Adaptive trends of sequence compositional complexity over pandemic time in the SARS-CoV-2 coronavirus

José L. Oliver[1,2,$,*], Pedro Bernaola-Galván[3], Francisco Perfectti[1,4], Cristina Gómez-Martín[1,2,5], Silvia Castiglione[6], Pasquale Raia[6], Miguel Verdú[7,$,*] & Andrés Moya[8,9,10,$,*]

*Correspondence: José L. Oliver (oliver@ugr.es), Miguel Verdú (Miguel.Verdu@ext.uv.es) and Andrés Moya (Andres.Moya@uv.es)

[$]These authors contributed equally: José L. Oliver, Miguel Verdú and Andrés Moya.

[1]Department of Genetics, Faculty of Sciences, University of Granada, 18071, Granada, Spain
[2]Laboratory of Bioinformatics, Institute of Biotechnology, Center of Biomedical Research, 18100, Granada, Spain
[3]Department of Applied Physics II and Institute Carlos I for Theoretical and Computational Physics, University of Málaga, 29071, Málaga, Spain
[4]Research Unit Modeling Nature, Universidad de Granada, 18071 Granada, Spain
[5]Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Pathology, Cancer Center Amsterdam , Amsterdam, Netherlands
[6]Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse, Università di Napoli Federico II, 80126, Napoli, Italy
[7]Centro de Investigaciones sobre Desertificación, Consejo Superior de Investigaciones Científicas (CSIC), University of València and Generalitat Valenciana, 46113, Valencia, Spain
[8]Institute of Integrative Systems Biology (I2Sysbio), University of València and Consejo Superior de Investigaciones Científicas (CSIC), 46980, Valencia, Spain
[9]Foundation for the Promotion of Sanitary and Biomedical Research of Valencian Community (FISABIO), 46020, Valencia, Spain
[10]CIBER in Epidemiology and Public Health, 28029, Madrid, Spain

Abstract

During the spread of the COVID-19 pandemic, the SARS-CoV-2 coronavirus underwent mutation and recombination events that altered its genome compositional structure, thus providing an unprecedented opportunity to check an evolutionary process in real time. The mutation rate is known to be lower than expected for neutral evolution, suggesting purifying selection and convergent evolution. We begin by summarizing the compositional heterogeneity of each viral genome by computing its Sequence Compositional Complexity (SCC). To analyze the full range of SCC diversity, we select random samples of high-quality coronavirus genomes covering the full span of the pandemic. We then search for evolutionary trends that could inform us on the adaptive process of the virus to its human host by computing the phylogenetic ridge regression of SCC against time (i.e., the collection date of each viral isolate). In early samples, we find no statistical support for any trend in SCC values, although the viral genome appears to evolve faster than Brownian Motion (BM) expectation. However, in samples taken after the emergence of high fitness variants, and despite the brief time span elapsed, a driven decreasing trend for SCC and an increasing one for its absolute evolutionary rate are detected, pointing to a role for purifying selection in the evolution of SCC in the coronavirus. The higher fitness of variant genomes leads to adaptive trends of SCC over pandemic time in the coronavirus.

Keywords: Adaptive trends, sequence compositional complexity, coronavirus evolution.

Introduction

Given the difficulties to observe evolution directly over long periods, test tube experiments revealed as a particularly powerful tool for examining evolutionary dynamics. The Richard Lenski’s long-term evolution experiment (LTEE) with a laboratory population of *Escherichia coli* sampled through 60,000 generations shows the relationships between rates of genomic evolution and organismal adaptation (Barrick *et al.*, 2009; Good *et al.*, 2017). Experimental evolution of a major evolutionary innovation (the origin of multicellularity) has been also carried out on both experimentally tractable model organisms (Ratcliff *et al.*, 2012), as well as in a unicellular relative of animals (Burnetti and Ratcliff, 2022). In the same way, computer simulations of digital organisms revealed important aspects of evolutionary dynamics (Adami, Ofria and Collier, 2000). Now, the outbreak of the COVID-19 pandemic provides an unprecedented opportunity to check for phylogenetic trends by analyzing a natural evolutionary process in real time, which could provide helpful information on the adaptive process of the viral genome to its human host.

Pioneering works showed that RNA viruses are excellent material for studies of evolutionary genomics (Domingo, Webster and Holland, 1999; Worobey and Holmes, 1999; Moya, Holmes and González-Candelas, 2004). Despite the difficulties of inferring reliable phylogenies of SARS-CoV-2 (Morel *et al.*, 2020; Pipes *et al.*, 2021), as well as the controversy surrounding the first days and location of the pandemic (Koopmans *et al.*, 2021; Worobey, 2021), the most parsimonious explanation for the origin of SARS-CoV-2 seems to lie in a zoonotic event (Holmes *et al.*, 2021; Balloux *et al.*, 2022). Direct bat-to-human spillover events may occur more often than reported, although most remain unknown (Sánchez *et al.*, 2022). Bats are known as the natural reservoirs of SARS-like CoVs, and early evidence exists for the recombinant origin of bat (SARS)-like coronaviruses (Hon *et al.*, 2008). A genomic comparison between these coronaviruses and SARS-CoV-2 has led to propose a bat origin of the COVID-19 outbreak (Zhang and Holmes, 2020). Indeed, a recombination event between the bat coronavirus and either an origin-unknown coronavirus (Ji *et al.*, 2020) or a pangolin virus (Zhang, Wu and Zhang, 2020) would lie at the origin of SARS-CoV-2. Bat RaTG13 virus best matches the overall codon usage pattern of SARS-CoV-2 in orf1ab, spike, and nucleocapsid genes, while the pangolin P1E virus has a more similar codon usage in the membrane gene (Gu *et al.*, 2020). Other intermediate hosts have been identified, such as RaTG15, and this knowledge is essential to prevent the further spread of the epidemic (Liu *et al.*, 2020).

Despite its proofreading mechanism and the brief time-lapse since its appearance, SARS-CoV-2 has accumulated an important amount of genomic and genetic variability (Elbe and Buckland-Merrett, 2017; Hatcher *et al.*, 2017; Hadfield *et al.*, 2018; Dorp *et al.*, 2020; Islam *et al.*, 2020; Hamed *et al.*, 2021; McBroome *et al.*, 2021), dramatically impacting viral nucleotide composition and genome organization. Synonymous and non-synonymous mutations (Banerjee *et al.*, 2020; Cai, Cai and Li, 2020; González-Candelas *et al.*, 2021), as well as mismatches and deletions in translated and untranslated regions (Islam *et al.*, 2020; Young *et al.*, 2020) have been tracked in the SARs-CoV-2 genome. This may be related to both its recombinational origin (Naqvi *et al.*, 2020) as well as mutation and additional recombination events accumulated later (Cyranoski, 2020; Jackson *et al.*, 2021).

