Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Virus Research 313 (2022) 198712

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Evolutionary shift from purifying selection towards divergent selection of SARS-CoV2 favors its invasion into multiple human organs

Amit K Maiti

Department of Genetics and Genomics, Mydnavar, 2645 Somerset Boulevard, Troy MI 48084, USA

ARTICLE INFO

Keywords:
SARS-CoV2
COVID-19
Mutation
Evolution
Divergent selection

ABSTRACT

SARS-CoV2 virus is believed to be originated from a closely related bat Coronavirus RaTG13 lineage and uses its key entry-point residues in S1 protein to attach with human ACE2 receptor. SARS-CoV2 could enter human from bat with its poorly developed entry-point residues much before its known appearance with slower mutation rate or recently with efficiently developed entry-point residues with higher mutation rate or through an intermediate host. Temporal analysis of SARS-CoV2 genome shows that its nucleotide substitution rate is as low as 27nt/year with an evolutionary rate of 9×10^{-4}/site/year, which is well within the range of other RNA virus (10^{-4} to 10^{-6}/site/year). TMRCA of SARS-CoV2 from bat RaTG13 lineage appears to be in between 9 and 14 years. Evolution of a critical entry-point residue Y493Q needs two substitutions with an intermediate virus carrying Y493H (Y>H>Q) but has not been identified in known twenty-nine bat CoV virus. Genetic codon analysis indicates that SARS-CoV2 evolution during propagation in human disobeys neutral evolution as nonsynonymous mutations surpass synonymous mutations with the increase of ω (d_n/d_s). Taken together, genetic data suggests that SARS-CoV2 is originated long time back before its appearance in human in 2019. Increase of ω signifies that SARs-CoV2 evolution is approaching towards diversifying selection from purifying selection predictably for its infection power to evade multiple human organs.

1. Introduction

Novel coronavirus SARS-CoV2 is believed to be originated in Wuhan, China in 2019 and has genomic identities to earlier SARS-CoV virus with 79.8% and with MERS-CoV virus with 59.1% (Ren et al., 2020; Zhou et al., 2020b). SARS-CoV2 shows highest homology with bat CoV virus RmYNO2 for most of the genome whereas shows weak homology with RBD (Receptor Binding Domain) region (Zhou et al., 2020b). Comparing highest identities of SARS-CoV2 genome with related CoV virus genome 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%)), Li et al. (2020) proposed that SARS-CoV2 arose from bat RaTG13 after gaining three insertions in the vicinity of RBM (Receptor Binding Motif) at RBD in S1 protein by exchanged recombination with Pan_SL_COV_GD of pangolin. Due to higher dissimilarities with Pan_SL_CoV_GD in other genes except RBD of S1 protein, they suggested that pangolin could not be an intermediate host of SARS-CoV2 but RaTG13 of bat (*Rhinolophus affinis*) is the most probable ancestors of SARS-CoV2 of human and, this views are supported by several phylogenetic studies (Forster et al., 2020; Gonzalez-Reiche et al., 2020; Latinne et al., 2020). Mainly based on the presence of furin cleavage site (PRRA motif) in SARS-CoV2 that is not observed in any other CoV virus, although present in some beta coronavirus (Hoffmann et al., 2020a). Andersen et al. (2020) suggested that SARS-CoV2 was originated naturally from the homologue of SARS-CoV2 that was circulating long time in the bat along with RaTG13 and, had a common ancestor. However, recent analysis assumes that RaTG13 lineage rather than RaTG13 alone could be common ancestor of SARS-CoV2 (Holmes et al., 2021).

S (spike) protein of SARS-CoV2 virus resides on their protein coat membrane and is cleaved into two small proteins S1 and S2 by the human host enzymes. The cleavage occurs at the two sites: one in between S1/S2 site by furin and other in S2 site by a serine protease, TMPRSS2 (Hoffmann et al., 2020b). S1 forms a claw like structure with seven key entry-point residues 449Y, 455L, 486F, 489Y, 493Q, 500T and, 501N at the RBM and, attaches with K31, E35, D38, M82 and K353 of the host ACE2 (Angiotensin Converting Enzyme 2) receptor (Wan et al., 2020; Wang et al., 2020) whereas S2 mediates membrane fusion with the host cell. Among these residues K31–493Q and K353–501N interactions are most important for SARS-CoV2 infection to human host and, provide more chemically favorable interactions than SARS-CoV

E-mail address: amit.maiti@mydnavar.com.

https://doi.org/10.1016/j.virusres.2022.198712
Received 15 December 2021; Received in revised form 10 February 2022; Accepted 13 February 2022
Available online 15 February 2022

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 (Wan et al., 2020).

Apart from uncertainty about the origin of SARS-CoV2, evidences are accumulating that RaTG13 of bat is zoonotically evolved to SARS-CoV2 of human without the recombination of RBD of S1 protein region (Boni et al., 2020) and, remains as a significant ancestral lineages of SARS-CoV2. Thus, an attempt to estimate evolutionary time frame, TMRCA (Time to the Most Common Recent Ancestor) from its nearest ancestor RaTG13 could give an insight into SARS-CoV2 origin which could even be longer if both share a common unknown ancestor.

Estimating the TMRCA to evolute SARS-CoV2 from RaTG13 is intricate and, depends on mutation rate with other factors. Especially, the RNA virus 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 (Temin 1989). 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 (Peck and Lauring, 2018). Although, all RNA virus RDRP do not possess proofreading activities, but Coronavirus have strong proofreading activities with very different sequences. As for example, SARS-CoV2 and Ebola RDRP are completely different in amino acid (aa) sequences (no significant similarities, data not shown) although SARS-CoV2 RDRP bears considerable identities with SARS-CoV (Ren et al., 2020; Zhou et al., 2020b). Thus, a general consensus about a mutation rate in SARS-CoV2 cannot be reached although the mutation rate for positive strand RNA virus have been estimated as 10^{-4} to 10^{-6}/s (substitution)/n(nucleotide)/c (cell infection). Cell infection estimates the viral generation (Holmes, 2009; Sanjuán et al., 2010b; Peck and Lauring, 2018).

The estimation of evolutionary time, TMRCA, from genetic data mostly are fitted to molecular clock theory (Zuckerkandl and Pauling, 1965) and later refined by neutral theory of evolution (Kimura, 1979). For RNA virus evolution, the neutral theory based on strict molecular clock appears to be correct for influenza A virus (Gojobori et al., 1990). Although, HIV1 evolution initially rejected the neutral theory (Posada and Crandall, 2001) but removal of datasets carrying recombination event from the nucleotide data showed HIV1 evolution also followed strict molecular clock (Liu et al., 2004). In oppose to molecular clock theory, rate of molecular evolution can vary over time and strict molecular clock can fail due to vast data (Lee et al., 2015) and in such cases, relaxed molecular clock could be fitted. However, such relaxed molecular clock (Ho and Duchêne 2014) like rate smoothing method (Sanderson 2002), Bayesian methods (Thorne et al., 1998; Huelsenbeck et al., 2000; Drummond and Suchard, 2010), likelihood distribution (Felsenstein, 1981; Paradis et al., 2013) or partitioned molecular clock (Thorne et al., 1998) have been widely used especially for phylogenetic separation of species. Local clocks (Yoder and Yang, 2000) or molecular dating (Lepage et al., 2007) techniques are also choice of methods that are used for species separations.

Here, I estimated the evolutionary rate of SARS-CoV2 virus in human by analyzing the nucleotide base substitutions of one sixty-six SARS-CoV2 genome from December 2019 to September 2020 in temporal manner. Comparing the evolutionary rate by synonymous amino acid changes following neutral evolution, I estimated that it would take approximately 9–14 years to appear as SARS-CoV2 in human from RaTG13 lineage of bat. Detail analysis of nucleotide evolution of RBD of S protein does not support recombination of S1 RBD region from Pan_SL_CoV_GD. Evolution of a key entry-point residue Y493Q indicates that Y must be mutated twice to give rise to Q as Y>H>Q but such an intermediate CoV virus is never identified in known twenty-nine bat CoV virus. Comparing the changes in genetic codons of synonymous and nonsynonymous amino acids I also show that SARS-CoV2 evolution from RaTG13 lineage followed strict molecular clock with neutral evolution but after its appearance in human, SARS-CoV2 does not follow neutral evolution. As the infection time progresses, the increase of ω (d_n/d_s, the rate of the proportion of nonsynonymous mutations to the synonymous mutations) value indicates that SARS-CoV2 is proceeding towards divergent selection from purifying selection predictably with the invasion of multiple organs of the human body before reaching saturation or equilibrium.

2. Materials and methods

2.1. Genomic sequences

SARS-CoV2 genomic sequences are obtained from covid-19 data portal (www.covid19dataportal.org; ENA browser, European nucleotide archive) of European institute and from NIH repository (https://www.ncbi.nlm.nih.gov/nuccore/) (Supplementary Table 1). Collection date and place of collection are recorded for each sequence and, these viral genomes are grouped by their collection date within 1st and 10th 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 were collected from different places in the world to maintain diversification.

2.2. Evolutionary tree construction and TMRCA determination

Evolutinary tree are constructed and TMRCA was determined using BEAST v1.10.4 package (Drummond and Rambaut, 2007; Drummond et al., 2012; Rambaut et al., 2018). Briefly whole genomic sequences of SARS-CoV2 (MN908947), bat CoV virus RaTG13 (MN996532.2), ZC45 (MG772933.1), ZXC21(MG772934.1), RmYNO2 (NMDC60013004–01, downloaded from China National Genome center)(Zhou et al., 2020b) and Pangolin CoV virus (Guangdong) (MT121216.1) are aligned with CLUSTAL Omega (www.ebi.ac.uk) to obtain nex format file for BEAUTI (a BEAST package). XML files were generated using strict molecular clock and relaxed molecular clock with standard parameters and GTR (Generalized Time Reversible) substitution model.

