The emergence of variants with increased fitness accelerates the slowdown of genome sequence heterogeneity 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,$,*]

[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)

Abstract

Since the outbreak of the COVID-19 pandemic, the SARS-CoV-2 coronavirus has accumulated an important amount of genetic and genomic variability through mutation and recombination events. To test evolutionary trends that could inform us on the adaptive process of the virus to its human host, we summarize all this sequence variability by computing the Sequence Compositional Complexity (*SCC*) in more than 23,000 high-quality coronavirus genome sequences from across the globe, covering the period spanning from the start of the pandemic in December 2019 to March 2022. In early samples, we found no statistical support for any trend in *SCC* values over time, although the virus as a whole appears to evolve faster than Brownian Motion expectation. However, in samples taken after the first Variant of Concern (VoC) with higher transmissibility (Alpha) emerges, and controlling for phylogenetic and sampling effects, we were able to detect a statistically significant trend for decreased *SCC* values over time. SARS-CoV-2 evolution towards lower values of genome heterogeneity is further intensified by the emergence of successive, widespread VoCs. Concomitantly to the temporal reduction in *SCC*, its absolute evolutionary rate kept increasing toward the present, meaning that the *SCC* decrease itself accelerated over time. As compared to Alpha or Delta variants, the currently dominant VoC, Omicron, shows much stronger trends in both *SCC* values and rates over time. These results indicate that the increases in fitness of variant genomes associated to a higher transmissibility leads to a reduction of their genome sequence heterogeneity, thus explaining the general slowdown of *SCC* along with the pandemic course.

Keywords: Coronavirus evolution, Variants of Concern (VoCs), evolutionary trends, genome sequence heterogeneity, evolutionary rate.

Introduction

Pioneer works showed that RNA viruses are excellent material for studies of evolutionary genomics (Domingo et al., 1999; Moya et al., 2004; Worobey and Holmes, 1999). Now, with the outbreak of the COVID-19 pandemic, this has become a key research topic. Despite 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 a zoonotic event (Holmes et al., 2021). Direct bat-to-human spillover events may occur more often than reported, although most remain unrecognized (Sánchez et al. 2021). Bats are known as the natural reservoirs of SARS-like CoVs (Li et al., 2005) 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 et al., 2020) would be at the origin of SARS-CoV-2. Bat RaTG13 virus best matched the overall codon usage pattern of SARS-CoV-2 in orf1ab, spike, and nucleocapsid genes, while the pangolin P1E virus had a more similar codon usage in the membrane gene (Gu et al., 2020). Other intermediate hosts have been identified, such as RaTG15, and its knowledge is essential to prevent 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 genetic and genomic variability (Elbe and Buckland-Merrett, 2017; Hadfield et al., 2018; Hamed et al., 2021; Hatcher et al., 2017; Li et al., 2020), 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; Patiño-Galindo et al., 2021). Noteworthy, RNA viruses can also accumulate high genetic variation during individual outbreaks (Pybus et al., 2015), showing mutation and evolutionary rates that may be 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 et al., 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; van Dorp et al., 2020; Wang et al., 2021; Zhou et al., 2020), as mutations provoking epitope loss and antibody escaping. These have been found mainly in evolved variants isolated from Europe and the Americas, which 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 course. Parallel mutations in multiple independent lineages and variants have been observed (van Dorp et al., 2020), which may indicate convergent evolution and that are of particular interest in the context of adaptation of SARS-CoV-2 to the human host (van Dorp et al., 2020). Other authors reported some sites under positive pressure in the nucleocapsid and spike genes (Benvenuto et al., 2020). All this research effort has allowed to track in real-time all these changes. 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, and can be 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 the nucleotides at the borders separating two domains, or by changing nucleotide frequencies in a given region, thus altering the number of domains or their compositional differences, consequently changing the resulting SCC value of the sequence (Bernaola-Galván et al., 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; Oliver et al., 2004, 2002; Román-Roldán et al., 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. (Bernaola-Galván et al., 2008; Bernardi, 2015; Bernardi et al., 1985; Oliver et al., 2004). 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 et al., 1998), an entropic measure of genome heterogeneity, meant as the number of domains and nucleotide differences among them, identified in a genome sequence through a proper segmentation algorithm (Bernaola-Galván et al., 1996). By using phylogenetic ridge regression, a method that has been able to reveal evolutionary trends in both macro- (Melchionna et al., 2019; Serio et al., 2019) and micro-organisms (Moya et al., 2020), we present here evidence for a long-term tendency of decreasing genome sequence heterogeneity in SARS-CoV-2. The trend is shared by the virus most important Variants of Concern (VoCs), Alpha or Delta, and greatly accelerated by the recent rise to dominance of Omicron (Du et al., 2022).