Recent phylogenetic estimates of the substitution rate of SARS-CoV-2 suggest that its genome accumulates around two mutations per month. However, Variants of Concern (VoCs) can have 15 or more defining mutations, and it is hypothesized that they emerged over the course of a few months, implying that they must have evolved faster for a period of time (Tay *et al.*, 2022). Noteworthy, RNA viruses can also accumulate high genetic variation during individual outbreaks (Pybus, Tatem and Lemey, 2015), showing mutation and evolutionary rates up to a million times higher than those of their hosts (Islam *et al.*, 2020).

Particularly interesting are those changes increasing viral fitness (Dorp *et al.*, 2020; Garvin *et al.*, 2020; Zhou *et al.*, 2020; Holmes *et al.*, 2021), such as mutations giving rise to epitope loss and antibody escape mechanisms. These have mainly been found in evolved variants isolated from Europe and the Americas, and have critical implications for SARS-CoV-2 fitness (transmission, pathogenesis, and immune interventions (Gupta and Mandal, 2020; Loucera *et al.*, 2022)). Some studies have shown that SARS-CoV-2 is acquiring mutations more slowly than expected for neutral evolution, suggesting that purifying selection is the dominant mode of evolution, at least during the initial phase of the pandemic time course. Parallel mutations in multiple independent lineages and variants have been observed (Dorp *et al.*, 2020), which may indicate convergent evolution, and are of particular interest in the context of adaptation of SARS-CoV-2 to the human host. Survival analysis of mutations in the SARS-CoV-2 genome revealed 27 of them were significantly associated with higher mortality of patients (Loucera *et al.*, 2022). Other authors have reported some sites under positive pressure in the nucleocapsid and spike genes (Benvenuto *et al.*, 2020). This impressive research effort has allowed tracking all these changes in real-time. The CoVizu^{e} project (https://filogeneti.ca/covizu/) provides a near real-time visualization of SARS-CoV-2 global diversity, the COVID-19 CG website (Chen *et al.*, 2021) tracks SARS-CoV-2 mutation and lineage by locations and dates of interest, while the CoV-Spectrum website (Chen *et al.*, 2022) supports the identification of new SARS-CoV-2 variants of concern and the tracking of known variants. Another recent developed tool (Sanderson, 2022) allows a visualization of mutation-annotated trees of millions SARS-CoV-2 sequences (https://cov2tree.org/).

Nucleotide compositional biases throughout the genome have been identified at all levels of the phylogenetic hierarchy, including RNA virus (Gaunt and Digard, 2022), being caused either by active selection or passive mutation pressure (Mooers and Holmes, 2000). The array of compositional domains in a genome can be potentially altered by most sequence changes (i.e., synonymous and non-synonymous nucleotide substitutions, insertions, deletions, recombination events, chromosome rearrangements, or genome reorganizations). Compositional domain structure can be altered either by changing nucleotide frequencies in a given region or by changing the nucleotides at the borders separating two domains, thus enlarging/shortening a given domain, or changing the number of domains (Bernaola-Galván, Román-Roldán and Oliver, 1996; Oliver *et al.*, 1999; Wen and Zhang, 2003; Keith, 2008). Ideally, a metric of nucleotide compositional heterogeneity should be able to summarize all the mutational and recombinational events accumulated by a genome sequence over time (Román-Roldán, Bernaola-Galván and Oliver, 1998; Bernaola-Galván *et al.*, 2004; Fearnhead and Vasilieou, 2009).

In many organisms, the patchy sequence structure formed by the array of compositional domains with different nucleotide composition (i.e., GC content) has been related to important biological features, as gene and repeat densities, timing of gene expression, recombination frequency, etc. (Bernardi *et al.*, 1985; Oliver *et al.*, 2004; Bernaola-Galván, Carpena and Oliver, 2008; Bernardi, 2015). Therefore, changes in sequence compositional heterogeneity may be relevant on evolutionary and epidemiological grounds. Specifically, the existence of evolutionary trends in the compositional complexity of the coronavirus could reveal adaptive processes of the virus to the human host.

To search for such trends, we computed the Sequence Compositional Complexity, or *SCC* (Román-Roldán, Bernaola-Galván and Oliver, 1998), an entropic measure of nucleotide compositional heterogeneity, representing the number of domains and nucleotide differences among them, identified in a genome sequence through a proper segmentation algorithm (Bernaola-Galván, Román-Roldán and Oliver, 1996). By using phylogenetic ridge regression, a method able to reveal both macro- (Melchionna *et al*., 2019; Serio *et al.*, 2019) and micro-evolutionary (Moya *et al.*, 2020) trends, we present here evidence for long-term adaptive tendencies of decreasing sequence compositional heterogeneity, and an increasing one for its evolutionary rate, in SARS-CoV-2. Both trends are shared by its most important VoCs (Alpha and Delta), being greatly accelerated by the recent rise to dominance of Omicron (Du, Gao and Wang, 2022).

Results

Sequence compositional complexity in the coronavirus

The first SARS-CoV-2 coronavirus genome sequence obtained at the onset of the pandemic (2019-12-30) was divided into eight compositional domains by the compositional segmentation algorithm (Bernaola-Galván, Román-Roldán and Oliver, 1996; Oliver *et al.*, 1999, 2004; Bernaola-Galván, Carpena and Oliver, 2008), resulting in a *SCC* value of 5.7 x 10E-3 bits by sequence position (Figure 1).

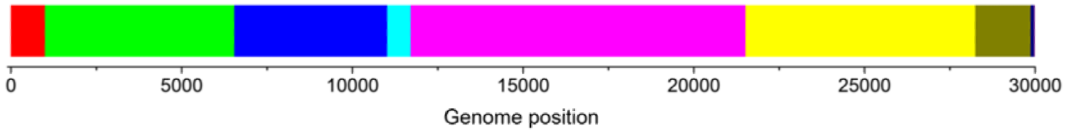

Figure 1: Compositional segmentation of the GISAID reference genome (hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30: https://www.gisaid.org/resources/hcov-19-reference-sequence/). Using an iterative segmentation algorithm (Bernaola-Galván, Román-Roldán and Oliver, 1996; Oliver *et al.*, 2004), the RNA sequence was divided into eight compositionally homogeneous segments with P value ≤ 0.05 (i.e., compositional domains). The genome position of domain borders is shown on the horizontal scale. Colors are used only to illustrate the differential nucleotide composition of each domain.

From then on, descendent coronaviruses have presented substantial variation in each domain's number, length, and nucleotide composition, which is reflected in a variety of *SCC* values. The number of segments ranges between 4 and 10, while the *SCC* do so between 2.71E-03 and 6.8E-03 bits by sequence position. The strain name, the collection date, and the *SCC* values for each analyzed genome are shown in Supplementary Tables S1-S18 available in the open repository Zenodo (https://doi.org/10.5281/zenodo.6844917).

