Identification of the host reservoir of SARS-CoV-2 and determining when it spilled over into humans

Vidyavathi Pamjula, Norval J.C Strachan and Francisco J. Perez-Reche

The School of Natural and Computing Sciences - Department of Physics,
University of Aberdeen, Scotland, UK.

Vidyavathi Pamjula: v.pamjula.19@abdn.ac.uk, PhD student. Corresponding author.
Norval J.C Strachan: n.strachan@abdn.ac.uk, Emeritus Professor.
Francisco J. Perez-Reche: fperez-reche@abdn.ac.uk, Reader.

1 Abstract

Since the emergence of SARS-CoV-2 in Wuhan in 2019 its host reservoir has not been established. Phylogenetic analysis was performed on whole genome sequences (WGS) of 71 coronaviruses and a Breda virus. A subset comprising two SARS-CoV-2 Wuhan viruses and 8 of the most closely related coronavirus sequences were used for host reservoir analysis using Bayesian Evolutionary Analysis Sampling Trees (BEAST). Within these genomes, 20 core genome fragments were combined into 2 groups each with similar clock rates (5.9×10^{-3} and 1.1×10^{-3} subs/site/year). Pooling the results from these fragment groups yielded a most recent common ancestor (MRCA) shared between SARS-COV-2 and the bat isolate RaTG13 around 2007 (95% HPD: 2003, 2011). Further, the host of the MRCA was most likely a bat (probability 0.64 - 0.87). Hence, the spillover into humans must have occurred at some point between 2007 and 2019 and bats may have been the most likely host reservoir.

2 Introduction

Emerging and re-emerging infectious diseases caused by viruses (e.g. severe acute respiratory syndrome (SARS), Middle East respiratory syndrome (MERS) [1], Influenza [2,3]), bacteria (e.g. Lyme disease [4], Cholera [5], Plague [6], *Escherichia coli* O157:H7 (E. coli) [7–9]), fungi (e.g. *Cryptococcus gatti* infection [10]) and parasites (e.g. malaria [11]) are still the leading causes of death globally [12]. They also raise concerns about global health, biosecurity and economic disruption [13,14]. Zoonoses comprise 60% of total infectious diseases whose agent is transmitted from animal hosts to humans [15].

Coronaviruses belong to the sub-family Coronavirinae and can be zoonotic [16–18]. They have caused two major human epidemics in the two decades preceding the SARS-CoV-2 pandemic: severe acute respiratory syndrome (SARS) in 2002-2003 and Middle East respiratory syndrome (MERS) in 2012. Research indicates SARS-CoV are closely related to viruses isolated from Chinese horseshoe bats (*Rhinolophus sp.)* which are considered to be the likely reservoir host [19–21]. *Vespertilio superans* bats are also thought to be the primary reservoir for MERS [22]. Intermediate hosts have been postulated for both of these epidemics. For example, coronavirus isolates from Himalayan palm civet cats were highly homologous to human SARS-CoV [17,23,24] as were dromedary camels for MERS-CoV [25–27]. However, whether civet-cats/dromedary camels were actually intermediate hosts remains an open question.

A human outbreak of Novel Coronavirus (SARS-CoV-2) with cases presenting symptoms of pneumonia was reported on December 8th 2019 in Wuhan, China. These cases were epidemiologically associated with a fresh seafood and wild animal market in Wuhan [28–30]. By January 7th 2020, the agent of this pneumonia was isolated from the respiratory epithelium of patients [31,32] and on 11th February WHO named this new coronavirus as severe acute respiratory syndrome coronavirus 2 (SARS- CoV-2) [33,34]. Subsequently, there was worldwide human transmission resulting in a global pandemic [35] and as of 4th October 2023, there has been 771,151,224 infections and 6,960,783 deaths reported [36]. The International Monetary Fund estimates cumulative output loss from the pandemic through to 2024 at $13.8 trillion [37].

Coronaviruses comprise four genera: alpha, beta, gamma and delta. Alpha and Betacoronaviruses are frequently found in bats and are mainly associated with infections in mammals [38]. The Betacoronavirus genus includes SARS-CoV-2, SARS CoV and MERS [34] which can cause respiratory, gastrointestinal, hepatic and central nervous system infections in humans [39]. Host species for the Gammacoronavirus genus include birds and Beluga whales. Deltacoronaviruses have been found in birds and mammals [40,41].

Studies utilizing whole genome sequencing have enabled an understanding of the phylogenies, transmission, genetic diversity and outbreak dynamics of zoonotic viruses including: SARS CoV [14,39]; MERS CoV [18,42]; Ebola [43]; HIV [44] and influenza [45,46]. These studies utilised Bayesian phylodynamic methods [47] to determine molecular clock rate(s) and likely host reservoirs as well as elucidating the temporal and geographical patterns of transmission.

Identification of the temporal signal between genome sequences is crucial to obtain timed phylogenies. The analysis of SARS-CoV-2 has proven challenging in this respect and no evidence of a temporal signal was obtained in previous studies [48]. This might have affected the accuracy of former estimates for the time to the most recent common ancestor. The lack of a temporal signal may be due to recombination and/or different molecular clock rates in different parts of the genome. To circumvent this problem, new methods need to be developed to identify genome fragments that exhibit a similar molecular clock rate across coronavirus genomes.

Bats have been postulated as a likely reservoir from which SARS-CoV-2 originated. Evidence pointing in this direction includes the close genetic relationship between a coronavirus isolated from *Rhinolophu*s bats in Yunnan, Southern China [49] and SARS-CoV [50]. At the whole genome sequence level, the closest available sequence to SARS-CoV-2 to date is the RaTG13 virus sampled from a *Rhinolophus affinis* bat (96.2% similarity [32,51]). Also, serological surveillance of people living in villages close to the natural habitat of bats in caves revealed 2.9% bat coronavirus seroprevalence in humans. This indicates that human exposure to bat-CoVs may be relatively common in China and that there is an opportunity for these viruses to spill over directly into humans without the need for an intermediate host [50]. However, there is the possibility of an intermediate host reservoir between bats and humans[51]. Several have been hypothesised for SARS-CoV-2. These include rodents [52], racoon dogs [53], pangolin [54,55] and other animals [56]. Despite significant efforts to identify an intermediate host, none has been clearly identified so far.

The zoonotic reservoir of SARS-CoV-2, the length of time the SARS-CoV-2 lineage has circulated in the host reservoir and the time when the first transmission occurred into humans remain unclear. This paper aims to: (1) identify the host reservoir of SARS-CoV-2, (2) to determine when SARS-CoV-2 spilled over from the host reservoir to humans and (3) to determine the phylogenetic relationship between SARS-CoV-2 and other coronaviruses. This will be achieved by conducting phylogenetic and Bayesian phylodynamic analysis utilizing a database of 71 coronavirus genomes and identification of core genome fragments with indistinguishable molecular clock rates.

3 Methods

Data

Coronavirus whole genome sequences (n = 71) and a Breda virus genome sequence were obtained from Genbank, the National Centre for Biotechnology Information (NCBI) and the Global Initiative on Sharing All Influenza Data (GISAID) for phylogenetic analysis (see Table S1). Metadata including year of isolation, host and country were also collected. Three sub-samples (selections) of these sequences were used to address the aims of this paper (Figure 1).

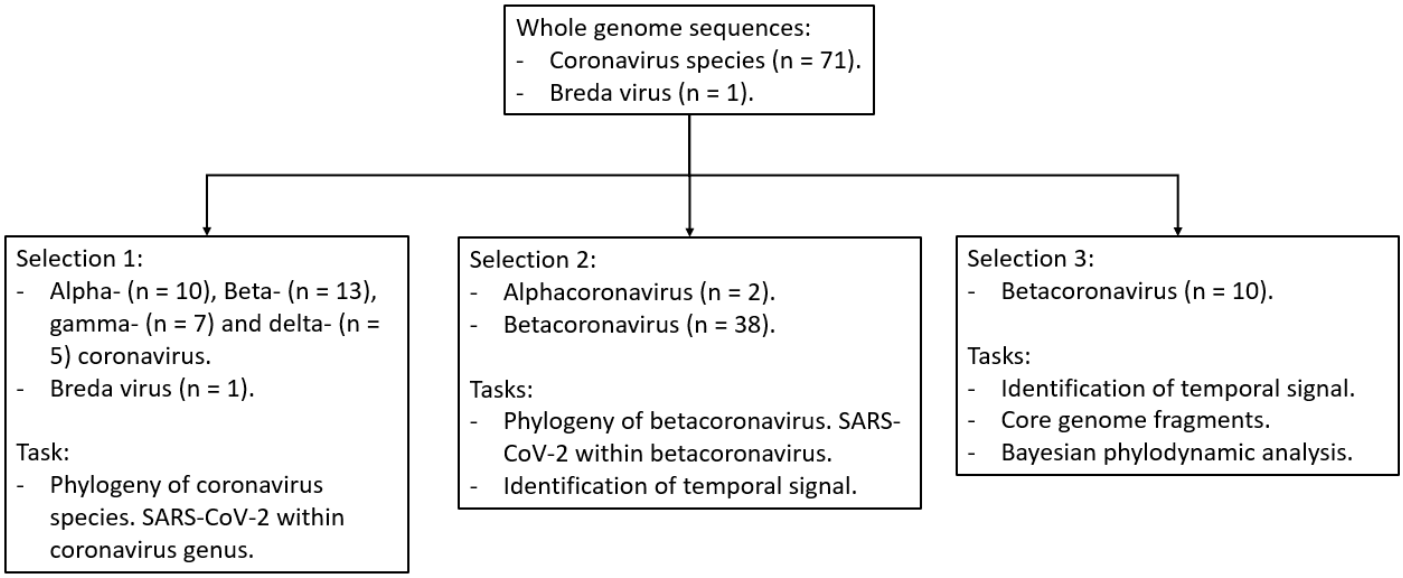


Figure 1: Genomes used in this study. Three selections of the 72 genomes in the original dataset were used for phylogenetic and phylodynamic analyses. Details on the genomes within each selection are given in Table S1.

