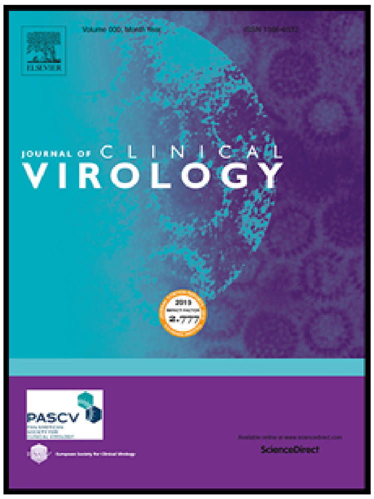
Journal Pre-proof



Sierra SARS-CoV-2 Sequence and Antiviral Resistance Analysis Program

Philip L. Tzou , Kaiming Tao , Malaya K. Sahoo ,
Sergei L. Kosakovsky Pond , Benjamin A. Pinsky ,
Robert W. Shafer

PII: S1386-6532(22)00255-4
DOI: https://doi.org/10.1016/j.jcv.2022.105323
Reference: JCV 105323

To appear in: *Journal of Clinical Virology*

Received date: 26 June 2022
Revised date: 11 October 2022
Accepted date: 21 October 2022

Please cite this article as: Philip L. Tzou , Kaiming Tao , Malaya K. Sahoo , Sergei L. Kosakovsky Pond , Benjamin A. Pinsky , Robert W. Shafer , Sierra SARS-CoV-2 Sequence and Antiviral Resistance Analysis Program, *Journal of Clinical Virology* (2022), doi: https://doi.org/10.1016/j.jcv.2022.105323

TITLE

Sierra SARS-CoV-2 Sequence and Antiviral Resistance Analysis Program

AUTHORS

Philip L. Tzou[1], Kaiming Tao[1], Malaya K. Sahoo[2], Sergei L. Kosakovsky Pond[3], Benjamin A. Pinsky[2], Robert W. Shafer[1]

AFFILIATIONS

[1]Division of Infectious Diseases, Department of Medicine, Stanford University, Stanford, CA, USA; [2]Department of Pathology, Stanford University, Stanford, CA, USA; [3]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, USA.

CORRESPONDING AUTHOR: Philip L. Tzou (philiptz@stanford.edu) and Robert W. Shafer (rshefer@stanford.edu)

WORD COUNT: 2,498

- Sierra SARS-CoV-2 provides quality control and annotation for viral genomic data.
- Sierra SARS-CoV-2 is highly concordant with established sequence analysis programs.
- 300 SARS-CoV-2 Spike mutations reduce susceptibility to monoclonal antibodies.
- Approximately 20 RdRP and Mpro mutations reduce antiviral susceptibility in vitro.
- Sierra SARS-CoV-2 has susceptibility data for 88% of Spike RBD mutation patterns.

ABSTRACT

Introduction: Although most laboratories are capable of employing established protocols to perform full-genome SARS-CoV-2 sequencing, many are unable to assess sequence quality, select appropriate mutation-detection thresholds, or report on the potential clinical significance of mutations in the targets of antiviral therapy. Methods: We describe the technical aspects and benchmark the performance of

Sierra SARS-CoV-2, a program designed to perform these functions on user-submitted FASTQ and FASTA sequence files and lists of Spike mutations. Sierra SARS-CoV-2 indicates which sequences contain an unexpectedly large number of unusual mutations and which mutations are associated with reduced susceptibility to clinical stage mAbs, the RdRP inhibitor remdesivir, or the Mpro inhibitor nirmatrelvir. Results: To assess the performance of Sierra SARS-CoV-2 on FASTQ files, we applied it to 600 representative FASTQ sequences and compared the results to the COVID-19 EDGE program. To assess its performance on FASTA files, we applied it to nearly one million representative FASTA sequences and compared the results to the GISAID mutation annotation. To assess its performance on mutations lists, we applied it to 13,578 distinct Spike RBD mutation patterns and showed that exactly or partially matching annotations were available for 88% of patterns. Conclusion: Sierra SARS-CoV-2 leverages previously published data to improve the quality control of submitted viral genomic data and to provide functional annotation on the impact of mutations in the targets of antiviral SARS-CoV-2 therapy. The program can be found at https://covdb.stanford.edu/sierra/sars2/ and its source code at https://github.com/hivdb/sierra-sars2.

Keywords: SARS-CoV-2; genomic sequencing; mutations; antiviral resistance

INTRODUCTION

SARS-CoV-2 sequencing is performed for surveillance as well as research and clinical purposes. The extent of sequencing for clinical purposes may increase as more SARS-CoV-2 inhibitors become available, particularly if resistance to these inhibitors arises. Although most laboratories are capable of performing full-genome SARS-CoV-2 sequencing employing established laboratory and sequence analysis protocols [1–8], many are unable to assess sequence quality, select appropriate mutation-detection thresholds, or report the potential clinical significance of SARS-CoV-2 mutations in the targets of antiviral therapy.

We previously briefly described a sequence analysis program called Sierra SARS-CoV-2 in a paper on the Stanford Coronavirus Antiviral Resistance Database (CoV-RDB) [9]. The program utilizes the same codebase as Sierra HIV [10,11], the Stanford HIV Drug Resistance Database sequence analysis program [10,11]. The program accepts three types of input: FASTQ files containing short reads from a deep sequencing instrument, FASTA sequences, and lists of Spike amino acid mutations.

To assess the performance of Sierra SARS-CoV-2 on FASTQ files, we applied it to two sets of sequences from the NCBI Sequence Read Archive (SRA) [12] and to sequences from a clinical laboratory. To assess its performance on consensus FASTA sequences, we applied it to 963,237 SARS-CoV-2 genome sequences from GISAID [13]. To assess its performance interpreting Spike mutations, we applied it to 13,578 distinct Spike receptor binding domain (RBD) amino acid mutation patterns from approximately 4.7 million SARS-CoV-2 GISAID sequences.

METHODS

Sierra SARS2-CoV-2 provides native support for FASTA sequences and lists of mutations, defined as amino acid differences from the Wuhan-Hu-1 reference sequence (GenBank accession NC_045512.2). Support for FASTQ files is provided through an auxiliary pipeline that converts FASTQ files to commadelimited files containing the frequency of each codon at each genomic position, i.e., codon frequency (CodFreq) files [9,11]. Table 1 summarizes SARS-CoV-2 output depending on whether CodFreq files, FASTA sequences, or mutation lists are submitted. Figure 1 illustrates the workflow of Sierra SARS2-CoV-2 for all three input types.

Generation of CodFreq files and consensus sequences

CodFreq files contain seven columns: gene, amino acid position, number of reads at a position, codon, number of reads for a codon, amino acid, and proportion of reads for a codon. For our application, the CodFreq format has several advantages over the commonly used variant call format (VCF) because CodFreq files can be interpreted without a reference sequence and used independently from the accompanying SAM/BAM file. CodFreq files can be used to generate a consensus FASTA sequence containing mixtures of codons above a user-specified threshold.

The CodFreq pipeline can be run on batched sequences using the Sierra SARS2-CoV-2 frontend or locally using a pre-built Docker image. A shell script is provided on GitHub for running the CodFreq pipeline from a local host (https://github.com/hivdb/codfreq). The frontend identifies paired-end files and prompts users to confirm the pairing. An advanced option is provided for users submitting primer information. The pipeline reports progress for each backend task.

The CodFreq pipeline includes the following steps (Supplementary Figure): (1) The Fastp program trims adapters, removes regions with low phred scores, and stitches paired reads; (2) MiniMap2 aligns FASTQ sequence reads to the reference sequence [14]; (3) Samtools converts the resulting SAM text file into a binary BAM file and a BAI index file [15]; (4) PySam reads the BAM file to determine the frequency of each codon at each position; and (5) PostAlign, a program we created, adjusts the placement of indels through a codon-aware process (https://github.com/hivdb/post-align).

Depending on user input, the programs Cutadapt or iVar are used to trim SARS-CoV-2 primers [16,17]. The CodFreq and BAM files are provided for users to download.

