Shell disorder analysis suggests that pangolins offered a window for a silent spread of an attenuated SARS-CoV-2 precursor among humans

Gerard Kian-Meng Goh,[1,*] A. Keith Dunker,[2] James A. Foster,[3,4] and Vladimir N. Uversky[5,6]

[1] Goh's BioComputing, Singapore, Republic of Singapore

[2] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana, USA.

[3] Department of Biological Sciences, University of Idaho, Moscow, Idaho, USA.

[4] Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho, USA

[5] Department of Molecular Medicine, USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA.

[6] Laboratory of New Methods in Biology, Institute for Biological Instrumentation of the Russian Academy of Sciences, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences", Pushchino, Moscow region, Russia

[*]Corresponding author

Email addresses:

GKMG: gohsbiocomputing@yahoo.com

AKD: kedunker@iupui.edu

JAF: foster@uidaho.edu

VNU: vuversky@health.usf.edu

Key words: pangolin;intrinsic; disorder; protein; nucleocapsid; Nipah; virulence; viral protein; protein structure; protein function, shell; covid;coronavirus; ebola;vaccine; immune; antibody;shell; nucleocapsid; nucleoprotein; matrix;attenuate;

Abstract

A model to predict the relative levels of respiratory and fecal-oral transmission potentials of coronaviruses (CoVs) by measuring the percentage of protein intrinsic disorder (PID) of the M (Membrane) and N (nucleoprotein) proteins in their outer and inner shells, respectively, was built before the MERS-CoV outbreak. Application of this model to the 2003 SARS-CoV indicated that this virus with M_{PID} = 8.6% and N_{PID} = 50.2% falls into group B, which consists of CoVs with intermediate levels of both fecal-oral and respiratory transmission potentials. Further validation of the model came with MERS-CoV (M_{PID} = 9%, N_{PID} = 44%) and SARS-CoV-2 (M_{PID} = 5.5%, N_{PID} = 48%) falling into the groups C and B, respectively. Group C contains CoVs with higher fecal-oral but lower respiratory transmission potentials. Unlike SARS-CoV, SARS-CoV-2 with M_{PID} = 5.5% has one of the hardest outer shells among CoVs. This shell hardness is believed to be responsible for high viral loads in the mucus and saliva making it more contagious than SARS-CoV. The hard shell is able to resist the anti-microbial enzymes in body fluids. Further searches have found that high rigidity of outer shell is characteristic for the CoVs of burrowing animals, such as rabbits (M_{PID} = 5.6%) and pangolins (M_{PID} = 5-6%), which are in contact with the buried feces. A closer inspection of pangolin-CoVs from 2017-19 reveals that these animals provided a unique window of opportunity for the entry of an attenuated SARS-CoV-2 precursor into the human population in 2017 or earlier, with the subsequent slow and silent spread as a mild cold that followed by its mutations into the current more virulent form. Evidence of this lies in the similarity of shell disorder and genetic proximity of the pangolin-CoVs to SARS-CoV-2 (~90%). A 2017 pangolin-CoV strain shows evidence of higher levels of attenuation and higher fecal-oral transmission associated with lower human infectivity via having lower N_{PID} (44.8%). Our shell disorder analysis also revealed that lower inner shell disorder is associated with the lesser virulence in a variety of viruses.

Introduction

COVID-19 and SARS-COV-2

In December 2019, it was noticed that patients in Wuhan were falling ill with severe cases of pneumonia. The culprit was quickly identified as a novel coronavirus closely related to the 2003 severe acute respiratory syndrome coronavirus (SARS-CoV). It was labelled SARS-CoV-2, and the disease it caused was named coronavirus infectious disease 2019 (COVID-19) (1). Just as viruses closely related to SARS-CoV can be found among animals such as horseshoe bats and civet car (2,3), close relatives of SARS-CoV-2 were found in bats (RATG13) and pangolins (4-10). In line with the debates on the actual identity of the animal intermediary of SARS-CoV-2(8,9,11), we are not only presenting here more evidence of the greater likelihood that pangolins served as an intermediary host, but are also able to detect the existence of attenuated strains of CoV closely related to SARS-CoV-2.

Protein intrinsic disorder of the viral shell and the modes of viral transmission

In 2011-12, before the outbreak of the Middle Eastern respiratory syndrome (MERS)-CoV, a model that studied the protein intrinsic disorder of the CoV shells was built (3,12). The model measured the level of intrinsic disorder in proteins comprising the outer and inner shells of CoVs, M (membrane protein found in the outer shell) and N (nucleoprotein constituting the inner shell(3,12)[3]. Upon doing so, the CoVs easily clustered into three groups based mainly on the N_{PID} (percentage of intrinsic disorder (PID) in N protein). The model predicted that SARS-CoV (M_{PID} = 8.6%; N_{PID} = 50.2%) would belong to Group B and have intermediate levels of respiratory and fecal-oral transmission potentials, whereas other CoVs, such as porcine transmissible epidemic gastroenteritis virus (TGEV: M_{PID} = 14%; N_{PID} = 43%), were expected to be in Group C, which includes CoVs with lower respiratory but higher fecal-oral transmission potentials (3,12). Then the MERS-CoV

came in 2012-13 (13), which presented a great opportunity to test the validity of the disorder-based viral transmission model. The model placed MERS-CoV to the group C(14), and, indeed, the MERS-CoV reservoir was later found to be among farm animals including camels, which are highly associated with fecal-oral transmission (13,15). Furthermore, MERS-CoV was found to be not easily transmissible among humans (13,14).

Later, another opportunity to test the model came, and this time it came in the form of the COVID-19 outbreak. Model suggested that the SARS-CoV-2 has to be in group B; i.e., a group with intermediate respiratory transmission, given its N_{PID} of 48% (16-18). Furthermore, something else strange and puzzling was seen in this virus: i.e., it has one of the hardest outer shells in the sizable sample of a wide variety of CoVs we had analyzed (17). This could account for the high levels of the SARS-CoV-2 contagiousness, as a hard outer shell is likely to make the virus more resistant to the antimicrobial enzymes found in body fluids, such as saliva and mucus (19,20). This should be manifested by higher viral load in the saliva and mucus. In agreement with these expectations, viral loads in the mucus and saliva for SAR-CoV-2 have been observed in clinical studies to be much higher than those of SAR-CoV(21).

Contagiousness, viral load, and virulence: An enigma

An interesting “competing” finding is the fact that SARS-CoV-2 Spike protein (S) binds to the host angiotensin converting enzyme-2 (ACE2) receptor more tightly than the SARS-CoV S protein by astonishing 20-30 times (10,22). While both “competing” findings are likely to be true, it is difficult to link the efficient binding of S to ACE2 with the high level of viral shedding as has been observed without necessitating a much higher viral load in the lungs. Why is then SARS-CoV-2 (case-fatality rate (CFR): 2-6%) not more virulent than SARS-CoV (CFR: 9-10%) (23-24) if we assume former's much higher viral load in the lungs, given that these two CoV are genetically close and therefore are likely to produce similar proteins. A more plausible answer would be that there could be a discrepancy between viral loads in the lungs and body fluids, as suggested by the shell disorder model.