Generated XML files are loaded in BEASTS to generate log and tree files. Log files are analyzed in TRACERv1.7.2 for coalescent time and TMRCA in both model. However, graphs from software output are reconstructed for clarity. Tree files are annotated in TreeAnnotator (a BEAST package) and trees for both model were visualized in FigTree v1.4.4 (https://tree.bio.ed.ac.uk)

2.3. Blast and alignments

Virus genome sequences are compared for identity differences using 2-nucleotide blasts (Needleman-Wunsch Global Align Nucleotide Sequences) at the 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 results the numbers of nucleotide identity differences are noted or counted removing the gaps and, other artifacts in alignments (Supplementary 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 as the base substitution rate in that month. Statistical significance ($p < 0.05$) is calculated by students' *t*-test for synonymous vs nonsynonymous aa changes by the nucleotides (nt) for each month.

2.4. Synonymous and nonsynonymous assignment of the codon

Synonymous and nonsynonymous aa changes are identified using blast (https://blast.ncbi.nlm.nih.gov/) and protein translation tools (www.expasy.ch). For identifying aa changes in SARS-CoV2, multiple SARS_CoV2 genome sequences (nucleotide) from September (1st–10th) isolates were aligned with CLUSTALW (https://npsa-prabi.ibcp.fr/)

with a reference genome of SARS-CoV2 (MN908947, 29,903nt). Mismatch nucleotide (nt) sequences are traced back to the aa translation of the same reference SARS-CoV2 genome and codon changes in synonymous and nonsynonymous aa are noted and listed (Supplementary Table 2). According to the type of nucleotide changes in a codon, α (transitions, UC, CU, GA, AG), ß (transversions, AU, UA, CG, GC) and Γ (transversions, UG, GU, CA, AC) are assigned. Each nucleotide mismatch at the various positions of a codon (1st, 2nd, 3rd) are deduced for each synonymous (α, ß, Γ) and nonsynonymous (α1, ß1 and Γ1) aa changes respectively. Statistical significance ($p < 0.05$) for 3rd position to 1st and 2nd position for nt of both synonymous and nonsynonymous aa changes are calculated by students' *t*-test.

Similarly, for RaTG13-SARS-CoV2 evolution, we used poly 1ab protein region of RaTG13 (acc no: MN996532.2; 266nt-21,555nt) and SARS-CoV2 (NC_045512.2, 266nt-21,555nt) for identification of synonymous and nonsynonymous aa changes. 689 homologous aa replacements are comparable in both genomes and substituted nucleotides in each codon of synonymous (583aa) or nonsynonymous (106aa) changes are assigned as earlier as α, ß, Γ and α1, ß1 and Γ1, respectively (Supplementary Table 3).

2.5. Estimation of evolutionary rate in SARS-CoV2 in human using neutral evolution

The evolutionary rate per site of SARS-CoV2 genome is estimated following three substitution model (3ST) (Kimura, 1981). SARS-CoV2 genome of 29,903nt mostly has coding sequences of 29,244nt that codes for 9748aa. The probability of transitions as P (base transitions) is calculated as the total number of transitions in each position of the genetic code to the total number of homologous codes compared (9748aa). For example, in 1st codon position, $P = (\alpha+\alpha1)/9748$. Simultaneously, Q (a transversion type of changes) at the 1st codon position should be $(\beta+ß1)/9748$ and R (other type of transversion) at the 1st site of codon position equals to $(\Gamma+\Gamma1)/9748$. In this way, subsequently P, Q and R are calculated in other positions (2nd, 3rd) of each codon (Supplementary Table 2).

When we put the P, Q and R value in the equation of 3ST base substitution model, we obtain the total rate of base substitution (K) per site.

$$K = -(1/4)ln[(1-2P--2Q)(1-2P--2R)(1--2Q--2R)]$$

The K is calculated for each position of the codon as for 1st codon position, $K = 0.0028$; For 2nd codon position, $K = 0.0027$ and for 3rd codon position, $K = 0.0070$. Using the formula of neutral theory of evolution, $K = 2T$knuc (T, time period in years, knuc=evolutionary rate per site), so, knuc=$K/2T$ ($T = 9$ months as 0.75 years for SARS-CoV2 from January, 2020 to September, 2020), thus the evolutionary rate at 1st codon position is knuc1=0.0027/(2*0.75)=0.001787=1.78E-03/site/year.

Similarly, the knuc2 and knuc3 for 2nd and 3rd codon position should be 1.8E-03 (0.00185)/site/year and 4.7E-03 (0.00467)/site/year

2.6. Estimation of evolutionary time TMRCA from RaTG13 using neutral evolution

Evolutionary time of SARS-CoV2 from RaTG13 is estimated using neutral evolution (Kimura, 1979; Nei and Gojobori 1986). For synonymous codons, α, ß, Γ, and for nonsynonymous codons α1, ß1 and Γ1 are calculated as earlier during estimation of evolutionary rate calculation for SARS-CoV2 in human. The total base substitution for each position at the codon are calculated using the same formula for neutral evolution K= -(1/4)ln[(1- 2P - 2Q)(1–2P - 2R)(1 - 2Q - 2R)] and the K corresponds to 0.0018, 0.0069 and 0.08 for 1st, 2nd and 3rd codon respectively (Supplementary Table 3).

The evolutionary time frames, TMRCA are calculated using the evolutionary rate in each codon position that had been earlier estimated for SARS-CoV2 in human (knuc1 for 1st codon position, 0.00178/site/year, for 2nd position (knuc2), 0.00185/site/year and for 3rd position (knuc3), 0.00467/site/year) by using same formula for neutral evolution, $K = 2T$knuc.

Hence, by using same formula ($T = K/2$knuc), we determined T, the time taken to evolute 1st codon corresponds to 3.3 years (=0.0018/2*0.0018), 1.9 years (0.0069/2*0.00178) and for 3rd codon position 9.57 years (0.08/2* 0.00467).

2.7. Estimation of w (d_n/d_s)

Synonymous and nonsynonymous amino acid changes are estimated according Nei and Gozobori (1986) with Zukes and cantor (1969); Yang and Nelson (2000) and Li et al. (1985). These estimations were essentially done using software PAML (Phylogenetic Analysis by Maximum Likelihood; https://abacus.gene.ucl.ac.uk/software/paml.html). Although all methods gave similar values for ω, I focused on Nei and Gojobori (1986) estimation method (Supplementary Table 5). Briefly, for RaTG13-SARS-CoV2 conversion, this estimation based on the respective sum of all 1-nt and 2nt (16 codons) changes of all synonymous (s_{dj}) and nonsynonymous (n_{dj}) codons as Sd ($\Sigma^r j=_1$sdj) and Nd ($\Sigma^r j=_1$ndj) are determined for r number of j-th codon. The proportion of synonymous (P_s) and nonsynonymous (P_n) codons are determined by S_d/S and N_d/N when S or N are the total expected alteration site in whole sequence length. The rate of synonymous (d_s) and nonsynonymous (d_n) changes are estimated by the formula $d = -3/4$ Log e (1-(4/3)p), when d is d_s or d_n and p is p_s or p_n. ω is calculated as d_n/d_s and designated as $\omega_{\text{-rtg13-sars}}$ for RaTG13 -SARS-CoV2 analysis.

For estimation of ω between SARS-CoV2, December and September isolates in human, we created an artificial SARS-CoV2 sequence of September isolates by incorporating 126 mutations that are identified in September and compared with reference SARS-CoV2 sequence of December 2019 in PAML

For ω analysis of RBD domain of S1 protein region of human SARS-CoV2, bat RaTG13 and pangolin Pan_SL_COV_GD, the aa sequences of (438–506aa) are calculated and compared in PAML.

2.8. Protein alignment

Alignments of ACE2 protein sequences from all animals and S1 protein region of RBD are done using CLUSTALW at https://npsa-prabi.ibcp.fr/.

3. Results

3.1. Estimation of base substitution rate in SARS-CoV2 in human

Although the exact ancestor of SARS-CoV2 is currently unknown but the RaTG13 lineage in bat CoV virus is likely to be its probable ancestor. The phylogenetic tree based on whole genomic sequences of most related bat CoV virus, RaTG13, SL_ZC45, SL_ZXC21, RmYNO2 and Pan_GD (Pangolin CoV virus, Pan_SL_CoV_GD) suggests that SARS-CoV2 is separated most recently and bears highest similarities with RaTG13 (**Fig. 1a**). Moreover, both strict and relaxed molecular clock model constructed similar tree with very little branching differences emphasizing SARS-CoV2 evolution could be traced with strict molecular clock **(Fig. 1b, c).**Similarly, TMRCA of SARS-CoV2 appears to be ~40 years from the most recent ancestors with evenly distributed coalescent time **(Fig. 1d,e).**

RaTG13 shows highest 96.3% nucleotide homology with SARS-CoV2 (Zhou et al., 2020b) implying that approximately 1106nt (29,903 x 3.7 (100%–96.3%)) has to be replaced over the years to evolute to SARS-CoV2 from RaTG13 lineage or even more if they shared a common ancestor. I estimated the evolutionary time frame, TMRCA from RaTG13, using genomic sequences of SARS-CoV2 after its emergence from collection date of December 2019 to September 2020 and, temporally analyzed to get an estimation that how rapidly the virus was changing (Supplementary Table 1). Pairwise sequence analysis of one