3.1 Phylogeny of coronavirus

A selection of 35 whole genome sequences representative of the Alpha, Beta, Gamma and Delta coronavirus genera and a Breda virus genome were used to determine the location of SARS-CoV-2 Wuhan viruses within the coronavirus genus (Table S1, Selection 1).

A subset of 40 sequences (Table S1, Selection 2) was used to determine the potential origin of the outbreak of SARS-CoV-2. These sequences comprise Betacoronaviruses (n=38, SARS-CoV-2, SARS-CoV, MERS- CoV, and viruses from a wide range of hosts) along with Alphacoronaviruses (n=2) as the outgroup.

Genome selections 1 and 2 were aligned using ClustalW [57,58].The alignment was used to generate phylogenies with MEGA X[59] using the neighbour-joining method with bootstrapping (1000 replications).

3.2 Identification of temporal signal from coronavirus genomic data

3.2.1 Whole Genome Sequences

40 WGS (Table S1, Selection 2) and 10 WGS (Table S1, Selection 3) were used to assess the temporal signal utilising root-to-tip regression by TempEST[60].

3.2.2 Core genome fragments

3.2.2.1 Identification of core genome fragments

Identification of core genome fragments was performed on a subset of the coronavirus genome sequences (n=10) (Table S1, Selection 3). Multiple sequence alignments for the 10 genomes were again obtained using ClustalW. Contiguous fragments with more than 200 consecutive nucleotide base positions (bp) across the 10 WGS were identified (Figure 2(a)). The identified core genome fragments were mapped to the reference sequence of SARS-CoV-2 (Gen Bank -NC-045512.2) to determine the non-structural proteins present.

3.2.2.2 Identification of core genome fragment groups with indistinguishable molecular clock rates

3.2.2.2.1 Pairs of fragments

The pairwise genetic distance between sequences i and j in a fragment f was measured by the number $D_f^{(i,j)}$ of nucleotide differences between them. Figure 2(b) shows the pairwise genetic distance, $D_A^{(1,2)}$ and $D_B^{(1,2)}$ between sequence 1 and sequence 2, for fragments A and B respectively.

The pairwise genetic distance $D_f^{(i,j)}$ was normalized to account for the fact that every fragment identified across 10 WGS had a different number of nucleotides. More explicitly, the pairwise distance between two sequences i and j corresponding to a fragment f was estimated by the proportion of sites that differ between the two sequences in this fragment, i.e.

$$d_f^{(i,j)} = \frac{D_f^{(i,j)}}{L_f}.$$

Here, $L_f \geq 200$ is the length of fragment f.

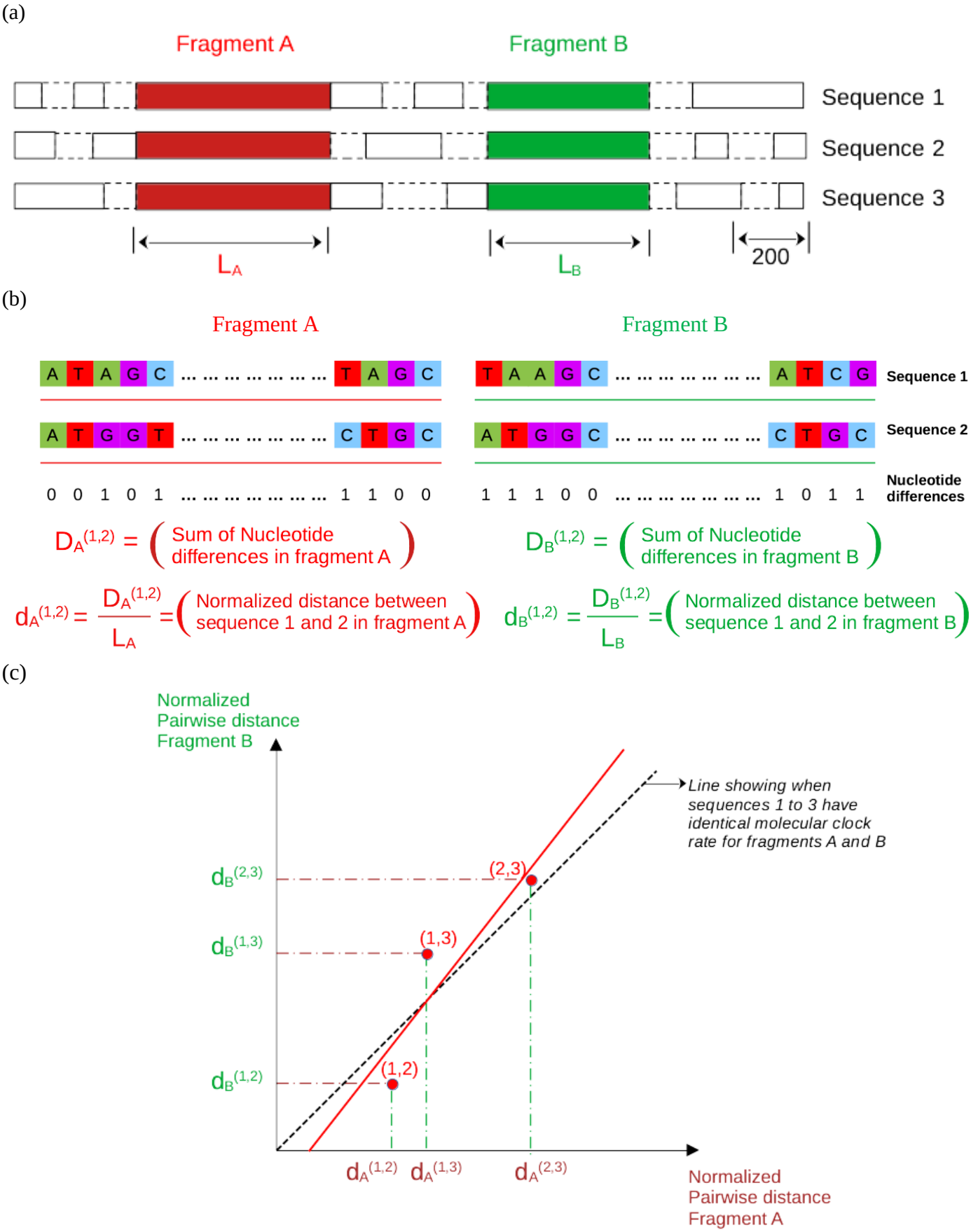


Figure 2: Method to identify genome fragment groups with indistinguishable molecular clock rates exemplified for three whole genome sequences. (a) Identification of contiguous DNA core fragments with more than 200 consecutive nucleotide base positions (rectangles with solid border indicate nucleotide fragments; dashed lines indicate genome gaps). This example identifies two fragments (*A* and *B*) from the three genomes. (b) Calculation of normalised pairwise genetic distance $d_A^{(1,2)}$ and $d_B^{(1,2)}$ between sequence 1 and sequence 2 in terms of fragment *A* and fragment *B* respectively. (c) Virtual example of linear regression analysis using normalized pairwise distances between fragments *A* and *B* across the three sequences. The dashed line indicates the ideal case in which fragments *A* and *B* have identical molecular clock rates.