Revisiting pangolins, this time, using shell disorder and AI

Nevertheless, the fact that SARS-CoV-2 S binds 20-30 times stronger to ACE2 than the SARS-CoV S protein (10,22) is an important piece of the puzzle, that we will be addressing in this paper, since it is saying that SARS-CoV-2 is highly adapted to human. Furthermore, a recent study has shown that part of the sequence pertaining to S may have come from an unrelated human enzyme(23). If so, how did it evolve with human? How and when did the virus enter humans? Thus far, these important questions have remained unresolved. We shall see that the data arising from the shell disorder model that combines empirical proteomic analysis with artificial intelligence (AI) has detected a unique window of opportunity, in which an attenuated precursor of SARS-CoV-2 could have entered the human population years ago and, thus, initiated a slow and silent spread before mutating to become virulent and more contagious form as currently seen.

Methods

Protein intrinsic disorder

An important concept that will be used constantly throughout this paper is protein intrinsic disorder. Protein intrinsic disorder refers to the lack of structures in parts or a whole functional protein[24]. Disorder play roles in the molecular recognition and protein-protein/DNA/RNA/polysaccharide binding (25,26). There are alternative names, such as natively/naturally unfolded/unstructured/flexible proteins(25).

Intrinsic disorder predictor and percentage of intrinsic disorder (PID) calculations

Multiple disorder predictors have been developed, and one the earliest such predictor is PONDR® VLXT (www.pondr.com)(27-29). This is a neural network (AI: Artificial Intelligence) that has

been trained on the sequences of known ordered and disordered proteins. Because PONDR® VLXT is highly sensitive to the local sequence peculiarities and protein-protein/DNA/RNA interactions(30-31), it has been highly successfully used to analyze viral proteins, especially shell proteins of a large variety of viruses, including human immunodeficiency virus (HIV), herpes simplex virus (HSV), hepatitis C virus (HCV), Nipah virus (NiV), Ebola virus (EBOV), 1918 HIN1 influenza A virus, CoVs, dengue virus (DENV) and flaviviruses eg yellow fever (YFV), Zika (ZIKV) (32-41).

Upon the reading of a protein sequence, PONDR® VLXT will provide intrinsic disorder predisposition scores between 0 and 1 for each amino-acid residue. Residues of 0.5 or above are those predicted to be disordered (27-29). An important ratio that will be repeatedly used in this study is PID (Percentage of Intrinsic Disorder), which is defined as the number of residues predicted to be disordered divided by the total number of residues in the protein and multiplied by 100%.

Other tools

The sequences were downloaded from UniProt (https://www.uniProt.org) or GenBank-NCBI (https://www.ncbi.nlm.nih.gov/protein). The sequences were used as inputs to the PONDR® VLXT server (https://www.pondr.com), and both the results and sequences were downloaded into a mySQL server using a program written in JAVA(32). Sequence similarities were evaluated by BLASTP available at NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) and phylogenetic trees were obtained from EMBI-EBI (https://www.ebi.ac.uk/Tools/msa/clustalo/) using the respective sequences as inputs. The N and M phylogenetic trees were annotated with PIDs using an open-source drawing software, GIMP (https://www.gimp.org/). The evolutionary pathways were illustrated using open-source platforms, OpenOffice Draw (https://www.openoffice.org/download/) and GIMP. Multivariate analysis used to calculate correlations was conducted using R statistical package(42) .

Results

Categorization of CoV based mainly on N_{PID}

The aforementioned categorization of CoVs based on disorder status of their outer and inner shells and their transmission modes is summarized in **Table 1**. While the categorization is done mainly based on N_{PID}, statistical analysis picked up a small increase in correlation between PID and levels of respiratory transmission potential when M_{PID} (r^2=0.80) was added as an independent variable in addition to the already used N_{PID} (r^2 =0.77)(17). This basically means that the statistics is able to detect a small contribution of M_{PID} to the determination of the categorization of CoV as seen in **Table 1**.

PID: Patterns of CoV evolution dependent on host behaviors

As reiterated, SARS-CoV-2 and SARS-CoV are in the group B, which includes CoVs that have both intermediate levels of fecal-oral and respiratory transmission potentials. However, SARS-CoV-2 was also observed to have exceptionally hard outer shell. In previous publication, the only other CoV detected to have a harder shell than SARS-CoV-2 was HCoV-HKU1. Since then, a search have been made to uncover other CoVs with similarly low M_{PID} values. One of these uncovered CoVs is rabbit-CoV (HKU14, see **Tables 1-2**) that has M_{PID} and N_{PID} of 5.4% and 52.2% respectively.

The other CoVs that were uncovered to be closely related to SARS-CoV-2 are bat-RATG13 and Pangolin-CoVs. With the exception of SARS-CoV and MERS-CoV that were added for comparative purposes, **Figure 1** highlights the CoVs that have the lowest M_{PID} values in our new CoV sample. The next CoVs that have higher M_{PID} values compared to SARS-CoV-2 are canine (Resp)-CoV (M_{PID} = 7%) and bovine-CoV (M_{PID} = 7.8%). An interesting note is that while pangolins and bats are generalized as to be in group C and B respectively, further details of their

virus categorization can be found in **Table 2**, which tells us that 3 of the 4 pangolin CoVs and only 1 of 4 bat CoV samples fall into group C with the rest being placed to the group B. We are beginning to see an emerging pattern of PIDs that is related to the evolutionary pressures faced by the various CoVs arising from the behaviors of their various animal hosts. This will be further analyzed in section below.

Low M_{PID}: Burrowing animals and contacts with fecal materials buried in the soil

A mystery immediately arises: How did SARS-CoV-2 acquire its hard outer shell? What is the evolutionary significance of a hard outer shell? An inspection of **Tables 1-2** and **Figure 1** suggests that the clue lies in rabbits, pangolins, and, perhaps, bats, since they have M_{PID} values of 5.7%, 4.5-6.3% and 4.1% (bat-RATG13 only) respectively. A more careful study, however, tells us that bat-CoVs have a wide range of M_{PID} values (4.1-17%), whereas all pangolin-CoVs in this sample have outer shells as hard as SARS-CoV-2. The hard outer shell is something inherent in pangolin-CoVs and rabbit-CoVs, but not necessarily in bat-CoVs. A new question is then: Why rabbits and pangolins? To answer this question, we need to look at the behaviors of these animals. Pangolins are ant-eaters that dig into the ground for ant and termite meals(43). Both rabbits and pangolins dig burrows in the ground to build their nests. As a result, they are likely to come into contact with feces that have been buried in the ground.