Temporal evolution of SCC over the coronavirus pandemic

To characterize the temporal evolution of *SCC* over the time course of the coronavirus pandemic, we downloaded from GISAID/Audacity (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017; Khare *et al.*, 2021) a series of random samples of high-quality genome sequences over consecutive time lapses, each starting at the outbreak of the COVID-19 (December 2019) and progressively including younger samples up to March 2022 (Table 1). In each random sample, we filtered and masked the genome sequences using the GenBank reference genome MN908947.3 to eliminate sequence oddities (Hodcroft, Domman, *et al.*, 2021). Non-duplicated genome sequences were aligned with *MAFFT* (Katoh and Standley, 2013), then inferring the best ML timetree using *IQ-TREE 2* (Minh *et al.*, 2020), which was then rooted to the GISAID reference genome (hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30). The proportion of variant genomes in each sample was determined with *Nextclade* (Aksamentov *et al.*, 2021) (Table 1, columns 5-8).

Finally, we sought temporal phylogenetic trends in *SCC* values and evolutionary rates by using the function *search.trend* in the *RRphylo* R package (Castiglione *et al.*, 2018), contrasting the realized slope of *SCC* versus time regression to a family of 1,000 slopes generated under the BM model of evolution, which models evolution with no trend in either the *SCC* or its evolutionary rate. We found that SARS-CoV-2 sequence compositional heterogeneity did not follow any trend in *SCC* during the first year of the pandemic time course, as indicated by the non-significant *SCC* against time regressions in any sample ending before December 2020 (Table 1). With the emergence of variants in December 2020 (s1573, Table 1), the SCC started to decrease significantly over time. In contrast to the decreasing trend observed for *SCC*, a clear tendency towards faster evolutionary rates occurred throughout the study period, indicating that the virus increased in variability early on but took on a monotonic trend in declining *SCC* as VoCs appeared. These results were robust to several sources of uncertainty, including those related to the algorithms used for multiple alignment or to infer phylogenetic trees (see the section 'Checking results reliability' in Supplementary Information). In summary, these analyses show that statistically significant trends for declining SCC began between the end of December 2020 (s1573) and March 2021 (s1871) corresponding with the emergence of the first VoC (Alpha), a path that continued with the successive emergence of other variants. This may suggest a role for purifying selection in the evolution of SCC in the coronavirus.

Table 1. Phylogenetic trends in random coronavirus samples downloaded from the GISAID database Audacity (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017; Khare *et al.*, 2021) covering the pandemic time range from December 2019 to March 2022. For each sample, the analyzed time range was from December 2019 to the date shown in the column 'Collection date'. Initial sample sizes were 500, 1,000, 2,000, or 3,000 genomes per sample, while the final sample size indicates the remaining genome sequences once duplicated sequences were discarded. Non-duplicated genomes in each sample were then aligned with *MAFFT* (Katoh and Standley, 2013) to the GenBank MN908947.3 reference genome sequence and masked to eliminate sequence oddities (Hodcroft, De Maio, *et al.*, 2021). The best ML timetree for each sample was inferred using IQ-TREE 2 (Minh *et al.*, 2020), which was rooted to the GISAID reference genome (hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30). The percentages of variant genomes were determined by *Nextclade* (Aksamentov *et al.*, 2021). The genome heterogeneity of each coronavirus genome was determined by computing its Sequence Compositional Complexity, or SCC (Román-Roldán, Bernaola-Galván and Oliver, 1998). Phylogenetic ridge regressions for SCC and its evolutionary rate were computed by the function *search.trend* from the *RRphylo* R package (Castiglione *et al.*, 2018). The estimated genomic value for each tip or node in the phylogenetic tree is regressed against age. The statistical significance of the ridge regression slope was tested against 1,000 slopes obtained after simulating a simple (i.e., no-trend) Brownian evolution of SCC in the phylogenetic tree. See Methods for further details.

		Sample size		**% Of main variants**			**Total of variants in the sample (%)**	***SCC* regression**		**Rate regression**	
Sample	**Collection date**	**Initial**	**Final**	**Alpha**	**Delta**	**Ómicron**		**Slope**	**P value**	**Slope**	**P value**
s726	mar-20	1,000	726	0.00	0.00	0.00	0.00	-97.10	0.250	88,684.70	0.002
s730		1,000	730	0.00	0.00	0.00	0.00	-65.20	0.268	145,566.20	0.001
s781		2,000	781	0.00	0.00	0.00	0.00	-9.67	0.426	125,140.00	0.001
s1170	jun-20	2,000	1,170	0.00	0.00	0.00	0.00	12.83	0.444	87,637.54	0.001
s1277	sep-20	2,000	1,277	0.00	0.00	0.00	0.00	-20.85	0.305	39,183.10	0.001
s1573	dec-20	2,000	1,573	4.32	0.00	0.00	4.83	-38.53	0.066	26,502.59	0.001
s1871	mar-21	2,000	1,871	50.03	0.00	0.00	57.03	-61.94	0.001	14,254.56	0.001
s498	may-21	500	498	56.43	0.00	0.00	64.65	-66.39	0.011	15,035.58	0.001
s496		500	496	64.52	1.41	0.00	73.79	-55.74	0.026	11,090.87	0.001
s987		1,000	987	57.85	0.20	0.00	67.17	-60.29	0.001	16,937.26	0.001
s980		1,000	980	65.31	0.82	0.00	75.41	-54.23	0.004	16,169.35	0.001
s1939		2,000	1,939	63.02	1.24	0.00	72.93	-41.74	0.010	13,044.29	0.001
s1974	jun-21	2,000	1,974	45.90	7.14	0.00	60.59	-34.83	0.016	18,624.45	0.001
s1985	sep-21	2,000	1,985	27.10	44.08	0.00	78.34	-19.30	0.131	11,688.97	0.001
s1994	dec-21	2,000	1,994	17.95	57.82	6.22	86.70	-20.93	0.060	7,495.37	0.001
s2347		3,000	2,347	18.41	46.66	0.00	73.11	-33.38	0.007	7,217.62	0.001
s1990	mar-22	2,000	1,990	14.32	51.41	18.29	87.28	-21.89	0.037	4,896.06	0.052
	TOTAL:	28,000	23,318								

Relative contributions of individual variants to the SARS-CoV-2 evolutionary trends

SCC trends of variants

We estimated the relative contribution of the three main VoCs (Alpha, Delta, and Omicron) to the trends in SARS-CoV2 evolution by picking samples both before (s726, s730) and after (s1871, s1990) their appearance. The trends for *SCC* and its evolutionary rate in sample s1990, which includes a sizeable number of Omicron genomes, are shown in Figure 2. In all these samples, we tested trends for variants individually (as well as for the samples' trees as a whole) while accounting for phylogenetic uncertainty, by randomly altering the phylogenetic topology and branch lengths 100 times per sample (see Methods, and Supplementary Information for details). These cautions seem to us necessary to ensure accuracy in the conclusions based on the SARs-CoV-2 phylogenies we inferred (Wertheim, Steel and Sanderson, 2022). In agreement with the previous analyses (seventeen consecutive bins, see Table 1), we found strong support for a decrease in *SCC* values through time along phylogenies including variants (s1871, s1990) and no support for any temporal trend in older samples. Just four out of the 200 random trees produced for samples s726 and s730 produced a trend in *SCC* evolution. The corresponding figure for the two younger samples is 186/200 significant and negative instances of declining *SCC* over time (Table 2). This ~50-fold increase in the likelihood of finding a consistent trend in declining *SCC* over time is shared unambiguously by all tested variants (Alpha, Delta, and Omicron; Table 3). Yet, Omicron shows a significantly stronger decline in *SCC* than the other variants (Table 3), suggesting that the trends starting with the appearance of the main variants became stronger with the emergence of Omicron by the end of 2021.