Identification of amino acid mutations and lineage assignment

Minimap2 and PostAlign are used to analyze FASTA sequences. Minimap2 aligns a query sequence to the reference sequence and saves the alignment in Pairwise mApping Format (PAF) files, which are then loaded into pairwise nucleotide alignments. PostAlign adjusts indels using a codon aware process and position-specific gap scores to increase consistency of indel placement in accordance with alignments of established SARS-CoV-2 variants. PostAlign also separates alignments into discrete genes, identifies mutations, and numbers them by gene. If complete genomes are submitted, the Pangolin program is used to assign the PANGO lineage [18].

Report generation and mutation annotation

The Sierra SARS-CoV-2 report contains sections summarizing sequence and mutation data (Supplementary File). The sequence summary reports the genes present in a sequence, areas in which sequence data are missing, the consensus sequence, and the assigned PANGO lineage. It contains a figure plotting read coverage along the sequence and read depths for Mpro, RdRp, and Spike genes. Dropdown menus enable users to interactively adjust the minimum number and proportion of reads for reporting non-consensus mutations.

Each mutation in a sequence is annotated with the following information: (1) The proportion and number of reads containing the mutation; (2) Whether the mutation is unusual, defined as having a global prevalence below 0.01% based on the open-source sequence analysis pipeline created by the Kosakovsky Pond laboratory [19]; (3) Whether the mutation is an mAb resistance mutation defined as a Spike mutation associated with reduced susceptibility to one or more clinical-stage mAbs; (4) Whether the mutation is a potential RdRp or Mpro resistance mutation; and (5) Comments for the most well studied Spike mutations associated with reduced mAb susceptibility and for most mutations associated with Mpro and RdRp inhibitor reduced susceptibility. The lists of mAb and potential RdRp and Mpro resistance mutations are updated monthly.

Each list of Spike mutations is also used to interrogate CoV-RDB for published mAb, convalescent plasma, and vaccinee plasma susceptibility data. There is an option to display just data from variants with exactly matching sets of mutations versus comprehensive results with data for variants with a subset or superset of the submitted mutations.

Generating a list of mAb resistance mutations

mAb resistance mutations were defined as Spike mutations with a median ≥5 fold reduction in susceptibility compared with wildtype according to CoV-RDB and/or having an escape fraction ≥0.1 in the deep mutational scanning (DMS) platform developed by the Bloom Laboratory at the University of Washington [20,21]. As of September 2022, there were 488 spike mutations meeting these criteria. Figure 2 illustrates the 160 RBD-associated mAb-resistance mutations having a prevalence ≥0.0001% with data on their neutralizing antibody susceptibilities, DMS escape fractions, and whether they were selected *in vitro* and/or *in vivo*.

Generating a list of mutations associated with potential small molecule inhibitor resistance

Mpro and RdRp mutations were classified as potential drug-resistance mutations if they met one of the following three criteria: (1) they were associated with 2.5-fold or higher reductions in susceptibility in either a biochemical assay or in cell culture; (2) they were selected during an in vitro passage experiment; or (3) they were selected in a person receiving an Mpro or RdRp inhibitor. Figure 3 illustrates that as of September 2022, 42 mutations at 28 positions were reported to be possibly associated with reduced susceptibility to Mpro inhibitors nirmatrelvir or ensitrelvir [22–35], and 11 mutations at 9 positions were reported to be possibly associated with reduced susceptibility to the RdRp inhibitor remdesivir [36–43].

Datasets used for benchmarking and validation

FASTQ files: Three sets of NGS files were used to compare the results of the CodFreq pipeline with the LANL EDGE COVID-19 program [3] including 200 randomly selected Illumina files obtained between March 2021 and March 2022 from the NCBI SARS-CoV-2 SRA portal, 200 randomly selected Oxford Nanopore Technology (ONT) files obtained March 2022 from the SRA portal, and 200 Illumina sequences from the Stanford University Hospital (SUH) Diagnostic Virology Laboratory between April 2021 and March 2022. Pangolin 4.0.5 classified 52.8% of the 600 sequences as Delta variants, 27.8% as Omicron variants, 11.6% as Alpha variants, and 7.8% as other variants. For the SRA sequences, we used the parameter –skip-technical to exclude adapters, primers, and bar-codes from the downloaded FASTQ file. The SUH sequences were generated using a recently published pipeline [44].

FASTA files: On March 25, 2022, a random set of 963,237 FASTA files was selected from 9,632,370 GISAID sequences[45].

Mutation data: The global prevalence of each Spike, Mpro and RdRp mutation was obtained from a publicly available quality controlled analysis pipeline created by the Kosakovsky Pond laboratory that contained 4,740,761 Spike, 5,328,735 Mpro, and 5,076,452 RdRp sequences containing 201,167 Spike, 5,404, Mpro and 32,788 RdRp distinct mutation patterns [46,47].

RESULTS

FASTQ files

To evaluate the CodFreq pipeline using FASTQ files, we tested the 200 NCBI SRA Illumina files, the 200 NCBI SRA ONT files and the 200 SUH Illumina files. For the 400 Illumina and 200 ONT sequences, we compared the consensus codon of each CodFreq file to the codon in the consensus FASTA file generated by the EDGE COVID-19 program (version 20220314). For both pipelines, a codon-level read depth ≥5 and a mutation-detection threshold of 50% were used.

Illumina sequences: For regions successfully aligned by both Sierra and EDGE, each program detected a mean 11.0, 1.7, and 0.19 amino acid mutations per sequence in Spike, RdRp, and Mpro. Of the 4,413 Spike mutations detected by either program, Sierra and EDGE detected the same mutation in 98.9% of cases; 0.7% were detected only by Sierra and 0.3% only by EDGE. Of 760 RdRp and Mpro mutations, Sierra and EDGE detected the same mutation in 98.4% of cases; 1.5% were detected only by Sierra and 0.1% only by EDGE. The 59 discordances in the three genes resulted from small differences in the threshold at which mutations were detected (n=49) and in placement of indels (n=10).

ONT sequences: For regions successfully aligned by both Sierra and EDGE, Sierra detected a mean 19.0, 1.7, and 0.48 mutations and EDGE detected a mean 18.7, 1.7, and 0.49 mutations per sequence in Spike, RdRp, and Mpro. Of the 3,855 Spike mutations detected by either program, Sierra and EDGE detected the same mutation in 94.3% of cases; 3.8% were detected only by Sierra and 2.0% only by EDGE. Of 448 detected RdRp and Mpro mutations, Sierra and EDGE detected the same mutation in 96.9% of cases; 2.0% were detected only by Sierra and 1.1% only by EDGE. The 235 discordances in the three genes resulted from small differences in the threshold at which mutations were detected (n=174) and in the placement of indels (n=61).

FASTA files

We compared the Spike, Mpro, and RdRp mutation lists generated by Minimap2 and PostAlign with the GISAID "AA substitutions" metadata, generated by the CoVSurver program [45] for 963,237

FASTA sequences. The list of mutations for Spike, Mpro, and RdRp genes identified by Sierra and GISAID were identical for 99.4%, 99.9% and 99.5% of sequences, respectively. However, there were differences in the placement of indels for Spike. Nearly all Spike differences were caused by indels at several positions, such as the Omicron BA.1 N-terminal domain deletion that has alternatively been placed at position 211 [48,49] or 212 [50–52]. The non-indel differences resulted from how mutations in regions surrounding missing sequence data were handled.

Mutation lists

Distribution of usual and unusual mutations: Figures 4A-C show the number of mutations in Spike, RdRP, and Mpro genes by the binned global prevalence of each mutation. Spike had 694 usual and 8,501 unusual non-indel mutations. RdRp had 300 usual and 4,192 unusual non-indel mutations. Mpro had 107 usual and 1,579 unusual non-indel mutations.

Figures 5A-C show the number of unusual mutations per sequence in Spike, RdRP, and Mpro. In Spike, 92.9% sequences had no unusual mutations, 6.7% had one, 0.4% had two, and <0.1% had three or more unusual mutations. In RdRp, 96.1% sequences had no unusual mutation, 3.8% had one and 0.1% had two or more unusual mutations. In Mpro, 99.2% sequences had no unusual mutation, 0.7% had one and 0.1% had two or more unusual mutations.

Figure 6 shows the numbers of usual and unusual Spike mutations at different mutation thresholds in the 200 NCBI Illumina and 200 NCBI ONT sequences. At mutation detection thresholds <50%, there was a markedly higher proportion of unusual mutations in ONT compared with Illumina sequences.

