot of other bat CoVs in the Beta-CoVs genus. In event 125 the NJ tree shows the SARS-CoV from the 2003 Chinese outbreak as a recombinant of bat coronavirus RaTG13 with 78.43% pairwise percentage identity as shown in Table 3. The pangolin CoVs also cluster closely with SARS-CoV with about 78.12% pairwise percentage identity. All of the recombination events were in the non-structural protein region of the genome.

R2: As a follow up to R1, we sought to check for recombination events within the different SARS-CoV-2 viruses available in Genbank at the time of the analysis and also computationally feasible to analyse. Out of the 422 events, 37 events satisfied the criteria of at least 4 methods having significant p values, Of the 37 events, 5 involved SARS-CoV-2 as a recombinant with bat CoVs and pangolin CoVs as major and minor parents as shown in events 12, 39, 112, 124 and 147, see also Table 2. A common feature among all the R2 events is that different “variants” of the SARS-CoV-2 virus as depicted by the different accession numbers show different bat and pangolin CoV parents. This might be an indication of ancient (and not recent) recombination events that happened over a long period of time as indicated by the interaction of the SARS-CoV-2 with a diverse number of other CoVs, which is not possible in a short space of time. In event 12 there is recombination between two SARS-CoV-2 sequences. This is important in describing the possibility of co-infection or super-infection, as the prerequisite for recombination is that two viral sequences have to be in the same space and time for recombination to occur. In event 39 and 124, there are recombination events between SARS-CoV-2 from the recent outbreak and SARS-CoV from the 2003 outbreak. This makes it possible to hypothesize that SARS-CoV is the backbone sequence for SARS-CoV-2 given that they have a 78.54% similarity as shown in Table 3. Hence it is reasonable to hypothesize that natural recombination and/or mutational events on a SARS-CoV backbone sequence might have led to the emergence of SARS-CoV-2, which may also provide basis for not to ruling out the possibility of laboratory manipulation of the SARS-CoV. Only event 147 was in the structural protein encoding region of the genome, specifically the nucleocapsid protein region of the genome.

Table 2. Recombination analysis with position and function of genomic regions

R1

Event	Recombinant, Genbank ID <i>Major Parent, Genbank ID</i> <i>Minor parent, Genbank ID</i>	No. of Methods	P-value range	*Position of breaking points Approx. genomic region	Known genomic function
26	SARS-CoV-2 sequence, NC_045512.2 <i>Pangolin coronavirus, MT040336</i> <i>SARS-CoV, NC_004718</i>	5	1.30x 10 ⁻⁰³ — 2.85x10 ⁻¹⁵	17581—18192 ORF1ab, nsp13 genomic region	RNA helicase, 5' triphosphatase (59)
125	SARS-CoV, NC_004718 <i>Pangolin coronavirus, MT040336</i> <i>Bat coronavirus RaTG13, MN996532.1.</i>	4	4.00x 10 ⁻⁰² — 3.10x10 ⁻⁰⁸	15105—15538 ORF1ab, nsp12 genomic region	Primer dependent RdRp (60)

R2

Event	Recombinant, Genbank ID <i>Major Parent, Genbank ID</i> <i>Minor parent, Genbank ID</i>	No. of Methods	P-value range	Position of breaking points Approx. genomic region	
12	SARS-CoV-2 sequence, MT188341 <i>SARS-CoV-2 sequence, MT293183</i> <i>BatCoV, NC_014470</i>	4	7.90 x 10 ⁻⁰³ —7.10 x10 ⁻⁴⁰	29810—45 ORF10 to ORF1ab, orf10 protein	Plays a role in virus assembly and release, and it involved in viral pathogenesis (61)
39	SARS-CoV,	5	1.30 x 10 ⁻	17752—	RNA helicase, 5'