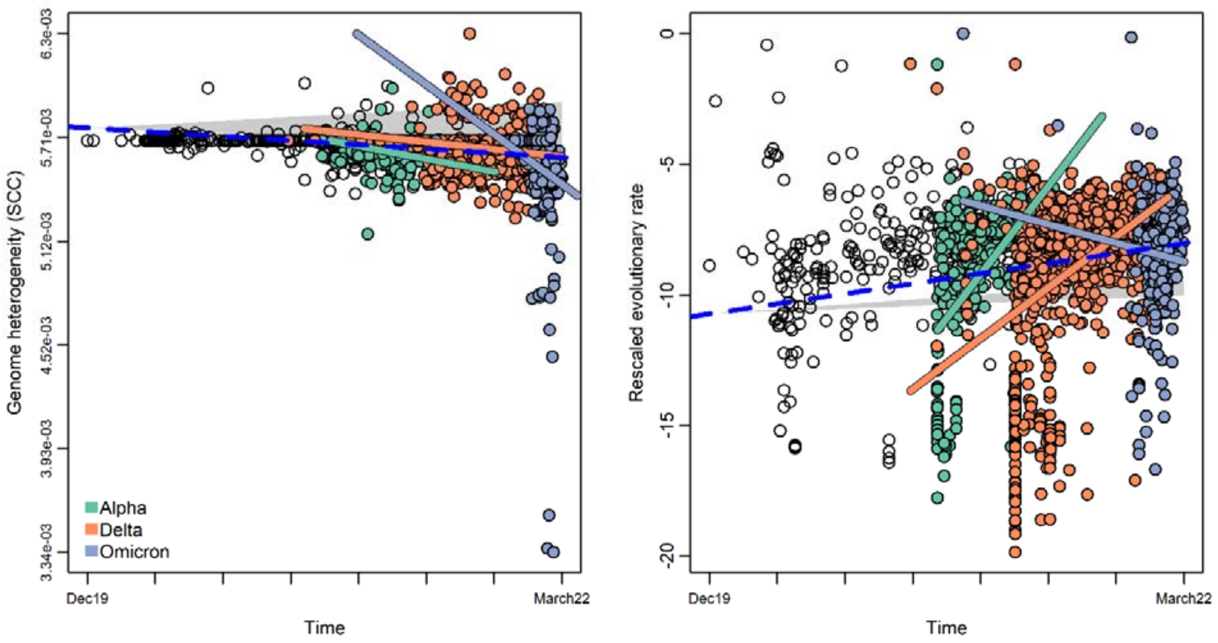

Figure 2: Phylogenetic ridge regressions for *SCC* (left) and its evolutionary rate (right) as detected by the *RRphylo* R package (Castiglione *et al.*, 2018) on the s1990 sample. For *SCC*, the estimated value for each tip in the phylogenetic tree is regressed (blue line) against its age (the phylogenetic time distance, meant mainly as the collection date of each virus isolate). The rescaled evolutionary rate was obtained by rescaling the absolute rate in the 0-1 range and then transforming to logs to compare to the BM expectation. The statistical significance of the ridge regression slopes was tested against

1,000 slopes obtained after simulating a simple Brownian evolution of the *SCC* in the phylogenetic tree. The 95% confidence intervals around each point produced according to the BM model of evolution are shown as shaded areas. Dots are colored according to the variant they belong to or left blank for strains collected before the appearance of variants.

We tested the difference in the slopes of *SCC* values versus time regression computed by grouping all the variants under a single group and the same figure for all other strains grouped together. The test was performed using the function *emtrends* available within the R package *emmeans* (Lenth, 2022). We found the slope for the group that includes all variants to be significantly larger than the slope for the other strains (estimate = -0.772 x 10^{-8}, P-value = 0.006), still pointing to the decisive effect of VoCs on *SCC* temporal trend.

Table 2. Percentages of significant results of *SCC* and *SCC* evolutionary rates versus time regressions performed on 100 randomly fixed (and subsampled for s1871 and s1990) phylogenetic trees. Higher/lower than BM = the percentage of simulation producing slopes significantly higher/lower than the BM expectation.

	SCC values		*SCC* evolutionary rates	
Sample	positive	negative	positive	negative
s726	0	4	88	0
s730	0	0	100	0
s1871	0	100	38	0
s1990	0	86	100	0

SCC evolutionary rates of variants

SCC evolutionary rate (absolute magnitude of the rate) tends to increase over time (Table 2). The slope of *SCC* rates through time regression for Omicron was always significantly lower than the slope computed for the rest of the tree (Table 3). This was also true for Alpha and Delta, although with much lower support.

Table 3. Percentages of significant results of *SCC* and *SCC* evolutionary rates versus time regressions performed on 100 randomly resolved (s1871 and s1990) phylogenetic trees. % Slope difference indicates the percentage of simulations producing significantly higher/lower slopes than the rest of the tree.

Sample	Variant		*SCC* values	*SCC* evolutionary rates
			% Slope difference	% Slope difference
s1871	Alpha	positive	0	0
		negative	99	39
s1990	Alpha	positive	0	0
		negative	100	0
	Delta	positive	0	0
		negative	91	27
	Omicron	positive	0	0
		negative	94	100

Discussion

Here we show that despite its short length (29,912 bp for the reference genome) and the brief time-lapse analyzed (28 months), the coronavirus RNA genome sequences can be segmented (Fig. 1 and Supplementary Tables S1-S18) into 4-10 compositional domains (~0.27 segments by kbp on average). Although such segment density is lower than in free-living organisms, like cyanobacteria where an average density of 0.47 segments by kbp was observed (Moya *et al.*, 2020), it may suffice for comparative evolutionary analyses of SCC in these genomes, which might shed light on the origin and evolution of the COVID-19 pandemic.

In early samples (i.e., collected before the emergence of variants), we found no statistical support for any trend in *SCC* values over time, although the virus as a whole appears to evolve faster than BM expectation. However, in samples taken after the first higher fitness VoC with higher transmissibility (Alpha) appeared in the GISAID database (December 2020), we started to detect statistically significant downward trends in *SCC* (Table 1). Concomitantly to the temporal decay in *SCC*, its absolute evolutionary rate kept increasing with time, meaning that the decline in *SCC* itself accelerated over time. In agreement with this notion, although declining *SCC* is an evolutionary path shared by variants, the nearly threefold increase in rates intensified after the appearance of the most recent VoC (Omicron) in late 2021, which shows a much faster decline in *SCC* than the other variants

(Table 3). These results indicate the existence of a driven, probably adaptive, trend in the variants toward a reduction of SCC.