Neutralizing susceptibility data in CoV-RDB for submitted RBD mutation patterns: The Spike mutation pattern dataset contained 13,578 distinct patterns of Spike RBD mutations. Each RBD mutation pattern was submitted to Sierra to determine the frequency for which complete or partial neutralizing susceptibility data was available in CoV-RDB [9](Figure 7). For 76.7% of sequences (1.3% of patterns), CoV-RDB contained data exactly matching the submitted mutation pattern. For 10.2% of sequences (86.6% of patterns), CoV-RDB contained data partially matching the submitted mutation pattern (i.e., CoV-RDB contained data for mutation patterns representing a subset, superset, or intersecting set of the mutations in the submitted mutation pattern). For 13.0% of sequences (12.0% of patterns), CoV-RDB contained no data matching the pattern of submitted mutations.

DISCUSSION

Sierra SARS-CoV-2 is an open-source web-based program that accepts FASTQ and FASTA files and lists of Spike mutations. Depending on the nature of the input data, it generates a consensus nucleotide sequence, assigns a sequence lineage, identifies amino acid mutations, and uses the mutations to interrogate a quality-controlled sequence analysis pipeline for global mutation prevalence data and CoV-RDB for data on SARS-CoV-2 susceptibility to antiviral agents and to plasma from previously infected and/or vaccinated persons.

We assessed the performance of Sierra SARS-CoV-2 using 600 FASTQ datasets, nearly one million FASTA sequences, and approximately 13,500 distinct Spike RBD mutation patterns. In the analysis of FASTQ sequences, Sierra SARS-CoV-2 and EDGE COVID-19 were highly concordant and in the analysis of FASTA sequences, Sierra SARS-CoV-2 and the GISAID mutation list were highly concordant. For both analyses, most discordances resulted from equally acceptable placements of several commonly occurring indels. An analysis of approximately 13,500 distinct Spike RBD mutation patterns, showed that exactly or partially matching annotation data were available for 88% of reported mutation patterns.

Sierra SARS-CoV-2 uses mutation prevalence data to identify sequences with an unexpectedly large number of unusual mutations. Indeed, only 0.1% of quality-controlled Spike sequences had three or more unusual mutations and only 0.1% of quality-controlled Mpro and RdRp sequences had two or more unusual mutations. Therefore, the presence of many unusual mutations in a sequence suggests the possibility of sequence artifact or possibly, although less likely, a novel variant.

Sierra SARS-CoV-2 uses published data to identify mutations potentially associated with reduced antiviral susceptibility. Although few major SARS-CoV-2 lineages circulate at any time, an increasing number of Omicron sub-variants containing different spike mutation patterns are now reported in many regions [53]. Therefore, a sequence analysis program that provides susceptibility data for mutation patterns, as well as for variants of concern has become increasingly relevant. Additionally, an increasing number of Mpro mutations associated with reduced nirmatrelvir susceptibility have been identified *in vitro*, although few have been reported in persons receiving nirmatrelvir.

In conclusion, Sierra SARS-CoV-2 is one of a few open-source analytic pipelines actively maintained and available through a web interface [3,6,7]. It uniquely leverages published data to improve the quality control of submitted viral genomic data and to provide functional annotation on the impact of mutations in the targets of antiviral therapy.

ACKNOWLEDGMENTS

PLT, KT, and RWS have been funded in part by a grant from the NIH/NIAID: AI136618. The funder played no role in this study.

RWS served on Gilead Sciences and Vir Biotechnologies/GlaxoSmithKline scientific advisory boards.

FIGURE LEGENDS

Figure 1

Sierra SARS-CoV-2 work flow for handling FASTQ files, FASTA files, and lists of SARS-CoV-2 mutations. Sierra provides native support for FASTA sequences and mutation lists. Support for FASTQ files is provided through an auxiliary pipeline that converts FASTQ files into CSV files containing the frequency of each codon at each position in a genome. The workflow for the auxiliary pipeline is shown in Supplementary Figure. The Supplementary File shows an example of the HTML output.

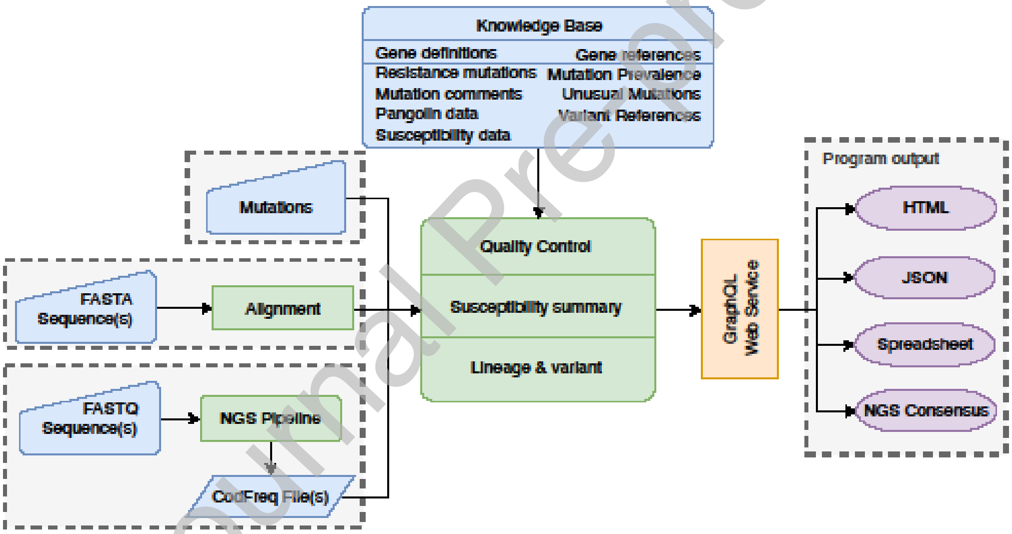


Figure 2

SARS-CoV-2 Spike RBD mAb-resistance mutations. The mAb-resistance mutations shown met one or more of the following criteria: (1) having a ≥5-fold reduction in susceptibility to a clinical stage mAb; (2) having a DMS escape fraction ≥0.1 and having a global prevalence >0.001%; (3) having been selected *in vitro* by an mAb; or (4) having been selected *in vivo* in a patient receiving an mAb or experiencing prolonged infection. A dark blue cell indicates a ≥25-fold reduction in susceptibility; a light blue cell indicates a 5-25-fold reduction in susceptibility; a white cell indicates a <5-fold reduction in susceptibility; and a gray cell indicates the absence of susceptibility data. Cells with a plus (+) symbol indicates that the mutation had a DMS escape fraction ≥0.1. Bold mutations with a yellow background represent the consensus for one or more variants of concern or of interest. The numbers in the "in vivo" column indicate the numbers of times the mutation was selected *in vivo* during prolonged infection or in a patient receiving an mAb. The numbers in the "*in vitro*" column indicate the number of times the mutation was reported to be selected during passage in the presence of an mAb.