	NC_004718 <i>Pangolin coronavirus, MT040336</i> <i>SARS-CoV-2 sequence, MT123293</i>		$10^{-8.40}$	18121 ORF1ab, nsp13 genomic region	triphosphatase (59)
112	SARS-CoV-2 sequence, MT039888 <i>Bat CoV, NC_025217</i> <i>Bat CoV-HKU4, NC_009019</i>	7	1.40×10^{-02} 1.14×10^{-08}	4706–4934 ORF1ab, nsp3 genomic region	Polypeptides cleaving, blocking host innate immune response, promoting cytokine expression (62)
124	SARS-CoV-2 sequence, MT163716 <i>Pangolin coronavirus, MT040335</i> <i>SARS-CoV, NC_004718</i>	4	4.90×10^{-02} 1.03×10^{-07}	15164–15584 ORF1ab, nsp12 genomic region	Primer dependent RdRp (60)
147	SARS-CoV-2 sequence, MT263421 <i>Bat CoV, NC_025217</i> <i>Bat CoV-HKU5, NC_009020</i>	5	1.60×10^{-03} 6.01×10^{-07}	28935–29371 ORF9, Nucleocapsid protein	An antagonist of interferon (IFN) and viral encoded repressor of RNA interference (63)

*Approximate genomic region was estimated from nucleotide positions based on the SARS-CoV-2 sequence, NC_045512.2 using the Genbank annotated format of the sequence.

Pairwise genetic distance Simplot of SARS-CoV-2 and other related CoV

To corroborate the findings from the RDP4 software in detecting recombination, we used the Similarity Plot. The Similarity Plot allows identification of one query sequence, generally the one suspected to be the recombinant or the mosaic, and the rest of the sequences are considered as reference sequences. The analysis indicated that the newly isolated SARS-CoV-2 represented by sequence SARS-CoV-2_NC_045512.2 in our analysis is more closely related to bat Coronavirus (RaTG13) further supporting preliminary reports that suggested a possible species jump from bats to humans to seed the current pandemic virus associated with COVID19 (see Figure 3). One pangolin CoV MT040336 also shows close identity to SARS-CoV-2 (85.1%) similarity as also shown in Table 3.

Table 3. Pairwise percentage identity of SARS-CoV-2 and other related CoV

Phylogenetic Incongruence testing

To confirm the findings from the recombination analysis (RDP4) and sequence similarity (Simplot), we used a more conclusive test, the SH test (48). The null hypothesis of the SH test states that the difference between trees (branch length, topology or likelihoods) is zero. In Table 4 the observed differences between the 60A1 and 60A2 and/or 473A1 and 473A2 were significantly greater than zero and the null hypothesis was rejected, thus declaring that these trees are significantly different i.e. incongruent ($p < 0.05$) as shown by the p-values (p-SH) indicating that there is substantial phylogenetic incongruence between the 60A1 and 60A2 and/or 473A1 and 473A2 trees. The incongruence between these trees allude to the fact that recombination plays a role in CoV evolution and in SARS CoV-2 in particular.

Table 4. Shimodaira-Hasegawa test for incongruence

60 A1 tree as reference

Tree	deltaL	bp-RELL	p-KH	p-SH	p-WKH	p-WSH	c-ELW	p-AU
60A1	0	1	1	1+	1	1	1	1
60 A2	988.9	0	0	0	0	0	0	0

60 A2 tree reference

Tree	deltaL	bp-RELL	p-KH	p-SH	p-WKH	p-WSH	c-ELW	p-AU
60 A1	1.99	0.16	0.15	0.15	0.15	0.15	0.22	0.17
60 A2	0	0.84	0.85	1+	0.85	0.85	0.78	0.83

473 A1 tree as reference

Tree	deltaL	bp-RELL	p-KH	p-SH	p-WKH	p-WSH	c-ELW	p-AU
473 A1	0	0.99	0.97	1	0.97	0.97	0.99	0.10
473 A2	233.63	0.01	0.03	0.03	0.03	0.03	0.00	0.05

473 A2 tree as reference

Tree	deltaL	bp-RELL	p-KH	p-SH	p-WKH	p-WSH	c-ELW	p-AU
473 A1	12.74	0.40	0.38	0.38	0.38	0.38	0.40	0.37
473 A2	0	0.60	0.62	1	0.62	0.62	0.60	0.63

deltaL: logL difference from the maximal logl in the set; **bp-RELL:** bootstrap proportion using RELL method (64); **p-KH:** p-value of one-sided (55); **p-SH:** p-value of Shimodaira-Hasegawa test (48); **p-WKH:** p-value of weighted KH test; **p-WSH:** p-value of weighted SH test; **c-ELW:** Expected Likelihood Weight (65); **p-AU:** p-value of approximately unbiased (AU) test (56); +: 95% confidence sets; -: significant exclusion

