The emergence of high-fitness variants accelerates the decay of genome heterogeneity in the 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,$,*]

[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

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

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

Keywords: Coronavirus evolution, phylogenetic evolutionary trends, genome heterogeneity, sequence compositional complexity.

E-mails:

José L. Oliver, oliver@ugr.es

Pedro Bernaola-Galván, rick@uma.es

Francisco Perfectti, fperfect@ugr.es

Cristina Gómez Martín, c.a.gomezmartin@amsterdamumc.nl

Silvia Castiglione, silviacastiglione2@gmail.com

Pasquale Raia, pasquale.raia@unina.it

Miguel Verdú, Miguel.Verdu@ext.uv.es

Andrés Moya, Andres.Moya@uv.es

Abstract

Since the outbreak of the COVID-19 pandemic, the SARS-CoV-2 coronavirus accumulated an important amount of genome compositional heterogeneity through mutation and recombination, which can be summarized by means of a measure of Sequence Compositional Complexity (SCC). To test evolutionary trends that could inform us on the adaptive process of the virus to its human host, we compute SCC in high-quality coronavirus genomes from across the globe, covering the full span of the pandemic. By using phylogenetic ridge regression, we find trends for SCC in the short time-span of SARS-CoV-2 pandemic expansion. In early samples, we find no statistical support for any trend in *SCC* values over time, although the virus genome appears to evolve faster than Brownian Motion expectation. However, in samples taken after the emergence of Variants of Concern with higher transmissibility, and controlling for phylogenetic and sampling effects, we detect a declining trend for *SCC* and an increasing one for its absolute evolutionary rate. This means that the decay in *SCC* itself accelerated over time, and that increasing fitness of variant genomes lead to a reduction of their genome sequence heterogeneity. Therefore, our work shows that phylogenetic trends, typical of macroevolutionary time scales, can be also revealed on the shorter time spans typical of viral genomes.

1 Introduction

Pioneering works showed that RNA viruses are excellent material for studies of evolutionary genomics (Domingo, Webster, and Holland 1999; Moya, Holmes, and González-Candelas 2004; Worobey and Holmes 1999). Now, with the outbreak of the COVID-19 pandemic, this has become a key research topic. Despite the difficulties of inferring reliable phylogenies of SARS-CoV-2 (Pipes et al. 2021; Morel et al. 2020), as well as the controversy surrounding the first days and location of the pandemic (Worobey 2021; Koopmans et al. 2021), the most parsimonious explanation for the origin of SARS-CoV-2 seems to lie in a zoonotic event (Holmes et al. 2021). Direct bat-to-human spillover events may occur more often than reported, although most remain unknown (Sánchez et al. 2021). 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 (Y. Z. 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 (T. 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 already accumulated an important amount of genomic and genetic variability (Elbe and Buckland-Merrett 2017; Hadfield et al. 2018; Hamed et al. 2021; Hatcher et al. 2017; Dorp et al. 2020; McBroome et al. 2021), which is due 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). Synonymous and non-synonymous mutations (Banerjee et al. 2020; Cai, Cai, and Li 2020), 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 sequence.

Particularly interesting changes are those increasing viral fitness (Holmes et al. 2021; Dorp et al. 2020; Zhou et al. 2020), 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 transmission, pathogenesis, and immune interventions (Gupta and Mandal 2020). 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 (Dorp et al. 2020). Other authors have reported some sites under positive pressure in the nucleocapsid and spike genes (Benvenuto et al. 2020). All this research effort has allowed tracking all these changes in real-time. The CoVizu[e] project (https://filogeneti.ca/covizu/) provides a visualization of SARS-CoV-2 global diversity of SARS-CoV-2 genomes.

Base composition varies at all levels of the phylogenetic hierarchy and throughout the genome, caused 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; Keith 2008; Oliver et al. 1999; Wen and Zhang 2003). Ideally, a genome sequence heterogeneity metric should be able to summarize all the mutational and recombinational events accumulated by a genome sequence over time (Bernaola-Galván et al. 2004; Fearnhead and Vasilieou 2009; Román-Roldán, Bernaola-Galván, and Oliver 1998).

