PNAS BRIEF REPORT | EVOLUTION OPEN ACCESS

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

Lucas Goiriz[a,b], Raúl Ruiz[a], Òscar Garibo-i-Orts[b], J. Alberto Conejero[b], and Guillermo Rodrigo[a,1]

Edited by Eugene Koonin, NIH, Bethesda, MD; received March 20, 2023; accepted June 11, 2023

The evolution of severe acute respiratory syndrome coronavirus 2 (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 a 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 wk of the pandemic in the United Kingdom. In particular, we found subdiffusion patterns for the Primal, Alpha, and Omicron variants but a weak superdiffusion pattern for the Delta variant. Our findings indicate that non-Brownian evolutionary motions occur in nature, thereby providing insight for viral phylodynamics.

anomalous diffusion | dynamic systems | virus evolution rate | stochastic process | whole genome sequencing

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). Although 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 the pandemic (11–13). To date, more than 10 million sequences are available. In this work, we conducted a study aimed at describing the evolution of a human virus in nature as a (non-)Brownian motion, considering the number of acquired mutations as the displacement of the viral particle from the origin (Fig. 1*A*). 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 theoretical foundations for viral evolution.

Results and Discussion

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

Author affiliations: [a]BioInstituto de Biología Integrativa de Sistemas, Consejo Superior de Investigaciones Científicas – Universitat de València, 46980 Paterna, Spain; and [b]Institut Universitari de Matemàtica Pura i Aplicada, Universitat Politècnica de València, 46022 Valencia, Spain

Author contributions: G.R. designed the research; L.G. performed the research with the help of Ò.G.O. under the supervision of G.R.; L.G., R.R., J.A.C., and G.R. analyzed the data; and G.R. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: guillermo.rodrigo@csic.es.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2303578120/-/DCSupplemental.

Published July 17, 2023.

$\xi(t)$ an integrative noise source whose properties shape the evolutionary motion.

Due to the large number of available SARS-CoV-2 sequences from the United Kingdom, our analysis was focused on this country. Using landmark multidimensional scaling (14), we obtained a representation of all available genotypes in a two-dimensional space (Fig. 1*B*), 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. 1*C*). 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 a 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., noncoding, synonymous, nonsynonymous, and indels), and the ratio between the number of nonsynonymous and synonymous substitutions per site (dN/dS) was estimated (Fig. 1*D*). The dN/dS signature (fluctuation around 1 over time) suggested evolution under purifying selection of a series of adapted variants.

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. 1*E*). 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 variants 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. 1*F*). The