Phylogenetic analysis

After demonstrating that recombination events play a major role in CoV evolution we sought to determine how the new 413 SARS-CoV-2 sequences would cluster together and also with previously known CoV sequences by constructing 3 phylogenetic trees as described in Figure 4, 5 and 6. Figure 4 shows 60 representative CoV sequences and how they are classified into the different genera. It also shows SARS-CoV-2 reference sequence being closely related to bat RaTG13 sequence in the sub-genus *Sarbecovirus*. Pangolin CoVs and other bat CoVs also cluster close to the SARS-CoV-2 virus. Figure 5 is an extension of Figure 4, with the addition of a collapsed branch of 413 SARS-CoV-2 sequences from the current outbreak. Figure 6 is an analysis of 690 SARS-CoV-2 complete sequences, with the tree rooted against the SARS-CoV-2 reference sequence (NC_045512.2).

Discussion

Recombination analysis

Our results show as been shown previously the likelihood of SARS-CoV-2 being a recombinant of pangolin and bat RaTG13 sequences (29, 35, 66, 67). Besides the events of interest reported in this paper, we also show in our analyses different other recombination events among CoV sequences, 28 and 37 significant events in R1 and R2 respectively (provided in supplementary material 2a and 2b). This is consistent with studies done elsewhere on the role played by recombination as a driving force in the evolution of CoVs (20, 23, 68). We hypothesize based on the different recombination events between SARS-CoV-2 and bat CoVs other than bat RaTG13 and pangolin CoVs other than pangolin CoV MT040336, that this multiplicity of events is not a recent occasion, but a natural event that happened over a longer period of time.

We report one event between two SARS-CoV-2 sequences (***SARS-CoV-2 sequence, MT188341, SARS-CoV-2 sequence, MT293183***). Bearing in mind that the prerequisite for recombination is that two viral sequences have to be in the same reservoir or host at the same space and time we postulate either co-infection with the two viral sequences, or superinfection after initial infection with one of the viral sequences, both scenarios have not been reported elsewhere. The possibility of recombination between the SARS-CoV-2 sequences poses the likelihood of the emergence of new and more virulent “strains” of the virus. Evidence of genetic diversity and rapid evolution of this novel CoV has been presented elsewhere (69), using an analysis of just 86 genomes.

The RDP4 analysis, well supported NJ trees, SH test and the SimPlot results all corroborate to the fact that recombination plays a role in CoV evolution. A1 trees generated from the original sequence alignments and A2 trees generated after removing the recombinant regions from the original sequences, were incongruent for the 60 CoV sequences and the 473 CoV sequences (i.e the trees showed different phylogenies), hence supporting recombination as an essential driving force in CoV evolution. The SimPlot as also shown by the pairwise percentage identity table shows the close relation between SARS-CoV-2, bat RaTG13 and pangolin MT040336 CoVs as previously alluded to.

In all the recombination events we also show that most events occurred in the non-structural protein encoding regions especially in the ORF1ab, which have different roles in replication of CoVs (4). Only one event, 147 in R2 occurred in part of the genome that encodes the nucleocapsid protein which is an antagonist of interferon (IFN) and viral encoded repressor of RNA interference (63).

Phylogenetic analysis

We also show the phylogenetic relatedness of the different genera of the CoVs in relation to SARS-CoV-2 reference sequence and in relation to 413 SARS-CoV-2 sequences. These results are in agreement with the recombination analyses mentioned above.

Phylogenetic analysis of 690 SARS-CoV-2 sequences available at the time of the study shows a distinct geographical distribution and evolution of the virus from the Chinese epicentre to two American epicentres past Middle east to Europe and then to Africa.