According to shell disorder model, rabbit-CoVs and pangolin-CoVs are associated with fecal-respiratory and fecal-oral transmissions, respectively: Buried feces

An apparent paradox then appears when we see that while rabbit-CoV is in group B (intermediate fecal-oral and respiratory transmission potentials), most of the pangolin-CoVs are placed into group C (more fecal-oral transmission potentials) as seen in **Tables 1-2** and **Figure 1 A-B**. Furthermore, a glance at the CoVs in group C makes us realized that pangolin-CoVs are different from CoVs of

farm animals, such as the well-studied porcine TGEV, which typically moves very rapidly among pigs via fecal-oral routes(3,12,44). Because TGEV does not have to remain in the environment for a long time, it has a soft outer shell (M_{PID} = 14%) but is clearly has greater fecal-oral potentials (N_{PID} = 43%). In contrast, both pangolin-CoVs and rabbit-CoV are likely to remain in the soil with feces for a long time, which necessitates a harder outer shell (pangolin M_{PID} = 4.5-6.3%; rabbit M_{PID} = 5.7%). The hard outer shells for rabbit and pangolin CoVs are not coincidental, as they are both burrowing animals exposed to the environments, in which the virus can remain active and buried along with the feces for a long time.

There are, however, obvious differences between the behaviors of pangolins and rabbits despite the fact that both lives in burrows, which could increase the chances of contact with feces that had been buried in the ground. Pangolins eat ants and termites by the use of their sticky tongues, which could accidentally touch fecal matters in the ground. Rabbits, in contrast, eat leafs that found above the ground. This is likely the reason that the model is detecting greater fecal-respiratory transmission for rabbits and greater fecal-oral transmission potential for pangolins, as seen by their N_{PID} values of 44-48% and 52.2% (respectively, Figure 1 B).

Phylogenetic and shell disorder analyses could uncover evolutionary pathways that would otherwise have been missed if phylogenetic analysis alone is used

Figure 2 provides phylogenetic trees of M and N proteins annotated with PIDs. We are able to observe that genetically distant CoVs could have the similar N or M PID because of co-evolution where different viruses could face the same evolutionary pressures. The mentioned case of pangolins and rabbits is an excellent example of this. We can see in Figure 2 that rabbits-CoV and pangolin-CoVs are genetically different but yet have similar M PID because of evolutionary pressures.

Bat-CoVs are genetically and M_{PID} diverse, but its N_{PID} values are around 46-49%: Evolutionary bottleneck related to the respiratory transmission potentials

Another interesting example is given by bats. **Table 2** shows that bat-CoVs are very genetically diverse, especially when one compares their sequence similarities to SARS-CoV-2, with the corresponding sequence similarities of pangolin-CoVs. Only RATG13, which is closely related to SARS-CoV-2, has a very hard outer shell (M_{PID} = 4.5%) as we can see in our bat samples (additional note: bat-RATG13 has 96% genomic similarity to SARS-CoV-2 (11)). Both bat M sequence similarities and M_{PID} values (M_{PID} = 4.5-17%) show a wide range of variability, whereas the levels of intrinsic disorder in N protein tell us a different story. Bat-CoVs are genetically diverse, as seen in **Table 2** (compare N and M sequence similarities) and **Figure 2B**, but the N_{PID} values are normally in the range of 46-49%. This characteristic is likely the result of the evolutionary pressure arising from the minimal levels of respiratory transmission potential necessary for the optimal infectivity among bats regardless of the CoV type.

Examining the possible evolutionary pathways for SARS-CoV and SARS-CoV-2 using both shell disorder and phylogenetic analysis

With a better understanding of differences in the evolutionary pressure arising from our knowledge of M_{PID} and N_{PID} values of the various CoVs, possible pathways of SARS-CoV and SARS-CoV-2 can be envisaged as seen in **Figure 3**. While the phylogenetic tree suggests pathways, the M and N disorder levels provide the approximate timeline necessary for the mutations to take place based on the differences in N_{PID} or M_{PID} values as indicators of the evolutionary pressure. **Figure 3** lists out some of the possible paths that SARS-CoV, SARS-CoV-2, and their precursors could have taken before reaching humans. We are able to observe that genetically dissimilar CoVs could have similar N_{PID} or M_{PID} because of they evolved under comparative conditions as we seen in the case of M_{PID} values of pangolin-CoVs and rabbit-CoV.

Path B: A quick transition without evolving much with pangolins – unlikely scenario

A previous study has dismissed the idea that the latest pangolin strain of CoV is a direct intermediary of the current SARS-CoV-2 on the basis of a genome-wide phylogenetic study(8). We believe that that study was incomplete for two reasons. Firstly, only one pangolin-CoV strain was used in the analysis. Secondly, tools that could take into account the peculiarities of evolutionary pressure, such as the shell disorder model, were not incorporated into the study. **Tables 1-2** and **Figure 1** show that the various pangolin-CoVs reveal signs of fecal-oral transmission traits despite having one strain (2019) that show close similarities to SARS-CoV-2. With their M_{PID} values remaining low, most pangolin-CoV strains tend to have low N_{PID} values as wells, which are a hallmark of the greater fecal-oral transmission potentials. This is consistent with the mentioned behaviors of pangolins, which dig the ground for ants or termites and use their sticky tongues to trap them as food[43]. It is not hard to see that fecal materials can easily enter their meals. For this reason, the two scenarios related to pangolins as SARS-CoV-2 intermediaries are put forth with the evolutionary pressures in mind. **Figure 3** illustrates path B as the most direct route, where a precursor virus enters the pangolins with N_{PID} values between 46-48% and low M_{PID}. In path B, the virus stays a short while among pangolins before moving to humans, as its current form (M_{PID} = 5.6% and N_{PID} = 48%). In this case, virus cannot stay long among pangolin because doing so would forces the virus to have decreased N_{PID} values, as fecal-oral evolutionary pressures set in.

Path C: A window for pangolin-CoV to enter the human population as an attenuated virus with low infectivity

Path C presents a more interesting scenario, where the precursor remains among pangolins for a long time and thus facing evolutionary pressures towards fecal-oral transmission with the lowering of N PID. In this scenario, however, an entry into the human population necessitate a quiet and slow

spread because the low N PID provides for low viral loads in both vital organs and body fluids. The shell disorder model suggests that in this instance the virus enters the human population as an attenuated strain of SAR-CoV-2. This could explain the reason that the SARS-CoV-2 is very adapted to humans. In other words, iIt was given a unique window of opportunity to evolve quietly with human for at least a few years.

Figure 2 and Table 2 tell us that while pangolin-CoV 2019 is very closely related to SARS-CoV-2 and is not likely to be an intermediary, there is a greater likelihood that an older pangolin-CoV strain, particularly a 2017 one, could have entered the human populations as seen in Figure 2B. Why is then the 2019 strain and Bat-RaTG13 that closely related to SARS-CoV-1 both genetically and in disorder of M and N? It is plausible that human SARS-CoV-2 re-entered the bat and pangolin populations as suggested in Figure 3C. This is not unimaginable as both animals could have encountered feces of infected humans in form of trash with fruit leftovers and ants.