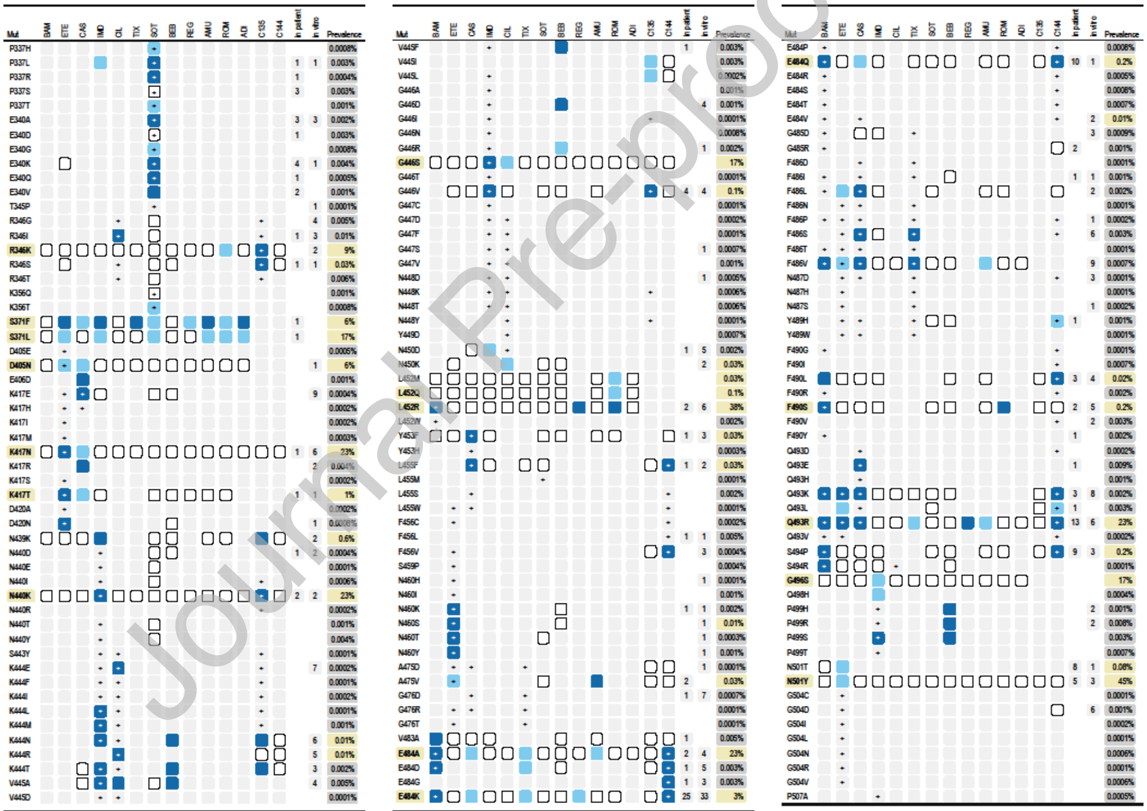


Figure 3

SARS-CoV-2 Mpro (A) and RdRp (B) resistance mutations. For Mpro, the figure shows which mutations are in the Mpro substrate binding pocket [34,35], which are associated with reduced susceptibility to nirmatrelvir (NTV) or ensitrelvir (ENS) either biochemically or in cell culture, which have been selected *in* *vitro*, the effect of mutations on Mpro fitness determined either biochemically or in cell culture, and the global mutation prevalence as of June 2022. For RdRp, the figure shows which mutations reduced susceptibility to remdesivir (RDV), which have been selected by RDV *in vitro* and *in vivo*, and the global mutation prevalence as of June 2022. A dark blue cell indicates ≥10-fold reduction in susceptibility; a light blue cell indicates 5-10-fold reduction; a very light blue cell indicates a 2.5-5-fold reduction; and a white cell indicates a <2.5-fold reduction. A gray cell indicates the absence of susceptibility data. *G15S is the consensus amino acid for the Lambda variant. †E166V has been reported in three persons receiving nirmatrelvir in the EPIC-HR study [22]. §Variable reductions in susceptibility were reported for this mutation in different studies. For RDV, S759A was evaluated only in combination with V792I; F480L and F557L were evaluated only in combination with each other.

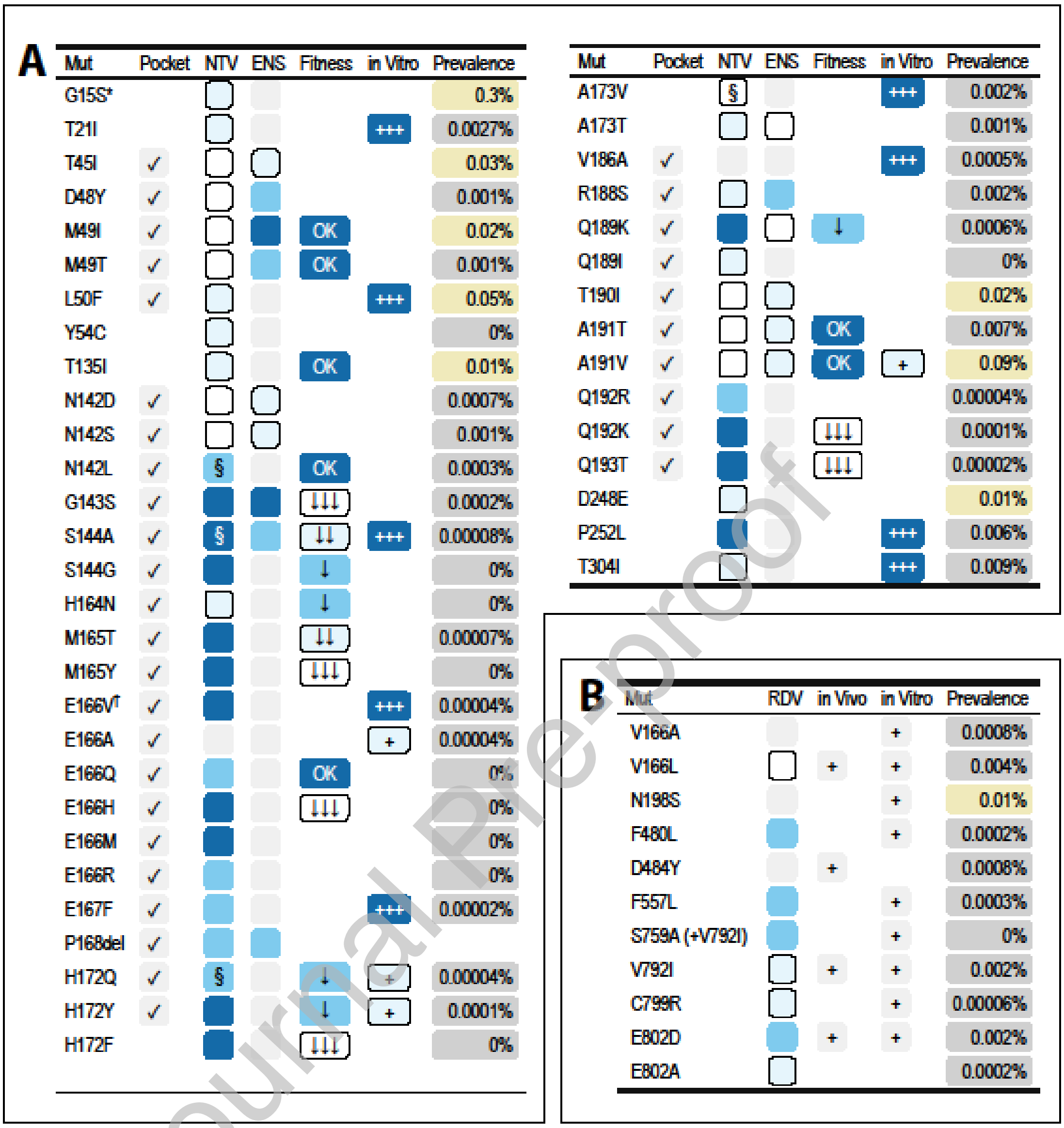


Figure 4

The numbers of Spike, RdRP, and Mpro mutations according to their global prevalence (A-C). The histograms represent the numbers of mutations on a $\log_{10}$ scale within five prevalence ranges (≥10%, 1%-10%, 0.1%-1%, 0.01%-0.1%, and <0.01%) in 4,740,761 quality-controlled sequences. Mutations that were never reported were not counted. The insets in each plot contain the actual numbers represented by the histograms.

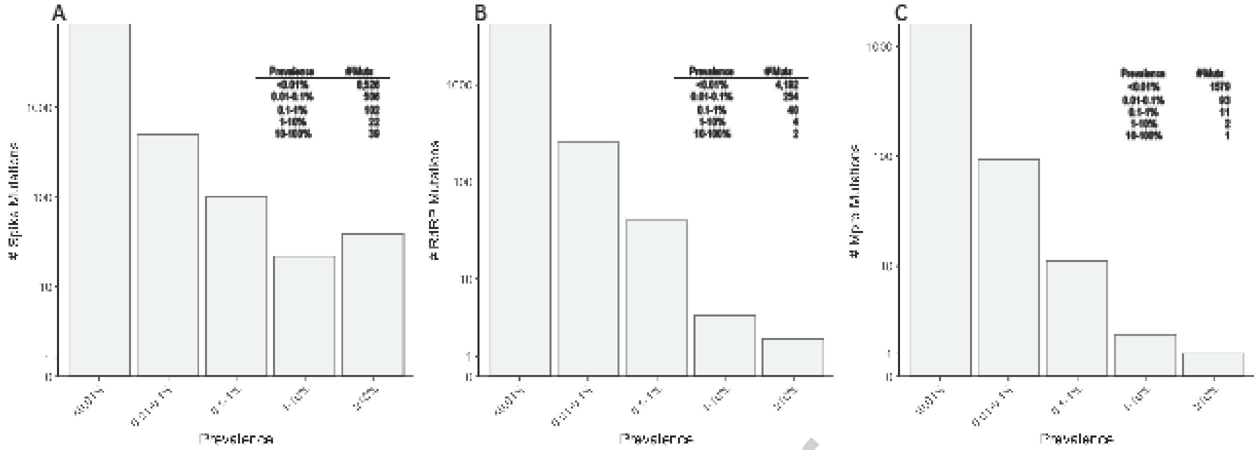


Figure 5

The distribution in the numbers of unusual mutations per sequence in Spike, RdRP, and Mpro in 4,740,761 quality-controlled sequences (A-C). The histograms represent the numbers of sequences on a $\log_{10}$ scale according to the number of unusual mutations per sequence. The insets in each plot contain the numbers represented by the first six histograms.