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

To this end, we computed the Sequence Compositional Complexity, or *SCC* (Román-Roldán, Bernaola-Galván, and Oliver 1998), an entropic measure of genome-wide 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- (Serio et al. 2019; Melchionna et al. 2019) and micro-evolutionary (Moya et al. 2020) trends, we present here evidence for a long-term tendency of decreasing genome sequence heterogeneity in SARS-CoV-2. The trend is shared by its most important VoCs (Alpha and Delta) and greatly accelerated by the recent rise to dominance of Omicron (Du, Gao, and Wang 2022).

2 Results

2.1 Genome heterogeneity 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 our compositional segmentation algorithm (Bernaola-Galván, Román-Roldán, and Oliver 1996; Oliver et al. 1999; Bernaola-Galván, Carpena, and Oliver 2008; Oliver et al. 2004), resulting in a *SCC* value of 5.7 x 10E-3 bits by sequence position (Figure 1).

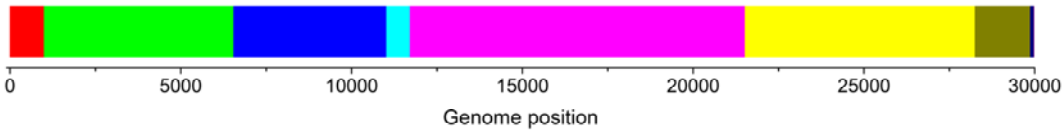

Figure 1: Compositional segmentation of the GISAID reference genome (hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30: https://www.gisaid.org/resources/hcov-19-reference-sequence/). Using an iterative segmentation algorithm (Bernaola-Galván, Román-Roldán, and Oliver 1996; Oliver et al. 2004), the RNA sequence was divided into eight compositionally homogeneous segments (i.e., compositional domains) with P value ≤ 0.05. 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 Zenodo (https://doi.org/10.5281/zenodo.6844917).

2.2 Temporal evolution of SCC over the coronavirus pandemic time course

To characterize the temporal evolution of *SCC* over the entire time course of the coronavirus pandemic (December 2019 to March 2022), we downloaded from GISAID/Audacity (Khare et al. 2021; Elbe and Buckland-Merrett 2017; Shu and McCauley 2017) 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 sample, we

filtered and masked these sequences using the GenBank reference genome MN908947.3 to eliminate sequence oddities (Hodcroft et al. 2021). Non-duplicated genomes were aligned with *MAFFT* (Katoh and Standley 2013), then inferring the best ML timetree 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 proportion of variant genomes in each sample was also determined (Table 1, columns 5-8).

Finally, we sought temporal trends in *SCC* values and evolutionary rates by using the function *search.trend* in the R package *RRphylo* (Silvia Castiglione et al. 2018), contrasting the realized slope of *SCC* versus time regression to a family of 1,000 slopes generated under the Brownian Motion (BM) model of evolution, which models evolution with no trend in either the *SCC* or its evolutionary rate. We found that SARS-CoV-2 genome sequence 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 genome sequence heterogeneity 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 Material). In summary, these analyses show that statistically significant trends for declining heterogeneity 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.

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 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 et al. 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 Materials & 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								

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

2.2.1 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 Materials and Methods, and Supplementary Material 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.