The emergence of VOCs has been associated to an episodic increase in the substitution rate of around 4-fold the background phylogenetic rate estimate (Tay *et al.*, 2022). It is also known that variant genomes have accumulated a higher proportion of adaptive mutations, which allows them to neutralize host resistance or escape host antibodies (Mlcochova *et al.*, 2021; Thorne *et al.*, 2021; Venkatakrishnan *et al.*, 2021), consequently gaining higher transmissibility (a paradigmatic example is the recent outbreak of the Omicron variant). The sudden increases in fitness of variant genomes, may be due to the gathering of co-mutations that become prevalent world-wide compared to single mutations, being largely responsible for their temporal changes in transmissibility and virulence (Ilmjärv *et al.*, 2021; Majumdar and Niyogi, 2021). In fact, more contagious and perhaps more virulent VoCs share mutations and deletions that have arisen recurrently in distinct genetic backgrounds (Richard *et al.*, 2021). We show here that these increases in fitness of variant genomes, associated with a higher transmissibility, lead to a reduction of their sequence compositional heterogeneity, thus explaining the general decay of *SCC* in line with the pandemic expansion. We conclude that the accelerated loss of SCC in the coronavirus is promoted by the rise of high viral fitness variants, leading to adaptation to the human host, a well-known process in other viruses (Bahir *et al.*, 2009). Further monitoring of the evolutionary trends in current and new co-mutations, variants, and recombinant lineages (Callaway, 2022; Ledford, 2022; Straten *et al.*, 2022) by means of the tools used here will enable to elucidate whether and to what extent the evolution of SCC in the virus impacts human health.

Methods

Data retrieval, filtering, masking and alignment

The sequences of the random samples of high-quality coronavirus genomes we retrieved from the GISAID/Audacity database (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017; Khare *et al.*, 2021) were compiled as EPI_SET_20220604yp, being available at https://doi.org/10.55876/gis8.220604yp. *MAFFT* (Katoh and Standley, 2013) was used to align each random sample to the genome sequence of the isolate Wuhan-Hu-1 (GenBank accession MN908947.3), then filtering and masking the alignments to avoid sequence oddities (Hodcroft, Domman, *et al.*, 2021). In order to check the reliability of our results (see the section 'Checking results reliability' in Supplementary Information), we also analyzed other 3,059 genomes of the SARS-CoV-2

Nextstrain global dataset (Hadfield *et al.*, 2018) downloaded from https://nextstrain.org/ncov/open/global?f_host=Homo%20sapiens on 2021-10-08.

Phylogenetic trees

The best ML timetree for each random sample in Table 1 was inferred using *IQ-TREE 2* (Minh *et al.*, 2020), using the GTR nucleotide substitution model (Tavaré, 1986; Rodríguez *et al.*, 1990) and the least square dating (*LSD2*) method (To *et al.*, 2016), finally rooting the timetree to the GISAID coronavirus reference genome (EPI_ISL_402124, hCoV-19/Wuhan/WIV04/2019, WIV04).

Compositional segmentation algorithm

To divide the coronavirus genome sequence into an array of compositionally homogeneous, non-overlapping domains, we used a heuristic, iterative segmentation algorithm (Bernaola-Galván, Román-Roldán and Oliver, 1996; Oliver *et al.*, 1999, 2004; Bernaola-Galván, Carpena and Oliver, 2008). We chose the Jensen-Shannon divergence as the divergence measure between adjacent segments, as it can be directly applied to symbolic nucleotide sequences. At each iteration, we used a significance threshold (s = 0.95) to split the sequence into two segments whose nucleotide composition is homogeneous at the chosen significance level, s. The process continued iteratively over the new resulting segments while sufficient significance continued to appear.

Computing the Sequence Compositional Complexity (SCC)

Once each coronavirus genome sequence was segmented into an array of statistically significant, homogeneous compositional domains, its nucleotide compositional heterogeneity was measured by computing the Sequence Compositional Complexity, or *SCC* (Román-Roldán, Bernaola-Galván and Oliver, 1998). *SCC* increased with both the number of domains in the genome and the degree of compositional differences among them. Thus, *SCC* is analogous to other biological complexity measures, particularly to that described by McShea and Brandon (McShea and Brandon, 2010), in which an organism is more complex if it has a greater number of parts and a higher differentiation among these parts. It should be emphasized that *SCC* is overly sensitive to any change in the RNA genome sequence, either nucleotide substitutions, indels, genome rearrangements, or recombination events, all of which could alter the number of domains or the differences in nucleotide frequencies among them.

Phylogenetic ridge regression

To search for trends in *SCC* values and evolutionary rates over time, phylogenetic ridge regression was applied using the *RRphylo* R package V. 2.5.8 (Castiglione *et al.*, 2018). The estimated *SCC* value for each tip or node in the phylogenetic tree was regressed against its age (the phylogenetic time distance, which represents the time distance between the first sequence ever of the virus and the collection date of individual virus isolates); the regression slope was then compared to BM expectation (which models evolution according to no trend in *SCC* values and rates over time) by generating 1,000 slopes simulating BM evolution on the phylogenetic tree, using the function *search.trend* (Castiglione *et al.*, 2019) in the *RRphylo* R package.

Comparing the effects of variants on the evolutionary trend

In order to explicitly test the effect of variants and to compare variants among each other we selected 4 different trees and SCC data (s730, a727, s1871, s1990) from Table 1. In each sample, we accounted for phylogenetic uncertainty by producing 100 dichotomous versions of the initial tree by removing polytomies applying the *RRphylo* function *fix.poly* (Castiglione *et al.*, 2018). This function randomly resolves polytomous clades by adding non-zero length branches to each new node and equally partitioning the evolutionary time attached to the new nodes below the dichotomized clade. Each randomly fixed tree was used to evaluate temporal trends in *SCC* and its evolutionary rates occurring on the entire tree and individual variants if present, by applying *search.trend*. Additionally, for the larger phylogenies (i.e., s1871 and s1990 lineage-wise trees) half of the tree was randomly sampled and half of the tips were removed. This way we avoided biasing the results due to different tree sizes.

Supplementary Information

Additional details regarding the methods used in this study are provided in the Supplementary Information and in the supplemental data files available in the open repository Zenodo (https://doi.org/10.5281/zenodo.6844917).

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. A complete list acknowledging all originating and submitting laboratories is available in the GISAID's

EpiCoV database (Elbe and Buckland-Merrett, 2017; Shu and McCauley, 2017; Khare *et al.*, 2021) (EPI_SET_ID: EPI_SET_20220604yp; DOI: https://doi.org/10.55876/gis8.220604yp). In the same way, we gratefully acknowledge the authors, originating and submitting laboratories of the genome sequences we used for the analysis of the SARS-CoV-2 *Nextstrain* global dataset (Hadfield *et al.*, 2018), downloaded on 2021-10-08; a complete acknowledgement list is shown in Supplementary Table S19 available in Zenodo (https://zenodo.org/record/6844917).

Author contributions

J.L.O., M.V. and A.M. designed research; J.L.O., P.B., F.P., C.G.M, S.C., P.R., M.V. and A.M. performed research. J.L.O., P.B., F.P., C.G.M, S.C., P.R., M.V. and A.M. analyzed data; J.L.O., M.V., A.M. and P.R. drafted the paper. All authors read and approved the final manuscript.

