Accelerated decline of genome heterogeneity in the SARS-CoV-2 coronavirus

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

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

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

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

[9]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

In the brief time since the outbreak of the COVID-19 pandemic, and despite its proofreading mechanism, the SARS-CoV-2 coronavirus has accumulated a significant amount of genetic variability through recombination and mutation events. To test evolutionary trends that could inform us on the adaptive process of the virus to its human host, we summarize all this variability in the Sequence Compositional Complexity (*SCC*), a measure of genome heterogeneity that captures the mutational and recombinational changes accumulated by a nucleotide sequence along time. Despite the brief time elapsed, we detected many differences in the number and length of compositional domains, as well as in their nucleotide frequencies, in more than 12,000 high-quality coronavirus genomes from across the globe. These differences in *SCC* are phylogenetically structured, as revealed by significant phylogenetic signal. Phylogenetic ridge regression shows that *SCC* followed a generalized decreasing trend along the ongoing process of pathogen evolution. In contrast, *SCC* evolutionary rate increased with time, showing that it accelerates toward the present. In addition, a low rate set of genomes was detected in all the genome groups, suggesting the existence of a stepwise distribution of rates, a strong indication of selection in favor of different dominant strains. Coronavirus variants reveal an exacerbation of this trend: non-significant *SCC* regression, low phylogenetic signal and, concomitantly, a threefold increase in the evolutionary rate. Altogether, these results show an accelerated decline of genome heterogeneity along with the SARS-CoV-2 pandemic expansion, a process that might be related to viral adaptation to the human host, perhaps paralleling the transformation of the current pandemic to epidemic.

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

Introduction

Pioneer works[1,2] showed that RNA viruses are an excellent material for studies of evolutionary genomics. Now, with the outbreak of the COVID-19 pandemic, this has become a key research topic. Despite its proofreading mechanism and the brief time-lapse, SARS-CoV-2 shows an important amount of genetic variability[3–5], which is due to both its recombinational origin[6] as well as mutation and additional recombination events accumulated along with the expansion of COVID-19 pandemic across the globe[7].

An unprecedented research effort has allowed to track in real-time all these changes along with pathogen evolution. Since the COVID-19 pandemic was declared by the World Health Organization (WHO) to be a public health emergency of international concern on March 2020 (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports), a massive amount of shared multidisciplinary information has been made available by the scientific community. Genome information about the coronavirus is available on websites as GISAID[4], Nextstrain[5], or NCBI virus[3]. The CoVizu[e] project (https://www.epicov.org/epi3/frontend#28b5af) supplied a near real-time visualization of SARS-CoV-2 global diversity of SARS-CoV-2 genomes.

To date, the most parsimonious explanation for the origin of SARS-CoV-2 is a zoonotic event[8]. Direct bat-to-human spillover events may occur more often than reported, although most remain unrecognized because of different causes[9]. Bats are known as the natural reservoirs of SARS-like CoVs[10]. Because of a comparison between these coronaviruses and SARS-CoV-2, a bat derivation for the outbreak was proposed[11]. Indeed, a recombination event between the bat coronavirus and either an origin-unknown coronavirus[12] or a pangolin virus[13,14] would be at the origin of SARS-CoV-2. Gu and co-workers[14] found that 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 membrane gene. Other intermediate hosts have been identified, such as RaTG15[15], knowledge of which is imperative to prevent further spread of the epidemic[16].

RNA viruses can accumulate high genetic variation during an individual outbreak[17], showing mutation and evolution rates that may be up to a million times higher than those of their hosts[18]. In the brief time since the COVID-19 pandemic appeared, and despite viral genomic proofreading mechanism, recombinations have accumulated[19] over multiple rounds of mutations, many of which have increased viral fitness[8,20–22]. Synonymous and non-synonymous mutations[23–25], as well as mismatches and deletions in translated and untranslated regions[18] have been tracked. Of particular interest are those non-synonymous mutations provoking epitope loss and antibody escaping found mainly in evolved variants isolated from Europe and the Americas, which have critical implications for SARS- CoV-2 transmission, pathogenesis, and immune interventions[26]. 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[27]. Parallel mutations in multiple independent lineages and variants have been observed[21,27], which may indicate convergent evolution and that are of particular interest in the context of adaptation of SARS-CoV-2 to the human host[21]. Other authors reported some sites under positive pressure in the nucleocapsid and spike genes[28]. Finally, genome rearrangements, as nucleotide deletions of different lengths, have been found, the major one affecting 382 nucleotides has been associated with a milder infection[29].