3.2.2.2.2 Identifying groups of fragments with indistinguishable clock rates

The normalized pairwise distance between pairs of fragments was used to identify groups of fragments with statistically indistinguishable molecular clock rates. In a hypothetical scenario in which two fragments A and B follow the same molecular clock for a pair of sequences i and j, it would be expected that $d_A^{(i,j)} = d_B^{(i,j)}$. If all the sequences evolve with the same molecular clock according to fragment A and B, the distances $(d_A^{(i,j)}, d_B^{(i,j)})$ between any pair of sequences (i, j) should fall along the line with zero intercept and slope 1 in the space (d_A, d_B). In general, the points $(d_A^{(i,j)}, d_B^{(i,j)})$ for all pairs of sequences will not exactly fall along the ideal straight line (see an example in Figure 2 (c)). In practice, a line was fitted to the points $(d_A^{(i,j)}, d_B^{(i,j)})$ for all pairs of sequences and the fragments A and B were assumed to follow the same molecular clock if the fitted line is close enough to the ideal line. More specifically, the fitted line was considered to agree with the ideal line if (i) the slope was statistically consistent with 1, (ii) the intercept was statistically compatible with 0 and (iii) the p-value for the correlation coefficient was smaller than 0.05. Only pairs of fragments satisfying these conditions were assumed to follow indistinguishable molecular clock rates. From a statistical viewpoint, these pairs of fragments are considered evolutionary synchronous. The statistical significance of the hypotheses slope = 1 and intercept = 0 were tested in terms of a 95% confidence interval. The fragment groups identified by this method were used in the BEAST analysis described below.

3.3 Bayesian phylodynamic analysis by BEAST

3.3.1 Bayesian phylodynamic analysis to estimate clock rate and timed phylogeny

Datasets comprising 10 WGS, all core genome fragments and the fragment groups with indistinguishable molecular clocks were used to conduct Bayesian phylodynamic analyses with BEAST (v.2.6.2)[47,61]. This involved three sub-models. The First is the DNA substitution sub-model which has 2 options: (1) Hasegawa-Kishino-Yano (HKY) and (2) General Time Reversible (GTR) model. The second is the population sub-model which has 4 options: (1) Coalescent Constant Population (CCP); (2) Coalescent Exponential Population (CEP); (3) Coalescent Bayesian Skyline (CBS) and (4) Yule Model. The third is the molecular clock sub-model with 2 options: (1) strict (SC) and (2) relaxed (RC) molecular clock. Both SC and RC had uninformed prior distributions. The clock rates obtained from this analysis were used to specify the clock rate priors in the subsequent host probability calculations [62–64] of section 3.3.3.

Models were run using the Markov Chain Monte Carlo (MCMC) method with 10^8 iterations after 10% burn-in and sampled once every 1000 iterations. All permutations of substitution, population and clock models (n=16) for each dataset (n=4) were performed. Hence, in total 64 runs were carried out in the BEAST analysis. Convergence was assessed using the effective sample size (ESS) criterion (ESS>100) implemented on Tracer v.1.71[65]. Tree Annotator was used to produce a Maximum Clade Credibility (MCC) tree and from this tMRCA of SARS-CoV-2 Wuhan, with 95% HPD was obtained. The timed phylogeny was visualized using Fig Tree v 1.4.4[66,67].

3.3.2 Model selection using Nested sampling

Nested sampling (NS) was used to identify the best fitting parsimonious models[68,69]. This method uses two tuning parameters: The number of points sampled from the prior (number of particles, set to n=32) and chain length set to 20,000 steps with sub-chain length of 10,000 steps. Only model runs that converged in the previous section were used in this analysis.

3.3.3 Determination of host reservoir probability, molecular clock rate and time to most recent common ancestor

To estimate the most probable host reservoir of SARS-CoV-2 and the time period of spill over from animal reservoirs to humans, host transition and evolutionary dynamics was performed using BEAST (v1.10.4) which implements a method to estimate cross-species transmission [70]. Models selected by NS together with priors using the clock rate obtained as described above were used in the analysis. Posterior distributions were obtained through MCMC analysis runs with chain lengths of 1×10^8 generations and convergence was assumed to be achieved when ESS >100. If ESS >100 was not achieved, the MCMC chain length was increased to 2×10^8 generations and the ESS was checked.

3.3.3.1 Determination of pooled estimate of tMRCA and associated uncertainty

All the estimates for the time to most recent common ancestor (tMRCA) and their associated uncertainties, obtained by the selected models with NS, were pooled using propagation of errors [71].

4 Results and discussion

4.1 Phylogenetic analysis

4.1.1 Classification of SARS-CoV-2 within Coronavirus

The phylogenetic analysis of coronavirus showed that the four different genera of coronavirus (alpha, beta, gamma and delta) belong to four distinct clades (Figure 3(a)). SARS-CoV-2 belongs to the Betacoronavirus clade as reported previously [34].

4.1.2 Phylogenetic relationship between SARS-CoV-2 Wuhan and other beta coronaviruses

Phylogenetic analysis (Figure 3 (b)) showed that all SARS-CoV-2 Wuhan genomes grouped together and that Bat CoV RaTG13 (obtained from bat *Rhinolophus affinis*) [23] was the most closely genetically related to SARS-CoV-2 . The next most closely related to SARS-CoV-2 was a virus sampled from Pangolin (*Manis javanica* – MP789)[72]. This agrees with previous research which found that RaTG13 [48,54] was the most closely related virus to SARS-CoV-2 with 96.2% whole genome homology [32,51]. These authors also found Pangolin-CoV as the next most closely related with 91.02% WGS similarity[54]. SARS-CoV are the next closest group to the SARS-CoV-2 clade with 79.5% WGS similarity, see Figure 3(b) [32,73].

MERS-CoV is phylogenetically more distant from the SARS-CoV-2 Wuhan clade than SARS-CoV and only shares 50% sequence identity, see Figure 3(b)[73,74]. The Alphacoronavirus samples are clustered together at the root of the tree as an outgroup, as expected.

(a) (b)

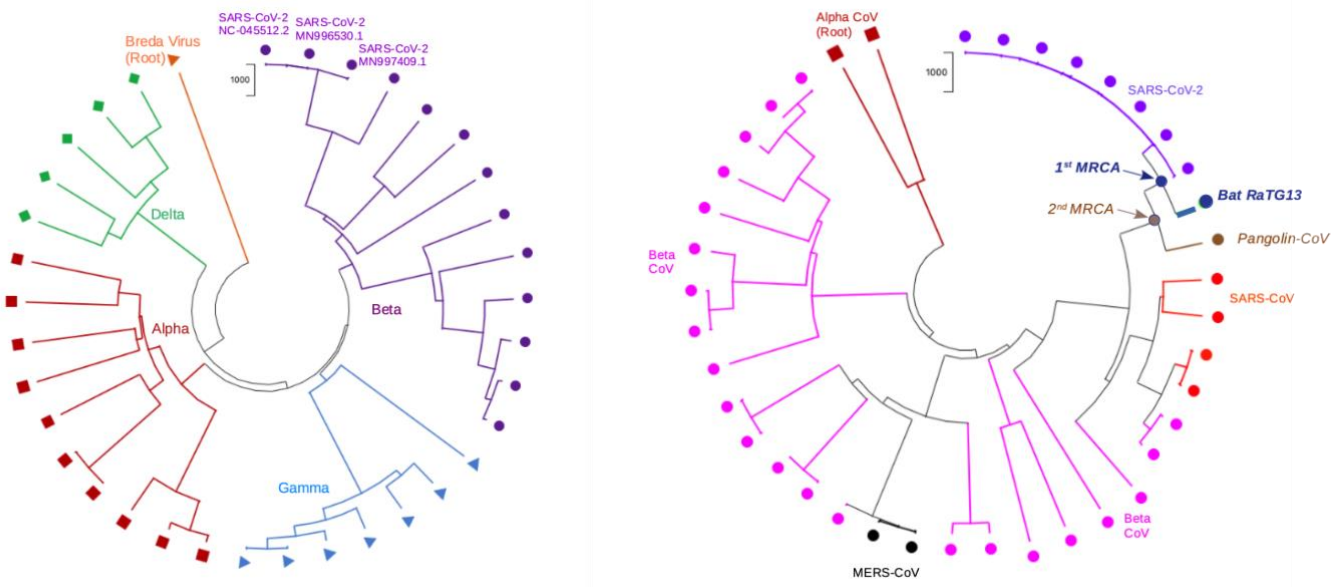


Figure 3: Phylogenetic analyses. (a) Neighbour-joining tree of 36 WGS representative of coronavirus (Alpha – maroon; Beta – purple; Gamma – blue and Delta – green). The tree is rooted with Breda virus (peach). Three genome sequences of SARS-CoV-2 (purple) cluster within the Betacoronavirus clade (see Table S1 - Data Selection 1) (b) Neighbour-joining tree of Betacoronavirus (n=38) with Alphacoronavirus samples (n=2, maroon) as root (see Table S1 - Data Selection 2); SARS-CoV-2 (purple); SARS-CoV (red), MERS-CoV (black), other Betacoronaviruses (Beta CoV, pink). BatCoV RaTG13 (indigo) is the 1st MRCA and Pangolin coronavirus (brown) is the 2nd MRCA to SARS-CoV-2 (purple). The scale bars provide 1000 bootstraps in the MEGA X.

