The Validity of critical pieces of evidence for the natural origin of SARS-CoV-2 is Dubious, and needed to be reconsidered.

Author: Daoyu Zhang

ABSTRACT

The origin of SARS-CoV-2, the agent that causes the global pandemic known as COVID-19, is of both heated Academic debate and political debate. As this directly affect policy decision and global politics, this matter must be considered with uttermost scrutiny.

The leading academic hypothesis of the origin was that of a natural recombination event between the Bat coronavirus RaTG13 and the pangolin coronavirus MP789, followed by adaptation in humans after zoonotic transfer.

However, this theory hinges critically on the validity of both RaTG13 and MP789, which require both strains to be able to be independently sequenced, tested and validated for infectivity of it's original host. Here we provide evidence that the validity of both strains are highly dubious and are incapable of sufficing the required conditions for both to be considered valid evidence for the hypothesis of a natural origin of SARS-CoV-2.

METHODS

Genomic and Proteomic data of RaTG13, MP789 and SARS-CoV-2 were obtained from GenBank, along with Bat coronaviruses

ZC45
ZXC21
AP040581.1
RsSHC014
SC2018
NP_828854.1
BtRs-BetaCoV/HuB2013
AVP78042.1
AVP78031.1
HKU3-8
AID16716.1
HKU3-12
HKU3-2
Bat SARS Cov Rs806/2006
HKU3-7
HKU3-13
HKU3-4

ACJ60703.1

ATO98169.1

Coronavirus BtRs-BetaCoV/YN2018D

Bat SARS CoV Rm1/2004

And the SARS Coronaviruses

SARS_ExoN1

BM48-31/BGR/2008

SARS_TW-GD1

SARS_Sino1-11

SARS_GD01

SARS coronavirus Rs_672/2006

SARS coronavirus GZ02

SARS coronavirus PC4-241

As control data.

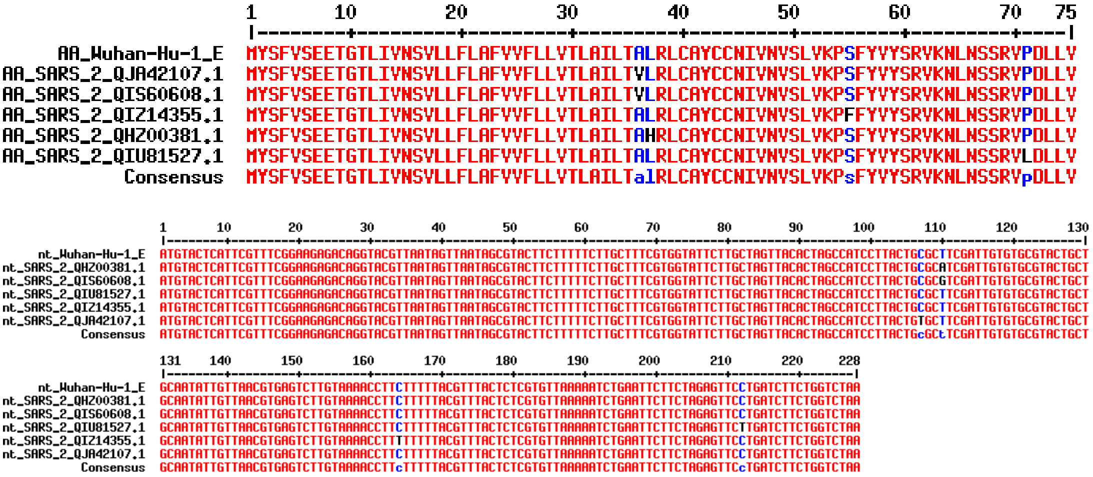
A Multalin Analysis was performed on the strains for the Amino Acid alignment data, while a BLAST analysis was performed on the nucleotide Data.

The RBM of Pangolin coronaviruses, GX-P1E, GX-P5E, GX-P4L, GX-P5L and GX-P2V were obtained from the relevant GenBank entries.

In addition, the binding affinity of The MP789 RBD and a Chimeric hACE2/pACE2 receptor, bearing the binding site amino acid residue of pACE2, was evaluated using the Rosetta protein structure prediction software.