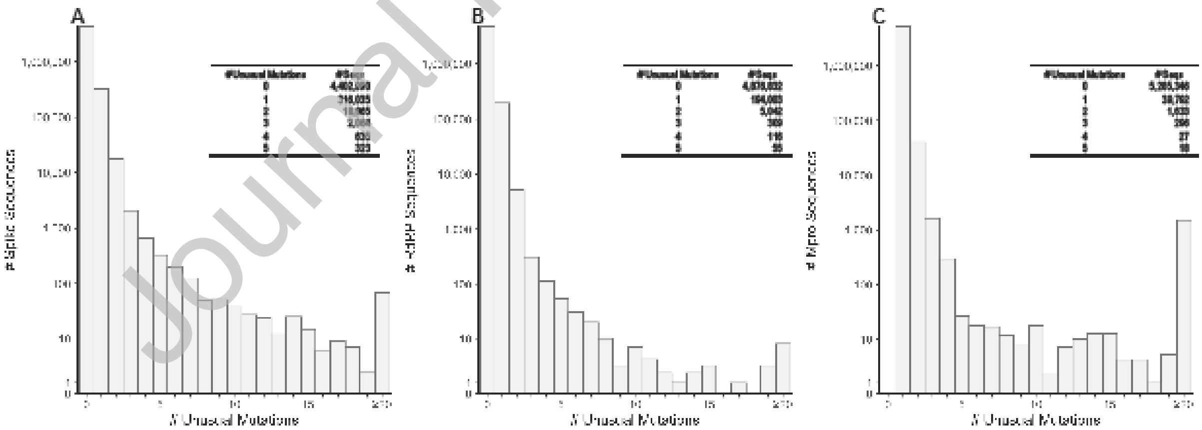


Figure 6

Box plots indicating the numbers of usual and unusual mutations per genome at different mutation thresholds for the 400 Illumina and 200 ONT sequences in the FASTQ dataset. The boxplots show the median and inter-quartile ranges (IQRs). The whiskers extend ±1.5 IQRs from the hinge. Regions for which the median read depth was <100 were excluded.

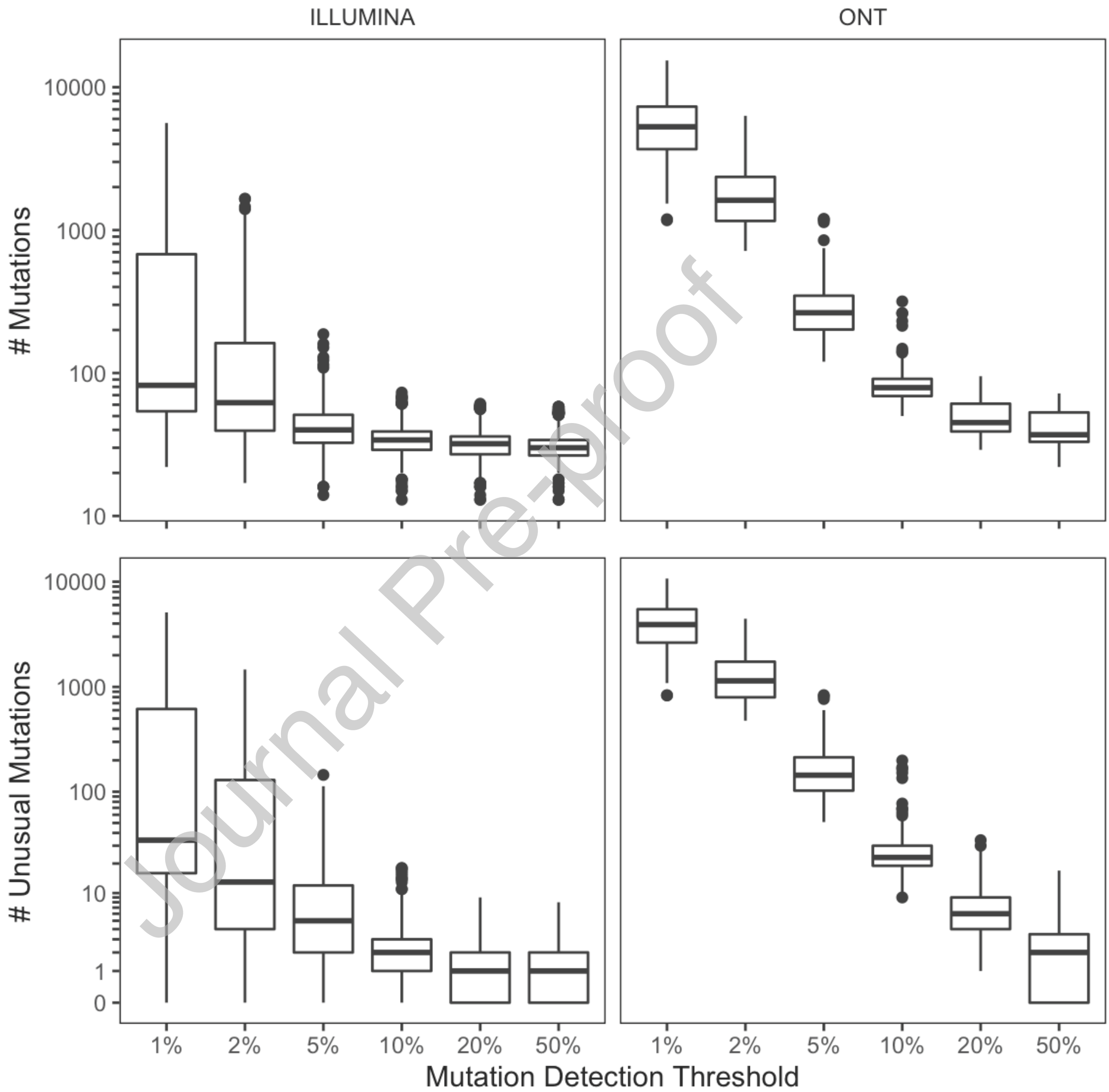


Figure 7

Availability of neutralizing susceptibility data in CoV-RDB for submitted sets of Spike receptor binding domain (RBD) mutations. The 13,578 unique patterns of RBD mutations, present in 4,740,761 sequences, were submitted to Sierra SARS-CoV-2. Exactly matching susceptibility data were available for 183 mutation patterns (1.3% of mutation patterns derived from 76.7% of sequences). Partially matching susceptibility data were available for 11,760 patterns (86.6% of patterns from 10.2% of sequences) including cases for which CoV-RDB contained data for a subset, superset, or intersecting set of mutation patterns. No matching susceptibility data were available for 1,635 mutation patterns (12.0% of patterns from 13.0% of sequences). Each of the five tables contain examples of the five scenarios: exact match, subset, superset, intersection, and no match with one column showing the submitted mutation pattern, another showing the closest CoV-RDB pattern, and the third showing the number of sequences (except for the tables showing the patterns that contained an exact match or no match in CoV-RDB).

