Integrated multi-scale modeling for synthetic genetic constructions

Lucas GOIRIZ BELTRÁN

Thesis Advisors:
Dr. Guillermo RODRIGO TÁRREGA
Prof. J. Alberto CONEJERO CASARES

Dummy PDF file

Dummy PDF file

Do not trust, verify

- vires in numeris -

Integrated multi-scale modeling for synthetic genetic constructions

Lucas Goiriz Beltrán

Abstract

The evolution of SARS-CoV-2 in humans has been monitored at an unprecedented level due to the public health crisis, yet the stochastic dynamics underlying such a process is dubious. Here, considering the number of acquired mutations as the displacement of the viral particle from the origin, we performed biostatistical analyses from numerous whole genome sequences on the basis of a time-dependent probabilistic mathematical model. We showed that a model with constant variant-dependent evolution rate and nonlinear mutational variance with time (i.e., anomalous diffusion) explained the SARS-CoV-2 evolutionary motion in humans during the first 120 weeks of pandemic in UK. In particular, we found subdiffusion patterns for the Primal, Alpha, and Omicron variants, while a weak superdiffusion pattern for the Delta variant. Our findings indicate that non-Brownian evolutionary motions occur in nature, thereby providing novel insight for viral phylodynamics.

Modelado multiescala integrativo para construcciones genéticas sintéticas

Lucas Goiriz Beltrán

Resumen

Ellipsis

Modelatge multiescala integratiu per a construccions genètiques sintètiques

Lucas Goiriz Beltrán

Resum

Ellipsis

Contents

Objectives

Ellipsis

Chapter 1

Introduction

cita interesante 1.
- El pavo al que cito

Ellipsis

Chapter 2

A variant-dependent molecular clock with anomalous diffusion models SARS-CoV-2 evolution in humans

Quieres chocolate?
- Roser, alguna que otra tarde

This work has been published in a peer-reviewed journal. See the full citation:

Goiriz L, Ruiz R, Garibo-i-Orts O, Conejero J.A, Rodrigo G. A variant-dependent molecular clock with anomalous diffusion models SARS-CoV-2 evolution in humans.

In this publication, I performed all the data analysis and figures, except Figure X where OGO provided substantial help. The results were discussed with RR, JAC and GR. GR designed the research.

2.1 Introduction

Viruses lie at the frontier of living and inert matter, as they lack own metabolism to sustain replication but are subject to Darwinian evolution (*i.e.*, mutation and selection) [1]. As fast evolving biological agents [2], they are ideal substrates from which to learn mechanisms that modulate genetic variation, as well as to test theoretical models of evolution. One important model is the molecular clock hypothesis, which dates back to early times of molecular biology and states that the rates at which genes accumulate mutations are constant with time [3, 4]. Neutral theory of molecular evolution predicts, in addition, that such clocks are Poissonian stochastic processes (*i.e.*, evolution seen as a Brownian motion with diffusivity such that mean and variance are equal) [5]. Despite the results from seminal studies of some viral genes are in tune [6, 7], the molecular clock hypothesis still raises controversy [4], as evolution appears as a highly volatile and vagary stochastic process due to environmental changes, transmission bottlenecks, and recombination and speciation events. Indeed, such a null model can be rejected in numerous cases [8], and overdispersed populations in genetic variation (*i.e.*, with larger variance than mean) seem common across phyla [9]. Nonetheless, without extensive monitoring of evolution in natural conditions for a reasonable period of time, it is difficult to describe the mathematical model underlying such stochastic dynamics.

The emergence (at the end of 2019) and global spread (during 2020) of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused the ongoing pandemic [10]. Due to the high impact on public health, huge efforts have been carried out worldwide to sequence the whole viral genome from different patients in real time of pandemic [11, 12, 13, 14, 15, 16, 17]. To date, more than 10 million sequences are available. Importantly, this has allowed tracking introductions and measuring the extent of virus transmission at the community level [11], assessing the effectiveness of containment strategies [12], realizing the impact of superspreading events [13], identifying novel variants of concern and their key mutations [14], monitoring spatiotemporal invasion dynamics of specific variants [15], disclosing co-infection occurrences with different variants [16], and discovering the circulation of recombinant viruses, which might lead to increased infectiousness or pathogenicity [17]. In addition, computational analyses have shown that SARS-CoV-2 primary evolves under purifying selection in humans, with some sites displaying adaptation [18]. They have also shown that the spectrum of acquired mutations is greatly asymmetric (*viz.*, dominated by C>U and G>U substitutions) [19]. Notwithstanding, a detailed study of the stochastic dynamics underlying such evolutionary motion to assess fundamental scaling laws and conditions for criticality has not been carried out.

In this work, we exploited the unprecedented monitoring of the evolution of a human virus in nature to conduct a study aimed at describing this process as a (non-)Brownian motion, considering the number of acquired mutations as the displacement of the viral particle from the origin (**Fig. 2.1**). For that, several biostatistical analyses over millions of whole genome sequences at the ensemble level were evaluated on the basis of a time-dependent probabilistic mathematical model, without relying on phylogeny. Of note, our results challenge the conventional molecular clock hypothesis by providing new theoretical foundations for viral evolution.

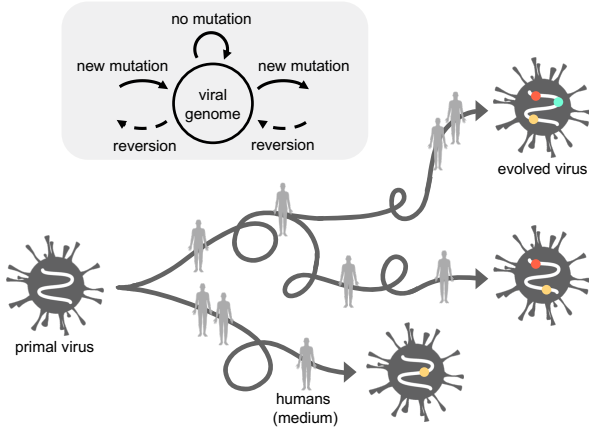

Figure 2.1: Schematics of the evolutionary motion of the virus (viewed as a stochastic process). Inset: associated state-transition diagram.

