Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Sierra SARS-CoV-2 sequence and antiviral resistance analysis program

Philip L. Tzou a,*, Kaimeing Tao a, Malaya K. Sahoo b, Sergei L. Kosakovsky Pond c, Benjamin A. Pinsky a, b, Robert W. Shafer a

a Division of Infectious Diseases, Department of Medicine, Stanford University, Stanford, CA, USA
b Department of Pathology, Stanford University, Stanford, CA, USA
c Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA

ARTICLE INFO

Keywords:
SARS-CoV-2
Genomic sequencing
Mutations
Antiviral resistance

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 interpreting Spike mutations, 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.

1. 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 interpreting Spike mutations, 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.

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.

2. 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 comma-delimited 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. Fig. 1 illustrates the workflow of Sierra SARS2-CoV-2 for all three input types.

2.1. 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 SARS-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.

### Table 1

| Feature                        | Input Type1 | FASTQ | FASTA | Mutations |
|--------------------------------|-------------|-------|-------|-----------|
| **Sequence summary**           |             | ✓     | ✓     | ✓         |
| Gene list                      | ✓           | ✓     | ✓     | ✓         |
| PANGO lineage2                 | ✓           | ✓     | ✓     | ✓         |
| Median read depth              | ✓           | ✓     | ✓     | ✓         |
| Interactive mutation detection thresholds | ✓         | ✓     | ✓     | ✓         |
| Consensus sequence with IUPAC nucleotides3 | ✓     | ✓     | ✓     | ✓         |
| **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 | ✓     | ✓     | ✓     | ✓         |

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

2.2. 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 converted 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].

2.3. 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.

2.4. 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 Fred Hutchinson Cancer Research Center [20,21]. As of September 2022, there were 488 spike mutations meeting these criteria. Fig. 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.

2.5. 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. Fig. 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].

2.6. Datasets used for benchmarking and validation

2.6.1. 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 \(--\text{skip-technical}\) to exclude adapters, primers, and bar-codes from the downloaded FASTQ file. The SUH sequences were generated using a recently published pipeline [44].

### 2.6.2. FASTA files

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

### 2.6.3. 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 32,798 RdRp distinct mutation patterns [46, 47].

### 3. Results

#### 3.1. 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 \(\geq 5\) and a mutation-detection threshold of 50% were used.

#### 3.1.1. 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 \((n=10)\).

#### 3.1.2. 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.

3.2. 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 CoVServer program [45] for 963,237 FASTA sequences. The list of mutations for Mpro, RdRp, 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.

3.3. Mutation lists

3.3.1. Distribution of usual and unusual mutations

Fig. 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.

Fig. 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.

Fig. 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.

3.3.2. 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] (Fig. 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.

4. 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 datasets, 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.
Fig. 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). 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).

Declaration of Competing Interest

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.

Acknowledgments

PLT, KT, and RWS have been funded in part by a grant from the NIH/NIAID: AI134384. The funder played no role in this study.

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jcv.2022.105323.

References

[1] C. Charre, G. Gieveva, M. Sabatier, H. Regue, G. Destra, S. Brun, G. Burfin, C. Scholtes, M. Forlin, M. Velleze, B. Lina, A. Bal, L. Josset, Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation, Virus Evol. 6 (2020), veea075, https://doi.org/10.1093/viruses/veea075.

[2] M. Simonetti, N. Zhang, L. Harbers, M.G. Milia, S. Brossa, T.T. Huong Nguyen, C.-C. Lo, M. Shakya, R. Connor, K. Davenport, M. Flynn, A.M. y Gutiérrez, S. Isern, S.F. Michael, L.L. Coffey, N.J. Loman, K.G. Andersen, An amplicon-based sequencing pipeline for SARS-CoV-2 lineage assignment for SARS-CoV-2 sequences, BMC Bioinf. 22 (2021), 373, https://doi.org/10.1093/bmcbi/biaa076.

[3] P.T. Truong Nguyen, I. Flynnsin, 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 Bioinf. 22 (2021), 373, https://doi.org/10.1093/biomedicines/biaa076.

[4] F.Z. Dezordi, A.M. da S. Neto, T. de L. Campos, P.M.C. Jeronimo, C.F. Aksenen, S. Shakya, R. Connor, K. Davenport, M. Flynn, A.M. y Gutiérrez, S. Isern, S.F. Michael, L.L. Coffey, N.J. Loman, K.G. Andersen, An amplicon-based sequencing pipeline for SARS-CoV-2 lineage assignment for SARS-CoV-2 sequences, BMC Bioinf. 22 (2021), 373, https://doi.org/10.1093/biomedicines/biaa076.