4.2 Identification of temporal signal from coronavirus genomic data

4.2.1 Whole Genome Sequences

Temporal analysis using TempEst utilising 40 WGS and 10 WGS (see Table S1 - Selections 2 and 3, respectively) failed to find a statistically significant correlation coefficient in terms of the root-to-tip divergence (Supplementary Figure 1, Table S2). This suggests a lack of temporal signal for both datasets, in agreement with other studies [48]. This may be due to the fact that these viruses are from a diverse coronavirus population with deep evolutionary histories.

4.2.2 Core genome fragments

4.2.2.1 Identification of core genome fragments

Twenty fragments with $L_f \geq 200$ were identified from the 10 WGS alignment of 32794 base pairs (bp) (Figure 4). The genomic location of the identified 20 fragments was found to be within a 266 to 21555 bp region which corresponds to two open reading frames 1ab (ORF's)[75–77]. The ORF1ab gene encodes non-structural proteins (nsps): nsp 1 to 11 and nsp 12 to 16, respectively (Table S3).

(a)

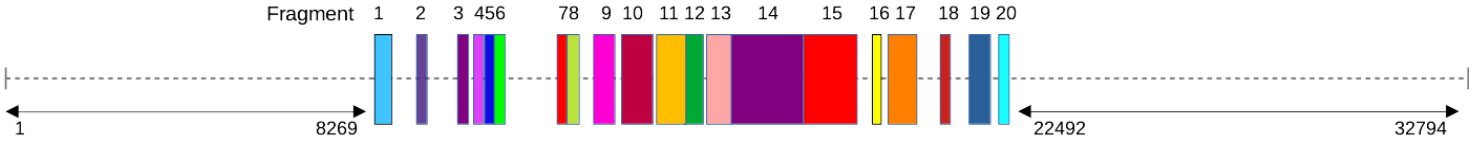


Figure 4: Scale diagram of the core genome fragments (n=20) with more than 200bp from the 10 WGS multiple sequence alignment (Tables S1, Selection 3 and S3).

4.2.2.2 Identification of core genome fragment groups with indistinguishable molecular clock rates

The 20 core genome fragments obtained above were analysed to form groups of fragments with indistinguishable molecular clock rates using the criteria in Section 3.2.2.2.2. Two groups of three fragments were identified with a mutually consistent temporal signal: {1,5,16} and {8,11,12} (Table 1). Scatterplots for the normalised pairwise distances for these fragments are presented in Supplementary Figures 2 and 3.

Table 1: Regression analysis parameters for core genome fragment groups which satisfy the conditions of indistinguishable molecular clock rate.

Fragment Group	**Fragment number**	**Fragment number**	**p-value**	**Slope**	**Slope (95% CI)**	**Intercept**	**Intercept (95% CI)**
1	Frag 1	Frag 5	1.64E-16	0.89	(0.75, 1.02)	1.25	(-1.18, 3.69)
	Frag 1	Frag 16	3.74E-24	0.92	(0.83, 1.01)	0.35	(-1.23, 1.93)
	Frag 5	Frag 16	3.46E-19	0.90	(0.78, 1.01)	1.25	(-0.77, 3.26)
2	Frag 8	Frag 11	9.55E-33	0.98	(0.92, 1.03)	0.04	(-0.76, 0.84)
	Frag 8	Frag 12	4.13E-25	0.94	(0.86, 1.03)	-0.43	(-1.62, 0.77)
	Frag 11	Frag 12	1.52E-31	0.97	(0.91, 1.03)	-0.50	(-1.34, 0.33)

4.3 Bayesian Phylodynamic analysis by BEAST

4.3.1 Model selection using Nested Sampling

Nested sampling found that the HKY substitution model provided the best fits for all the datasets. In contrast, the best combination of population and molecular clock models was found to be data specific. This resulted in six evolutionary models being selected (Table 2). For the 10 WGS dataset, the best fitting model was one with coalescent exponential population and relaxed molecular clock (CEP-RC). Nested sampling was unable to discriminate between the fits of two combinations of models when using the 20 core fragments dataset: Coalescent constant population with relaxed clock (CCP-RC) and coalescent exponential population with relaxed clock (CEP-RC). Both model combinations are taken as the best fits for this dataset. The best fit to the evolution of the fragment group {8,11,12} was also obtained for two combinations of models which are statistically indistinguishable: the coalescent population model combined with a strict clock model (CCP-SC) and a coalescent constant population model combined with a relaxed clock model (CCP-RC). The combination CCP-SC also provided the best fit to the evolution of the fragment group {1,5,16}. By estimating maximum Effective Sample Size (Max ESS) using Nested sampling, increasing the number of particles to 32 in the parameter space at each iteration, yields more precision to reach better convergence. This provides higher values of Max ESS estimate for six evolutionary models (Supplementary Figure 4).

Table 2. Results of the Bayesian phylodynamic and host probability analyses for the selected 6 best fitting models identified by nested sampling. The final row provides the pooled values from the 6 models.

Data and evolutionary model	Molecular clock rate [95% CI]	Year of 1st MRCA [95% CI]	1st MRCA host probability (human, pangolin, bat)	Year of 2nd MRCA [95% CI]	2nd MRCA host probability (human, pangolin, bat)	ESS
10WGS, CEP-RC	0.0029 [0.0018, 0.0049]	2006 [2000, 2011]	(0.16, 0.02, 0.82)	1998 [1987, 2007]	(0.14, 0.08, 0.79)	445
20 FG, CCP-RC	0.0039 [0.0022, 0.0079]	2009 [2002, 2013]	(0.13, 0.01, 0.86)	2003 [1985, 2011]	(0.12, 0.07, 0.81)	124
20 FG, CEP-RC	0.0059 [0.0041, 0.0095]	2010 [2005, 2013]	(0.11, 0.02, 0.87)	2005 [1996, 2011]	(0.10, 0.08, 0.82)	818
{8,11,12} FG, CCP-RC	0.0058 [0.0032, 0.0116]	2010 [2004, 2013]	(0.14, 0.02, 0.84)	2006 [1995, 2012]	(0.13, 0.06, 0.81)	1651
{1,5,16} FG, CCP-SC	0.0023 [0.0020, 0.0030]	2005 [2000, 2009]	(0.28, 0.01, 0.70)	1984 [1973, 1996]	(0.26, 0.16, 0.58)	62571
{8,11,12} FG, CCP-SC	0.0011 [0.0010, 0.0014]	1998 [1992, 2004]	(0.34, 0.02, 0.64)	1972 [1959, 1985]	(0.29, 0.18, 0.52)	62671
Pooled values	0.0037 [0.0017, 0.0057]	2007 [2003, 2011]	(0.19 [0.16, 0.23], 0.02 [0.02, 0.02], 0.79 [0.75, 0.83])	1995 [1986, 2004]	(0.17 [0.14, 0.21], 0.11 [0.08, 0.13], 0.72 [0.67, 0.78])	21380

Footnote: The molecular clock rate distribution in Table S5 was used to specify the rate of uniform prior, mean with 95% CI, in obtaining the results of Table 2.

4.3.2 Clock rate, time to the most recent common ancestor and host reservoir probability

Clock rate

Clock rates for the six selected evolutionary models range from 1.1 to 5.9× 10^{-3} subs/site/year (Table 2). The clock rate of the two models using a strict clock ({1,5,16}-CCP-SC and {8,11,12}-CCP-SC) tends to be slower on average than that of models with a relaxed clock. However, only the model based on the fragment group {8,11,12} is statistically slower than the rest (Supplementary Figure 5(a)). Pooling the values of the molecular clock rate from the six models yielded 0.0037 (0.0017 - 0.0057) subs/site/year. This estimate is compatible with the molecular clock rate of SARS, 0.00169 (0.00131 - 0.00205) subs/site/year found previously by Boni et al[48]. In the same work, however, the authors used a slower estimate of the molecular clock rate, 0.00055 subs/site/year, for their phylogenetic dating methods. This estimate was obtained for non-recombinant regions of a 68-genome sarbecovirus alignment using the molecular clock rate of MERS-CoV (0.00078 subs/site/year) and HCoV-OC43 (0.00024 subs/site/year) to define a prior. Such estimates were used to circumvent the lack of temporal signal in their dataset. There are several reasons that could explain the difference between the current studies pooled estimate of the molecular clock rate and the estimate obtained by Boni et al.,[48]. These include the selection of genomes and the parts of the genome used. Another possibility is that the estimated molecular clock rate is influenced by the relatively slow clock rate of the prior based on MERS-CoV and HCoV-OC43. In the current study, by estimating the clock rate distribution for each of the six datasets (Table S5) yields a faster pooled estimate of molecular clock rate.