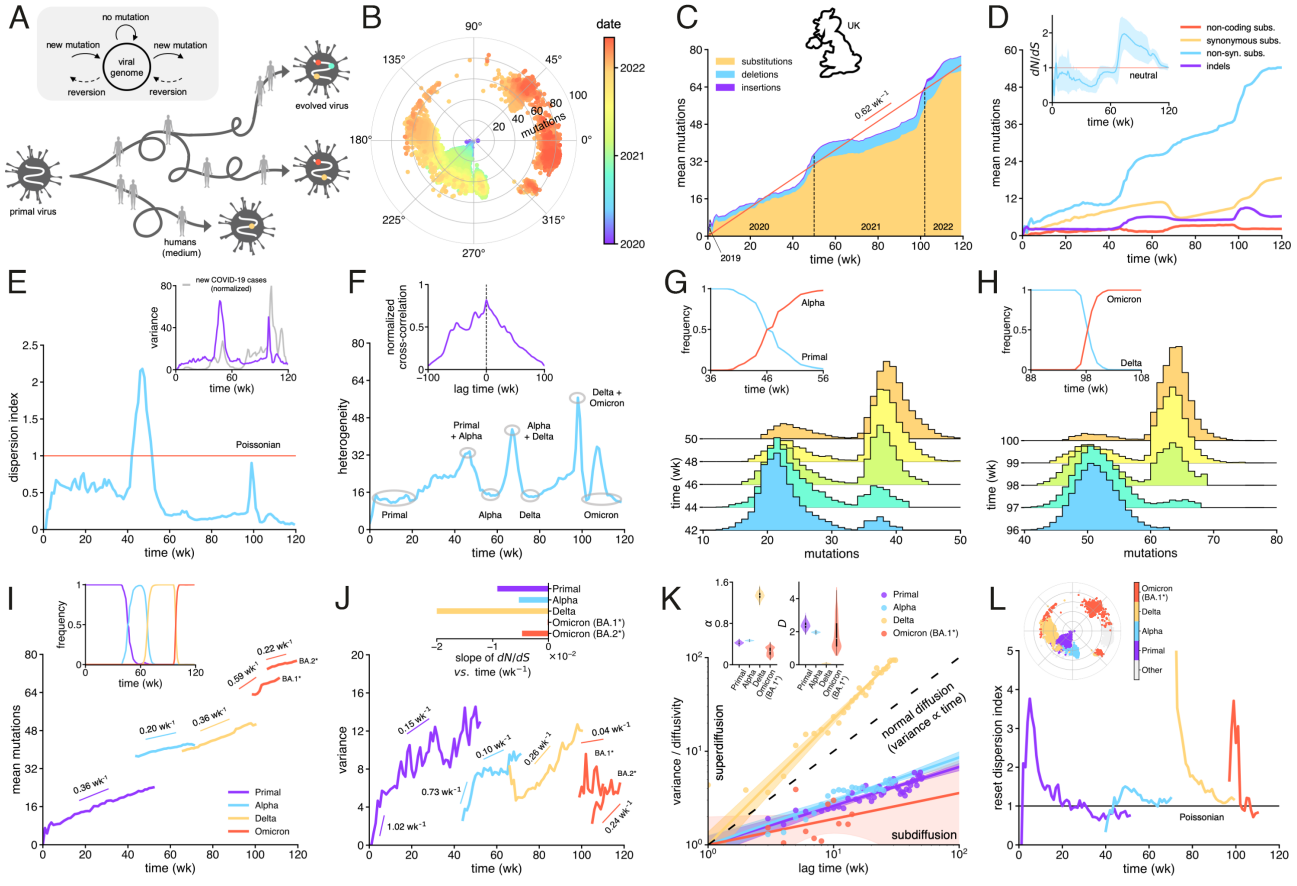

Fig. 1. Characterization of SARS-CoV-2 evolution in the United Kingdom. (*A*) Schematics of the evolutionary motion of the virus (viewed as a stochastic process). *Inset*: Associated state-transition diagram. (*B*) 2D projection of all viral sequences colored by date. (*C*) Time course of the mean number of accumulated mutations in the viral genome, distinguishing between substitutions, deletions, and insertions. Linear regression over the total shown in red ($R^2 = 0.95$). (*D*) Time course of the mean number of noncoding substitutions, synonymous substitutions, nonsynonymous substitutions, and indels. *Inset*: dN/dS with time (mean ± SD). (*E*) Time course of the dispersion index (variance/mean). *Inset*: Variance with time; the normalized number of new COVID-19 cases is superimposed. (*F*) 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. (*G* and *H*) Probability-based histograms of the number of accumulated mutations during the transition from Primal to Alpha (*G*) and Delta to Omicron (*H*). *Insets*: Population frequency of the variants with time. (*I*) 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. (*J*) 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. (*K*) Representation of the rescaled variance normalized by the diffusivity (D) with time in the 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% CIs of the regression lines. *Inset*: Distributions of values for each diffusion parameter (violin plots) obtained by bootstrapping. (*L*) Time course of the reset dispersion index per variant. *Inset*: 2D projection of all viral sequences colored by variant (projection as in panel *B*). The variant-specific analyses were restricted to the time period in which their population frequency was at least 10%.

number of new coronavirus disease 2019 (COVID-19) cases also correlated with the variance (Fig. 1*E*, *Inset*). Representing the distributions of accumulated mutations with time, we disclosed a bimodal behavior during such transitions (Fig. 1 *G* and *H*), explaining the increased dispersion. The invading genotypes carried about 15 to 20 more mutations on average. Moreover, the transition from Alpha to Delta generated only 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 (13).