Path A: No such attenuating pathway seen for the 2003 SARS-CoV in civet

This opportunity is not seen in the civet cat in the case of SARS-CoV as shown in path A in **Figure 3**. There are no tell-tale signs of fecal-oral transmission potentials in the civet cat CoV that we saw in pangolins. Nor should we expect this possibility based on the behaviors of the civet cats (3). Therefore, as seen in **Figure 3A**, the civet SARS-CoV must have entered the civet population to evolve within it for a moderate timespan to acquire a little more respiratory transmission potential; i.e., slight increase in N_{PID} before entering into the human population. Apparently, SARS-CoV did not have a chance to evolve and adapt to humans before being detected by the medical community.

Paths D-E: No such attenuating pathway seen for the SARS-CoV-2 in other animals

There are also other possible intermediaries for SARS-CoV-2, such as rabbits and cow (**Figure 3 D-E**). For such animal intermediaries, specific evolutionary pressures will be encountered, but no

inherent fecal-oral pressures are seen in these animals, based on the high N_{PID} values of their respective CoVs, even though rabbits have a hard outer shell comparable to that of SARS-CoV-2. In any case, SARS-CoV-2 is so genetically different from their respective CoVs that it is difficult to envisage that SARS-CoV-2 did not pick up any of CoV genetic materials along the way, if any of these animals had served as an intermediary, even if temporary for a short time.

Discussion

Basis of links between modes of transmission and N disorder: Viral load in body fluid

The shell disorder model is a spinoff from its parent research that involved the study of HIV and the lack of the effective anti-HIV vaccine. It was carefully contrived in 2011-12 from what is known about the behavior of farm animals and their CoVs, especially those of porcine (3,12,44). Before the 2003 SARS outbreak, CoVs were not of medical interest, but veterinary data for animal coronaviruses were plentiful, because of their threat to the farming community. The reason that there are correlations between N disorder and mode of transmission has to do with the necessity of having a minimal viral load in the body fluids; i.e., mucus and saliva, before respiratory transmission becomes viable[17].

Correlations between virulence and N disorder: Viral loads in vital organs

While our shell model provides a link between N disorder and mode of transmission, other studies have provided correlations between N disorder and viral virulence. The studies included a wide variety of related and non-related viruses, such as NiV, flaviviruses, DENV, and EBOV (3,18,38-41). There are also hints of links between N disorder and virulence in SARS-CoV and SARS-CoV-2. While there is still much we need to know about SARS-CoV and SARS-CoV-2, there is much stronger evidence of correlations between virulence and inner shell disorder in a fairly large number

of viruses as aforementioned. This has paved a way for a novel strategy in the development of COVID-19 vaccine using shell disorder as describe in our previous paper (18).

"Trojan horse" immune evasion: Quick replication before immune detection

We have described the correlations between inner shell disorder and virulence as a "Trojan horse" immune evasion strategy, because it involves rapid replication of the virus before the host immune system is able to detect its presence. This, however, often backfires on the virus by leading to the death of the host, since large viral loads in vital organs, such as the lungs can, of course, kill the host(3,18,38-41). The reason that inner shell disorder provides important means for the rapid replication is that inner shell plays vital role in the replication process and disorder is important for greater efficiency of protein-protein/RNA/DNA binding (26,45).

Inner shell disorder defines more efficient protein-protein/DNA/RNA interactions and therefore more efficient viral replication

The inner shells of many viruses play very similar roles. In the case of CoV N protein, it assists in the transportation of other viral proteins to regions near the host Golgi apparatus and ER (endoplasmic reticulum), where N helps with the packaging of the viral particles(46). Similarly in the case of EBOV, its NP (nucleoprotein) builds a structure that is involved in the transportation of viral proteins to the ER(47). All this requires protein-protein/DNA/RNA interactions (48). The greater disorder provides important means for greater binding efficiency especially to host proteins/DNA/RNA(17,18).

Links among inner shell disorder, virulence, and respiratory transmission: Viral loads in body fluids *vs*. vital organs

Virulence is associated with viral load in vital organs, such as the lungs, whereas respiratory transmission viability is highly dependent on the viral load in the mucus and saliva. While the viral load in vital organs is likely to correlate with the viral load in the body fluids, a heavy viral load in vital organs does not necessitate a heavy viral load in body fluids, as the latter contains anti-microbial and anti-viral enzymes (19,20). The ability of the virus to resist such enzymes is dependent on the hardness of the outer shell. The observed greater hardness of M (5.9%) in the case of SARS-CoV-2 is likely to confer greater resistance to such anti-microbial and anti-viral enzymes. This itself could account for the observed large viral load in body fluids. This is why SARS-CoV-2 is much more contagious than the 2003 SARS-CoV, which has a higher M_{PID} of 8.6%. The question is then: Why isn't the immense viral load in body fluids of COVID-19 patients not translated into the higher virulence, given that the CFR of SARS-CoV-2 is 2-6% and that of SARS-CoV being 9-10%? The answer has to do with the predicted discrepancy of viral load in the body fluids and vital organs. The shell disorder model tells us that this discrepancy comes not just from the differences in M_{PID} values but also the slightly lower SARS-CoV-2 N_{PID} (48% vs. 50.2%) that translates into a slightly lower ”Trojan horse” effect.

Pangolins offered a window of opportunity for a SARS-CoV-2 precursor to be attenuated with greater fecal-oral potentials before entering the human population

With all this in mind we are now ready to re-visit the roles of pangolins. We have mentioned that pangolins provided a unique window of opportunity, not seen in other animals, for a SARS-CoV-2 precursor to enter the human population quietly. Upon the inspection of the various pangolin-CoV strains dating back to 2017, the shell disorder model that we have just described detected strains that resemble attenuated versions of SARS-CoV-2, given reductions in pangolin-CoV N_{PID} values as seen in **Table 2** and **Figures 2C-3B**. Even more interestingly, while two previous strains have N_{PID} around 46%, which sits at the borderline of group B and C (**Figure 1**), one of the 2017 strains has an N_{PID} of 44.8% (see **Table 2** and **Figures 2C-3B**) that places this strain squarely into the group of

CoVs with higher fecal-oral transmission; i.e., the group C. Clinical data from both veterinary and medical communities have shown that while CoVs in group C can move rapidly via fecal-oral routes in farm animals, such as pigs or camels (3,12,44), it is likely to spread slowly within human populations, as it will not have sufficient viral loads to spread rapidly via respiratory modes. We have seen this in the case of MERS-CoV spread among humans(3,18).

Attenuated SARS-CoV-2 strains found in 2017-18 pangolin samples

The shell model not only makes the case for a strain that spreads slowly within the human population, but also points at the strain that had moved into the human population as an attenuated form of virus. Again, the strain in the focus here is the 2017 strain of pangolin-CoV with the N_{PID} of 44.8% (**Table 2**). We know that this is an attenuated strain for a number of reasons. Firstly, as seen in **Table 2**, all pangolin-CoV strains are genetically close to SARS-CoV-2 (pangolin-CoV 2019, SARS-CoV, and MERS-CoV have about 90%, 80% and 50% genetic similarity to SARS-CoV-2, respectively). Secondly, while SARS-CoV-2 and SARS-CoV are consistent with the inner shell attenuation theory based on the difference in their N_{PID} values (48% *vs*. 50%) and CFRs (2-6% *vs*. 9-10%) (**Figure 1C**), the wealth of our knowledge comes from other viruses and their established correlations between their degree of virulence and inner shell disorder. These fairly large variety of viruses include flaviviruses, DENV, NiV. and EBOV(38-41). Summarily, we should be reminded that respiratory transmission requires a maximal viral load in mucus and saliva, whereas greater virulence comes with the higher viral load in vital organs.