2.2 Results

We sought to characterize the mean and variance (mean squared displacement) of the overall stochastic process by which the observable viral genome accumulates mutations with time (since the emergence in Wuhan, China). This was modeled in a continuous form by the Langevin equation $\frac{dm(t)}{dt} = \kappa + \xi(t)$, where $m(t)$ is the total number of mutations in the genome at time t, κ the evolution rate (which could be time-dependent), and $\xi(t)$ an integrative noise source whose properties shape the evolutionary motion.

Due to the large number of available SARS-CoV-2 sequences from United Kingdom (UK), our analysis was focused on this country. Using landmark multidimensional scaling [20], we obtained a representation of all available

genotypes in a two-dimensional space (**Fig. 2.2**), which served to appreciate the virus evolution as a complex diffusion process. In this polar plot, the radius represented the number of mutations and the angle encompassed the rest of sequence variation. To characterize the stochastic process, we first quantified the rate at which the viral genome accumulates a mean number of mutations with time. Considering all types of mutations and discretizing time by weeks (*i.e.*, all sequences available in a week were pooled together), we obtained a macroscopic evolution rate of 0.62 wk^{-1} (Pearson's correlation with no intercept, $P < 10^{-4}$; **Fig. 2.3a**). Substitutions were much more frequent than insertions and deletions (indels). However, at some points (at the end of 2020 and of 2021), an acceleration in the evolution rate was observed, thereby deviating from a molecular clock model with constant rate. Yet, without phylogenetic inference, this picture just reflected population changes and not strict evolutionary paths. In addition, mutations were classified according to their type (*viz.*, non-coding, synonymous, non-synonymous, and indels), and the ratio between the number of nonsynonymous and synonymous substitutions per site (dN/dS) was estimated (**Fig. 2.3b**). The dN/dS signature (fluctuation around 1 over time) suggested evolution under purifying selection of a series of adapted variants.

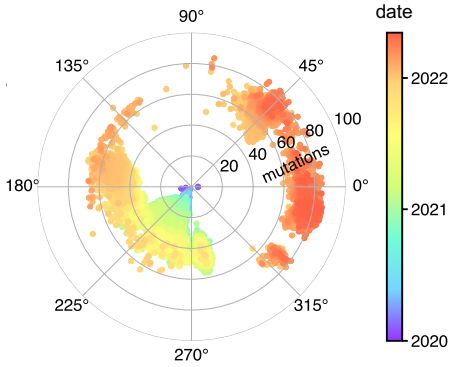

Figure 2.2: 2D projection of all viral sequences colored by date.

To test whether the accumulation of mutations in SARS-CoV-2 was a Poissonian stochastic process, we also calculated the variance and the dispersion index, understood as the ratio between variance and mean (**Fig. 2.4a**). The study of the variance is often overlooked, despite it is essential to comprehend the evolutionary motion. We found a largely sub-Poissonian dynamics (*i.e.*, dispersion index < 1) with two main dispersion bursts at the times at which the evolution rate was accelerated. To inspect the origin of such a dynamic profile, we performed a sequence classification into variants. For simplicity, four vari-

ants were considered, *viz.*, Primal, Alpha, Delta, and Omicron. We realized that the first dispersion burst corresponded to the transition from Primal to Alpha, while the second to the transition from Delta to Omicron (**Fig. 2.4b**). The number of new coronavirus disease 2019 (COVID-19) cases also correlated with the variance (inset of **Fig. 2.4a**). Representing the distributions of accumulated mutations with time, we disclosed a bimodal behavior during such transitions (**Fig. 2.5a, b**), explaining the increased dispersion. The invading genotypes carried about 15-20 more mutations on average. Moreover, the transition from Alpha to Delta only generated a slight dispersion signal because both variants carried a similar number of mutations. Arguably, outlier SARS-CoV-2 genotypes in the existing distribution at a time led to the emergence of new variants, and the observed accelerations in evolution rate came from the inherent stochasticity of the evolutionary motion followed by rapid, mostly deterministic invasion events once a particular genotype acquired a selective advantage, such as higher transmissibility [15].

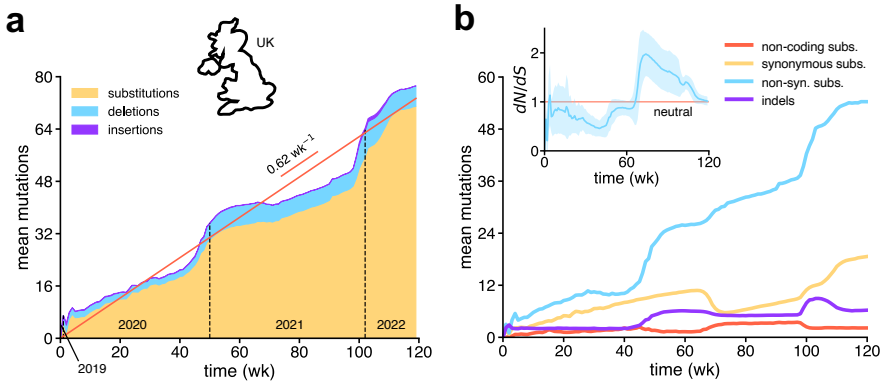

Figure 2.3: a) Time-course of the mean number of accumulated mutations in the viral genome, distinguishing between substitutions, deletions, and insertions. Linear regression over the total shown in red ($R^2 = 0.95$). b) Time-course of the mean number of non-coding substitutions, synonymous substitutions, non-synonymous substitutions, and indels. Inset: dN/dS with time (mean plus/minus standard deviation).

Due to the virus population reset caused by the invasion of a new variant, we calculated the time-dependent statistics per variant. The analyses conducted for each variant were independent from each other by considering subsets of properly annotated sequences (*i.e.*, no evolutionary paths between variants were considered). Of note, the evolution rates of Primal, Alpha, and Delta were substantially lower (up to 0.36 wk^{-1}) than the inferred macroscopic value of 0.62 wk^{-1} (**Fig. 2.6a**), in agreement with previous estimates following phylogenetic methods [21]. In the analyzed dataset, the Omicron population

was composed of two lineages with sufficient dissimilarity, *viz.*, BA.1 and BA.2 (BA.1 displaced Delta and BA.2 displaced BA.1). Performing a decomposition, we observed that BA.1 evolved faster than BA.2 in UK. Collectively, the mean evolutionary motion was well captured by $\langle m(t) \rangle = \kappa t$ with a constant variant-dependent rate, considering $\langle \xi(t) \rangle = 0$ (Pearson's correlations, $P < 10^{-4}$ in all cases).

