PNAS INAUGURAL ARTICLE | MICROBIOLOGY OPEN ACCESS

Incipient functional SARS-CoV-2 diversification identified through neural network haplotype maps

Soledad Delgado[a,1], Pilar Somovilla[b,c], Cristina Ferrer-Orta[d], Brenda Martínez-González[e,f], Sergi Vázquez-Monteagudo[d], Javier Muñoz-Flores[g], María Eugenia Soria[b,f], Carlos García-Crespo[b], Ana Isabel de Ávila[b], Antoni Durán-Pastor[e], Ignacio Gadea[f], Cecilio López-Galíndez[h], Federico Moran[i], Ramon Lorenzo-Redondo[j], Nuria Verdaguer[d], Celia Perales[e,f,1], and Esteban Domingo[b,1]

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2020.
Contributed by Esteban Domingo; received October 16, 2023; accepted January 8, 2024; reviewed by Karla Kirkegaard and Peter F. Stadler

Since its introduction in the human population, SARS-CoV-2 has evolved into multiple clades, but the events in its intrahost diversification are not well understood. Here, we compare three-dimensional (3D) self-organized neural haplotype maps (SOMs) of SARS-CoV-2 from thirty individual nasopharyngeal diagnostic samples obtained within a 19-day interval in Madrid (Spain), at the time of transition between clades 19 and 20. SOMs have been trained with the haplotype repertoire present in the mutant spectra of the nsp12- and spike (S)-coding regions. Each SOM consisted of a dominant neuron (displaying the maximum frequency), surrounded by a low-frequency neuron cloud. The sequence of the master (dominant) neuron was either identical to that of the reference Wuhan-Hu-1 genome or differed from it at one nucleotide position. Six different deviant haplotype sequences were identified among the master neurons. Some of the substitutions in the neural clouds affected critical sites of the nsp12-nsp8-nsp7 polymerase complex and resulted in altered kinetics of RNA synthesis in an in vitro primer extension assay. Thus, the analysis has identified mutations that are relevant to modification of viral RNA synthesis, present in the mutant clouds of SARS-CoV-2 quasispecies. These mutations most likely occurred during intrahost diversification in several COVID-19 patients, during an initial stage of the pandemic, and within a brief time period.

COVID-19 | clade transition | viral quasispecies | self-organized maps | intrahost evolution

Significance

The study establishes haplotype self-organized maps (SOMs) as a means to dissect the composition of complex viral mutant spectra and to identify low-frequency haplotypes. In combination with biochemical assays, the procedure provides a means to detect in infected patients functionally relevant mutations before the latter become dominant, and are recorded in virus data banks. For SARS-CoV-2, the SOM analysis has documented minority haplotypes whose corresponding encoded viral RNA-dependent RNA polymerase exhibits differences in viral RNA synthesis.

The initial expectation was that the genetic variability of SARS-CoV-2 might be limited due to its genome encoding an error-correcting exonuclease (Exo N) activity (1–5), and to the restricted mutation tolerance expected for large RNA genomes (6). Limited intrahost diversity is still portrayed as a feature of SARS-CoV-2 (7, 8). Yet, several studies have reported remarkable genetic heterogeneity in samples of this virus and of other coronaviruses due to point mutations and deletions (9–25). Recently, a cutoff mutation and deletion frequency of 0.1% was attained in ultradeep sequencing analysis of the mutant spectra of SARS-CoV-2 present in nasopharyngeal samples of patients diagnosed with COVID-19 (26). The attainment of this resolution was facilitated by the combined result of high clean read coverage using the MiSeq Illumina platform for amplicons of the nsp12- and S-coding region (26) and application of SeekDeep bioinformatics pipeline (27) for data processing; several experimental and bioinformatics controls established the reliability of such mutation and deletion identification level (26) (see also *Materials and Methods*). With this increased sensitivity, the mutant spectrum composition of six amplicons from virus isolated from 30 patients was determined. The number of different mutations detected with the 0.1% mutation frequency cutoff was 50- to 100-fold larger than the number identified with a 0.5% mutation frequency cutoff. The number of variable sites amounted to 38% of the genomic positions analyzed (26) which is a remarkable intrahost SARS-CoV-2 genetic heterogeneity whose biological significance begs being investigated.

A mutation rate of around 10^{-6} mutations introduced per nucleotide and infection cycle for SARS-CoV-2 has been inferred from mutation frequencies (28). Such mutation rate for an RNA virus—which is congruous with an active ExoN-mediated proofreading (29)—is not incompatible either with a heterogeneous mutant spectrum within a viral isolate or with rapid virus evolution in the field. Intrahost heterogeneity and rate of evolution are both influenced by the number of rounds of productive infection in each infected individual, negative and positive selection (at any stage during and after genome replication), and number of transmission events that involve population bottlenecks, in addition to the virus mutation rate (30, 31).

Author contributions: S.D., N.V., C.P., and E.D. designed research; S.D., C.F.-O., B.M.-G., S.V.-M., and M.E.S. performed research; S.D., F.M., R.L.-R., N.V., C.P., and E.D. contributed new reagents/analytic tools; S.D., P.S., C.F.-O., B.M.-G., S.V.-M., J.M.-F., M.E.S., C.G.-C., A.I.d.Á., A.D.-P., I.G., C.L.-G, F.M., R.L.-R., N.V., C.P., and E.D. analyzed data; and S.D., N.V., C.P., and E.D. wrote the paper.

Reviewers: K.K., Stanford University; and P.F.S., University of Leipzig.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: mariasoledad.delgado@upm.es, celia.perales@cnb.csic.es, or edomingo@cbm.csic.es.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2317851121/-/DCSupplemental.

Published February 28, 2024.

The question we address in the present study is whether low-frequency mutant genomes that arise in infected individuals may play a significant role in intrahost viral genetic and functional diversification. The response to this question requires penetration into the mutant spectrum composition of closely related isolates within the same disease episode, in search of the critical, inceptive events that lead to the renewal of dominant genomic sequences, with their corresponding mutant clouds. Information is needed on haplotype composition, abundance, modifications, as well as similarities and differences among haplotype clouds in virus from different patients. This information has to be complemented by an evaluation of the structural impact of the observed amino acid substitutions and biochemical assays to test possible functional differences.

To these aims, we have generated three-dimensional (3D) self-organized neural network maps (SOMs) (32, 33) of viral haplotypes within individual isolates following an approach that was previously applied to the analysis of HIV type 1, hepatitis C virus, and viroid populations (34–36). The comparison has been carried out with diagnostic nasopharyngeal samples from thirty patients who were infected during the first COVID-19 wave. The patients were admitted to the Fundación Jiménez Díaz Hospital (Madrid, Spain) within a 19-d interval in April 2020, when a strict confinement of the Spanish population had been implemented (strict lockdown measures in Spain starting on March 14, 2020, published in the Royal Decree 463/2020). This was at the time of transition between the original clades 19A and 19B and the following clades 20A, 20B, and 20C. Virus from clade 20 acquired the spike D614G substitution and replaced almost entirely the clade 19 viruses (37).

Six composite SOMs have been obtained for a total of four amplicons of the nsp12 (polymerase)-coding region and two amplicons of the S-coding region of SARS-CoV-2. Seven SOM classes were distinguished according to the composition of the master (dominant) neuron in relation to the initial Wuhan-Hu-1 genome. Functional diversification was evidenced by the presence in the neural cloud of virus from several patients of amino acid substitutions in nsp12 that altered the kinetics of RNA synthesis, catalyzed by the nsp12-nsp8-nsp7 polymerase complex. The results identify intrahost genetic and functional diversification of SARS-CoV-2, emphasizing the value of intrahost mutation repertoire analyses.

Results

Clade Identification through SARS-CoV-2 Consensus Genomic Sequences. Nasopharyngeal swabs from thirty COVID-19 patients were collected from April 3 until April 22, 2020, in Madrid (Spain). The time of sample collection, patient demographics, risk factors for COVID-19, clinical profile at the time of diagnosis, and cycle threshold (C_T) of sample RNA (22, 38) are described in *SI Appendix*, Table S1 (https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). The isolates were divided into mild (n = 10), moderate (n = 10), and severe (exitus) (n = 10), according to the severity of associated COVID-19. The C_T values for the 30 isolates were the following: 25.3 ± 3.9 (range 18.4 to 28.5) for mild, 21.8 ± 2.4 (range 18.3 to 25.8) for moderate, and 20.4 ± 2.9 (range 15.6 to 25.4) for exitus. They were assigned to clade 19 or 20, based on the deduced consensus sequence of the nsp12- and S-coding regions (22, 26), and the entire genome sequence of 10 of the isolates (Fig. 1 and *SI Appendix*, Table S2 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). Thus, the isolates correspond to the transition between clades 19 and 20 of the early phase of the COVID-19 pandemic.

Construction of Self-organized Neuron Network Maps (SOMs) for SARS-CoV-2 Mutant Spectra. The ultradeep sequencing data covered four amplicons of the nsp12-coding region (amplicons A1 to A4, spanning nucleotides 14,534 to 16,054, that encode amino acids 366 to 871 of nsp12) and two amplicons of the S-coding region (amplicons A5 and A6, spanning nucleotides 22,872 to 23,645, that encode amino acids 438 to 694 of S). Together, the two regions analyzed comprise 7.8% of the SARS-CoV-2 genomic residues and include major functional domains of nsp12 (polymerase) and S (*SI Appendix*, Fig. S1 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). Residues are numbered according to the SARS-CoV-2 reference Wuhan-Hu-1 genome (NCBI accession number NC_045512.2).