DISCUSSIONS

The E protein of RaTG13, ZC45/ZXC21 and SARS-CoV-2



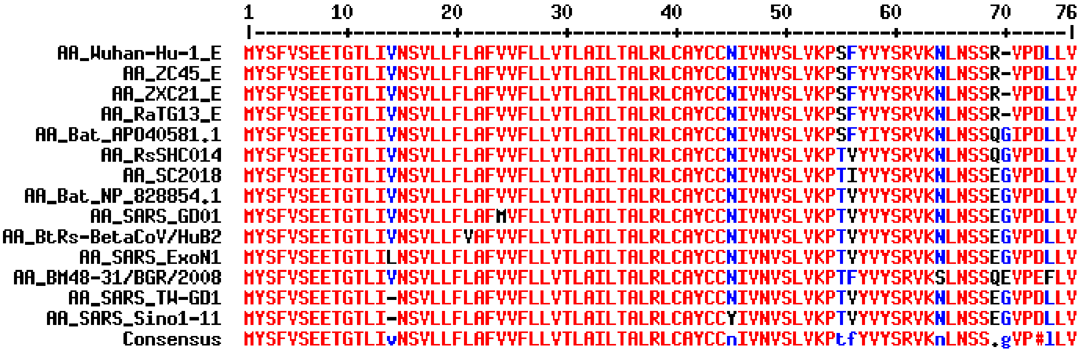
```
                      1        10        20        30        40        50        60        70    75
                      |--------+---------+---------+---------+---------+---------+---------+----|
     AA_Wuhan-Hu-1_E  MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSRVPDLLV
AA_SARS_2_QJA42107.1  MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTVLRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSRVPDLLV
AA_SARS_2_QIS60608.1  MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTVLRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSRVPDLLV
AA_SARS_2_QIZ14355.1  MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPFFYVYSRVKNLNSSRVPDLLV
AA_SARS_2_QHZ00381.1  MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTAHRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSRVPDLLV
AA_SARS_2_QIU81527.1  MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSRVLDLLV
           Consensus  MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTalRLCAYCCNIVNVSLVKPsFYVYSRVKNLNSSRVpDLLV
```

```
                      1        10        20        30        40        50        60        70        80        90        100       110       120       130
                      |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
     nt_Wuhan-Hu-1_E  ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCT
nt_SARS_2_QHZ00381.1  ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGCGCATCGATTGTGTGCGTACTGCT
nt_SARS_2_QIS60608.1  ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGCGCGTCGATTGTGTGCGTACTGCT
nt_SARS_2_QIU81527.1  ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCT
nt_SARS_2_QIZ14355.1  ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCT
nt_SARS_2_QJA42107.1  ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGTGCTTCGATTGTGTGCGTACTGCT
           Consensus  ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGcGCtTCGATTGTGTGCGTACTGCT

                      131     140       150       160       170       180       190       200       210       220      228
                      |--------+---------+---------+---------+---------+---------+---------+---------+---------+-------|
     nt_Wuhan-Hu-1_E  GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTCTGGTCTAA
nt_SARS_2_QHZ00381.1  GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTCTGGTCTAA
nt_SARS_2_QIS60608.1  GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTCTGGTCTAA
nt_SARS_2_QIU81527.1  GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCTTGATCTTCTGGTCTAA
nt_SARS_2_QIZ14355.1  GCAATATTGTTAACGTGAGTCTTGTAAAACCTTTTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTCTGGTCTAA
nt_SARS_2_QJA42107.1  GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTCTGGTCTAA
           Consensus  GCAATATTGTTAACGTGAGTCTTGTAAAACCTTcTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCcTGATCTTCTGGTCTAA
```

Fig.1ab: The E protein sequence alignment data of SARS-COV-2 strains WuHan-Hu-1, QJA42107.1,QIS60608.1,QIZ14355.1,QHZ00381.1 and QIU81527.1.

In order to establish the mutation rate of the E protein of strains related to SARS-COV-2, the alignment of the Amino Acid sequence of the different strains of SARS-COV-2 were performed. Alignment result indicate that there have been a minimum of 5 single nucleotide substitutions within the E gene of SARS-COV-2, 4 of which caused an amino acid change, since the start of it’s spread within humans.



```
                     1       10        20        30        40        50        60        70    76
                     |--------+---------+---------+---------+---------+---------+---------+-----|
     AA_Wuhan-Hu-1_E MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSR-VPDLLV
           AA_ZC45_E MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSR-VPDLLV
          AA_ZXC21_E MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSR-VPDLLV
          AA_RaTG13_E MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSRVKNLNSSR-VPDLLV
   AA_Bat_AP040581.1 MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYIYSRVKNLNSSQGIPDLLV
         AA_RsSHC014 MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPTVYVYSRVKNLNSSQGVPDLLV
           AA_SC2018 MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPTIYVYSRVKNLNSSEGVPDLLV
  AA_Bat_NP_828854.1 MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPTVYVYSRVKNLNSSEGVPDLLV
        AA_SARS_GD01 MYSFVSEETGTLIVNSVLLFLAFMVFLLVTLAILTALRLCAYCCNIVNVSLVKPTVYVYSRVKNLNSSEGVPDLLV
AA_BtRs-BetaCoV/HuB2 MYSFVSEETGTLIVNSVLLFVAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPTVYVYSRVKNLNSSEGVPDLLV
       AA_SARS_ExoN1 MYSFVSEETGTLILNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPTVYVYSRVKNLNSSEGVPDLLV
    AA_BM48-31/BGR/2008 MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPTFYVYSRVKSLNSSQEVPEFLV
      AA_SARS_TW-GD1 MYSFVSEETGTLI-NSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPTVYVYSRVKNLNSSEGVPDLLV
    AA_SARS_Sino1-11 MYSFVSEETGTLI-NSVLLFLAFVVFLLVTLAILTALRLCAYCCYIVNVSLVKPTVYVYSRVKNLNSSEGVPDLLV
           Consensus MYSFVSEETGTLIvNSVLLFLAFVVFLLVTLAILTALRLCAYCCnIVNVSLVKPtfYVYSRVKnLNSS.gVP#lLV
```

Fig.2a:The E protein sequence, of SARS-COV-2, when compared to ZC45, ZXC21, RaTG13, other Bat coronaviruses and the SARS coronaviruses.

This alignment data, along with that of current strains of SARS-CoV-2, clearly show high amino acid sequence variability both within the Bat host and within the Human host.

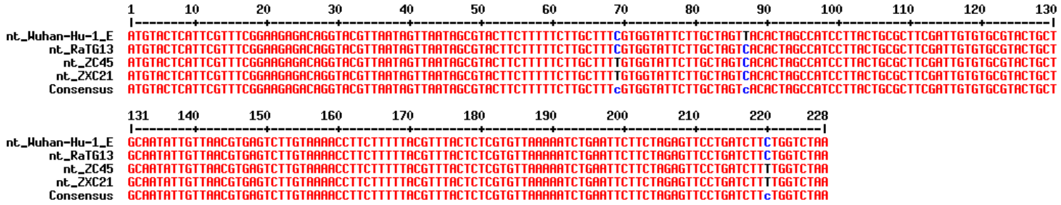
Of the four new Amino acid mutations within the current strains of SARS-COV-2, three of which were novel—the change lands within places that are not known to change previously. This brings up the total amino acid variability within the E protein up to 13 out of 75.

However, despite the high variability within the E protein, the Amino Acid sequence of the E protein within WuHan-Hu-1, the first published genome of SARS-CoV-2, was exactly the same as both ZC45, ZXC21 and RaTG13—indicating a highly conserved protein across this lineage of Coronaviruses. A level of Conservation that is known to not hold in either Bats or Humans.

What is the E protein?