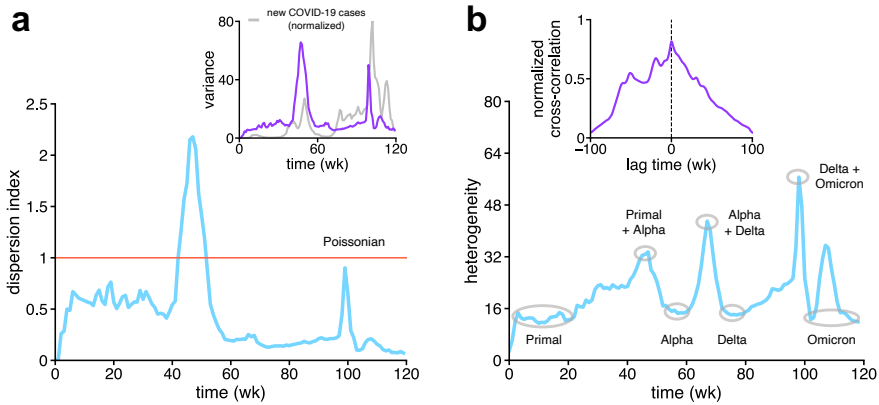

Figure 2.4: a) Time-course of the dispersion index (variance/mean). Inset: variance with time; the normalized number of new COVID-19 cases is superimposed. b) Time-course of the degree of heterogeneity (mean Hamming distance), showing the different stages of the virus population in terms of variants. Inset: normalized cross-correlation between variance and heterogeneity with time.

In addition, we found significant nonlinear dependencies of the variance with time in all cases (Fisher-Snedecor's F tests, $P < 10^{-4}$ for Primal, Alpha, and Delta, $P = 0.027$ for Omicron BA.1, and $P = 0.0003$ for Omicron BA.2; **Fig. 2.6b**), which indicated a stochastic behavior with anomalous diffusion [22]. In other words, SARS-CoV-2 underwent a non-Brownian evolutionary motion. This exciting result entailed that the explorations of the genotypic space by the virus to discover new phenotypes at different times were not fully uncorrelated within a clade. To provide a quantitative picture of the process, we fitted $\langle \Delta m(t)^2 \rangle$ to the general expression Dt^α, where D is the diffusion coefficient and α the diffusion exponent (this could be derived considering $\langle \xi(t)\xi(t') \rangle = \frac{1}{2} D\alpha(\alpha - 1) \left| t - t' \right|^{\alpha - 2}$ as the autocorrelation function of the noise source). We found subdiffusion ($\alpha = 0.42$, $\alpha = 0.47$, and $\alpha = 0.28$, respectively) in the cases of Primal, Alpha, and Omicron BA.1, while weak superdiffusion ($\alpha = 1.34$) in the case of Delta (Pearson's correlations in log scale, $P < 10^{-4}$ for Primal, Alpha, and Delta and $P = 0.020$ for Omicron BA.1; **Fig. 2.7a**). Although not plotted, we also found subdiffusion for Omicron BA.2 ($\alpha = 0.37$). The robustness of these results was assessed by bootstrapping,

i.e., performing a sampling with replacement of the sequences available each week in the original dataset and recomputing the dynamic profile of the variance. This also allowed dealing with the sequence pseudoreplication issue due to a shared history. Tolerable uncertainties for the diffusion parameters were noticed (inset of **Fig. 2.7a**).

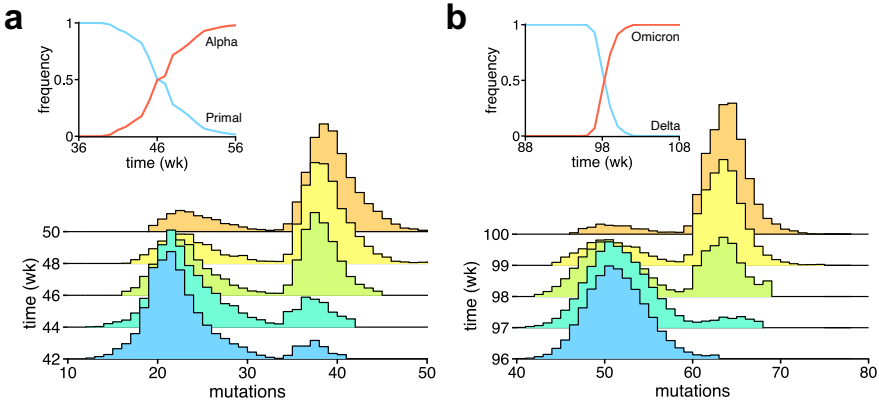

Figure 2.5: Probability-based histograms of the number of accumulated mutations during the transition from Primal to Alpha (a) and Delta to Omicron (b). Insets: population frequency of the variants with time.

To inspect the origin of anomalous diffusion in evolutionary motion, the rate at which the dN/dS ratio changed with time was analyzed per variant (inset of **Fig. 2.6b**). We observed a decreasing trend in all cases, more pronounced for Delta. This suggested that Delta evolved by accumulating more synonymous mutations per site than the other variants. If these mutations were neutral [23], the evolved genotypes of Primal, Alpha, and Omicron BA.1 would be more constrained as a result of their non-synonymous mutations, thereby explaining, at least in part, the observed subdiffusion patterns. Furthermore, we calculated a reset dispersion index, considering the accumulation of mutations since the appearance of the variant of study (*i.e.*, each time a new variant invades the population, the number of mutations is reset). At long times, we found values in the neighborhood of 1, revealing an asymptotic Poissonian behavior following this metric (**Fig. 2.7b**).