Point mutations and deletions were counted relative to the Wuhan-Hu-1 sequence; haplotypes were weighted according to their frequency. The 3D irregular codification (35) was used to transform the nucleotide sequences into numerical vectors of the same size and to weigh the mutations according to their type. Starting from a SOM with the neurons arranged in a 2D grid, the training algorithm used the codified sequences to iteratively modify the prototype vectors of the neurons to fit them to the input data space. The grid size, and therefore the number of neurons of the SOM, was chosen for each amplicon (procedure detailed in *Materials and Methods*). Master neuron refers to the neuron with the largest haplotype frequency (visualized as the highest peak in the 3D SOM) for each amplicon and patient (abbreviated as amplicon-patient).

Patient-to-patient Variation of 3D Self-organized Maps. The 3D SOM pattern for amplicon-patient was comparable in that it included a master neuron with a single haplotype which was surrounded by a cloud of neurons of lower frequency, each with one or more haplotypes. The haplotype of the master neuron was either identical in sequence to the Wuhan-Hu-1 reference genome (termed SOM class Wu) or it was at a one nucleotide Hamming distance (the number of positions at which two nucleotide sequences of the same length differ) from it; the latter occurred in the haplotype of master neurons of several patients in four of the amplicons analyzed, corresponding to six out of the twelve SOM classes that have been distinguished (Table 1). Composite SOMs were displayed for all amplicon-patients that belong to the same SOM class, with the distribution of peak height abundance (Fig. 2). SOMs for the individual amplicon-patients on which the composite maps are based, are depicted in *SI Appendix*, Figs. S2–S7 (https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). In spike amplicon A6, the isolates were divided into those that have amino acid D614 as dominant (five isolates of clade 19) and those that have amino acid G614 as dominant (twenty-five isolates of clade 20). The cloud of minority neurons of four of the five clade 19 isolates includes the G614 variation (39). In six of the twelve composite maps, the distribution of peak abundances followed an exponential decay function, that reflected the abundance of low-frequency peaks. In the other cases, the data could not be adjusted to this type of function because peaks were spread over a larger peak height distribution (right panels in Fig. 2) (*Discussion*). The differences in minority peak distribution are particularly clear among spike amplicons of different SOM classes (Fig. 2), suggesting differences in replicative features revealed by SOMs.

While the master neurons included only one haplotype, a maximum of six different haplotypes was present in the neurons of the mutant cloud, with neurons that are either unique to one SOM class or that are shared with clouds of other SOM classes (*SI Appendix*, Fig. S8 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). Thus, different trajectories of intrahost SARS-CoV-2 diversification,

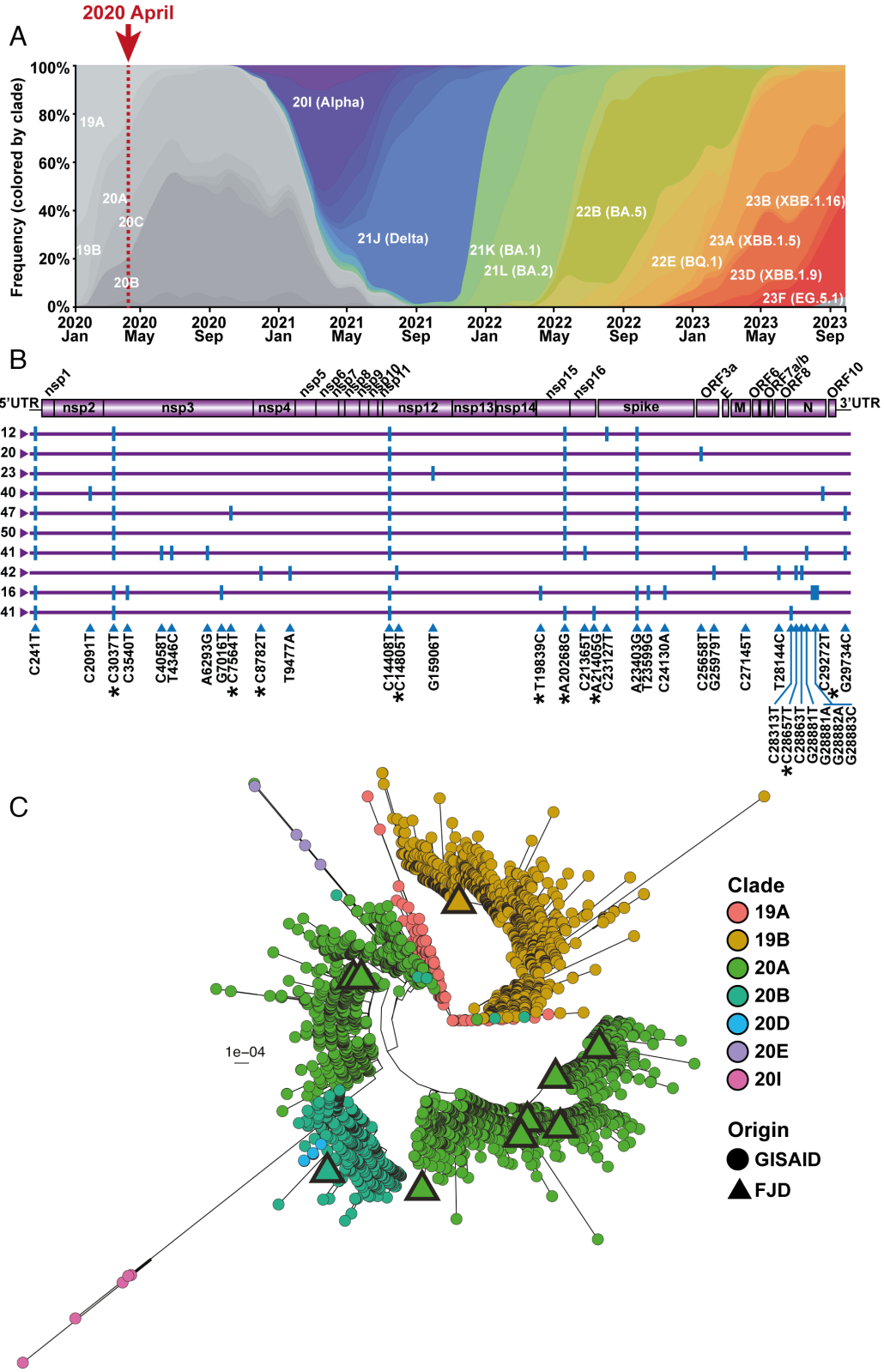

Fig. 1. Epidemiological context of the Madrid COVID-19 patient cohort and characterization of the SARS-CoV-2 genomes. (*A*) Successive COVID-19 waves in the world from January 2020 until September 2023 (source Nextstrain, https://nextstrain.org/ncov/gisaid/global). The vertical red arrow and discontinuous line indicate the time at which the samples for the present analysis were obtained (isolation dates, sample characteristics, and clinical profiles of the patients are summarized in *SI Appendix*, Table S1 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). (*B*) Scheme of the SARS-CoV-2 genome with the names of encoded proteins flanked by the 5′UTR and the 3′UTR (*Top*). The horizontal lines below the genome depict the consensus genomic nucleotide sequences of ten isolates from the Madrid cohort (patient code given on the left), with indication of mutations relative to the Wuhan-Hu-1 reference genome (*Bottom* of the alignment). Synonymous mutations are indicated by an asterisk. Mutation C28313T, which is indicated as nonsynonymous for the coding of protein N, is synonymous for ORF9b (amino acid P10). Mutations and amino acid substitutions are summarized in *SI Appendix*, Table S3 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw. (*C*) Maximum likelihood phylogeny tree of SARS-CoV-2 entire genome sequences, derived using IQ Tree v2.0.5. It includes 4,802 high-quality (high coverage and low N number) sequences from the GISAID database (https://gisaid.org), collected between January 1, 2020, and April 30, 2020 (circles), and 10 full genome consensus sequences of virus from the Madrid cohort (triangles; sequences displayed in part B). Clades are color coded as indicated on the right. The tree was rooted using the SARS-CoV-2 reference genome NC_045512.2.