Results

Genome heterogeneity in the coronavirus

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

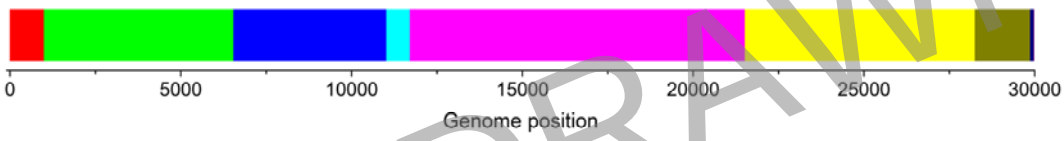

Figure 1. Compositional segmentation of the GISAID reference genome (hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30). Using an iterative segmentation algorithm (Bernaola-Galván et al., 1996; Oliver et al., 2004), the RNA sequence was divided into eight compositionally homogeneous segments (i.e., 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 present a lot of variation in each domain's number, length and nucleotide composition, which is reflected in higher or lower *SCC* values in individual genomes. The number of segments ranges between 4 and 9, while the *SCC* do it between 2.71E-03 and 6.8E-03 bits. The strain name, the collection date, and the *SCC* values for each analyzed genome are shown in Supplementary Tables S1-S18.

Temporal evolution of SCC over the coronavirus pandemic course

To characterize the temporal evolution of *SCC* over the entire range of the coronavirus pandemic (December 2019 to March 2022), we downloaded from GISAID/Audacity (GISAID, 2020) 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). We then filtered, masked and aligned these sequences to the reference genome (see Methods). For each of these samples, we determined the proportion of variants (Table 1, columns 5-8) and inferred an ML phylogenetic tree by means of IQ TREE 2 (Minh et al., 2020). Finally, we sought for temporal trends in *SCC* values and evolutionary rates by using the function *search.trend* in the R package RRphylo (Castiglione et al., 2018), contrasting the realized slope of *SCC* versus time regression to a family of 1000 slopes generated under the Brownian motion 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

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 takes place throughout the study period, indicating that the virus increased in variability early on, but took on a monotonic trend for decreasing *SCC* as VoCs appeared. These results were robust to several sources of uncertainty, including those related with the algorithms used for multiple alignment or to infer phylogenetic trees (see Supplementary Information). In summary, these analyses show that statistically significant trends for declining heterogeneity began in between the end of December 2020 (s1573) and March 2021 (s1871) in coincidence with the emergence of the first VoC (Alpha), a path continued over the successive emergence of other variants.

Table 1. Phylogenetic trends in coronavirus random samples downloaded from the *GISAID* database (Elbe and Buckland-Merrett, 2017; Koehorst et al., 2017; Shu and McCauley, 2017) 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, 1000, 2000 or 3000 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 by *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 Methods and Supplementary Information 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	1000	726	0.00	0.00	0.00	0.00	-97.10	0.250	88,684.70	0.002
s730		1000	730	0.00	0.00	0.00	0.00	-65.20	0.268	145,566.20	0.001
s781		2000	781	0.00	0.00	0.00	0.00	-9.67	0.426	125,140.00	0.001
s1170	jun-20	2000	1170	0.00	0.00	0.00	0.00	12.83	0.444	87,637.54	0.001
s1277	sep-20	2000	1277	0.00	0.00	0.00	0.00	-20.85	0.305	39,183.10	0.001
s1573	dec-20	2000	1573	4.32	0.00	0.00	4.83	-38.53	0.066	26,502.59	0.001
s1871	mar-21	2000	1871	50.03	0.00	0.00	57.03	-61.94	0.001	14,254.56	0.001
s498		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	may-21	1000	987	57.85	0.20	0.00	67.17	-60.29	0.001	16,937.26	0.001
s980		1000	980	65.31	0.82	0.00	75.41	-54.23	0.004	16,169.35	0.001
s1939		2000	1939	63.02	1.24	0.00	72.93	-41.74	0.010	13,044.29	0.001
s1974	jun-21	2000	1974	45.90	7.14	0.00	60.59	-34.83	0.016	18,624.45	0.001
s1985	sep-21	2000	1985	27.10	44.08	0.00	78.34	-19.30	0.131	11,688.97	0.001
s1994	dec-21	2000	1994	17.95	57.82	6.22	86.70	-20.93	0.060	7,495.37	0.001
s2347		3000	2347	18.41	46.66	0.00	73.11	-33.38	0.007	7,217.62	0.001
s1990	mar-22	2000	1990	14.32	51.41	18.29	87.28	-21.89	0.037	4,896.06	0.052
	TOTAL:	28000	23318								

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

SCC trends of variants

We estimated the relative contribution of the three most important 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 the sample s1990, which includes a sizeable number of Omicron genomes, are shown in Figure 2. On 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). In agreement with the previous (seventeen consecutive bins, see Table 1) analysis, 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 4 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 *SCC* decrease over time (Table 2). This ~50-fold increase in the likelihood to find a consistent trend for *SCC* decline over time is shared unambiguously by all tested variants (Alpha, Delta, and Omicron) (Table 3). Yet, Omicron shows significantly stronger decline in *SCC* than the other variants (Table 3) suggesting that the trends initiated with the appearance of main variants became more intense with the emergence of Omicron by the end of 2021.