We hypothesize migration of SARS CoV-2 sequences in a clockwise fashion when referring to Figure 6, from the Chinese/USA sequences in the early days of the epidemic, a migration into the Middle East (No UAE sequences yet) in purple then the rise of an epicentre in the USA, represented by the first green cluster, then followed by a spillage of these sequences into Europe and Africa (only sequences from South Africa were available), probably before the travel bans. The second green cluster is also exclusively USA sequences, which may represent another geographical epidemic epicentre in the same country given the size of the USA. The basis of this theory is that travel bans were not put into effect for about 3 months into the epidemic, which allowed mingling of sequences from distant parts of the world, as represented by the bulk of sequences in frequently visited countries and also in popular transit airports in countries such as Turkey, Istanbul. It will be interesting to have sequences from Dubai or any UAE states to qualify this hypothesis. It was estimated that the overall risk of SARS-CoV-2/COVID-19 importation/introduction to Africa was lower than that to Europe (1% vs 11%, respectively) (70).

Caveats and limitations to understanding CoV evolution

Our current knowledge of CoVs is limited and mostly focused on 7 human, medically important and human closely related CoVs that have been associated with cross species infection, while the rest of the other plethora of CoVs biology is largely understudied and unknown. According to the virus pathogen database (<https://www.viprbrc.org/brc/home.spg?decorator=vipr>) there are about 4000 CoV sequences from different animal and avian species. Hence assumptions made from studying a limited number of CoVs cannot be necessarily generalised and applied to all CoVs. The same logic applies to just SARS-CoV-2 sequences from the current outbreak. Many SARS-CoV-2 sequences are still being deposited into Genbank and other repositories, and until a threshold number of representative sequences are attained, the SARS-CoV-2 community of researchers will remain underpowered to make assumptions closest to the reality of what happened in the emergence and evolution of this group of viruses.

Conclusion

The possibility of recombination between the SARS-CoV-2 sequences poses the likelihood of the emergence of new and more or less virulent “strains” of the virus. We believe that the future of science lies in our ability to be able to use computational based methods to predict the genetic sequences of infectious agents of the next epidemics. The addition of more SARS-CoV-2 sequences is essential in

achieving this, as it has a bearing on the understanding of the origin, evolution and clinical outcome prediction of given viral genomes. More SARS-CoV-2 sequences are needed to elucidate our understanding of this family of viruses.

Declarations

Ethics approval and consent to participate

No applicable

Consent for publication

No applicable

Availability of data and materials

Four sequence data sets were used 1) 60 representative and reference CoV sequences 2) 413 SARS-CoV-2 complete genome sequences from the current outbreak 3) 473 CoV sequences, which were a combination of datasets 1 and 2 and 4) All the 690 SARS-CoV-2 complete genome sequences that were available from Genbank database at the time of the analyses. All the sequences including the reference SARS-CoV-2 sequence (**NC_045512.2**) were downloaded from Genbank (<https://www.ncbi.nlm.nih.gov/genbank/>) and the NCBI Virus resource (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/>), see also supplementary materials provided. Because of the nature of data formats generated in the recombination analysis RDPv4.5 software has to be installed to view the data, and can be downloaded from <http://web.cbio.uct.ac.za/~darren/rdp.html>, the files are available as supplementary material 2a and 2b.

Competing interests

The authors declare that they have no competing interests.

Funding

Dr Alltalents T. Murahwa is the recipient of a Post Doc fellowship from the South African National Research Foundation. This work is based upon research supported by the South African Research Chairs Initiative of the Department of Science and Technology of South Africa and NRF.

Authors' contributions

Alltalents T. Murahwa (ATM) conceptualised the study did the recombination analysis and the writing of the manuscript. Harris Onywera (HO) and Fredrick Nindo (FN) did the evolutionary analysis, phylogenetic

tree constructions and the phylogenetic incongruence test.

Acknowledgements

We thank Prof Anna-Lise Williamson for the help rendered with proof reading, we thank Oliver Charity for helping with computational services for part of the SimPlot analysis.

Abbreviations

CoVs: Coronaviruses

SARS-CoV-2: Severe acute respiratory syndrome coronavirus-2

RDP: Recombination detection program

deltaL: logL difference from the maximal logL in the set

bp-RELL: bootstrap proportion using RELL method

p-KH: p-value of one-sided

p-SH: p-value of Shimodaira-Hasegawa test

p-WKH: p-value of weighted KH test

p-WSH: p-value of weighted SH test

c-ELW: Expected Likelihood Weight

p-AU: p-value of approximately unbiased (AU)

NJ: Neighbour joining

ORF: Open reading frame

References

1. Hu D, Zhu C, Ai L, He T, Wang Y, Ye F, et al. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg Microbes Infect.* 2018;7(1):154-.
2. Cui J, Han N, Streicker D, Li G, Tang X, Shi Z, et al. Evolutionary relationships between bat coronaviruses and their hosts. *Emerg Infect Dis.* 2007;13(10):1526-32.

3. Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. *Nature reviews Microbiology*. 2009;7(6):439-50.
4. Chen Y, Liu Q, Guo D. Emerging coronaviruses: Genome structure, replication, and pathogenesis. *J Med Virol*. 2020;92(4):418-23.
5. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol*. 2016;24(6):490-502.
6. Marra MA, Jones SJM, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YSN, et al. The Genome Sequence of the SARS-Associated Coronavirus. *Science*. 2003;300(5624):1399-404.
7. Gao J, Lu G, Qi J, Li Y, Wu Y, Deng Y, et al. Structure of the Fusion Core and Inhibition of Fusion by a Heptad Repeat Peptide Derived from the S Protein of Middle East Respiratory Syndrome Coronavirus. *J Virol*. 2013;87(24):13134-40.
8. Erles K, Toomey C, Brooks HW, Brownlie J. Detection of a group 2 coronavirus in dogs with canine infectious respiratory disease. *Virology*. 2003;310(2):216-23.
9. Pedersen NC, Evermann JF, McKeirnan AJ, Ott RL. Pathogenicity studies of feline coronavirus isolates 79-1146 and 79-1683. *Am J Vet Res*. 1984;45(12):2580-5.
10. Lai MMC, Cavanagh D. The Molecular Biology of Coronaviruses. In: Maramorosch K, Murphy FA, Shatkin AJ, editors. *Adv Virus Res*. 48: Academic Press; 1997. p. 1-100.
11. Weiner LP. Pathogenesis of Demyelination Induced by a Mouse Hepatitis. *Arch Neurol*. 1973;28(5):298-303.
12. Bridger JC, Caul EO, Egglestone SI. Replication of an enteric bovine coronavirus in intestinal organ cultures. *Arch Virol*. 1978;57(1):43-51.
13. Ismail MM, Tang Y, Saif YM. Pathogenicity of Turkey Coronavirus in Turkeys and Chickens. *Avian Dis*. 2003;47(3):515-22.
14. Wevers BA, van der Hoek L. Recently Discovered Human Coronaviruses. *Clin Lab Med*. 2009;29(4):715-24.
15. Kin N, Miszczak F, Lin W, Gouilh MA, Vabret A, Consortium E. Genomic Analysis of 15 Human Coronaviruses OC43 (HCoV-OC43s) Circulating in France from 2001 to 2013 Reveals a High Intra-Specific Diversity with New Recombinant Genotypes. *Viruses*. 2015;7(5):2358-77.
16. Monto AS. Medical reviews. Coronaviruses. *The Yale journal of biology and medicine*. 1974;47(4):234-51.
17. Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *The Lancet*. 2003;361(9366):1319-25.
18. Raj VS, Osterhaus ADME, Fouchier RAM, Haagmans BL. MERS: emergence of a novel human coronavirus. *Curr Opin Virol*. 2014;5:58-62.
19. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. *Nature Reviews Microbiology*. 2019;17(3):181-92.

20. Li X, Song Y, Wong G, Cui J. Bat origin of a new human coronavirus: there and back again. *Science China Life Sciences*. 2020;63(3):461-2.
21. Li W, Shi Z, Yu M, Ren W, Smith C, Epstein JH, et al. Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science*. 2005;310(5748):676-9.
22. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host & Microbe*. 2020;27(3):325-8.
23. Xu X, Chen P, Wang J, Feng J, Zhou H, Li X, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Science China Life sciences*. 2020;63(3):457-60.
24. Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. *J Med Virol*. 2020;92(4):455-9.
25. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. *bioRxiv*. 2020:2020.01.22.914952.
26. Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect*. 2020;9(1):221-36.
27. Wei X, Li X, Cui J. Evolutionary perspectives on novel coronaviruses identified in pneumonia cases in China. *National Science Review*. 2020;7(2):239-42.
28. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-3.
29. Paraskevis D, Kostaki EG, Magiorkinis G, Panayiotakopoulos G, Sourvinos G, Tsiodras S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infect Genet Evol*. 2020;79:104212.
30. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet (London, England)*. 2020;395(10224):565-74.
31. Gralinski LE, Menachery VD. Return of the Coronavirus: 2019-nCoV. *Viruses*. 2020;12(2).
32. Wong MC, Javornik Cregeen SJ, Ajami NJ, Petrosino JF. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv*. 2020:2020.02.07.939207.
33. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv*. 2020:2020.02.17.951335.
34. Lam TT-Y, Shum MH-H, Zhu H-C, Tong Y-G, Ni X-B, Liao Y-S, et al. Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv*. 2020:2020.02.13.945485.
35. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nat Med*. 2020;26(4):450-2.