[5] P.T. Truong Nguyen, I. Flynnsin, 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 Bioinf. 22 (2021), 373, https://doi.org/10.1093/biomedicines/biaa076.

[6] J. Phelan, W. Deelder, D. Ward, S. Campino, M.L. Hibberd, T.G. Clark, COVID-19: A web platform for the analysis of SARS-CoV-2 sequencing data, Nucleic. Acids. Res. 40 (2012) D54–D56, https://doi.org/10.1093/nar/gkr854.

[7] 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–D56, https://doi.org/10.1093/nar/gkr854.

[8] R.K. Gupta, R.L. Hamers, P.R. Harrigan, M.R. Jordan, R. Kantor, D.A. Katzenstein, D.R. Kursitzke, F. Maldarelli, D. Otelet, 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.

[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, V. Hill, J.T. McCrone, A. Nekrutenko, A.J. Greaney, K. Abu-Dahab, B. Taylor, C. Yeats, L. du Plessis, D. Maloney, C. Scheepers, J.N. Bhiman, J. Everatt, D.G. Amoako, J.E. San, J. Giandhari, A. De Klerk, 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.

[11] Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data to vision to reality, Euro Surveill. 22 (2017), 20494, https://doi.org/10.2807/1560-7917.ES.2017.22.30494.

[12] N.D. Grubaugh, K. Gangavarapu, J. Quick, N.L. Matteson, J.G. De Jesus, B.J. Main, J. McCauley, 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.

[13] Y. Li, B. Handhaker, A. Wysock, T. Fenniell, 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/btp452.

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

[15] 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 intra-host viral diversity using PrimaSeq and iVar, Genome Biol. 20 (2019), 8, https://doi.org/10.1186/s13059-018-1611-7.

[16] A. 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. Meded, S.W. Atwood, D.M. Amsen, E.C. Holmes, O.G. Pybus, A. Rambaut, Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, Virus Evol. 7 (2021), https://doi.org/10.1093/viruses/veya064.

[17] P.L. Tzou, L.S. Kosakovskiy Pond, S. Aria-Vion, 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.

[18] 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–D56, https://doi.org/10.1093/nar/gkr854.

[19] A. 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. Meded, S.W. Atwood, D.M. Amsen, E.C. Holmes, O.G. Pybus, A. Rambaut, Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool, Virus Evol. 7 (2021), https://doi.org/10.1093/viruses/veya064.
E. Heilmann, F. Costacurta, A. Volland, D. von Laer, SARS-CoV-2 3CLpro mutations

V.M. de Oliveira, M.F. Ibrahim, X. Sun, R. Hilgenfeld, J. Shen, H172Y mutation

A.M. Shaqra, S.N. Zvornicanin, Q.Y.J. Huang, G.J. Lockbaum, M. Knapp, V.M. Sasi, S. Ullrich, J. Ton, S.E. Fry, J. Johansen-Leete, R.J. Payne, C. Nitsche, C.J. G.D. Noske, E. de S. Silva, M.O. de Godoy, I. Dolci, R.S. Fernandes, R.V.C. Guido, P.

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 drug confer resistance to nirmatrelvir, (2022) 2022.06.07.503099. https://doi.org/10.1101/2022.06.07.503099.

S. Betani, H. Mohri, B. Culbertson, S.J. Hong, Y. Duan, M.I. Luck, M.K. Annamalai, Y. Guo, Z. Sheng, A.-C. Uhlemann, S.P. Goff, Y. Sabe, 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.

Y. Zhou, K.A. Gammeltoft, L.V. Ryberg, L.V. Pham, U. Fahnoe, A. Binderup, C.R.D. Hernandez, A. Offergaard, C. Fernandes-Attard, G.J.H. Peters, S. Ramirez, J. Bulb, J.M. Gottweia, 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.

B. Lindenbach, A. Pyle, C.B. Wilen, O. Ogbuagu, A.L. Greninger, A. Iwasaki, W. Virgin, D. Corti, G. Snell, D. Veesler, Structural basis of SARS-CoV-2 Omicron RBD evolution, (2022) 2022.09.15.507787. https://doi.org/10.1126/science.abn7737.

L. Robertson, T. de Oliveira, S.L. Kosakovsky Pond, The emergence and ongoing evolution and variation of SARS-CoV-2 in the UK, (2022) 2022.01.27.22269965. https://doi.org/10.1101/2022.01.27.22269965.