2.3 Discussion

The observation of patterns of anomalous diffusion in biology has opened new avenues of research [22]. Intriguingly, recent studies in which the physical movement of single SARS-CoV-2 virions was monitored throughout the in-

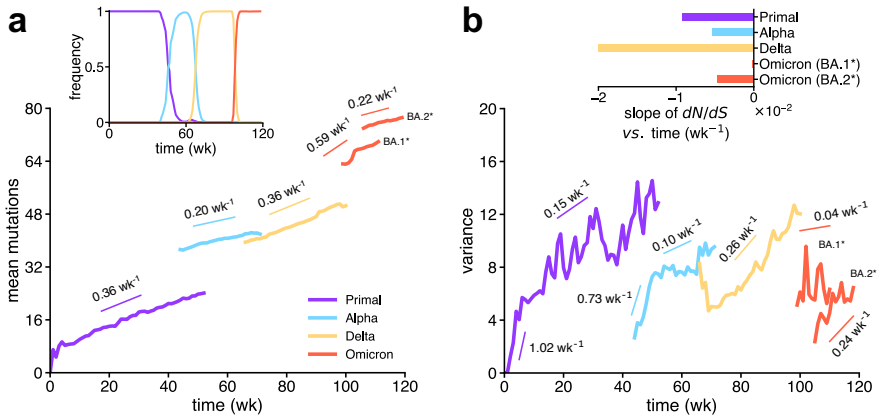

Figure 2.6: a) Time-course of the mean number of accumulated mutations per variant (Omicron decomposed into BA.1 and BA.2). Linear regressions shown in each case ($R^2 \geq 0.90$). Inset: population frequency of the variants with time. b) Time-course of the variance per variant. Piecewise linear regressions shown in each case. Inset: slope of dN/dS with time for each variant obtained by linear regression.

fectious cycle highlighted transient and variant-dependent directionality and confinement outside and inside the cell [24, 25], indicating deviation from a pure Brownian motion. Here, we have presented a new application domain in evolution. Of note, we uncovered that a probabilistic model with constant variant-dependent evolution rate and nonlinear mutational variance with time explained the SARS-CoV-2 evolutionary motion in humans during the first 120 weeks of pandemic in UK. This model might be used to refine phylodynamic approaches aimed at understanding the spread and adaptation of the virus. As shown, canonical descriptions based on the Poisson distribution do not accurately capture the observed dispersion at all time points. Notwithstanding, it is worth to note the bias in this type of studies caused by the fact that most of the sequenced viral genomes came from symptomatic infected people. Another issue is the imbalance in sequencing effort among countries, which prevents performing comprehensive analyses at a global scale. Further studies are required to assess the potential impact of movement and contact restrictions, vaccination, and self-diagnostic testing on the observed dynamic patterns. Overall, we anticipate deep implications of our data-driven results for future evolutionary and genomic studies, especially when dealing with fast evolving biological agents such as viruses.

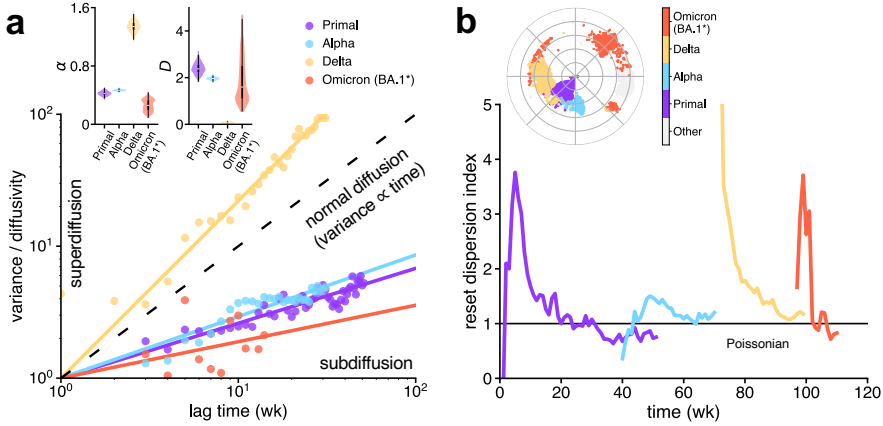

Figure 2.7: a) Representation of the rescaled variance normalized by the diffusivity (D) with time in log scale (points correspond to data; for Primal, Alpha, Delta, and Omicron BA.1, R^2 = 0.87, 0.88, 0.94, and 0.37, respectively, relative to Pearson's correlations in log scale). The slope of the fitted lines (α) defines the type of diffusion ($\alpha > 1$ superdiffusion, $\alpha < 1$ subdiffusion). Shaded areas represent the 95% confidence intervals of the regression lines. Inset: distributions of values for each diffusion parameter (violin plots) obtained by bootstrapping. b) Time-course of the reset dispersion index per variant. Inset: 2D projection of all viral sequences colored by variant (projection as in **Fig. 2.2**). The variant-specific analyses were restricted to the time period in which their population frequency was at least 10%.

2.4 Materials and Methods

Whole-genome sequencing data

The nucleotide sequences of the SARS-CoV-2 genomes used in this study were retrieved from the GISAID database (https://www.gisaid.org). As of May 2022, 10791877 sequences and the corresponding metadata were downloaded. Our analysis was restricted to the data from UK, which consists of 2735543 sequences.

Pairwise sequence alignments

The nucleotide sequences of the SARS-CoV-2 genomes were aligned against a reference genome using Multiple Alignment using Fast Fourier Transform (MAFFT) [26]. The results were collected in Clustal format. In this work, the reference sequence (root) was `hCoV-19/Wuhan/IVDC-HB-01/2019` (`EPI_ISL_402119`), which has 100% identity with the GeneBank reference genome (`NC_045512.2`) as shown by a Clustal Omega alignment.