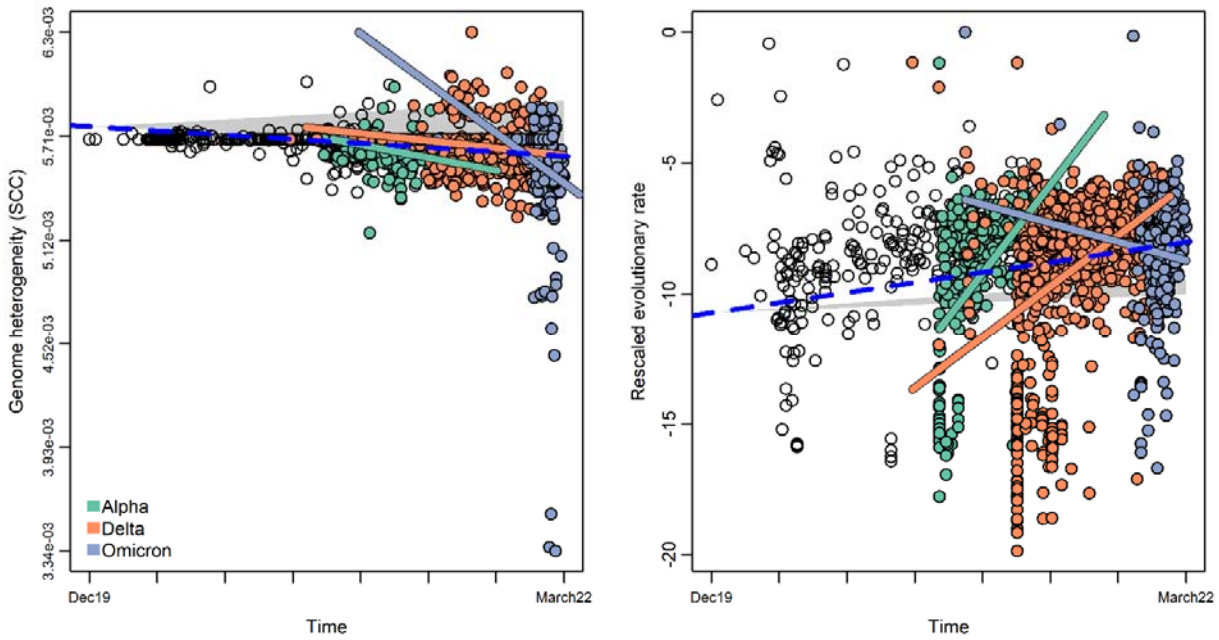

Figure 2: Phylogenetic ridge regressions for *SCC* (left) and its evolutionary rate (right) as detected by the *RRphylo* R package (Silvia Castiglione et al. 2018) on the s1990 sample. For *SCC*, the estimated value for each tip in the phylogenetic tree is regressed (blue line) against its age (the phylogenetic 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 = -0.772×10^{-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

2.2.2 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

3 Discussion

Here we show that despite its short length (29,912 bp for the reference genome) and the short time-lapse analyzed (28 months), the coronavirus RNA genome sequences can be segmented (Fig. 1) into 4-9 compositional domains (~0.27 segments by kbp on average). Although such segment density is lower than in free-living organisms, like cyanobacteria where we observe an average density of 0.47 segments by kbp (Moya et al. 2020), it may suffice for comparative evolutionary analyses of compositional sequence heterogeneity 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 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 genome sequence heterogeneity. Furthermore, the emergence of VOCs may be also associated to an episodic increase in the substitution rate of around 4-fold the background phylogenetic rate estimate (Tay et al. 2022). It is well established that variant genomes have accumulated a higher proportion of adaptive mutations, which allows them to neutralize host resistance or escape host antibodies (Thorne et al. 2021; Venkatakrishnan et al. 2021; Mlcochova 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 also 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 genome sequence heterogeneity, thus explaining the general decay of *SCC* in line with the pandemic expansion.

We conclude that the accelerated loss of genome heterogeneity 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 (Ledford 2022; Straten et al. 2022; Callaway 2022) by means of the tools used here will enable to elucidate whether and to what extent the evolution of genome sequence heterogeneity in the virus impacts human health.

4 Materials and Methods

4.1 Data retrieval, filtering, masking and alignment

We retrieved random samples of high-quality coronavirus genome sequences (EPI_SET_20220604yp, available at https://doi.org/10.55876/gis8.220604yp), from the GISAID/Audacity database (Khare et al. 2021; Elbe and Buckland-Merrett 2017; Shu and McCauley 2017). *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 et al. 2021). In order to check results reliability (see Supplementary Material), 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.

4.2 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).

4.3 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; Bernaola-Galván, Carpena, and Oliver 2008; Oliver et al. 2004). 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.

4.4 Computing the Sequence compositional complexity (SCC)

Once each coronavirus genome sequence was segmented into an array of statistically significant, homogeneous compositional domains, its genome sequence 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 segments 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 highly 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 compositional nucleotide differences among them.

4.5 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 (Silvia 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* (S Castiglione et al. 2019) in the *RRphylo* R package.

4.6 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* (Silvia 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.

5 Supplementary Material

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

6 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.

