On The Origin of SARS-COV2 Virus

Amit K Maiti[1]

[1]Affiliation not available

June 30, 2020

Abstract

SARS-COV2 virus is originated from a closely related bat Coronavirus RaTG13 after gaining insertions by exchanged recombination with pangolin virus Pan_SL_COV_GD. SARS-COV2 uses its entry-point key residues in S1 protein to attach with ACE2 receptor to infect human. The evolution of SARS-COV2 could include any of these three possibilities: it entered human from bat early with its poorly developed entry-point residues and remained silent for long time with slower mutation rate to evade human immune system but eventually perfected them for widespread infectivity; or recently with efficiently developed entry-point residues having more infective power but adapted with higher mutation rate to evade human immune system; or recently through an intermediate host having human like conditions where it mutated both its entry-point residues as well as immune evading system to survive easily in human. RaTG13 shows 96.3% identity with SARS-COV2 genome of 29903 base implying that it substituted ˜1106 nucleotides to become present-day virus. Using pairwise sequence analysis of eighty-three SARS-COV2 genome from December, 2019, we show that its mutation rate in human is as low as 36 nucleotides per year that would take approximately 30 years to emerge as SARS-COV2 from bat RaTG13. Furthermore, a critical entry-point residue 493Q that binds with K31 residue of ACE2 is evoluted from RaTG13 amino acid Y, which needs the code must be mutated twice with an intermediate virus carrying amino acid H (Y>H>Q). However, such an intermediate COV virus with 493H has not been identified in bat or pangolin. Taken together, absence of any evidence of silent presence of SARS-COV2 in human for a long time or very high mutation rate or an intermediate host or virus emphasizes that either such an intermediate host or virus must be still obscure in nature or the emergence of SARS-COV2 is urguably doubtful.

Introduction

Novel coronavirus SARS-COV2 created pandemic by creating Covid-19 disease and believed to be originated in Wuhan, China in 2019. SARS-COV2 bears genomic identity to earlier SARS-COV virus with 79.8% and with MERS-COV virus with 59.1% (1, 2). Although Bat (*Rhinolophus affinis* from Wunnan) could be considered as a natural reservoir for this group of Coronavirus, an intermediate host of SARS-COV2 is much expected in between bat and human host. Genomic similarities from isolate of SARS-COV2 like virus from pangolin suggests that it could serve as an intermediate host (3). With a detail study comparing the genomic sequences that bears highest identities to related virus of Bat (ZC-45 (87.7%), RaTG13 (96.3%)), Pangolin (Pan-SL_CoV_GD (Guangdong, China) (91.2%), Pan_SL_CoV_GX (Guanxhi) (85.4%)) with SARS-COV2(4), they proposed that SARS-COV2 arose from Bat RaTG13 and gained three insertions in the vicinity of RBM (Receptor Binding Motif) at RBD (Receptor Binding Domain) in S1 region by exchanged recombination with Pan_SL_COV_GD genome of pangolin from Guangdong. However, due to higher dissimilarities with Pan-COV genomic sequences, they suggested that pangolin could not be an intermediate host of SARS-COV2 but RatG13 is the most probable ancestors of SARS-COV2 of human.

S (spike) protein of the SARS-COV2 virus resides on their protein coat membrane and is cleaved into two small proteins S1 and S2 by the human host enzymes. S1 forms a claw like structure and attaches with the host ACE2 (Angiotensin Converting Enzyme 2) receptor with five key entry point residues whereas S2 mediates membrane fusion with the host cell. The cleavage of the S protein occurs at the two sites:

one in between S1/S2 site by furin and other in S2 site by a serine protease, TMPRSS2 (5, 6). The critical residues 449Y, 455L, 486F, 489Y, 493Q, 500T and 501N at the RBM in RBD in S1 of SARS-COV2 binds with K31, E35, D38, M82 and K353 of human ACE2 (7). Among these residues K31-493Q and K353-501N interactions are most important for SARS-COV2 infection to human host and provide more chemically favorable interaction than SARS-COV K31-479L/N (homologue of SARS-COV2 493Q) and K353-487S/T (homologue of 501N) binding, which gave SARS-COV2 more infection power over SARS-COV (4, 7). Recently, another mutation D614G is observed only in more virulent SARS-COV2 strain that is believed to be the cause of a widespread pandemic in Europe and USA with much more infectivity(8, 9). This mutation creates an extra serine protease cleavage site at the S1/S2 junction of the spike protein and facilitate further infectivity in Caucasians with a Del C (rs35074065) genotypic background in the intergenic region between TMPRSS2 and MX1 gene (9). Zhang et al (2020) showed that 614G mutated protein reduces S1 shedding and increase infectivity (10).

Until now, it is believed that SARS-COV2 is originated in bat and gained three insertions by recombination with interchanging genetic materials from Pan_SL_COV_GD of Guangdong. For the evolution of SARS-COV2 three hypothesis can be predicted, 1) SARS-COV2 entered human early without all required mutation at these key entry-point residues at RBD with a poor efficiency and then spent silently long time in human host, adapted to evade host immune system with slower mutation rate, eventually perfected its entry-point residues and attained widespread infectivity; or 2) it gained all required mutations in those entry-point residues to infect human efficiently with widespread infectivity then adapted to evade the immune system with higher mutation rate ; or 3) entered an intermediate host from bat that have human like conditions, then entered human and adapted easily without spending long time. Here we will discuss all these possibilities by comparing their genomic sequence identities, and the existence of probable intermediate host by tracking the evolution of key entry-point residues in RBD in S1 protein. We estimated the mutation rate of SARS-COV2 in human host and calculated the time frame for evolution of SARS-COV2 from bat RaTG13 and its mutational constraints that led to select them to infect, survive and become virulent in human.