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 of 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. 1*I*), in agreement with previous estimates following phylogenetic methods (15). 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 the United Kingdom. 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).

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. 1*J*), which indicated a stochastic behavior with anomalous diffusion (16). 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 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 $\langle \xi(t)\xi(t')\rangle = \frac{1}{2}D\alpha(\alpha-1)|t-t'|^{\alpha-2}$ 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. 1*K*). 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 (Fig. 1*K*, *Inset*).

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 (Fig. 1*J*, *Inset*). 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 (17), the evolved genotypes of Primal, Alpha, and Omicron BA.1 would be more constrained as a result of their nonsynonymous 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. 1*L*).

The observation of patterns of anomalous diffusion in biology has opened new avenues of research (16). Intriguingly, recent studies in which the physical movement of single SARS-CoV-2 virions was monitored throughout the infectious cycle highlighted transient and variant-dependent directionality and confinement outside and inside the cell (18, 19), indicating deviation from a pure Brownian motion. Here, we have presented an application domain in evolution. 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.

Methods

Whole genome sequencing data were retrieved from the Global Initiative on Sharing All Influenza Data (GISAID). All top–down biostatistical and bioinformatic analyses were carried out in Python. Detailed descriptions are provided in *SI Appendix*.

Data, Materials, and Software Availability. All study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We thank R. Montagud-Martínez and N. Firbas for useful discussions. Work supported by CSIC PTI Global Health (SGL2021-03-040) through the NextGenerationEU Fund (reg. 2020/2094) and Generalitat Valenciana (ACIF/2021/183, GVA-COVID19/2021/100, and GVA-COVID19/2021/036). Funding for open access by Universitat Politècnica de València.

1. E. V. Koonin, V. V. Dolja, A virocentric perspective on the evolution of life. *Curr. Opin. Virol.* **3**, 546–557 (2013).
2. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).
3. F. J. Ayala, Molecular clock mirages. *BioEssays* **21**, 71–75 (1999).
4. S. Kumar, Molecular clocks: Four decades of evolution. *Nat. Rev. Genet.* **6**, 654–662 (2005).
5. M. Kimura, Molecular evolutionary clock and the neutral theory. *J. Mol. Evol.* **26**, 24–33 (1987).
6. T. Gojobori, E. N. Moriyama, M. Kimura, Molecular clock of viral evolution, and the neutral theory. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 10015–10018 (1990).
7. T. Leitner, J. Albert, The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 10752–10757 (1999).
8. G. M. Jenkins, A. Rambaut, O. G. Pybus, E. C. Holmes, Rates of molecular evolution in RNA viruses: A quantitative phylogenetic analysis. *J. Mol. Evol.* **54**, 156–165 (2002).
9. T. Bedford, I. Wapinski, D. L. Hartl, Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. *Genetics* **179**, 977–984 (2008).
10. J. Li, S. Lai, G. F. Gao, W. Shi, The emergence, genomic diversity and global spread of SARS-CoV-2. *Nature* **600**, 408–418 (2021).
11. T. Bedford *et al.*, Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020).
12. M. G. López *et al.*, 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 (2021).
13. M. U. Kraemer *et al.*, Spatiotemporal invasion dynamics of SARS-CoV-2 lineage B.1.1.7 emergence. *Science* **373**, 889–895 (2021).
14. V. de Silva, J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction" in *Advances in Neural Information Processing Systems 15*, S. Becker *et al.*, Eds. (MIT Press, 2003), pp. 721–728.
15. M. Ghafari *et al.*, Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. *Mol. Biol. Evol.* **39**, msac009 (2022).
16. C. Manzo, M. F. Garcia-Parajo, A review of progress in single particle tracking: From methods to biophysical insights. *Rep. Prog. Phys.* **78**, 124601 (2015).
17. N. De Maio *et al.*, Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, evab087 (2021).
18. S. M. Christie *et al.*, Single-virus tracking reveals variant SARS-CoV-2 spike proteins induce ACE2-independent membrane interactions. *Sci. Adv.* **8**, eabo3977 (2022).
19. A. J. Kreutzberger *et al.*, SARS-CoV-2 requires acidic pH to infect cells. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2209514119 (2022).