7 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 (Khare et al. 2021; Elbe and Buckland-Merrett 2017; Shu and McCauley 2017) (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).

8 Data availability

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

9 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.

10 Competing interests

The authors declare no competing interests.

11 References

Bahir, Iris, Menachem Fromer, Yosef Prat, and Michal Linial. 2009. "Viral Adaptation to Host: A Proteome-Based Analysis of Codon Usage and Amino Acid Preferences." *Molecular Systems Biology* 5 (1): 311. https://doi.org/10.1038/msb.2009.71.

Banerjee, Anindita, Rakesh Sarkar, Suvrotoa Mitra, Mahadeb Lo, Shanta Dutta, and Mamta Chawla-Sarkar. 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* 1 (1): e20735. https://doi.org/10.2196/20735.

Benvenuto, Domenico, Marta Giovanetti, Alessandra Ciccozzi, Silvia Spoto, Silvia Angeletti, and Massimo Ciccozzi. 2020. "The 2019-New Coronavirus Epidemic: Evidence for Virus Evolution." *Journal of Medical Virology* 92 (4): 455–59. https://doi.org/10.1002/jmv.25688.

Bernaola-Galván, Pedro, Pedro Carpena, and Jose L. Oliver. 2008. "A Standalone Version of IsoFinder for the Computational Prediction of Isochores in Genome Sequences." *ArXiv Preprint ArXiv:0806.1292*, 1–7. http://arxiv.org/abs/0806.1292.

Bernaola-Galván, Pedro, José L. Oliver, Pedro Carpena, Oliver Clay, and Giorgio Bernardi. 2004. "Quantifying Intrachromosomal GC Heterogeneity in Prokaryotic Genomes." *Gene* 333 (May): 121–33. https://doi.org/10.1016/j.gene.2004.02.042.

Bernaola-Galván, Pedro, Ramon Román-Roldán, and Jose L. Oliver. 1996. “Compositional Segmentation and Long-Range Fractal Correlations in DNA Sequences.” *Physical Review E* 53 (5): 5181–89. https://doi.org/10.1103/PhysRevE.53.5181.

Bernardi, G, B Olofsson, J Filipski, M Zerial, J Salinas, G Cuny, M Meunier-Rotival, and F Rodier. 1985. “The Mosaic Genome of Warm-Blooded Vertebrates.” *Science* 228 (4702): 953–58. https://doi.org/10.1126/science.4001930.

Bernardi, Giorgio. 2015. “Chromosome Architecture and Genome Organization.” Edited by Dmitry I Nurminsky. *PLoS ONE* 10 (11): e0143739. https://doi.org/10.1371/journal.pone.0143739.

Cai, Hugh Y., Kimberly K. Cai, and Julang Li. 2020. “Identification of Novel Missense Mutations in a Large Number of Recent SARS-CoV-2 Genome Sequences.” *Journal General Medical Research* 2 (May). https://doi.org/10.20944/preprints202004.0482.v1.

Callaway, Ewen. 2022. “Are COVID Surges Becoming More Predictable? New Omicron Variants Offer a Hint.” *Nature* 605 (7909): 204–6. https://doi.org/10.1038/d41586-022-01240-x.

Castiglione, S, C Serio, A Mondanaro, M Di Febbraro, A Profico, G Girardi, and P Raia. 2019. “Simultaneous Detection of Macroevolutionary Patterns in Phenotypic Means and Rate of Change with and within Phylogenetic Trees Including Extinct Species.” *PLoS ONE* 14 (1). https://doi.org/10.1371/journal.pone.0210101.

Castiglione, Silvia, Gianmarco Tesone, Martina Piccolo, Marina Melchionna, Alessandro Mondanaro, Carmela Serio, Mirko Di Febbraro, and Pasquale Raia. 2018. “A New Method for Testing Evolutionary Rate Variation and Shifts in Phenotypic Evolution.” *Methods in Ecology and Evolution* 9 (4): 974–83. https://doi.org/10.1111/2041-210X.12954.

Cyranoski, David. 2020. “Profile of a Killer: The Complex Biology Powering the Coronavirus Pandemic.” *Nature*. https://doi.org/10.1038/d41586-020-01315-7.

Domingo, Esteban., Robert G. Webster, and John J. Holland. 1999. *Origin and Evolution of Viruses*. Academic Press.