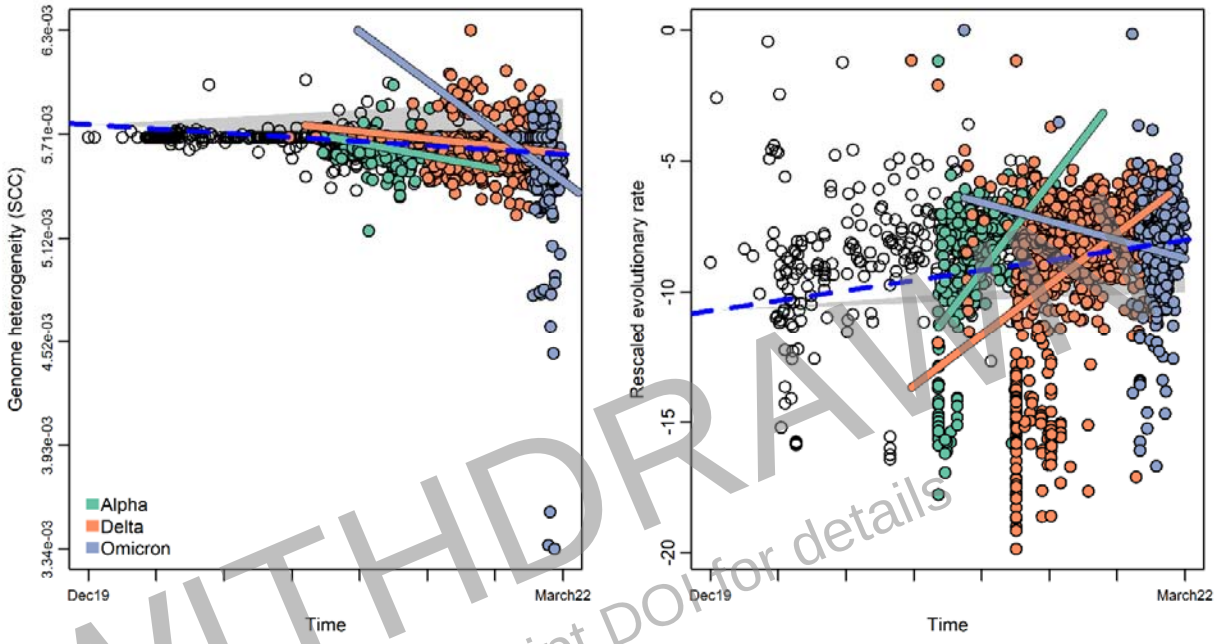

Figure 2. Phylogenetic ridge regressions for *SCC* (left) and its evolutionary rate (right) as detected by the RRphylo R package (Castiglione et al., 2018) on the s1990 sample. For *SCC*, the estimated value for each tip in the phylogenetic tree is regressed (blue line) against its age (the phylogenetic 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 Brownian motion 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 Brownian motion 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 by using the function *emtrends* available within the R package *emmeans* (Lenth, 2022). We found the slope for the group including all variant 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 Brownian Motion 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 rates (absolute magnitude of the rate) showed a tendency to increase through time (Table 2). The slope of *SCC* rates through time regression for Omicron was always significantly lower than the slope computed for rest of the tree (Table 3). This is 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 short time-lapse analyzed (28 months), the coronavirus RNA genomes can be divided 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 observed 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 Brownian Motion 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 decreasing trends in *SCC* (Table 1). Concomitantly to the temporal reduction in *SCC*, its absolute evolutionary rate kept increasing toward the present, meaning that the *SCC* decrease itself accelerated over time. In agreement with this notion, although the *SCC* decrease is an evolutionary path shared by variants, the nearly threefold increase in rates becomes more intense after the appearance of the most recent VoC (Omicron) on later 2021, which shows much faster decrease 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. 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 a higher transmissibility (a paradigmatic example is the recent outbreak of the Omicron variant). The sudden increases in fitness of variant genomes, mainly due to the gathering of co-mutations, which become prevalent world-wide compared to single mutations, are 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 to a higher transmissibility leads to a reduction of their genome sequence heterogeneity, thus explaining the general slowdown of *SCC* along 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 (Callaway, 2022; Ledford, 2022; Straten et al., 2022) by means of the tools used here will allow elucidating whether and in what extension the evolution of genome sequence heterogeneity in the virus impacts human health.