Most sequence changes (i.e., synonymous and non-synonymous nucleotide substitutions, insertions, deletions, recombination events, chromosome rearrangements, genome reorganizations…) can potentially alter the array of compositional domains in a genome. These domains can be changed either by altering nucleotide frequencies in a given region or by changing the nucleotides at the borders separating two putative domains, thus enlarging, or shortening a given domain[30–34]. A good metric of genome heterogeneity should be able to summarize the mutational and recombinational events accumulated by a genome sequence over time[35–39].

In many organisms, the patchy sequence structure formed by the array of domains with different nucleotide composition has been related to important biological features, i.e., gene and repeat densities, the timing of gene expression, recombination frequency, etc.[37,40–42]. Therefore, changes in genome heterogeneity may be relevant on evolutionary and epidemiological grounds. Specifically, evolutionary trends on genome heterogeneity could reveal adaptative processes of the virus to the human host.

To this end, we computed the Sequence Compositional Complexity *SCC*[39], 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[30]. We present evidence of a considerable amount of phylogenetically structured compositional heterogeneity in SARS-CoV-2 genomes, showing an evolutionary trend along with pandemic expansion. Phylogenetic ridge regressions of *SCC* and evolutionary rates against time (i.e., virus collection date) in genome samples of the general virus population and viral variants reveal trends toward the loss of genome compositional heterogeneity, while the evolutionary rate accelerates as the pathogen evolves in the human host.

Results

Genome heterogeneity in the coronavirus

The first SARS-CoV-2 coronavirus genome obtained from the start of the pandemics (2019-12-30) was divided into eight compositional domains by the iterative segmentation algorithm[30,31,37,42], resulting in a *SCC* value of 5.7 x 10^{-3} bits (Figure 1). Since that time, descendent coronaviruses present a lot of variation in each domain's number, length, nucleotide composition, and so in *SCC* genome heterogeneity values (Supplementary Tables S1-S16).

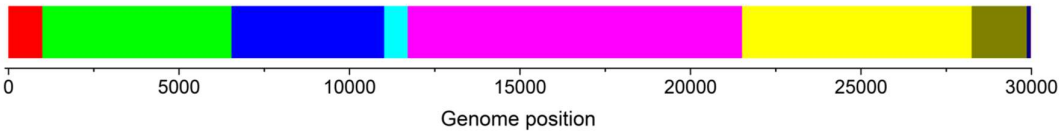

Figure 1. Compositional segmentation of the GISAID reference genome (hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30). Using an iterative segmentation algorithm[30,31,37,42], the RNA sequence was divided into eight compositionally homogeneous segments (or compositional domains) with *P* value ≤ 0.05. The genome position of domain borders is shown on the horizontal scale. Colors illustrate the different nucleotide compositions at each domain.

We analyzed more than 12,000 high-quality, complete coronavirus sequences of different genome groups: general virus population (which includes genomes from different waves), Variants of Concern (VoCs) and Variants of Interest (VoIs). The number of segments ranged between 5 and 10 while the *SCC* did it between 2.60^{-03} and 6.31^{-03} bits in the different genome groups (Supplementary Table S1). The strain name, the collection date, the *SCC* values, and the number of segments for each analyzed genome are shown in detail in Supplementary Tables S2-S16.

Phylogenetic signal of SCC

First, an ML phylogenetic tree was inferred for each sample, then computing the phylogenetic signal[43] for *SCC* (Table 1). Interestingly, the phylogenetic signal values (*K*) were clearly lower in the samples of variant genomes, becoming even non-significant in most of them, as compared to the samples from the general virus population.

Table 1. Phylogenetic signal[43] (*K*) obtained in distinct groups of coronavirus genome sequences. The *phylosignal* R package[44] was used.

Genome group	Sample	N	*K*	P value
General virus population	s300_1	297	2.59E-01	0.121
	s300_2	299	6.48E-01	0.012
	s500_1	498	6.45E-01	0.014
	s500_2	496	5.48E-01	0.020
	s1000_1	987	6.11E-01	0.004
	s1000_2	980	5.91E-01	0.009
Variants of Concern (VoCs)	Alpha	928	2.94E-01	0.031
	Beta	954	5.81E-04	0.082
	Gamma	943	1.82E-05	0.071
	Delta	817	5.68E-07	0.835
	Delta Plus	908	1.04E-06	0.344
Variants of Interest (VoIs)	Epsilon	990	1.46E-08	0.481
	Eta	789	1.28E-08	0.508
	Lambda	629	3.86E-06	0.001
	Mu	729	9.21E-08	0.482

Decreasing trends for genome heterogeneity

The ridge regression of *SCC* against age shows a highly significant, decreasing trend. Figure 2 shows the regression obtained for the sample s1000_1. Similar decreasing trends were seen in the other samples from the general virus population; interestingly, non-significant slopes were obtained in most of the variant samples (Table 2).