Dorp, Lucy van, Mislav Acman, Damien Richard, Liam P. Shaw, Charlotte E. Ford, Louise Ormond, Christopher J. Owen, et al. 2020. “Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2.” *Infection, Genetics and Evolution*, May, 104351. https://doi.org/10.1016/j.meegid.2020.104351.

Du, Pei, George Fu Gao, and Qihui Wang. 2022. “The Mysterious Origins of the Omicron Variant of

SARS-CoV-2.” *The Innovation* 3 (2): 100206. https://doi.org/10.1016/J.XINN.2022.100206.

Elbe, Stefan, and Gemma Buckland-Merrett. 2017. “Data, Disease and Diplomacy: GISAID’s Innovative Contribution to Global Health.” *Global Challenges* 1 (1): 33–46. https://doi.org/10.1002/gch2.1018.

Fearnhead, Paul, and Despina Vasilieou. 2009. “Bayesian Analysis of Isochores.” *Journal of the American Statistical Association*. http://eprints.lancs.ac.uk/26253/1/Isochorejasa.pdf.

Gu, Haogao, Daniel K W Chu, Malik Peiris, and Leo L M Poon. 2020. “Multivariate Analyses of Codon Usage of SARS-CoV-2 and Other Betacoronaviruses.” *Virus Evolution* 6 (1). https://doi.org/10.1093/ve/veaa032.

Gupta, Aayatti Mallick, and Sukhendu Mandal. 2020. “Non-Synonymous Mutations of SARS-Cov-2 Leads Epitope Loss and Segregates Its Variants,” May. https://doi.org/10.21203/RS.3.RS-29581/V1.

Hadfield, James, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. 2018. “Nextstrain: Real-Time Tracking of Pathogen Evolution.” *Bioinformatics (Oxford, England)* 34 (23): 4121–23. https://doi.org/10.1093/bioinformatics/bty407.

Hamed, Samira M., Walid F. Elkhatib, Ahmed S. Khairalla, and Ayman M. Noreddin. 2021. “Global Dynamics of SARS-CoV-2 Clades and Their Relation to COVID-19 Epidemiology.” *Scientific Reports* 11 (1): 8435. https://doi.org/10.1038/s41598-021-87713-x.

Hatcher, Eneida L, Sergey A Zhdanov, Yiming Bao, Olga Blinkova, Eric P Nawrocki, Yuri Ostapchuck, Alejandro A. Schaffer, and J. Rodney Brister. 2017. “Virus Variation Resource-Improved Response to Emergent Viral Outbreaks.” *Nucleic Acids Research* 45 (D1): D482–90. https://doi.org/10.1093/nar/gkw1065.

Hodcroft, Emma B., Daryl B. Domman, Daniel J. Snyder, Kasopefoluwa Oguntuyo, Maarten Van Diest, Kenneth H. Densmore, Kurt C. Schwalm, 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*, February. https://doi.org/10.1101/2021.02.12.21251658.

Holmes, Edward C, Stephen A Goldstein, Angela L Rasmussen, David L Robertson, Alexander Crits-

Christoph, Joel O Wertheim, Simon J Anthony, et al. 2021. “The Origins of SARS-CoV-2: A Critical Review.” *Cell*. https://doi.org/10.1016/j.cell.2021.08.017.

Hon, Chung-Chau, Tsan-Yuk Lam, Zheng-Li Shi, Alexei J. Drummond, Chi-Wai Yip, Fanya Zeng, Pui-Yi Lam, and Frederick Chi-Ching Leung. 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* 82 (4): 1819–26. https://doi.org/10.1128/jvi.01926-07.

Ilmjärv, Sten, Fabien Abdul, Silvia Acosta-Gutiérrez, Carolina Estarellas, Ioannis Galdadas, Marina Casimir, Marco Alessandrini, Francesco Luigi Gervasio, and Karl Heinz Krause. 2021. “Concurrent Mutations in RNA-Dependent RNA Polymerase and Spike Protein Emerged as the Epidemiologically Most Successful SARS-CoV-2 Variant.” *Scientific Reports* 11 (1): 13705. https://doi.org/10.1038/s41598-021-91662-w.