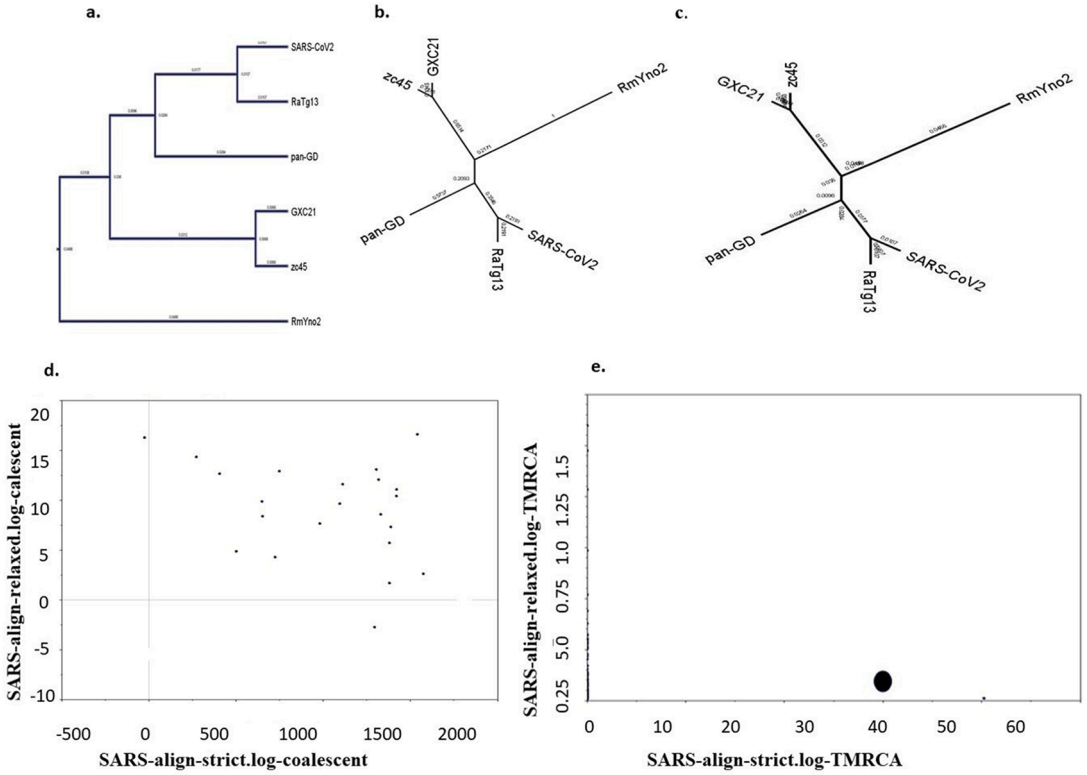


Fig. 1. Evolutinary neighbors of SARS-CoV2. a. Phylogenetic tree of SARS-CoV2 with bat and pangolin CoV virus. b. SARS-CoV2 evolutionary tree assuming strict molecular clock and c. relaxed molecular clock model. SARS-CoV2 has minimum distance of branching in both model from other neighboring bat virus ZXC-21, ZC45 and RmYNO2 and pangolin virus (Pan-GD (Guangdong); Pan_SL_CoV_GD). d. The coalescent time SARS-CoV2 took to evolute. e) The TMRCA calculated is equivalent to ~40 years from most recent ancestors.

hundred and sixty-six SARS-CoV2 genomic sequences with reference genome, I calculated the average base substitution rate (Fig. 2) of the virus. The average nucleotide changes occurred ~2.22bp/month and the typical average number of nucleotide substituted from December 2019 to September (1st–10th) for 9 months is 20.16nt (~20 nucleotides). A simple extension of this calculation of this observed base substitutions with a selection constraint at this rate in human host gives us 27nt (2.22 x 12) substitutions per year, which ultimately gives $\sim 9.00^{-4}$nt/site/yr (27nt/29,903nt/0.75 years).

Although the total average nucleotide substitution increases temporally from January 2020, the average substitution rate is almost similar in all months (Fig. 3a) suggesting that evolutionary rate of SARS-CoV2 in human are more or less constant in these months (Fig. 3a, b, c).

Estimation of synonymous and nonsynonymous substitution rates indicate that synonymous aa changes are much greater (avg synonymous aa/sites/yr (0.000591689) / avg nonsynonymous aa/sites/yr (0.000340282) =1.735) (Supplementary Table 1) than nonsynonymous aa changes. It is noted that in January, 2020 isolates when SARS-CoV2 in its initial phase of emergence, the synonymous and nonsynonymous aa changes are not significant ($p = 0.23$) (Supplementary Table 4A). After January, in all months until September 2020 p-values are highly significant ($p < 0.05$). Although this apparent discrepency is unknown, the

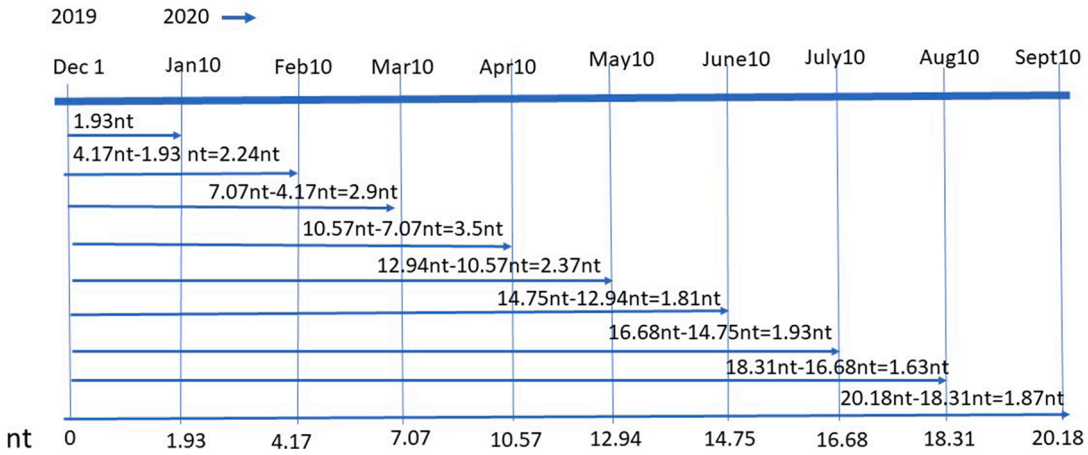
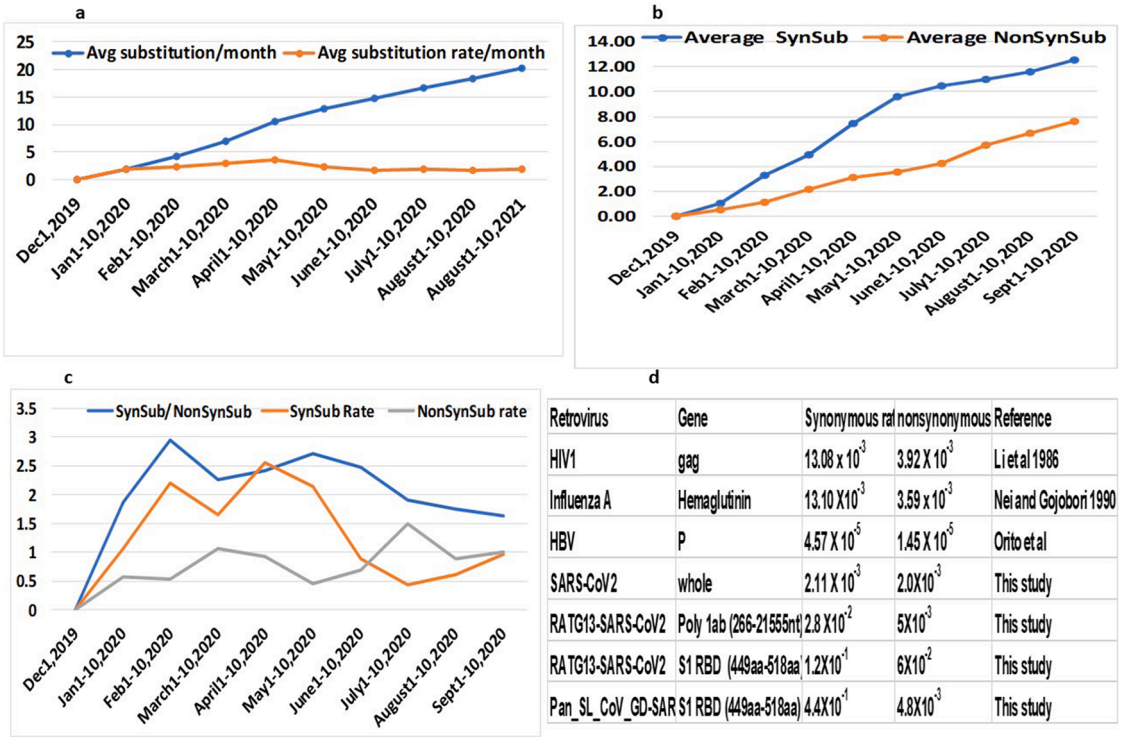


Fig. 2. Average Nucleotide changes in SARS-CoV2 from January, 2020 to September, 2020. Each genomic sequences of SARS-CoV2 that is collected within 1st 10 days of each month were compared with 2-sequences BLAST with reference genome and nucleotide differences were counted except artfacts and gaps. Rate of substitution for each months was considered by substracting the nt differences of previous month.



Retrovirus	Gene	Synonymous rat	nonsynonymous	Reference
HIV1	gag	13.08×10^{-3}	3.92×10^{-3}	Li et al 1986
Influenza A	Hemaglutinin	13.10×10^{-3}	3.59×10^{-3}	Nei and Gojobori 1990
HBV	P	4.57×10^{-5}	1.45×10^{-5}	Orito et al
SARS-CoV2	whole	2.11×10^{-3}	2.0×10^{-3}	This study
RATG13-SARS-CoV2	Poly 1ab (266-21555nt)	2.8×10^{-2}	5×10^{-3}	This study
RATG13-SARS-CoV2	S1 RBD (449aa-518aa)	1.2×10^{-1}	6×10^{-2}	This study
Pan_SL_CoV_GD-SAR	S1 RBD (449aa-518aa)	4.4×10^{-1}	4.8×10^{-3}	This study

Fig. 3. Average rate of nucleotide substitution in various months. (a)Comparison of total and rate of substitution of nt. (b)Average synonymous and nonsynonymous amino acid changes in NT level. Average synonymous substitution is always higher than nonsynonymous substitution in all months. (c) Rate of synonymous substitution is also higher in all months (except April) than nonsynonymous substition. (d)Rate of synonymous and nonsynonymous substition are compared with other RNA virus.(synSub, Synonymous substitution; NonSynSub; Nonsynonymous substitution).

greater synonymous aa changes over nonsynonymous aa changes emphasizes that SARS-CoV2 evolution is following the neutral evolution with strict molecular clock, which assumes that substitution rate will be equal to the mutation rate (Kimura, 1979).