Table 1. Haplotype composition of the master neuron in the nsp12 and spike amplicons that define seven SOM classes

Amplicon and SOM class[*]			Number of patients in which the viral haplotype occurred[‡]	
			Located in the master neuron[§]	Same location as the master neuron, but in the cloud[§]
nsp12 A1	Wu	Mutation (amino acid substitution) relative to the Wuhan-Hu-1 genome[†]	26[¶]	3[¶]
nsp12 A1	Wu-Mutant (1)	C14805T (Syn)	4[#]	1[#]
nsp12 A2	Wu	-	30[||]	0[||]
nsp12 A3	Wu	-	30[**]	0[**]
nsp12 A4	Wu	-	29[††]	1[††]
nsp12 A4	Wu-Mutant (2)	G15906T (Q822H)	1[‡‡]	0[‡‡]
spike A5	Wu	-	25[§§]	1[§§]
spike A5	Wu-Mutant (3)	C23127T (A522V)	1[¶¶]	0[¶¶]
spike A5	Wu-Mutant (4)	T23042C (S494P)	1[##]	5[##]
spike A6	Wu	-	4[***]	1[***]
spike A6	Wu-Mutant (5)	A23403G (D614G)	25[†††]	3[†††]
spike A6	Wu-Mutant (6)	C23380T (Syn)	1[‡‡‡]	1[‡‡‡]

[*]SARS-CoV-2 amplicons are those depicted in *SI Appendix*, Fig. S1 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw. SOM class refers to the distinction between SOMs whose haplotype in the master neuron had identical nucleotide sequence than the corresponding genomic region of the Wuhan-Hu-1 (Wu), versus those that differed from the Wuhan-Hu-1 sequence [Wu-Mutant (1), Wu-Mutant (2), Wu-Mutant (3), Wu-Mutant (4), Wu-Mutant (5), and Wu-Mutant (6)].
[†]The specific mutation that defines each SOM class. The sequence of the haplotypes is from residue 14,534 to 14,920; 14,911 to 15,299; 15,288 to 15693; and 15,669 to 16,054 for nsp12 A1, nsp12 A2, nsp12 A3, and nsp12 A4, respectively; from residue 22,872 to 23,268 and 23,259 to 23,645 for spike A5 and spike A6, respectively (numbering according to the Wuhan-Hu-1 genome, NCBI accession number NC_045512.2).
[‡]Occurrence in different patients of the Madrid cohort. Each master neuron includes a single haplotype (see footnote †).
[§]Number of patients in whose virus the haplotype of the SOM class indicated in the second column was present.
[¶]Found in all patients except Pt013. In Pt165, Pt168, and Pt342, the haplotype was in the neuron cloud.
[#]Found in patients Pt013, Pt165, Pt168, and Pt342. In Pt416, the haplotype was in the neuron cloud.
[||]Found in all patients.
[**]Found in all patients.
[††]Found in all patients.
[‡‡]Found in patient Pt023.
[§§]Found in all patients except Pt012 (no amplification was obtained for patients Pt036, Pt416, and Pt427; see *SI Appendix*, Fig. S6 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). In Pt417, the haplotype was in the neuron cloud.
[¶¶]Found in patient Pt012.
[##]Found in patients Pt130, Pt142, Pt150, Pt354, Pt417, and Pt426. Except for Pt417, in all other patients, the haplotype was in the neuron cloud.
[***]Found in patients Pt013, Pt106, Pt138, Pt165, and Pt342. For Pt106, the haplotype was in the neuron cloud.
[†††]Found in all patients, except Pt165 and Pt168. For Pt013, Pt138, and Pt342, the haplotype was in the neuron cloud. For Pt168, the haplotype was in the cloud but in a different neuron.
[‡‡‡]Found in Pt165 and Pt168. For Pt165, the haplotype was in the neuron cloud.
[§§§]The code of the patients, date of SARS-CoV-2 isolation, and clinical profiles are compiled in *SI Appendix*, Table S1 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw.

relative to the Wuhan-Hu-1 virus, were evidenced by distinct haplotype maps within a short time period and within a limited geographical area during an early stage of the COVID-19 pandemic.

Functional Modifications due to Amino Acid Differences Present in the Neuron Cloud of Several Isolates. The SOM analysis has revealed the presence of several minority amino acid substitutions in nsp12 (they are compiled in *SI Appendix*, Table S3 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). Amplicon A1 of the nsp12-coding region spans the critical fingers (F) domain of the viral polymerase (*SI Appendix*, Fig. S1 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). The amino acid substitutions in the mutant cloud of SOM class Wu-Mutant (1) (*SI Appendix*, Fig. S9 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw) are concentrated at the nsp12/nsp8 and nsp12/nsp7-nsp8 contact surfaces (Fig. 3*A*). In particular, substitutions F396L, S384P, L372P, V405A, and D545G are located close to the contact interface nsp12-nsp8$_F$ (Fig. 3*A*). We have previously shown that amino acid changes in this region, and in particular those that disrupted direct nsp12-nsp8$_F$ interactions or that weakened the hydrophobic contacts near the nsp12-nsp8$_F$ interface, including L372P, significantly reduced the RdRp activity of nsp12 (40). Additional nsp12 substitutions, whose position in the replication complex suggested that they could lead to alterations in RdRp activity, include T409A, V410A, D445G, and F441L located at the fingers close to the thumb subdomain, near the nsp12-nsp7 interface (Fig. 3*A*), and K430R, S433G, F419L, D421G, F422L, V424A, and S425P, also within the fingers, close to the thumb subdomain and near the nsp12-nsp8$_T$ interface (Fig. 3*A*).

Substitutions F419L, D421G, or F441L were selected for biochemical studies of RNA synthesis, both because of their location in the polymerase structure, and because they were present in the neuron clouds of all patients. We expressed and purified the RdRp complexes with wild-type nonstructural proteins nsp7, nsp8, complexed with nsp12 containing either F419L, D421G, or F441L. The polymerization activity of the wild-type and mutant polymerase complexes was measured in primer extension assays, using a fluorescently labeled RNA primer (20-mer) annealed to an unlabeled RNA template (28-mer; Fig. 3*B*), following previously described protocols (40, 41). The kinetic data show that substitutions F419L, D421G, and F441L clearly impaired the RNA synthesis activity of the polymerase complex at both 37 °C and 33 °C (Fig. 3 *C* and *D*). In particular, D421G gave rise to an almost inactive enzyme. None of the enzymes with these amino acid substitutions consumed more than 5% of the primer in the first minute, neither at 37°C nor at 33°C. Primer consumption at the end of the experiment (minute 60) was about 60 and 75% for the F441L and F419L, respectively (Fig. 3 *C* and *D*). These

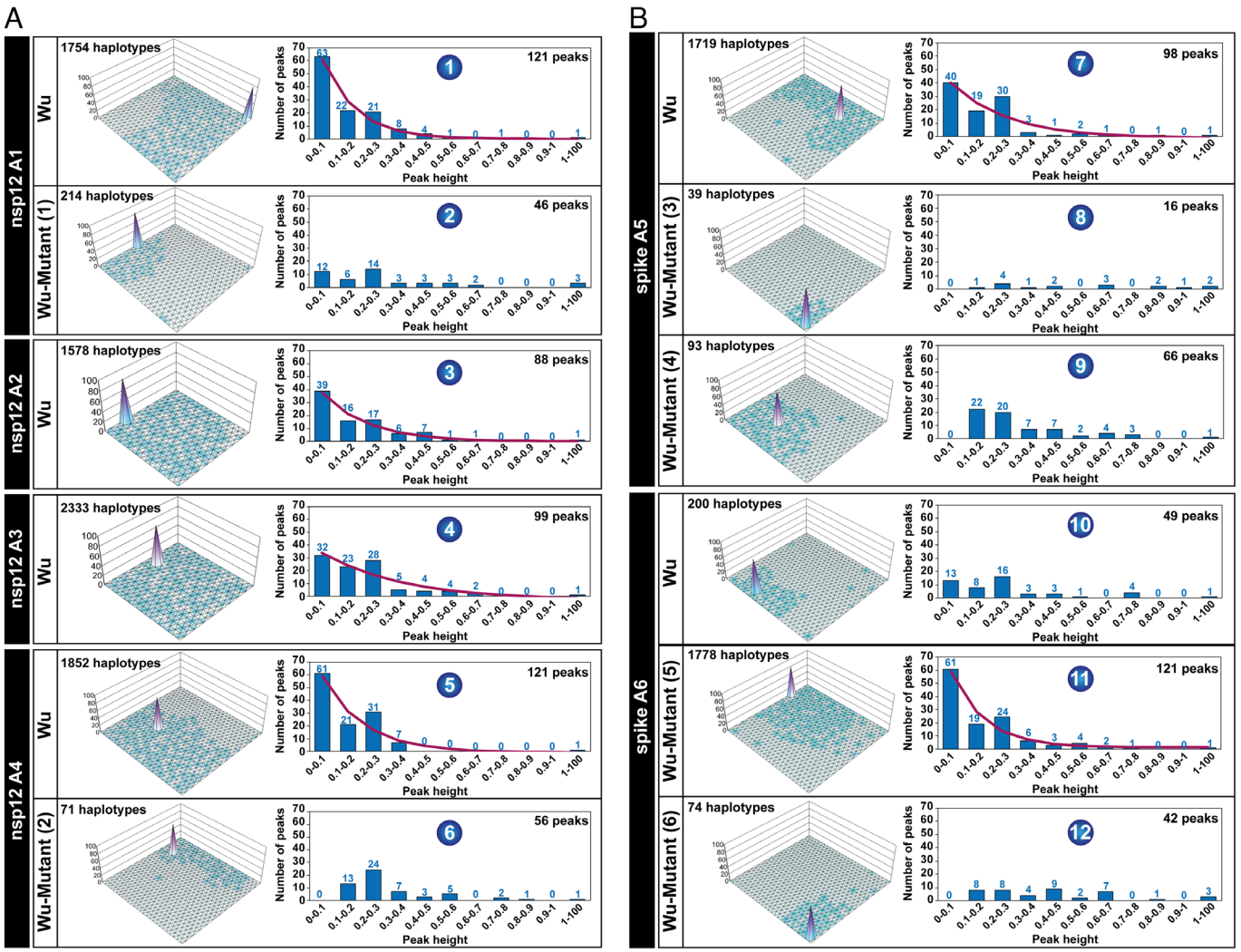

Fig. 2. Composite self-organized haplotype maps (SOMs) of amplicons of the nsp12-coding region (*A*) and of the spike (S)-coding region (*B*) of SARS-CoV-2 from the Madrid cohort. The coding region, amplicon number, and SOM class are indicated in the boxes at the *Left* of each panel. The number of haplotypes that compose each composite map and the number of neurons and the peak size distribution (plots on the *Right*) are shown. For plot numbers 1, 3, 4, 5, 7, and 11 (numbers encircled in each plot), the distributions could be fitted to an exponential decay function (purple lines) of the form $y = y_0 + A \times e^{(-x/t)}$ ($R^2 \geq 0.9$). However, for plot numbers 2, 8, and 10, the distributions could not fit this function optimally ($R^2 < 0.9$). For plot numbers 6, 9, and 12, the method did not converge. The SOMs for each individual patient are depicted in *SI Appendix*, Figs. S2–S7 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw).

two substitutions are found in the fingers, close to thumb subdomain. In the wild-type nsp12, the phenylalanine side chains participate in hydrophobic contacts that are established between these two subdomains (Fig. 3*A*). In particular, F441 plays an essential role as part of the fingertips, maintaining the closed right-hand conformation of the polymerase. The substitution of F by L at both positions would imply a weakening of these interactions. Considering the third substitution tested, amino acid D421 in the wild-type RdRp is salt bridged with the $nsp8_T$ K97 side chain. The D421G substitution results in the disruption of this electrostatic interaction, weakening the nsp12-$nsp8_T$ contacts (Fig. 3*A*).