Islam, M. Rafiul, M. Nazmul Hoque, M. Shaminur Rahman, A. S.M.Rubayet Ul Alam, Masuda Akther, J. Akter Puspo, Salma Akter, Munawar Sultana, Keith A. Crandall, and M. Anwar Hossain. 2020. “Genome-Wide Analysis of SARS-CoV-2 Virus Strains Circulating Worldwide Implicates Heterogeneity.” *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-70812-6.

Jackson, Ben, Maciej F. Boni, Matthew J. Bull, Amy Colleran, Rachel M. Colquhoun, Alistair C. Darby, Sam Haldenby, et al. 2021. “Generation and Transmission of Interlineage Recombinants in the SARS-CoV-2 Pandemic.” *Cell* 184 (20): 5179-5188.e8. https://doi.org/10.1016/j.cell.2021.08.014.

Ji, Wei, Wei Wang, Xiaofang Zhao, Junjie Zai, and Xingguang Li. 2020. “Cross-Species Transmission of the Newly Identified Coronavirus 2019-NCoV.” *Journal of Medical Virology* 92 (4): 433–40. https://doi.org/10.1002/jmv.25682.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Keith, Jonathan M. 2008. “Sequence Segmentation.” *Methods in Molecular Biology (Clifton, N.J.)* 452 (January): 207–29. https://doi.org/10.1007/978-1-60327-159-2_11.

Khare, Shruti, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, et al. 2021. “GISAID’s Role in Pandemic Response.” *China CDC Weekly* 3 (49): 1049–51.

https://doi.org/10.46234/ccdcw2021.255.

Koopmans, Marion, Peter Daszak, Vladimir G. Dedkov, Dominic E. Dwyer, Elmoubasher Farag, Thea K. Fischer, David T. S. Hayman, et al. 2021. “Origins of SARS-CoV-2: Window Is Closing for Key Scientific Studies.” *Nature* 596 (7873): 482–85. https://doi.org/10.1038/d41586-021-02263-6.

Ledford, Heidi. 2022. “The next Variant: Three Key Questions about What’s after Omicron.” *Nature* 603 (7900): 212–13. https://doi.org/10.1038/d41586-022-00510-y.

Lenth, Russell V. 2022. “Emmeans: Estimated Marginal Means, Aka Least-Squares Means. Https://Github.Com/Rvlenth/Emmeans.” https://github.com/rvlenth/emmeans.

Liu, Zhixin, Xiao Xiao, Xiuli Wei, Jian Li, Jing Yang, Huabing Tan, Jianyong Zhu, Qiwei Zhang, Jianguo Wu, and Long Liu. 2020. “Composition and Divergence of Coronavirus Spike Proteins and Host ACE2 Receptors Predict Potential Intermediate Hosts of SARS-CoV-2.” *Journal of Medical Virology* 92 (6): 595–601. https://doi.org/10.1002/jmv.25726.

Majumdar, Parinita, and Sougata Niyogi. 2021. “SARS-CoV-2 Mutations: The Biological Trackway towards Viral Fitness.” *Epidemiology and Infection*. Cambridge University Press. https://doi.org/10.1017/S0950268821001060.

McBroome, Jakob, Bryan Thornlow, Angie S Hinrichs, Alexander Kramer, Nicola De Maio, Nick Goldman, David Haussler, Russell Corbett-Detig, and Yatish Turakhia. 2021. “A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees.” Edited by Jian Lu. *Molecular Biology and Evolution* 38 (12): 5819–24. https://doi.org/10.1093/molbev/msab264.

McShea, Daniel W., and Robert N. Brandon. 2010. *Biology’s First Law□: The Tendency for Diversity and Complexity to Increase in Evolutionary Systems*. University of Chicago Press. 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.

Melchionna, M, A Mondanaro, C Serio, S Castiglione, M Di Febbraro, L Rook, J A F Diniz-Filho, et al. 2019. “Macroevolutionary Trends of Brain Mass in Primates.” *Biological Journal of the Linnean Society* 129 (1): 14–25. https://doi.org/10.1093/biolinnean/blz161.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams,

Arndt Von Haeseler, Robert Lanfear, and Emma Teeling. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Mlcochova, Petra, Steven Kemp, Mahesh Shanker Dhar, Guido Papa, Bo Meng, Isabella A.T.M. Ferreira, Rawlings Datir, et al. 2021. “SARS-CoV-2 B.1.617.2 Delta Variant Replication and Immune Evasion.” *Nature*, September, 1–8. https://doi.org/10.1038/s41586-021-03944-y.