Based on strict molecular clock, the estimation of the TMRCA to evolute SARS-CoV2 from RaTG13 indicates (1106nt/27nt) 40.96 years to replace 1106nt of RaTG13. Moreover. both Strict and relaxed molecular clock using GTR model with BEAST software also indicated similar time frame (~40 years) as TMRCA of SARS-CoV2 (Fig. 1d). Duchene et al. (2014) suggested that the evolutionary time-scales in RNA viruses require to consider substitution rates as a dynamic, rather than as a static, especially when molecular data sampled over a short timeframe that are often appear to evolve at higher rates than those sampled over a longer time period. The inadequate 'correction' of multiple substitutions at single nucleotide sites is necessary to explain the evolutionary time-scale of a RNA virus, hence three orders of magnitude lower than that estimated using virus samples should be used (Worobey et al., 2010; Ho and Duchêne, 2014). Thus, after all corrections, it would take approximately 13.6 (40.96yr/3) years to evolve SARS-CoV2 from RaTG13 lineage or more from a common ancestor at this rate of evolution that are occurring in present day SARS-CoV2 virus in human.

3.2. Estimation of evolutionary rate of SARS-CoV2 in human using strict molecular clock with neutral evolution

Molecular clock based on neutral evolution suggests that synonymous mutations that are not in a selection constraint always must be greater than nonsynonymous mutations (Kimura 1979; Gojobori et al., 1990). As the synonymous aa changes are nonconstrained, the spontaneous nucleotide changes at the 3rd position of a codon that leads to mostly synonymous aa changes and, that would be greater than the evolutionary rate of other two positions of the codon (72%) (1st position gives 5% and 2nd position is zero as all 2nd position nt changes gives rise to nonsynonymous mutation) (Nei and Gojobori 1986). I estimated the evolutionary rate of SARS-CoV2 genomes in human in September isolates (1st–10th) in comparing with SARS-CoV2 reference genome (NC_045512) by counting the synonymous and nonsynonymous aa changes. This enables that the nucleotide changes are preserved in the population after 9 months of propagation from its appearance in December 2019 and the proportion of synonymous to nonsynonymous changes are maintained. In contrast, random mutation counting would include any nt changes in these months without considering deleterious nonsynonymous mutation in a strain also abolishes synonymous mutations. For both synonymous and nonsynonymous aa, the nt changes in 3rd position is significantly higher than 1st ($p_{hsyn\text{-}3rd\text{-}1st}$= 2.24E-43; $p_{hnonsyn\text{-}3rd\text{-}1st}$= 1.17028E-07 or 2nd ($p_{hsyn\text{-}3rd\text{-}2nd}$=0; $p_{hnonsyn\text{-}3rd\text{-}2nd}$= 3.24E-07) position of the codon (Supplementary Table 4B). I identified 126 amino acid changes (63 synonymous and 63 nonsynonymous aa changes) in SARS-CoV2 genome (29,903nt, 9748aa) and estimated the evolutionary rate, knuc for the 1st (0.00178/site/year), 2nd (0.0185/site/year) and 3rd (0.00467/site/year) codon positions using the value of T as the time taken to change these aa equal to 9 months (0.75 years) (Supplementary Table 2). Obviously the knuc for 3rd position is much higher than other two positions of the codon and most changes in 3rd position leads to synonymous aa changes (Nei and Gojobori 1986). These knuc values for each codon position are used for TMRCA calculations for the transition of RaTG13 to SARS-CoV2

3.3. Estimation of evolutionary time TMRCA of SARS-CoV2 from bat RaTG13

We calculated the TMRCA to evolute SARS-CoV2 from RaTG13 lineage that would provide insight into the origin of SARS-CoV2. For comparing mismatch nucleotide sequences between two virus and estimation of TMRCA using strict molecular clock following neutral

evolution, we excluded S protein region from the whole genomic sequences of SARS-CoV2 and RaTG13 virus as this region is ambiguously believed to be inserted by recombination from Pan_SL_CoV_GD (Li et al., 2020). In addition, Liu et al. (2004) showed that when recombination regions were removed from dataset, HIV1 evolution based on only mismatch nucleotide analysis exactly followed strict molecular clock. I considered SARS-CoV2 poly 1ab gene (266nt-21,555nt, 7097aa) including RDRP gene for estimating synonymous and nonsynonymous changes of aa among 7096 aa (one is omitted due to a codon insertion). Total 689 aa are changed and comparable with 106aa and 583aa are nonsynonymous and synonymous aa respectively. The synonymous aa changes are also much and significantly higher (583/106=5.5; p = 2.0839E-197) (Supplementary Table 4B) than nonsynonymous aa changes and, could be fitted to strict molecular cloak (Kimura, 1979). The nt changes in 3rd position of the codon for both synonymous and nonsynonymous aa are to 1st ($p_{rsyn\text{-}3rd\text{-}1st}$=1.77E-304; $p_{rnonsyn\text{-}3rd\text{-}1st}$=0.00062184) and 2nd ($p_{rsyn\text{-}3rd\text{-}2nd}$=0; $p_{rnonsyn\text{-}3rd\text{-}2nd}$=3.09461E-06) position are highly significant. After estimation of the probability of transitions (P) and transversions (Q and R) of these aa and using the evolutionary rate at each position of the codon in SARS-CoV2, I determined the evolutionary time frame T (in years) following neutral evolution. The time for the nucleotide in 1st, 2nd and 3rd position of the codon are estimated using the knuc1, knuc2, knuc3 values (Fig. 4a) that are obtained for SARS-CoV2 evolution in human for each position and are 3.31 years, 1.9 years and 9.6 years respectively. As the 3rd codon replacements mostly (72%) stands for synonymous aa changes (Nei and Gojobori 1986) and evolution under zero constraints, 13.3 years ((9.6/72) X 100) could be an reasonable TMRCA to arise SARS-CoV2 from RaTG13 or more than 9.6–13.3 years from an unknown ancestor of both of them. However, this time frame is in close approximation of 13.6 years that has been estimated by using strict molecular clock based on substitution model, Thus, TMRCA determination by three independent ways (direct determination by base substitution rate, codon replacement and whole genomic sequence by BEAST software using strict and relaxed molecular clock) points out similar timeframe for SARS-CoV2 emergence (9.6–13.6) years.

3.4. SARS-CoV2 is proceeding towards divergent selection from purifying selection

I estimated the proportion of synonymous and nonsynonymous differences, ω (dn/ds) in SARS-CoV2 in human propagation and from RaTG13 to SARS-CoV2 evolution. Although the ω is below (1< ω) in both the cases (Fig. 4b) for RaTG13-SARS-CoV2 (ω-rtg13-sars =0.04) and SARS-CoV2 in human (ω-sars-hu=0.30) (Fig. 4c) signifying purifying or negative selection, but they have large differences in values. In fact, SARS-CoV2 in human has more than 7 fold (ω-sars-hu/ ω-rtg13-sars=0.30/0.04=7.5) higher ω value than RaTG13-SARS-CoV2 implying that selection constraints are tremendously driving the SARS-CoV2 evolution in human from purifying selection towards diversifying selection (approaching ω >=1) (Yang and Bielawski 2000). The purifying selection often could be a force for viral evolution as they proceed towards fixation by deleting the nonsynonymous and deleterious mutations (Kryazhimskiy and Plotkin 2008; Lin et al., 2019) and, it appears that SARS-CoV2 is proceeding such evolutionary trajectories towards diversifying selection.

3.5. Nucleotide evolution does not support recombination of S1 protein RBD region from pangolin CoV virus

RBD (438aa-506aa) of S1 protein contains key entry-point residues to attach human ACE2 receptor in SARS-CoV2 and, is believed to be originated from recombination from the Pan_SL_CoV_GD (Li et al., 2020; Patiño-Galindo et al., 2021). The main reason of such a conclusion came from aa alignment (**Fig. 5a, b)** of this region of SARS-CoV2 with Pan_SL_CoV_GD and RaTG13 where only 1 nonsynonymous amino acid differs with pangolin virus whereas 17 aa differs with RaTG13. Almost all other evidence of phylogenetic reports are based on protein