36. Wu C-I, Poo M-m. Moral imperative for the immediate release of 2019-nCoV sequence data. *National Science Review*. 2020.
37. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*. 2017;20(4):1160-6.
38. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*. 2009;25(9):1189-91.
39. Martin D, Rybicki E. RDP: detection of recombination amongst aligned sequences. *Bioinformatics*. 2000;16(6):562-3.
40. Martin DP, Posada D, Crandall KA, Williamson C. A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints. *AIDS Res Hum Retroviruses*. 2005;21(1):98-102.
41. Martin DP, Williamson C, Posada D. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics*. 2005;21(2):260-2.
42. Padidam M, Sawyer S, Fauquet CM. Possible Emergence of New Geminiviruses by Frequent Recombination. *Virology*. 1999;265(2):218-25.
43. Smith JM. Analyzing the mosaic structure of genes. *J Mol Evol*. 1992;34(2):126-9.
44. G W. Phylogenetic profiles: A graphical method for detecting genetic recombination in homologous sequences. *J Mol Biol Evol*. (1998).15::326-35.
45. Forslund K, Huson DH, Moulton V. VisRD—visual recombination detection. *Bioinformatics*. 2004;20(18):3654-5.
46. Holmes E.C. W, M. & Rambaut, A. . Phylogenetic evidence for recombination in dengue virus. *J Mol Biol and Evol* (1999).16:405-9.
47. Gibbs MJ, Armstrong JS, Gibbs AJ. Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*. 2000;16(7):573-82.
48. Shimodaira H, Hasegawa M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol*. 1999;16(8):1114-.
49. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res*. 2016;44(W1):W232-5.
50. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
51. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res*. 2015;43(W1):W580-4.
52. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, et al. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res*. 2013;41(Web Server issue):W597-600.
53. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*. 2017;14(6):587-9.

54. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol.* 2013;30(5):1188-95.
55. Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol.* 1989;29(2):170-9.
56. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. *Syst Biol.* 2002;51(3):492-508.
57. Xia. X. DAMBE6: New tools for microbial genomics, phylogenetics and molecular evolution. . *J Hered.* 2017;108:431-7.
58. Tamura K, and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. . *Mol Biol Evol.* 1993;10:512-26.
59. Jia Z, Yan L, Ren Z, Wu L, Wang J, Guo J, et al. Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Res.* 2019;47(12):6538-50.
60. Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nature Communications.* 2019;10(1):2342.
61. DeDiego ML, Álvarez E, Almazán F, Rejas MT, Lamirande E, Roberts A, et al. A Severe Acute Respiratory Syndrome Coronavirus That Lacks the E Gene Is Attenuated In Vitro and In Vivo. *J Virol.* 2007;81(4):1701-13.
62. Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res.* 2018;149:58-74.
63. Cui L, Wang H, Ji Y, Yang J, Xu S, Huang X, et al. The Nucleocapsid Protein of Coronaviruses Acts as a Viral Suppressor of RNA Silencing in Mammalian Cells. *J Virol.* 2015;89(17):9029-43.
64. Kishino H, Miyata T, Hasegawa M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol.* 1990;31(2):151-60.
65. Strimmer K, Rambaut A. Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci.* 2002;269(1487):137-42.
66. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. *National Science Review.* 2020.
67. Xiao K, Zhai J, Feng Y, Zhou N, Zhang X, Zou J-J, et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature.* 2020.
68. Li X, Zai J, Zhao Q, Nie Q, Li Y, Foley BT, et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J Med Virol.* 2020.
69. Phan T. Genetic diversity and evolution of SARS-CoV-2. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases.* 2020;81:104260.
70. Gilbert M, Pullano G, Pinotti F, Valdano E, Poletto C, Boëlle P-Y, et al. Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. *The Lancet.*

71. Xia X. DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. . Mol Biol Evol. 2013. ;30 (7):1720-8.
72. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406-25.