The E protein, or Envelope protein of Coronaviruses are the protein that is located on the inside of the Envelope of the virus—it helps to assemble the virion during maturation and neither contact Host cell proteins nor Host surface receptors during the formation and transmission of the virion. Therefore, the E protein does not affect host selection—as indicated by it’s high variability both within Bat_CoVs, SARS-CoVs and SARS-CoV-2.

The E gene of ZC45, ZXC21, RaTG13 and SARS-CoV-2.



```
                1       10        20        30        40        50        60        70        80        90       100       110       120       130
                |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
nt_Wuhan-Hu-1_E ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTTACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCT
      nt_RaTG13 ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTCGTGGTATTCTTGCTAGTCACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCT
        nt_ZC45 ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTTGTGGTATTCTTGCTAGTCACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCT
       nt_ZXC21 ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTTGTGGTATTCTTGCTAGTCACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCT
      Consensus ATGTACTCATTCGTTTCGGAAGAGACAGGTACGTTAATAGTTAATAGCGTACTTCTTTTTCTTGCTTTcGTGGTATTCTTGCTAGTcACACTAGCCATCCTTACTGCGCTTCGATTGTGTGCGTACTGCT

                131    140       150       160       170       180       190       200       210       220     228
                |--------+---------+---------+---------+---------+---------+---------+---------+---------+-------|
nt_Wuhan-Hu-1_E GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTCTGGTCTAA
      nt_RaTG13 GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTCTGGTCTAA
        nt_ZC45 GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTTTGGTCTAA
       nt_ZXC21 GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTTTGGTCTAA
      Consensus GCAATATTGTTAACGTGAGTCTTGTAAAACCTTCTTTTTACGTTTACTCTCGTGTTAAAAATCTGAATTCTTCTAGAGTTCCTGATCTTcTGGTCTAA
```

Fig.2b: The gene encoding the Envelope(E) protein of ZC45,ZXC21RaTG13 and SARS-CoV-2.

The difference from SARS-CoV-2 to RaTG13 is 1nt, while the difference from RaTG13 to ZC45 is 2nt.

From the multalin result, we can tell that the mutation rate between RaTG13 and SARS-CoV-2 was off—the first SARS-CoV-2 isolate, WuHan-Hu-1/MN908947.1, was submitted at 12 January 2020, while the mutated forms of the E protein, QHZ00381.1, was submitted at 11 February

2020,QJA42107.1, 17 April 2020, QIS60608.1 and QIU81527.1, 15 April 2020, QIZ14355.1, 13 April 2020.

What does this mean?

By averaging the dates of all four discreet mutations, we can establish the average mutation rate of the E gene was one nucleotide substitution per 2.6 months – While the collection time of RaTG13 was allegedly at 21 July 2013—which is about 6 years and 5 months before Wuhan-HU-1, or 77 months earlier. If Natural evolution have accounted for the evolutionary distance of the E protein between SARS-COV-2 and RaTG13, we should have seen 29.6 nucleotide substitutions between RaTG13 and SARS-CoV-2, and the Protein sequence of the E protein should not be identical.

The ZC45-RaTG13 connection.

The E protein of ZC45/ZXC21 and RaTG13 are identical, and the Nucleotide sequence coding for the two proteins are also identical save for two single nucleotide substitutions. This could be the sign of shared ancestry—However, A blast search on the RaTG13/SC45 nucleotides reveal that they are comparing ZC45 and RaTG13 reveal that there were only 21597 out of 29855 nucleotides that can be aligned with each other, and of the 21597 nucleotides that can be aligned, there were only 19227 nucleotides, 89% total, that were the same.

Table 1: the BLAST result heading of RaTG13 and ZC45

Bat coronavirus RaTG13, complete genome
Sequence ID: MN996532.1Length: 29855Number of Matches: 2
Range 1: 1 to 21563GenBankGraphicsNext MatchPrevious Match
Alignment statistics for match #1
Score Expect Identities Gaps Strand
26679 bits(14447) 0.0 19227/21597(89%) 80/21597(0%) Plus/Plus

In order to deduce the chance of which such similarity between the E gene sequences being the result of natural evolution, the number of permissible mutations within the Betacoronavirus genome must first be established using the Level of protein sequence conservation, which was established by a BLAST comparison between the ORF1ab polyprotein of two betacoronaviruses of different lineages: MERS-COV and SARS-CoV.

Table 2: The BLAST result heading of MERS and SARS

orf1ab [SARS coronavirus BJ182-4]
Sequence ID: ACB69882.1Length: 7073Number of Matches: 3
Range 1: 1235 to 7072GenPeptGraphics
Next Match
Previous Match
Alignment statistics for match #1 Score Expect Method Identities Positives Gaps

Score	Expect	Method	Identities	Positives	Gaps
5999 bits(15564)	0.0	Compositional matrix adjust.	3027/5985(51%)	4019/5985(67%)	215/5985(3%)

Score	Expect	Method	Identities	Positives	Gaps
172 bits(435)	8e-38	Compositional matrix adjust.	215/877(25%)	372/877(42%)	86/877(9%)
Score	Expect	Method	Identities	Positives	Gaps
152 bits(384)	8e-32	Compositional matrix adjust.	73/155(47%)	104/155(67%)	0/155(0%)

The total number of identical Amino Acids between the two ORF1ab polyproteins= 3027+215+73=3315 out of 7073 total.

As identical amino acids typically tolerate mutation at the 3rd place of the codon, the total number of nucleotides that can tolerate mutations for Betacoronaviruses are therefore 3315+3*(7073-3315)=14589 out of 21219 nucleotides, or 68% of total.

Calculating the chance of which the evolutionary distance between ZC45/ZXC21 and RaTG13 changing only 2 nucleotides within their E genes

As the part of two genomes that can not be aligned are typically more different than the part that can be aligned, We could use a conservative estimate for the total number of nucleotide substitutions between ZC45 and RaTG13: 29855-(19227/21597)*29855=3276.21.

Using the figure of ORF1ab, the range of which these 3276.21 substitutions could land on becomes 29855*(14589/21219)=20526.63 nucleotides.