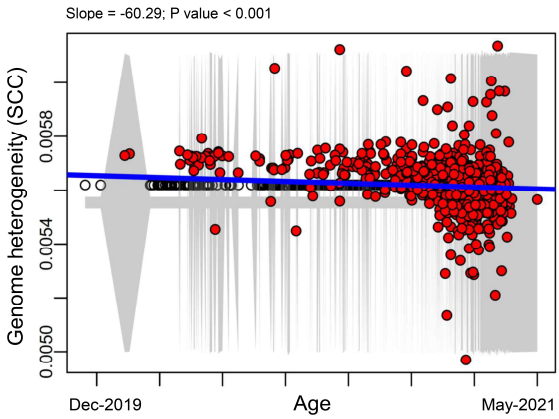

Figure 2. The phylogenetic trend for genome heterogeneity (*SCC*) was detected by the RRphylo R package[45,46] on the s1000_1 random sample. The estimated *SCC* value for each tip (red circles) or node (white circles) in the phylogenetic tree is regressed (blue line) against its age (the phylogenetic time distance, meant mainly as the collection date of each virus isolate). The statistical significance of the ridge regression slope 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 *SCC* values produced according to the Brownian motion model of evolution are shown as shaded areas.

Evolutionary rate

The ridge regression for the evolutionary rate of *SCC* obtained in the sample s1000_1 is shown in Figure 3. In contrast to the decreasing trend observed for *SCC* (Figure 2), an increasing trend was observed for its evolutionary rate (higher rate towards the present).

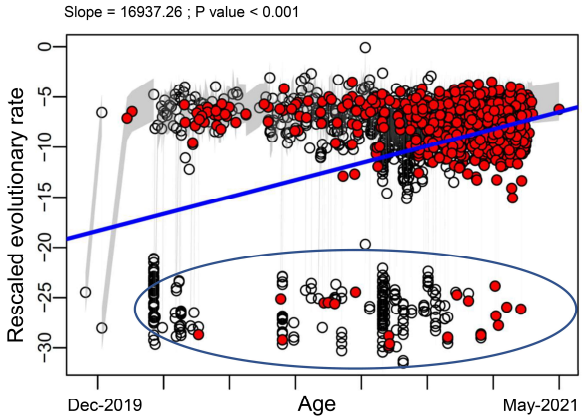

Figure 3. The phylogenetic trend for the evolutionary rate of *SCC* was detected by using the RRphylo R package [45,46] on the s1000_1 random sample. The rescaled evolutionary rate was obtained by rescaling the absolute rate in the 0-1 range and then transforming to logs to compare to the Brownian motion expectation. The statistical significance of the ridge regression slope was tested against 1,000 slopes obtained after simulating a simple Brownian evolution in the phylogenetic tree. The 95% confidence intervals around *SCC* values produced according to the Brownian motion model of evolution are shown as shaded areas. The oval highlights the low rate set of genomes.

Table 2 shows that the slopes for evolutionary rate were positive and highly significant in all the genome groups (except for the Gamma variant sample where there are a marginal significance).

High- and low-evolutionary-rate genomes

Despite the general positive slopes for evolutionary rate (e.g., Figure 3), a conspicuous set of genomes with a low evolutionary rate also appears. This set is indicated by an oval in Figure 3 for the sample s1000_1, but similar sets appear in all the samples, even in the variants. However, these genomes never form a separate clade on the tree, appearing instead scattered over different branches of the tree, and always spanning a wide range of ages.

Table 2. Phylogenetic trends in both genome heterogeneity (*SCC*) and its evolutionary rate as detected by the function *search.trend* from the RRphylo R package[45,46] on coronavirus samples. The estimated SCC value for each tip or node in the phylogenetic tree for each sample was regressed against age. The significance of the ridge regression slope was then evaluated against 1,000 slopes obtained after simulating a simple (i.e., no-trend) Brownian evolution of the trait in the phylogenetic tree.