Thus, nsp12 substitutions in the mutant cloud that marks a nascent diversification within infected individuals have functional consequences. A similar conclusion can be reached with the substitutions in S identified in neural clouds, in addition to the phenotypic impact of S substitution D614G that distinguishes clade 19 from clade 20 isolates (39) (*Discussion*).

In conclusion, 3D SOM analyses have detected differences in haplotype composition among amplicons from SARS-CoV-2 RNAs from individual patients who were infected at an early time point in the COVID-19 pandemic. Substitutions in nsp12 in intrahost neuron clouds of the virus affected the kinetics of in vitro RNA synthesis, suggesting functional flexibility embodied in the mutant spectrum of virus from individual infections.

Discussion

Viral quasispecies pose an important challenge concerning the characterization and biological significance of different genome subpopulations that coexist in a virus isolate. Several bioinformatics procedures have been developed to investigate the internal organization of complex mutant spectra (42–45). SARS-CoV-2 contributes to the challenge because isolates sampled from different cohorts are composed of mutant spectra (14–23). Reaching a detection limit of 0.1% mutation frequency cutoff revealed an enormously complex swarm of mutant genomes (26). Reliability of the mutations and deletions with such a mutation frequency cutoff is supported by the clean read coverage attained, and by the types of mutations and their location in the three codon positions, when compared with those scored with a 0.5% mutation frequency cutoff (*Materials and Methods*).

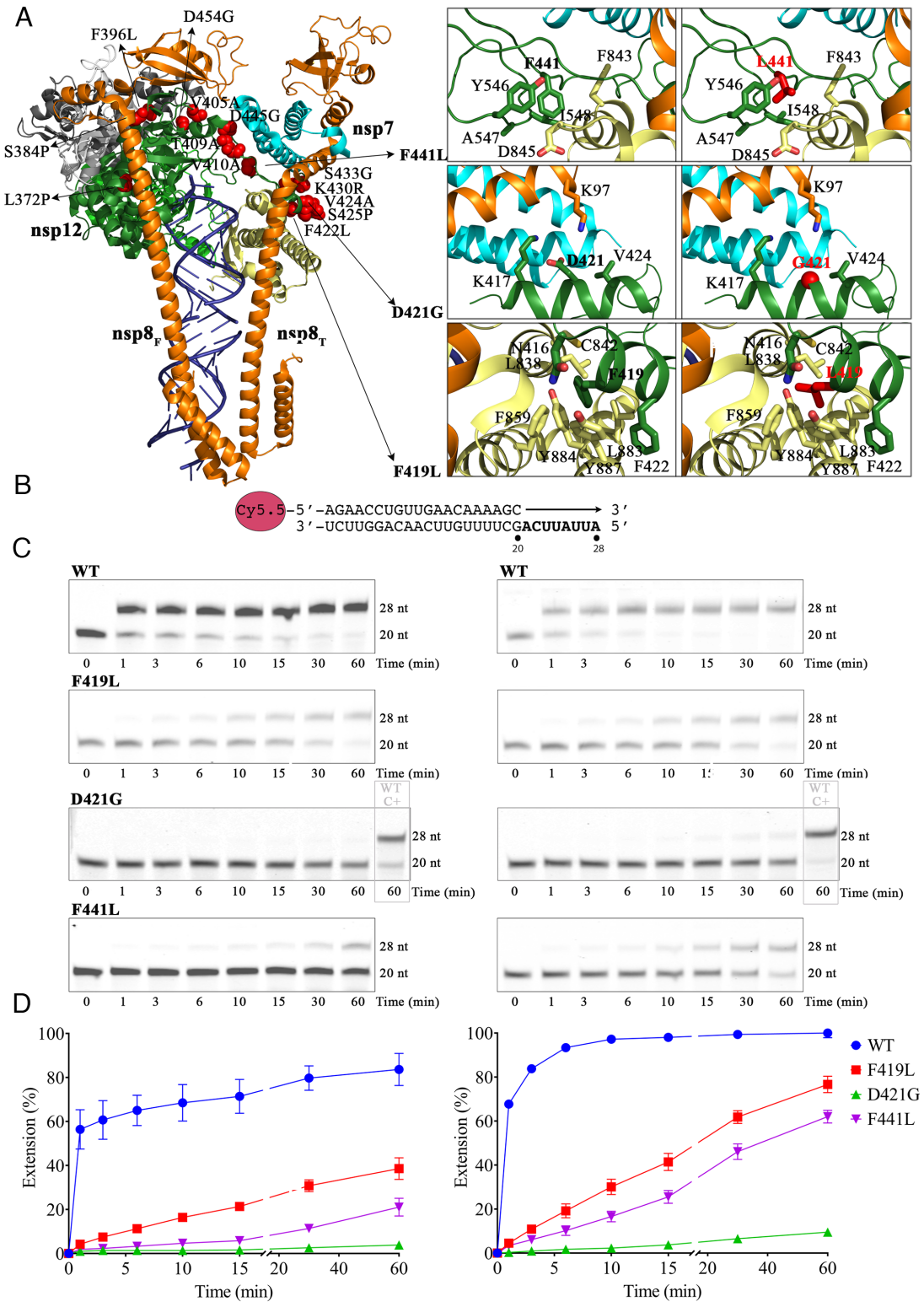

Fig. 3. Location and functional implications of the nsp12 amino acid substitutions detected in haplotypes of Wu-Mutant SOM master neurons. (*A*) The *Left* panel shows a cartoon representation of the nsp12-nsp8-nsp7-RNA complex (PDB code 6YYT), with nsp12 colored by domains (NiRAN in gray, interface in white, and the RdRp in dark green, green, and dark yellow for the fingers, palm, and thumb subdomains, respectively); the RNA is depicted in dark blue, and the cofactors nsp8 and nsp7 are colored in orange and cyan, respectively. Amino acid substitutions, whose position suggests that they could lead to alterations in RdRp activity, are shown as red spheres and are labeled. The right panels show close-up views highlighting the interactions of nsp12 substitutions (F419L, D421G, and F441L), selected for primer extension assays. (*B*) The template/primer duplex used in the assays. The arrow indicates the location and direction of primer extension. (*C*) Primer extension reactions of the nsp12-nsp7-nsp8 replication complexes, wild-type (WT) and selected nsp12 substitutions (indicated on the chromatograms shown on the *Left*). All reactions were performed at 37 °C (*Left*) and 33 °C (*Right*); samples were collected at different times (0, 1, 3, 6, 10, 15, 30, and 60 min), and the reaction products were separated by denaturing PAGE, 18% polyacrylamide, 7 M urea/Tris Borate-EDTA (TBE). (*D*) Graphs comparing the polymerization activities of WT and mutated nsp12 proteins, measured at 37 °C (*Left* graph) and 33 °C (*Right* graph). In ordinate, elongation is represented as normalized percentages of the extended product; to normalize these data, the RNA elongated at 33 °C by the WT complex was taken as 100%. The abscissa has been segmented to better visualize the first minutes of the elongation reactions. The first 17 min are shown in the first section of the abscissa and from min 17 to 60 in the second section. At both temperatures, the differences between WT and each of the mutants were statistically significant ($P < 0.001$; Dunnett test). Procedures are described in (*Materials and Methods*).

How frequently variants of SARS-CoV-2 originate during infection, and how many of them carry functionally relevant changes—that can nurture epidemiologically dominant viruses displaying alternative biological features—remains largely unexplored. This is partly due to lack of studies that have compared intrahost heterogeneity among related clinical isolates with sufficient penetration into the composition of their mutant spectra.

For visualizing relationships among large datasets, as those provided by mutant spectra (and deduced haplotypes) of viral quasispecies, SOM has advantages over other multidimensional scaling algorithms (46–48). Each SOM neuron represents a set of haplotypes related by similarity. Furthermore, SOM organizes neurons into a 2D grid, thus ordering the haplotypes of the dataset into this low-dimensional structure. This regular mesh is used as a foundation to generate a 3D graph (surface or landscape) by assigning to each neuron a numerical value associated with a property of the data that has not been used in the SOM training.