Assuming that the E proteins of Bat coronaviruses of lineage ZC45/ZXC21-RaTG13 was perfectly conserved (e.g. no amino acid substitutions are tolerated), since there was no Tryptophan(W) within the E proteins in neither proteins, and the Start codon must be ATG for Methionine, This gives a total number of places where a mutation can be accepted within the E gene being 75-1=74 nucleotides.

Getting the first two mutations to land within the E gene will require an average of 2*(20526.63/74)=544.77Substitutions, which leaves the other 3276.21-544.77=2731.44 nucleotide substitutions to land on the places other than the E gene. The chance of which all the other 2731.44 nucleotide substitutions did not land on the E gene is ((20526.63-74)/20526.63)^2731.44=5.197056e-5, or 1 in 19241.66.

In the other way, the chance of the otherwise extremely distant ZC45/ZXC21 and RaTG13 to have only 2 different nucleotides on the E gene that encodes the same exact amino acid sequence, should both strains of the viruses being the result of natural evolution, is less than 1 in 19241.66

What does this mean for Articles that uses RaTG13 as “Evidence” for the purported natural origin of SARS-CoV-2?

From prior calculation, we concluded that, due to the abnormally similar E protein genes and identical E proteins between the both geographically and phylogenetically very distant ZC45/ZXC21 and RaTG13 being nearly impossible of being the result of natural evolution, one of the viruses must be unnatural. Since RaTG13 was submitted at 27 January 2020, AFTER the outbreak, without being independently sequenced by any institutions or scientists other than the Wuhan Institute of Virology(WIV) where the sequence was first submitted, the Validity of RaTG13 can not be confirmed by independent research and should therefore be excluded from all credible researches on the origins of SARS-CoV-2.

What about MP789, the famed pangolin Coronavirus?

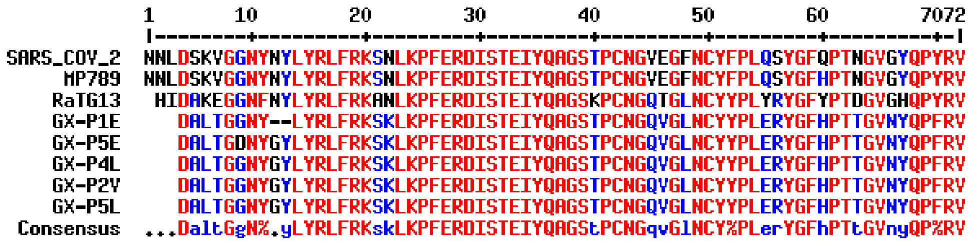
In order for a strain of virus to be considered to be valid as evidence for studies that affect policy-making decisions, the genetic sequence must be decisively concluded to be from the same virus, be a viable virus that can be physically reproduced, be independently sequenced, and it must be able to infect it's original host of it was allegedly first isolated from.

Is the MP789 virus a viable virus that can be physically reproduced?

The only sequence data for the MP789 coronavirus, MT084071.1, was submitted at 13 February 2020 by SCSFRI, Guangzhou. Again AFTER the outbreak.
A quick check on the FASTA sequence on GenBank revealed that 1872 Nucleotides out of a total of 27989, were marked as "N"—nucleotides that were missing from the complete sequence. The missing nucleotides occurred uniformly across the entire sequence, and major gaps, each more than 100nt long, splits the entire sequence into 12 long segments while up to 21 more minor gaps fragments the genomic sequence even more.
This mean that the entire MP789 sequence was fragmented and incomplete—there is no chance that such an incomplete sequence could be conceivably reproduced within any laboratories to generate a viable virus for assaying the infectivity and pathogenicity of the live virus within it's alleged original host.
The fact that the genome being incomplete, also mean that live examples of the MP789 coronavirus does not exist anywhere in the world—If such a live sample exist, the sequence should have been complete since the live example could be easily sequenced.

Can the alleged MP789 Coronavirus infect it's original host, the pangolins?

In order to answer this question, the Receptor Binding Motif(RBM) of the MP789 Coronavirus must be able to bind the pangolin ACE2 receptor—Which were never confirmed since the alleged "discovery" of the MP789 coronavirus fragments from pangolin metagenome data that were announced at 13 February 2020.
In order to find out the possibility that the MP789 coronavirus could conceivably bind to the pangolin ACE2 receptor, the RBM sequence of such a virus must be sufficiently similar to that of the existing pangolin Coronaviruses GX-P1E, GX-P5E, GX-P4L, GX-P5L and GX-P2V.



```
            1        10        20        30        40        50        60       7072
            |--------+---------+---------+---------+---------+---------+--------+-|
SARS_COV_2  NNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV
     MP789  NNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFHPTNGVGYQPYRV
    RaTG13   HIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPCNGQTGLNCYYPLYRYGFYPTDGVGHQPYRV
    GX-P1E     DALTGGNY--LYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRV
    GX-P5E     DALTGDNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRV
    GX-P4L     DALTGGNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRV
    GX-P2V     DALTGGNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRV
    GX-P5L     DALTGGNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRV
 Consensus  ...DaltGgN%.yLYRLFRKskLKPFERDISTEIYQAGStPCNGqvGlNCY%PLerYGFhPTtGVnyQP%RV
```

Fig 3: Alignment data for the RBM of GX-P1E, GX-P5E, GX-P4L, GX-P5L and GX-P2V, MP789 and SARS-CoV-2.