Construction of a functional dataset

For each sequence, the number of mutations (substitutions, insertions, and deletions) with respect to the reference SARS-CoV-2 genome were counted. This information was retrieved from the MAFFT output alignments. In addition, the sequence collection dates and Pangolin lineages were retrieved from the metadata. Next, the sequences with unreliable recorded dates, whose unresolved base content surpassed 1% (proportion of `Ns`), or whose size was below 25 kb were discarded, as they were considered of low quality. Duplicated entries in the dataset were also removed. Furthermore, sequences isolated from non-human hosts were discarded. Then, the variant names were assigned where applicable based on the Pangolin lineage annotation. The Pangolin lineage-to-variant mapping was performed with information available at the Cov-lineages initiative (https://cov-lineages.org). All sequences dated earlier than 21 February 2021 were annotated as Primal variant (*i.e.*, the original SARS-CoV-2 variant from Wuhan). The sequences were grouped by weeks. Finally, for each week, mutation outliers were filtered out to avoid artifacts in the calculated statistical parameters. These outliers could originate from incorrect date annotations, aberrant evolutionary trajectories, or sudden point introduction [27, 28]. For each week, if the number of sequences was greater than 20, a Generalized Extreme Studentized Deviate (GESD) many-outlier procedure was applied [29]. Otherwise, a filtering based on interquartile ranges was performed (*i.e.*, the upper/lower bound was equal to the first/third quartile point plus/minus the interquartile range).

All data analyses were performed in Python using the libraries pandas (https://pandas.pydata.org), NumPy (https://numpy.org), SciPy (https://scipy.org), scikit-learn (https://scikit-learn.org), and Biopython (https://biopython.org)

Landmark multidimensional scaling

Landmark multidimensional scaling (LMDS) is a variation of classical multidimensional scaling (MDS) [30] used to analyze and visualize dissimilarities between items based on a set of pairwise distance measures. The technique uses a small number of landmark items to compute their pairwise distances and estimate the distances between the remaining items, which are then mapped into a low-dimensional space using MDS [20]. LMDS presents several advantages over traditional MDS, including reduced computational complexity, scalability, and flexibility, making it an appropriate tool for analyzing large and complex datasets.

LMDS and principal component analysis (PCA) are both techniques used for dimensionality reduction, but they have some fundamental differences in

their goals and methods. In contrast to LMDS, PCA is used to identify the underlying structure in a dataset by finding the principal components that explain the most variance in the data, capturing this way the most important patterns in the data [31]. The interpretation of the coordinates in the projection space of both techniques is different as well, with PCA representing the patterns of variation and MDS representing the similarities and dissimilarities among the data points.

To obtain a representation of all available sequences in a two-dimensional (2D) space, a procedure based on landmark multidimensional scaling (LMDS) was followed [20]. For that, the Hamming distance between any two sequences was calculated (given by the number of mutations that separate each other). Metric axioms (minimality, symmetry, and triangle inequality) hold for the Hamming distance, so LMDS can be applied. The following sequences were used as landmarks:

hCoV-19/England/CAMC-C42AEA/2020	hCoV-19/England/LSPA-2E824E8/2021
hCoV-19/England/PHEC-YYBI3UW/2021	hCoV-19/Scotland/NORT-YBF4CD/2021
hCoV-19/England/LSPA-3DC3179/2022	hCoV-19/England/ALDP-3A3CE1D/2022
hCoV-19/England/ALDP-2E0DCFC/2021	hCoV-19/Wales/PHWC-PYDUBM/2021
hCoV-19/England/QEUH-F8AA01/2021	hCoV-19/Scotland/QEUH-9AD0C0/2020

Table 2.1: Sequence identifiers of landmarks employed during LMDS analysis.

Polar coordinates were used to project the sequences in a 2D space. For the sake of interpretability, the radius was directly the total number of mutations from root and the angle was obtained from the coordinates generated by LMDS. It is worth noting that the position of two sequences in the projection plane will be determined by their similarity according to the Hamming distance. Consequently, if two sequences have the same number of mutations but these mutations are different, they will be located in different regions of the plane (`i.e.`, their angles will be different and the radius will be the same).

Calculation of statistical parameters

For each set of SARS-CoV-2 sequences in a week, the mean and variance of the number of accumulated mutations (including substitutions and indels) were computed. For this computation, only the number of mutations was considered, not their type (*i.e.*, different sequences with the same number of mutations counted the same). Then, the dispersion index was calculated, defined as the ratio between variance and mean. In addition, the mean Hamming distance between all sequence pairs in a week was computed to realize the extent of sequence heterogeneity. To compute the normalized cross-correlation

between mutational variance and sequence heterogeneity, the `correlate` function from NumPy was used, having previously divided the statistical parameters by their norm (using the `linalg.norm` function). Finally, the number of COVID-19 cases in UK was retrieved from the database Our World in Data (https://ourworldindata.org/). The weekly number of new cases was computed. Probability-based histograms of the total number of mutations were obtained with the NumPy `histogram` function (`density=True`).

The calculation of mutational mean and variance was also performed per variant. This was done for the major lineages Primal, Alpha, Delta, and Omicron [32], considering only the time period in which the variant represented at least the 10% of the population. This limit was applied to avoid artifacts in the calculated statistical parameters due to a low number of sequences. In the case of Omicron, the calculation was performed for the sublineages BA.1 and BA.2 because of their great difference in mutations [33]. For each variant, a reset dispersion index was also calculated, defined as the ratio between variance and the mean number of accumulated mutations since the first appearance of the variant in the population (*i.e.*, each time a new variant invades the population, the number of mutations is reset). To some extent, this is in tune with the definition of a founder genotype for each clade from which to start counting as done in ref. [34].