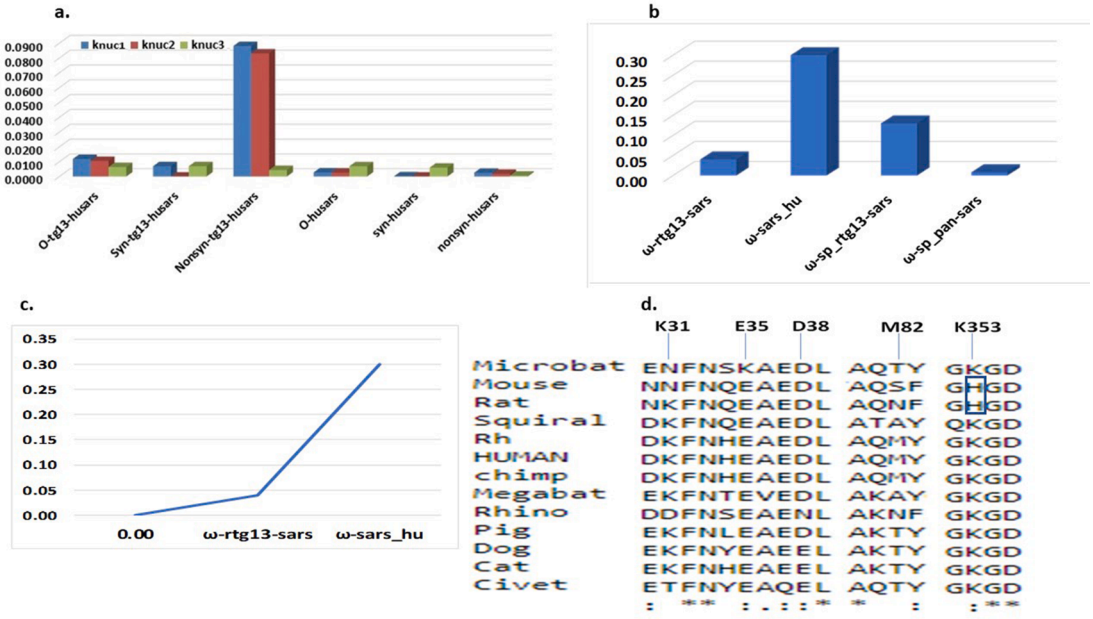


Fig. 4. Evolutionary rate in codon position and ω (d_n/d_s) value in various conversion. (a) Evolutionary rate (k_{nuc}) at each codon position (e.g. 1st (knuc1), 2nd (knuc2) and 3rd (knuc3)) in Overall (O), Ssynonymous (Syn) and nonsynonymous(Nonsyn) aa changes in SARS-CoV2 in human propagation from December, 2019-September, 2020 and from RaTG13 of bat to the appearance in human. (b) ω of various conversion are compared. It is notable that if RBD domain of S1 protein (sp) had come from pangolin virus into SARS-CoV2, Pan_SL_CoV_GD by recombination and then accumulated mutation, the ω value is very low (ω < 0.0078) compared to other conversion implying that RBD region might originated zoonotically from RaTG13 of bat. (c) Sudden increase of ω implying increase of nonsynonymous substitution over synonymous substitution in SARS-CoV2 during human propagation. (d) Alignment of five key entry point residues in ACE2 protein of various animals. SARS-CoV2 shows poor infectivity due to absence of K353 residue in mouse and rat (shown in blue box) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

sequences that made this view stronger (Forster et al., 2020; Latinne et al., 2020; Li et al., 2020). However, the nucleotide alignment in this region shows that both RaTG13 and Pan_SL_CoV_GD have 58 and 41 nucleotide changes respectively from SARS-CoV2 in 69 codons in this region (Fig. 6, Supplementary Table 5). Among 58 codons of RaTG13-SARS-CoV2, 41 codons are synonymous and 17 codons (41aa+11aa=58) are nonsynonymous whereas among 41 codons of Pan_SL_CoV_GD-SARS-CoV2, 40 codons are synonymous and only 1 codon are nonsynonymous (40aa +1aa). At the nucleotide level, if 41nt replacement can occur after recombination of Pan_SL_CoV_GD RBD region into SARS-CoV2 leading to mostly synonymous aa changes, 58nt replacement with synonymous and nonsynonymous changes could also occur naturally in RaTG13 to SARS-CoV2.

However, the large differences observed in the number of non-synonymous aa (1aa vs 17aa) in these two cases but not with synonymous aa changes (40aa vs 41aa). Although arguably it is possible to assume that S1 region could arise by recombination from Pan_SL_CoV_GD and then synonymous (neutral) mutations were accumulated in course of SARS-CoV2 evolution when neutral theory states that the rate of synonymous or silent substitutions is usually larger than that of nonsynonymous (i.e., aa-altering) substitutions (Gojobori et al., 1990). But if we observe the synonymous and nonsynonymous mutation rate in other RNA virus [Table 3d], only 2-to-3-fold differences generally occurred and, not with such a large difference as 1 vs. 17 non-synonymous aa. It is again questionable after recombination why 40 synonymous aa changes could occur in this region with only one non-synonymous aa change unless this region is refractory to nucleotide changes. But the present data of SARS-CoV2 mutation profile does not support that this region is refractory to nucleotide changes to generate nonsynonymous mutation as Long et al. (2020) and Islam et al. (2020) identified numerous nonsynonymous mutations (C480F, Y495S, L517F and G476S, V483A,Y508H respectively) in this region in SARS-CoV2. Nevertheless, when ω is calculated for this region of S1 protein (438aa-506aa), the Pan_SL_CoV_GD (after recombination followed by natural synonymous mutation) shows very low value (ω = 0.008) than other conversions (Fig. 4b, Supplementary Tables 6 and 7) (Lin et al., 2019). Such a low ω compared to other evolution cast doubt on whether this region of S1 protein at all came by recombination and, indicate that it might come from RaTG13 itself zoonotically and accumulated both 41 neutral and 17 nonsynonymous mutation.

In addition, the concept of recombination in this region of S1 protein from Pan_SL_CoV_GD ambiguously determined with Simplot analysis (Li et al., 2020) meant for similar sequence identification but not recombination although Wu et al. (2020), who first noted recombination of this region compared only several bat CoV virus with SARS-CoV2 but not with any pangolin CoV virus. Moreover, using extensive recombination breakpoint analysis at the S1 protein of SARS-CoV2, Pan_SL_CoV_GD and other bat coronavirus, Boni et al. (2020) and MacLean et al. (2021) recently showed that S1 RBD region is unlikely came from Pan_SL_CoV_GD by recombination, instead, SARS-CoV2 naturally evoluted from RaTG13 or other related bat virus.

3.6. 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 (Wan et al., 2020). Among them, K31 and, K353 of ACE2 are the most important residues for effective SARS-CoV2 binding. Analysis of these residues in ACE2 receptors in various animals suggests that mouse and, rat possess poor ACE2 receptors (H353 instead of K353; also, mouse has N31 instead of K31) (**Fig. 4d**)(Wan et al., 2020). SARS-CoV2 does not infect mouse or rat. Furthermore, cloning and infectivity experiments showed that Civet cats possessed K353 in ACE2 receptor but T31 instead of K31 and allowed a moderate SARS-CoV2 infection. These results indicate that K353 may be the most crucial residue for SARS-CoV2 attachment. Other animals like chimp, rhesus monkey, monkey, cat, dog and pig possess both K31 and, K353 residues in their ACE2 receptor as an excellent attachment point for SARS-CoV2 and can efficiently serve as an intermediate host before infecting human. Although these animals are artificially infectible with SARS-CoV2, none of them naturally harbored SARS-CoV2 or its nearby genetically related CoV virus. Thus, a conjecture remains whether an intermediate host between human and bat would be existed or be explored in future.

3.7. Evolution of SARS-CoV2 entry-point residues interacting with ACE2 receptor

K353–501N attachment site of human ACE2-SARS-CoV2 is the most efficient and crucial entry-point. In RaTG13 of bat from where SARS-CoV2 is believed to be originated, the homologue at 501N position is aaD (code GAU). An aa changes from D (code GAU) to N (code AAU) at this position in SARS-CoV2 enables them to infect human host. Thus, a single substitution in 1st codon from G>A nucleotide could give rise aaN from aaD at the 501 position in the RBD of SARS-CoV2 for K353–501N salt bridge formation for important attachment site and, almost gave RaTG13 a passport to infect human efficiently. Very recently it is

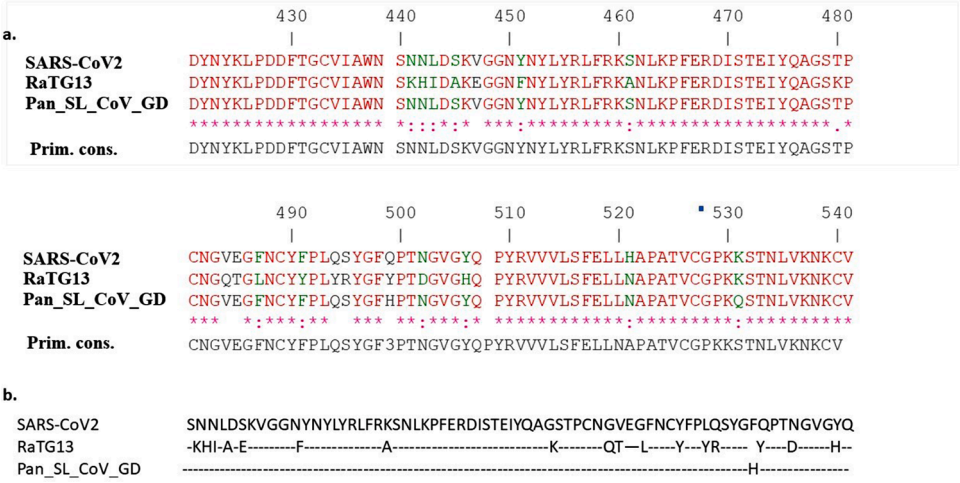


Fig. 5. Amino acid alignment of RBD of SARS-CoV2 with RaTG13 and Pan_SL_CoV_GD. (a)Color coded alignment are shown with marked amino acid residues that are mismatched. (b) The predicted inserted region are shown to emphasize that only one nonsynonymous aa is replaced if it originated from Pan_SL_GD_CoV after recombination whereas 17 amino acid are replaced if it naturally evoluted from RaTG13.- denotes identical.