Figure 1

Unrooted Neighbour Joining tree of event 26 of R1 analysis showing in red SARS-CoV-2 sequence, NC_045512.2 as the recombinant sequence and in green the Pangolin coronavirus, MT040336 as the major parent and the SARS-CoV, NC_004718 as the minor parent, as also shown in Table 2.

Figure 2

Unrooted Neighbour Joining tree of event 125 of R1 analysis showing in red SARS-CoV, NC_004718 as the recombinant sequence and in green the Pangolin coronavirus, MT040336 as the major parent and the Bat coronavirus RaTG13, MN996532.1 as the minor parent, as also shown in Table 2.

Figure 3

Simplot of 9 aligned sequences of 32638 sites generated with DAMBE (67) Using SARS-CoV-2_NC_045512.2 as query sequence with sliding window size of nucleotides and 556 step size of 50 under the TN93 distance (55) model, see also supplementary material 4 for a second version of the SimPlot showing genomic regions.

Figure 4

60 CoV reference sequences. Phylogenetic analysis of 60 CoV whole genome sequences. The sequences were aligned using MAFFT (Waterhouse, Procter et al. 2009). This unrooted tree was generated using The Neighbor-Joining (NJ) method, a very popular method due to its good balance between accuracy and efficiency (68). The Newick format of the tree was uploaded and modified in iTOL <https://itol.embl.de/tree/>. It can be seen from the tree that the SARS-CoV-2 reference sequence is a Betacoronavirus more closely related to Bat Coronavirus (RaTG13), to Pangolin CoVs, and also to SARS CoV sequence from the 2003 outbreak in Asia (NC004718).

Figure 5

413 SARS-CoV-2 sequences and 60 CoV reference sequences. Phylogenetic analysis of 473 CoV complete genome sequences (413 SARS-CoV-2 sequences from the current outbreak plus 60 569 previously known Coronavirus reference sequences), obtained from the NCBI Virus resource. The sequences were aligned using MAFFT (Waterhouse, Procter et al. 2009). This tree was generated using the Neighbor-Joining (NJ) method (68). The Newick format of the tree was uploaded and modified in iTOL <https://itol.embl.de/tree/>. It can be seen from the tree that the SARS-CoV-2 sequences are Betacoronavirus more closely related to Bat Coronavirus (RaTG13), to Pangolin CoVs, and also to SARS CoV sequence from the 2003 outbreak in Asia (NC004718). MERS-CoV (JX869059) 5 also known as Human betacoronavirus 2c EMC/2012 from the 2012 outbreak.

Figure 6

Phylogenetic analysis of 690 SARS-CoV-2 complete genome sequences obtained from the NCBI Virus resource. The sequences were aligned using MAFFT (38). This tree was generated using the Neighbor-Joining (NJ) method (68). The Newick format of the tree was uploaded and modified in iTOL <https://itol.embl.de/tree/>. For clarity of node labels see Supplementary material 3c, which is a Newick format of the tree and can be viewed in iTOL. The leaves labelled in red represent a clade of USA and Chinese SARS-CoV-2 sequences, while green are USA sequences only. Purple represent Middle East (Pakistan, Iran and Turkey) sequences. The blue is a mixture of mainly USA, South African (in black) here representing Africa, European, and Australian and South Korean sequences.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial3a.nwk](#)
- [Supplementarymaterial3c.nwk](#)
- [Supplementarymaterial3b.nwk](#)
- [Supplementarymaterial4.pptx](#)
- [Supplementarymaterial1a.txt](#)
- [Supplementarymaterial2a.rdp](#)
- [Supplementarymaterial1c.txt](#)
- [Supplementarymaterial1b.txt](#)
- [Supplementarymaterial1d.txt](#)
- [Supplementarymaterial2b.rdp](#)