Time to the most recent common ancestor

All of the 6 models selected by nested sampling gave statistically compatible estimates for the time to the 1st MRCA of SARS-CoV-2 and BatCoV RaTG13 (Table 2 and Supplementary Figure 5(b)). Pooling the results from the six models (Table S4(a)), the time to the 1st MRCA was found to be 13.5 ± 4.1 years which corresponds to the year 2007 (2003 - 2011) (Supplementary Figure 6). This estimate for the time to the 1st MRCA is more recent than the estimates given in Boni et al.,[48] : 1948 (1879 – 1999), 1969 (1930 – 2000) and 1982 (1948 – 2009). This is expected due to the slower molecular clock rate used, as discussed above.

The time to the 2nd MRCA between Pangolin-CoV and the SARS-CoV-2 Wuhan/bat RaTG13 lineage was also statistically similar across the six models (Supplementary Figure 5(b)). Pooled results from the six models estimated it to be 25.4 ± 8.9 years, i.e., [1995 (1986 - 2004)], (Table 2, Table S4(b)). This estimate is also more recent than 1851(1730 - 1958) determined by Boni et al.,[48] .

Host reservoir probability

The most likely host reservoir of the 1st MRCA between SARS-CoV-2 and BatCoV RaTG13 was a bat with a probability of 0.79 [0.75, 0.83] (pooled over the six selected models, Table 2). This indicates bat as potentially the natural zoonotic origin of SARS-CoV-2. The pooled probability of origin for humans was 0.19 [0.16, 0.23] and pangolin 0.017 [0.015, 0.019] (Supplementary Figures 7 and 8). From the estimated time for the 1st MRCA and the likelihood of bat as a natural reservoir, it can be concluded that the ancestor of SARS-CoV-2 spilled over from bats to humans sometime between 2007 and 2019. It is unclear whether that was directly into humans or via an intermediate animal reservoir.

Evidence of bats as the natural reservoir

The results (Table 2) agree with the finding that phylogenetically close coronaviruses to SARS-CoV-2 have been circulating in bats for many decades[48,78]. Despite the likelihood of a spill-over of SARS-CoV-2 from the bat reservoir to humans, there is a lack of intermediate sequences that prevents a precise understanding of the spill over in terms of whether a potential intermediate host exists (WHO, 2021). The first recorded human cases with severe pneumonia were at Jin Yin-Tan hospital in Wuhan in December, 2019 and full genomes were sequenced at Wuhan Institute of Virology (WIV) (Zhou et al., 2020). However, the high prevalence of asymptomatic and unreported cases[79] and the low surveillance prior to the Wuhan outbreak make the identification of an intermediate host, even if one existed, very challenging.

5 Conclusion

Coronaviruses closely related to SARS-CoV-2 appear to have emerged from bats at some point within the 13 years before the SARS-CoV-2 outbreak. This is more recent than the estimates from previous studies which suggested 40 to 70 years [48]. Bat was found to be the most likely natural reservoir for SARS-CoV-2. There may be an intermediate host reservoir between bats and humans but there is insufficient information to demonstrate this hypothesis. To be more precise about the spillover of SARS-CoV-2 into humans, more sequencing and epidemiological evidence would be required. Unfortunately, this is difficult to achieve retrospectively and would have required an internationally coordinated surveillance program prior to the pandemic. To prevent future zoonotic pandemics, such a surveillance system will be required. This will require collaboration between epidemiologists, geneticists, medics, mathematical modelers and social scientists.

Acknowledgements

FJPR acknowledges funding from a Medical Research Council Fellowship (MR/W021455/1).

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Author contributions

VP contributed to the ideas behind this work, carried out the analysis and led the writing of the paper. FJPR and NJCS contributed to the ideas behind the work, advised on the analysis and edited drafts of the paper.

Competing interests

The authors declare no competing interests.