We have seen attenuated viruses becoming virulent before in polio vaccine

Attenuated viruses mutating to a virulent strain have been seen before. An excellent example of this is given by the case of polio vaccines. There are two main polio vaccine type. The first type is the Salk vaccine that is made up of inactivated viruses (proteins) that confers only short-term immunity upon

several booster shots (2,3). A second type is the Sabin vaccine, which is an attenuated virus that provides lifetime immunity to those receiving it (2,3). This, however, comes with a price. It has been long known that the inactivated vaccine has mutated to the virulent types (3,49). It should therefore not be difficult to envisage an attenuated SARS-CoV-2 that had entered into the human population a few years ago before mutating into its current virulent form by acquiring greater disorder at the N protein.

Conclusions

How did SARS-CoV-2 adapt so well to humans? Through pangolins.

There is a general consensus among scientists that SARS-CoV-2 is somehow very highly adapted to humans, so much so that it is suggested that the virus must have entered the human population a long time ago (10,23). This is where the mystery begins. When and how did SARS-CoV-2 enter the human population? How did it remain in the population for so long without the medical community detecting its presence? The shell disorder analysis has important answers to all these questions. A precursor virus probably entered the human population around 2017 or before. This virus was an attenuated form of SARS-CoV-2 that moved slowly, perhaps, within localized and limited communities as a mild cold before mutating to its current virulent form. Pangolins are uniquely suited to facilitate such pathway through their behaviors that support fecal-oral transmission of viruses. This also adds to the overwhelming evidence that the SARS-CoV-2 is not man-made.

Have we discovered a vaccine strain of SARS-CoV-2? Precaution advised

As mentioned, the shell disorder model has detected attenuated SARS-CoV-2 strains based on their close genetic proximity to the current virus and their lower levels of N disorder. Of particular interest is a 2017 pangolin-CoV strain seen Table 2 with N_{PID} of 44.8%. Can we suggest that this strain can be used as a vaccine strain of SARS-CoV-2? Although it is tempting to give a positive answer to this question, it is needless to say that extra precaution needs to be taken, when such information is used in

vaccine development. Firstly, as already mentioned, attenuated vaccines have the tendency to mutate to its virulent types. This is an inherent risk for all attenuated vaccines. Perhaps, additional mutations are needed to force the N_{PID} to go to even lower values, so as to lessen the chances of the mentioned converting mutations. Secondly, we do not know if even with the N_{PID} of 44.8% virus is sufficiently attenuated as a vaccine strain. Currently, we do not have adequate information to address these questions, and subsequent careful animal and clinical studies are required.

Author Contributions

GKMG conceived the idea, collected, and analyzed the data, and wrote the first draft. VNU helped with the collection and analysis of literature data, reviewed and revised the draft. AKD and JAF reviewed the manuscript and provided the resources necessary for the research.

Conflicts of Interest

GKMG is an independent researcher and the owner of Goh's BioComputing, Singapore.
GKMG has also written a book ("Viral Shapeshifters: Strange Behaviors of HIV and Other Viruses") on a related subject. The authors have no other potential conflict of interests.

References

1. WHO, Novel coronavirus (2019-nCoV) https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

2. Acheson, N.H. *Fundamentals of Molecular Virology*; Wiley: Hoboken, NJ, USA, 2007.

3. Goh, G.K. *Viral Shapeshifters: Strange Behaviors of HIV and Other Viruses*, 1st ed.; Simplicity Research Institute: Singapore, 2017.

4. Li, X.; Zai, J.; Zhao, Q.; Nie, Q.; Li, Y; Foley, B.T.; Chaillon. A. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J. Med. Virol*. **2020,** 11, 10.1002 doi: 10.1002/jmv.25731.

5. Zhou, P; Yang, X.L.; Wang, X.G.; Hu, B.; Zhang, L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020,** 579, 270-273. doi: 10.1038/s41586-020-2012-7

6. Fan, H.H.; Wang, L.Q/;, Liu, W.L.; An, X.P.; Liu, Z.D.; He, X.Q.; Song, L.H.; Tong, Y.G. Repurposing of clinically approved drugs for treatment of coronavirus disease 2019 in a 2019-novel coronavirus (2019-nCoV) related coronavirus model. *Chin. Med. J. (Engl)*. **2020,** 6. doi: 10.1097/CM9.0000000000000797.

7. Ji, W.; Wang, W.; Zhao, X.; Zai, J.; Li, X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. *J. Med. Virol*. **2020,** 92, 433-440. doi: 10.1002/jmv.25682.

8. Liu, P., Jiang, J.Z., Wan, X.F., et al. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)?. *PLoS Pathog*. **2020,** *16,* e1008421. doi:10.1371/journal.ppat.1008421.

9. Zhang, T., Wu, Q., Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol*. **2020,** *30,* 1346-1351.e2. doi:10.1016/j.cub.2020.03.022

10. Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F. The proximal origin of SARS-CoV-2. *Nat. Med*. **2020,** *26,* 450-452. doi:10.1038/s41591-020-0820-9

11. Mousavizadeh, L.; Ghasemi, S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. *J. Microbiol. Immunol. Infect*. **2020**. doi: 10.1016/j.jmii.2020.03.022.

12. Goh, G.K.; Dunker, A.K.; Uversky, V.N. Understanding viral transmission behavior via protein intrinsic disorder prediction: Coronaviruses. *J. Pathog.* **2012**, *2012*, 738590

13. WHO, Middle Eastern respiratory syndrome coronavirus (MERS-CoV) (https://www.who.int/ news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov)

14. Goh, G.K.; Dunker, A.K.; Uversky, V.N. Prediction of Intrinsic Disorder in MERS CoV/HCoV-EMC Supports a High Oral-Fecal Transmission. *PLoS Curr.* **2013**, *5*.

15. Ferguson, M.; Van Kerkhov, M.D. Identification of MERS-CoV in dromedary camels. *Lancet Infect. Dis*. *2014,* 14,93-4.

16. Goh, G.K.-M.; Dunker, A.K.; Foster, J.A.; Uversky, V.N. Rigidity of the outer shell predicted by a protein disorder model sheds light on COVID-19 (Wuhan-2019- nCoV) infectivity. *Biomolecules* **2020**, *10*, 331.

17. Goh, G.K.-M.; Dunker, A.K.; Foster, J.A.; Uversky, V.N. Shell disorder analysis predicts greater resilience of SARS-CoV-2 (COVID-19) outside the body and in body fluids. *Microb. Pathog.* **2020**, *144*, 4177.