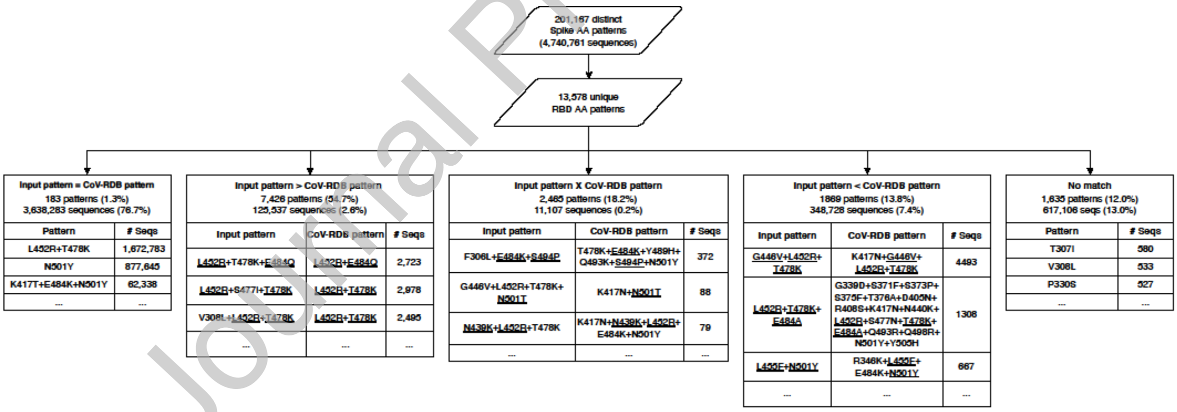


Table 1. Overview of the Sierra SARS2-CoV-2 Analysis Report

Feature	Input Type[1] FASTQ	FASTA	Mutations
Sequence summary			
Gene list	✓	✓	
PANGO lineage[2]	✓	✓	
Median read depth	✓		
Interactive mutation detection thresholds	✓		
Consensus sequence with IUPAC nucleotides[3]	✓		
Sequence quality assessment			
List of unsequenced regions	✓	✓	
List of unusual mutations	✓	✓	
List of low-coverage regions	✓		
Mutation summaries			
Prevalence of each mutation in a sample	✓		
mAb susceptibility summaries	✓	✓	✓
Mutation-specific annotation	✓	✓	✓
Convalescent and vaccinee plasma susceptibility data	✓	✓	✓

Footnote: [1]FASTQ indicates the raw data associated with an NGS platform, most commonly Illumina and Oxford Nanopore Technologies; FASTA sequences are usually derived from the consensus of NGS data. Mutations indicate user submitted amino acid differences from the consensus Wuhan-Hu-1 Spike sequence. [2]PANGO – Phylogenetic Assignment of Named Global Outbreak. [3]IUPAC – International Union of Pure and Applied Chemistry representation of nucleotide ambiguities or mixtures

REFERENCES

[1] C. Charre, C. Ginevra, M. Sabatier, H. Regue, G. Destras, S. Brun, G. Burfin, C. Scholtes, F. Morfin, M. Valette, B. Lina, A. Bal, L. Josset, Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation, Virus Evol. 6 (2020) veaa075. https://doi.org/10.1093/ve/veaa075.

[2] M. Simonetti, N. Zhang, L. Harbers, M.G. Milia, S. Brossa, T.T. Huong Nguyen, F. Cerutti, E. Berrino, A. Sapino, M. Bienko, A. Sottile, V. Ghisetti, N. Crosetto, COVseq is a cost-effective workflow for mass-scale SARS-CoV-2 genomic surveillance, Nat Commun. 12 (2021) 3903. https://doi.org/10.1038/s41467-021-24078-9.

[3] C.-C. Lo, M. Shakya, R. Connor, K. Davenport, M. Flynn, A.M. y Gutiérrez, B. Hu, P.-E. Li, E.P. Jackson, Y. Xu, P.S.G. Chain, EDGE COVID-19: a web platform to generate submission-ready genomes from SARS-CoV-2 sequencing efforts, Bioinformatics. (2022) btac176. https://doi.org/10.1093/bioinformatics/btac176.

[4] P.T. Truong Nguyen, I. Plyusnin, T. Sironen, O. Vapalahti, R. Kant, T. Smura, HAVoC, a bioinformatic pipeline for reference-based consensus assembly and lineage assignment for SARS-CoV-2 sequences, BMC Bioinformatics. 22 (2021) 373. https://doi.org/10.1186/s12859-021-04294-2.

[5] F.Z. Dezordi, A.M. da S. Neto, T. de L. Campos, P.M.C. Jeronimo, C.F. Aksenen, S.P. Almeida, G.L. Wallau, null On Behalf Of The Fiocruz Covid-Genomic Surveillance Network, ViralFlow: A Versatile Automated Workflow for SARS-CoV-2 Genome Assembly, Lineage Assignment, Mutations and Intrahost Variant Detection, Viruses. 14 (2022) 217. https://doi.org/10.3390/v14020217.

[6] J. Phelan, W. Deelder, D. Ward, S. Campino, M.L. Hibberd, T.G. Clark, COVID-profiler: a webserver for the analysis of SARS-CoV-2 sequencing data, BMC Bioinformatics. 23 (2022) 137. https://doi.org/10.1186/s12859-022-04632-y.

[7] W. Maier, S. Bray, M. van den Beek, D. Bouvier, N. Coraor, M. Miladi, B. Singh, J.R. De Argila, D. Baker, N. Roach, S. Gladman, F. Coppens, D.P. Martin, A. Lonie, B. Grüning, S.L. Kosakovsky Pond, A. Nekrutenko, Ready-to-use public infrastructure for global SARS-CoV-2 monitoring, Nat Biotechnol. (2021) 1–2. https://doi.org/10.1038/s41587-021-01069-1.

[8] R.R.M. Oliveira, T.C. Negri, G. Nunes, I. Medeiros, G. Araújo, F. de O. Silva, J.E.S. de Souza, R. Alves, G. Oliveira, PipeCoV: a pipeline for SARS-CoV-2 genome assembly, annotation and variant identification, PeerJ. 10 (2022) e13300. https://doi.org/10.7717/peerj.13300.

[9] P.L. Tzou, K. Tao, S.L.K. Pond, R.W. Shafer, Coronavirus Resistance Database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons, PLOS ONE. 17 (2022) e0261045. https://doi.org/10.1371/journal.pone.0261045.

[10] R. Paredes, P.L. Tzou, G. van Zyl, G. Barrow, R. Camacho, S. Carmona, P.M. Grant, R.K. Gupta, R.L. Hamers, P.R. Harrigan, M.R. Jordan, R. Kantor, D.A. Katzenstein, D.R. Kuritzkes, F. Maldarelli, D. Otelea, C.L. Wallis, J.M. Schapiro, R.W. Shafer, Collaborative update of a rule-based expert system for HIV-1 genotypic resistance test interpretation, PLoS ONE. 12 (2017) e0181357. https://doi.org/10.1371/journal.pone.0181357.

[11] P.L. Tzou, S.L. Kosakovsky Pond, S. Avila-Rios, S.P. Holmes, R. Kantor, R.W. Shafer, Analysis of unusual and signature APOBEC-mutations in HIV-1 pol next-generation sequences, PLoS One. 15 (2020) e0225352. https://doi.org/10.1371/journal.pone.0225352.

[12] Y. Kodama, M. Shumway, R. Leinonen, International Nucleotide Sequence Database Collaboration, The Sequence Read Archive: explosive growth of sequencing data, Nucleic Acids Res. 40 (2012) D54-56. https://doi.org/10.1093/nar/gkr854.

[13] Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality, Euro Surveill. 22 (2017) 30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

[14] H. Li, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics. 34 (2018) 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

[15] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, Bioinformatics. 25 (2009) 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

[16] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet.Journal. 17 (2011) 10–12. https://doi.org/10.14806/ej.17.1.200.

[17] N.D. Grubaugh, K. Gangavarapu, J. Quick, N.L. Matteson, J.G. De Jesus, B.J. Main, A.L. Tan, L.M. Paul, D.E. Brackney, S. Grewal, N. Gurfield, K.K.A. Van Rompay, S. Isern, S.F. Michael, L.L. Coffey, N.J. Loman, K.G. Andersen, An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar, Genome Biology. 20 (2019) 8. https://doi.org/10.1186/s13059-018-1618-7.

[18] Á. O'Toole, E. Scher, A. Underwood, B. Jackson, V. Hill, J.T. McCrone, R. Colquhoun, C. Ruis, K. Abu-Dahab, B. Taylor, C. Yeats, L. du Plessis, D. Maloney, N. Medd, S.W. Attwood, D.M. Aanensen, E.C. Holmes, O.G. Pybus, A. Rambaut, Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, Virus Evolution. 7 (2021). https://doi.org/10.1093/ve/veab064.