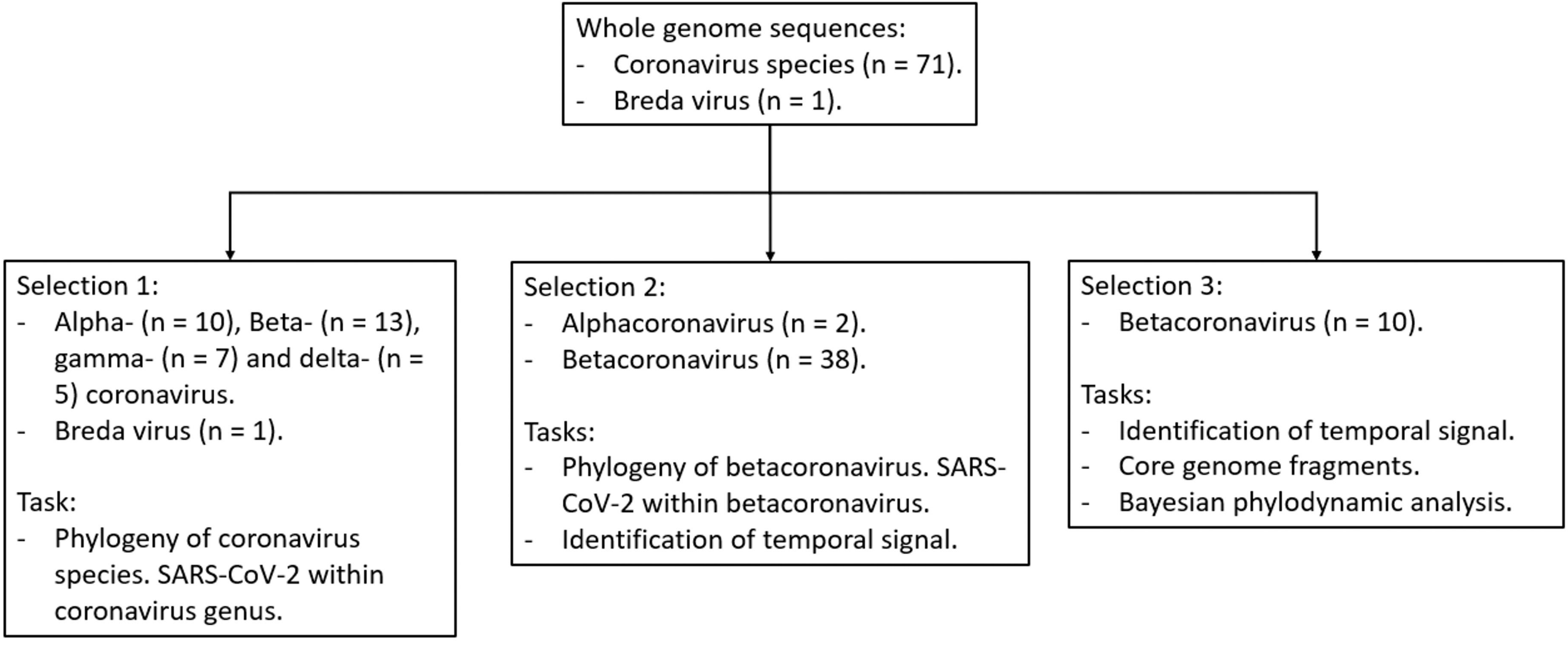
6 References

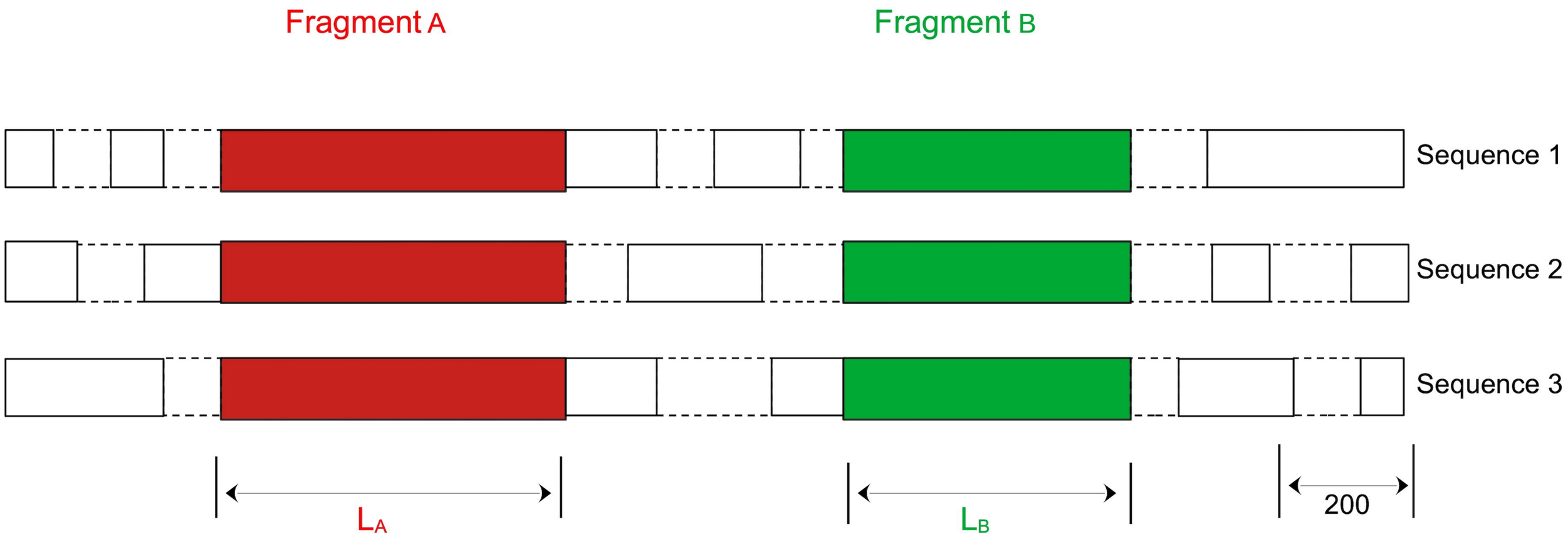
1. WHO EMRO | Sixty-first session 2. Current situation in the Region.
2. Kilbourne, E. D. Influenza pandemics of the 20th century. *Emerg Infect Dis* **12**, 9–14 (2006).
3. Yen, H. L. & Webster, R. G. Pandemic influenza as a current threat. *Curr Top Microbiol Immunol* **333**, 3–24 (2009).
4. Feder, H. M. *et al.* A Critical Appraisal of 'Chronic Lyme Disease'. (2007).
5. Kaper, J. B., Morris, J. G. & Levine, M. M. Cholera. *Clin Microbiol Rev* **8**, 48–86 (1995).
6. Zietz, B. P. & Dunkelberg, H. The history of the plague and the research on the causative agent Yersinia pestis. *Int J Hyg Environ Health* **207**, 165–178 (2004).
7. Franz, E. *et al.* Phylogeographic Analysis Reveals Multiple International transmission Events Have Driven the Global Emergence of Escherichia coli O157:H7. *Clinical Infectious Diseases* **69**, 428–437 (2019).
8. Strachan, N. J. C., Doyle, M. P., Kasuga, F., Rotariu, O. & Ogden, I. D. Dose response modelling of Escherichia coli O157 incorporating data from foodborne and environmental outbreaks. *Int J Food Microbiol* **103**, 35–47 (2005).
9. Strachan, N. J. C., Dunn, G. M., Locking, M. E., Reid, T. M. S. & Ogden, I. D. Escherichia coli O157: burger bug or environmental pathogen? *Int J Food Microbiol* **112**, 129–137 (2006).
10. C-A Chen, S., Meyer, W. & Sorrell, T. C. Cryptococcus gattii Infections. (2014) doi:10.1128/CMR.00126-13.
11. McFee, R. B. EMERGING INFECTIOUS DISEASES – OVERVIEW. *Disease-a-Month* **64**, 163 (2018).
12. Bloom, D. E., Kuhn, M. & Prettner, K. *Modern Infectious Diseases: Macroeconomic Impacts and Policy Responses*. www.iza.org (2020).
13. Orellana, C. Laboratory-acquired SARS raises worries on biosafety. *Lancet Infect Dis* **4**, 64 (2004).
14. Cheng, V. C. C., Lau, S. K. P., Woo, P. C. Y. & Kwok, Y. Y. Severe Acute Respiratory Syndrome Coronavirus as an Agent of Emerging and Reemerging Infection. *Clin Microbiol Rev* **20**, 660 (2007).
15. Taylor, L. H., Latham, S. M. & Woolhouse, M. E. J. Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences* **356**, 983 (2001).
16. Peiris, J. S. M., Guan, Y. & Yuen, K. Y. Severe acute respiratory syndrome. *Nature Medicine 2004 10:12* **10**, S88–S97 (2004).
17. Cui, J., Li, F. & Shi, Z. L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).
18. Memish, Z. A. *et al.* Middle East Respiratory Syndrome Coronavirus in Bats, Saudi Arabia. *Emerg Infect Dis* **19**, 1819 (2013).
19. Yin, Y. & Wunderink, R. G. MERS, SARS and other coronaviruses as causes of pneumonia. *Respirology* **23**, 130 (2018).
20. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science (1979)* **310**, 676–679 (2005).
21. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog.* **13**, e1006698 (2017).
22. Yang, L. *et al.* MERS–Related Betacoronavirus in Vespertilio superans Bats, China. *Emerg Infect Dis* **20**, 1260 (2014).
23. Ge, X. Y. *et al.* Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature* **503**, 535 (2013).
24. Guan, Y. *et al.* Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China. *Science (1979)* **302**, 276–278 (2003).

25. Meyer, B. *et al.* Antibodies against MERS Coronavirus in Dromedary Camels, United Arab Emirates, 2003 and 2013. *Emerg Infect Dis* **20**, 552 (2014).
26. Hemida, M. G. *et al.* MERS Coronavirus in Dromedary Camel Herd, Saudi Arabia. *Emerg Infect Dis* **20**, 1231 (2014).
27. Stalin Raj, V. *et al.* Isolation of MERS Coronavirus from a Dromedary Camel, Qatar, 2014. *Emerg Infect Dis* **20**, 1339 (2014).
28. WHO-convened global study of origins of SARS-CoV-2. https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part.
29. Li, X. *et al.* Transmission dynamics and evolutionary history of 2019-nCoV. *J Med Virol* **92**, 501 (2020).
30. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine* **382**, 1199–1207 (2020).
31. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265 (2020).
32. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature 2020 579:7798* **579**, 270–273 (2020).
33. Gorbalenya, A. E. *et al.* Severe acute respiratory syndrome-related coronavirus: The species and its viruses-a statement of the Coronavirus Study Group. doi:10.1101/2020.02.07.937862.
34. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology 2020 5:4* **5**, 536–544 (2020).
35. Cucinotta, D. & Vanelli, M. WHO Declares COVID-19 a Pandemic. *Acta Bio Medica : Atenei Parmensis* **91**, 157 (2020).
36. WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data. https://covid19.who.int/.
37. One Health Approach - Prevent the Next Pandemic | World Bank. https://www.worldbank.org/en/news/feature/2022/10/24/one-health-approach-can-prevent-the-next-pandemic.
38. Zhou, Z., Qiu, Y. & Ge, X. The taxonomy, host range and pathogenicity of coronaviruses and other viruses in the Nidovirales order. *Animal Diseases 2021 1:1* **1**, 1–28 (2021).
39. Wang, L. F. *et al.* Review of Bats and SARS. *Emerg Infect Dis* **12**, 1834 (2006).
40. Woo, P. C. Y. *et al.* Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. *J Virol* **86**, 3995–4008 (2012).
41. Woo, P. C. Y. *et al.* Molecular diversity of coronaviruses in bats. *Virology* **351**, 180–187 (2006).
42. Mohd, H. A., Al-Tawfiq, J. A. & Memish, Z. A. Middle East Respiratory Syndrome Coronavirus (MERS-CoV) origin and animal reservoir Susanna Lau. *Virol J* **13**, 1–7 (2016).
43. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369 (2014).
44. Faria, N. R. *et al.* The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56 (2014).
45. Smith, G. J. D. *et al.* From the Cover: Dating the emergence of pandemic influenza viruses. *Proc Natl Acad Sci U S A* **106**, 11709 (2009).
46. Dos Reis, M., Hay, A. J. & Goldstein, R. A. Using non-homogeneous models of nucleotide substitution to identify host shift events: Application to the origin of the 1918 'spanish' influenza pandemic virus. *J Mol Evol* **69**, 333–345 (2009).
47. Drummond, A. J. & Bouckaert, R. R. *Bayesian Evolutionary Analysis with BEAST*. (Cambridge University Press, 2015). doi:10.1017/CBO9781139095112.
48. Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology 2020 5:11* **5**, 1408–1417 (2020).