Specifically, if there are N_k sequences in the k^{th} week, the mean number of accumulated mutations in that week, denoted by $\langle m_k \rangle$, is calculated as

$$\langle m_k \rangle = \frac{1}{N_k} \sum_{i=1}^{N_k} m_{k,i}$$

where $m_{k,i}$ is the number of mutations of the i^{th} sequence in the k^{th} week. And the variance, denoted by $\langle \Delta m_k^2 \rangle$, is calculated as

$$\langle \Delta m_k^2 \rangle = \frac{1}{N_k - 1} \sum_{i=1}^{N_k} (m_{k,i} - \langle m_k \rangle)^2$$

To perform the calculations for a particular variant v, only the sequences annotated as such were considered (note that one sequence is linked at most to one particular variant). If there are $N_{v,k}$ sequences in the k^{th} week for variant v, the mean number of accumulated mutations in that week, denoted by $\langle m_{v,k} \rangle$ is calculated as

$$\langle m_{v,k} \rangle = \frac{1}{N_{v,k}} \sum_{i=1}^{N_{v,k}} m_{v,k,i}$$

where $m_{v,k,i}$ is the number of mutations of the i^{th} sequence in the k^{th} week

annotated as variant v. And the variance, denoted by $\left\langle \Delta m^2_{v,k} \right\rangle$, is calculated as

$$\left\langle \Delta m^2_{v,k} \right\rangle = \frac{1}{N_{v_k} - 1} \sum_{i=1}^{N_{v,k}} \left(m_{v,k,i} - \langle m_{v,k} \rangle \right)^2$$

Certainly, $N_k = \sum_{v \in V} N_{v,k} + N_{\emptyset,k}$, where V is the set of variants and $N_{\emptyset,k}$ denotes the number of sequences that are not linked to any variant of V in the k^{th} week.

Moreover, the dispersion index in the k^{th} week (ρ_k) is calculated as

$$\rho_k = \frac{\left\langle \Delta m^2_{v,k} \right\rangle}{\langle m_{v,k} \rangle}$$

and the reset dispersion index in the k^{th} week (ρ_k^{reset}) As

$$\rho_k = \frac{\left\langle \Delta m^2_{v,k} \right\rangle}{\langle m_{v,k} \rangle - \mu_v}$$

where μ_v is the mean number of mutations in the week in which $\left\langle \Delta m^2_{v,k} \right\rangle$ is minimal. This condition corresponded to the first appearance of the variant of study in the population for Primal (first instance 26 Jan 2020), Alpha (first instance 25 Oct 2020), and Omicron BA.1 (first instance 21 Nov 2021), according to our functional dataset. In the case of Delta, however, the date of minimal variance did not coincide with the date of first instance, but rather with the moment at which the AY.4 lineage became dominant (16 May 2021) after a first period of time of selection within the Delta population, so this date was used for the mutation count reset.

Global *vs.* variant-based analysis

The global analysis was carried out considering all available sequences from our functional dataset (*i.e.*, pooling together the sequences even if they corresponded to different variants to compute the mean and variance for each week). This analysis served to appreciate the overall evolutionary trajectory by which the observable viral genome accumulates mutations with time in UK. That is, it allowed having a bird's eye perspective. More in detail, we could calculate a macroscopic evolution rate by linear regression between mean and time, and we could evaluate the dispersion dynamics of the resulting mutation distribution by representing the variance/mean ratio with time.

By contrast, the variant-based analysis was carried out considering only the sequences corresponding to a given variant, according to the annotation. This

was done for Primal, Alpha, Delta, and Omicron (distinguishing also between the BA.1 and BA.2 lineages). Our analysis showed alternation of periods of evolution at lower rates and bursts of dispersion due to invasion events. We acknowledge that previous work already showed changes in the evolution rate with time in the particular case of SARS-CoV-2 [21, 35]. However, because our study was not based on phylogeny, we were able to process all available sequences in the database, gaining accuracy. In addition, and most importantly, because the dynamic profile of the variance was also analyzed, we were able to disclose anomalous diffusion patterns. This is a notable result that may contribute to change our understanding of virus evolution.

Comparison with phylogenetic methods

While in this work we are interested in how a viral population evolves, phylogenetic approaches mainly focus on genotypic differences in order to reconstruct evolutionary paths. Phylogenetic methods have been applied to produce estimates of the SARS-CoV-2 evolution rate, reporting values of 0.3-0.4 wk^{-1} during the first year of pandemic [21, 36]. These values are in tune with our calculations in the case of Primal. This congruence suggested us the formation with time of a sufficiently heterogenous viral population. Indeed, using the metric of heterogeneity, once a variant was fixed, the divergence between two arbitrary sequences of the population was about 11-15 mutations.

Further phylogenetic inferences have pointed out a transient increase of the evolution rate concomitant with the emergence of new invading variants (e.g., Alpha) [35]. However, by restricting the study to the sequences within the clade, the evolution rate did not appear to increase but rather to be maintained or even reduced (e.g., in the cases of Alpha and Delta). According to our analysis and also others in the field following non-phylogenetic approaches [34], Alpha evolved a bit slower than Primal and Delta did at a similar rate. Despite the uncertainty associated with the emergence of new variants [27], viral population- based studies are useful to understand the mutation-selection dynamics.

Categorization of mutations

For each viral sequence present in the functional dataset, the set of substitutions with respect to root were broken down into several categories: non-coding substitutions (*i.e.*, substitutions that fall on non-coding regions of the genome), synonymous substitutions (*i.e.*, substitutions that fall on coding regions but do not trigger amino acid changes), and non-synonymous substitutions (*i.e.*, substitutions that trigger amino acid changes in coding regions). Insertions and deletions were counted into a unique variable called indels. Fi-

nally, the weekly mean and variance in terms of number of non-coding substitutions, synonymous substitutions, non-synonymous substitutions, and indels were computed.

Estimation of natural selection signatures

The ratio between the number of nonsynonymous and synonimous substitutions per site (dN/dS) was used to realize the sense of natural selection during SARS-CoV-2 evolution, as it is a suitable statistical parameter to estimate the balance between positive (adaptive), neutral, and negative (purifying) selection acting on a set of protein-coding genes [37]. A simple method was employed to estimate the dN/dS ratio for each viral sequence [38], assuming

i that the total length of the protein-coding genes was constant (equal to that of the reference genome)

ii that the four nucleotides had equal frequencies

iii that the substitution events were random

First, the total number of synonymous (S) and non-synonymous (N) sites was estimated. Given that the probability of maintaining the same amino acid sequence is 5% if the substitution occurs at the first position of the codon, 0% if it occurs at the second position, and 72% if it occurs at the third position, it turns out that $S \approx (0.05{+}0.72)R$ and $N \approx 3(R{-}S)$, where R is the total length in base pairs of the protein-coding genes. Then, the proportion of synonymous (p_S) and non-synonymous (p_N) substitutions per site were computed. Second, these proportions were corrected to account for multiple potential changes at the same site. The genetic distance of synonymous (d_S) and non-synonymous (d_N) substitutions per site was estimated using the Jukes-Cantor formula, that is, $d_S = -\frac{3}{4}\ln\left(1-\frac{4}{3}p_S\right)$ and $d_N = -\frac{3}{4}\ln\left(1-\frac{4}{3}p_N\right)$.