Funding

This project was funded by grants from the Spanish Minister of Science, Innovation and Universities (former Spanish Minister of Economy and Competitiveness) to J.L.O. (Project AGL2017-88702-C2-2-R) and A.M. (Project PID2019-105969GB-I00), a grant from Generalitat Valenciana to A.M. (Project Prometeo/2018/A/133) and co-financed by the European Regional Development Fund (ERDF). The most time-demanding computations were done on the servers of the Laboratory of Bioinformatics, Dept. of Genetics & Institute of Biotechnology, Center of Biomedical Research, 18100, Granada, Spain.

Availability of data and materials

The data underlying this article are available in the open repository Zenodo at https://zenodo.org/, and can be accessed with https://zenodo.org/record/6844917.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare no competing interests.

References

Adami, C., Ofria, C. and Collier, T. C. (2000) 'Evolution of biological complexity.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 97(9), pp. 4463–8. doi: 10.1073/pnas.97.9.4463.

Aksamentov, I. *et al.* (2021) 'Nextclade: clade assignment, mutation calling and quality control for viral genomes', *Journal of Open Source Software*, 6(67), p. 3773. doi: 10.21105/joss.03773.

Bahir, I. *et al.* (2009) 'Viral adaptation to host: A proteome-based analysis of codon usage and amino acid preferences', *Molecular Systems Biology*. John Wiley & Sons, Ltd, 5(1), p. 311. doi: 10.1038/msb.2009.71.

Balloux, F. *et al.* (2022) 'The past, current and future epidemiological dynamic of SARS-CoV-2', *Oxford Open Immunology*. Oxford Academic, 3(1). doi: 10.1093/oxfimm/iqac003.

Banerjee, A. *et al.* (2020) 'The Novel Coronavirus Enigma: Phylogeny and Analyses of Coevolving Mutations Among the SARS-CoV-2 Viruses Circulating in India', *JMIR Bioinform Biotech 2020;1(1):e20735 https://bioinform.jmir.org/2020/1/e20735*. JMIR Bioinformatics and Biotechnology, 1(1), p. e20735. doi: 10.2196/20735.

Barrick, J. *et al.* (2009) 'Genome evolution and adaptation in a long-term experiment with Escherichia coli', *Nature*. Nature Publishing Group, 461(7268), pp. 1243–1247. doi: 10.1038/nature08480.

Benvenuto, D. *et al.* (2020) 'The 2019-new coronavirus epidemic: Evidence for virus evolution', *Journal of Medical Virology*. John Wiley and Sons Inc., 92(4), pp. 455–459. doi: 10.1002/jmv.25688.

Bernaola-Galván, P. *et al.* (2004) 'Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes', *Gene*, 333, pp. 121–133. doi: 10.1016/j.gene.2004.02.042.

Bernaola-Galván, P., Carpena, P. and Oliver, J. L. (2008) 'A standalone version of IsoFinder for the computational prediction of isochores in genome sequences', *arXiv preprint arXiv:0806.1292*, pp. 1–7. Available at: http://arxiv.org/abs/0806.1292 (Accessed: 26 March 2013).

Bernaola-Galván, P., Román-Roldán, R. and Oliver, J. L. (1996) 'Compositional segmentation and long-range fractal correlations in DNA sequences.', *Physical review E*, 53(5), pp. 5181–5189. doi: 10.1103/PhysRevE.53.5181.

Bernardi, G. *et al.* (1985) 'The mosaic genome of warm-blooded vertebrates', *Science*, 228(4702), pp. 953–958. doi: 10.1126/science.4001930.

Bernardi, G. (2015) 'Chromosome architecture and genome organization', *PLoS ONE*. Edited by D. I. Nurminsky. Public Library of Science, 10(11), p. e0143739. doi: 10.1371/journal.pone.0143739.

Burnetti, A. and Ratcliff, W. C. (2022) 'Experimental evolution is not just for model organisms', *PLoS Biology*. Public Library of Science, 20(3), p. e3001587. doi: 10.1371/journal.pbio.3001587.

Cai, H. Y., Cai, K. K. and Li, J. (2020) 'Identification of Novel Missense Mutations in a Large Number of Recent SARS-CoV-2 Genome Sequences', *Journal General Medical Research*, 2. doi: 10.20944/preprints202004.0482.v1.

Callaway, E. (2022) 'Are COVID surges becoming more predictable? New Omicron variants offer a hint', *Nature*, 605(7909), pp. 204–206. doi: 10.1038/d41586-022-01240-x.

Castiglione, S. *et al.* (2018) 'A new method for testing evolutionary rate variation and shifts in phenotypic evolution', *Methods in Ecology and Evolution*. Edited by N. Cooper. John Wiley & Sons, Ltd, 9(4), pp. 974–983. doi: 10.1111/2041-210X.12954.

Castiglione, S. *et al.* (2019) 'Simultaneous detection of macroevolutionary patterns in phenotypic means and rate of change with and within phylogenetic trees including extinct species', *PLoS ONE*. Public Library of Science, 14(1). doi: 10.1371/journal.pone.0210101.

Chen, A. T. *et al.* (2021) 'Covid-19 cg enables sars-cov-2 mutation and lineage tracking by locations and dates of interest', *eLife*, 10, pp. 1–15. doi: 10.7554/eLife.63409.

Chen, C. *et al.* (2022) 'CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants', *Bioinformatics*. Edited by C. Alkan. Oxford Academic, 38(6), pp. 1735–1737. doi: 10.1093/bioinformatics/btab856.

Cyranoski, D. (2020) 'Profile of a killer: the complex biology powering the coronavirus pandemic', *Nature*, pp. 22–26. doi: 10.1038/d41586-020-01315-7.

Domingo, E., Webster, R. G. and Holland, J. J. (1999) *Origin and evolution of viruses*. Academic Press.

Dorp, L. van *et al.* (2020) 'Emergence of genomic diversity and recurrent mutations in SARS-CoV-2', *Infection, Genetics and Evolution*. Elsevier, p. 104351. doi: 10.1016/j.meegid.2020.104351.

Du, P., Gao, G. F. and Wang, Q. (2022) 'The mysterious origins of the Omicron variant of SARS-CoV-2', *The Innovation*. Cell Press, 3(2), p. 100206. doi: 10.1016/J.XINN.2022.100206.

Elbe, S. and Buckland-Merrett, G. (2017) 'Data, disease and diplomacy: GISAID's innovative contribution to global health', *Global Challenges*. John Wiley & Sons, Ltd, 1(1), pp. 33–46. doi: 10.1002/gch2.1018.

Fearnhead, P. and Vasilieou, D. (2009) 'Bayesian Analysis of Isochores.', *Journal of the American Statistical Association*. Available at: http://eprints.lancs.ac.uk/26253/1/Isochorejasa.pdf (Accessed: 5 June 2013).

Garvin, M. R. *et al.* (2020) 'Potentially adaptive SARS-CoV-2 mutations discovered with novel spatiotemporal and explainable AI models', *Genome Biology*. BioMed Central, 21(1), p. 304. doi: 10.1186/s13059-020-02191-0.