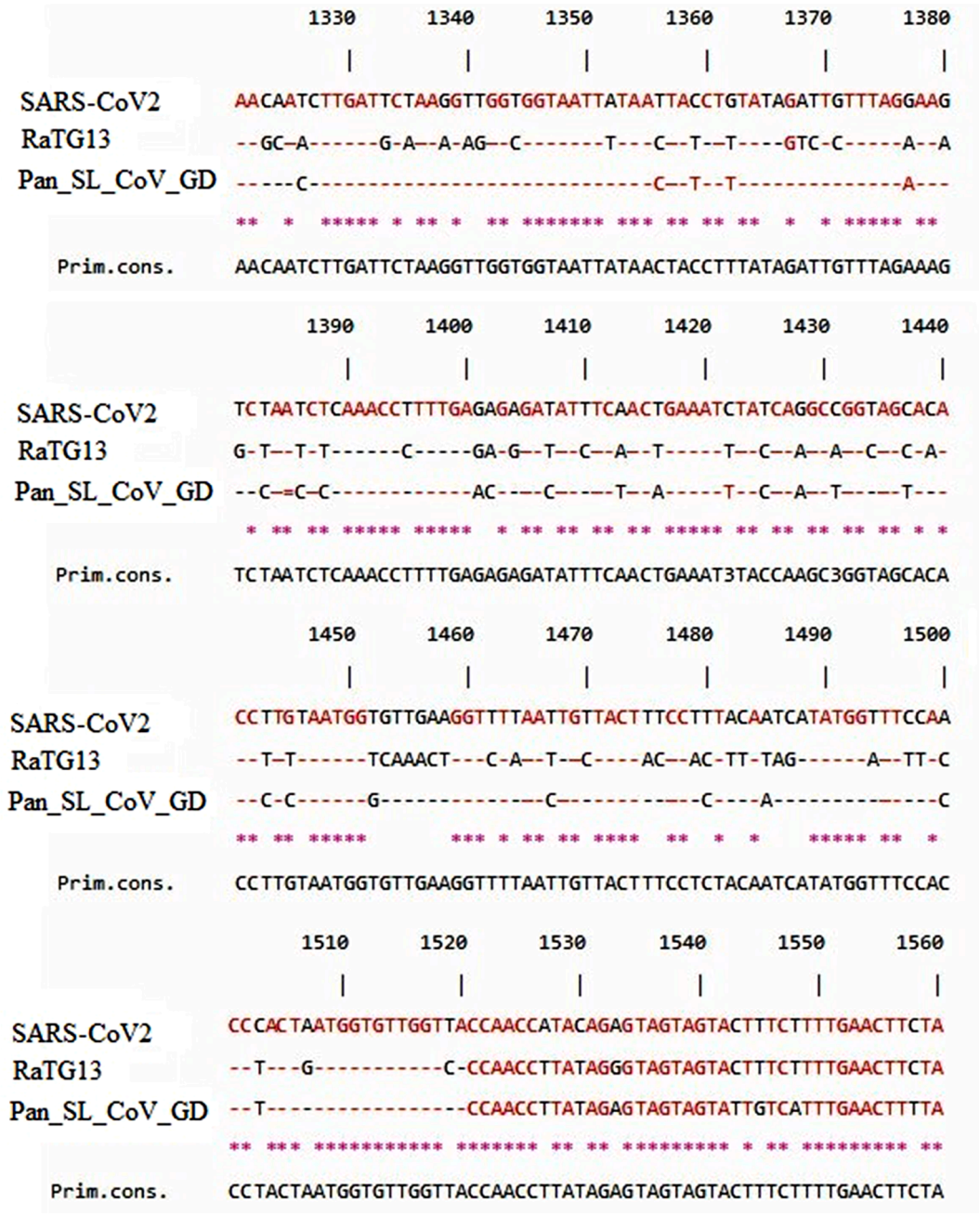


Fig. 6. Nucleotide alignment of the same RBD region of SARS-CoV2, RaTG13 and Pan_SL_CoV_GD. There are at least 28 NT changes although only one non-synonymous aa replacement implying that aa identities between SARS-CoV2 and Pan_SL_CoV_GD are apparent.

proposed that 486F residue could be highest energetic contributor for SARS-CoV2 binding (Liu et al., 2021). A series of crystalographic and modeling reports (Wang et al., 2020; Zhu et al., 2021) suggests that 493 and 501 are most important residues for SARS-CoV2 attachment to ACE2 receptor. Especially in the evolutionary point of view all SARS-CoV2 virulent strains (alpha, beta, gamma, delta, omicron) contains a new mutation in 501 residue with more infective power.

Similarly, 493Q residue in SARS-CoV2 for K31–493Q interaction, which is the second most important entry-point attachment is evolved from amino acid Y, which is present in RaTG13 of bat. However, Y is coded by UAU 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 virus, it would lead to nonsense (stop) code (UAA) and immaturely terminate S protein formation and become uninfectible. 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 bat virus strain. Until now whole genomic sequences from twenty-nine types of bat CoV virus are known and analyzed (Forster et al., 2020; Latinne et al., 2020; Li et al., 2020) but no such ancestral viral strain was identified with a 493H in RBD. Although the etiology of predicted 493H binding affinity with K31 of ACE2 receptor is unknown but presumed to have some binding activities because it had to be existed before transforming to be a 493Q residue. The possibility of simultaneously arising

double mutation cannot be ruled out but the probability of such a double mistake in a replication cycle in two neighboring nucleotides is extremely rare. 493H carrying region might come from Pan_SL_CoV_GD by recombination as believed by Li et al. (2020) but the recombination event is ambiguously projected as it is ruled out (Boni et al., 2020). Nevertheless, it also assumes that similar transformation of 493 Y>H>Q occurred in pangolin. Thus, beside known bat CoV virus, there should be a SARS-CoV2 ancestor virus carrying 493H that remains to be identified or the existence of such an ancestor virus still could be explored in bat.

For other remaining entry-point residues of 449Y, 455L, 486F, 489Y and 500T, three residues 455L, 489Y and 500T of SARS-CoV2 did not need any nucleotides substitutions. But 449Y of SARS-CoV2 needs a single nucleotide substitution from bat RaTG13 (aaF>aaY, UUU > UAU, 2nd codon, U>A). Similarly, for 486F (aaL > aaF, CUA >CUU, 3rd codon A>U). Thus, in 449Y and 486F both cases, a single nucleotide substitution from bat RaTG13 could give rise to these SARS-CoV2 entry-point residues.

4. Discussion

Effective vaccines or medicines development and, testing them in animals, it is necessary to know the origin of this virus and, how it is behaving in the human population by changing their genetic profile with evolutionary dynamics. The creation of new mutations with evolutionary trajectories driving these pandemic could be replicated *in vitro* that could predict the reversion of vaccine effect and, its virulency as shown in other RNA virus (Stern et al., 2017).

Most of the RNA virus followed neutral evolution with strict molecular clock, such as Influenza A virus (Gojobori et al., 1990), HIV1 virus (Liu et al., 2004), Rous sarcoma virus and Herpes simplex virus (Sanjuán et al., 2010b). As I observed synonymous aa changes are much greater in both SARS-CoV2 in human than from RaTG13 lineage-SARS-CoV2, the evolution of these viruses is predictably following neutral evolution with strict molecular clock. Here I also did not consider any relaxed molecular clock as I did not compare here divergence of species from related bat coronavirus nor I focused on phylogenetic relationship with other related virus. Although, it follows neutral evolution, the impact of phenotypic changes of less non-synonymous changes over synonymous aa are much greater in the evolutionary context of evolution. My analysis suggests that the mutation rate of SARS-CoV2 ($\sim$9.0 X 10^{-4}/site/year) is slightly lower than other RNA virus as it has been estimated in the order of 10^{-3} to 10^{-4}/site/year (Holmes, 2009; Duchêne et al., 2014) or 10^{-4} to 10^{-6} /site/nucleotide/cell infection (Peck and Lauring, 2018). In HIV1virus, the mutation rate for synonymous and nonsynonymous aa changes are (13.2 X10^{-3}/site/year) and, (6.9 X 10^{-3}/site/year) respectively (Li et al., 1988) (Fig. 3d).

It is surprising that SARS-CoV2 appears to be undergoing higher number of generations as it created pandemic in world human population, yet the mutation rate is lower than other RNA viruses. Thus SARS-CoV2 mutation rate could be better explained by the rate upon context dependent (Peck and Lauring, 2018). Mutation frequency can be confounded by selection and genetic drift. In that case mutation rate could be lowered as the deleterious mutation drives the mutation rate lower as a principle criteria for purifying or negative selection (Peck and Lauring, 2018). Between two models as speed vs adaptability of viral mutation rate, here it appears that SARS-CoV2 evolution fits with adaptability model, which 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 (Orr, 2000; Sanjuán, 2010a; Sanjuán et al., 2010b; Peck and Lauring, 2018).

When purifying selection occurs in the population, the ω value is low (ω <1), but in case of higher value (ω =>1) significant divergent and aggressive evolution are warrant. The ω value from RaTG13 lineage-SARS-CoV2 evolution (ω =0.04) reached to much higher value (ω=0.30) after its appearance and, propagation in human suggesting that the virus is proceeding towards divergent selection (ω =>1) from purifying selection (ω =<1). A virus undergoes divergent selection when novel conditions are created by (i) change of environment over short time, (ii) heterogenous environment with multiple niches (Elena and Sanjuán 2005). SARS-CoV2 attaches with key entry-point residues in ACE2 receptor that is highly expressed in pneumocytes in lung, endocytes in gut and nasal goblet cells but also expresses in almost all tissues in the human body with a moderate level (Ziegler et al., 2020). Thus, SARS-CoV2 has ability to invade multiple organs having ACE2 receptor expressions and faces constant challenges from tissue specific immune surveillance with multiple niches of environmental conditions. Invasion of SARS-CoV2 into various human organs is predictably increasing the nonsynonymous mutation over synonymous mutations that are persisted over the deleterious mutation and, increases its selection fitness (Elena and Sanjuán, 2005; Peck and Lauring, 2018). *In vitro* studies confirmed the attainment of high ω value for HIV1 env gene when propagated in multiple tissue specific propagation (Sanjuán et al., 2004). Indeed, it has been shown that the turnover rate of the type of infected cells (tissue specificity) is positively associated with the rate of HIV1 viral evolution (Hicks and Duffy, 2014).