3D SOMs revealed the presence of a dominant neuron in the SARS-CoV-2 isolates, that was surrounded by a low-frequency neuron cloud, and defined twelve SOM classes (Table 1 and Fig. 2). In the same patient different SOM classes coexisted in different amplicons (Fig. 2 and *SI Appendix*, Figs. S2–S7 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw), suggesting that diversification forces and constraints may intervene differentially among genomic regions of viral RNA subsets within the same isolate. The cloud of Wu-Mutant 3D SOMs occupied a location in sequence space that was clearly distinct from the location occupied by the cloud of the Wu SOMs. This is expected because most cloud components are directly connected with the master that dominates them. Comparison of the number of neuron peaks among peak height intervals indicates differences in peak height distribution (Fig. 2). Additional work is necessary to interpret such differences, and also the presence of different SOM classes in virus of the same patient, as well as to obtain information on mutations that are present in the same genomic molecule (mutational linkage).

SOMs directed our search toward mutations that were potentially relevant regarding protein function. Specifically, for nsp12 A1, SOMs revealed 4 master neurons whose positions differed from the position of the master neurons present in virus of the rest of the patients. Of the 18 substitutions shared by the mutant clouds of these 4 master neurons (*SI Appendix*, Fig. S9 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw), substitutions F419L, D421G, and F441L were initially selected to be tested in functional assays because of their location in the three-dimensional structure of the polymerase complex. Additionally, these 4 substitutions were present in the mutant cloud of all isolates of our cohort. According to the three-dimensional structure of the nsp12-nsp8-nsp7 complex (Fig. 3*A*), F419L and F441L result in loss of intramolecular interactions between nsp12 subdomains, while D421G weakens intermolecular nsp12-nsp7 interactions. These substitutions resulted in altered kinetics of RNA synthesis in an in vitro primer-extension assay (Fig. 3 *B*–*D*). In all cases, the result was a drastic decrease in polymerase activity. These results correlate well with previously obtained biochemical data showing that weakening of interactions at or near the nsp12-nsp8 interface significantly affected RdRp activity (40). Interestingly, substitution L372F which enhanced the RNA synthesis in the biochemical assays (40) has been also found in the mutant cloud of one of the patients (Pt045) from the cohort. This underlines the possibility that nsp12 substitutions, which arise as minorities in infected patients, and which are at a very short mutational distance from the dominant sequence, may either enhance or diminish RNA synthesis activity.

Potential functional effects of amino acid substitutions during incipient diversification can be predicted for the S protein, in addition to the well-established fitness-enhancing effects of substitution D614G (39). The haplotype of the master neuron in amplicon A5 of S included C, rather than U at genomic position 23,042 [SOM Wu-Mutant (4) in Table 1]. This mutation results in amino acid substitution S494P, located at the RNA binding domain of S, and it has been shown to enhance the affinity of S for ACE2 (49), and to confer the virus potential for immune escape (50–52). The haplotype of the master neuron in amplicon A5 of S included U, rather than C, at genomic position 23,127 [SOM Wu-Mutant (3) in Table 1]. This mutation results in amino acid substitution A522V which maps also in the RNA binding domain of S. This substitution has been shown to enhance monoclonal antibody-mediated neutralization of SARS-CoV-2 (53) (All mutations and amino acid substitutions identified in the present study are listed in *SI Appendix*, Table S3 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw).

The effect of an amino acid substitution may be sequence context dependent. SARS-CoV-2 spike substitutions Q498R and N501Y were more favorable for enhancing ACE2 receptor binding when they were placed in the context of the S sequence of Omicron than the S sequence of earlier variants (54). Substitution Q498R is present in the mutant cloud of patients Pt130 and Pt354 of our cohort (*SI Appendix*, Table S3 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). Similar observations on context dependence effect of substitutions have been made with other RNA viruses. For example, poliovirus polymerase substitution G64D, and its counterpart G62D in the foot-and-mouth disease virus polymerase, have distinct consequences for polymerase function and antiviral resistance in the two viruses (55–57).

SOM maps constructed with high-resolution deep sequencing data may untangle haplotype composition modifications that underlie the process of SARS-CoV-2 diversification throughout different COVID-19 waves that may emerge as the virus continues replicating in the human population. This procedure may be applied to other complex microbial pathogens to gain insights in their short-term evolutionary mechanisms at the molecular level.

Materials and Methods

Patient Cohort and Sample Collection. The study has been performed with SARS-CoV-2 sequences retrieved from diagnostic, nasopharyngeal samples of 30 patients admitted to the Fundación Jiménez Díaz Hospital (FJD, Madrid, Spain) from April 3 to April 22, 2020, during the first COVID-19 outbreak in Spain. Positive SARS-CoV-2 diagnosis was based on a specific real-time RT-PCR analysis (38). COVID-19 severity classification was according to the following criteria: mild (no hospitalization; n = 10); moderate (hospitalization without admission in intensive care unit; n = 10); and severe (hospitalization with admission to the intensive care unit, and exitus; n = 10); comorbidities were equally distributed among the three severity groups. Patient code, virus isolation date, and clinical profiles [allowed by the Ethics Committee and the Institutional Review Board (ECIRB) of the FJD hospital (no. PIC-087-20-FJD)] are compiled in *SI Appendix*, Table S1 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). The regulation of the ECIRB follows the regulation ("Orden") number SAS 3470/2009 and the Helsinki's Declaration of the World Medical Association on medical investigation with humans. Furthermore, according to Spanish regulations, patients' informed consent was waved during the first disease wave for emergency reasons; this is stated in article 58 of LIB (law of biomedical investigation) (14/2007) and article 24 of RD (Royal Decree) 1716/2011. The samples were anonymized prior to the present study. Isolates were analyzed directly from the diagnostic samples, without cell culture amplification. Strict precautions were taken to avoid cross-contaminations among samples at the time of their collection and subsequent laboratory procedures until the samples reached the Illumina sequencing apparatus.

SARS-CoV-2 RNA Amplification. Total RNA was extracted from 140 μL of the nasopharyngeal swab sample, with the QIAamp Viral RNA Mini Kit (250) from Qiagen. The specific primers to amplify residues 14,534 to 16,054 of the nsp12 (polymerase)-coding region (which encode amino acids 366 to 871) and residues 22,872 to 23,645 of the S-coding region (which encode amino acids 438 to 694) are listed in *SI Appendix*, Table S3 (https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). Nucleotide and amino acids are numbered according to the Wuhan-Hu-1 isolate (NCBI reference sequence NC_045512.2). The procedure used for the amplifications using the Transcriptor One Step RT-PCR kit (Roche Applied Science) has been previously described (22, 23, 26). In brief, each amplification mixture included the RNA preparation (5 μL; 12.5% of the total volume), 5× buffer (10 μL), a solution that contained the forward oligonucleotide primer (2 μL), a solution with the reverse primer (2 μL) (50 ng/μL each), Transcriptor reverse transcriptase and Taq polymerase (1 μL). RNA was reverse transcribed at 50 °C for 30 min; the reaction was followed by an initial denaturing step at 94 °C for 7 min and then by 35 to 45 cycles of a denaturing step at 94 °C for 10 s, annealing at 46 to 48 °C for 30 s, and an extension step at 68 °C for 40 s; the final extension at 68 °C was carried out for 7 min. To ensure the absence of cross-contaminations, amplifications in the absence of viral RNA were run in parallel; no evidence of contamination was obtained with these negative controls. The amplification products were analyzed by 2% agarose gel electrophoresis, using Gene Ruler 1 Kb Plus DNA ladder (Thermo Scientific) as molar mass standard. Then, they were purified (QIAquick Gel Extraction Kit, Qiagen), quantified (Qubit dsDNA Assay kit, Thermo Fisher Scientific), and analyzed for quality (TapeStation System, Agilent Technologies), prior to deep sequencing using the Illumina MiSeq platform. To ensure that there was no limitation of template molecules, amplifications by RT-PCR were carried out with dilutions of 1:10, 1:100, and 1:1,000 of the initial RNA preparations; only when the amplification with the 1:1,000 dilution of template produced a visible DNA band, the ultradeep sequencing analysis was performed using the undiluted template (22, 26).

Sequencing and Ultradeep Sequencing. Phylogenetic and Bioinformatics Analyses. Entire SARS-CoV-2 genomic sequences were determined using the COVIDSeq assay, with the MiSeq sequencing platform (Illumina). BaseSpace (Illumina) was used for quality check and Bio-IT Processor (version: 0x04261818) for the alignment relative to the reference NC_045512.2. Only those sequences that were qualified as "good" according to the QC overall score were used. For maximum likelihood (ML) phylogenetic analysis, genome sequences were aligned using MAFFT v7.453 software (58). ML phylogeny was inferred with IQ-Tree v2.0.5 (59) using its Model Finder function (60) before each analysis, to estimate the nucleotide substitution model best-fitted for each dataset, by means of Bayesian information criterion (BIC). We assessed the tree topology for each phylogeny with the Shimodaira–Hasegawa approximate likelihood-ratio test (SH-aLRT) (61). TreeTime v0.7.6 (62) was used for the ancestral reconstruction of most likely sequences of internal nodes of the tree and their clades.