Mooers, Aø, and Ec Holmes. 2000. “The Evolution of Base Composition and Phylogenetic Inference.” *Trends in Ecology & Evolution* 15 (9): 365–69. http://www.ncbi.nlm.nih.gov/pubmed/10931668.

Morel, Benoit, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas Hübner, Sarah Lutteropp, Dora Serdari, et al. 2020. “Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult.” *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msaa314.

Moya, Andrés, Edward C. Holmes, and Fernando González-Candelas. 2004. “The Population Genetics and Evolutionary Epidemiology of RNA Viruses.” *Nature Reviews Microbiology*. Nat Rev Microbiol. https://doi.org/10.1038/nrmicro863.

Moya, Andrés, José L. Oliver, Miguel Verdú, Luis Delaye, Vicente Arnau, Pedro Bernaola-Galván, Rebeca de la Fuente, et al. 2020. “Driven Progressive Evolution of Genome Sequence Complexity in Cyanobacteria.” *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-76014-4.

Naqvi, Ahmad Abu Turab, Kisa Fatima, Taj Mohammad, Urooj Fatima, Indrakant K. Singh, Archana Singh, Shaikh Muhammad Atif, Gururao Hariprasad, Gulam Mustafa Hasan, and Md Imtaiyaz Hassan. 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. https://doi.org/10.1016/j.bbadis.2020.165878.

Oliver, Jose L., Pedro Carpena, Michael Hackenberg, and Pedro Bernaola-Galván. 2004. “IsoFinder: Computational Prediction of Isochores in Genome Sequences.” *Nucleic Acids Res* 32 (Web Server issue): W287-92. https://doi.org/10.1093/nar/gkh399.

Oliver, Jose L., Ramon Román-Roldán, Javier Pérez, and Pedro Bernaola-Galván. 1999. “SEGMENT: Identifying Compositional Domains in DNA Sequences.” *Bioinformatics* 15 (12): 974–79. http://bioinformatics.oxfordjournals.org/content/15/12/974.short.

Pipes, Lenore, Hongru Wang, John P Huelsenbeck, and Rasmus Nielsen. 2021. “Assessing

Uncertainty in the Rooting of the SARS-CoV-2 Phylogeny." Edited by Harmit Malik. *Molecular Biology and Evolution* 38 (4): 1537–43. https://doi.org/10.1093/molbev/msaa316.

Pybus, Oliver G., Andrew J. Tatem, and Philippe Lemey. 2015. "Virus Evolution and Transmission in an Ever More Connected World." *Proceedings of the Royal Society B: Biological Sciences*. Royal Society of London. https://doi.org/10.1098/rspb.2014.2878.

Richard, Damien, Liam P Shaw, Rob Lanfear, Mislav Acman, Christopher J Owen, Cedric Cs Tan, Lucy Van Dorp, and François Balloux. 2021. "A Phylogeny-Based Metric for Estimating Changes in Transmissibility from Recurrent Mutations in SARS-CoV-2." https://doi.org/10.1101/2021.05.06.442903.

Rodríguez, Francisco, Jose L. Oliver, Antonio Marín, and Juan R. Medina. 1990. "The General Stochastic Model of Nucleotide Substitution." *Journal of Theoretical Biology*. https://doi.org/10.1016/S0022-5193(05)80104-3.

Román-Roldán, Ramon, Pedro Bernaola-Galván, and Jose L. Oliver. 1998. "Sequence Compositional Complexity of DNA through an Entropic Segmentation Method." *Physical Review Letters* 80 (6): 1344–47. http://link.aps.org/doi/10.1103/PhysRevLett.80.1344.

Serio, Carmela, Silvia Castiglione, Gianmarco Tesone, Martina Piccolo, Marina Melchionna, Alessandro Mondanaro, Mirko Di Febbraro, and Pasquale Raia. 2019. "Macroevolution of Toothed Whales Exceptional Relative Brain Size." *Evolutionary Biology* 46 (4): 332–42. https://doi.org/10.1007/s11692-019-09485-7.