Methods

Genomic Sequences

SARS-COV2 genomic sequences are obtained from covid-19 data portal (www.covid19dataportal.org; ENA browser (European nucleotide archive) of European institute. Collection date and place of collection are recorded for each sequence, and these viral genomes are grouped by their collection date within 1^{st} and 10^{th} of each month to use for analysis so that sequences should represent gaps of at least approximately of one month. Also, in each month group, SARS-COV2 genomes those were collected in different places in the world were used to analyzed to maintain diversification. URL of each of these sequences are catalogued in **Suppl Table 1** .

Blast and Alignments

Virus genome sequences are compared for identity differences using 2-nucleotide blasts (Needleman-Wunsch Global Align Nucleotide Sequences) and are done in NCBI website using the SARS-COV2 reference genome (NC_045512, Wuhan-Hu-1). This genome has 100% identity with the genome that was collected on 12/01/2019 (MN908947). From the blast result identity differences in nucleotides are noted or counted over the gaps and other artifacts in alignments [**Suppl. Table 1**]. Average nucleotide differences are calculated for each month by using mean differences in nucleotides of all the genome collected in that month. Average nucleotide difference of a month group over the average nucleotide difference of previous month is considered the mutation rate in that month.

Global blast with 300bp flanking sequences of rs35074065 is done in Ensembl website (www.ensembl.org). ACE2 amino acid (aa) homology percentage for each animal with human is obtained from pre-aligned sequences for orthologues groups in Ensembl. Alignments of ACE2 protein sequences from all animals are done using CLUSTALW at https://npsa-prabi.ibcp.fr/.

Other Analysis and Database Information

Regulatory motifs for rs35074065 were obtained from ensemble database (www.ensembl.org). Hi-C information was obtained from UCSC database (ucsc.genome.edu)(11). Protein binding motifs are predicted at MAST (Motif Alignment and Search Tools; (http://meme-suite.org/) using the method of Bailey et al (1998) (12). eQTL and gene expression information were obtained from GTex portal (gtexportal.org). Nucleotides of SARS-COV2 sequences were translated to protein at www.expasy.ch.

Results

SARS-COV2 Could Take Approximately 30 Years to Emerge From Bat to Human Host

Estimating the time frame to evolute SARS-COV2 from RaTG13 is intricate and depends on mutation rate and other factors. Especially the Retrovirus evolution is complicated as it depends on the forces that drive the mutation rate per site nucleotide in the genome for its extra step of reverse transcription. The optional mutation rate is context dependent at which rate the errors are made during replication of the viral genome. Apart from depending on the size of the genome, it also depends on the fidelity of RDRP (RNA Directed RNA Polymerase), proofreading activity and selection pressure (13). RDRP could be very different for each Retrovirus, as for example, SARS-COV2 and Ebola RDRP are completely different (no significant similarities, data not shown) but SARS-COV2 RDRP bears considerable identity with SARS-COV (1, 2). Again, all Retrovirus do not possess proofreading activities, but Coronavirus have strong proofreading activities. Thus, a general consensus about a mutation rate in SARS-COV2 cannot be reached although the mutation rate for positive strand Retrovirus have been estimated as 10^{-4} to 10^{-6}/s(substitution)/n(nucleotide) /c (cell infection). Cell infection estimates the viral generation) (13, 14).

RaTG13 of bat is believed to be the ancestor of SARS-COV2 that bears 96.3% nucleotide identities, which overall corresponds ˜1106 nucleotides (100-96.3=3.7/100 x 29903) substitutions assuming the genome size of SARS-COV2 is 29903bases (2). Thus, a huge number (˜1106) of nucleotide substitutions occurred in RATG13 of bat to become present-day SARS-COV2 of human.

After the emergence of SARS-COV2 since December, 2019 a large number of genomic sequences are deposited in various database and several reports about their phylogeny has been elucidated(4, 15-17). Pairwise sequence analysis of eighty-three SARS-COV2 genomic sequences from collection date of December, 2019 to April 2020 by BLAST with reference genome, we calculated the average mutation rate [**Fig. 1A**] of the virus to get an estimation that how rapidly the virus was changing. The average nucleotide changes occurred ˜2 bp/month [**Fig1B**] in January to 4.89 bp/month in April. The typical average nucleotide substituted from December 2019 to April (1^{st}-10th) for 4 months is 11.94 ˜12 nucleotides. If this observed mutation after selection continues at this rate in human host, a simple extension of this calculation gives us 36 nucleotide (12 x 12/4) substitutions per year, which ultimately takes 30.7 years (1106 nucleotide/36) to evolute present-day SARS-COV2 from RaTG13 of bat.