For the ultradeep sequencing of the amplification products of amplicons A1 to A4 of the nsp12-coding region and amplicons A5 and A6 of the S-coding region (*SI Appendix*, Fig. S1 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw), DNA pools (4×10^9 molecules/μL) were purified (employing Kapa Pure Beads from Kapa Biosystems, Roche), quantified (using Qubit dsDNA Assay kit, Thermo Fisher Scientific), and adjusted to a concentration of 1.5 ng/μL. Each DNA pool was indexed employing the SeqCap Adapter Kit A/B (Roche) (24 Index), concomitantly with DNA processing using the Kapa Hyper Prep kit (Kapa Biosystems, Roche). Each DNA was quantified using the LightCycler 480 system and then subjected to deep sequencing with the MiSeq Illumina platform, with Miseq Reagent kit v3 (2 × 300 bp mode, with the 600 cycle kit). Sequences were processed from the Fastq data by applying the SeekDeep bioinformatics pipeline (27), using the following options: --extraExtractorCmds=-- checkRevComplementForPrimers –primerNumOfMismatches 3" "–extraProcessClusterCmds=--fracCutOff 0.005 –rescueExcludedOneOffLowFreqHaplotypes".

Quality Controls and Reproducibility. Several parameters were measured to test the adequacy of the bioinformatics pipeline to determine mutant spectra of SARS-CoV-2. According to Illumina specifications, 88.93% of the residues that were obtained matched a quality score Q > 30 (https://emea.illumina.com/systems/sequencing-platforms/miseq/specifications.html). The average number of clean reads per amplicon was 110,074 (range 89,201 to 129,807). Such a clean read coverage allowed reaching a cutoff mutation frequency of 0.1%. Controls that such cutoff level faithfully gathers the authentic mutant spectrum composition in the sample have been previously detailed (26). Briefly, the main evidence is the following: i) except for one mutation, the remaining 96 mutations and 10 deletions identified with a 0.5% mutation frequency cutoff were also represented using the 0.1% cutoff; ii) both resolution levels described the same mutational biases and ranking of mutations in the three codon positions; and iii) both cutoff values yielded a similar percentage of amino acid substitutions represented in isolates of outbreak.info database. Additional arguments were previously detailed (26).

The following test of sample representability of haplotype composition was performed (depicted in *SI Appendix*, Fig. S10 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw): a preparation of viral RNA from SARS-CoV-2 USA-WA1/2020 was amplified twice to determine the mutant spectrum composition of amplicons A1, A2, and A3 of the nsp12-coding region (replicas A and B). In addition, amplicons A1, A2, and A3 from replica B were also sequenced in a different run. Out of 32 different haplotypes identified, 13 were shared by the three replicas, and 8 were shared by two replicas (results in *SI Appendix*, Fig. S10 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). As an additional control, the six amplicons from 8 patients (Pt012, Pt013, Pt023, Pt138, Pt165, Pt168, Pt342, Pt417) were sequenced a second time. The analysis of clean reads of both SARS-CoV-2 quasispecies [called run (a) and run (b)] yielded a robust similar number of point mutations, and their frequencies ($P < 0.0001$ in all cases; Pearson correlation test) (results in *SI Appendix*, Fig. S11 in https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw), providing an additional argument in favor of the reliability of using the 0.1% mutation frequency cutoff.

SOM Derivation. A FASTA file was prepared for each amplicon-patient; the file included the sequence of the haplotypes identified in the amplicon, jointly with the haplotypes present in the sequence of isolate Wuhan-Hu-1 (NCBI reference sequence NC_045512.2). The sequence of each haplotype was labeled with a name and with the frequency of identical sequences that define it. Six files including all unique haplotypes of each amplicon were prepared. These files contained the haplotype sequences, without name or frequency, and they provided the basis for the SOM training datasets. The haplotype frequency label was used to build the 3D maps derived from the SOMs.

The SOM neural network model exhibits an architecture consisting of neurons arranged in a rectangular 2D grid that defines neighbor relationships between them (32, 33). Each neuron has a prototype vector with the same nature and dimension as the input dataset. The SOM training algorithm iteratively processes the input vectors of the dataset and modifies the prototype vectors of the neurons so that they finally represent groups of similar input samples (vector quantization). In training, each input vector is associated with the best matching unit (*bmu*), which is the neuron with the closest prototype vector to the input sample. The modification of the prototype vectors is done in such a way that the distances between samples in the input high-dimensional space are most faithfully reflected between *bmus* in the 2D space of the grid (vector projection or multidimensional scaling). By combining vector quantization and vector projection, the SOM training algorithm orders the input dataset samples on the 2D grid. As a mathematical model, SOM is based on the calculation of distances between input samples and prototype vectors of neurons, which is usually done using Euclidean distance. In this work, the training datasets are composed of haplotype sequences, so a previous transformation into numerical vectors was necessary. This was done using the 3D irregular encoding (35), which weights each transition with distance 1, each transversion with distance 2, and each deletion with distance 1.06 to any nucleotide.

The prototype vectors of the neurons in a SOM are initiated with random values, which introduces a stochastic factor. Therefore, multiple SOMs with the same grid size were trained, and the one that best represents the input space based on a quality metric was selected. To obtain a single SOM for each amplicon dataset, five groups of 10 SOMs were trained, each group with a square grid size. The size of the grid in each group was $\lceil \sqrt{X} \rceil$, $\lceil \sqrt{2X} \rceil$, $\lceil \sqrt{3X} \rceil$, $\lceil \sqrt{4X} \rceil$, and $\lceil \sqrt{5X} \rceil$, where X was the number of samples in the training dataset; to address the stochastic initialization of the prototype vectors, for each group of 10 SOMs of the same size, the one that produced the lowest Kaski–Lagus error was selected (63). The average and deviation of these differences were 0.030 ± 0.016, 0.029 ± 0.018, 0.026 ± 0.016, and 0.026 ± 0.015 for amplicons A1, A2, A3, and A4 of the nsp12-coding region, respectively, and 0.018 ± 0.011 and 0.033 ± 0.023

for amplicons A5 y A6 of the S-coding region, respectively. These values indicate that similar 3D SOM maps were obtained from the same massive sequencing data in independent trainings. From the resulting 5 SOMs, the one with the greatest gain in *bmu's* and the least loss in *non-bmus* (neurons that did not represent any sample) with respect to the size of the SOM immediately below, was selected. This criterion aimed at attaining a balance between having sufficient dispersion of the sequences in the grid and avoiding an excessive number of *non-bmu* neurons. For training all the SOMs, we employed a square grid architecture and hexagonal neighborhood connections. The initial neighborhood area had a radius equal to the number of grid rows minus 1, and this area decreased by 1 at the end of each five epochs. The learning factor alpha(t) was defined as $0.1 \times [1 - t/(\text{total number of iterations})]$. The total number of iterations was determined by multiplying the total number of epochs by the number of samples in the dataset. The total number of epochs was calculated as the initial radius of the neighborhood area multiplied by 5, allowing the algorithm to iterate until the neighborhood area only affected the best matching unit, plus 5 additional epochs for fine-tuning.

In the 3D maps derived from a SOM, the third dimension represents the sum of the frequencies of the haplotypes associated with each *bmu*. For all 3D maps, a frequency normalization was applied to ensure that the sum of frequencies of all peaks in the map was 100. From the SOM selected for each amplicon, both the frequency maps per patient and the composite maps per SOM class were generated (two patients belonged to the same SOM class if they had the same master neuron in their respective 3D SOM maps).

Biochemical assays. The pRSFDuet-1 plasmid [(14his-nsp8-nsp7)(nsp12), from the Wuhan isolate] optimized for expression in *Escherichia coli*, and obtained from Addgene (Plasmid #165451) (64), was used to coexpress 14his-nsp8-nsp7-nsp12 complex. Mutations in the nsp12-coding region were introduced by site mutagenesis. The primers used are summarized in *SI Appendix*, Table S4 (https://saco.csic.es/index.php/s/smeN9oSSgMMxDLw). The nsp12 wild-type and mutants F419L, D421G, and F441L were expressed and purified as previously described (40). Briefly, the four constructs were expressed in *E. coli* BL21 Star (DE3) with 0.1 mM IPTG induction at 20 °C in overnight cultures. Cells were resuspended in 50 mL lysis buffer [50 mM Na-HEPES pH 8, 500 mM NaCl, 10 mM imidazole supplemented with 1 complete EDTA-free protease inhibitor (Roche) tablet, and 10 μg/mL DNase I] and lysed with a cell disruptor in 3 passes at 1.4 kbar at 4 °C. The lysate was cleared by centrifugation at 40,000 × g and 4 °C for 30 min; the supernatant was filtered and loaded into an immobilized metal-affinity chromatography column (IMAC) (HisTrap HP 5mL, Cytiva) in 50 mM Na-HEPES pH 8, 500 mM NaCl, and 10 mM imidazole and eluted in a linear gradient from 5 to 100% with 500 mM imidazole. Fractions containing his-nsp8, nsp7, and nsp12, diluted with 4 volumes of 50 mM Na-HEPES pH 8, were loaded into an anion exchange chromatography column (HiTrapQ HP 5 mL, Cytiva) in 50 mM Na-HEPES pH 8, 150 mM NaCl) and eluted in a linear gradient from 0 to 100% with 1M NaCl. An additional purification step by size-exclusion chromatography was carried out, using a Superdex 200 increase 10/300 GL (GE Healthcare) size-exclusion chromatography column carefully cleaned of RNAses, and equilibrated with RNase-free buffer 50 mM Na-HEPES pH 8, 300 mM NaCl, and 1 mM MgCl2. Peak fractions containing the his-nsp8-nsp7-nsp12 complex were collected and concentrated to around 5 mg/mL with Amicon Ultra 30 kDa MWCO centrifugal filters (Millipore), flash-frozen in liquid nitrogen, and stored at −80 °C until use.