Methods

Data retrieving, filtering, masking and alignment

We retrieved random samples (see Table 1) of high-quality coronavirus genome sequences from the GISAID/Audacity database (Elbe and Buckland-Merrett, 2017; Koehorst et al., 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 (MN908947.3), then filtering and masking the alignments to avoid sequence oddities (Hodcroft et al., 2021).

Phylogenetic trees

The best ML timetree for each random sample in Table 1 was inferred by means of IQ-TREE 2 (Minh et al., 2020), using the GTR nucleotide substitution model (Rodríguez et al., 1990; Tavaré, 1986) 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 et al., 2008, 1996; Oliver et al., 2004, 1999). We choose 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 statistically significant segments. The process continues iteratively over the new resulting segments while sufficient significance continues appearing.

Computing the Sequence compositional complexity (SCC)

Once each coronavirus genome sequence is 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 et al., 1998). *SCC* increases/decreases with both the number of segments and the degree of compositional differences among them. In this way, *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 sensible to any change in the RNA genome sequence, either nucleotide substitutions, indels, genome rearrangements or recombination events.

Phylogenetic ridge regression

To search for trends in *SCC* values and evolutionary rates over time, phylogenetic ridge regression was applied by 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 is 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 Brownian Motion (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 test explicitly the effect of variants and to compare variants among each other we selected 4 different trees and SCC data (s730, a727, s1871, s1990) from the entire dataset (Table 1). On 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*. 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 the presence of temporal trends in *SCC* and *SCC* 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 because of different tree sizes.

Additional details regarding the methods used in this study are provided in the Supplementary Information.

Data availability

All data generated or analyzed during this study are included in this published article (and its Supplementary Information files).

References

Bahir I, Fromer M, Prat Y, Linial M. 2009. Viral adaptation to host: A proteome-based analysis of codon usage and amino acid preferences. *Mol Syst Biol* **5**:311. doi:10.1038/msb.2009.71

Banerjee A, Sarkar R, Mitra S, Lo M, Dutta S, Chawla-Sarkar M. 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**:e20735. doi:10.2196/20735

Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. 2020. The 2019-new coronavirus epidemic: Evidence for virus evolution. *J Med Virol* **92**:455–459. doi:10.1002/jmv.25688

Bernaola-Galván P, Carpena P, Oliver JL. 2008. A standalone version of IsoFinder for the computational prediction of isochores in genome sequences. *arXiv Prepr arXiv08061292* 1–7.

Bernaola-Galván P, Oliver JL, Carpena P, Clay O, Bernardi G. 2004. Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes. *Gene* **333**:121–133. doi:10.1016/j.gene.2004.02.042

Bernaola-Galván P, Román-Roldán R, Oliver JL. 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys Rev E* **53**:5181–5189. doi:10.1103/PhysRevE.53.5181

Bernardi G. 2015. Chromosome architecture and genome organization. *PLoS One* **10**:e0143739. doi:10.1371/journal.pone.0143739

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. *Science (80-)* **228**:953–958. doi:10.1126/science.4001930

Cai HY, Cai KK, Li J. 2020. Identification of Novel Missense Mutations in a Large Number of Recent SARS-CoV-2 Genome Sequences. *J Gen Med Res* **2**. doi:10.20944/preprints202004.0482.v1

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

Castiglione S, Serio C, Mondanaro A, Di Febbraro M, Profico A, Girardi G, Raia P. 2019. Simultaneous detection of macroevolutionary patterns in phenotypic means and rate of change with and within phylogenetic trees including extinct species. *PLoS One* **14**.

doi:10.1371/journal.pone.0210101

Castiglione S, Tesone G, Piccolo M, Melchionna M, Mondanaro A, Serio C, Di Febbraro M, Raia P. 2018. A new method for testing evolutionary rate variation and shifts in phenotypic evolution. *Methods Ecol Evol* **9**:974–983. doi:10.1111/2041-210X.12954

Cyranoski D. 2020. Profile of a killer: the complex biology powering the coronavirus pandemic. *Nature*. doi:10.1038/d41586-020-01315-7

Domingo E, Webster RG, Holland JJ. 1999. Origin and evolution of viruses. Academic Press.

Du P, Gao GF, Wang Q. 2022. The mysterious origins of the Omicron variant of SARS-CoV-2. *Innov* **3**:100206. doi:10.1016/J.XINN.2022.100206

Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Challenges* **1**:33–46. doi:10.1002/gch2.1018

Fearnhead P, Vasilieou D. 2009. Bayesian Analysis of Isochores. *J Am Stat Assoc*.