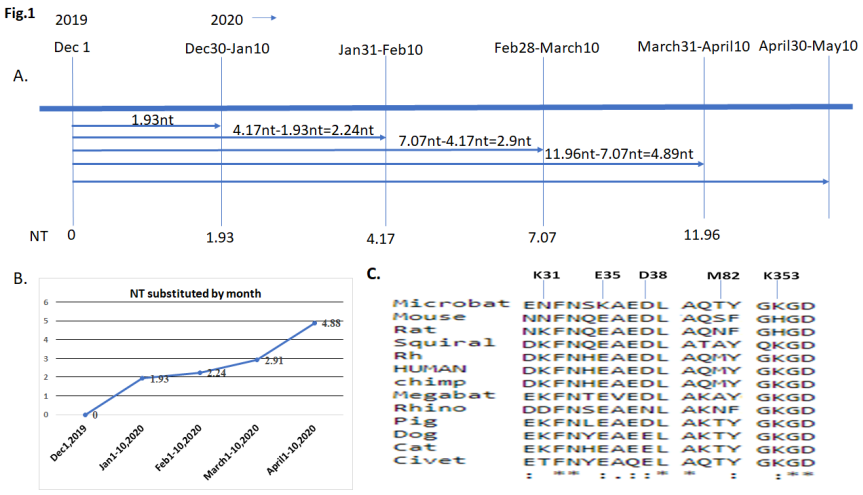


Figure 1: **Estimation of mutation rate of SARS-COV2 in human. A)** Average nucleotide differences for each month were calculated by mean of all sequences analyzed for that month. In April, it substituted 11.94 (˜12) nucleotides from December, 2019. **B**) Mutation rate is plotted against each month. **C**) Alignment of five key entry point residues in ACE2 protein of various animals. SARS-COV2 shows poor infectivity due to absence of K353 in mouse and rat.

Unavailability of Intermediate Host between Bat and Human

SARS-COV2 virus uses key entry-point residues of RBD in S1 protein to bind with the ACE2 receptor of human through K31, E35, D38, M82 and K353 (7). Among them, K31 and K353 are the most important residues for effective SARS-COV2 binding. Analysis of these residues in ACE2 receptors in various animals [**Fig. 1C**] suggests that mouse and rat possess poor ACE2 receptors (H353 in both animals instead of K353; also mouse has N31 instead of K31) for SARS-COV2 attachment (7). By cloning and infectivity experiments, they also showed that Civet cats, T31 instead of K31 but with intact K353 in ACE2 receptor allows a moderate SARS-COV2 infection but not mouse or rat (absence of K353) and indicated that K353 of ACE2 may be the most crucial residue in terms of SARS-COV2 attachment. Other animals like Chimp, Rhesus monkey, monkey, cat, dog and pig have high identity with human ACE2 receptor protein sequence [**Suppl Materials 3C**] and possess both K31 and K353 residues in their ACE2 receptor [**Fig.1C**] that could serve as an excellent attachment point for SARS-COV2 RBM and could efficiently serve as an intermediate host before infecting human. Although these animals are artificially infectible with SARS-COV2 virus, none of these animals are found to be naturally harbored any SARS-COV2 or its nearby genetically related COV virus. Thus, the conjecture remains to be elucidated whether such an intermediate host between human and bat would be existed or be explored in future in nature.

Evolution of SARS-COV2 Entry-point Residues Interacting with ACE2 Receptor

K31-493Q and K353-501N attachment site of human ACE2-SARS-COV2 respectively are the most efficient virus-host entry-point and civet cat experiment suggests that K353-501N is most crucial entry-point between these two attachment site (7). In RaTG13 of bat from where SARS-COV2 is believed to be originated, the homologue at 501N position is aa D (code GAU). Thus, an amino acid changes from D (code GAU) to N (code AAU) at this position in SARS-COV2 enables them to infect human host. D is also present at the same homologous position in pangolin virus Pan_SL_COV_GD. Thus, a single substitution in 1stcodon from G>A nucleotide could give rise aa N from aa D at the 501 position in the RBD of SARS-COV2 for K353-501N salt bridge formation and gave the important attachment site to entry into human host.

Similarly, 493Q residue in SARS-COV2 for K31-493Q interaction, which is the second most important

entry-point attachment is evoluted from amino acid Y, which is present in both RaTG13 of bat and Pan_-SL_COV_GD of pangolin and can come from either of these two virus. However, Y is coded by UAU in both animals and to become Q (code CAA) of SARS-COV2, the codon needs to mutate at least twice i.e. mutation in two nucleotides in 1st and 3rd codon. The 1st codon must be U>C mutation and the second mutation at the 3rd codon could be U>A. If the 3rd codon mutation occurred earlier than 1st codon mutation in the bat or pangolin virus, it would lead to nonsense (stop) code (UAA) and immaturely terminate S protein formation. Thus, 1st codon mutation (U>C) had to be created earlier than 3rd codon mutation for survival of this present-day virus. Eventually, 1st codon mutation (U>C) would create intermediate code CAU in ancestors of SARS-COV2 virus that would code for H (Histidine) at this position. Thus, the conversion of Y > Q had to be in the course of pathway Y >H>Q. In that case, 493H carrying intermediate ancestor virus must be existed in any of the related virus strain. Until now sequences from twenty-six types of bat and eight types of pangolin COV virus are known and analyzed (4, 8, 15, 16) but no such ancestral viral strain was identified with a 493H in RBD. Thus, besides these two animals, **there must be an intermediate host with SARS-COV2 ancestors carrying 493H virus that remains to be identified**unless the existence of such an ancestor virus still could be explored in bat or Pangolin.