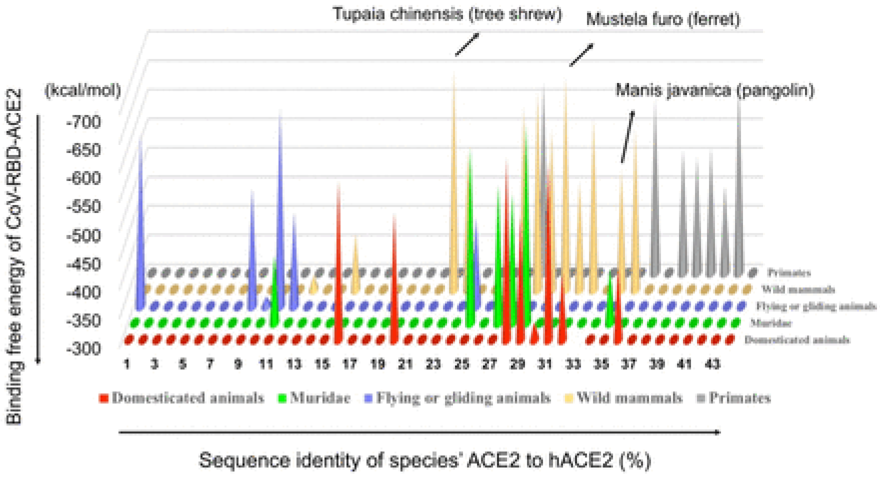
From alignment, we can clearly deduce the existence of a consensus sequence between all other known pangolin Coronaviruses before MP789, GX-P1E, GX-P5E, GX-P4L, GX-P5L and GX-P2V, the Consensus being

DALTGgNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPFRV.

This sequence is highly conserved across all pangolin Coronaviruses, with the only substitution being G447D for GX-P5E. Although GX-P1E have two amino acid deletions, it too maintained the full identity of the consensus sequence. On the rest of the RBM.

By comparing the sequences, we can tell that MP789 is only about 77% similar to the consensus sequence of all other pangolin coronaviruses—where RaTG13 also have 77% similarity to.

From a previous study[1], of which the RaTG13 RBD were docked to different animal receptors, It was concluded that the RaTG13 RBD have a binding free energy of -504.76KCal/mol to the pangolin ACE2 receptor.



The top two best binding free energies for tree shrew (−675.71 kcal/mol) and ferret (−663.57 kcal/mol) were highlighted with pangolin as a comparison (−504.76 kcal/mol). The horizonal axis represents the sequence identity of ACE2 protein of animal species with human ACE2 in amino acid level computed by BLAST tool. The vertical axis represents five animal groups classified by domesticated animals (red color), muridae (green color), flying or gliding animals (blue color), other wild mammals (sand color) and primates (grey color). The Z axis represents the binding free energies of RaTG13-CoV-RBD complexed with ACE2 proteins computed by MM/PBSA method.

Fig 4a.The binding energy of the RaTG13 RBD to receptors from different animals. The Human ACE2 Receptor scores the highest binding affinity 0f -682.62KCal/mol, followed by the Tree Shrew(-675.71 Kcal/mol) and the Ferret(-663.57Kcal/mol).

Since the similarity between the RaTG13 RBD to that of the pangolin coronaviruses was 77% and the pACE2 binding affinity was about -504.76kCal/mol, we could reasonably estimate that MP789 with the same levels of similarity, should have similar binding profiles and energies to the pACE2 receptor.

Computational study for the binding affinity of the MP789 RBD to the pACE2 receptor.

In order to further validate this hypothesis, a computational study, using the Rosetta protein structural modeling software, were conducted on the binding affinity of the MP789 RBD to the hACE2 receptor.

In order to ensure the free energy calculations are limited to Binding energies only, Chimeric hACE2/pACE2 receptors are constructed using Homology Based Modeling, by swapping out the sequence of the part of the hACE2 protein that binds the ACE2 RBD with that of the pACE2 protein. Similarily, Chimeric MP789 RBD is constructed by swapping out the Receptor Binding Motif(RBM) of the SARS-COV-2 RBD with the sequence from MP789.

The proteins were docked and the free energies of the resulting complex were minimized, before a total free energy reading was taken.

As a control, the total free energy of SARS-COV-2 and hACE2, when separated, were also measured as the standard for a binding affinity of 0.

Table 3: The total free energies of the binding experiments, in Rosetta Energy Units(R.E.U.)

Test condition	Energy(R.E.U)
SARS-COV-2-ACE2-RBD+hACE2	-522.530
Chimeric MP789-ACE2-RBD+Chimeric pACE2	-498.16
SARS-COV-2-ACE2-RBD and hACE2, separeted	-502.69

Since the canonical binding free energy of the SARS-COV-2 RBD to hACE2 was determined to be -904.76Kcal/Mol, by the same previous study[1], and the Rosetta Energy Unit scales only with total molecular mass and number of residues within a protein (of which were the same across two different experiments) according to the Rosetta manual, The scale of the R.E.U for this particular experiment was determined to be -904.76/(-522.419-(-502.690))=45.85Kcal/Mol.

Using the scale obtained from the control experiment calculation, the binding energy between MP789-RBD and pACE2 was calculated to be (-498.16-(-502.690))=4.53 R.E.U =+207.7005+-500Kcal/mol, with a maximum binding affinity of -293.2995Kcal/mol and a minumum binding energy of +707.7005 Kcal/Mol. None of which could lead to In-Vivo infection as indicated with the same computational study using Bat_CoV as a control on the hACE2 receptor.

A positive binding free energy indicate that the proteins will not dock—Which is a surprise considering the similar levels of similarity of RaTG13/pangolin Consensus and MP789/pangolin Consensus sequences.