Gaunt, E. R. and Digard, P. (2022) 'Compositional biases in RNA viruses: Causes, consequences and applications', *Wiley Interdisciplinary Reviews: RNA*. John Wiley & Sons, Ltd, 13(2), p. e1679. doi: 10.1002/wrna.1679.

González-Candelas, F. *et al.* (2021) 'One year into the pandemic: Short-term evolution of SARS-CoV-2 and emergence of new lineages', *Infection, Genetics and Evolution*. Elsevier, p. 104869. doi: 10.1016/j.meegid.2021.104869.

Good, B. H. *et al.* (2017) 'The dynamics of molecular evolution over 60,000 generations', *Nature*. Nature Research, 551(7678), pp. 45–50. doi: 10.1038/nature24287.

Gu, H. *et al.* (2020) 'Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses', *Virus Evolution*. Oxford Academic, 6(1). doi: 10.1093/ve/veaa032.

Gupta, A. M. and Mandal, S. (2020) 'Non-synonymous Mutations of SARS-Cov-2 Leads Epitope Loss and Segregates its Variants'. doi: 10.21203/RS.3.RS-29581/V1.

Hadfield, J. *et al.* (2018) 'Nextstrain: real-time tracking of pathogen evolution.', *Bioinformatics (Oxford, England)*. Oxford University Press, 34(23), pp. 4121–4123. doi: 10.1093/bioinformatics/bty407.

Hamed, S. M. *et al.* (2021) 'Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology', *Scientific Reports*. Nature Publishing Group, 11(1), p. 8435. doi: 10.1038/s41598-021-87713-x.

Hatcher, E. L. *et al.* (2017) 'Virus Variation Resource-improved response to emergent viral outbreaks', *Nucleic Acids Research*, 45(D1), pp. D482–D490. doi: 10.1093/nar/gkw1065.

Hodcroft, E. B., Domman, D. B., *et al.* (2021) 'Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting amino acid position 677.', *medRxiv: the preprint server for health sciences*. Cold Spring Harbor Laboratory Preprints. doi: 10.1101/2021.02.12.21251658.

Hodcroft, E. B., De Maio, N., *et al.* (2021) 'Want to track pandemic variants faster? Fix the bioinformatics bottleneck', *Nature*. Nature Research, pp. 30–33. doi: 10.1038/d41586-021-00525-x.

Holmes, E. C. *et al.* (2021) 'The Origins of SARS-CoV-2: A Critical Review', *Cell*. doi: 10.1016/j.cell.2021.08.017.

Hon, C.-C. *et al.* (2008) 'Evidence of the Recombinant Origin of a Bat Severe Acute Respiratory Syndrome (SARS)-Like Coronavirus and Its Implications on the Direct Ancestor of SARS Coronavirus', *Journal of Virology*. American Society for Microbiology, 82(4), pp. 1819–1826. doi: 10.1128/jvi.01926-07.

Ilmjärv, S. *et al.* (2021) 'Concurrent mutations in RNA-dependent RNA polymerase and spike protein emerged as the epidemiologically most successful SARS-CoV-2 variant', *Scientific Reports*. Nature Publishing Group, 11(1), p. 13705. doi: 10.1038/s41598-021-91662-w.

Islam, M. R. *et al.* (2020) 'Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity', *Scientific Reports*. Nature Research, 10(1). doi: 10.1038/s41598-020-70812-6.

Jackson, B. *et al.* (2021) 'Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic', *Cell*. Cell Press, 184(20), pp. 5179-5188.e8. doi: 10.1016/j.cell.2021.08.014.

Ji, W. *et al.* (2020) 'Cross-species transmission of the newly identified coronavirus 2019-nCoV', *Journal of Medical Virology*. John Wiley and Sons Inc., 92(4), pp. 433–440. doi: 10.1002/jmv.25682.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: Improvements in performance and usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. doi: 10.1093/molbev/mst010.

Keith, J. M. (2008) 'Sequence segmentation.', *Methods in molecular biology (Clifton, N.J.)*, 452, pp. 207–29. doi: 10.1007/978-1-60327-159-2_11.

Khare, S. *et al.* (2021) 'GISAID's Role in Pandemic Response', *China CDC Weekly*. China CDC Weekly, 3(49), pp. 1049–1051. doi: 10.46234/ccdcw2021.255.

Koopmans, M. *et al.* (2021) 'Origins of SARS-CoV-2: window is closing for key scientific studies', *Nature*. Nature Publishing Group, 596(7873), pp. 482–485. doi: 10.1038/d41586-021-02263-6.

Ledford, H. (2022) 'The next variant: three key questions about what's after Omicron', *Nature*, 603(7900), pp. 212–213. doi: 10.1038/d41586-022-00510-y.

Lenth, R. V (2022) 'emmeans: Estimated Marginal Means, aka Least-Squares Means. https://github.com/rvlenth/emmeans'. Available at: https://github.com/rvlenth/emmeans.

Liu, Z. *et al.* (2020) 'Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2', *Journal of Medical Virology*. John Wiley and Sons Inc., 92(6), pp. 595–601. doi: 10.1002/jmv.25726.

Loucera, C. *et al.* (2022) 'Assessing the Impact of SARS-CoV-2 Lineages and Mutations on Patient Survival', *Viruses*. Multidisciplinary Digital Publishing Institute, 14(9), p. 1893. doi: 10.3390/v14091893.

Majumdar, P. and Niyogi, S. (2021) 'SARS-CoV-2 mutations: The biological trackway towards viral fitness', *Epidemiology and Infection*. Cambridge University Press, p. e110. doi: 10.1017/S0950268821001060.

McBroome, J. *et al.* (2021) 'A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees', *Molecular Biology and Evolution*. Edited by J. Lu. Oxford Academic, 38(12), pp. 5819–5824. doi: 10.1093/molbev/msab264.

McShea, D. W. and Brandon, R. N. (2010) *Biology's first law□: the tendency for diversity and* *complexity to increase in evolutionary systems*. University of Chicago Press. Available at: https://books.google.es/books?hl=es&lr=&id=u0_OEwg9HckC&oi=fnd&pg=PR7&dq=McShea.+D.N.,+Brandon,+R.N.+2010.+Biology's+first+law.+Chicago+University+Press,+Chicago&ots=mp1Z3FxjMj&sig=JFwspSz9Zx0wlwT193QpQez4LF0&redir_esc=y#v=onepage&q&f=false (Accessed: 27 October 2018).

Melchionna, M. *et al.* (2019) 'Macroevolutionary trends of brain mass in Primates', *Biological Journal of the Linnean Society*. Oxford Academic, 129(1), pp. 14–25. doi: 10.1093/biolinnean/blz161.

Minh, B. Q. *et al.* (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*. Oxford University Press, 37(5), pp. 1530–1534. doi: 10.1093/molbev/msaa015.

Mlcochova, P. *et al.* (2021) 'SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion', *Nature*. Nature Publishing Group, pp. 1–8. doi: 10.1038/s41586-021-03944-y.