			SCC regression		*SCC* rate regression	
Genome group	Sample	N	Slope	P value	Slope	P value
General virus population	s300_1	297	-68.92	0.019	9753.63	0.001
	s300_2	299	-71.03	0.017	8775.98	0.001
	s500_1	498	-66.39	0.011	15035.58	0.001
	s500_2	496	-55.74	0.026	11090.87	0.001
	s1000_1	987	-60.29	0.001	16937.26	0.001
	s1000_2	980	-54.23	0.004	16169.35	0.001
Variants of Concern (VoCs)	Alpha	928	-19.68	0.388	99079.25	0.001
	Beta	954	-4.49	0.499	36174.80	0.001
	Gamma	943	-4.90	0.465	38232.75	0.110
	Delta	817	26.77	0.260	20797.94	0.001
	Delta Plus	908	-10.67	0.426	40993.06	0.001
Variants of Interest (VoIs)	Epsilon	990	-42.09	0.146	45270.91	0.001
	Eta	789	37.68	0.117	15899.56	0.001
	Lambda	629	-38.51	0.224	48375.11	0.001
	Mu	729	45.22	0.162	52634.16	0.001

Discussion

Despite its relatively short length (29,912 bp for the GISAID reference genome (hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30) and the short time-lapse analyzed in this study (from December 2019 to October 2021), we found that the coronavirus RNA genomes were segmented into 5-10 compositional domains (~0.27 segments by kbp on average). Although such segment density is lower than in free-living organisms (e.g., cyanobacteria, where an average density of 0.47 segments by kbp was observed[47], it may suffice to do a comparative evolutionary analysis of the compositional heterogeneity of these genomes, which would shed light on the origin and evolution of the COVID-19 pandemic.

Using a genome heterogeneity measure (*SCC*) to characterize the genome sequence of each coronavirus, we were able to detect a consistent phylogenetic signal in all but one of the samples from the general virus population (Table 1), while only two variant samples (Alpha and Lambda) show a significant phylogenetic signal. According to earlier observations, which equate low phylogenetic signal with evolutionary lability and rapid evolutionary change[43,48], the loss of phylogenetic signal could be related to the high evolutionary rate we observed in most variant genomes (Table 2).

In most SARS-CoV-2 samples, the ridge regression of *SCC* against age shows a highly significant, negative slope (Figure 2 and Table 2), thus indicating a generalized declining trend along the ongoing process of pathogen evolution. However, the evolutionary rate of *SCC* shows positive slopes, thus indicating that the rates increase (i.e., they are faster than the Brownian motion expectation) toward the present. Therefore, an accelerated decline of sequence heterogeneity over time exists in the coronavirus.

The biological meaning of high, increasing evolutionary rates (Figure 3 and Table 2) indicate a fast changing (i.e., still adapting) virus genome. However, the co-existence in all the samples of a set of genomes with a low evolutionary rate (e.g., Figure 3) might suggest the existence of a stepwise distribution of rates, which will be a strong indication of selection in favor of different dominant strains (Prof. Pasquale Raia, personal communication).

The analysis of virus variants reveals non-significant regression slopes for *SCC*, low phylogenetic signals, and, concomitantly, a threefold increase in evolutionary rates (Table 2), thus indicating a further acceleration in the loss of sequence heterogeneity in the variants. These results may indicate the existence of a driven, adaptive trend in the variants. It is known that variant genomes have accumulated a higher proportion of adaptive mutations, allowing them to neutralize host resistance or escape host antibodies[49–51], consequently gaining higher transmissibility. In fact, more contagious and perhaps more virulent Variants of Concern (VoCs) share mutations and deletions that have arisen recurrently in distinct genetic backgrounds[52]. The higher number of adaptive changes might have altered the compositional evolutionary dynamics of variant genomes, disrupting the slowly decreasing trend in genome heterogeneity we observed in genomes from the general virus population.

A caution over our results could be the specific protocol we followed to select, filter, and mask the sampled genomes, as well as on the particular algorithm we used to infer phylogenetic trees. Therefore, we repeat our analyses using collected coronavirus samples and trees inferred by other groups; using these data, we found qualitatively comparable results. An example was the analysis of the SARS-CoV-2 Nextstrain global dataset containing 3059 genomes sampled between December 2019 and October 2021[5] , and using the ML phylodynamic tree obtained by these authors by means of the

TreeTime software[53]. Interestingly, we obtained a decreasing, although marginally significant trend, for genome heterogeneity (slope = -0.01, P value ≤ 0.122), as well as a slight, increasing, but highly significant trend for the evolutionary rate (slope = 0.89, P value ≤ 0.001).

The accelerated loss of genome heterogeneity in the coronavirus revealed in this work might be related to viral attenuation[54] leading to adaptation to the human host, a well-known process in other viruses[55], perhaps paralleling the ongoing transformation of the current pandemic into an epidemic. Further monitoring of current and new variants will allow checking these hypotheses to elucidate how virus evolution impacts human health.