In order to investigate further, a binding model between the SARS-CoV-2 RBD and the aforementioned chimeric pangolin ACE2 was performed (since the proteins will not dock), using PDB/6lzg as a template, in order to elucidate the reason behind the failure of the two proteins to properly dock with each other

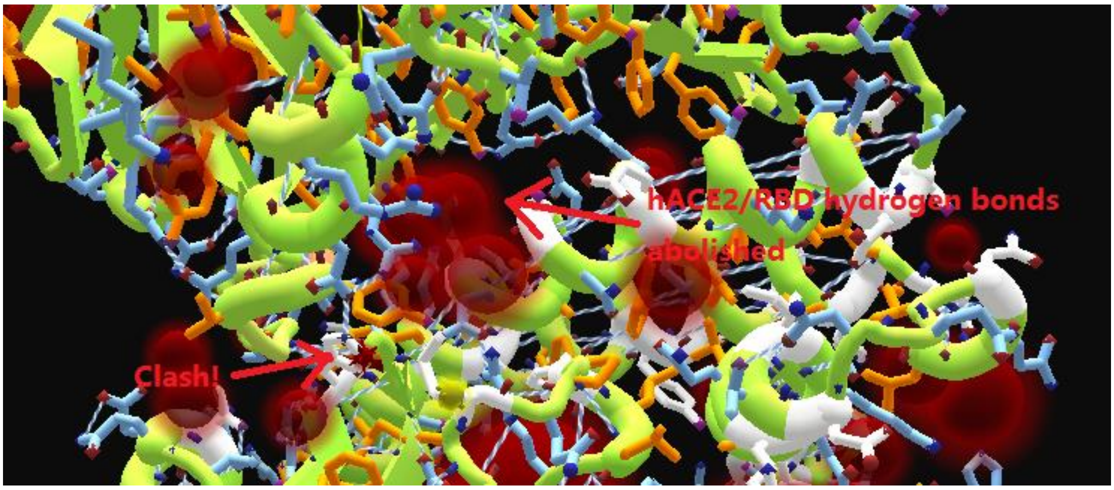


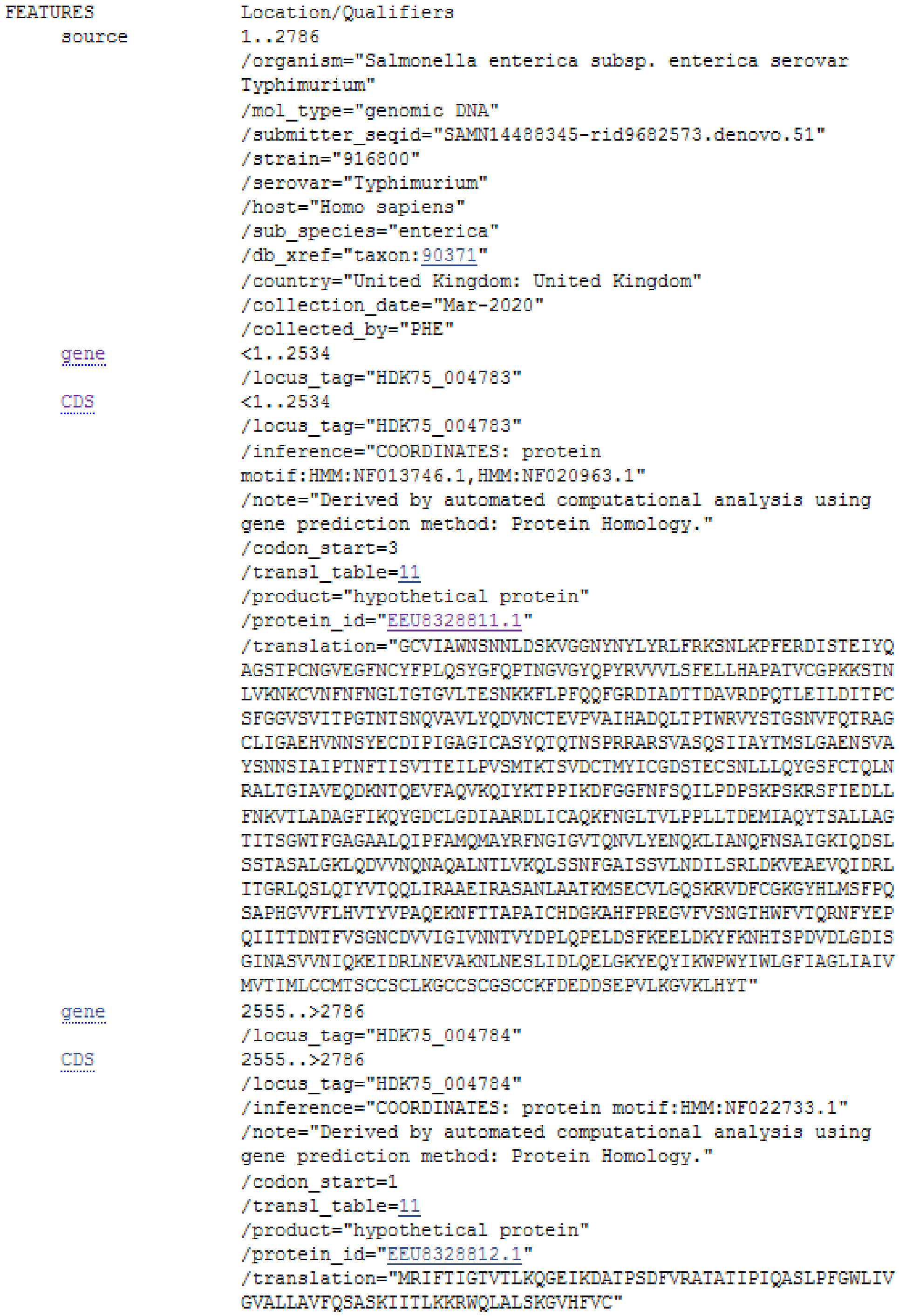
Fig. 4b: the docking conformation of the SARS-CoV-2 RBD to pACE2 receptor. In the MP789 RBD, a mutation of Q498H in MP789 further abolished one of the binding interactions between the two proteins.

From closer structural analysis, it turn out that a major clash between Y505 of SARS-COV-2/MP789 and H354 of the pangolin ACE2 receptor where a Glycine was present in the Human ACE2 receptor at the location, along with the abolishment of two(three if counting Q498H) of the four major interactions between the hACE2 and SARS-COV-2 inMP789/pACE2, completely abolishes binding of the SARS_COV_2 ACE2 RBD, and in extension, the MP789 RBD to the pangolin ACE2 receptor.

What does it mean for the research using MP789 as evidence for the origin of the ACE2 RBD of SARS-CoV-2?