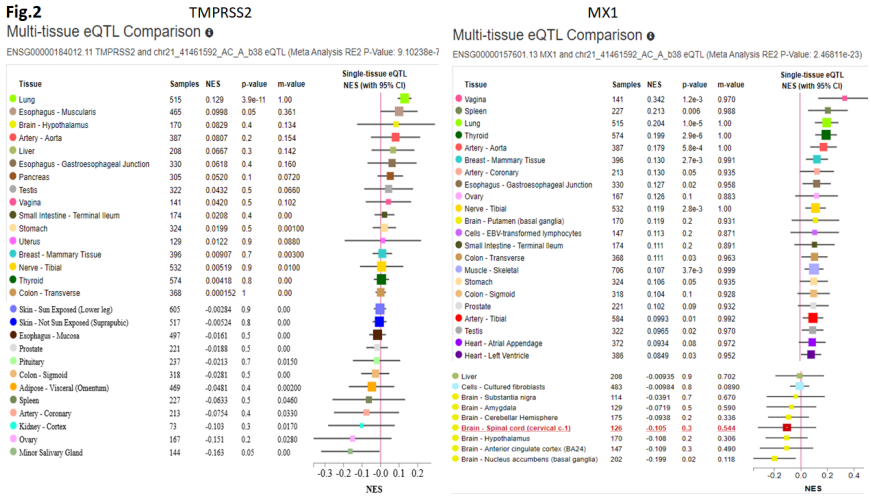


Figure 2: rs35074065 are in eQTL with MX1 and TMPRSS2 that influences expression of these genes in various human tissues. Highest expression of MX1 and TMPRESS2 can explain more infective power of a D614G SARS-COV2 strain in Del C genotype carrying patients in Caucasians.

For other remaining entry-point residues of 449Y, 455L, 486F, 489Y and 500T, three residues 455L, 489Y and 500T of SARS-COV2 are identical to both RatG13 and Pan_SL_COV_GD and did not need any nucleotides substitution. But 449Y of SARS-COV2 (RaTG1, aa F; Pan_SL_COV_GD, aa Y) could come directly from Pangolin (aaY>aaY) or by a single nucleotide substitution from bat (aaF>aaY, UUU > UAU, 2ndcodon, U>A). Similarly, for 486F (RaTG13 ,aa L; Pan_SL_COV_GD , aa F), it can directly come from Pangolin or by a single nucleotide substitution from RaTG13 (aaL > aaF, CUA >CUU, 3rd codon A>U). Thus, in 449Y and 486F both cases, a single nucleotide substitution from bat can give rise to SARS-COV2 entry-point residues or they may come directly from Pangolin by recombination(4).

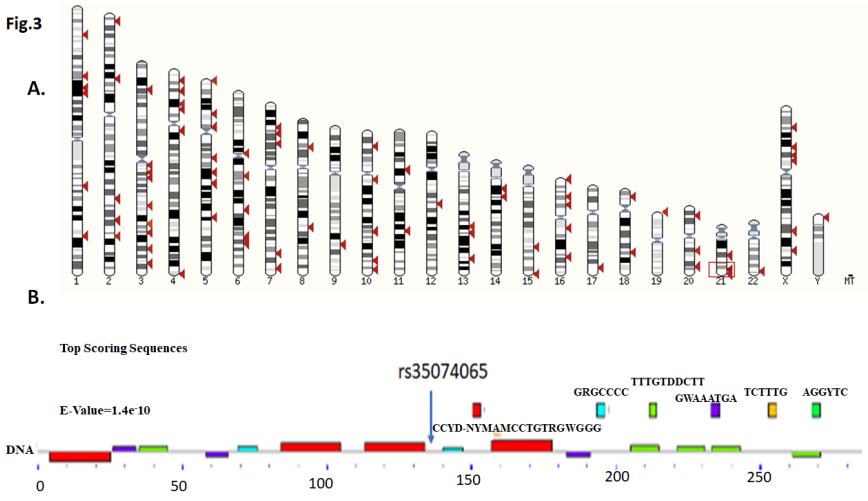


Figure 3:rs35074065 carrying genetic region contains a global regulatory element. a) Blast of 300bp flanking region of this SNP with human genome gives multiple hit. B) Important regulatory motifs are the conspicuous feature of these flanking nucleotide sequences carrying rs35074065.

Attainment of Virulence of SARS-COV2

After the emergence of SARS-COV2 in Wuhan, a strain was evolved with more infective power (8). Genomic analysis shows that this strain bears a nonsynonymous mutation (D614G) at the S1/S2 boundary that can generate extra TMPRSS2 serine protease cleavage site (9). However, it is predicted that people with an SNP (Del C) at the intergenic region between TMPRSS2 and MX1 gene apparently are infected more as this deletion is prevalent in Europe and United states and also in Indian subcontinent than other parts of the world (MAF, Minor Allele Frequency of Caucasian (CEU) 0.49; Indian 0.35; African, 0.005 and Chinese, 0.006; www.ensembl.org). This SNP is in cis- eQTL for both TMPRSS2 and MX1 gene and increase their expression in human lungs and other tissues [**Fig.2**]. Further analysis suggests that this SNP region is H3K27AC layered (ucsc.genome.edu) with regulatory region. Hi-C interactions confirms this region contains a TAD (Topologically Associated Domain) and promote interaction of this SNP region with MX1 and TMPRSS2 promoter [**Suppl. materials 3A**]. The flanking region of this SNP contain two regulatory motifs – a CTCF binding region and promoter flanking region. Immediate flanking nucleotides consist of a protein binding motif (GWAAATGA) [**Fig.3B, Suppl. Materials 3B**]. Most conspicuous feature is that 300bp flanking sequences of this SNP are identified at several genomic locations implying that these sequences may act as a global regulatory element [**Fig.3A, Suppl. Materials 2**]. It appears that Del C SNP is a strong regulatory element and modulate the expression of TMPRSS2 and MX1 gene and these proteins may have a major role in controlling the infectivity of SARS-COV2 in Caucasians and Indians. With extensive experiments, recently Zhang et al (2020) showed that 614G mutated protein increase the number of binding sites by shedding the S1 protein and increase infectivity(10).