49. Ge, X. Y. *et al.* Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft. *Virol Sin* **31**, 31 (2016).
50. Wang, N. *et al.* Serological Evidence of Bat SARS-Related Coronavirus Infection in Humans, China. *Virol Sin* **33**, 104 (2018).
51. Zhang, Y. Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* **181**, 223 (2020).
52. Huang, Y. *et al.* Sars-cov-2: Origin, intermediate host and allergenicity features and hypotheses. *Healthcare (Switzerland)* **9**, (2021).
53. Mallapaty, S. COVID-origins study links raccoon dogs to Wuhan market: what scientists think. *Nature* **615**, 771–772 (2023).
54. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology* **30**, 1346 (2020).
55. Lam, T. T. Y. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature 2020 583:7815* **583**, 282–285 (2020).
56. Wardeh, M., Baylis, M. & Blagrove, M. S. C. Predicting mammalian hosts in which novel coronaviruses can be generated. *Nature Communications 2021 12:1* **12**, 1–12 (2021).
57. Higgins, D. G., Thompson, J. D. & Gibson, T. J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266**, 383–400 (1996).
58. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680 (1994).
59. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547 (2018).
60. Rambaut, A., Lam, T. T., Carvalho, L. M. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, (2016).
61. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019).
62. Baele, G. & Lemey, P. Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. *Bioinformatics* **29**, 1970–1979 (2013).
63. Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A. & Lemey, P. Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics. *Mol Biol Evol* **30**, 239 (2013).
64. Duchene, S. *et al.* Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations. *Mol Biol Evol* **37**, 3363–3379 (2020).
65. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901–904 (2018).
66. FigTree. http://tree.bio.ed.ac.uk/software/figtree/.
67. FigTree v. 1.4.4. (Rambaut 2018). Genetic variation parameters such as... | Download Scientific Diagram. https://www.researchgate.net/figure/FigTree-v-144-Rambaut-2018-Genetic-variation-parameters-such-as-observed-alleles_fig1_354454604.
68. Skilling, J. Nested sampling for general Bayesian computation. *https://doi.org/10.1214/06-BA127* **1**, 833–859 (2006).
69. Russel, P. M., Brewer, B. J., Klaere, S. & Bouckaert, R. R. Model Selection and Parameter Inference in Phylogenetics Using Nested Sampling. *Syst Biol* **68**, 219–233 (2019).
70. Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G. & Lemey, P. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, (2013).
71. Fantner, G. A brief introduction to error analysis and propagation. (2011).
72. Liu, P. *et al.* Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PLoS Pathog* **16**, (2020).
73. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565–574 (2020).

74. Gralinski, L. E. & Menachery, V. D. Return of the Coronavirus: 2019-nCoV. *Viruses* **12**, (2020).
75. Boopathi, S., Poma, A. B. & Kolandaivel, P. Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment. *J Biomol Struct Dyn* 1 (2020) doi:10.1080/07391102.2020.1758788.
76. Naqvi, A. A. T. *et al.* Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochim Biophys Acta Mol Basis Dis* **1866**, 165878 (2020).
77. Subissi, L. *et al.* SARS-CoV ORF1b-encoded nonstructural proteins 12–16: Replicative enzymes as antiviral targets. *Antiviral Res* **101**, 122 (2014).
78. Wang, H., Pipes, L. & Nielsen, R. Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. *Virus Evol* **7**, (2021).
79. Pérez-Reche, F. J., Forbes, K. J. & Strachan, N. J. C. Importance of untested infectious individuals for interventions to suppress COVID-19. *Scientific Reports 2021 11:1* **11**, 1–13 (2021).


Whole genome sequences:
- Coronavirus species (n = 71).
- Breda virus (n = 1).
Selection 1:
- Alpha- (n = 10), Beta- (n = 13), gamma- (n = 7) and delta- (n = 5) coronavirus.
- Breda virus (n = 1).
Task:
- Phylogeny of coronavirus species. SARS-CoV-2 within coronavirus genus.
Selection 2:
- Alphacoronavirus (n = 2).
- Betacoronavirus (n = 38).
Tasks:
- Phylogeny of betacoronavirus. SARS-CoV-2 within betacoronavirus.
- Identification of temporal signal.
Selection 3:
- Betacoronavirus (n = 10).
Tasks:
- Identification of temporal signal.
- Core genome fragments.
- Bayesian phylodynamic analysis.


Fragment A
Fragment B
Sequence 1
Sequence 2
Sequence 3
L_A
L_B
200

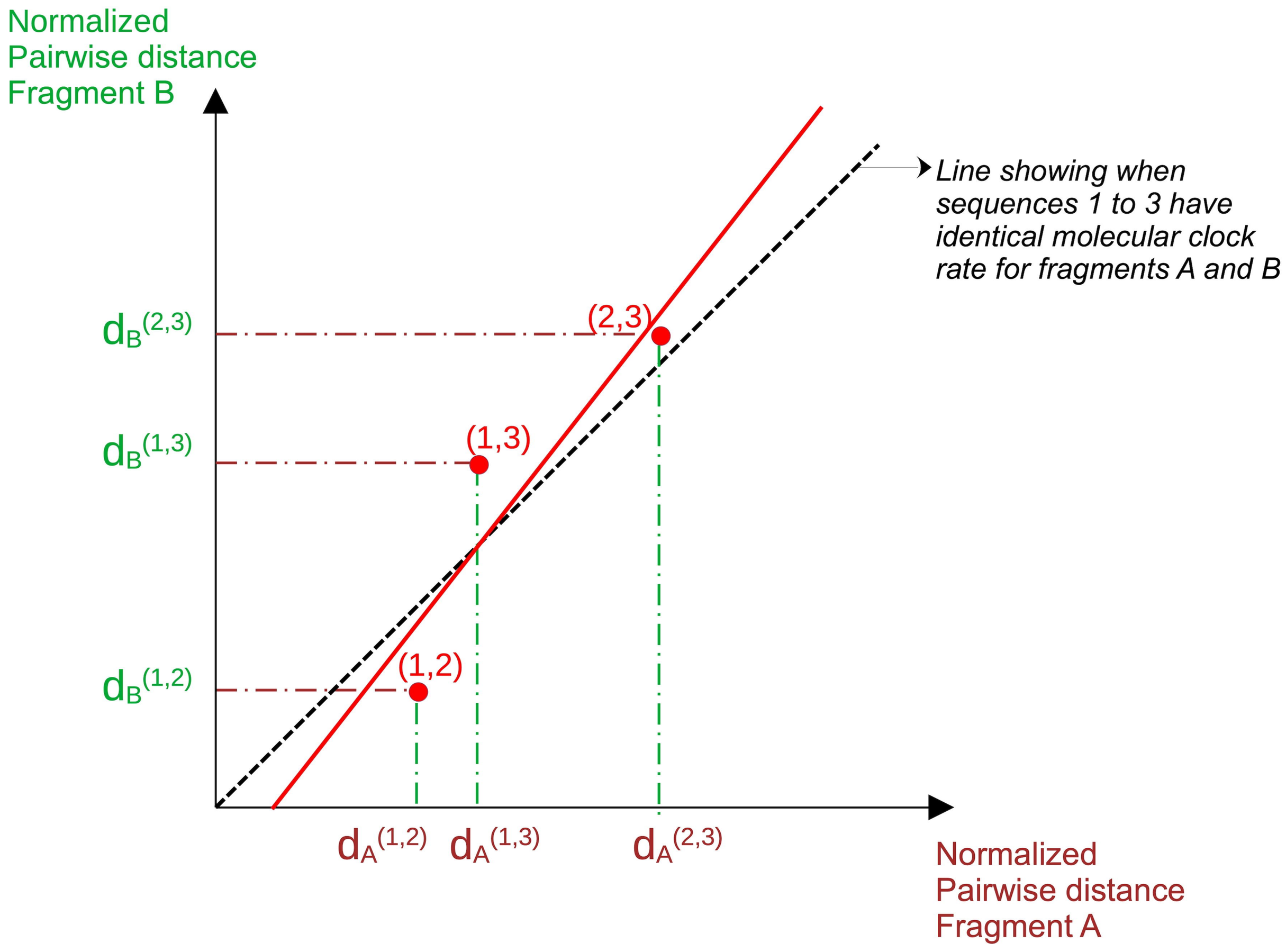
	Fragment A	Fragment B
Sequence 1	A T A G C … … … … … … … T A G C	T A A G C … … … … … … … A T C G
Sequence 2	A T G G T … … … … … … … C T G C	A T G G C … … … … … … … C T G C
Nucleotide differences	0 0 1 0 1 … … … … … … … 1 1 0 0	1 1 1 0 0 … … … … … … … 1 0 1 1