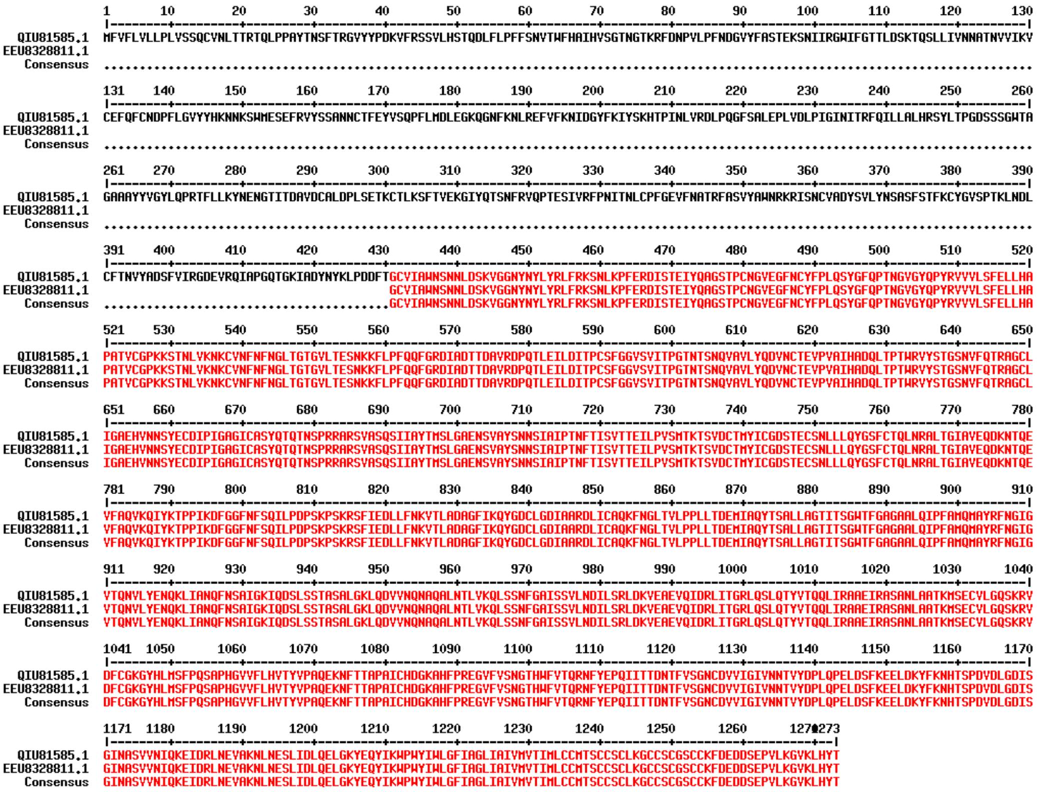
By using both homology based analysis and computational analysis, we have determined that the RBD of the MP789 Coronavirus will not bind to the pangolin ACE2 receptor in the level of affinity that would constitute an In Vivo infection for a virus with such an RBD in pangolins, this, as long with the fact that the MP789 sequence is both incomplete, fragmented and are never sequenced independently by a scientist or an institution other than the original institute who have submitted it only after the beginning of the SARS-CoV-2 outbreak, argues strongly against the validity of MP789 as an evidence for the study on the origin of SARS-CoV-2.

As metagenomic data is prone to contamination, alongside with the fact that the particular sequence was only submitted at 13 February 2020 and couldn’t have been sequenced a month earlier (as RNA is prone to degradation within tissue samples, especially once the sample have been taken out of storage and the sequencing of the sample have started), we can not rule out a condition where such a sequence may have been arisen via sample contamination in the lab by a Coronavirus RNA fragment that are similar to SARS-CoV-2, or even a direct contamination by SARS-CoV-2, which have already contaminated a sample of Salmonella Enterica Typimurium being analyzed in the U.K. in the form of Hypothetical Protein EEU8328811.1.



```
FEATURES             Location/Qualifiers
     source          1..2786
                     /organism="Salmonella enterica subsp. enterica serovar
                     Typhimurium"
                     /mol_type="genomic DNA"
                     /submitter_seqid="SAMN14488345-rid9682573.denovo.51"
                     /strain="916800"
                     /serovar="Typhimurium"
                     /host="Homo sapiens"
                     /sub_species="enterica"
                     /db_xref="taxon:90371"
                     /country="United Kingdom: United Kingdom"
                     /collection_date="Mar-2020"
                     /collected_by="PHE"
     gene            <1..2534
                     /locus_tag="HDK75_004783"
     CDS             <1..2534
                     /locus_tag="HDK75_004783"
                     /inference="COORDINATES: protein
                     motif:HMM:NF013746.1,HMM:NF020963.1"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Protein Homology."
                     /codon_start=3
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="EEU8328811.1"
                     /translation="GCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQ
                     AGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTN
                     LVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPC
                     SFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAG
                     CLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVA
                     YSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLN
                     RALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLL
                     FNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAG
                     TITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSL
                     SSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRL
                     ITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQ
                     SAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEP
                     QIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDIS
                     GINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIV
                     MVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT"
     gene            2555..>2786
                     /locus_tag="HDK75_004784"
     CDS             2555..>2786
                     /locus_tag="HDK75_004784"
                     /inference="COORDINATES: protein motif:HMM:NF022733.1"
                     /note="Derived by automated computational analysis using
                     gene prediction method: Protein Homology."
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="EEU8328812.1"
                     /translation="MRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIV
                     GVALLAVFQSASKIITLKKRWQLALSKGVHFVC"
```