GISAID. 2020. GISAID Initiative. *Adv Virus Res* **2008**:1–7.

Gu H, Chu DKW, Peiris M, Poon LLM. 2020. Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. *Virus Evol* **6**. doi:10.1093/ve/veaa032

Gupta AM, 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, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**:4121–4123. doi:10.1093/bioinformatics/bty407

Hamed SM, Elkhatib WF, Khairalla AS, Noreddin AM. 2021. Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. *Sci Rep* **11**:8435. doi:10.1038/s41598-021-87713-x

Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, Schaffer AA, Rodney Brister J. 2017. Virus Variation Resource-improved response to emergent viral outbreaks. *Nucleic Acids Res* **45**:D482–D490. doi:10.1093/nar/gkw1065

Hodcroft EB, Domman DB, Snyder DJ, Oguntuyo K, Van Diest M, Densmore KH, Schwalm KC, Femling J, Carroll JL, Scott RS, Whyte MM, Edwards MD, Hull NC, Kevil CG, Vanchiere JA,

Lee B, Dinwiddie DL, Cooper VS, Kamil JP. 2021. Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting amino acid position 677. *medRxiv Prepr Serv Heal Sci*. doi:10.1101/2021.02.12.21251658

Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, Anthony SJ, Barclay WS, Boni MF, Doherty PC, Farrar J, Geoghegan JL, Jiang X, Leibowitz JL, Neil SJD, Skern T, Weiss SR, Worobey M, Andersen KG, Garry RF, Rambaut A. 2021. The Origins of SARS-CoV-2: A Critical Review. *Cell*. doi:10.1016/j.cell.2021.08.017

Hon C-C, Lam T-Y, Shi Z-L, Drummond AJ, Yip C-W, Zeng F, Lam P-Y, Leung FC-C. 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. *J Virol* **82**:1819–1826. doi:10.1128/jvi.01926-07

Ilmjärv S, Abdul F, Acosta-Gutiérrez S, Estarellas C, Galdadas I, Casimir M, Alessandrini M, Gervasio FL, Krause KH. 2021. Concurrent mutations in RNA-dependent RNA polymerase and spike protein emerged as the epidemiologically most successful SARS-CoV-2 variant. *Sci Rep* **11**:13705. doi:10.1038/s41598-021-91662-w

Islam MR, Hoque MN, Rahman MS, Alam ASMRU, Akther M, Puspo JA, Akter S, Sultana M, Crandall KA, Hossain MA. 2020. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. *Sci Rep* **10**. doi:10.1038/s41598-020-70812-6

Jackson B, Boni MF, Bull MJ, Colleran A, Colquhoun RM, Darby AC, Haldenby S, Hill V, Lucaci A, McCrone JT, Nicholls SM, O'Toole Á, Pacchiarini N, Poplawski R, Scher E, Todd F, Webster HJ, Whitehead M, Wierzbicki C, Loman NJ, Connor TR, Robertson DL, Pybus OG, Rambaut A. 2021. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell* **184**:5179-5188.e8. doi:10.1016/j.cell.2021.08.014

Ji W, Wang W, Zhao X, Zai J, Li X. 2020. Cross-species transmission of the newly identified coronavirus 2019-nCoV. *J Med Virol* **92**:433–440. doi:10.1002/jmv.25682

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

Keith JM. 2008. Sequence segmentation. *Methods Mol Biol* **452**:207–29. doi:10.1007/978-1-60327-159-2_11

Koehorst J, van Dam J, Saccenti E, Martins-Santos V, Suarez-Diez M, Schaap P. 2017. GISAID

Global Initiative on Sharing All Influenza Data. Phylogeny of SARS-like betacoronaviruses including novel coronavirus (nCoV). *Oxford* **34**:1401–1403. doi:10.1093/BIOINFORMATICS

Koopmans M, Daszak P, Dedkov VG, Dwyer DE, Farag E, Fischer TK, Hayman DTS, Leendertz F, Maeda K, Nguyen-Viet H, Watson J. 2021. Origins of SARS-CoV-2: window is closing for key scientific studies. *Nature* **596**: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**: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.