Data and methods

Data retrieving

To search for coronavirus phylogenetic trends, we retrieved coronavirus genome sequences from the GISAID database[4,56,57]. At the time of writing, this database contains more than 4 million SARS-CoV-2 entries with complete collection dates, which we used as a proxy for the appearance over time of each strain (https://www.epicov.org/epi3/frontend#1ba380). To sample this big database, we used two approaches for random sampling. The first one merely consists of randomizing the database, then ordering and extracting each time an elite formed by the first 300, 500, or 1000… genome entries (Supplementary Tables S2-S16). A second approach uses a Python script (https://github.com/cris12gm/covid19/blob/master/getRandomSamples.py) to get random samples stratified by date.

We retrieved random samples for different genome groups of SARS-CoV-2 sequences: the general virus population, which includes genomes from different pandemic waves, Variants of Concern (VoCs) and Variants of Interest (VoIs). The number of genomes analyzed, and the ranges of collection date, number of segments and *SCC* values in the different genome groups are summarized in the Supplementary Table S1. We used the quality filters provided by the GISAID webpage to retrieve only high-quality genome sequences (only entries with complete collection date, larger than

29000 nt, with < 1% Ns and <0.05% unique aminoacid mutations (i.e., not seen in other sequences in database) and not insertions/deletions unless verified by submitter). The virus name, the collection date (spanning from December 2019 to October 2021), the genome heterogeneity (*SCC*) value, and the number of segments in the coronavirus samples analyzed here are detailed in Supplementary Tables S2-S16.

Genomic information on the official reference sequence employed by GISAID (EPI_ISL_402124, hCoV-19/Wuhan/WIV04/2019, WIV04) can be found at https://www.epicov.org/epi3/frontend#628435. We used this sequence as a root when inferring a phylogeny for each coronavirus sample. An updated genomic map of the isolate Wuhan-Hu-1 MN908947.3 we used for filtering and masking alignments of SARS-CoV-2 sequences is shown at https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3?report=graph. Note that although WIV04 is 12 nucleotides shorter than Wuhan-Hu-1 at the 3' end, the two sequences are identical in practical terms, particularly the 5' UTR is the same length, and the coding regions are identical. Therefore, the coordinates and relative changes are the same whichever sequence is used[58].

Filtering and masking

We followed the steps that have been recognized so far as useful for filtering and masking alignments of SARS-CoV-2 sequences, in this way avoiding sequence oddities[59]. First, we aligned the dataset of each sample to the genome sequence of the isolate Wuhan-Hu-1 (MN908947.3) using MAFFT[60], following the detailed protocol at https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473. We then mask the alignment using the *Python* script '*mask_alignment_using_vcf.py*' and following the detailed protocols at https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480 and https://github.com/W-L/ProblematicSites_SARS-CoV2. In this way, we avoid oddities in the SARS-CoV-2 genome sequences from GISAID, such as alignment ends, which are affected by low coverage and high rate of apparent sequencing/mapping errors, recurrent or systematic sequencing errors, or homoplasic, recombination, and hypervariable sites.

Multiple alignment and phylogeny

The masking of some sites in the alignment provokes that some of the sequences in the initial random sample become identical to others. Upon eliminating such duplicates, we realigned each sample with MAFFT[60] using default options. To solve polytomies, we used the function *fix.poly* from the RRphylo package V. 2.5.8[45,46]. We then infer the best ML trees for each sample by means of the software IQ-TREE 2[61], using the GTR nucleotide substitution model[62,63] and additional options suggested by the software (i.e., GTR+F+R2). We also used the least square dating (LSD2) method[64] to build a time-scaled tree. Finally, we rooted the obtained timetree to the GISAID coronavirus reference genome (hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30).

Coronavirus genome heterogeneity: the Sequence Compositional Complexity (SCC)

We measure genome heterogeneity by computing each genome's Sequence Compositional Complexity, or *SCC*[39]. This was a two-step process: the nucleotide sequence was first segmented into homogeneous, statistically significant compositional domains, then computing *SCC*.