Synthetic RNA oligonucleotides were ordered from Integrated DNA Technologies (IDT, Coralville, IA, USA). For RNA primer/template duplex formation, 1 μM primer fluorescently labeled with cyanine 5.5 (5′-Cy5. 5-AGAACCUGUUGAACAAAAGC-3′) and 4 μM template (5′-AUUAUUCAGCUUUUGUUCAACAGGUUCU-3′) were mixed in 50 mM NaCl and heated at 95 °C for 10 min, then slowly cooled down to 10 °C, and stored at −20 °C until use.

In vitro RNA polymerase activity assays were based on previous experiments (40). Briefly, primer extension assays were carried out in 20 μL reactions for each time point. The reaction buffer contained 5 nM annealed RNA and 1.475 μM RdRp complex in 20 mM Na-HEPES pH 7.8, 50 mM NaCl, 20 mM DTT, 5 mM MgCl2, 0.01 U RiboLock (Thermo Scientific). Reactions were pre-incubated first in the absence of rNTPs at 37 °C for 3 min, then triggered with the addition of rNTPs, further incubated at 37 °C or 33 °C. Reactions were stopped at different time points with the addition of 40 μL quenching buffer [8 M urea, 90 mM Tris base, and 10 mM EDTA (ethylenediaminetetraacetic acid), 0.02% SDS (sodium dodecyl sulfate), and 0.1% bromophenol blue]. Quenched reaction mixtures were heated at 95 °C for at least 15 min to completely denature RNA and were then loaded into a 18% polyacrylamide (37.5:1 acrylamide:bisacrylamide) denaturing urea-PAGE gel [7 M urea, 0.5 X TBE (50 mM Tris base, 50 mM boric acid, and 1 mM EDTA) buffer] and ran in 0.5 X TBE for 100 min at 150 V. Gels were scanned using an Odyssey 9120 Infrared (Li-Cor Biosciences) imaging system with Image Studio software and analyzed using the Fiji package (65) to quantify the amount of extended and nonextended RNA primers. Data obtained from independent triplicate reactions were used to calculate the percentage of primer extension at each time point. Data analysis was performed with GraphPad Prism (version 9.3.0 for Windows, GraphPad Software, San Diego, CA, www.graphpad.com).

Statistics. The significance of the difference between the number of unique and shared neuron peaks among amplicon-patients was calculated using the proportion test with the Yates continuity correction. The significance of the differences in the number of neuron peaks among amplicon-patients was calculated using the *t* test since we previously established that the data followed a normal distribution according to the Shapiro–Wilk test. The compactness of the mutant clouds in the composite maps for each SOM class was analyzed to assess whether it corresponds to the same compactness in the input space. The mean and deviation of the Hamming distance between all the haplotypes, and the mean and deviation of the minimum distance in the grid between the *bmus* of all the haplotypes (with distance 1 for each neighborhood connection), were calculated. Compactness was estimated in the input space using a sphere with a radius equal to the Hamming mean plus the deviation and in the output space using a circle with a radius equal to the mean plus the deviation in the grid. The results obtained for the four amplicons with SOM Wu-Mutant classes validated that the compactness of the mutant clouds in the 3D maps did not correspond to the compactness of the haplotypes in the input space, being basically related to the number of haplotypes and the number of *bmus* of each SOM class. The significance of the difference between RNA elongation rates was calculated with the Dunnett test.

Data, Materials, and Software Availability. All genomic sequences for the phylogenetic analysis and associated metadata are published in Global Initiative on Sharing All Influenza Data (GISAID's EpiCoV database) (https://gisaid.org) (GISAID Identifier: EPI_SET_231009qw; 10.55876/gis8.231009qw) (66). The Fastq files for all SARS-CoV-2 ultradeep sequences used in the present study are available at ENA (ID PRJEB48766) (67). The genomic sequences obtained using COVIDSeq, and the ultradeep sequences (repetition of analysis of virus from patients Pt012, Pt013, Pt023, Pt138, Pt165, Pt168, Pt342, and Pt417) are available at ENA (ID PRJEB70805) (68).

ACKNOWLEDGMENTS. Institutional support from Global Management Solutions and valuable discussions are acknowledged. The work at UPM was supported by grants PID2019-104903RB-100 and PID2022-139908OB-I00, funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe. The work at FJD was supported by project PI21/00139 from Instituto de Salud Carlos III, co-funded by the European Union. The work at CSIC was supported by PID2020-113888RB-I00, PID2020-117976GB-I00, and 202220I116 from Ministerio de Ciencia e Innovación and S2018/BAA-4370 (PLATESA2 from Comunidad de Madrid/FEDER). This research work was also funded by the European Commission – NextGenerationEU (regulation EU 2020/2094), through the CSIC's Global Health Platform (PTI Salud Global), and Fundació La Marató de TV3 (project 525/C/2021), grants 202136-30 and 202136-31. Institutional grants from the Fundación Ramón Areces and Banco Santander to the CBMSO are also acknowledged. The team at CBMSO belongs to the Global Virus Network (GVN). Work at Centro Nacional de Microbiología (ISCIII) was supported by grants SAF2016-77894-R from MINECO and PI13/02269 from ISCIII. The work at Universidad Complutense Madrid has been supported by research grant CTQ2017-87864-C2-2-P, from MINECO. B.M.-G. is supported by predoctoral contract PFIS FI19/00119 from ISCIII, cofinanced by Fondo Social Europeo (FSE). A.D.-P. is supported by the contract 13-2022-008566 cofinanced by the Comunidad de Madrid, through the Programa Investigo, en el marco del Plan de Recuperación, Transformación y Resiliencia, financed by the European Union–Next Generation EU. P.S. is supported by postdoctoral contract "Margarita Salas" CA1/RSUE/2021 from MCIU. C.G.-C. is supported by predoctoral contract PRE2018-083422 from MCIU.

Author affiliations: [a]Departamento de Sistemas Informáticos, Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, Universidad Politécnica de Madrid, Madrid 28031, Spain; [b]Microbes in Health and Welfare Program, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Consejo Superior de Investigaciones Científicas, Madrid 28049, Spain; [c]Departamento de Biología Molecular, Universidad Autónoma de Madrid, Madrid 28049, Spain; [d]Structural and Molecular Biology Department, Institut de Biología Molecular de Barcelona, Consejo Superior de Investigaciones Científicas, Barcelona 08028, Spain; [e]Department of Molecular and Cell Biology, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas, Madrid 28049, Spain; [f]Department of Clinical Microbiology, Instituto de Investigación Sanitaria-Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid, Madrid 28040, Spain; [g]Global Management Solutions S.L., Torre Picasso, Madrid 28020, Spain; [h]Unidad de Virología Molecular, Laboratorio de Referencia e Investigación en retrovirus, Centro Nacional de Microbiología, Instituto de salud Carlos III, Majadahonda 28222, Spain; [i]Departamento de Bioquímica y Biología Molecular, Universidad Complutense de Madrid, Madrid 28040, Spain; and [j]Department of Medicine, Division of Infectious Diseases, Northwestern University Feinberg School of Medicine, Center for Pathogen Genomics and Microbial Evolution, Northwestern University Havey Institute for Global Health, Chicago, IL 60611