Discussion

We investigated here the origin of SARS-COV2 virus that created pandemic in all over the world with considerable morbidity. In near future the chance of getting a vaccine is far from reality and much more morbidity is expected. It is imperative that currently people must depend on the various medicines only with a trial and error basis. To develop effective vaccines and medicines and for testing them in animals, it is necessary to know the origin of this virus and their intermediate host if any existed before its emergence as a major human infecting virus.

Among the three possibilities predicted earlier whether SARS-COV2 directly came from RaTG13 from bat

using defective entry-point residues in RBD with poor infective power with less efficiency as it is observed in civet cats (7) or forceful infection in mice (18) and remained silent for long time but highly adapted to replicate with slower mutation rate and survive in a specialized immune system of the human body. Eventually, the entry point residues have been modified and perfected to attain widespread infectivity; or it gained the efficient entry-point mutations first to bind with ACE2 receptor to enter human body and then it perfected itself to adapt with higher mutation rate and evade the human immune system; or it entered to an intermediate human like host from bat with defective entry-point residues and adapted long time, then entered human host recently and survived easily with optimum mutation fitness. In all cases, after adaptation in human host, it gained more virulence by further substitution followed by selection pressure.

Our analysis indicates the occurrence of extremely low frequency of SARS-COV2 mutation in the human host. Mutation frequency can be confounded by selection and genetic drift. In optimal mutational fitness, mutation frequency is generally biased towards nonlethal mutations and most mutations are either beneficial or neutral, thus may dramatically underestimates mutation frequency. In that case mutation rate could be lowered as the deleterious mutation drives the mutation rate lower (13). Between two models as speed vs adaptability of viral mutation rate, here it appears that SAR-COV2 evolution fits with adaptability model. Adaptability model states that after a long adaptation to evade immune system, the selection pressure is relatively low and the supply of beneficial mutation frequency is reduced, thus population favors a low mutation rate. When the mutation reaches to an optimum level simply because selection is acting on it long time within the context of immune escape to reach the maximum mutation fitness (13, 14, 19, 20).

If SARS-COV2 has come directly from bat as it is presumed, it would take a very long time to evolute as a present-day SARS-COV2 virus in the human host. Only assumption that permits this kind of viral association in the human respiratory tract by staying as a silent virus and then gained the virulence after a long time of adaptation. In the last 6 months the emergence of a new strain with more infective power has been demonstrated (10). Such a creation of a strain with more infective power also suggests that SARS-COV2 might not reside in the human host very long time without revealing its existence even in very mild form when human immune system tend to defeat its very existence.

However, our analysis has some limitations. It is unknown why the mutation rates are almost double (4.89nt) in April than other previous months. A biased sampling of a particular variant strain could represent repetitively over other low mutating strain or inclusion of a single genome consists of a 17base deletion or as expected by increasing generations in April than previous months for widespread infectivity. Although, the continuation of this increasing trend could not be verified due to unavailability of SARS-COV2 genomic sequences beyond April. Also, we wanted to assess here the average mutation rate in SARS-COV2 virus and not a strain specific by assuming all strains are capable of infecting human efficiently and undergoing substitutions to evolute to become a better strain. We also did not separate out the synonymous or nonsynonymous mutations although nonsynonymous mutation selection would have been much stringent. Another important consideration is that we did not observe any recombination or big insertions (except one that is collected in Washington in April) in these four months and frequent occurrence of those could increase the mutation rate that can occur any time. However, such an event could be very rare in an optimally mutationally fitted virus and may not add much weightage in overall mutation rate in the long run. Lastly, we estimated the mutation rate of SARS-COV2 in human host but extended it to calculate the time taken by this virus from bat RaTG13. Although such an estimation may need extensive experimental study in bat system as there would be different selection pressure than human. Nevertheless, to take less time to evolute in bat than human (<30years) could presume bat system must have higher mutation rate than human which further assume that it has to face much more challenging environment in bat than human but that is not expected as RaTG13 is native (long time adaptation) to bat.

Among the key entry-point residues in SARS-COV2 455L, 459Y and 500T are same in both RaTG13 and Pan_SL_COV_GD, thus they can come from any of them. The most important residue for SARS-COV2 interaction with human ACE2 is K353 that binds with 501N and can evolute by conversion of D (aspartic acid) to N (aspargine) by a single nucleotide mutation (G>A). It is also to be noted that a single nucleotide

mutation almost gave RaTG13 a passport to infect human efficiently.

But 493Q needs mutation in two nucleotides in 1st(U>C) and 3rd codon (U>A) sequentially either it would generate a nonsense codon (UAA). Again if 1st codon mutation occurred before 3rd codon it would code Histidine (H) by CAU. Thus, 493H carrying intermediate ancestor of SARS-COV2 virus must exists in bat or pangolin or in any other intermediate host.