S. Li, Y. Cao, X. Zhang, S. Wu, S. Luo, C. Hu, Z. Chen, Q. Yang, Z. Wang, Naturally occurring mutations of SARS-CoV-2 main protease confer resistance to nirmatrelvir, (2022) 2022.06.28.497978. https://doi.org/10.1126/science.aaz2282.

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 nonnucleotide hydrophobic bond, (2022) 2022.07.31.502215. https://doi.org/10.1101/2022.07.31.502215.

G.D. Noske, E. de S. Silva, M.O. de Godoy, I. Dolci, R.S. Fernandes, R.V.C. Guido, P. Sjo, G. Oliva, A.S. Gural, Structural basis of nirmatrelvir and remdesivir resistance profiles against SARS-CoV-2 Main Protease naturally occurring polymorphisms, (2022) 2022.08.31.501607. https://doi.org/10.1126/science.aaz4022.

T.N. Starr, A.J. Greaney, S.K. Hilton, D. Ellis, K.H.D. Crawford, A.S. Dingens, M.

P.S.-W. Yeung, H. Wang, M. Sibai, D. Solis, F. Yamamoto, N. Iwai, B. Jiang, N. Hasegawa, K. Nakamura, B. Huang, M. Mar, C. Kai, K.O. Mfuh, J.A. Miller, C. Huang, M.K. Sahoo, J.L. Zehnder, B.A. Pinky, 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.

S. Khare, C. Gurry, L. Freitas, M.R. Schultz, G. Bach, D. Aliado, N. Akite, J. Ho, R. T. Lee, W. Yee, G.C. Carction Team, S. Maurer-Stroh, GISAID’s Role in Pandemic Response, China CDC Weekly 3 (2021) 1049-1051, https://doi.org/10.46236/ccw2021.055.

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.

D.P. Martin, S. Weaver, H. Tegally, J.E. San, D.S. Shank, E. Wilkinson, A.G. Lucaci, J. Giandhari, S. Naidoo, Y. Pillay, L. Singh, R.J. Lessells, R.K. Gupta, J. O. Wertheim, A. Nekrubenko, B. Murrell, G.W. Harkins, P. Leemey, O.A. Maclean, D. L. Robertson, T. de Oliveira, S.L. Kosakovsky Pond, The emergence and ongoing convergent evolution of the SARS-CoV-2 NS510Y lineages, Cell 184 (2021), e7, 10.1016/j.cell.2021.07.093.

M. McCallum, N. Czudnowski, L.E. Rosen, S.K. Zepeda, J.E. Bowen, A.C. Walls, K. Hauer, A. Joshi, C. Stewart, J.R. Dillen, A.E. Powell, T.I. Croll, J. Nic, H. W. Virgin, D. Cirt, G. Smell, D. Veesler, Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement, Science (2022), eabc6652. https://doi.org/10.1126/science.abc6652.

K. Kangavarapu, A.A. Latif, J.I. Mullen, A. Alkuzzeyw, E. Hufbauer, T. Tsueng, E. Haug, M. Zeller, C.M. Acesve, K. Zaies, M. Cano, J. Zhou, Z. Qian, R. Sattler, N.L. Matteson, J.J. Levy, M.A. Suchard, C. Wu, A.J. Su, K.G. Andersen, L.D. Hughes, Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants, (2022) 2022.01.25.477724. https://doi.org/10.1126/science.aab4263.

S. Martinot, A. Jary, S. Safi-Kremer, V. Leducu, H. Delagreviere, M. Garnier, J. Pacanowski, A. Mekinian, F. Pireneze, P. Tiberghien, V. Calvez, C. Humbrecht, A.-G. Morand, K. Lacombe, Emerging RNA-dependent RNA Polymerase Mutations in a Remdesivir-Bred Cell Immunodeficient Patient With Protracted Coronavirus Disease 2019, Clin. Infect. Dis. 73 (2021) e1762-e1765, https://doi.org/10.1093/cid/ciaa474.

A. Heyeg, Remedivir resistance in transplant recipients with persistent COVID-19, Clin. Infect. Dis. 2022, ciac769, https://doi.org/10.1093/cid/ciac769.

J.H. Hogan, R. Duerer, D. Dimantino, C. Marier, S.E. Hochman, S. Mehta, G. Wang, A. Heyeg, Remedivir resistance in transplant recipients with persistent COVID-19, Clin. Infect. Dis. 2022, ciac769, https://doi.org/10.1093/cid/ciac769.