18. Goh, G.; Dunker, A.K.; Foster, J.; Uversky, V. A Novel Strategy for the Development of Vaccines for SARS-CoV-2 (COVID-19) and Other Viruses Using AI and Viral Shell Disorder. *Preprints* **2020**, 2020050116 https://www.preprints.org/manuscript/202005.0116/v1

19. Malamud, D., Abrams, W., Barber C., Weissman D., Rehtanz M., Golub E. Antiviral activities in saliva. *Adv. Dent. Res.* **2011,** *23,* 34–37

20. Cole A.M., Dewan P., Ganz T. Innate antimicrobial activity of nasa secretions. *Infect. Immun.* **1999,** *67,* 3267–3275

21. Wolfel, R.; Corman, V.M.; Guggemos, W.; et al. Virological asscessmant of hospitalized patients with COVID-19. *J. Mol. Bio.* **2019***,* 431, 1650-1670.

22. Shang, J.; Ye, G.; Shi, K.; Hu, B; et al. Structural basis of receptor recovnition by SARS-CoV-2. *Nature* **2020,** 367,14444-14448. doi: 10.1038/s41586-020-2179-y.

23. Anand, P., Puranik, A., Aravamudan, M., Venkatakrishnan, A.J., Soundararajan, V.. SARS-CoV-2 strategically mimics proteolytic activation of human EnaC. *Elife* **2020***, 9,* e58603. doi:10.7554/eLife.58603

24. Tompa, P. Intrinsically unstructured proteins. *Trends Biochem. Sci.* **2002**, *27*, 527–533.

25. Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* **2000**, *41*, 415–427.

26. Wright, P.E.; Dyson, H.J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **1999**, *293*, 321–331.

27. Li, X.; Romero, P.; Rani, M.; Dunker, A.K.; Obradovic, Z. Predicting protein disorder for N-, C-, and internal regions. *Genome Inform. Ser. Workshop Genome Inform*. **1999,** 10, 30-40.

28. Garner, E.; Romero, P.; Dunker, A.K.; Brown, C.; Obradovic, Z. Predicting Binding Regions within Disordered Proteins. *Genome Inf.* **1999**, *10*, 41–50.

29. Romero, P.; Obradovic, Z.; Li, X.; Garner, E.C.; Brown, C.J.; Dunker, A.K. Sequence complexity of disordered protein. *Proteins* **2001**, *42*, 38–48.

30. Cheng, Y.; Oldfield, C.J.; Meng, J.; Romero, P.; Uversky, V.N.; Dunker, A.K. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* **2007**, *46*, 13468–13477.

31. Oldfield, C.J.; Cheng, Y.; Cortese, M.S.; Romero, P.; Uversky, V.N.; Dunker, A.K. Coupled folding and binding
with alpha-helix-forming molecular recognition elements. *Biochemistry.* **2005**, *44*, 12454–12470. [CrossRef]

32. Goh, G.K.; Dunker, A.K.; Uversky, V.N. Protein intrinsic disorder toolbox for comparative analysis of viral proteins. *BMC Genom.* **2008**, *9* (Suppl. 2).

33. Xue, B.; Williams, R.W.; Oldfield, C.J.; Meng Goh, G. qK.; Dunker, A.; Uversky, V.N. Viral disorder or disordered viruses: Do viral proteins possess unique features? *Protein Pept. Lett.* **2010**, *17*, 932–951. [CrossRef]

34. Goh, G.K.; Dunker, A.K.; Uversky, V.N. A comparative analysis of viral matrix proteins using disorder predictors. *Virol. J.* **2008**, *5*, 126.

35. Goh, G.K.; Dunker, A.K.; Uversky, V.N. Protein intrinsic disorder and influenza virulence: The 1918 H1N1 and H5N1 viruses. *Virol. J.* **2009**, *6*, 69.

36. Goh, G.K.; Dunker, A.K.; Uversky, V.N. Shell disorder, immune evasion and transmission behaviors among human and animal retroviruses. *Mol. Biosyst.* **2015**, *11*, 2312–2323.

37. Goh, G.K.-M.; Dunker, A.K.; Foster, J.A.; Uversky, V.N. HIV Vaccine Mystery and Viral Shell Disorder. *Biomolecules* **2019**, *9*, 178.

38. Goh, G.K.; Dunker, A.K.; Uversky, V.N. Correlating Flavivirus virulence and levels of intrinsic disorder in shell proteins: Protective roles vs. immune evasion. *Mol. Biosyst.* **2016**, *12*, 1881–1891.

39. Goh, G.K.-M.; Dunker, A.K.; Foster, J.A.; Uversky, V.N. Zika and flavivirus shell disorder: Virulence and fetal morbidity. *Biomolecules* **2019**, *9*, 710.

40. Goh, G.K.-M.; Dunker, A.K.; Foster, J.A.; Uversky, V.N. Nipah shell disorder: Mode of tramsmission and virulence . *Microb. Pathog.* **2020**, *141*, 3976.

41 Goh, G.K.; Dunker, A.K.; Uversky, V.N. Detection of links between Ebola nucleocapsidand virulence using disorder analysis. *Mol. Biosyst.* **2015**, *11*, 2337–2344.

42. R Core Team, R: A language and environment for statistical computing. Vienna, Austria. 2019.

43. Coulson, I.M.; Heath, M.E. Foraging behaviour and ecology of the Cape pangolin (Manix temminckii) in north-west Zimbabwe. A*frican J. Ecol.* **2020**, *35*, 361-369. doi:10.1111/j.1365-2028.1997.101-89101.x.

44. Lin, C.M., Gao, X., Tomochir, O., Vlasova, A.N., Oeselli, M.A., Saif, L. Antigenic

relationships among procine epidemic diarrhea virus and transmissible gastroenteritis virus strains. *J. Virol.* **2015,** *89,* 3332–3342

45 Macosay-Castillo, M.; Marvelli, G; Guharoy, M.; et al. The balancing act of intrinsically disordeeered proteins enabling functions while minimizing promiscuity. *J. Mo.l Bio*. **2019,** 431, 1650-1670.

46. McBride, R.; van Zyl, M,; Fielding, B.C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **2014,** 6: 2991-3018. doi: 10.3390/v6082991.

47. Wantanabe, S.; Noda, T.; Kawaoka, Y. Functional mapping of the nucleoprotein of the ebola virus. *J. Vir*. **2006,** 80, 3743-51.

48. Habchi, J.; Longhi, S. Structural disorder within paramyxovirus nucleoproteins and phosphoproteins. *Mol Biosyst* **2012**, *8*, 69-81.

49. Hanley, K.A. The double-edged sword: How evolution can make or break a live-attenuated virus vaccine. *Evolution (N Y)* **2011,** *4,* 635-643. doi:10.1007/s12052-011-0365-y

50. Mousavizadeh, L.; Ghasemi, S. Genotype and phenotype of COVID-19: Their roles in pathogenesis. *J. Microbiol. Immunol. Infect*. **2020**. doi: 10.1016/j.jmii.2020.03.022.