However, although a genetic drift might come into play in these conversion from Y >H>Q but such a drift can occur only after entering it into human or intermediate host. The silent presence of SARS-COV2 related virus is not documented in human for long time. Also, no evidence has supported the notion that any such primate population are endangered/suffered due to a recent viral attack. The mutation must be inside a host but there is a possibility that this intermediate host no longer exists (wiped out) any more in nature or yet to be explored. Also, with the current genomic and amino acid sequences of SARS-COV2 having 493Q and 501N in the RBM suggests that SARS-COV2 could infect any of the primate or higher order mammals as intermediate host having K31 and K353 residues in their ACE2 receptor gene. But till date, none of them are shown to naturally harbor SARS-COV2 or any closely related virus. Li et al (2004) (4) also suggested that such an intermediate host can never be identified. Although, it is impossible to conclude that such an intermediate host can never be found, a systematic investigation can be continued to search for such a host.

Taken together, our analysis do not satisfy any of these conditions such as absence of any evidence of silent presence of SARS-COV2 virus in human for a long time that would take approximately 30 years to evolute as a present day SARS-COV2 or very high mutation rate or a must needed intermediate host carrying intermediate virus with 493H. Taken together, the absence of any intermediate host or virus between bat and human and inability to stay long time silently in human host also can lead to believe that SARS-COV2 would have been more easier to be created unnaturally.

Conflict of Interest: None

Funding : None

Acknowledgement

I am indebted to Mr. U. Biswas for inspiring me and helpful discussion to do this work.

References

1. Ren LL, Wang YM, Wu ZQ, Xiang ZC, Guo L, Xu T, et al. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chin Med J (Engl). 2020.

2. Zhou P, Yang XL, Wang XG, 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.

3. Zhang T, Wu Q, Zhang Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. Curr Biol. 2020;30(7):1346-51.e2.

4. Li X, Giorgi EH, Marichannegowda MH, Foley B, Xiao C, Kong X-P, et al. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci. Adv2020.

5. Hoffmann M, Kleine-Weber H, Pöhlmann S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol Cell. 2020;78(4):779-84.e5.

6. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020.

7. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol. 2020;94(7).

8. Korber B, Fisher WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRXIV; 2020.DOI: https://doi.org/10.1101/2020.04.29.069054

9. Bhattacharyya C, Das C, Ghosh A, Singh AK, Mukherjee S, Majumder PP, et al. Global Spread of

SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that

Regulate Expression of TMPRSS2 and MX1 Genes. biorXiv; 2020. DOI: https://doi.org/10.1101/2020.05.04.075911

10. Zhang L, Jackson C, Mou H, Ojha A, Rangarajan ES, Izard T, et al. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRXIV; 2020. doi: https://doi.org/10.1101/2020.06.12.148726

11. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80.

12. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics. 1998;14(1):48-54.

13. Peck KM, Lauring AS. Complexities of Viral Mutation Rates. J Virol. 2018;92(14).

14. Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. J Virol. 2010;84(19):9733-48.

15. Latinne A, Hu B, Olival KJ, Zhu G, Zhang L, Li H, et al. Origin and cross-species transmission of bat coronaviruses in China. BioRXIV2020. DOI: https://doi.org/10.1101/2020.05.31.116061

16. Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A. 2020;117(17):9241-3.

17. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, et al. Introductions and early spread of SARS-CoV-2 in the New York City area. Science. 2020.

18. Yang XH, Deng W, Tong Z, Liu YX, Zhang LF, Zhu H, et al. Mice transgenic for human angiotensin-converting enzyme 2 provide a model for SARS coronavirus infection. Comp Med. 2007;57(5):450-9.

19. Orr HA. The rate of adaptation in asexuals. Genetics. 2000;155(2):961-8.

20. Sanjuán R. Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. Philos Trans R Soc Lond B Biol Sci. 2010;365(1548):1975-82.

Suppl. Table 1 : **List of SARS-COV2 Genomes and their accession no, collection date and place of collection**