1. P. V'Kovski, A. Kratzel, S. Steiner, H. Stalder, V. Thiel, Coronavirus biology and replication: Implications for SARS-CoV-2. *Nat. Rev. Microbiol.* **19**, 155–170 (2021).
2. M. R. Denison, R. L. Graham, E. F. Donaldson, L. D. Eckerle, R. S. Baric, Coronaviruses: An RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.* **8**, 270–279 (2011).
3. N. S. Ogando *et al.*, The enzymatic activity of the nsp14 exoribonuclease is critical for replication of MERS-CoV and SARS-CoV-2. *J. Virol.* **94**, e01246–20 (2020).
4. S. Lin *et al.*, Crystal structure of SARS-CoV-2 nsp10 bound to nsp14–ExoN domain reveals an exoribonuclease with both structural and functional integrity. *Nucleic Acids Res.* **49**, 5382–5392 (2021).
5. J. Gribble *et al.*, The coronavirus proofreading exoribonuclease mediates extensive viral recombination. *PLoS Pathog.* **17**, e1009226 (2021).
6. P. Schuster, P. F. Stadler, Virus evolution on fitness landscapes. *Curr. Top Microbiol. Immunol.* **439**, 1–94 (2023).
7. K. M. Braun *et al.*, Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. *PLoS Pathog.* **17**, e1009849 (2021).
8. P. V. Markov *et al.*, The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* **21**, 361–379 (2023).
9. D. Xu, Z. Zhang, F. S. Wang, SARS-associated coronavirus quasispecies in individual patients. *N. Engl. J. Med.* **350**, 1366–1367 (2004).
10. J. W. Tang *et al.*, The large 386-nt deletion in SARS-associated coronavirus: Evidence for quasispecies? *J. Infect Dis.* **194**, 808–813 (2006).
11. J. Liu *et al.*, SARS transmission pattern in Singapore reassessed by viral sequence variation analysis. *PLoS Med.* **2**, e43 (2005).
12. D. Park *et al.*, Analysis of intrapatient heterogeneity uncovers the microevolution of Middle East respiratory syndrome coronavirus. *Cold Spring Harb. Mol. Case Stud.* **2**, a001214 (2016).
13. M. K. Borucki *et al.*, Middle East respiratory syndrome coronavirus intra-host populations are characterized by numerous high frequency variants. *PLoS One* **11**, e0146251 (2016).
14. T. Karamitros *et al.*, SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. *J. Clin. Virol.* **131**, 104585 (2020).
15. M. R. Capobianchi *et al.*, Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. *Clin. Microbiol. Infect* **26**, 954–956 (2020).
16. C. Andres *et al.*, Naturally occurring SARS-CoV-2 gene deletions close to the spike S1/S2 cleavage site in the viral quasispecies of COVID19 patients. *Emerg. Microbes Infect.* **9**, 1900–1911 (2020).
17. M. Rueca *et al.*, Compartmentalized replication of SARS-Cov-2 in upper vs. lower respiratory tract assessed by whole genome quasispecies analysis. *Microorganisms* **8**, E1302 (2020).
18. H. A. Al Khatib *et al.*, Within-host diversity of SARS-CoV-2 in COVID-19 patients with variable disease severities. *Front. Cell Infect. Microbiol.* **10**, 575613 (2020).
19. J. Gregori *et al.*, Host-dependent editing of SARS-CoV-2 in COVID-19 patients. *Emerg. Microbes Infect.* **10**, 1777–1789 (2021).
20. F. Sun *et al.*, SARS-CoV-2 quasispecies provides an advantage mutation pool for the epidemic variants. *Microbiol. Spectr.* **9**, e0026121 (2021).
21. D. Khateeb *et al.*, SARS-CoV-2 variants with reduced infectivity and varied sensitivity to the BNT162b2 vaccine are developed during the course of infection. *PLoS Pathog.* **18**, e1010242 (2022).
22. B. Martinez-Gonzalez *et al.*, SARS-CoV-2 point mutation and deletion spectra and their association with different disease outcomes. *Microbiol. Spectr.* **10**, e0022122 (2022).
23. B. Martinez-Gonzalez *et al.*, Vaccine-breakthrough infections with SARS-CoV-2 Alpha mirror mutations in Delta Plus, Iota and Omicron. *J. Clin. Invest.* **132**, e157700 (2022).
24. H. Zhao *et al.*, Plasticity in structure and assembly of SARS-CoV-2 nucleocapsid protein. *PNAS Nexus* **1**, pgac049 (2022).
25. A. Jary *et al.*, Evolution of viral quasispecies during SARS-CoV-2 infection. *Clin. Microbiol. Infect.* **26**, 1560.e1–1560.e4 (2020).
26. B. Martinez-Gonzalez *et al.*, SARS-CoV-2 mutant spectra at different depth levels reveal an overwhelming abundance of low frequency mutations. *Pathogens* **11**, 662 (2022).
27. N. J. Hathaway, C. M. Parobek, J. J. Juliano, J. A. Bailey, SeekDeep: Single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Res.* **46**, e21 (2018).
28. M. Amicone *et al.*, Mutation rate of SARS-CoV-2 and emergence of mutators during experimental evolution. *Evol. Med. Public Health* **10**, 142–155 (2022).
29. L. D. Eckerle, X. Lu, S. M. Sperry, L. Choi, M. R. Denison, High fidelity of murine hepatitis virus replication is decreased in nsp14 exoribonuclease mutants. *J. Virol.* **81**, 12135–12144 (2007).
30. R. Sender *et al.*, The total number and mass of SARS-CoV-2 virions. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2024815118 (2021).
31. E. Domingo, C. Garcia-Crespo, R. Lobo-Vega, C. Perales, Mutation rates, mutation frequencies, and proofreading-repair activities in RNA virus genetics. *Viruses* **13**, 1882 (2021).
32. T. Kohonen *et al.*, Self organization of a massive document collection. *IEEE Trans. Neural. Netw.* **11**, 574–585 (2000).
33. T. Kohonen, *Self-Organizing Maps* (Springer-Verlag, 2001).
34. R. Lorenzo-Redondo, S. Delgado, F. Moran, C. Lopez-Galindez, Realistic three dimensional fitness landscapes generated by self organizing maps for the analysis of experimental HIV-1 evolution. *PLoS One* **9**, e88579 (2014).
35. S. Delgado, F. Moran, A. Mora, J. J. Merelo, C. Briones, A novel representation of genomic sequences for taxonomic clustering and visualization by means of self-organizing maps. *Bioinformatics* **31**, 736–744 (2015).
36. S. Delgado *et al.*, A two-level, intramutant spectrum haplotype profile of hepatitis C virus revealed by self-organized maps. *Microbiol. Spectr.* **9**, e0145921 (2021).
37. M. G. Lopez *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).
38. M. E. Soria *et al.*, High SARS-CoV-2 viral load is associated with a worse clinical outcome of COVID-19 disease. *Access Microbiol.* **3**, 000259 (2021).
39. B. Korber *et al.*, Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812–827.e9 (2020).
40. C. Ferrer-Orta *et al.*, Point mutations at specific sites of the nsp12-nsp8 interface dramatically affect the RNA polymerization activity of SARS CoV-2. Under review (2023).
41. D. S. Ferrero *et al.*, Supramolecular arrangement of the full-length Zika virus NS5. *PLoS Pathog.* **15**, e1007656 (2019).
42. P. Baccam, R. J. Thompson, O. Fedrigo, S. Carpenter, J. L. Cornette, PAQ: Partition analysis of quasispecies. *Bioinformatics* **17**, 16–22 (2001).
43. P. Skums *et al.*, QUENTIN: Reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics* **34**, 163–170 (2018).
44. S. Ahn, Z. Ke, H. Vikalo, Viral quasispecies reconstruction via tensor factorization with successive read removal. *Bioinformatics* **34**, i23–i31 (2018).
45. R. Henningsson, G. Moratorio, A. V. Borderia, M. Vignuzzi, M. Fontes, DISSEQT-DIStribution-based modeling of SEQuence space Time dynamics. *Virus Evol.* **5**, vez028 (2019).
46. R. C. Tillquist, M. E. Lladser, Low-dimensional representation of genomic sequences. *J. Math. Biol.* **79**, 1–29 (2019).
47. L. H. Nguyen, S. Holmes, Ten quick tips for effective dimensionality reduction. *PLoS Comput. Biol.* **15**, e1006907 (2019).
48. R. Henningsson, M. Fontes, SMSSVD: SubMatrix selection singular value decomposition. *Bioinformatics* **35**, 478–486 (2019).
49. S. Chakraborty, Evolutionary and structural analysis elucidates mutations on SARS-CoV2 spike protein with altered human ACE2 binding affinity. *Biochem. Biophys. Res. Commun.* **534**, 374–380 (2021).
50. J. Chen, K. Gao, R. Wang, G. W. Wei, Revealing the threat of emerging SARS-CoV-2 mutations to antibody therapies. *J. Mol. Biol.* **433**, 167155 (2021).
51. M. Alenquer *et al.*, Signatures in SARS-CoV-2 spike protein conferring escape to neutralizing antibodies. *PLoS Pathog.* **17**, e1009772 (2021).
52. F. Grabowski, G. Preibisch, S. Gizinski, M. Kochanczyk, T. Lipniacki, SARS-CoV-2 variant of concern 202012/01 has about twofold replicative advantage and acquires concerning mutations. *Viruses* **13**, 392 (2021).
53. Q. Li *et al.*, The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell* **182**, 1284–1294.e9 (2020).
54. T. N. Starr *et al.*, Deep mutational scans for ACE2 binding, RBD expression, and antibody escape in the SARS-CoV-2 Omicron BA.1 and BA.2 receptor-binding domains. *PLoS Pathog.* **18**, e1010951 (2022).
55. J. K. Pfeiffer, K. Kirkegaard, Increased fidelity reduces poliovirus fitness under selective pressure in mice. *PLoS Pathog.* **1**, 102–110 (2005).
56. M. Vignuzzi, J. K. Stone, J. J. Arnold, C. E. Cameron, R. Andino, Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344–348 (2006).
57. C. Ferrer-Orta *et al.*, Structure of foot-and-mouth disease virus mutant polymerases with reduced sensitivity to ribavirin. *J. Virol.* **84**, 6188–6199 (2010).
58. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
59. B. Q. Minh *et al.*, IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
60. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
61. M. Anisimova, O. Gascuel, Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55**, 539–552 (2006).
62. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus. Evol.* **4**, vex042 (2018).
63. S. Kaski, K. Lagus, "Comparing self-organized maps" in *Artificial Neural Networks- ICAN 1996. Lecture Notes in Computer Science*, C. Von der Malsburg, W. von Seelen, J. C. Vorbrueggen, B. Sendhoff, Eds. (Springer, Berlin, Heidelberg, 1996), vol. 1112, pp. 809–814.
64. C. Madru *et al.*, Fast and efficient purification of SARS-CoV-2 RNA dependent RNA polymerase complex expressed in Escherichia coli. *PLoS One* **16**, e0250610 (2021).
65. J. Schindelin *et al.*, Fiji: An open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
66. R. Lorenzo-Redondo *et al.*, Incipient functional SARS-CoV-2 diversification identified through neural network haplotype maps. GISAID's EpiCoV Database. https://doi.org/10.55876/gis8.231009qw. Deposited 23 October 2023.
67. C. Perales *et al.*, SARS-CoV-2 point mutation and deletion spectra, and their association with different disease outcome. European Nucleotide Archive (ENA). https://www.ebi.ac.uk/ena/browser/view/PRJEB48766. Deposited 13 November 2021.
68. C. Perales *et al.*, Incipient functional SARS-CoV-2 diversification identified through neural network haplotype maps. European Nucleotide Archive (ENA). https://www.ebi.ac.uk/ena/browser/view/PRJEB70805. Deposited 1 December 2023.