Sequence segmentation

We divided a given nucleotide sequence into an array of compositionally homogeneous, non-overlapping domains using a heuristic, iterative segmentation algorithm[30,31,37,42]. In brief, a sliding cursor is moved along the sequence, and the position that optimizes an appropriate measure of compositional divergence between the left and right parts is selected. We choose the Jensen-Shannon divergence (equations (1) and (2) in ref[30]) as the divergence measure, as it can be directly applied to symbolic nucleotide sequences. If the divergence is statistically significant (at a given significance level, *s*), the sequence is split into two segments. Note that the resulting segments are more homogeneous than the original sequence. The two resulting segments are then independently subjected to a new round of segmentation. The process continues iteratively over the new resulting segments while sufficient significance continues appearing. Since Shannon entropy is invariant under symbol interchange, the segmentation algorithm and the *SCC* values derived from it, are invariable to sequence orientation,

Note that the statistical significance level *s* is the probability that the difference between adjacent domains is not due to statistical fluctuations. By changing this parameter, one can obtain the underlying distribution of segment lengths and nucleotide compositions at distinct levels of detail[65], thus fulfilling one of the key requirements to compute a complexity measure[66]. Recent improvements to this segmentation algorithm also allow segmenting long-range correlated sequences[65]. Implementation details, source codes, and executable binaries for different operating systems can be downloaded from: https://github.com/bioinfoUGR/segment and https://github.com/bioinfoUGR/isofinder.

The result is the segmentation of the original sequence into an array of contiguous, non-overlapping segments (or compositional domains) which are compositionally homogeneous at the chosen significance level (see Figure 1).

Computing *SCC*

Once a sequence is segmented into an array of homogeneous compositional domains, a reliable measure of Sequence Compositional Complexity or *SCC*[39] was computed:

$$SCC = H(S) - \sum_{i=1}^{n} \frac{G_i}{G} H(S_i) \qquad [1]$$

where S denotes the whole genome sequence and G its length, G_i the length of the ith domain S_i. $H(\cdot) = -\sum f log_2 f$ is the Shannon entropy of the distribution of relative frequencies of symbol occurrences, f, in the corresponding (sub)sequence. It should be noted that the above expression is the same as the one used in the segmentation process, applying it to the tentative two new subsequences (n = 2) to be obtained in each step. Thus, the two steps of the *SCC* computation are based on the same theoretical background. Note that 1) this measure is 0 if no segments are found in the sequence (the sequence is compositionally homogeneous, i.e., a random sequence) and 2) increases with both the number of segments and the degree of compositional differences among them. In this way, the *SCC* measure is analogous to the measure used by McShea and Brandon[67] to obtain complexity estimates on morphological characters: an organism is more complex if it has a greater number of parts and a higher differentiation among these parts. It should also be emphasized the high sensibility of our measure of sequence heterogeneity. Only one nucleotide substitution or one little indel often suffices to alter

the number, the length, or the nucleotide frequencies of the compositional domains, and therefore the resulting value for *SCC*.

Phylogenetic signal

The phylogenetic signal can be defined as the tendency for related species to resemble each other more than species randomly drawn from a phylogenetic tree. So, high values of the phylogenetic signal indicate that closely related species in the phylogeny tend to be more similar than expected by chance. It measures how trait variation, in our case *SCC*, is correlated with the phylogenetic relatedness of species[68,69]. Here we used Blomberg's *K*[43] to measure the phylogenetic signal. This metric is for continuous characters only and it uses an explicit model of trait evolution, the Brownian motion (BM) model. So, the expected variance behind this scenario is the summed branch length from the root to each species represented by a variance-covariance matrix. Blomberg's *K* measures phylogenetic signal by quantifying the observed trait variance relative to the trait variance expected under BM.

Phylogenetic ridge regression

Evolutionary trends for *SCC* were determined using the *RRphylo* R package V. 2.5.8[45,46]. The estimated *SCC* value for each tip or node in the phylogenetic tree is regressed against its age (the phylogenetic time distance, meant mainly as the collection date of each virus isolate). 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 our phylogenetic tree with the *search.trend* function of this package.

Evolutionary rate

The evolutionary rate of *SCC* was also computed by *search.trend* function of the RRphylo package[45,46]. However, to search for trends in evolutionary rate, its comparison to the expectation of the Brownian motion model is needed. To this end, the absolute evolutionary rate needs to be rescaled in the 0-1 range and then transformed to logs (Prof. Pasquale Raia, personal communication). Note that the time distance is expressed as the distance from the tree root (+1 for mathematical reasons). The

statistical significance of the ridge regression slope was tested against 1,000 slopes obtained after simulating a simple Brownian evolution in the phylogenetic tree.

Data availability

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

References

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

2. Moya, A., Holmes, E. C. & González-Candelas, F. The population genetics and evolutionary epidemiology of RNA viruses. *Nature Reviews Microbiology* vol. 2 279–288 (2004).

3. Hatcher, E. L. *et al.* Virus Variation Resource-improved response to emergent viral outbreaks. *Nucleic Acids Research* **45**, D482–D490 (2017).

4. Koehorst, J. *et al.* GISAID Global Initiative on Sharing All Influenza Data. Phylogeny of SARS-like betacoronaviruses including novel coronavirus (nCoV). *Oxford* **34**, 1401–1403 (2017).

5. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics (Oxford, England)* **34**, 4121–4123 (2018).

6. Naqvi, A. A. T. *et al.* Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochimica et Biophysica Acta - Molecular Basis of Disease* vol. 1866 165878 (2020).

7. Patiño-Galindo, J. Á. *et al.* Recombination and lineage-specific mutations linked to the emergence of SARS-CoV-2. *Genome Medicine* **13**, 124 (2021).

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

9. Sánchez, C. A. *et al.* A strategy to assess spillover risk of bat SARS-related coronaviruses in Southeast Asia. *medRxiv* 2021.09.09.21263359 (2021) doi:10.1101/2021.09.09.21263359.

10. Li, W. *et al.* Bats are natural reservoirs of SARS-like coronaviruses. *Science* **310**, 676–679 (2005).

11. Zhang, Y. Z. & Holmes, E. C. A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* **181**, 223–227 (2020).

12. Ji, W., Wang, W., Zhao, X., Zai, J. & Li, X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. *Journal of Medical Virology* **92**, 433–440 (2020).

13. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology* **30**, 1346-1351.e2 (2020).

14. Gu, H., Chu, D. K. W., Peiris, M. & Poon, L. L. M. Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses. *Virus Evolution* **6**, (2020).

15. Guo, H. *et al.* Identification of a novel lineage bat SARS-related coronaviruses that use bat 1 ACE2 receptor 2. *bioRxiv* 2021.05.21.445091 (2021) doi:10.1101/2021.05.21.445091.

16. Liu, Z. *et al.* Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. *Journal of Medical Virology* **92**, 595–601 (2020).

17. Pybus, O. G., Tatem, A. J. & Lemey, P. Virus evolution and transmission in an ever more connected world. *Proceedings of the Royal Society B: Biological Sciences* vol. 282 (2015).

18. Islam, M. R. *et al.* Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. *Scientific Reports* **10**, (2020).

19. Cyranoski, D. Profile of a killer: the complex biology powering the coronavirus pandemic. *Nature* **581**, 22–26 (2020).

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

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

22. Mishra, A. *et al.* Mutation landscape of SARS-CoV-2 reveals three mutually exclusive clusters of leading and trailing single nucleotide substitutions. *bioRxiv* 2020.05.07.082768 (2020) doi:10.1101/2020.05.07.082768.

23. Banerjee, A. *et al.* The novel Coronavirus enigma: Phylogeny and mutation analyses of SARS-CoV-2 viruses circulating in India during early 2020. *bioRxiv* 2020.05.25.114199 (2020) doi:10.1101/2020.05.25.114199.

24. Eskier, D., Karakülah, G., Suner, A. & Oktay, Y. RdRp mutations are associated with SARS-CoV-2 genome evolution. *bioRxiv* 2020.05.20.104885 (2020) doi:10.1101/2020.05.20.104885.

25. Cai, H. Y., Cai, K. K. & Li, J. Identification of Novel Missense Mutations in a Large Number of Recent SARS-CoV-2 Genome Sequences. *Journal General Medical Research* **2**, (2020).

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

27. Wright, E. S., Lakdawala, S. S. & Cooper, V. S. SARS-CoV-2 genome evolution exposes early human adaptations. *bioRxiv* 2020.05.26.117069 (2020) doi:10.1101/2020.05.26.117069.

28. Benvenuto, D. *et al.* The 2019-new coronavirus epidemic: Evidence for virus evolution. *Journal of Medical Virology* **92**, 455–459 (2020).

29. Young, B. E. *et al.* 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 (2020).

30. Bernaola-Galván, P., Román-Roldán, R. & Oliver, J. L. Compositional segmentation and long-range fractal correlations in DNA sequences. *Physical review E* **53**, 5181–5189 (1996).

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

32. Wen, S.-Y. & Zhang, C.-T. Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. *Biochemical and Biophysical Research Communications* **311**, 215–222 (2003).

33. Keith, J. M. Sequence segmentation. *Methods in molecular biology (Clifton, N.J.)* **452**, 207–29 (2008).

34. Li, W. Delineating relative homogeneous G+C domains in DNA sequences. *Gene* **276**, 57–72 (2001).

35. Bernaola-Galván, P., Oliver, J. L., Carpena, P., Clay, O. & Bernardi, G. Quantifying intrachromosomal GC heterogeneity in prokaryotic genomes. *Gene* **333**, (2004).

36. Oliver, J. L. *et al.* Isochore chromosome maps of the human genome. in *Gene* vol. 300 117–127 (Gene, 2002).

37. Oliver, J., Carpena, P., Hackenberg, M. & Bernaola-Galván, P. IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res* **32**, W287-92 (2004).

38. Fearnhead, P. & Vasilieou, D. Bayesian Analysis of Isochores. *Journal of the American Statistical Association* (2009).

39. Román-Roldán, R., Bernaola-Galván, P. & Oliver, J. L. Sequence compositional complexity of DNA through an entropic segmentation method. *Physical Review Letters* **80**, 1344–1347 (1998).

40. Bernardi, G. *et al.* The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958 (1985).

41. Bernardi, G. Chromosome architecture and genome organization. *PLoS ONE* **10**, e0143739 (2015).

42. Bernaola-Galván, P., Carpena, P. & Oliver, J. A standalone version of IsoFinder for the computational prediction of isochores in genome sequences. *arXiv preprint arXiv:0806.1292* 1–7 (2008).

43. Blomberg, S., Garland, T. & Ives, A. R. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).

44. Keck, F., Rimet, F., Bouchez, A. & Franc, A. Phylosignal: An R package to measure, test, and explore the phylogenetic signal. *Ecology and Evolution* **6**, 2774–2780 (2016).

45. Castiglione, S. *et al.* A new, fast method to search for morphological convergence with shape data. *PLoS ONE* **14**, e0226949 (2019).

46. Castiglione, S. *et al.* Simultaneous detection of macroevolutionary patterns in phenotypic means and rate of change with and within phylogenetic trees including extinct species. *PLoS ONE* **14**, (2019).

47. Moya, A. *et al.* Driven progressive evolution of genome sequence complexity in Cyanobacteria. *Scientific Reports* **10**, (2020).

48. Rheindt, F. E., Grafe, T. U. & Abouheif, E. *Rapidly evolving traits and the comparative method: how important is testing for phylogenetic signal?* (2004).

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

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

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

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

53. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evolution* **4**, (2018).

54. Badgett, M. R., Auer, A., Carmichael, L. E., Parrish, C. R. & Bull, J. J. Evolutionary Dynamics of Viral Attenuation. *Journal of Virology* **76**, 10524 (2002).

55. Bahir, I., Fromer, M., Prat, Y. & Linial, M. Viral adaptation to host: A proteome-based analysis of codon usage and amino acid preferences. *Molecular Systems Biology* **5**, 311 (2009).

56. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33–46 (2017).

57. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* vol. 22 (2017).

58. Singer, J., Gifford, R., Cotten, M. & Robertson, D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. *Preprints* 2020060225 (2020) doi:10.20944/PREPRINTS202006.0225.V1.

59. Hodcroft, E. B. *et al.* Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* vol. 591 30–33 (2021).

60. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

61. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

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

63. Rodríguez, F., Oliver, J. L., Marín, A. & Medina, J. R. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* vol. 142 485–501 (1990).

64. To, T. H., Jung, M., Lycett, S. & Gascuel, O. Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology* **65**, 82–97 (2016).

65. Bernaola-Galván, P. *et al.* Segmentation of time series with long-range fractal correlations. *The European Physical Journal B* **85**, 211 (2012).

66. Gell-Mann, M. & Lloyd, S. Information measures, effective complexity, and total information. *Complexity* (1996) doi:10.1002/(SICI)1099-0526(199609/10)2:1<44::AID-CPLX10>3.0.CO;2-X.

67. McShea, D. W. & Brandon, R. N. *Biology's first law : the tendency for diversity and complexity to increase in evolutionary systems*. (University of Chicago Press, 2010).

68. Revell, L. J., Harmon, L. J. & Collar, D. C. Phylogenetic signal, evolutionary process, and rate. *Systematic biology* **57**, 591–601 (2008).

69. Revell, L. J. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* **1**, 319–329 (2010).

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). Special thanks are due to Professor Pasquale Raia and Dr. Silvia Castiglione for its advice in applying the RRphylo package to coronavirus data, particularly the development of a new function (*fix.poly*) to resolve polytomies in the tree, and also for sharing code to rescale the evolutionary rate. 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 S17 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 S18.

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

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, M.V. and A.M. performed research. J.L.O., P.B., F.P., C.G.M, M.V. and A.M. analyzed data; J.L.O., M.V., and A.M. 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.

Additional information

Supplementary information:

- Supplementary Tables S1-S16: File SupplementaryTables_S1_S16.xlsx
- Supplementary Table S17: File SupplementaryTableS17.pdf
- Supplementary Table S18: File SupplementaryTableS18.tsv