Collection Date	Reference genome	identity with reference genome	NT differences	special feature	Link with Accession No	Place of collection
12/1/1919	NC_045512.2		0	None	https://www.ebi.ac.uk/ena/browser/view/MN908947	Wuhan, China
Dec 30, 2019-Jan10, 2020						
1/1/2020	NC_045512.2	100%	0.00	None	https://www.ebi.ac.uk/ena/browser/view/MN988668	Wuhan, China
1/1/2020	NC_045512.2	100%	0.00	None	https://www.ebi.ac.uk/ena/browser/view/MN988669	Wuhan,China
1/1/2020	NC_045512.2	99%	5.00	None	https://www.ebi.ac.uk/ena/browser/view/LC522973	Japan
1/2/2020	NC_045512.2	99%	4.00	None	https://www.ebi.ac.uk/ena/browser/view/LC522972	Japan
1/1/2020	NC_045512.2	99%	2.00	None	https://www.ebi.ac.uk/ena/browser/view/MN996531	Wuhan,China
1/1/2020	NC_045512.2	99%	2.00	None	https://www.ebi.ac.uk/ena/browser/view/MN996527	Wuhan,China
12/30/2019	NC_045512.2	99%	2.00	None	https://www.ebi.ac.uk/ena/browser/view/MN996529	Wuhan,China
12/30/2019	NC_045512.2	100%	0.00	None	https://www.ebi.ac.uk/ena/browser/view/MN996528	Wuhan,China
1/1/2020	NC_045512.2	99%	3.00	None	https://www.ebi.ac.uk/ena/browser/view/LC529905	Japan
8/1/2020	NC_045512.3	100%	0.00	None	https://www.ebi.ac.uk/ena/browser/view/MT093631	China
8/1/2020	NC_045512.4	99%	4.00	None	https://www.ebi.ac.uk/ena/browser/view/LC522974	Japan
12/30/2019	NC_045512.5	100%	0.00	None	https://www.ebi.ac.uk/ena/browser/view/MN996530	Wuhan,China
12/30/2019	NC_045512.6	100%	3.00	None	https://www.ebi.ac.uk/ena/browser/view/LR757995	Wuhan,China
12/26/2019	NC_045512.7	100%	2.00	None	https://www.ebi.ac.uk/ena/browser/view/LR757998	Wuhan,China
Total NT difference			27			
Average NT difference			1.93			
std			1.73			
Feb, 1st-Feb, 2020						
2/6/2020	NC_045512.2	99%	7	None	https://www.ebi.ac.uk/ena/browser/view/MT093571	Sweeden
2/10/2020	NC_045512.2	99%	1	None	https://www.ebi.ac.uk/ena/browser/view/MT106053	CA, USA
2/1/2020	NC_045512.2	99%	2	None	https://www.ebi.ac.uk/ena/browser/view/MT365031	Hongkong
2/1/2020	NC_045512.2	99%	2	None	https://www.ebi.ac.uk/ena/browser/view/MT276597	Israel
2/10/2020	NC_045512.2	99%	6	None	https://www.ebi.ac.uk/ena/browser/view/LC528232	Japan
2/6/2020	NC_045512.2	99%	4	None	https://www.ebi.ac.uk/ena/browser/view/MT106052	CA, USA
2/2/2020	NC_045512.2	99%	2	None	https://www.ebi.ac.uk/ena/browser/view/MT106053	Japan
2/2/2020	NC_045512.2	99%	9	None	https://www.ebi.ac.uk/ena/browser/view/MT121215	sanzhai, China
2/3/2020	NC_045512.2	99%	4	None	https://www.ebi.ac.uk/ena/browser/view/LC542976	Japan
2/10/2020	NC_045512.2	99%	7	None	https://www.ebi.ac.uk/ena/browser/view/MT106054	TX, USA
2/5/2020	NC_045512.3	99%	2	None	https://www.ebi.ac.uk/ena/browser/view/MT066176	Taiwan
2/5/2020	NC_045512.4	99%	4	None	https://www.ebi.ac.uk/ena/browser/view/MT374101	Taiwan
1/31/2020	NC_045512.5	99%	1	None	https://www.ebi.ac.uk/ena/browser/view/MT039887	WI,USA
1/31/2020	NC_045512.6	99%	4	None	https://www.ebi.ac.uk/ena/browser/view/MT365032	Hongkong
1/31/2020	NC_045512.7	99%	3	None	https://www.ebi.ac.uk/ena/browser/view/MT365030	Hongkong
1/31/2020	NC_045512.8	99%	6	None	https://www.ebi.ac.uk/ena/browser/view/MT050493	Kerala, India
2/5/2020	NC_045512.9	99%	3	None	https://www.ebi.ac.uk/ena/browser/view/MT123290	GuangDong, China
2/6/2020	NC_045512.10	99%	8	None	https://www.ebi.ac.uk/ena/browser/view/MT198652	Valencia, Spain
Total NT difference			75			
Average NT difference			4.17			
std			2.46			
Differences in NT			2.24			

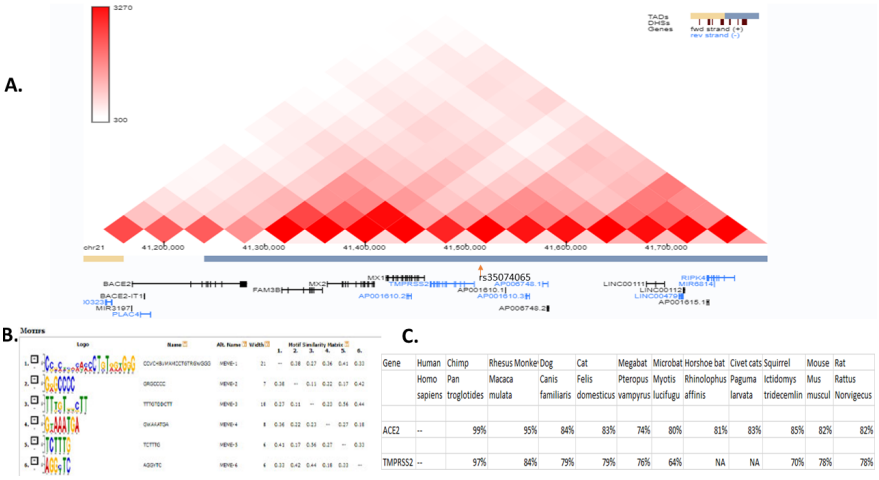
4/1/2020	NC_045512.4	99%	10.00	None	https://www.ebi.ac.uk/ena/browser/view/MT358666	WA, USA
4/1/2020	NC_045512.5	99%	12.00	None	https://www.ebi.ac.uk/ena/browser/view/MT358744	WA, USA
4/1/2020	NC_045512.6	99%	6.00	None	https://www.ebi.ac.uk/ena/browser/view/MT358743	WA, USA
4/7/2020	NC_045512.7	99%	11.00	None	https://www.ebi.ac.uk/ena/browser/view/MT375463	WA, USA
4/7/2020	NC_045512.8	99%	8.00	None	https://www.ebi.ac.uk/ena/browser/view/MT375470	WA, USA
4/1/2020	NC_045512.9	99%	11.00	None	https://www.ebi.ac.uk/ena/browser/view/MT350257	VA, USA
4/1/2020	NC_045512.10	99%	12.00	None	https://www.ebi.ac.uk/ena/browser/view/MT358650	WA, USA
4/6/2020	NC_045512.4	99%	15.00	None	https://www.ncbi.nlm.nih.gov/nuccore/MT535509.1	UT,USA
4/6/2020	NC_045512.5	99%	13.00	None	https://www.ncbi.nlm.nih.gov/nuccore/MT535508.1	UT,USA
4/5/2020	NC_045512.6	99%	12.00	None	https://www.ncbi.nlm.nih.gov/nuccore/MT535507.1	UT,USA
Total NT difference			275			
Average NT difference			11.96			
std			5.33			
Differences in NT			4.88			