Our estimation of evolutionary time taken by the SARS-CoV2 from bat RaTG13 lineage using strict molecular clock based on phylogenetic model, base substitution model and neutral evolution are in close proximation of time frame of 9.6 years to 13.6 years. SARS-CoV2 might directly came from RaTG13 of bat or through intermediate host with efficient entry-point residues but highly adapted to replicate with slower mutation rate and, survive in a specialized immune system of the human body. After adaptation in human host, it gained more virulence by further substitution followed by selection pressure. Virulency of SARS-CoV2 is further evidenced by a strain containing D614G mutation believed to be the reason of widespread infection in USA and Europe. 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 (Korber et al., 2020). Zhang et al. (2020) showed that 614G mutated protein reduces S1 shedding and increase infectivity. However, van Dorp et al. (2020) did not find any evidence of a particular strain of SARS-CoV2 that has over-infecting ability in the population although Long et al. (2020) convincingly showed the widespread infectivity of this strain in southern USA population. However, very recently a super-infective b.1.1.7 strain carries a set of spike protein mutations including D614G, N501Y and P681H (Volz et al., 2021). This N501Y (AAT->TAT) conversion is believed to provide further stronger K353–501Y attachment to enable more infective power. P681H residue in the furin cleavage site in this strain predictably influences cleavage with improved function. However, these three mutations were observed independently but not together in one strain like b.1.1.7. I also detected P681H in a strain (isolated in New Mexico, USA, in September isolate,) in several people (e.g., Acc no. MW075768) and D614G independently. It appears that other spike mutations are selected on D614G carrying background to give better adaptability in different geographical places to evade challenging human immune environment.

Y493Q conversion needs mutation in two nucleotides but could occur through a genetic drift. However, such an intermediate CoV virus are yet to be identified in bat and, also the silent presence of SARS-CoV2 related virus is neither documented in human for long time nor with any primate population that are suffered due to this viral attack (recent mink infection is by a mutated SARS-CoV2). Although, with the current genomic and aa sequences of SARS-CoV2 having 493Q and 501N in the S1 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. Li et al. (2020) also suggested that

such an intermediate host can never be identified. However, systematic investigations to search for such an intermediate host are expected in future investigations.

Based on phylogenetic analysis several reports suggests the emergence of RBD by recombination event from pangolin RBD. Similar phylogenetic and recombination breakpoint analysis also opposes the recombination event for RBD in SARS-CoV2 and suggests that it had originated naturally by mutation from its ancestor lineage (Boni et al., 2020; MacLean et al., 2021). My detail analysis with nucleotide sequences does not support recombination event in RBD region in SARS-CoV2 and favors the view that RBD of SARS-CoV2 could naturally be evolved from its ancestor lineage.

Limitations in my study may involve a biased sampling of a particular variant strain that could represent repetitively over other low mutating strain or inclusion of a single genome consists of repeated mismatch. Also, we wanted to assess here the average mutation rate in SARS-CoV2 virus undergoing substitutions to evolve to become a better strain. Another important consideration is that I did not observe any recombination or big insertions in a particular strain as frequent occurrence of those event could increase the mutation rate further. In September isolates, during substitution rate estimation, synonymous rate is higher than nonsynonymous mutations, but they are almost equal during evolutionary rate estimation as later is accounted for proportionate calculation in the same sequences. Lastly, we estimated the evolutionary rate of SARS-CoV2 in human host and, those values are used to calculate the TMRCA from bat RaTG13 lineage. Nevertheless, to take less time than our estimation time to evolute in bat than human (<9–14 years) could presume that bat system must have higher mutation rate than human which further assumes that it had to face much more challenging environment in bat than human. Although this is not expected as RaTG13 is a native virus (long time adaptation) or circulating in bat for long time (Andersen et al., 2020) in the same environment with same immune system. Although, the sequence analysis before the pandemic from RaTG13 lineage is done with available sequences in nearest ancestor lineage of SARS-CoV2 (e.g. RaTG13) and the 1st SARS-CoV2 sequences but the analysis in human is performed during the pandemic in a short time scale in terms of evolutionary perspectives. However, earlier standardized suggestions (Sanjuán et al., 2010b; Sanjuán and Domingo-Calap, 2016) are followed for evolutionary time-frame analysis to obtain conclusion as accurate as possible. It is also true that demography specific selection pressure should made analysis complicated although has not been enquired here. Lastly, we estimated mutation rate using RaTG13 sequence as an ancestor of SARS-CoV2 although several studies support that RaTG13 lineage rather than RaTG13 alone could be the ancestor of SARS-CoV2 (Holmes et al., 2021; MacLean et al., 2021). However, estimation here may not be accurate estimation for SARS-CoV2 emergence from contemporary lineage of its ancestors but provide a tractable time point from RaTG13.

SARS-CoV2 evolution comprises any of these three possibilities: it entered human host from bat early with its poorly developed entry-point residues much before its known appearance and remained silent for long time with slower mutation rate to evade human immune system or recently with efficiently developed entry-point residues having more infective power but adapted with higher mutation rate or recently through an intermediate host having human like conditions. Taken together, our analysis predicts the presence of SARS-CoV2 virus in human for a long time (9–14 years) even it could be silent because it does not satisfy any of other two conditions such as very high mutation rate of SARS-CoV2 or a must needed intermediate host carrying intermediate virus with 493H. Furthermore, after its known appearance in human, it is proceeding towards aggressive and divergent selection predictably due to invasion into multiple organs in diverse population of the world. Due to excessive constrained immune and environmental challenges, SARS-CoV2 is evolving rapidly with proportionately more nonsynonymous aa changes over synonymous aa and is acquiring excessive selective fitness to be stable to create pandemic. Moreover, nucleotide evolution suggests that SARS-CoV2 S1 RBD region is likely to be originated naturally from RaTG13 of bat without the recombination of S1 RBD region from pangolin CoV virus.

Funding information

This work is not funded

Availability of data and material (data transparency)

All data will be available after publication

Code availability (software application or custom code)

Not applicable

Ethics approval (include appropriate approvals or waivers)

NA

Consent to participate (include appropriate statements)

All authors consent to participate

Consent for publication

All author gave consent to publish

Author statement

I, Amit K Maiti, have full consent to publish this paper. There is no financial interest.

CRediT authorship contribution statement

Amit K Maiti: Conceptualization, Formal analysis, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

I am indebted to Dr. S. Gupta for critically reading the manuscript and U. Biswas for assistance in preparing the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.virusres.2022.198712.

References

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F., 2020. The proximal origin of SARS-CoV-2. Nat. Med. 26 (4), 450–452.
Boni, M.F., Lemey, P., Jiang, X., Lam, T.T., Perry, B.W., Castoe, T.A., Rambaut, A., Robertson, D.L., 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. 5 (11), 1408–1417.
Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.
Drummond, A.J., Suchard, M.A., 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 8, 114.
Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29 (8), 1969–1973.