Li W, Shi Z, Yu M, Ren W, Smith C, Epstein J, Wang H, Crameri G, Hu Z, Zhang H, Zhang J, McEachern J, Field H, Daszak P, Eaton B, Zhang S, Wang L. 2005. Bats are natural reservoirs of SARS-like coronaviruses. *Science (80-)* **310**:676–679. doi:10.1126/science.1118391

Li X, Giorg EE, Marichannegowda MH, Foley B, Xiao C, Kong XP, Chen Y, Gnanakaran S, Korber B, Gao F. 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv* **6**. doi:10.1126/sciadv.abb9153

Liu Z, Xiao X, Wei X, Li J, Yang J, Tan H, Zhu J, Zhang Q, Wu J, Liu L. 2020. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. *J Med Virol* **92**:595–601. doi:10.1002/jmv.25726

Majumdar P, Niyogi S. 2021. SARS-CoV-2 mutations: The biological trackway towards viral fitness. *Epidemiol Infect*. doi:10.1017/S0950268821001060

McShea DW, Brandon RN. 2010. Biology's first law□: the tendency for diversity and complexity to increase in evolutionary systems. University of Chicago Press.

Melchionna M, Mondanaro A, Serio C, Castiglione S, Di Febbraro M, Rook L, Diniz-Filho JAF, Manzi G, Profico A, Sansalone G, Raia P. 2019. Macroevolutionary trends of brain mass in Primates. *Biol J Linn Soc* **129**:14–25. doi:10.1093/biolinnean/blz161

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R, Teeling E. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**:1530–1534. doi:10.1093/molbev/msaa015

Mlcochova P, Kemp S, Dhar Mahesh Shanker, Papa G, Meng B, Ferreira IATM, Datir R, Collier DA,

Albecka A, Singh S, Pandey R, Brown J, Zhou J, Goonawardane N, Mishra S, Whittaker C, Mellan T, Marwal R, Datta M, Sengupta S, Ponnusamy K, Radhakrishnan VS, Abdullahi A, Charles O, Chattopadhyay P, Devi P, Caputo D, Peacock T, Wattal DC, Goel N, Satwik A, Vaishya R, Agarwal M, Chauhan H, Dikid T, Gogia H, Lall H, Verma K, Dhar Mahesh S., Singh MK, Soni N, Meena N, Madan P, Singh P, Sharma Ramesh, Sharma Rajeev, Kabra S, Kumar S, Kumari S, Sharma U, Chaudhary U, Sivasubbu S, Scaria V, Wattal C, Oberoi JK, Raveendran R, Datta S, Das S, Maitra A, Chinnaswamy S, Biswas NK, Parida A, Raghav SK, Prasad P, Sarin A, Mayor S, Ramakrishnan U, Palakodeti D, Seshasayee ASN, Thangaraj K, Bashyam MD, Dalal A, Bhat M, Shouche Y, Pillai A, Abraham P, Atul PV, Cherian SS, Desai AS, Pattabiraman C, Manjunatha M V., Mani RS, Udupi GA, Nandicoori V, Bharadwaj K, Tallapaka, Sowpati DT, Kawabata R, Morizako N, Sadamasu K, Asakura H, Nagashima M, Yoshimura K, Ito J, Kimura I, Uriu K, Kosugi Y, Suganami M, Oide A, Yokoyama M, Chiba M, Saito A, Butlertanaka EP, Tanaka YL, Ikeda T, Motozono C, Nasser H, Shimizu R, Yuan Y, Kitazato K, Hasebe H, Nakagawa S, Wu J, Takahashi M, Fukuhara T, Shimizu K, Tsushima K, Kubo H, Shirakawa K, Kazuma Y, Nomura R, Horisawa Y, Takaori-Kondo A, Tokunaga K, Ozono S, Baker S, Dougan G, Hess C, Kingston N, Lehner PJ, Lyons PA, Matheson NJ, Owehand WH, Saunders C, Summers C, Thaventhiran JED, Toshner M, Weekes MP, Maxwell P, Shaw A, Bucke A, Calder J, Canna L, Domingo J, Elmer A, Fuller S, Harris J, Hewitt S, Kennet J, Jose S, Kourampa J, Meadows A, O'Brien C, Price J, Publico C, Rastall R, Ribeiro C, Rowlands J, Ruffolo V, Tordesillas H, Bullman B, Dunmore BJ, Fawke S, Gräf S, Hodgson J, Huang C, Hunter K, Jones E, Legchenko E, Matara C, Martin J, Mescia F, O'Donnell C, Pointon L, Pond N, Shih J, Sutcliffe R, Tilly T, Treacy C, Tong Z, Wood J, Wylot M, Bergamaschi L, Betancourt A, Bower G, Cossetti C, De Sa A, Epping M, Fawke S, Gleadall N, Grenfell R, Hinch A, Huhn O, Jackson S, Jarvis I, Krishna B, Lewis D, Marsden J, Nice F, Okecha G, Omarjee O, Perera M, Potts M, Richoz N, Romashova V, Yarkoni NS, Sharma Rahul, Stefanucci L, Stephens J, Strezlecki M, Turner L, De Bie EMDD, Bunclark K, Josipovic M, Mackay M, Michael A, Rossi S, Selvan M, Spencer S, Yong C, Allison J, Butcher H, Caputo D, Clapham-Riley D, Dewhurst E, Furlong A, Graves B, Gray J, Ivers T, Kasanicki M, Le Gresley E, Linger R, Meloy S, Muldoon F, Ovington N, Papadia S, Phelan I, Stark H, Stirrups KE, Townsend P, Walker N, Webster J, Scholtes I, Hein S, King R, Mavousian A, Lee JH, Bassi J, Silacci-Fegni C, Saliba C, Pinto D, Irie T, Yoshida I, Hamilton WL, Sato K, Bhatt S, Flaxman S, James LC, Corti D, Piccoli L, Barclay WS, Rakshit P, Agrawal A, Gupta RK. 2021. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* 1–8. doi:10.1038/s41586-021-03944-y