Figure legends

Figure 1. N and M PIDs of the various CoVs A. M PID of the CoVs with lowest M PID found in the sample. B. N PIDs of the CoVs with hardest outer shell (M) C. N PIDs of pangolin-CoV with comparison to SARS-CoV-2 (SARS2) and 2003 SARS-CoV (SARS). Case-fatality rates (CFR) of SARS-CoV and SARS-CoV-2 are added in C. Civet-SARS-CoV and MERS-CoV were shown in B only as references. Pangolin-CoV 2017(**) has been identified as a possible vaccine strain .

Figure 2. Phylogenetic trees of CoV N and M with disorder (PID) annotations. **A**. M Phylogenetic tree B. N Phylogenetic tree. Pangolin-CoV strains can be cross-referenced and identified using the respective N PIDs or M PIDs in Table 2.

Figure 3. Possible pathways of SARS-CoV/SARS-CoV-2 and its precursors in various animals. A) Civet cat and SAR-CoV B) Pangolins and SARS-CoV-2- Path B, where the CoV did not have a chance to evolve with the pangolins C. Pangolins and SARS-CoV-2- Path C, where the CoV has evolved with pangolins for a long time D. Rabbits and SARS-Cov-2 (Hypothetical purpose only) E. Bovine and SARS-CoV-2 (Hypothetical purpose only). The timeline for each pathway is adjusted for the evolutionary pressures seen in the shell disorder model. Paths A and C are highly plausible based on phylogenetic and shell disorder models.

Table 1. Categorization of coronaviruses by mainly N PID to predict levels of respiratory and fecal-oral transmission potentials ($p < 0.001$, $r^2 = 0.8$) .

Shell Disorder Group	Coronavirus	UniProt(U) /Genbank(G) Accession Code (M Proteins)[a]	UniProt(U) Genbank(G) Accession Code (N)[a]	M PID	N PID	Remarks
A	HCoV-229E	P15422	P15130	23	56	Higher levels of respiratory transmission lower levels of fecal-oral transmission
	IBV(Avian)	P69606	Q8JMI6	10	56	
B	Bovine	P69704	Q8V432(U)	7.8	53.1	Intermediate levels of respiratory and fecal-oral transmission
	Rabbit	H9AA37(U)	H9AA59(U)	5.7	52.2	
	PEDV (Porcine)	P59771(U)	Q07499(U)	8	51	
	Canine (Resp.)	A3E2F6(U)	A3E2F7(U)	7	50.5	
	HCoV-OC43	Q4VID2(U)	P33469(U)	7	51	
	SARS-CoV	P59596(U) Q6Q1R9(U)	P59595(U) Q6Q1R8(U)	8.6	50.2	
	HCoV-NL63	P0DTC5(U)	P0DTC9(U)	11	49	
	SARS-Cov-2	A3EXD6(U)	Q3LZX4(U)	5.9	48.2	
	Bats[b]			11.2±5.3	47.7±0.9	
C	MHV(Murine)[c]	Q9JEB4(U)	P03416(U)	8	46.8	Lower levels of respiratory transmission higher levels of fecal-oral transmission
	Pangolin[d]	QIA428617(G)	QIA48630(G)	5.6±0.9	46.6±1.6[e]	
	MERS-CoV	K0BU37(U)	K0BVN3(U)	9.1	44.3	
	TGEV(Porcine)	P09175(U)	P04134(U)	14	43	
	Canine(Ent.)	B8RIR2(U)	Q04700(U)	8	40	
	HCoV-HKU1	Q14EA7(U)	Q0ZME3(U)	4.5	37.4	

[a]UniProt(U): https://www.uniProt.org); (G)GenBank-NCBI (https://www.ncbi.nlm.nih.gov/protein).
[B]More details on the bat samples can be found in Table 2. 3 out of 4 bat-CoVs are in group B. Note: Large standard deviation can be seen for N PID as denoted by "±"
[c]MHV sits at the borderline and is placed in group C for convenience.
[d]More details on the pangolin samples can be found in Table 2. 3 out of 4 bat-CoVs are in group C. Standard deviation is denoted by "±".

Table 2. Grouping of pangolin-CoVs and Bat-CoVs by mainly N PID with SARSCoV and SARS-CoV-2 as references.

Coronavirus	M PID (%)	Accession: UniProt(U) GenBank(G)	Sequence Similarity (%)	N PID (%)	Accession UniProt(U) GenBank(G)	Sequence Similarity	Group
SARS-CoV-2	5.9	P0DTC5(U)	100	48.2	P0DTC9(U)	100	B
SARS-CoV	8.6	P59596(U)	90.5	50.24	56	90.5	B
Civet-SARS-CoV	8.6	QZ3TE9(U)	90.1			90.01	B
Pangolin-CoV	5.6±0.9[a]			46.6±1.6[a]	53		
2019	6.3	QIG55948(G)	98.2	48.7	QIG55953(G)	98	B
2018	4.5	QIQ54051(G)	97.7	46.3	QIQ54056(G)	93.8	C
2017**	**5.9**	QIA48617(G)	**98.2**	**44.9**	**QIA48630(G)**	**94**	**C**
				46.5	QIA48656(G)	93.32	C
Bat-CoV	11.2±15[a]	Q9JEB4		47.7±0.9[a]			
RATG13	4.1	QHR63303(G)	99.6	48.5	QHR63308(G)	99.1	B
512	15.3	Q0Q463(U)	35.5	46.5	Q0Q462(U)	29.4	C
HKU3	7.7	Q3LZX9(U)	91	48	Q3LZX4(U)	89.6	B
HKU4	16.4	A3EXA0(U)	42.7	48.5	A3EXA1	51.1	B
HKU5	11.8	A3EXD6(U	44.7	47.1	A3EXD7(U)	47.9	B

[a]Standard deviation is denoted by "±".
** Possible vaccine strain for SARS-CoV-2 detected