Mooers, A. and Holmes, E. (2000) 'The evolution of base composition and phylogenetic inference.', *Trends in ecology & evolution*, 15(9), pp. 365–369. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10931668.

Morel, B. *et al.* (2020) 'Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msaa314.

Moya, A. *et al.* (2020) 'Driven progressive evolution of genome sequence complexity in Cyanobacteria', *Scientific Reports*, 10(1). doi: 10.1038/s41598-020-76014-4.

Moya, A., Holmes, E. C. and González-Candelas, F. (2004) 'The population genetics and evolutionary epidemiology of RNA viruses', *Nature Reviews Microbiology*. Nat Rev Microbiol, pp. 279–288. doi: 10.1038/nrmicro863.

Naqvi, A. A. T. *et al.* (2020) 'Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach', *Biochimica et Biophysica Acta - Molecular Basis of Disease*. Elsevier B.V., p. 165878. doi: 10.1016/j.bbadis.2020.165878.

Oliver, J. L. *et al.* (1999) 'SEGMENT: identifying compositional domains in DNA sequences', *Bioinformatics*, 15(12), pp. 974–9. Available at:

http://bioinformatics.oxfordjournals.org/content/15/12/974.short (Accessed: 22 March 2013).

Oliver, J. L. *et al.* (2004) 'IsoFinder: computational prediction of isochores in genome sequences.', *Nucleic Acids Res*, 32(Web Server issue), pp. W287-92. doi: 10.1093/nar/gkh399.

Pipes, L. *et al.* (2021) 'Assessing Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny', *Molecular Biology and Evolution*. Edited by H. Malik. Oxford Academic, 38(4), pp. 1537–1543. doi: 10.1093/molbev/msaa316.

Pybus, O. G., Tatem, A. J. and Lemey, P. (2015) 'Virus evolution and transmission in an ever more connected world', *Proceedings of the Royal Society B: Biological Sciences*. Royal Society of London. doi: 10.1098/rspb.2014.2878.

Ratcliff, W. C. *et al.* (2012) 'Experimental evolution of multicellularity', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 109(5), pp. 1595–1600. doi: 10.1073/pnas.1115323109.

Richard, D. *et al.* (2021) 'A phylogeny-based metric for estimating changes in transmissibility from recurrent mutations in SARS-CoV-2'. doi: 10.1101/2021.05.06.442903.

Rodríguez, F. *et al.* (1990) 'The general stochastic model of nucleotide substitution', *Journal of Theoretical Biology*, pp. 485–501. doi: 10.1016/S0022-5193(05)80104-3.

Román-Roldán, R., Bernaola-Galván, P. and Oliver, J. L. (1998) 'Sequence compositional complexity of DNA through an entropic segmentation method', *Physical Review Letters*, 80(6), pp. 1344–1347. Available at: http://link.aps.org/doi/10.1103/PhysRevLett.80.1344 (Accessed: 22 March 2013).

Sánchez, C. A. *et al.* (2022) 'A strategy to assess spillover risk of bat SARS-related coronaviruses in Southeast Asia', *Nature Communications*. Nature Publishing Group, 13(1), p. 4380. doi: 10.1038/s41467-022-31860-w.

Sanderson, T. (2022) 'Taxonium: a web-based tool for exploring large phylogenetic trees', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2022.06.03.494608. doi: 10.1101/2022.06.03.494608.

Serio, C. *et al.* (2019) 'Macroevolution of Toothed Whales Exceptional Relative Brain Size', *Evolutionary Biology*. Springer, 46(4), pp. 332–342. doi: 10.1007/s11692-019-09485-7.

Shu, Y. and McCauley, J. (2017) 'GISAID: Global initiative on sharing all influenza data – from vision to reality', *Eurosurveillance*. European Centre for Disease Prevention and Control, p. 30494. doi: 10.2807/1560-7917.ES.2017.22.13.30494.

Straten, K. van der *et al.* (2022) 'Mapping the antigenic diversification of SARS-CoV-2', *medRxiv*. Cold Spring Harbor Laboratory Press, p. 2022.01.03.21268582. doi: 10.1101/2022.01.03.21268582.

Tavaré, S. (1986) 'Some probabilistic and statistical problems in the analysis of DNA sequences', *Lectures on mathematics in the life sciences*, 17(2), pp. 57–86.

Tay, J. H. *et al.* (2022) 'The Emergence of SARS-CoV-2 Variants of Concern Is Driven by Acceleration of the Substitution Rate.', *Molecular biology and evolution*. Mol Biol Evol, 39(2). doi: 10.1093/molbev/msac013.

Thorne, L. G. *et al.* (2021) 'Evolution of enhanced innate immune evasion by the SARS-CoV-2 B.1.1.7 UK variant'. doi: 10.1101/2021.06.06.446826.

To, T. H. *et al.* (2016) 'Fast Dating Using Least-Squares Criteria and Algorithms', *Systematic Biology*. Oxford University Press, 65(1), pp. 82–97. doi: 10.1093/sysbio/syv068.

Venkatakrishnan, A. J. *et al.* (2021) 'Antigenic minimalism of SARS-CoV-2 is linked to surges in COVID-19 community transmission and vaccine breakthrough infections'. doi: 10.1101/2021.05.23.21257668.

Wen, S.-Y. and Zhang, C.-T. (2003) 'Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis', *Biochemical and Biophysical Research Communications*, 311(1), pp. 215–222. doi: 10.1016/j.bbrc.2003.09.198.

Wertheim, J. O., Steel, M. and Sanderson, M. J. (2022) 'Accuracy in Near-Perfect Virus Phylogenies', *Systematic Biology*. Edited by V. Savolainen. Oxford Academic, 71(2), pp. 426–438. doi: 10.1093/sysbio/syab069.

Worobey, M. (2021) 'Dissecting the early COVID-19 cases in Wuhan', *Science*. American Association for the Advancement of Science, 374(6572), pp. 1202–1204. doi: 10.1126/science.abm4454.

Worobey, M. and Holmes, E. C. (1999) 'Evolutionary aspects of recombination in RNS viruses', *Journal of General Virology*. Microbiology Society, 80(May), pp. 2535–2543. doi: 10.1099/0022-1317-80-10-2535.

Young, B. E. *et al.* (2020) 'Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study.', *Lancet (London, England)*. Elsevier, 396(10251), pp. 603–611. doi: 10.1016/S0140-6736(20)31757-8.

Zhang, T., Wu, Q. and Zhang, Z. (2020) 'Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak', *Current Biology*. Cell Press, 30(7), pp. 1346-1351.e2. doi: 10.1016/j.cub.2020.03.022.

Zhang, Y. Z. and Holmes, E. C. (2020) 'A Genomic Perspective on the Origin and Emergence of SARS-CoV-2', *Cell*. Cell Press, 181(2), pp. 223–227. doi: 10.1016/j.cell.2020.03.035.

Zhou, P. *et al.* (2020) 'Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin', *Nature*. Cold Spring Harbor Laboratory, p. 2020.01.22.914952. doi: 10.1101/2020.01.22.914952.