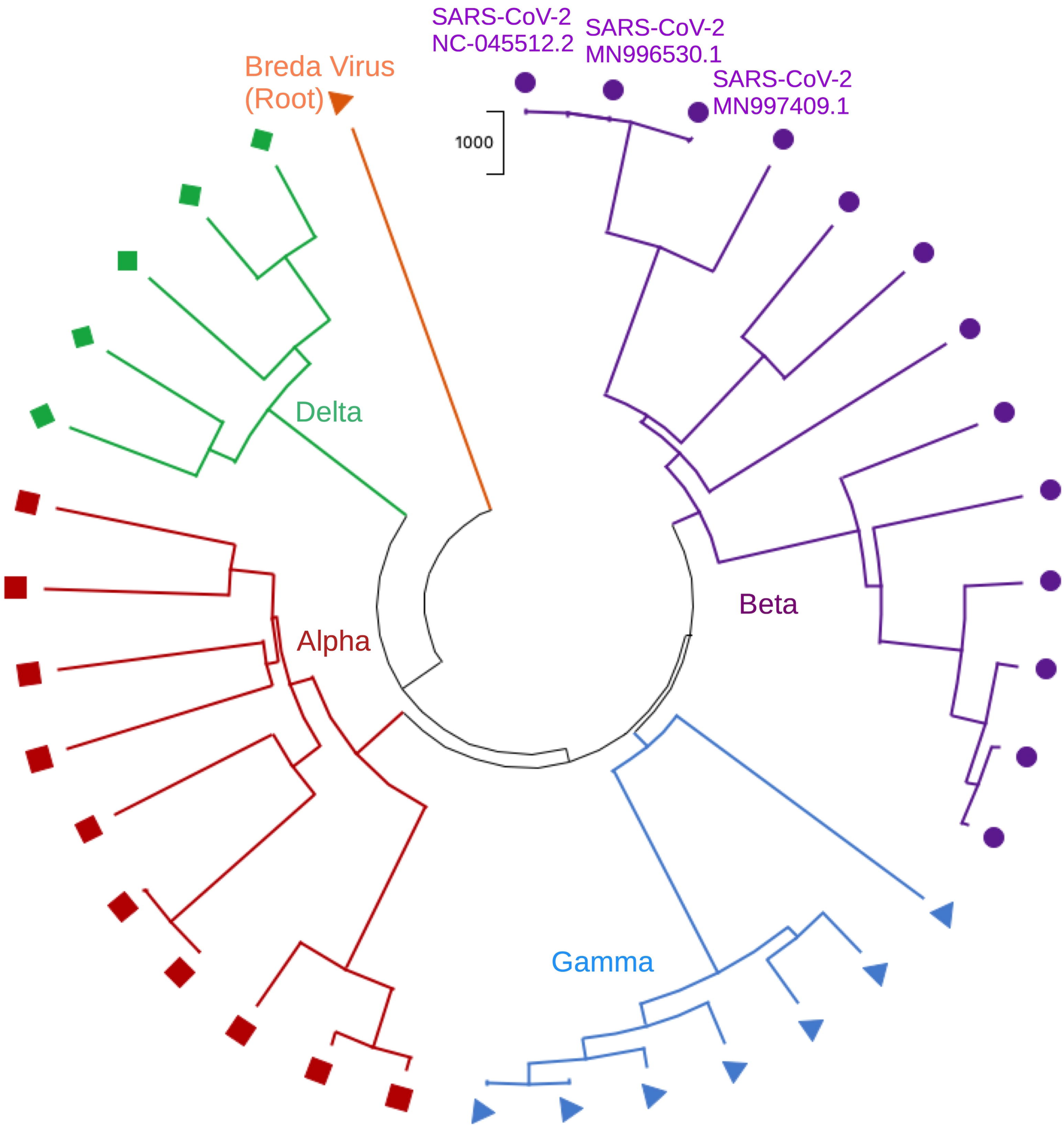
$$D_A^{(1,2)} = \left(\text{Sum of Nucleotide differences in fragment A} \right)$$

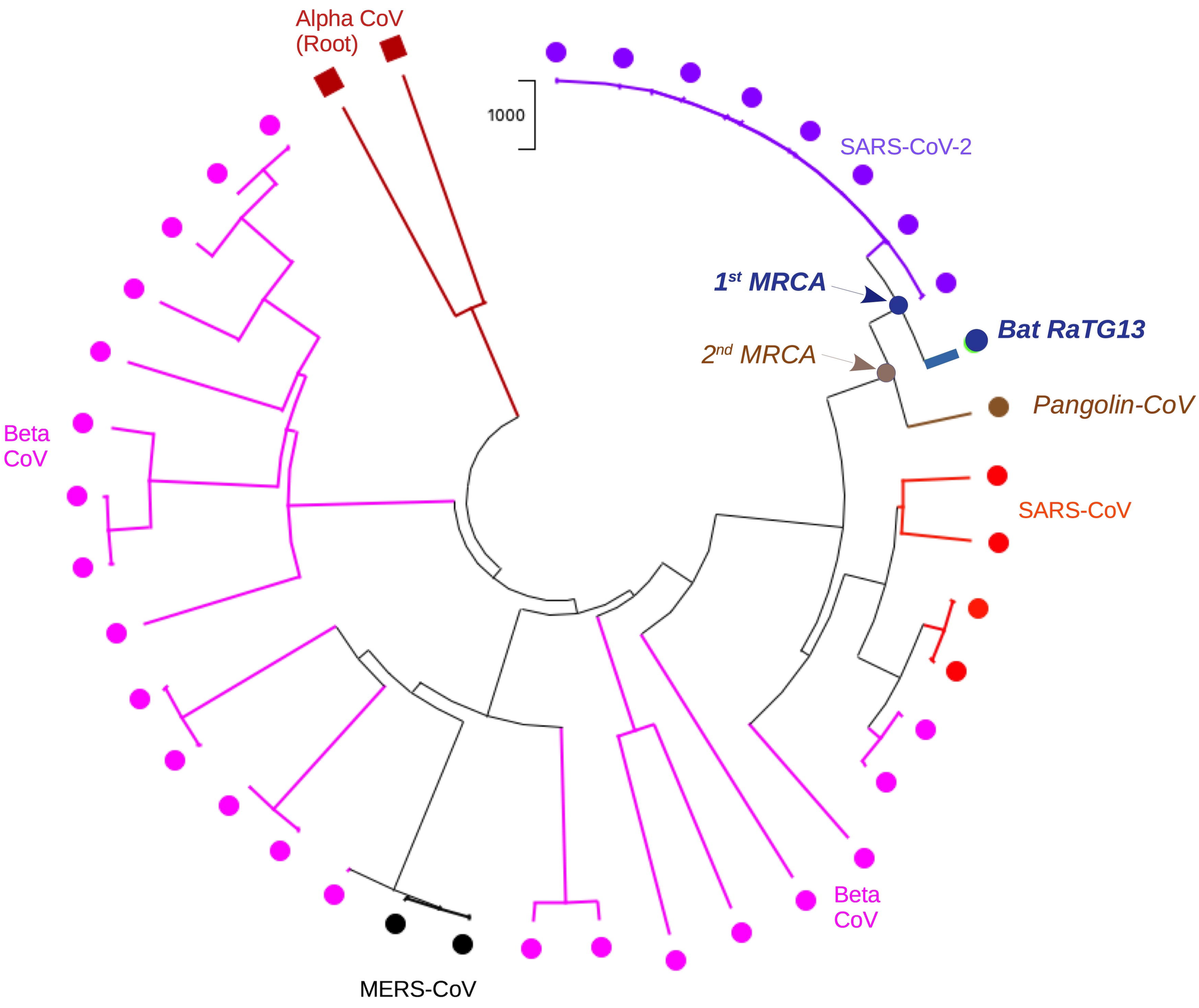
$$D_B^{(1,2)} = \left(\text{Sum of Nucleotide differences in fragment B} \right)$$

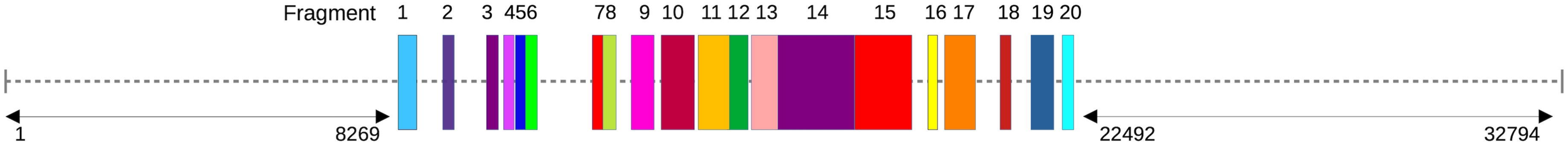
$$d_A^{(1,2)} = \frac{D_A^{(1,2)}}{L_A} = \left(\text{Normalized distance between sequence 1 and 2 in fragment A} \right)$$

$$d_B^{(1,2)} = \frac{D_B^{(1,2)}}{L_B} = \left(\text{Normalized distance between sequence 1 and 2 in fragment B} \right)$$


Normalized
Pairwise distance
Fragment B
$d_B^{(2,3)}$
$d_B^{(1,3)}$
$d_B^{(1,2)}$
(2,3)
(1,3)
(1,2)
Line showing when sequences 1 to 3 have identical molecular clock rate for fragments A and B
$d_A^{(1,2)}$
$d_A^{(1,3)}$
$d_A^{(2,3)}$
Normalized
Pairwise distance
Fragment A


SARS-CoV-2
NC-045512.2
SARS-CoV-2
MN996530.1
SARS-CoV-2
MN997409.1
Breda Virus
(Root)
1000
Delta
Beta
Alpha
Gamma


Alpha CoV
(Root)
1000
SARS-CoV-2
1st MRCA
Bat RaTG13
2nd MRCA
Pangolin-CoV
Beta
CoV
SARS-CoV
Beta
CoV
MERS-CoV


Fragment
1
2
3
456
78
9
10
11
12
13
14
15
16
17
18
19
20
1
8269
22492
32794