Figure 1

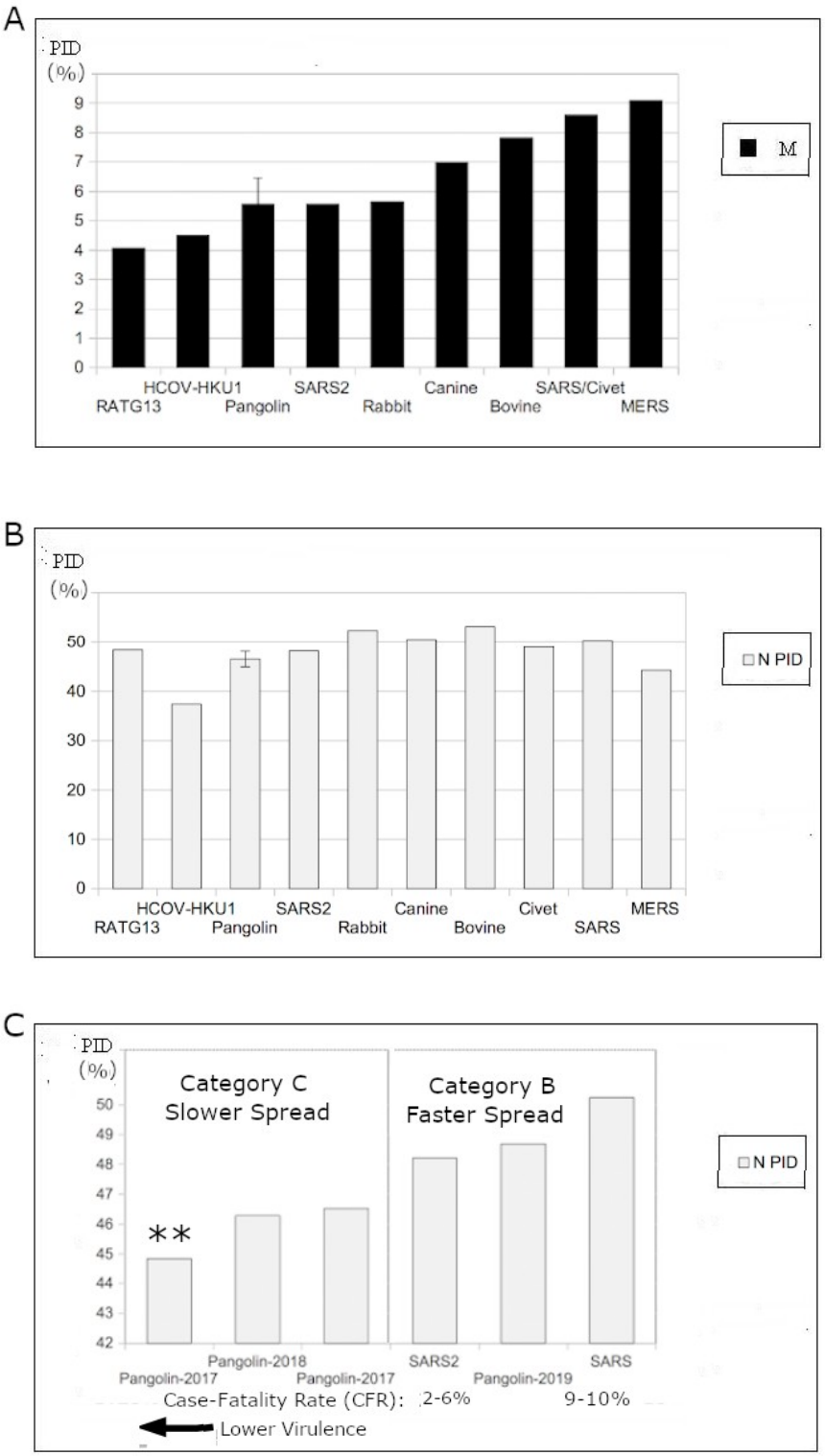
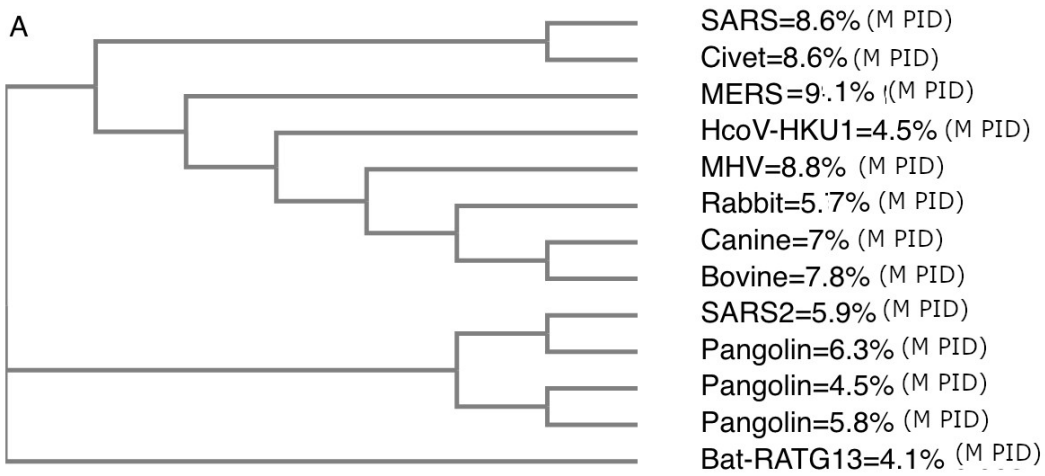


Figure 2



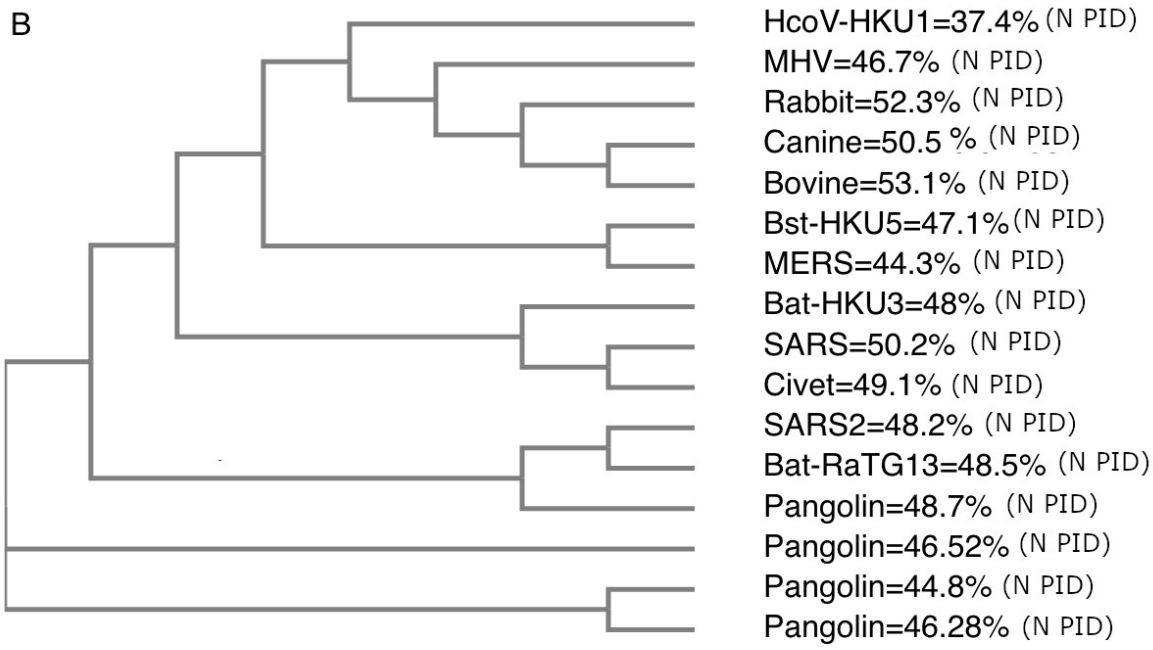


Figure 3