[19] D.P. Martin, S. Lytras, A.G. Lucaci, W. Maier, B. Grüning, S.D. Shank, S. Weaver, O.A. MacLean, R.J. Orton, P. Lemey, M.F. Boni, H. Tegally, G.W. Harkins, C. Scheepers, J.N. Bhiman, J. Everatt, D.G. Amoako, J.E. San, J. Giandhari, A. Sigal, C. Williamson, N. Hsiao, A. von Gottberg, A. De Klerk, R.W. Shafer, D.L. Robertson, R.J. Wilkinson, B.T. Sewell, R. Lessells, A. Nekrutenko, A.J. Greaney, T.N. Starr, J.D. Bloom, B. Murrell, E. Wilkinson, R.K. Gupta, T. de Oliveira, S.L. Kosakovsky Pond, Selection Analysis Identifies Clusters of Unusual Mutational Changes in Omicron Lineage BA.1 That Likely Impact Spike Function, Molecular Biology and Evolution. 39 (2022) msac061. https://doi.org/10.1093/molbev/msac061.

[20] T.N. Starr, A.J. Greaney, S.K. Hilton, D. Ellis, K.H.D. Crawford, A.S. Dingens, M.J. Navarro, J.E. Bowen, M.A. Tortorici, A.C. Walls, N.P. King, D. Veesler, J.D. Bloom, Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding, Cell. 182 (2020) 1295-1310.e20. https://doi.org/10.1016/j.cell.2020.08.012.

[21] T.N. Starr, A.J. Greaney, A. Addetia, W.W. Hannon, M.C. Choudhary, A.S. Dingens, J.Z. Li, J.D. Bloom, Prospective mapping of viral mutations that escape antibodies used to treat COVID-19, Science. 371 (2021) 850–854. https://doi.org/10.1126/science.abf9302.

[22] FDA, Fact sheet for healthcare providers: Emergency use authorization for Paxlovid, (2021). https://www.fda.gov/media/155050/download (accessed February 12, 2022).

[23] S. Iketani, H. Mohri, B. Culbertson, S.J. Hong, Y. Duan, M.I. Luck, M.K. Annavajhala, Y. Guo, Z. Sheng, A.-C. Uhlemann, S.P. Goff, Y. Sabo, H. Yang, A. Chavez, D.D. Ho,

Multiple pathways for SARS-CoV-2 resistance to nirmatrelvir, (2022) 2022.08.07.499047. https://doi.org/10.1101/2022.08.07.499047.
[24] Y. Zhou, K.A. Gammeltoft, L.A. Ryberg, L.V. Pham, U. Fahnoe, A. Binderup, C.R.D. Hernandez, A. Offersgaard, C. Fernandez-Antunez, G.H.J. Peters, S. Ramirez, J. Bukh, J.M. Gottwein, Nirmatrelvir Resistant SARS-CoV-2 Variants with High Fitness in Vitro, (2022) 2022.06.06.494921. https://doi.org/10.1101/2022.06.06.494921.
[25] D. Jochmans, C. Liu, K. Donckers, A. Stoycheva, S. Boland, S.K. Stevens, C.D. Vita, B. Vanmechelen, P. Maes, B.S. Trüeb, N. Ebert, V. Thiel, S.D. Jonghe, L. Vangeel, D. Bardiot, A. Jekle, L.M. Blatt, L. Beigelman, J.A. Symons, P. Raboisson, P. Chaltin, A. Marchand, J. Neyts, J. Deval, K. Vandyck, The substitutions L50F, E166A and L167F in SARS-CoV-2 3CLpro are selected by a protease inhibitor in vitro and confer resistance to nirmatrelvir, (2022) 2022.06.07.495116. https://doi.org/10.1101/2022.06.07.495116.
[26] E. Heilmann, F. Costacurta, A. Volland, D. von Laer, SARS-CoV-2 3CLpro mutations confer resistance to Paxlovid (nirmatrelvir/ritonavir) in a VSV-based, non-gain-of-function system, (2022) 2022.07.02.495455. https://doi.org/10.1101/2022.07.02.495455.
[27] S. Iketani, S.J. Hong, J. Sheng, F. Bahari, B. Culbertson, F.F. Atanaki, A.K. Aditham, A.F. Kratz, M.I. Luck, R. Tian, S.P. Goff, H. Montazeri, Y. Sabo, D.D. Ho, A. Chavez, Functional map of SARS-CoV-2 3CL protease reveals tolerant and immutable sites, Cell Host & Microbe. (2022). https://doi.org/10.1016/j.chom.2022.08.003.
[28] G.D. Noske, E. de S. Silva, M.O. de Godoy, I. Dolci, R.S. Fernandes, R.V.C. Guido, P. Sjö, G. Oliva, A.S. Godoy, Structural basis of nirmatrelvir and ensitrelvir resistance profiles against SARS-CoV-2 Main Protease naturally occurring polymorphisms, (2022) 2022.08.31.506107. https://doi.org/10.1101/2022.08.31.506107.
[29] Y. Hu, E.M. Lewandowski, H. Tan, R.T. Morgan, X. Zhang, L.M.C. Jacobs, S.G. Butler, M.V. Mongora, J. Choy, Y. Chen, J. Wang, Naturally occurring mutations of SARS-CoV-2 main protease confer drug resistance to nirmatrelvir, (2022) 2022.06.28.497978. https://doi.org/10.1101/2022.06.28.497978.
[30] V.M. de Oliveira, M.F. Ibrahim, X. Sun, R. Hilgenfeld, J. Shen, H172Y mutation perturbs the S1 pocket and nirmatrelvir binding of SARS-CoV-2 main protease through a nonnative hydrogen bond, (2022) 2022.07.31.502215. https://doi.org/10.1101/2022.07.31.502215.
[31] V.M. Sasi, S. Ullrich, J. Ton, S.E. Fry, J. Johansen-Leete, R.J. Payne, C. Nitsche, C.J. Jackson, Predicting antiviral resistance mutations in SARS-CoV-2 main protease with computational and experimental screening, (2022) 2022.08.24.505060. https://doi.org/10.1101/2022.08.24.505060.
[32] S.A. Moghadasi, E. Heilmann, S.N. Moraes, F.L. Kearns, D. von Laer, R.E. Amaro, R.S. Harris, Transmissible SARS-CoV-2 variants with resistance to clinical protease inhibitors, (2022) 2022.08.07.503099. https://doi.org/10.1101/2022.08.07.503099.
[33] J. Ou, E.M. Lewandowski, Y. Hu, A.A. Lipinski, R.T. Morgan, L.M.C. Jacobs, X. Zhang, M.J. Bikowitz, P. Langlais, H. Tan, J. Wang, Y. Chen, J.S. Choy, A yeast-based system to study SARS-CoV-2 Mpro structure and to identify nirmatrelvir resistant mutations, (2022) 2022.08.06.503039. https://doi.org/10.1101/2022.08.06.503039.
[34] A.M. Shaqra, S.N. Zvornicanin, Q.Y.J. Huang, G.J. Lockbaum, M. Knapp, L. Tandeske, D.T. Bakan, J. Flynn, D.N.A. Bolon, S. Moquin, D. Dovala, N. Kurt Yilmaz, C.A. Schiffer, Defining the substrate envelope of SARS-CoV-2 main protease to predict and avoid drug resistance, Nat Commun. 13 (2022) 3556. https://doi.org/10.1038/s41467-022-31210-w.

[35] K.S. Yang, S.Z. Leeuwon, S. Xu, W.R. Liu, Evolutionary and Structural Insights about Potential SARS-CoV-2 Evasion of Nirmatrelvir, J. Med. Chem. 65 (2022) 8686–8698. https://doi.org/10.1021/acs.jmedchem.2c00404.

[36] S. Gandhi, J. Klein, A.J. Robertson, M.A. Peña-Hernández, M.J. Lin, P. Roychoudhury, P. Lu, J. Fournier, D. Ferguson, S.A.K. Mohamed Bakhash, M. Catherine Muenker, A. Srivathsan, E.A. Wunder, N. Kerantzas, W. Wang, B. Lindenbach, A. Pyle, C.B. Wilen, O. Ogbuagu, A.L. Greninger, A. Iwasaki, W.L. Schulz, A.I. Ko, De novo emergence of a remdesivir resistance mutation during treatment of persistent SARS-CoV-2 infection in an immunocompromised patient: a case report, Nat Commun. 13 (2022) 1547. https://doi.org/10.1038/s41467-022-29104-y.