Supplemenatary Materials 2: Blast of rs35074065 with human genome showing list multiple genomic regions and their locations. A brief list of alignment of this SNP flanking sequences in various genes.

Show All entries | Show/hide columns (2 hidden) | Filter

Genomic Location	Overlapping Gene(s)	Orientation	Query start	Query end	Length	Score	E-val	%ID
21:41461433-41461733		Forward	1	300	301	575.0	2.4e-164	99.67
21:28376958-28377176	AF165147.1	Forward	59	262	209	300.0	1.1e-81	88.52
18:7910294-7910442	PTPRM	Reverse	131	277	149	241.0	7.5e-64	91.28
10:129412898-129413072		Forward	59	262	204	234.0	9.9e-62	77.94
20:9392176-9392320	PLCB4	Reverse	119	262	145	221.0	1.1e-57	87.59
17:76357393-76357537	PRPSAP1	Forward	119	262	145	219.0	3.3e-57	88.28
5:32623657-32623806		Forward	105	255	152	217.0	1.5e-56	86.18
1:76302786-76302936	ST8GALNAC3	Reverse	114	262	151	217.0	1.9e-56	87.42
6:78915947-78916132	IRAK1BP1	Reverse	57	241	186	214.0	1.3e-55	79.03
13:98789000-98789152		Forward	119	262	153	209.0	3.7e-54	86.27
2:171553832-171553969	CYBRD1	Reverse	141	277	138	208.0	7.9e-54	88.41
5:127976442-127976587	LINC01184	Reverse	119	262	146	207.0	1.3e-53	86.30
13:75273753-75273892		Reverse	138	276	140	207.0	1.9e-53	87.14
6:137623626-137623767		Forward	119	259	142	206.0	2.8e-53	86.62
3:15084833-15084934	SIAH2-AS1	Forward	59	259	202	205.0	7.2e-53	74.75
X:133421377-133421502		Reverse	138	262	129	205.0	7.4e-53	92.06
3:173955822-173955964	NLGN1	Reverse	144	285	143	201.0	6.7e-52	84.62
12:67397260-67397375		Reverse	162	277	116	197.0	1.1e-50	92.24
X:37848481-37848606	AF241726.2	Reverse	138	262	126	192.0	3.7e-49	88.10
8:27113306-27113438		Forward	138	277	141	189.0	2.9e-48	83.69
5:44695293-44695415		Reverse	141	262	123	189.0	3.6e-48	88.62
5:86477287-86477432		Reverse	127	276	151	189.0	4.1e-48	82.12
14:34335840-34335963	EGLN3	Reverse	140	262	124	189.0	4.1e-48	88.71
8:100811839-100811986		Reverse	136	281	128	188.0	6.2e-48	89.06
X:39884863-39884979		Reverse	161	277	117	188.0	8.4e-48	89.74
13:76856066-76856204		Reverse	140	277	140	187.0	1.2e-47	86.43
3:96650487-96650609		Reverse	141	262	123	185.0	4.9e-47	88.62
X:19641019-19641143	SH3KBP1	Reverse	138	262	126	184.0	1.4e-46	88.10
10:126156538-126156655	ADAM12	Forward	160	277	118	184.0	1.7e-46	87.29
19:261156-261311		Forward	59	213	156	183.0	1.9e-46	80.77
10:45451795-45451954	AL731567.1	Reverse	59	213	157	183.0	2.9e-46	80.89
4:32104656-32104780	LINC02506	Forward	138	261	125	182.0	3.5e-46	87.20
7:152711809-152711925		Forward	161	277	117	180.0	2.3e-45	86.32
Y:3707860-3707976		Forward	161	277	117	178.0	1.1e-44	86.32

Supplemenatary Materials 3: Regulatory properties of rs35074065. a) Hi-C image showing the TAD that indicates the interaction of rs35074065 with TMPRSS2 and MX1 promoter. B) rs35074065 flanking DNA sequences show regulatory motifs that ay involve in gene regulation. C) Percent homology of ACE2 receptor amino acid sequences with human ACE2 protein. NA-sequence not available.



Gene	Human	Chimp	Rhesus Monkey	Dog	Cat	Megabat	Microbat	Horshoe bat	Civet cats	Squirrel	Mouse	Rat
	Homo sapiens	Pan troglotides	Macaca mulata	Canis familiaris	Felis domesticus	Pteropus vampyrus	Myotis lucifugu	Rhinolophus affinis	Paguma larvata	Ictidomys tridecemlin	Mus muscul	Rattus Norvigecus
ACE2	--	99%	95%	84%	83%	74%	80%	81%	83%	85%	82%	82%
TMPRSS2	--	97%	84%	79%	79%	76%	64%	NA	NA	70%	78%	78%