Mathematical modeling of evolutionary motion

Being $m(t)$ the number of accumulated mutations in the viral genome at time t, the stochastic differential equation $\frac{dm(t)}{dt} = \kappa + \xi(t)$ governs the dynamics of the system, where κ is the evolution rate and $\xi(t)$ is an integrative noise source. The statistical properties of the noise source are $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = \frac{1}{2} D\alpha\left(\alpha-1\right)|t-t'|^{\alpha-2}$, which corresponds to a scenario of fractional Brownian motion [39]. In this formulation, D is the diffusion coefficient and α the diffusion exponent. On the one hand, the solution for the mean evolutionary motion is $\langle m(t) \rangle = \kappa t$, which is compatible to the molecular clock hypothesis [3]. Linear regressions were performed between the calculated mean number of

mutations in the viral sequences and time using the LinearRegression function from scikit-learn. This was done for the global data and also for the major lineages Primal, Alpha, Delta, and Omicron (BA.1 and BA.2).

On the other hand, the solution for the variance is $\langle \Delta m(t)^2 \rangle = Dt^\alpha$. However, this solution comes from a fixed initial condition (*i.e.*, no variability at $t = 0$). The calculated variance from the viral sequences was fitted to the general expression $\langle \Delta m(t)^2 \rangle = \sigma_0^2 + Dt^\alpha$, where σ_0^2 is a parameter that accounts for the initial variance in the population (σ_0^2 was directly computed from the set of sequences). Nonlinear regressions were performed between $\langle \Delta m(t)^2 \rangle - \sigma_0^2$ and time using the `curve_fit` function from SciPy. This was done for the global data and also for the major lineages Primal, Alpha, Delta, and Omicron (BA.1). In the case of Delta, the variance analysis was restricted to the AY.4 sublineage, which was the dominant in UK after a first period of time in which other sublineages coexisted (the fixation of the AY.4 sublineage led to a decrease in variance).

Brownian *vs.* non-Brownian motion

A Brownian motion is a continuous-space and continuous-time model to describe the stochastic movement of a free particle. Here, we considered a viral particle that moves in the space of sequences. If we denote by $\Delta m(t)$ the deviation from the mean behavior in terms of number of mutations, *i.e.*, $\Delta m(t) = m(t) - \kappa t$, we can write

$$\frac{d\Delta m(t)}{dt} = \xi(t)$$

For $\Delta m(t)$ to be a Brownian motion, it must have a null mean displacement, *i.e.*, $\langle \Delta m(t) \rangle = 0$, and a mean squared displacement (*i.e.*, variance) proportional to time, *i.e.*, $\langle \Delta m(t)^2 \rangle \propto t$. A null mean displacement generally holds because typical noise sources obey $\langle \xi(t) \rangle = 0$. In a scenario in which the mean squared displacement does not scale linearly with time, let us say $\langle \Delta m(t)^2 \rangle \propto t^\alpha$ with $\alpha \neq 1$, the motion is said to be non-Brownian (the term anomalous diffusion is also used to refer to this scenario) [22].

The properties of the noise source $\xi(t)$ determine the type of stochastic movement. In the typical case of Gaussian white noise, *i.e.*, $\langle \xi(t) \rangle = 0$ and $\langle \xi(t)\xi(t') \rangle = D\delta(t - t')$, where $\delta(t)$ is the Dirac delta function, it turns out that

$$\begin{aligned}\left\langle \Delta m(t)^2 \right\rangle &= \left\langle \int_0^t \int_0^t \xi(r)\xi(s) dr ds \right\rangle \\ &= \int_0^t \int_0^t \left\langle \xi(r)\xi(s) \right\rangle dr ds \\ &= D \int_0^t \int_0^t \left\langle \delta(r-s) \right\rangle dr ds \\ &= Dt\end{aligned}$$

However, considering $\langle \xi(t)\xi(t') \rangle = \frac{1}{2} D\alpha(\alpha-1) \left| t-t' \right|^{\alpha-2}$, it turns out that

$$\begin{aligned}\left\langle \Delta m(t)^2 \right\rangle &= \left\langle \int_0^t \int_0^t \xi(r)\xi(s) dr ds \right\rangle \\ &= \int_0^t \int_0^t \frac{1}{2} D\alpha(\alpha-1) \left| t-s \right|^{\alpha-2} dr ds \\ &= \frac{1}{2} D\alpha(\alpha-1) \int_0^t \left[\int_0^s (s-r)^{\alpha-2} dr + \int_s^t (r-s)^{\alpha-2} dr \right] ds \\ &= \frac{1}{2} D\alpha \int_0^t \left[s^{\alpha-1} + (t-s)^{\alpha-1} \right] ds \\ &= Dt^{\alpha}\end{aligned}$$

Therefore, the analysis of $\left\langle \Delta m(t)^2 \right\rangle$ with time is instrumental to assess the nature of the stochastic movement. Previous evolutionary studies of viruses mainly focused on the mean behavior [8, 21, 34, 35, 36], i.e., evaluating the relationship $\langle m(t) \rangle = \kappa t$, so our study is pertinent due to the completeness achieved.

Statistical significance of the diffusion parameters

Bootstrapping was applied to assess the robustness of the estimations of D and α. This approach resamples the original dataset with replacement to generate new bootstrap datasets. We can then fit the same model to each of these bootstrap datasets, obtaining a distribution of model parameters. This distribution can be used to estimate the variability of the model parameters and to evaluate the robustness of the model to small changes in the original dataset [40]. In this work, a random sampling with replacement of the sequences was performed each week. The sampling size was defined as the 50% of the total number of sequences available in each week in the original dataset (*i.e.*, if there are 100 sequences in a week, the bootstrap sample size for that week is 50, what is called subsampling). This was done for all 120 weeks in an

independent manner. We chose a sample size that was large enough to capture the key characteristics of the original dataset, but small enough to make the bootstrap procedure computationally feasible and robust to observations with a disproportionate impact on the results. With the new bootstrap dataset, the mean and variance were calculated. This procedure was repeated 1000 times. As a result, a distribution of values for each diffusion parameter was obtained, having performed 1000 independent regressions. This was done for the major variants Primal, Alpha, Delta, and Omicron (BA.1).