Mooers A, Holmes E. 2000. The evolution of base composition and phylogenetic inference. *Trends Ecol Evol* **15**:365–369.

Moya A, Holmes EC, González-Candelas F. 2004. The population genetics and evolutionary epidemiology of RNA viruses. *Nat Rev Microbiol*. doi:10.1038/nrmicro863

Moya A, Oliver JL, Verdú M, Delaye L, Arnau V, Bernaola-Galván P, de la Fuente R, Díaz W, Gómez-Martín C, González FM, Latorre A, Lebrón R, Román-Roldán R. 2020. Driven progressive evolution of genome sequence complexity in Cyanobacteria. *Sci Rep* **10**. doi:10.1038/s41598-020-76014-4

Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, Atif SM, Hariprasad G, Hasan GM, Hassan MI. 2020. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochim Biophys Acta - Mol Basis Dis*. doi:10.1016/j.bbadis.2020.165878

Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P. 2004. IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res* **32**:W287-92. doi:10.1093/nar/gkh399

Oliver JL, Carpena P, Román-Roldán R, Mata-Balaguer T, Mejías-Romero A, Hackenberg M, Bernaola-Galván P. 2002. Isochore chromosome maps of the human genome. *Gene* **300**:117–127. doi:10.1016/S0378-1119(02)01034-X

Oliver JL, Román-Roldán R, Pérez J, Bernaola-Galván P. 1999. SEGMENT: identifying compositional domains in DNA sequences. *Bioinformatics* **15**:974–9.

Patiño-Galindo JÁ, Filip I, Chowdhury R, Maranas CD, Sorger PK, AlQuraishi M, Rabadan R. 2021. Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2. *Genome Med* **13**:124. doi:10.1186/s13073-021-00943-6

Pybus OG, Tatem AJ, Lemey P. 2015. Virus evolution and transmission in an ever more connected world. *Proc R Soc B Biol Sci*. doi:10.1098/rspb.2014.2878

Richard D, Shaw LP, Lanfear R, Acman M, Owen CJ, Tan CC, Van Dorp L, Balloux F. 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, Oliver JL, Marín A, Medina JR. 1990. The general stochastic model of nucleotide substitution. *J Theor Biol*. doi:10.1016/S0022-5193(05)80104-3

Román-Roldán R, Bernaola-Galván P, Oliver JL. 1998. Sequence compositional complexity of DNA through an entropic segmentation method. *Phys Rev Lett* **80**:1344–1347.

Serio C, Castiglione S, Tesone G, Piccolo M, Melchionna M, Mondanaro A, Di Febbraro M, Raia P. 2019. Macroevolution of Toothed Whales Exceptional Relative Brain Size. *Evol Biol* **46**:332–342. doi:10.1007/s11692-019-09485-7

Shu Y, McCauley J. 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*. doi:10.2807/1560-7917.ES.2017.22.13.30494

Straten K van der, Guerra D, Gils MJ van, Bontjer I, Caniels TG, Willigen HDG van, Wynberg E, Poniman M, Burger JA, Bouhuijs JH, Rijswijk J van, Lavell AHA, Appelman B, Sikkens JJ, Bomers MK, Han AX, Nichols BE, Prins M, Vennema H, Reusken C, Jong MD de, Bree GJ de, Russell CA, Eggink D, Sanders RW. 2022. Mapping the antigenic diversification of SARS-CoV-2. *medRxiv* 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. *Lect Math life Sci* **17**:57–86.