Duchêne, S., Holmes, E.C., Ho, S.Y., 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc. Biol. Sci. 281 (1786), 1–7.
Elena, S.F., Sanjuán, R., 2005. Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. J. Virol. 79 (18), 11555–11558.
Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17 (6), 368–376.
Forster, P., Forster, L., Renfrew, C., Forster, M., 2020. Phylogenetic network analysis of SARS-CoV-2 genomes. Proc. Natl. Acad. Sci. U. S. A. 117 (17), 9241–9243.
Gojobori, T., Moriyama, E.N., Kimura, M., 1990. Molecular clock of viral evolution, and the neutral theory. Proc. Natl. Acad. Sci. U. S. A. 87 (24), 10015–10018.
Gonzalez-Reiche, A.S., Hernandez, M.M., Sullivan, M.J., Ciferri, B., Alshammary, H., Obla, A., Fabre, S., Kleiner, G., Polanco, J., Khan, Z., Alburquerque, B., van de Guchte, A., Dutta, J., Francoeur, N., Melo, B.S., Oussenko, I., Deikus, G., Soto, J., Sridhar, S.H., Wang, Y.C., Twyman, K., Kasarskis, A., Altman, D.R., Smith, M., Sebra, R., Aberg, J., Krammer, F., García-Sastre, A., Luksza, M., Patel, G., Paniz-Mondolfi, A., Gitman, M., Sordillo, E.M, Simon, V, Van Bakel, H., 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369 (6501), 297–301.
Hicks, A.L., Duffy, S., 2014. Cell tropism predicts long-term nucleotide substitution rates of mammalian RNA viruses. PLoS Pathog. 10 (1), e1003838.
Ho, S.Y., Duchêne, S., 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Mol. Ecol. 23 (24), 5947–5965.
Hoffmann, M., Kleine-Weber, H., Pöhlmann, S., 2020a. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung cells. Mol. Cell 78 (4), 779–784 e775.
Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A., Müller, M.A., Drosten, C., Pöhlmann, S., 2020b. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181 (2), 271–280.
Holmes, E.C., 2009. RNA virus genomics: a world of possibilities. J. Clin. Investig. 119 (9), 2488–2495.
Holmes, E.C., Goldstein, S.A., Rasmussen, A.L., Robertson, D.L., Crits-Christoph, A., Wertheim, J.O., Anthony, S.J., Barclay, W.S., Boni, M.F., Doherty, P.C., Farrar, J., Geoghegan, J.L., Jiang, X., Leibowitz, J.L., Neil, S.J.D., Skern, T., Weiss, S.R., Worobey, M., Andersen, K.G., Garry, R.F., Rambaut, A., 2021. The origins of SARS-CoV-2: a critical review. Cell 184 (19), 4848–4856.
Holmes, E.C., 2009. The Evolution and Emergence of RNA Viruses. Oxford University Press, New York, NY.
Huelsenbeck, J.P., Larget, B., Swofford, D., 2000. A compound poisson process for relaxing the molecular clock. Genetics 154 (4), 1879–1892.
Islam, M.R., Hoque, M.N., Rahman, M.S., Alam, A.S.M.R., Akther, M., Puspo, J.A., Akter, S., Sultana, M., Crandall, K.A., Hossain, M.A., 2020. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci. Rep. 10 (1), 14004.
Kimura, M., 1979. The Neutral Theory of Molecular Evolution. Scientific American 241 (5), 98–100.
Kimura, M., 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. U. S. A. 78 (1), 454–458.
Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., Hastie, K.M., Parker, M.D., Partridge, D.G., Evans, C.M., Freeman, T.M., de Silva, T.I., McDanal, C., Perez, L.G., Tang, H., Moon-Walker, A., Whelan, S.P., LaBranche, C.C., Saphire, E.O., Montefiori, D.C., Group, S.C.-G., 2020. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 Virus. Cell 182 (4), 812–827 e819.
Kryazhimskiy, S., Plotkin, J.B., 2008. The population genetics of dN/dS. PLos Genet. 4 (12), e1000304.
Latinne, A., Hu, B., Olival, K.J., Zhu, G., Zhang, L., Li, H., Chmura, A.A., Field, H.E., Zambrana-Torrelio, C., Epstein, J.H., Li, B., Zhang, W., Wang, L.F., Shi, Z.L., Daszak, P., 2020. Origin and cross-species transmission of bat coronaviruses in China. Nat. Commun. 11 (1), 4235.
Lee, H.J., Rodrigue, N., Thorne, J.L., 2015. Relaxing the molecular clock to different degrees for different substitution types. Mol. Biol. Evol. 32 (8), 1948–1961.
Lepage, T., Bryant, D., Philippe, H., Lartillot, N., 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24 (12), 2669–2680.
Li, W.H., Tanimura, M., Sharp, P.M., 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol. Biol. Evol. 5 (4), 313–330.
Li, W.H., Wu, C.I., Luo, C.C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. 2 (2), 150–174.
Li, X., Giorgi, E.E., Marichannegowda, M.H., Foley, B., Xiao, C., Kong, X.P., Chen, Y., Gnanakaran, S., Korber, B., Gao, F., 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci. Adv. 6 (27), eabb9153.
Lin, J.J., Bhattacharjee, M.J., Yu, C.P., Tseng, Y.Y., Li, W.H., 2019. Many human RNA viruses show extraordinarily stringent selective constraints on protein evolution. Proc. Natl. Acad. Sci. U. S. A. 116 (38), 19009–19018.
Liu, Y., Hu, G., Wang, Y., Ren, W., Zhao, X., Ji, F., Zhu, Y., Feng, F., Gong, M., Ju, X., Cai, X., Lan, J., Guo, J., Xie, M., Dong, L., Zhu, Z., Na, J., Wu, J., Lan, X., Xie, Y., Wang, X., Yuan, Z., Zhang, R., Ding, Q., 2021. Functional and genetic analysis of viral receptor ACE2 orthologs reveals a broad potential host range of SARS-CoV-2. Proc. Natl. Acad. Sci. U. S. A. 118 (12), e2025373118-27.
Liu, Y., Nickle, D.C., Shriner, D., Jensen, M.A., Learn, G.H., Mittler, J.E., Mullins, J.I., 2004. Molecular clock-like evolution of human immunodeficiency virus type 1. Virology 329 (1), 101–108.
Long, S.W., Olsen, R.J., Christensen, P.A., Bernard, D.W., Davis, J.J., Shukla, M., Nguyen, M., Saavedra, M.O., Yerramilli, P., Pruitt, L., Subedi, S., Kuo, H.C., Hendrickson, H., Eskandari, G., Nguyen, H.A.T., Long, J.H., Kumaraswami, M., Goike, J., Boutz, D., Gollihar, J., McLellan, J.S., Chou, C.W., Javanmardi, K., Finkelstein, I.J., Musser, J.M., 2020. Molecular architecture of early dissemination and massive second wave of the SARS-CoV-2 virus in a major metropolitan area. mBio 11 (6), e01031–17.
MacLean, O.A., Lytras, S., Weaver, S., Singer, J.B., Boni, M.F., Lemey, P., Kosakovsky Pond, S.L., Robertson, D.L., 2021. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLoS Biol. 19 (3), e3001115.
Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3 (5), 418–426.
Orr, H.A., 2000. The rate of adaptation in asexuals. Genetics 155 (2), 961–968.
Paradis, E., Tedesco, P.A., Hugueny, B., 2013. Quantifying variation in speciation and extinction rates with clade data. Evolution 67 (12), 3617–3627.
Patiño-Galindo, J., Filip, I., Chowdhury, R., Maranas, C., Sorger, P., AlQuraishi, M., Rabadan, R., 2021. Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2. Genome Medicine 13, 124.
Peck, K.M., Lauring, A.S., 2018. Complexities of viral mutation rates. J. Virol. 92 (14), e01031–17.
Posada, D., Crandall, K.A., 2001. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. 18 (6), 897–906.
Rambaut, A., Drummond, A.J., Xie, D., Baele, G., Suchard, M.A., 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. Syst. Biol. 67 (5), 901–904.
Ren, L.L., Wang, Y.M., Wu, Z.Q., Xiang, Z.C., Guo, L., Xu, T., Jiang, Y.Z., Xiong, Y., Li, Y. J., Li, X.W., Li, H., Fan, G.H., Gu, X.Y., Xiao, Y., Gao, H., Xu, J.Y., Yang, F., Wang, X. M., Wu, C., Chen, L., Liu, Y.W., Liu, B., Yang, J., Wang, X.R., Dong, J., Li, L., Huang, C.L., Zhao, J.P., Hu, Y., Cheng, Z.S., Liu, L.L., Qian, Z.H., Qin, C., Jin, Q., Cao, B., Wang, J.W., 2020. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chin. Med. J. (Engl.) 133 (9), 1015–1024.
Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19 (1), 101–109.
Sanjuán, R., 2010a. 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. 365 (1548), 1975–1982.
Sanjuán, R., Codoñer, F.M., Moya, A., Elena, S.F., 2004. Natural selection and the organ-specific differentiation of HIV-1 V3 hypervariable region. Evolution 58 (6), 1185–1194.
Sanjuán, R., Domingo-Calap, P., 2016. Mechanisms of viral mutation. Cell. Mol. Life Sci. 73 (23), 4433–4448.
Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010b. Viral mutation rates. J. Virol. 84 (19), 9733–9748.
Stern, A., Yeh, M.T., Zinger, T., Smith, M., Wright, C., Ling, G., Nielsen, R., Macadam, A., Andino, R., 2017. The evolutionary pathway to virulence of an RNA virus. Cell 169 (1), 35–46.e19.
Temin, H.M., 1989. Retrovirus variation and evolution. Genome 31 (1), 17–22.
Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15 (12), 1647–1657.
van Dorp, L., Richard, D., Tan, C.C.S., Shaw, L.P., Acman, M., Balloux, F., 2020. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. Nat. Commun. 11 (1), 5986.
Volz, E., Mishra, S., Chand, M., Barret, J.C., Johnsson, R., Geidelberg, L, Hindsley, W.L, COVID-19 Genomics UK (COG-UK) consortium, et al., 2021. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 593 (7858), 266–269.
Wan, Y., Shang, J., Graham, R., Baric, R.S., Li, F., 2020. Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. 94 (7), e00127–20.
Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K. Y., Zhou, H., Yan, J., Qi, J., 2020. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 181 (4), 894–904.e9.
Worobey, M., Telfer, P., Souquière, S., Hunter, M., Coleman, C.A., Metzger, M.J., Reed, P., Makuwa, M., Hearn, G., Honarvar, S., Roques, P., Apetrei, C., Kazanji, M., Marx, P.A., 2010. Island biogeography reveals the deep history of SIV. Science 329 (5998), 1487.
Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C., Zhang, Y.Z., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579 (7798), 265–269.
Yang, Z., Bielawski, J.P., 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15 (12), 496–503.
Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17 (1), 32–43.
Yoder, A.D., Yang, Z., 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17 (7), 1081–1090.
Zhang, L., Jackson, C.B., Mou, H., Ojha, A., Peng, H., Quinlan, B.D., Rangarajan, E.S., Pan, A., Vanderheiden, A., Suthar, M.S., Li, W., Izard, T., Rader, C., Farzan, M., Choe, H., 2020. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat. Commun. 11 (1), 6013.
Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E. C., Hughes, A.C., Bi, Y., Shi, W., 2020b. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. Curr. Biol. 30 (11), 2196–2203 e2193.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., Shi, Z.L., 2020b. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579 (7798), 270–273.

Zhu, X., Mannar, D., Srivastava, S.S., Berezuk, A.M., Demers, J.P., Saville, J.W., Leopold, K., Li, W., Dimitrov, D.S., Tuttle, K.S., Zhou, S., Chittori, S., Subramaniam, S., 2021. Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. PLoS Biol. 19 (4), e3001237.

Ziegler, C.G.K., Allon, S.J., Nyquist, S.K., Mbano, I.M., Miao, V.N., Tzouanas, C.N., Cao, Y., Yousif, A.S., Bals, J., Hauser, B.M., Feldman, J., Muus, C., Wadsworth, M.H., Kazer, S.W., Hughes, T.K., Doran, B., Gatter, G.J., Vukovic, M., Taliaferro, F., Mead, B.E., Guo, Z., Wang, J.P., Gras, D., Plaisant, M., Ansari, M., Angelidis, I., Adler, H., Sucre, J.M.S., Taylor, C.J., Lin, B., Waghray, A., Mitsialis, V., Dwyer, D.F., Buchheit, K.M., Boyce, J.A., Barrett, N.A., Laidlaw, T.M., Carroll, S.L., Colonna, L., Tkachev, V., Peterson, C.W., Yu, A., Zheng, H.B., Gideon, H.P., Winchell, C.G., Lin, P. L., Bingle, C.D., Snapper, S.B., Kropski, J.A., Theis, F.J., Schiller, H.B., Zaragosi, L.E., Barbry, P., Leslie, A., Kiem, H.P., Flynn, J.L., Fortune, S.M., Berger, B., Finberg, R.W., Kean, L.S., Garber, M., Schmidt, A.G., Lingwood, D., Shalek, A.K., Ordovas-Montanes, J., 2020. SARS-CoV-2 receptor ACE2 is an interferon-stimulated gene in human airway epithelial cells and is detected in specific cell subsets across tissues. Cell 181 (5), 1016–1035 e1019.

Zuckerkandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8 (2), 357–366.

Zukes, T.H., Cantor, C.R., Munro, H.N., 1969. Evolution of protein molecules. In: Mammalian Protein Metabolism III, 2. Academic Press, New York, pp. 1–132.