References

[1] Koonin E, Dolja V (2013) A virocentric perspective on the evolution of life. *Curr Opin Virol*, 3: 536–557.

[2] Drake J, Charlesworth B, Charlesworth D, Crow J (1998) Rates of spontaneous mutation. *Genetics*, 148: 1667–1686.

[3] Ayala F (1999) Molecular clock mirages. *BioEssays*, 21: 71–75.

[4] Kumar S (2005) Molecular clocks: four decades of evolution. *Nat Rev Genet*, 6: 654–662.

[5] Kimura M (1987) Molecular evolutionary clock and the neutral theory. *J Mol Evol*, 26: 24–33.

[6] Gojobori T, Moriyama E, Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. *Proc Natl Acad Sci USA*, 87: 10015–10018.

[7] Leitner T, Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci USA*, 96: 10752–10757.

[8] Jenkins G, Rambaut A, Pybus O, Holmes E (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol*, 54: 156–165.

[9] Bedford T, Wapinski I, Hartl D (2008) Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. *Genetics*, 179: 977–984.

[10] Li J, Lai S, Gao G, Shi W (2021) The emergence, genomic diversity and global spread of SARS-CoV-2. *Nature*, 600: 408–418.

[11] Bedford T, *et al.* (2020) Cryptic transmission of SARS-CoV-2 in Washington state. *Science*, 370: 571–575.

[12] López M, *et al.* (2021) The first wave of the COVID-19 epidemic in Spain was associated with early introductions and fast spread of a dominating genetic variant. *Nat Genet*, 53: 1405–1414.

[13] Lemieux J, *et al.* (2021) Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science*, 371: eabe3261.

[14] Tegally H, *et al.* (2021) Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592: 438–443.

[15] Kraemer M, *et al.* (2021) Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science*, 373: 889–895.

[16] Rockett R, *et al.* (2022) Co-infection with SARS-CoV-2 Omicron and Delta variants revealed by genomic surveillance. *Nat Commun*, 13: 2745.

[17] Jackson B, *et al.* (2021) Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell*, 184: 5179–5188.

[18] Rochman N, *et al.* (2021) Ongoing global and regional adaptive evolution of SARS-CoV-2. *Proc Natl Acad Sci USA*, 118: e2104241118.

[19] Yi K, *et al.* (2021) Mutational spectrum of SARS-CoV-2 during the global pandemic. *Exp Mol Med*, 53: 1229–1237.

[20] de Silva V, Tenenbaum J, Global versus local methods in nonlinear dimensionality reduction. In: *Advances in Neural Information Processing Systems* (2003), 721–728.

[21] Ghafari M, *et al.* (2022) Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. *Mol Biol Evol*, 39: msac009.

[22] Manzo C, Garcia-Parajo M (2015) A review of progress in single particle tracking: from methods to biophysical insights. *Rep Prog Phys*, 78: 124601.

[23] De Maio N, *et al.* (2021) Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol Evol*, 13: evab087.

[24] Christie S, *et al.* (2022) Single-virus tracking reveals variant SARS-CoV-2 spike proteins induce ACE2-independent membrane interactions. *Sci Adv*, 8: eabo3977.

[25] Kreutzberger A, *et al.* (2022) SARS-CoV-2 requires acidic pH to infect cells. *Proc Natl Acad Sci USA*, 119: e2209514119.

[26] Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on Fast Fourier transform. *Nucleic Acids Res*, 30: 3059–3066.

[27] Hill V, *et al.* (2022) The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol*, 8: veac080.

[28] Michaelsen T, *et al.* (2022) Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark. *Genome Med*, 14: 47.

[29] Rosner B (1983) Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*, 25: 165–172.

[30] Mead A (1992) Review of the development of multidimensional scaling methods. *J R Stat Soc D*, 41: 27–39.

[31] Jolliffe I, Cadima J (2016) Principal component analysis: A review and recent developments. *Philos Trans R A*, 374: 20150202.

[32] da Costa C, de Freitas C, Alves C, Lameira J (2022) Assessment of mutations on RBD in the Spike protein of SARS-CoV-2 Alpha, Delta and Omicron variants. *Sci Rep*, 12: 8540.

[33] Kumar S, Karuppanan K, Subramaniam G (2022) Omicron (BA.1) and sub-variants (BA.1.1, BA.2, and BA.3) of SARS-CoV-2 spike infectivity and pathogenicity: A comparative sequence and structural-based computational assessment. *J Med Virol*, 94: 4780–4791.

[34] Neher R (2022) Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. *Virus Evol*, 8: veac113.

[35] Tay J, Porter A, Wirth W, Duchene S (2022) The emergence of SARS-CoV-2 variants of concern is driven by acceleration of the substitution rate. *Mol Biol Evol*, 39: msac013.

[36] Wang S, *et al.* (2022) Molecular evolutionary characteristics of SARS-CoV-2 emerging in the United States. *Med Virol*, 94: 310–317.

[37] Kryazhimskiy S, Plotkin J (2008) The population genetics of dN/dS. *PLoS Genet*, 4: e1000304.

[38] Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3: 418–426.

[39] Kursawe J, Schulz J, Metzler R (2013) Transient aging in fractional Brownian and Langevin-equation motion. *Phys Rev E*, 88: 062124.

[40] Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Statist*, 7: 1–26.

Chapter 3

Paper 2

cita interesante 3.
- El pavo al que cito

Ellipsis.

References

Chapter 4

Paper 3

cita interesante 4.
- El pavo al que cito

Ellipsis

References

Chapter 5

General conclusions

cita interesante 5.
- El pavo al que cito

Ellipsis

Acknowledgements

Ellipsis