Thorne LG, Bouhaddou M, Reuschl A-K, Zuliani-Alvarez L, Polacco B, Pelin A, Batra J, Whelan MV, Ummadi M, Rojc A, Turner J, Obernier K, Braberg H, Soucheray M, Richards A, Chen K-H, Harjai B, Memon D, Hosmillo M, Hiatt J, Jahun A, Goodfellow IG, Fabius JM, Shokat K, Jura N, Verba K, Noursadeghi M, Beltrao P, Swaney DL, Garcia-Sastre A, Jolly C, Towers GJ, Krogan NJ. 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 TH, Jung M, Lycett S, Gascuel O. 2016. Fast Dating Using Least-Squares Criteria and Algorithms. *Syst Biol* **65**:82–97. doi:10.1093/sysbio/syv068

van Dorp L, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, Owen CJ, Pang J, Tan CCS, Boshier FAT, Ortiz AT, Balloux F. 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect Genet Evol* 104351. doi:10.1016/j.meegid.2020.104351

Venkatakrishnan AJ, 2+ A, Lenehan P, Ghosh P, Suratekar R, Siroha A, Chowdhury DR, O'horo JC, Yao JD, Pritt BS, Norgan A, Hurt RT, Badley AD, Halamka JD, Soundararajan V. 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

Wang R, Chen J, Wei G-W. 2021. Mechanisms of SARS-CoV-2 Evolution Revealing Vaccine-

Resistant Mutations in Europe and America. *J Phys Chem Lett* **12**:11850–11857. doi:10.1021/acs.jpclett.1c03380

Wen S-Y, Zhang C-T. 2003. Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. *Biochem Biophys Res Commun* **311**:215–222. doi:10.1016/j.bbrc.2003.09.198

Worobey M. 2021. Dissecting the early COVID-19 cases in Wuhan. *Science (80-)* **374**:1202–1204. doi:10.1126/science.abm4454

Worobey M, Holmes EC. 1999. Evolutionary aspects of recombination in RNS viruses. *J Gen Virol* **80**:2535–2543. doi:10.1099/0022-1317-80-10-2535

Young BE, Fong S-W, Chan Y-H, Mak T-M, Ang LW, Anderson DE, Lee CY-P, Amrun SN, Lee B, Goh YS, Su YCF, Wei WE, Kalimuddin S, Chai LYA, Pada S, Tan SY, Sun L, Parthasarathy P, Chen YYC, Barkham T, Lin RTP, Maurer-Stroh S, Leo Y-S, Wang L-F, Renia L, Lee VJ, Smith GJD, Lye DC, Ng LFP. 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**:603–611. doi:10.1016/S0140-6736(20)31757-8

Zhang T, Wu Q, Zhang Z. 2020. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr Biol* **30**:1346-1351.e2. doi:10.1016/j.cub.2020.03.022

Zhang YZ, Holmes EC. 2020. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* **181**:223–227. doi:10.1016/j.cell.2020.03.035

Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. 2020. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. *Nature* 2020.01.22.914952. doi:10.1101/2020.01.22.914952

Acknowledgements

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. We also gratefully acknowledge both the originating and submitting laboratories for the sequence data in GISAID EpiCoV on which these analyses are based. Supplementary Table S19 shows a complete list acknowledging all originating and submitting laboratories. In the same way, we gratefully acknowledge the authors, originating and submitting laboratories of the genetic sequences we used for the analysis of the Nextstrain sample; a complete list is shown in Supplementary Table S20.

Author information

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

Affiliations

Department of Genetics, Faculty of Sciences, University of Granada, 18071, Granada, Spain

José L. Oliver, Francisco Perfectti & Cristina Gómez-Martín

Laboratory of Bioinformatics, Institute of Biotechnology, Center of Biomedical Research, 18100, Granada, Spain

José L. Oliver & Cristina Gómez-Martín

Research Unit Modeling Nature, Universidad de Granada, 18071 Granada, SPAIN

Francisco Perfectti

Department of Applied Physics II and Institute Carlos I for Theoretical and Computational Physics, University of Málaga, 29071, Málaga, Spain

Pedro Bernaola-Galván

Amsterdam UMC, Vrije Universiteit Amsterdam, Department of Pathology, Cancer Center Amsterdam, Amsterdam, Netherlands

Cristina Gómez-Martín

Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse, Università di Napoli Federico II, 80126, Napoli, Italy

Pasquale Raia & Silvia Castiglione

Centro de Investigaciones sobre Desertificación, Consejo Superior de Investigaciones Científicas (CSIC), University of València and Generalitat Valenciana, 46113, Valencia, Spain

Miguel Verdú

Institute of Integrative Systems Biology (I2Sysbio), University of València and Consejo Superior de Investigaciones Científicas (CSIC), 46980, Valencia, Spain

Andrés Moya

Foundation for the Promotion of Sanitary and Biomedical Research of Valencian Community (FISABIO), 46020, Valencia, Spain

Andrés Moya

CIBER in Epidemiology and Public Health, 28029, Madrid, Spain

Andrés Moya

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.

Corresponding authors

Correspondence to José L. Oliver, Miguel Verdú and Andrés Moya.

Competing interests

The authors declare no competing interests.