Shu, Yuelong, and John McCauley. 2017. "GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality." *Eurosurveillance*. European Centre for Disease Prevention and Control. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

Straten, Karlijn van der, Denise Guerra, Marit J. van Gils, Ilja Bontjer, Tom G. Caniels, Hugo D.G. van Willigen, Elke Wynberg, et al. 2022. "Mapping the Antigenic Diversification of SARS-CoV-2." *MedRxiv*, January, 2022.01.03.21268582. https://doi.org/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): 57–86.

Tay, John H, Ashleigh F Porter, Wytamma Wirth, and Sebastian Duchene. 2022. "The Emergence of SARS-CoV-2 Variants of Concern Is Driven by Acceleration of the Substitution Rate."

Molecular Biology and Evolution 39 (2). https://doi.org/10.1093/molbev/msac013.

Thorne, Lucy G, Mehdi Bouhaddou, Ann-Kathrin Reuschl, Lorena Zuliani-Alvarez, Ben Polacco, Adrian Pelin, Jyoti Batra, et al. 2021. "Evolution of Enhanced Innate Immune Evasion by the SARS-CoV-2 B.1.1.7 UK Variant." https://doi.org/10.1101/2021.06.06.446826.

To, Thu Hien, Matthieu Jung, Samantha Lycett, and Olivier Gascuel. 2016. "Fast Dating Using Least-Squares Criteria and Algorithms." *Systematic Biology* 65 (1): 82–97. https://doi.org/10.1093/sysbio/syv068.

Venkatakrishnan, A J, Anand 2+, Patrick Lenehan, Pritha Ghosh, Rohit Suratekar, Abhishek Siroha, Dibyendu Roy Chowdhury, et al. 2021. "Antigenic Minimalism of SARS-CoV-2 Is Linked to Surges in COVID-19 Community Transmission and Vaccine Breakthrough Infections." https://doi.org/10.1101/2021.05.23.21257668.

Wen, Sheng-Yun, and Chun-Ting Zhang. 2003. "Identification of Isochore Boundaries in the Human Genome Using the Technique of Wavelet Multiresolution Analysis." *Biochemical and Biophysical Research Communications* 311 (1): 215–22. https://doi.org/10.1016/j.bbrc.2003.09.198.

Wertheim, Joel O, Mike Steel, and Michael J Sanderson. 2022. "Accuracy in Near-Perfect Virus Phylogenies." Edited by Vincent Savolainen. *Systematic Biology* 71 (2): 426–38. https://doi.org/10.1093/sysbio/syab069.

Worobey, Michael. 2021. "Dissecting the Early COVID-19 Cases in Wuhan." *Science* 374 (6572): 1202–4. https://doi.org/10.1126/science.abm4454.

Worobey, Michael, and Edward C. Holmes. 1999. "Evolutionary Aspects of Recombination in RNS Viruses." *Journal of General Virology* 80 (May): 2535–43. https://doi.org/10.1099/0022-1317-80-10-2535.

Young, Barnaby E, Siew-Wai Fong, Yi-Hao Chan, Tze-Minn Mak, Li Wei Ang, Danielle E Anderson, Cheryl Yi-Pin Lee, 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)* 396 (10251): 603–11. https://doi.org/10.1016/S0140-6736(20)31757-8.

Zhang, Tao, Qunfu Wu, and Zhigang Zhang. 2020. "Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak." *Current Biology* 30 (7): 1346-1351.e2. https://doi.org/10.1016/j.cub.2020.03.022.

Zhang, Yong Zhen, and Edward C. Holmes. 2020. “A Genomic Perspective on the Origin and Emergence of SARS-CoV-2.” *Cell* 181 (2): 223–27. https://doi.org/10.1016/j.cell.2020.03.035.

Zhou, Peng, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, et al. 2020. “Discovery of a Novel Coronavirus Associated with the Recent Pneumonia Outbreak in Humans and Its Potential Bat Origin.” *Nature*, January, 2020.01.22.914952. https://doi.org/10.1101/2020.01.22.914952.