Fig.5a: The original description of EEU8328811.1, which have since been removed due to being realized as being the result of contamination.



```
             1       10        20        30        40        50        60        70        80        90       100       110       120       130
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKV
EEU8328811.1 
   Consensus ..................................................................................................................................

             131    140       150       160       170       180       190       200       210       220       230       240       250       260
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 CEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQILLALHRSYLTPGDSSSGWTA
EEU8328811.1 
   Consensus ..................................................................................................................................

             261    270       280       290       300       310       320       330       340       350       360       370       380       390
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 GAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDL
EEU8328811.1 
   Consensus ..................................................................................................................................

             391    400       410       420       430       440       450       460       470       480       490       500       510       520
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 CFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHA
EEU8328811.1                                         GCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHA
   Consensus ........................................GCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHA

             521    530       540       550       560       570       580       590       600       610       620       630       640       650
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 PATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCL
EEU8328811.1 PATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCL
   Consensus PATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCL

             651    660       670       680       690       700       710       720       730       740       750       760       770       780
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 IGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE
EEU8328811.1 IGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE
   Consensus IGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE

             781    790       800       810       820       830       840       850       860       870       880       890       900       910
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIG
EEU8328811.1 VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIG
   Consensus VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIG

             911    920       930       940       950       960       970       980       990      1000      1010      1020      1030      1040
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 VTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRV
EEU8328811.1 VTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRV
   Consensus VTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRV

             1041  1050      1060      1070      1080      1090      1100      1110      1120      1130      1140      1150      1160      1170
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  QIU81585.1 DFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDIS
EEU8328811.1 DFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDIS
   Consensus DFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDIS

             1171  1180      1190      1200      1210      1220      1230      1240      1250      1260      127●273
             |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+--|
  QIU81585.1 GINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
EEU8328811.1 GINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
   Consensus GINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT
```

Fig.5b: The sequence of EEU8328811.1, in comparison with the SARS-CoV-2 Spike protein, QIU81585.1.

We concluded that all such studies on the origins of SARS-CoV-2 using the MP789 sequence as evidence should be revised on the basis of both lack of validity of the MP789 sequence and possible data contamination with SARS-CoV-2 during the sequencing of the original sample, of which were never independently sequenced since it's first publication by SCSFRI, Guangzhou.

Conclusions

By using sequence analysis and computational-based analysis, the validity of both RaTG13 and MP789 as evidence for deducing the origin of SARS-CoV-2 were discredited on the basis of both the lack of independent verifiability and the lack of credibility of the sequences on a molecular basis. Unless such samples can be independently sequenced and verified by an institution, scientist or a group of scientists without connection to nor conflict of interest with the original publisher of the sequences, any study that uses such sequences as evidence to deduce the origin of SARS-CoV-2 should be discredited and rejected for use as basis for policy-making decisions.

References:

[1] Computational analysis suggests putative intermediate animal hosts of the SARS-CoV-2
Peng Chu, Zheng Zhou, Zhichen Gao, Ruiqi Cai, Sijin Wu, Zhaolin Sun, Shuyuan Chen, Yongliang Yang
doi: https://doi.org/10.1101/2020.04.04.025080
https://www.biorxiv.org/content/10.1101/2020.04.04.025080v1

APPENDIX: The hACE2 and pACE2 sequences used by the computational study

```
>XP_017505752.1 PREDICTED: angiotensin-converting enzyme 2 [Manis javanica]
MSGSSWLLLSLVAVTAAQSTSDEEAKTFLEKFNSEAEELSYQSSLASWNYNTNITDENVQKMNVAGAKWS
TFYEEQSKIAKNYQLQNIQNDTIKRQLQALQLSGSSALSADKNQRLNTILNTMSTIYSTGKVCNPGNPQE
CSLLEPGLDNIMESSKDYNERLWAWEGWRSEVGKQLRPLYEEYVVLKNEMARANHYEDYGDYWRGDYEAE
GANGYNYSRDHLIEDVEHIFTQIKPLYEHLHAYVRAKLMDNYPSHISPTGCLPAHLLGDMWGRFWTNLYP
LTVPFRQKPNIDVTDAMVNQTWDANRIFKEAEKFFVSVGLPKMTQTFWENSMLTEPGDGRKVVCHPTAWD
LGKHDFRIKMCTKVTMDDFLTAHHEMGHIQYDMAYAMQPYLLRNGANEGFHEAVGEIMSLSAATPKHLKN
IGLLPPDFYEDNETEINFLLKQALTIVGTLPFTYMLEKWRWMVFSGQIPKEQWMKKWWEMKREIVGVVEP
VPHDETYCDPASLFHVANDYSFIRYYTRTIYQFQFQEALCQTAKHEGPLHKCDISNSAEAGQKLLQMLSL
GKSKPWTLALERVVGTKNMDVRPLLNYFEPLLTWLKEQNKNSFVGWNTDWSPYAAQSIKVRISLKSALGE
KAYEWNDSEMYLFRSSVAYAMREYFSKVKKQTIPFEDECVRVSDLKPRVSFIFFVTLPKNVSAVIPRAEV
EEAIRISRSRINDAFRLDDNSLEFLGIQPTLQPPYQPPVTIWLIVFGVVMGVVVVGIVVLIFTGIRDRKK
KDQARSEQNPYASVDLSKGENNPGFQNVDDVQTSF
```

```
>AAQ89076.1 ACE2 [Homo sapiens]
MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQNMNNAGDKWS
AFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTILNTMSTIYSTGKVCNPDNPQE
CLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLYEEYVVLKNEMARANHYEDYGDYWRGDYEVN
GVDGYDYSRGQLIEDVEHTFEEIKPLYEHLHAYVRAKLMNAYPSYISPIGCLPAHLLGDMWGRFWTNLYS
LTVPFGQKPNIDVTDAMVDQAWDAQRIFKEAEKFFVSVGLPNMTQGFWENSMLTDPGNVQKAVCHPTAW
D
LGKGDFRILMCTKVTMDDFLTAHHEMGHIQYDMAYAAQPFLLRNGANEGFHEAVGEIMSLSAATPKHLKS
IGLLSPDFQEDNETEINFLLKQALTIVGTLPFTYMLEKWRWMVFKGEIPKDQWMKKWWEMKREIVGVVEP
VPHDETYCDPASLFHVSDDYSFIRYYTRTLYQFQFQEALCQAAKHEGPLHKCDISNSTEAGQKLL
```