[37] J.I. Hogan, R. Duerr, D. Dimartino, C. Marier, S.E. Hochman, S. Mehta, G. Wang, A. Heguy, Remdesivir resistance in transplant recipients with persistent COVID-19, Clinical Infectious Diseases. (2022) ciac769. https://doi.org/10.1093/cid/ciac769.

[38] M. Martinot, A. Jary, S. Fafi-Kremer, V. Leducq, H. Delagreverie, M. Garnier, J. Pacanowski, A. Mékinian, F. Pirenne, P. Tiberghien, V. Calvez, C. Humbrecht, A.-G. Marcelin, K. Lacombe, Emerging RNA-Dependent RNA Polymerase Mutation in a Remdesivir-Treated B-cell Immunodeficient Patient With Protracted Coronavirus Disease 2019, Clin Infect Dis. 73 (2021) e1762–e1765. https://doi.org/10.1093/cid/ciaa1474.

[39] A. Heyer, T. Günther, A. Robitaille, M. Lütgehetmann, M.M. Addo, D. Jarczak, S. Kluge, M. Aepfelbacher, J. Schulze zur Wiesch, N. Fischer, A. Grundhoff, Remdesivir-induced emergence of SARS-CoV2 variants in patients with prolonged infection, Cell Reports Medicine. 3 (2022) 100735. https://doi.org/10.1016/j.xcrm.2022.100735.

[40] L.J. Stevens, A.J. Pruijssers, H.W. Lee, C.J. Gordon, E.P. Tchesnokov, J. Gribble, A.S. George, T.M. Hughes, X. Lu, J. Li, J.K. Perry, D.P. Porter, T. Cihlar, T.P. Sheahan, R.S. Baric, M. Götte, M.R. Denison, Distinct genetic determinants and mechanisms of SARS-CoV-2 resistance to remdesivir, (2022) 2022.01.25.477724. https://doi.org/10.1101/2022.01.25.477724.

[41] L. Checkmahomed, J. Carbonneau, V. Du Pont, N.C. Riola, J.K. Perry, J. Li, B. Paré, S.M. Simpson, M.A. Smith, D.P. Porter, G. Boivin, In Vitro Selection of Remdesivir-Resistant SARS-CoV-2 Demonstrates High Barrier to Resistance, Antimicrobial Agents and Chemotherapy. 0 (n.d.) e00198-22. https://doi.org/10.1128/aac.00198-22.

[42] A.M. Szemiel, A. Merits, R.J. Orton, O.A. MacLean, R.M. Pinto, A. Wickenhagen, G. Lieber, M.L. Turnbull, S. Wang, W. Furnon, N.M. Suarez, D. Mair, A. da S. Filipe, B.J. Willett, S.J. Wilson, A.H. Patel, E.C. Thomson, M. Palmarini, A. Kohl, M.E. Stewart, In vitro selection of Remdesivir resistance suggests evolutionary predictability of SARS-CoV-2, PLOS Pathogens. 17 (2021) e1009929. https://doi.org/10.1371/journal.ppat.1009929.

[43] M.L. Agostini, E.L. Andres, A.C. Sims, R.L. Graham, T.P. Sheahan, X. Lu, E.C. Smith, J.B. Case, J.Y. Feng, R. Jordan, A.S. Ray, T. Cihlar, D. Siegel, R.L. Mackman, M.O. Clarke, R.S. Baric, M.R. Denison, Coronavirus Susceptibility to the Antiviral Remdesivir (GS-5734) Is Mediated by the Viral Polymerase and the Proofreading Exoribonuclease, MBio. 9 (2018) e00221-18. https://doi.org/10.1128/mBio.00221-18.

[44] P.S.-W. Yeung, H. Wang, M. Sibai, D. Solis, F. Yamamoto, N. Iwai, B. Jiang, N. Hammond, B. Truong, S. Bihon, S. Santos, M. Mar, C. Mai, K.O. Mfuh, J.A. Miller, C. Huang, M.K. Sahoo, J.L. Zehnder, B.A. Pinsky, Evaluation of a Rapid and Accessible Reverse Transcription-Quantitative PCR Approach for SARS-CoV-2 Variant of Concern Identification, J Clin Microbiol. 60 (n.d.) e00178-22. https://doi.org/10.1128/jcm.00178-22.

[45] S. Khare, C. Gurry, L. Freitas, M.B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R.T. Lee, W. Yeo, G.C. Curation Team, S. Maurer-Stroh, GISAID's Role in Pandemic Response, China CDC Wkly. 3 (2021) 1049–1051. https://doi.org/10.46234/ccdcw2021.255.

[46] S.L.K. Pond, SARS-CoV-2-variation/variation-new at master · spond/SARS-CoV-2-variation, GitHub. (2022). https://github.com/spond/SARS-CoV-2-variation (accessed October 3, 2022).

[47] D.P. Martin, S. Weaver, H. Tegally, J.E. San, S.D. Shank, E. Wilkinson, A.G. Lucaci, J. Giandhari, S. Naidoo, Y. Pillay, L. Singh, R.J. Lessells, R.K. Gupta, J.O. Wertheim, A. Nekturenko, B. Murrell, G.W. Harkins, P. Lemey, O.A. MacLean, D.L. Robertson, T. de Oliveira, S.L. Kosakovsky Pond, The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages, Cell. 184 (2021) 5189-5200.e7. https://doi.org/10.1016/j.cell.2021.09.003.

[48] E.B. Hodcroft, CoVariants: SARS-CoV-2 Mutations and Variants of Interest., (2021). https://covariants.org/ (accessed May 29, 2022).

[49] M. McCallum, N. Czudnochowski, L.E. Rosen, S.K. Zepeda, J.E. Bowen, A.C. Walls, K. Hauser, A. Joshi, C. Stewart, J.R. Dillen, A.E. Powell, T.I. Croll, J. Nix, H.W. Virgin, D. Corti, G. Snell, D. Veesler, Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement, Science. (2022) eabn8652. https://doi.org/10.1126/science.abn8652.

[50] K. Gangavarapu, A.A. Latiff, J.L. Mullen, M. Alkuzweny, E. Hufbauer, G. Tsueng, E. Haag, M. Zeller, C.M. Aceves, K. Zaiets, M. Cano, J. Zhou, Z. Qian, R. Sattler, N.L. Matteson, J.I. Levy, M.A. Suchard, C. Wu, A.I. Su, K.G. Andersen, L.D. Hughes, Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations, (2022) 2022.01.27.22269965. https://doi.org/10.1101/2022.01.27.22269965.

[51] W. Yin, Y. Xu, P. Xu, X. Cao, C. Wu, C. Gu, X. He, X. Wang, S. Huang, Q. Yuan, K. Wu, W. Hu, Z. Huang, J. Liu, Z. Wang, F. Jia, K. Xia, P. Liu, X. Wang, B. Song, J. Zheng, H. Jiang, X. Cheng, Y. Jiang, S.-J. Deng, H.E. Xu, Structures of the Omicron spike trimer with ACE2 and an anti-Omicron antibody, Science. 375 (2022) 1048–1053. https://doi.org/10.1126/science.abn8863.

[52] N. Ikemura, S. Taminishi, T. Inaba, T. Arimori, D. Motooka, K. Katoh, Y. Kirita, Y. Higuchi, S. Li, T. Suzuki, Y. Itoh, Y. Ozaki, S. Nakamura, S. Matoba, D.M. Standley, T. Okamoto, J. Takagi, A. Hoshino, An engineered ACE2 decoy neutralizes the SARS-CoV-2 Omicron variant and confers protection against infection in vivo, Science Translational Medicine. 0 (2022) eabn7737. https://doi.org/10.1126/scitranslmed.abn7737.

[53] Y. Cao, F. Jian, J. Wang, Y. Yu, W. Song, A. Yisimayi, J. Wang, R. An, N. Zhang, Y. Wang, P. Wang, L. Zhao, H. Sun, L. Yu, S. Yang, X. Niu, T. Xiao, Q. Gu, F. Shao, X. Hao, Y. Xu, R. Jin, Y. Wang, X.S. Xie, Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution, (2022) 2022.09.15.507787. https://doi.org/10.1101/2022.09.15.507787.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: