Virologica Sinica 39 (2024) 156–168

Contents lists available at ScienceDirect

Virologica Sinica

journal homepage: www.keaipublishing.com/en/journals/virologica-sinica
www.virosin.org

Research Article

RNA barcode segments for SARS-CoV-2 identification from HCoVs and SARSr-CoV-2 lineages

Changqiao You [a,1], Shuai Jiang [a,1], Yunyun Ding [a,1], Shunxing Ye [b], Xiaoxiao Zou [a], Hongming Zhang [a], Zeqi Li [a], Fenglin Chen [a], Yongliang Li [a,*], Xingyi Ge [a,*], Xinhong Guo [a,*]

[a] College of Biology, Hunan University, Changsha, 410082, China
[b] College of Bioscience and Biotechnology, Hunan Agricultural University, Changsha, 410128, China

ARTICLE INFO

Keywords:
RNA barcode segments
SARS-CoV-2 variants and related lineages
HCoVs
Genetic tests
Complete genome sequences

ABSTRACT

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the pathogen responsible for coronavirus disease 2019 (COVID-19), continues to evolve, giving rise to more variants and global reinfections. Previous research has demonstrated that barcode segments can effectively and cost-efficiently identify specific species within closely related populations. In this study, we designed and tested RNA barcode segments based on genetic evolutionary relationships to facilitate the efficient and accurate identification of SARS-CoV-2 from extensive virus samples, including human coronaviruses (HCoVs) and SARSr-CoV-2 lineages. Nucleotide sequences sourced from NCBI and GISAID were meticulously selected and curated to construct training sets, encompassing 1733 complete genome sequences of HCoVs and SARSr-CoV-2 lineages. Through genetic-level species testing, we validated the accuracy and reliability of the barcode segments for identifying SARS-CoV-2. Subsequently, 75 main and subordinate species-specific barcode segments for SARS-CoV-2, located in *ORF1ab*, *S*, *E*, *ORF7a*, and *N* coding sequences, were intercepted and screened based on single-nucleotide polymorphism sites and weighted scores. Post-testing, these segments exhibited high recall rates (nearly 100%), specificity (almost 30% at the nucleotide level), and precision (100%) performance on identification. They were eventually visualized using one and two-dimensional combined barcodes and deposited in an online database (http://virusbarcodedatabase.top/). The successful integration of barcoding technology in SARS-CoV-2 identification provides valuable insights for future studies involving complete genome sequence polymorphism analysis. Moreover, this cost-effective and efficient identification approach also provides valuable reference for future research endeavors related to virus surveillance.

1. Introduction

The causative agent of coronavirus disease 2019 (COVID-19) is severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), categorized as a positive-sense, single-stranded RNA viruses (Kirtipal et al., 2020). It falls under the subgenus *Sarbecovirus*, genus *Betacoronavirus*, in the family *Coronaviridae*. Over the past few years, significant challenges related to SARS-CoV-2, such as genome sequencing (Nimavat et al., 2021), protein structure prediction (Swanson et al., 2020), and genetic lineage construction (Peng et al., 2020), have gradually been addressed. Until now, identification techniques for detecting SARS-CoV-2 infections in populations primarily include the following: first, nucleotide sampling and PCR testing of large populations or individuals in specific areas (Chaimayo et al., 2020; Meng et al., 2022); and second, utilizing chemiluminescence technology to detect the serum-specific antibody IgM and IgG levels in populations (Selingerova et al., 2021). However, these technologies demand a significant investment of time and human resources, and they exhibit limited repeatability and accuracy in identifying viral infections (Meng et al., 2022; Selingerova et al., 2021).

SARS-CoV-2 has diverged into numerous subtypes, and experts anticipate its persistence in humans, akin to influenza viruses (Ullah et al., 2021). Since the discovery of the first human coronavirus (HCoV), seven HCoVs have emerged as significant threats to human society, making the investigation of genetic differences and similarities a vibrant area of study in virology (Kirtipal et al., 2020; Zhou et al., 2021). Additionally, strains within SARS-CoV-2 related (SARSr-CoV-2) lineages exhibit a high degree of sequence identity with SARS-CoV-2, posing challenges for its identification. Regrettably, insufficient work has been

* Corresponding authors.
E-mail addresses: lyl13618481357@hnu.edu.cn (Y. Li), xyge@hnu.edu.cn (X. Ge), gxh@hnu.edu.cn (X. Guo).
[1] Changqiao You, Shuai Jiang and Yunyun Ding contributed equally to this work.

https://doi.org/10.1016/j.virs.2024.01.006
Received 8 April 2023; Accepted 17 January 2024
Available online 20 January 2024

done in the technical realm of leveraging HCoVs and SARSr-CoV-2 genetic diversity for SARS-CoV-2 identification (Cosar et al., 2022). Databases serving as the foundation for identifying SARS-CoV-2 are also rare; this data scarcity significantly hinders the progress of SARS-CoV-2 identification technologies and SARS-CoV-2 population genetics research (Tan et al., 2023). Scientists have endeavored to develop effective methods for identifying SARS-CoV-2. However, due to the limited quantity of sequencing data in early studies, successfully recovering genetic markers specific to a particular species remained challenging. Frequently, the identified genetic markers were lengthy and required simultaneous utilization, resulting in a lack of adaptability (Lam et al., 2020; Cohen-Aharonov et al., 2022). Despite the widespread use of SARS-CoV-2 vaccinations, the emergence of new mutant strains (BF. 7, BQ. 1.1, BA. 2.75, XBB.1.16, XBB.1.9.2 and XBB.2.3, etc.) continues to pose a global threat to individual safety (Amiral and Seghatchian, 2022; GISAID, 2023; WHO, 2023). Therefore, the development of a universal and consistent identification system for coronavirus is crucial to assist medical professionals in effectively responding to future crises.

Paul Hebert initially proposed barcoding technology in 2003 (Hebert et al., 2003). Barcode segments leverage high-throughput sequencing techniques to mine genetic information about species from sequence data, providing more specific molecular genetic markers than conventional biometrics (Sheth and Thaker, 2017). The use of these highly discriminative markers enables rapid, accurate, and high-throughput analysis, allowing for the identification and detection of species in complex settings and the determination of species-wide mutations. In recent years, the application of similar barcoding technology has steadily expanded, with numerous research studies demonstrating its potential for exceptional breakthroughs in viral detection (Minervina et al., 2022; Westhaus et al., 2020; Lam et al., 2020; Langat et al., 2021). Lam et al. sequenced a virus population in tissue samples of pangolins, extracting species-specific markers such as virus isolate (GX/P2V), reads, and contigs to identify SARS-CoV-2-related coronaviruses in pangolins (Lam et al., 2020). Langat et al. employed metabarcoding to profile RNA viruses in the Biting Midges (*Ceratopodidae*) and identify the insect hosts associated with these viruses (Langat et al., 2021). Consequently, screening barcode segments based on barcoding technology (or similar techniques) can assist researchers in understanding the mechanisms of virus evolution and transmission, facilitating rapid responses to large-scale virus infections (Guan et al., 2020). Furthermore, barcoding technology is anticipated to deliver more precise and valuable information for future viral research and vaccine development (Guan et al., 2020; Lago et al., 2020; Mahima et al., 2022). Hence, barcode segment research holds immense research significance and value.

Following established design principles, this study integrated and refined the RNA barcode research technique for SARS-CoV-2. The goal was to develop barcode segments, evaluate their identification accuracy, reliability, and generalizability, all in the interest of efficiently identifying SARS-CoV-2 within unknown virus samples. Based on the principles of genetic similarity and evolution among various variants of SARS-CoV-2, HCoVs, and SARSr-CoV-2 lineages, we constructed multiple distinct test sets for SARS-CoV-2. These sets underwent assessments involving nucleic acid and sequence polymorphism, genetic distance (GEDI) matrix analysis, and phylogenetic tree analysis. Subsequently, we utilized single-nucleotide polymorphism (SNP) sites of SARS-CoV-2 and tools such as the basic local alignment search tool (BLAST) to acquire and filter high-quality barcode segments. These segments were ultimately evaluated for their identification capabilities using multiple test datasets (Zhou et al., 2021). Finally, the obtained SARS-CoV-2 species-specific barcode segments, along with their related information, were stored within visual barcodes and uploaded to an online database. Researchers could conveniently access this information by scanning these barcodes using mobile electronic devices or by visiting the provided web address. The visual barcodes and the barcode segment database established in this study would greatly facilitate the dissemination of barcode technology-related concepts, expand researchers' understanding of SARS-CoV-2, and provide new insights for the efficient identification of SARS-CoV-2 or other species.

2. Materials and methods

2.1. Establishment of SDs

The raw sequence data collected for training sets required additional authoritative annotations to ensure consistent localization of segments within distinct viral strains (Wu et al., 2020). Common barcode screening procedures included tests for nucleotide site diversity of the target species, focusing primarily on SNP sites and overall sequences. Additionally, phylogenetic tree construction was employed to validate the capacity and accuracy of barcoding technology for species identification (Lam et al., 2020; Blois et al., 2022; Gogoi et al., 2020). The assessment of barcode identification skills frequently relied on GEDI both across and within species, with larger disparities indicating more accurate species identification (Jiang et al., 2022; Li et al., 2021).

Building upon aforementioned concepts, we conducted a retrieval and screening process involving the reference (Ref) strain and six major variants of SARS-CoV-2 (Alpha B.1.1.7, Belta B.1.351, Delta B.1.617.2, Gamma P.1, Lambda C.37, Omicron B.1.1.529) (Singhal, 2022). Additionally, we included SARSr-CoV-2 lineages collected from bats and pangolins (Hu et al., 2021; Zhou et al., 2021), along with the complete genome sequences of the remaining six HCoVs (HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1, SARS-CoV and MERS-CoV) as training sets. These sequences have been published in public research and are archived at the NCBI (Schoch et al., 2020) and GISAID databases (Shu and McCauley, 2017) (Table 1). Accession and version numbers

Table 1
Basic information of training sequence data.

Strains and SDs	Accession and version numbers	Training sets	Number of sequences	Average length
Ref*,#	NC_045512.2	Set1	1	29,903
Alpha B.1.1.7*,#	Details in Supplementary Table S1	Set1	844	29,843
Beta B.1.351*,#	Details in Supplementary Table S1	Set1	8	29,559
Delta B.1.617.2*,#	Details in Supplementary Table S1	Set1	52	29,727
Gamma P.1*,#	Details in Supplementary Table S1	Set1	2	29,806
Lambda C.37*,#	Details in Supplementary Table S1	Set1	3	29,784
Omicron B.1.1.529*,#	Details in Supplementary Table S1	Set1	26	29,303
SARSr-CoV-2 lineages*	Details in Supplementary Table S1	Set2	17	29,141
SARS-CoV-2 (SDI)*	Details in Supplementary Table S1	Set3	936	29,818
HCoV-NL63*	Details in Supplementary Table S1	Set3	54	27,540
HCoV-229E*	Details in Supplementary Table S1	Set3	35	27,366
HCoV-HKU1*	Details in Supplementary Table S1	Set3	41	29,849
HCoV-OC43*	Details in Supplementary Table S1	Set3	74	30,663
MERS-CoV*	Details in Supplementary Table S1	Set3	421	30,081
SARS-CoV*	Details in Supplementary Table S1	Set3	155	29,715
SDII	Details in Supplementary Table S1	Set3	1733	29,782

"#" indicates that the strains included in SDI; "*" indicates that the strains included in SDII.

were meticulously recorded for reference (Supplementary Table S1). The commonly-used SARS-CoV-2 Ref strain, initially submitted in January 2020 (Wu et al., 2020), along with its variants, was integrated into an independent sequence database I (SDI) for individual testing. Furthermore, all sequences related to HCoVs and SARSr-CoV-2 lineages were consolidated into a unified SD (SDII) for comprehensive testing. The selected barcode segments demonstrated their resilience to SARS-CoV-2 mutations, ensuring rapid identification within the strains in SDII. The downloaded sequence files underwent batch processing in a standardized manner, being spliced together into SDI (comprising the Ref strain and six major variants of SARS-CoV-2) and SDII (comprising all training sets) (Supplementary File 1−4), simulating SARS-CoV-2 intraspecific and interspecific levels respectively.

Totally, 1733 sequences were retrieved from NCBI and GISAID. SDI encompassed 936 sequences (Ref:1; Alpha: 844; Beta: 8; Delta: 52; Gamma: 2; Lambda: 3; Omicron: 26); SDII comprised 17 SARSr-CoV-2, 54 HCoV-NL63, 35 HCoV-229E, 41 HCoV-HKU1, 74 HCoV-OC43, 421 MERS-CoV, 155 SARS-CoV and 936 SARS-CoV-2 sequences, respectively. The average sequence length ranged from 27,366 bp (HCoV-229E and HCoV-NL63) to 30,663 bp (HCoV-OC43), with the length difference between SARS-CoV-2 sequences (in SDI) with the higher sequencing quality being below 600 bp.

To ensure the accuracy of SDs and mitigate potential errors stemming from degenerate bases (formation in nucleotide sequences: RYMKSWHBVDN) and sequencing errors in partial sequences (Grantham et al., 1980; Linhart and Shamir, 2005), Python codes (Supplementary File 3−5) were employed for their removal prior to aligning SDs. This procedural step not only heightened the accuracy of SDs after alignment but also addressed compatibility issues with software. SDs were stored in the FASTA format. The workflow is illustrated in Fig. 1.

2.2. Bioinformatic analysis of SDs

SDs containing extensive sequences with an average length outstripping 25,000 bp underwent alignment using the multiple alignment program (MAFFT) v7.0 on the Linux operating system (Rozewicki et al., 2019). To ensure high alignment accuracy, the optional parameters were configured as follows: “GAPS: Deletion”, “Scoring matrix for nucleotide sequence: 200PAM/k = 2”, “Gap opening penalty: 1.53” and “Offset value: 0.0”. Subsequent to alignment, SDs were dealt with by the molecular evolutionary genetics analysis (MEGA) v11 (Tamura et al., 2021), with global parameters set as: “Data type: Nucleotide sequences” and “Genetic code: Standard”. In the “Analysis” interface, SNP of SDs were detected and described using the functions of “Nucleotide Composition”, “Nucleotide Pair Frequencies (directional-16 pairs)” and “CpG Finding”. The Ref strain and six variants of SARS-CoV-2 (SDI) were assorted into seven taxa to generate the interspecific and intraspecific distance matrices using the “group mean distance Kimura 2-parameter (K2P)” model (Ghoyounchi et al., 2017). Similarly, HCoVs and SARSr-CoV-2 strains in SDII were divided into eight separate taxa to cultivate commensurable GEDI matrices. The GEDI matrices were visualized as heatmaps with R language. The DnaSP6 software provided the “Conserved Regions” function (the dynamic parameter defined: “given the observed S″) to search conserved regions in coding sequences (CDSs) (Rozas et al., 2017). However, handling the data output format of complicated nucleotide sequences within SDs in DnaSP6 proved unsatisfactory, impeding the enumeration of conserved-regions-related information. To overcome this, custom format optimization programs (Supplementary File 6−7) were developed to automatically retrieve the number, length, *P* value and distribution information of conserved regions in SARS-CoV-2 CDSs. Significantly, the prerequisite for the analysis of RNA sequences in DnaSP6 was the wholesale removal of all degenerate bases.

2.3. Construction of phylogenetic trees

The presence of latent interspecific gene flows among strains existing in sizeable SDs (storing over 1000 sequences) disrupted the assumption of similar nucleotide substitution rates between distant homologous variants (Sylla et al., 2009). To address this, we optimized the

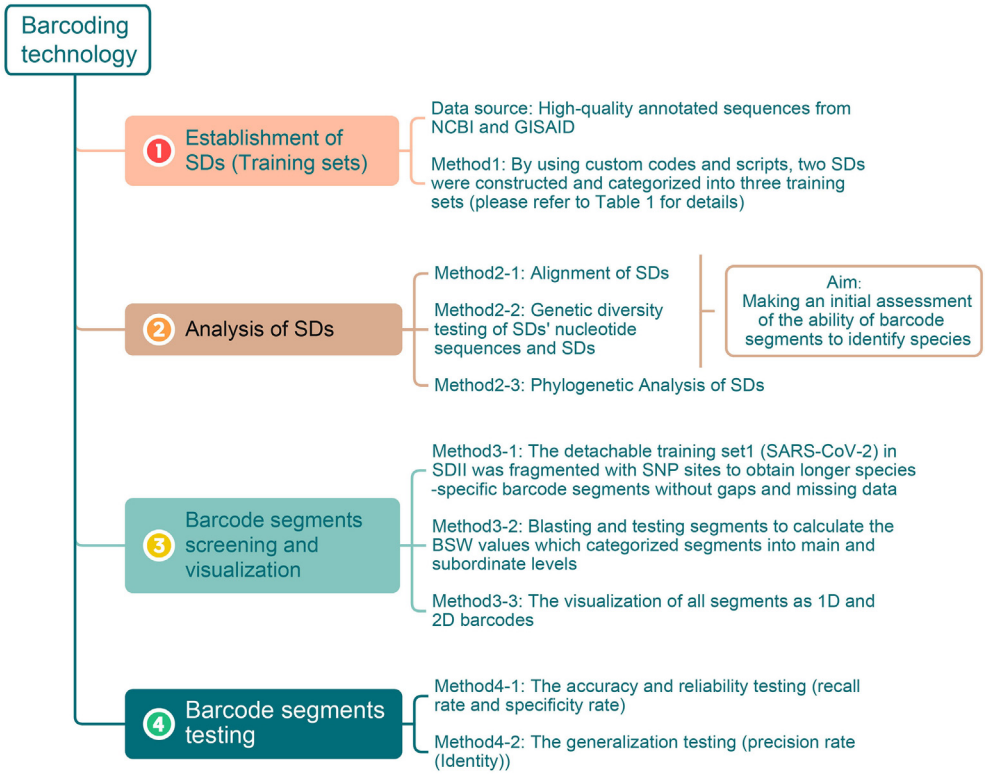

Fig. 1. The flowchart of barcoding technology.

neighbor-joining (NJ) method using MEGA software to construct phylogenetic trees (Tamura et al., 2021). The substitution model utilized the K2P algorithm, which accounted for both base transitional and transversional substitutions and provided a reliable estimation of sequence diversity. The resulting phylogenetic trees saved in the NEWICK format, with bootstrap values exceeding 70% to indicate statistically supported branches. The interactive tree of life (iTOL) (https://itol.embl.de/) online platform (Letunic and Bork, 2021) was used to visualize the phylogenetic trees.

2.4. Barcode segments screening and visualization

By employing the "Define Sequence Sets" and "Polymorphic Sites" functions in DnaSP6 software, sequences related to SARS-CoV-2 in SDII were consolidated into a detachable dataset, enabling the retrieval of SNP sites. Subsequently, this dataset was fragmented into preliminary segments by utilizing two adjacent SNP sites (SNP sites shared conceptual and quantitative similarities with variable sites, Jiang et al., 2022). Preliminary segments containing gaps, missing data, or those shorter than 40 bp were excluded from subsequent analysis (Li et al., 2021). Then, these high-integrity (high-quality) segments were subjected to BLAST using standard nucleotide databases available on NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) (nucleotide collection consisting of GenBank + EMBL + DDBJ + PDB + RefSeq, number of sequences: 89,603,884, update date: 2023/01/12) (Schoch et al., 2020). The alignment results comprised only extremely comparable sequences, with a maximum of 5000 aligned sequences displayed. For accurate and reliable identification of SARS-CoV-2, it was crucial to confirm that all 5000 total BLAST values were consistent with the highest total score. Owing to variations in segment lengths and gaps in the SDII after alignment (Rozas et al., 2017), the "Conserved DNA regions" function in DnaSP6 was used to assess the conservation of barcode segments and verify their accuracy.

Based on the highest barcode segment weight (BSW) scores, we evaluated the optimal species-specific segments from SARS-CoV-2 CDSs, which were then visually mapped into one and two dimensional (1D and 2D) combined barcodes. The 1D and comb-like barcodes were generated using the Python "Barcode" library, distinguishing A, T (U), C, G, AT (AU) base pairs (BPs), and GC BPs distinguished in purple, red, green, blue, long comb, and short comb, respectively. The 2D barcode was generated on https://cli.im/, allowing storage and dynamic conversion of 1D barcodes and their associated information. Electronic mobile devices scanning the 2D barcode could access the barcode information, and the 2D image remained accessible even with partial loss, with an error tolerance of 30%.

2.5. The accuracy and reliability testing of barcode segments

To assess the accuracy and reliability of the barcode segments, we constructed three test sets (Test set1 set2-1 and set2-2) and conducted quantitative validation using recall rate and specificity metrics (Table 2). Variants of concern (VOCs) for SARS-CoV-2 were classified and defined by reputable sources, including the WHO and GISAID (GISAID, 2023; Wang et al., 2023; WHO, 2023). Complete and high-coverage genome sequences of 22 currently circulating VOCs were screened from GISAID, considering their latest collection dates (Supplementary Table S2). These sequences constituted test set1, which was used to evaluate the recall rate of barcode segments towards new variants (recall rate: 1 − the number of error sites/the length of segments/the number of variants) (Supplementary Table S2). SARS-CoV-2 variants and sublineages with more recent collection dates often exhibited higher intraspecific diversity, leading to more convincing results when testing barcode segments. Given the high genome sequence identity shared between SARS-CoV-2, SARSr-CoV-2 and SARS-CoV, we included SARSr-CoV-2 and SARS-CoV sequences (Supplementary Table S1) in test set2 to evaluate the specificity of barcode segments (1 − the number of identical sites/length/the number of strains) in differentiating between these two strains and SARS-CoV-2 (Supplementary Table S2). Test set1 and set2 were merged using the script (Supplementary File 2) and aligned with MAFFT (Katoh et al., 2019). Accession IDs, associated variants, and collection dates for the sequences in these test sets were provided in Supplementary Table S1, and the number of gaps in the alignment regions was provided in Supplementary Table S2.

2.6. The generalization testing of barcode segments

SDs with highly similar internal nucleic acid structures might lead to overfitting of identification accuracy and confine the applicability of segments in different monitoring environments (Shariat et al., 2010). Hence, additional testing was required to assess the generalization ability of barcode segments in identifying other virus strains or variants beyond the SDII, especially those of unknown homologous human viruses. To accomplish this, we utilized the BLAST service available in various databases, including GISAID's EpiCoV (SARS-CoV-2 database, https://gisaid.org/) (Shu and McCauley, 2017), Virus Pathogen Resource (ViPR, reference and representative virus genomes database, https://www.bv-brc.org/app/Homology) (Pickett et al., 2012), National Gene Science Data Center [NGDC, *Coronaviridae* famliy (SARS-CoV-2 strains were independent of this database), *Poxviridae* family and Monkeypox virus genomes database, https://ngdc.cncb.ac.cn/blast/blastn] (CNCB-NGDC

Table 2
Basic information of testing sets.

Test sets	Virus types	Lineage (sublineage) names	Accession and version numbers	Time span (collection date)	Main (optimal) and subordinate (non-optimal) barcode segments
Test set1 (new variants)	New (currently circulating) SARS-CoV-2 variants	Details in Supplementary Table S2	Details in Supplementary Table S2	Details in Supplementary Table S2	Average recall rate: 99.96%
Test set2-1 (SARS-CoV)	SARS-CoV	–	Details in Supplementary Table S2	From 2003-04 to 2008-12	Average specificity: 29.73%
Test set2-2 (SARSr-CoV-2)	SARSr-CoV-2 lineages	Details in Supplementary Table S2	Details in Supplementary Table S2	Details in Supplementary Table S2	Average specificity: 29.03%
Test set3-1 (GISAID's EpiCoV)	All SARS-CoV-2 variants	–	–	From 2023 to 04-01 to 2023-05-02	Average precision rate (Identity): all 100% SARS-CoV-2
Test set3-2 (ViPR)	Reference and representative virus genomes	–	–	From 2019-12 to 2021-11	Average precision rate (Identity): all 100% SARS-CoV-2
Test set3-3 (NGCD)	*Coronaviridae*, *Poxviridae*, Monkeypox virus	–	–	Up to the 7th of June	No results
Test set3-4 (NCBI)	All Influenza viruses	–	–	Up to the 7th of June	No significant similarity found

"–" indicates that the sequences in these test sets either cover the entire database or that some test sets are composed of species from a single lineage.

Members and Partners, 2023), and Influenza Virus BLAST on NCBI (the All Influenza viruses database, https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi) (Bao et al., 2008). The BLAST parameters for these databases were consistently set to "Optimize for highly similar sequences".

2.7. Constructing the online sharing platform

The online database was established by leveraging system network environment architectures (Table 3) and integrating various programming languages (Agosto-Arroyo et al., 2017). It served as a platform for sharing comprehensive information related to SARS-CoV-2 and COVID-19. Moreover, the database offered useful online tools for barcode design and segment alignment. To ensure data preservation and efficient accessibility, SARS-CoV-2 barcodes-related data was stored and transmitted to an elastic compute service, which was a cloud server provided by Alibaba cloud computing company. This database was publicly accessible at http://virusbarcodedatabase.top/.

3. Results

3.1. Nucleotide polymorphism

After filtrating gaps or missing data (e.g., "-" in FASTA files), we found that SDs had general numerical and genetic patterns. In SDI, the AT BP content was 62.1%, while the corresponding GC BP content was 37.9% (Fig. 2; Supplementary Table S3, sheet 1), forming the basis for the creation of CpG islands (Supplementary Table S3, sheet 2) (Trávníček et al., 2019). Moreover, the contents of the four bases exhibited similar quantitative patterns at different coding positions of codons, indicating that the distribution of bases in codon sites did not impact the codon usage bias and the number of synonymous codons in SARS-CoV-2 (Trávníček et al., 2019). The sheet 1 in Supplementary Table S3 contained additional general information on the BP content of SDs.

The ratio of ln[transitional pairs (si)] to ln[transversional pairs (sv)] values (referred to as R values) for HCoVs and SARSr-CoV-2 lineages ranged from 1.05 (MERS-CoV) to 1.48 (HCoV-NL63), revealing that base substitutions in these lineages were predominately in the form of si as opposed to sv (Fig. 3). Consequently, the above strains were less susceptible to nucleotide substitution saturation and exhibited less evolutionary noise, facilitating the construction of phylogenetic trees and the acquisition of accurate genetic information. SARS-CoV-2 strains, particularly the Lambda and Omicron variants, displayed fewer base substitutions (Fig. 3) throughout the entire genome (identical pairs proportion >99.80%, Supplementary Table S3, sheet 3) compared to other strains in SDII. This stability in SARS-CoV-2 made it more suitable for excavating species-specific sites rather than some systematic error sites caused by factors like high-throughput sequence problems (Meacham et al., 2011). Furthermore, a higher proportion of identical pairs (>25,000 identical pairs) provided sufficient sequence space for extracting barcode segments, enabling longer segment lengths and yielding higher BLAST scores for the barcodes (as mentioned in subsequent results).

3.2. Genetic features of SDs

In SDI, the average interspecific GEDI (0.0010) was 1.25 times that of the average intraspecific GEDI (0.0008), and intraspecific differences

Table 3
The configuration of network environment architecture.

Software	Open-source download address
Nginx 1.18	https://nginx.org/en/
MySQL 5.6	https://www.mysql.com/cn/
Pure-Ftpd 1.0.49	https://www.pureftpd.org/project/pure-ftpd/
PHP 5.6	https://www.php.net/
phpMyAdmin 4.4	https://www.phpmyadmin.net/

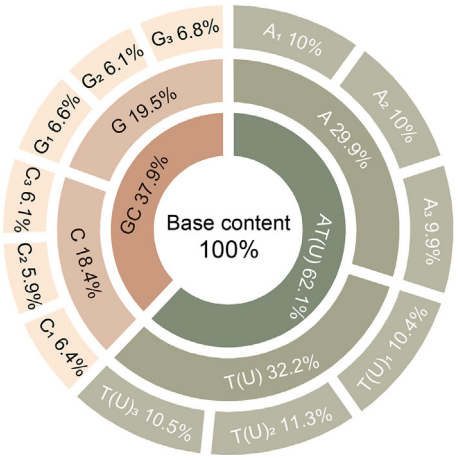

Fig. 2. The sunburst plot of the average nucleotide content in SARS-CoV-2. The inner circle represents the average content of the AT(U) and GC BPs. The middle circle represents the average content of the four bases. The outer circle represents the average content of the four bases at three positions of codons.

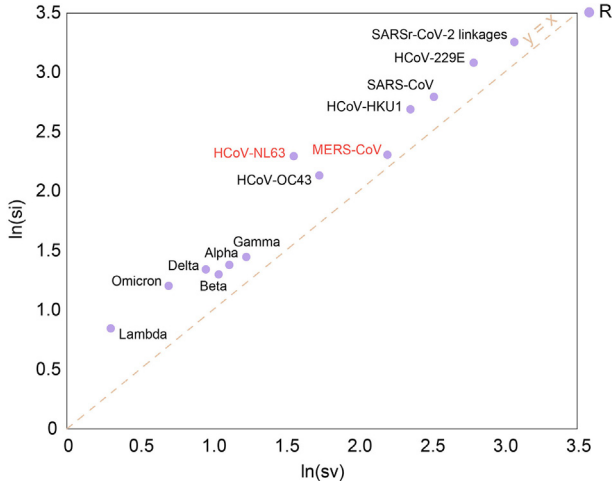

Fig. 3. Nucleotide pair frequencies of HCoVs and SARSr-CoV-2 lineages. The common logarithmic treatment is applied since the si and sv values of viral strains differ significantly. The brown dashed diagonal line (x = y) divides the coordinate system into upper and lower regions. The R-value [the ratio of ln(si) and ln(sv)] anchor point is above the line (R value > 1), suggesting that the species' base substitution form is biased toward si, otherwise the form is biased toward sv (R value < 1). The degree of bias increases as the vertical distance between the anchor point and the diagonal increases. Si, transitional pairs; sv, transversional pairs.

observed in SARS-CoV-2, resulting from sequencing errors and minor variations among variants, could be disregarded (Fig. 4; Supplementary Table S3, sheet 4). Differing in other variants, the Ref, as the initial strain of SARS-CoV-2, preserved mutual genetic characteristics owing to relatively small GEDI differences with the variants. Thus, the Ref was considered as the benchmark sequence for screening barcode segments. By consulting location annotations on NCBI, the relative positions of all CDSs in the SDI were sequentially aligned and recorded to summarize the quantity of conserved regions in each CDS (Fig. 4). Notably, the *ORF1ab* and *S* CDSs accounted for a substantial portion of conserved regions, and their translation products were replicases + non-structural proteins and spike proteins, respectively. Nevertheless, no conserved regions were measured for *M*, *ORF7b*, *ORF8* and *ORF10* with shorter sequence lengths. The distribution of conserved regions revealed a

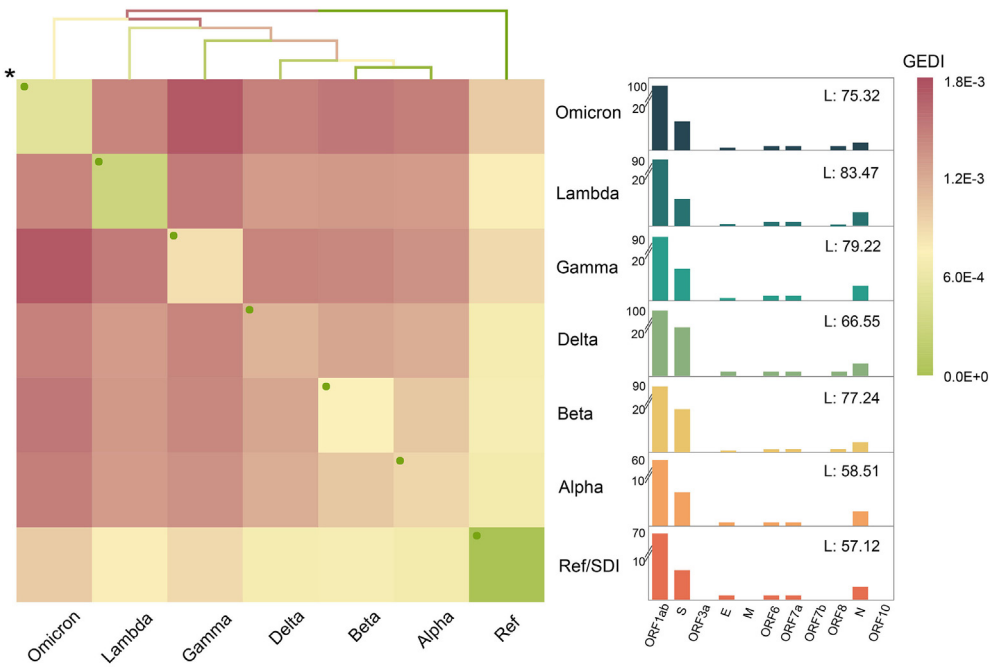

Fig. 4. GEDI matrix and the distribution of conserved regions in CDSs of SARS-CoV-2. The diagonal line of the matrix heat map with "*" and small green spots indicates the average intraspecific GEDI. L indicates the average length of the conserved regions (length unit: bp). Ref/SDI (Ref and SDI) are the group names on the vertical axes on the left and right sides, respectively.

significant decrease in their number (SDI: 76; Alpha: 71) and average length (SDI: 57.12 bp; Alpha: 58.51 bp; Delta: 66.55 bp) within the variants with an increasing sequence volume (SDI: 936; Alpha: 844; Delta: 52) (Supplementary Table S3, sheet 5). Consequently, as the number of sequences increased, the impact of sequencing-generated gaps or missing data on the generation of conserved regions became more pronounced. Furthermore, conserved regions were predominantly distributed in *ORF1ab*, *S*, and *N* regions, with a smaller number of regions distributed in *E*, *ORF6*, *ORF7a*, and *ORF8* regions (conserved region distribution in SDI, Gamma, and Alpha strains was absent in the *ORF8* region). Considering the recognition accuracy of barcode segments for all SARS-CoV-2 variants, it could be inferred that the obtained segments were more likely to be distributed in *ORF1ab*, *S*, *E*, *ORF6*, *ORF7a*, and *N* regions. The length and relative positions of conserved regions in SDI were depicted in Supplementary Fig. S1 (P value $<$ 0.008). The maximum P value for different variants could be found in Supplementary Table S3, sheet 5.

The GEDI matrix demonstrated that the interspecific GEDI of all strains (0.3624) was 12 times greater than the intraspecific GEDI (0.0296) (Fig. 5; Supplementary Table S3, sheet 4). This significant difference in GEDI between both allowed for the identification of each strain through independent genetic markers. Even with respect to SARS-CoV-2 and SARSr-CoV-2 lineages, which exhibited the minimal interspecific GEDI (0.0923), their interspecific GEDI was still higher than the intraspecific GEDI of SARS-CoV-2 (0.0008). In gene flow tests, the haplotype diversity (Hd) of all strains exceeded 0.9882 (except SARS-CoV: 0.9571) (Fig. 5; Supplementary Table S3, sheet 6). SDII's overall nucleotide diversity (π) was 0.2457, but all strains' π values were below 0.1070. This illustrated that the variation direction among strains was inconsistent, and the internal differentiation regions and fundamental variable sites tended to be concentrated. Genetic differentiation between strains were evident without apparent gene flows, particularly in the case of SARS-CoV-2 (self-contained unit in SDII, π: 0.0008).

3.3. Phylogenetic trees of SDI

The independent phylogenetic tree of SARS-CoV-2 (SDI) amplified the minor genetic differences between variants (Fig. 6). With the exception of the two external nodes with longer branch lengths (>0.0014), the branches representing the Ref, Alpha and Beta variants were relatively shortish (<0.0011) and distantly related to Gamma, Delta, Lambda and Omicron variants. Notably, the Delta and Omicron variants exhibited the most divergent branch lengths, mostly exceeding 0.0009 or even 0.0011. Moreover, the relative length of the sequences correlated with the Hd and π values of variants, suggesting the presence of specific recognition sites or segments with significant distinguishing features in SARS-CoV-2 sequences. The evolutionary process from the Alpha to the Omicron variant involved a gradual reduction in π and Hd (Supplementary Table S3, sheet 6), resulting in genetic variability of SARS-CoV-2 and a progressive stabilization of its internal genetic features, which assured the identification stability and "shelf life" of the barcodes.

3.4. Phylogenetic trees of SDII

The developmental tree of SDII exhibited significant evolutionary distance differences, highlighting the divergence among the strains (Fig. 7). The HCoV-229E and HCoV-NL63 strains, which appeared earliest, displayed a branch length interval of 0.24–0.48. The considerable genetic diversity within these strains suggested their potential role as a gene pool sharing certain genetic characteristics with other HCoVs. Assuming the effective application of barcoding technology for SARS-CoV-2, the analogous technical framework could be extended to other human viruses with rather distant phylogenetic relationships. Likewise, MERS-CoV (branch length interval: 0.20−0.36), SARS-CoV (branch length interval: 0.20−0.36) and SARSr-CoV-2 lineages (branch length interval: 0.20−0.32) could also serve as natural gene pools for SARS-CoV-2 by virtue of their rich genetic diversity (no further elaboration in this dissertation) (Zhou et al., 2021). Furthermore, most MERS-CoV

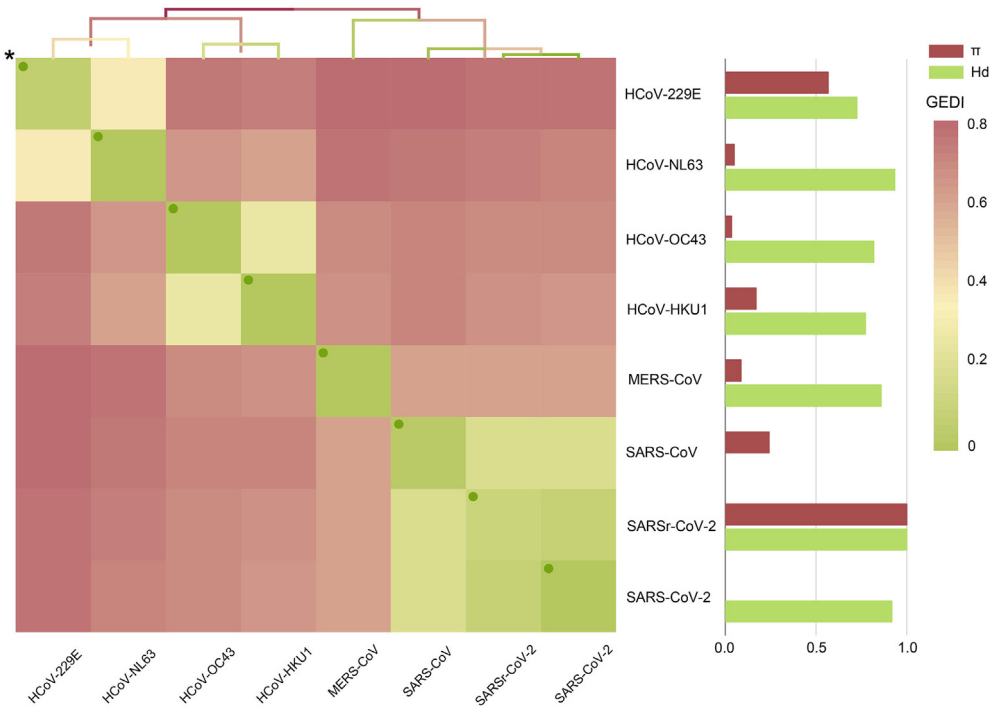

Fig. 5. GEDI matrix and gene flow of SDII. The diagonal line of the matrix heat map with "*" and small green spots indicates the average intraspecific GEDI. π and Hd are normalized values (The normalized interval is "[0, 1]"). Hd, Haplotype (gene) diversity; π, Nucleotide diversity.

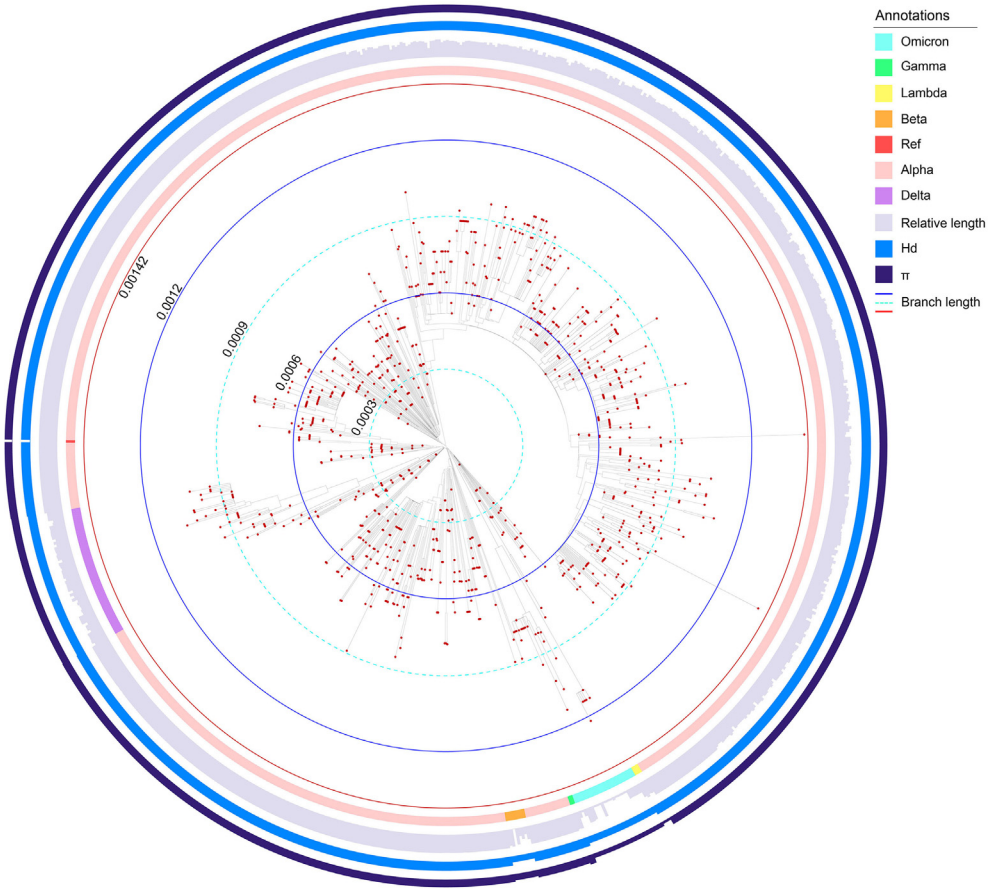

Fig. 6. Phylogenetic tree of SDI (SARS-CoV-2). Considering the order of magnitude of the complete genome length of variants (29,086 bp-29903 bp), a normalized relative sequence length is used to replace the original length value, highlighting the length disparities between various variants. Hd, Haplotype (gene) diversity; π, Nucleotide diversity.

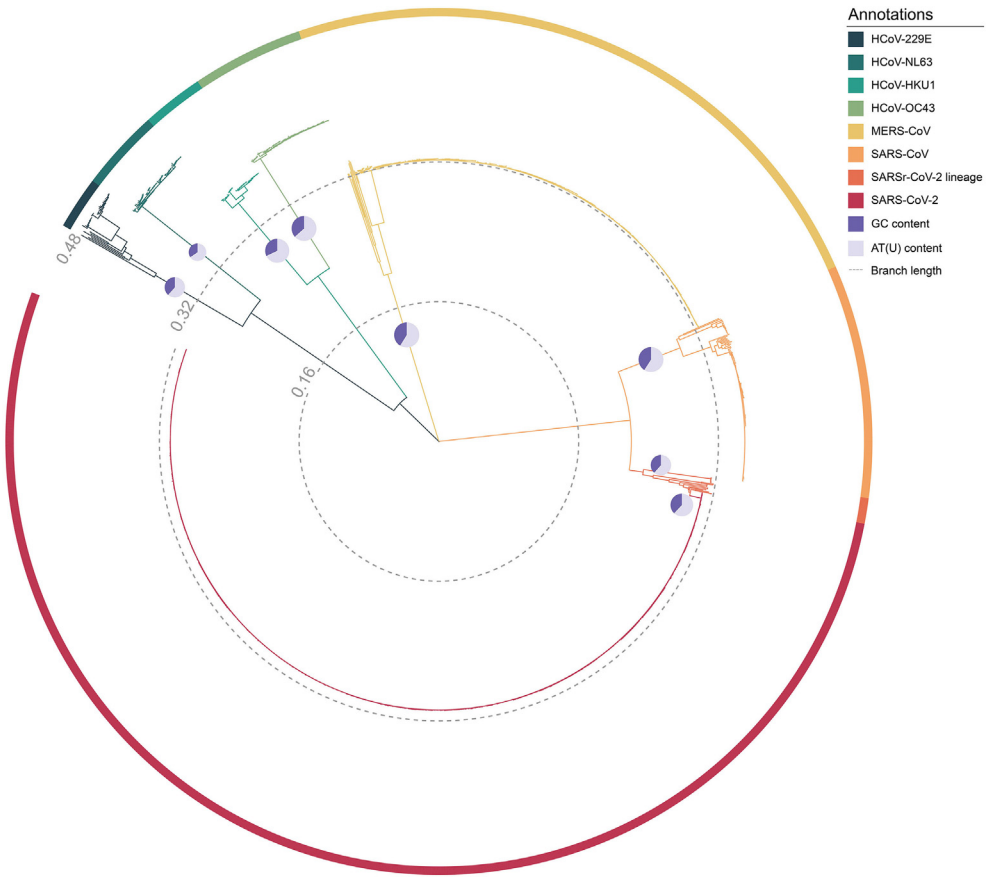

Fig. 7. Phylogenetic tree of SDII (HCoVs and SARrS-CoV-2 lineages). The branch length in the phylogenetic tree is equivalent to the evolutionary distances of species. The radius of the circle representing the base content is positively correlated with the sequence length of the viral strain. The greater the difference in area between the 2 semicircles of the pie chart on each main branch, the more biased the AT or GC BP content is, and the more unstable the genetic structure of the species.

and SARS-CoV strains exhibited a smooth distribution trend of evolutionary distances (no significant variation in branch length within variants, $P < 0.05$), and SARS-CoV-2 displayed comparable levels of branch lengths to these strains (Fig. 7). Hence, considering the genetic resemblance between SARS-CoV and SARS-CoV-2, we speculated that SARS-CoV-2 has undergone mutations. The phylogenetic tree trunk (colored in orange) revealed a close evolutionary relationship among SARS-CoV, SARSr-CoV-2, and SARS-CoV-2 (Guruprasad, 2021), and the above three were the primary targets for barcode identification. HCoV-HKU1 was found to be a closely related strain to HCoV-OC43, exhibiting an average branch length difference of less than 0.04 between them.

The mean GC BP content across strains ranged from 32.0% (HCoV-HKU1) to 41.2% (MERS-CoV), with substantial heterogeneity in content levels (Fig. 7). One-Way ANOVA analysis ($P < 0.05$) revealed no significant correlation between GC BP content and sequence length for all strains. Therefore, viewing GC BP content alone as a genetic differentiation identification indicator was susceptible to sequencing quality and awkward to differentiate on a segment-by-segment basis.

3.5. Visualization of barcode segments

Using the DnaSP6 software, a total of 2118 SNP sites related to SARS-CoV-2 were identified in the SDII dataset (Supplementary Table S3, sheet 7). Poor-quality segments were removed based on detailed screening criteria mentioned in Section 2.4. Eventually, 75 barcodes containing SNP sites (distribution in CDSs: 64 in *ORF1ab*; 6 in *S*; 3 in *N*; 1 in *E*; 1 in *ORF7a*, Fig. 8) were screened by BLAST (all BSW scores $\geq$4.05, all *P* values $\leq$ 0.008, Supplementary Table S4). The distribution of BSW

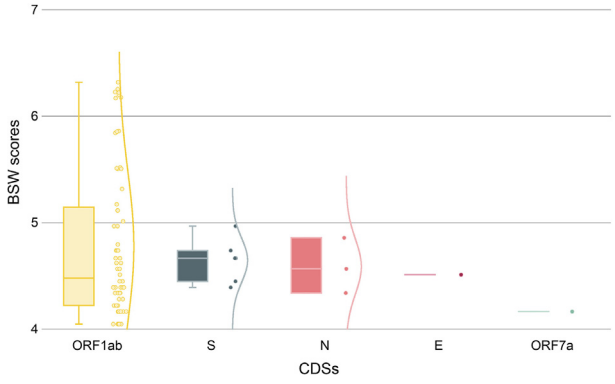

Fig. 8. The box plots and scatter plots of BSW values for barcode segments in CDSs. The curve on the right side of the scatter plot represents the fitted normal distribution curve of the BSW values for segments. The central line inside the box represents the median BSW value of the barcode segments within CDSs.

scores in the *S* and *N* CDSs followed an approximately normal distribution, whereas *ORF1ab* exhibited a noticeable inclination towards higher BSW scores (Fig. 8). On the other hand, the median BSW scores were highest in the *S* and *N* CDSs (4.67; 4.56), followed by *E* with a slightly lower median score of 4.51. The segments in the *ORF1ab* and *ORF7a* CDSs displayed lower median BSW scores of 4.48 and 4.16, respectively. Hence, when disregarding the influence of small sample sizes, it was observed that certain segments derived from *ORF1ab*, which had higher BSW scores, exhibited improved identification capabilities, making them more suitable for identification tasks in complex scenarios. Segments located within the *S* CDS displayed superior stability in identification,

making them more suitable for large-scale batch identification. Besides, length range of segments was from 44 to 113 bp, with an average length of 57.07 bp, and the accuracy of identification precision of barcodes ameliorated with rising barcode segment lengths and total BLAST scores (Supplementary Fig. S2).

Eventually, five main (optimal) species-specific segments were designed as combinatorial barcodes (Fig. 9A and B) with the highest BSW scores, including: *ORF1ab* (113 bp, 6.32 scores), *S* (60 bp, 4.97 scores), *E* (52 bp, 4.51 scores), *ORF7a* (46 bp, 4.16 scores), *N* (58 bp, 4.86 scores) (Fig. 9B). The scanning test demonstrated that the visual 2D code was user-friendly, and that it presented intuitively and understandably the sequence's length and bases composition in text format. The other subordinate (non-optimal) barcodes were also clearly visualized in combined barcodes (Supplementary Fig. S3A and B) for alternative choice.

3.6. Barcode segments testing

The results from test sets demonstrated that all main and subordinate barcode segments exhibited a considerable level of accuracy and a strong generalizability in identifying SARS-CoV-2 (Table 2). In test set1, the average recall rate of main and subordinate barcode segments reached an impressive 99.96%, indicating their ability to accurately identify new variants. Even when testing sequences with a small number of gaps, all segments maintained high coverage, with a recall rate exceeding 99.90% (Supplementary Table S2). Test set2 primarily aimed to assess the specificity of SARS-CoV-2 species-specific barcode segments in identifying genetically SARS-CoV-2 like CoVs (close kinship with SARS-CoV-2, Fig. 7).

In test set2-1, the main and subordinate barcode segments exhibited an average specificity of 29.73% when identifying SARS-CoV. In other words, the segments exhibited a nucleotide difference of up to 20%–30% compared to SARS-CoV. Similarly, in test set2-2, the average specificity of the main and subordinate barcode segments in identifying SARSr-CoV-2 was found to be 29.03%, with a nucleotide difference of approximately 20%–30% between the segments and SARSr-CoV-2 lineages. Therefore, even the closest relatives, SARS-CoV and SARSr-CoV-2, could not be misidentified as SARS-CoV-2.

Test set3 focused on investigating the generalization ability of SARS-CoV-2 species-specific barcode segments in identifying SARS-CoV-2 within a fresh and complicated big-data environments such as SDs. Both test set3-1 and test set3-2 demonstrated that all main and subordinate barcode segments achieved a perfect identification precision rate (Identity) of 100% for SARS-CoV-2. In test set3-1, the time span of collection dates for the strains in the BLAST results ranged from April 1, 2023, to May 2, 2023. In test set3-2, the time span covered from December 2019 to November 2021. These time spans encompassed the most severe phases of the epidemic and the majority of currently circulating variants, providing temporal evidence of the stability of barcode segments and barcoding technology (Singhal, 2022; Wang et al., 2023). Furthermore, test set3-3 and test set3-4 showed that all main and subordinate barcode segments yielded "No significant results found" in the high similarity identification against Monkeypox, Influenza, *Poxviridae* family viruses, and other viruses within *Coronaviridae*. This suggested that the species-specific barcode segments of SARS-CoV-2 remain unaffected by the presence of these viruses, ensuring accurate identification even during the prevalence of influenza and monkeypox diseases in 2023 (Daniels and McCauley, 2023; Peng et al., 2023).

The specificity of identifying strains closely related to SARS-CoV-2 using a few barcode segments was found to be low, and this was influenced by the positioning of the segments in various CDSs (Table 4). In light of this, we recommended the combined use of multiple barcode segments, which would significantly enhance the robustness of identification. Comparing the identification levels of the barcode segments in test set2-1 and test set2-2, we evidently found that segments in the *E* and *N* CDSs exhibited lower average specificity rates (5.77%, 14.73%), while segments in the *S* and *ORF7a* CDSs showed significantly higher rates compared to other CDSs (48.28%, 45.66%), including *ORF1ab* (28.41%). The subordinate barcode segments also followed a similar trend in specificity distribution (Supplementary Table S2). Therefore, drawing inspiration from Li et al.'s approach to enhance species-specific marker identification (Li et al., 2021), it was recommended to use multiple

Table 4
The identification specificity of main and subordinate barcode segments in Test set2.

CDSs	Specificity in test set2-1 (SARS-CoV)	Specificity in test set2-2 (SARSr-CoV-2)	Average specificity in test sets
ORF1ab	29.04%	27.77%	28.41%
S	45.20%	51.35%	48.28%
E	7.69%	3.85%	5.77%
ORF7a	47.83%	43.48%	45.66%
N	14.76%	14.70%	14.73%

Fig. 9. The major visual dynamic 2D barcode for barcode segments. **A** The 2D barcode. People obtain 1D barcodes and basic information about barcode segments by scanning. **B** The basic information screenshot of barcode segments with 2D barcode scanning. The barcode segments use the standard representation of DNA sequences, which is denoted as ACGT.

barcode segments simultaneously in special circumstances to significantly improve identification tolerance. For instance, performing BLAST using both the barcode segments in the *E* and *ORF7a* CDSs was equivalent to conducting two identifications on the target species, which could be particularly effective in complex microbial ecological environments, leading to remarkable results.

3.7. Online database

The online platform (http://virusbarcodedatabase.top/) housed essential information and data pertaining to SDs and species-specific barcode segments discussed in this article. To enhance the dissemination of COVID-19 informatics, the database offered genome functional annotation and lineage information for SARS-CoV-2, along with real-time news updates and literature resources relevant to COVID-19. The platform's BLAST online tool utilized a barcode segment alignment database comprising both main and subordinate barcode segments. Users had the capability to directly import sequencing data and adjust the "Percent Identity" parameter, allowing them to obtain results with varying levels of accuracy. Furthermore, the platform featured tools for creating and visualizing barcode segments. Incidentally, in accordance with the user's sequencing data volume or testing requirements, we recommended an alternative BLAST alignment method. This entailed integrating the nucleotide sequences from barcode segments obtained through visual barcode scanning (Fig. 9A and B; Supplementary Fig. S3A and B) (Cotuțiu et al., 2022) and sample sequences sets into the same SD for sequence alignment using offline software such as MEGA. This approach facilitated an intuitional and rapid assessment of the identity and differences of nucleotide sites within one SD (Cotuțiu et al., 2022).

4. Discussion

Over the past two decades since the introduction of the barcoding technology concept, barcode segments have demonstrated outstanding performance in categorizing and identifying species across animals, plants and microorganisms. Essentially, serving as a molecular genetic marker comparable in function to microsatellites, barcoding technology has proven valuable for the swift identification of viruses and the assessment of mutation degrees. Research focused on the identification the *Theaceae* and *Orchidaceae* plants has validated that adopting bioinformatics approaches to screen species-specific barcode segments proved efficacious in identifying these species accompanying the analysis of GEDI and phylogenetic trees. Additionally, Substantial studies have confirmed that the usage of molecular genetic markers (e.g., barcode segments) could rapidly identify viruses at both interspecific and intraspecific levels.

In this study, we acquired 75 barcode segments distributed across *ORF1ab*, *S*, *E*, *ORF7a* and *N* CDSs through big-data and whole-genome filtering. These segments have consistently demonstrated effective identification of SARS-CoV-2 from other HCoVs and SARSr-CoV-2 lineages. Importantly, they remained unaffected by mutations within SARS-CoV-2 itself, as confirmed by GEDI and phylogenetic analysis. The visually designed SARS-CoV-2 species-specific combinatorial barcodes exhibited several favorable features in quick initiation and convenient access to the related information, which favored research personnel in uncovering novel genetic evolution patterns within and between species. Furthermore, these barcodes offered valuable insights for rapid identification of SARS-CoV-2 variants (including recombinants), SARSr-CoV-2 lineages, HCoVs, and even all viruses. Additionally, our online platform for barcode sharing aimed to widely disseminate barcoding techniques to the public, addressing the imperative for the rapid and efficient identification of SARS-CoV-2 and promoting advancements in molecular biology.

Prior notions and techniques for creating SDs were relatively straightforward and had limited generalizability (Guo et al., 2016; Li et al., 2021; Mahima et al., 2022). For the intention of promoting the quality of SDs, we originally captured the NCBI Ref sequence and other sequences formally published in periodicals. Meanwhile, researchers could extract relative position information of CDSs within these complete genome sequences to serve as a reference for the subsequent gene localization of barcode segments. Apart from genetic relevance, the six significant SARS-CoV-2 variants (Table 1) considered in the article also had a widespread global dissemination and damage (Hu et al., 2021), and the database stored a certain number of public sequences with excellent sequencing quality, so the barcode segments derived from the aforementioned variants became more representative. In this study, we improved the establishment of the training sets (SDs) based on the work of Guan et al. (2020). Our training sets involved HCoVs and strains from SARSr-CoV-2 lineages exhibiting a high degree of genomic sequence similarity with SARS-CoV-2 within the scope of species-specific barcode screening. The creation of training sets for classifying SDs also held innovative value. Developing training sets not only facilitated the modularization of the internal structure of SDs, enabling independent development and modification, but it also ensured the database's scalability to meet evolving needs and increasing data volumes (Raphael et al., 2017). Some databases or software might encounter compatibility issues when analyzing RNA viral nucleotide sequences. Therefore, following the format used by NCBI for sequence publication (AGCT), we made partial adjustments to the format of the output files (e.g., base U→base T). Despite altering of base form, the software analysis of RNA viruses had no effect on the findings of the research on nucleotide composition or genetic evolutionary relationship (Rozas et al., 2017; Tamura et al., 2021), according to comparative experiments and the conclusion of Kirtipal et al. (2020). Thus, researchers only needed to consider the presence of degenerate bases. In addition, we delineated two methods for optimizing the internal structure of the SDs based on several tests. Firstly, to mitigate the impact of genetic noise and alignment errors, degenerate bases were eliminated from the nucleotide sequences. Secondly, genetic studies on a particular species might well be separated into interspecific and intraspecific categories, particularly for viruses with multiple variants (Carvalho et al., 2022).

The tests on nucleotide polymorphism and genetic diversity of SDs served distinct research purposes. The former aimed not only to recapitulate the basic information of SDs but also highlighted the potential degree of variation in SARS-CoV-2 and the applied potency of barcodes from the perspective of evolutionary terms by contrasting base substitutions content between two SDs. Notably, the GC BP content of SARS-CoV-2 was comparable to that of other SARSr-CoV-2 and HCoVs strains (30%–40%) and exhibited similar genetic traits (Zhang et al., 2020). Employing the acquisition concepts of SARS-CoV-2 barcode segments to other SARSr-CoV-2 lineages and HCoVs might thus prove efficacious. The latter served an important function in locating barcode segments utilizing SNP sites. While the barcodes retrieved for the key genes of *Clerodendrum* species in Gogoi's research had certain reliability, ensuring the universality of the barcode segments tested for the SDs with a limited number of species sequences became challenging (Gogoi et al., 2020). Nevertheless, retrieval of conserved regions narrowed the search scope of the momentous barcodes in SARS-CoV-2 CDSs, guaranteeing the applicability of the barcodes to all SARS-CoV-2 variants. Meanwhile, to prevent poor identification accuracy, the interspecific and intraspecific GEDI of SDs were compared to preliminarily forecast identification accuracy and confirm again the minor influence of sequencing or variation factors on barcode screening. Moreover, based on repeated testing of datasets in SDs (Kumar et al., 2018; Hall, 2013), MAFFT was selected for extensive multi-sequence alignment (Katoh et al., 2019), and MEGA was used to create NJ evolution trees. The parameters were adjusted in accordance with the software's user manual.

Studies have indicated that certain phylogenetic traits of HCoVs were closer to those of species within the genus *Betacoronavirus* (SARS-CoV-2 and SARSr-CoV-2 lineages belonged to this genus, while HCoV-NL63 and HCoV-229E belong to the genus *Alpacoronavirus*) (Kirtipal et al., 2020; Guruprasad, 2021). From the perspective of social harmfulness, strains in

training sets also tended to be more closely related to each other, making it more feasible to designate training set2 and set3 as extra-specific taxa of SARS-CoV-2. Moreover, in comparison to *Sulfolobus* spindle-shaped viruses (Zhang et al., 2020), *Betacoronavirus* and HCoVs species had more SNP sites, which was indeed helpful in discovering barcode segments. The phylogenetic status of SARS-CoV-2 involved intraspecific and interspecific levels (SDI and SDII). The SDI evolutionary tree highlighted the internal distinctions of SARS-CoV-2, whereas the SDII evolutionary tree compared SARS-CoV-2 to other HCoVs and SARSr-CoV-2 lineages to explore their genetic evolution rules. The phylogenetic results of SDs demonstrated that the internal nodes in trees and the clustering algorithm for adjacency matrix (Chu et al., 2022) had little effect on the classification of certain species, which intuitively displayed excessively long branches in the tree due to poor sequencing quality of a few sequences. Internal nodes of the tree (Fig. 6) had a similar taxonomic effect to SNP sites, further validating that the barcodes screened based on SNP sites showed a positive capability for identifying viral strains with comparable genetic links (MERS-CoV, SARS-CoV and SARSr-CoV-2 lineages) or remote relationships (HCoV-229E and HCoV-NL63). This article's research on gene flow focused on the in-depth mining and comparison of genetic diversity information of HCoVs; therefore, there was no need for additional in-depth research on the calculation of invasive polymorphism and interspecific diversity in neutral evolution models (Welch et al., 2008) or on evolutionary trends relating to temporal evolution and geographical distribution (Sylla et al., 2009).

We employed multiple algorithms to enhance the reliability and robustness of barcode segments for identification purposes. The identification process was optimized by identifying SNP sites across a wide range of species in the SDII, reducing the margin of error (Blois et al., 2022; Fujito et al., 2021). Through GEDI analysis, we designated SARS-CoV-2 as a separate taxon in SDII to limit the loss of internal conserved nucleotide sites caused by alignment algorithms, hence making it simpler to locate additional barcode segments. This study firstly introduced the innovative concept of BSW values, which were derived from logarithmic weighting of results from NCBI and DnaSP6 (Rozas et al., 2017; Schoch et al., 2020) and provided a valuable method for assessing the precision and reliability of species identification using barcode segments. The distribution of barcode segments (*E*, *N*, *ORF1ab*, *ORF7a*, *S*) fell within the range of conserved regions in SDI (Fig. 4), verifying the reasonableness of the Ref strain as a screening criterion for barcode segments (discussed in Section 3.2), which constituted a crucial theoretical breakthrough in barcode segments design for viruses (Blois et al., 2022). *ORF7a* counteracted the antiviral effect of Serine Incorporator 5, facilitating SARS-CoV-2 entry by blocking virus-cell fusion (Timilsina et al., 2022). Variable sites were predominantly located in the *E* and *S* CDSs, while the *N* CDS exhibited lower variation levels (Zhou et al., 2021). Functionally, the spike protein encoded by the *S* CDS had a high affinity for the host's ACE2 receptor, making individuals more susceptible to symptoms; nucleocapsid and envelope proteins encoded by the *N* and *E* CDSs were crucial for viral assembly, with mutations in these proteins influencing the pathogenicity of SARS-CoV-2 (Rodríguez-Hernández and Sanz-Moreno, 2020). Accordingly, the internal barcode segments of the aforementioned CDSs associated with pathogenicity could monitor, to some extent, the pathogenic alterations of the virus (such as Omicron). Saini and Badua et al. (2021) found that the variation rate within *ORF1ab* was substantially lower than in the rest of CDSs in SARS-CoV-2 (Saini et al., 2021), and nucleotide mutation hotspots in *ORF1ab* were at position 11,083 (5′−3′). Consequently, selecting a segment longer than 100 bp that ended between the 15,570 (±50 bp) and 15,682 (±50 bp) nucleotide sites was unaffected by the invalidity of the barcode caused by the SARS-CoV-2 mutation. Additionally, we provided all main and subordinate barcode segments to facilitate the simultaneous identification of multiple segments within the same CDS, catering to the users' demand for accuracy.

The constructed test sets in this study encompassed a diverse range of strains and species in the database, representing significant genetic characteristics. This allowed for comprehensive evaluation of the accuracy, stability, and generalization ability of barcode segments. In comparison to the barcode identification accuracy of 94% reported by Guan et al. (2020), the barcode identification accuracy achieved in this study was 100%, with a recall rate exceeding 99.96%. Notably, there was also an average specificity difference of over 29% when distinguishing strains with high sequence identity, such as SARS-CoV and SARSr-CoV-2 lineages (Table 2). Test sets 1−3 were specifically designed to complement previous studies on barcode segments (Guan et al., 2020) and assessed the performance of the barcode segments in identifying new variants of SARS-CoV-2 resulting from recombination events, genetically related strains to SARS-CoV-2, and authoritative virus databases. Test set 3 specifically targeted the highly prominent pathogens in current domestic and international epidemics, including monkeypox and influenza (Daniels and McCauley, 2023; Peng et al., 2023), to evaluate the performance of the barcode segments. By combining the validation of accuracy presented in Tables 2 and it became evident that the potential recombination events (Tiecco et al., 2022; Markov et al., 2023) between the aforementioned strains and SARS-CoV-2 had a minimal impact on the identification capability of the barcode segments (Table 4). In Zhou et al.'s study (2021) on the evolutionary origins of SARS-CoV-2 and related viruses, it was observed that SARS-CoV-2 exhibited high sequence homology with related viruses in the E and N CDSs, while the sequence homology was lower in the *ORF1ab*, *S*, and *ORF7a* CDSs. This finding was consistent with the specificity results obtained from our barcode segments (Table 4).

Researchers interested in barcoding technology considered the screening of high-quality segments and efficient public access to visual information crucial. Based on previous design of barcodes (Gogoi et al., 2020; Li et al., 2021), we employed distinct colors and patterns in 1D barcodes to clearly differentiate bases and BPs. 2D barcodes were eye-catching and dynamic, allowing real-time content updates without impacting the scanning experience of the client. In comparison to the barcode segment BLAST function offered by the online database in this study, the sequences stored in the 2D codes could be promptly accessed through scanning, enabling users to effortlessly construct customized SDs for the identification of known species or the detection of new species.

In this study, despite the fact that SARS-CoV-2 could be promptly and reliably identified using barcode segments derived from genetics and scoring tests, unresolved issues remained to be highlighted. Firstly, as the sequences expanded, more mutation sites were detected by the software, resulting in the interception of shorter barcode segments (Blois et al., 2022). Hence, barcode screening and identification accuracy criteria were evolving. In reality, the rising number of identifiable barcode segments of a species had a favorable effect on identification, as it could enhance the screening threshold and eliminate longer false barcodes resulting from external instability factors (e.g., sequencing quality). This article advocated the incorporation of genetic test results into barcode screening criteria. After obtaining convincing test results, the screening requirements were consistently based on the BSW values, ensuring a "100% identity" between the segments and the species. RNA barcoding technology provided the potential for batch identification by formulating particular barcode rules for various species (Gong et al., 2021). Secondly, there is potential for further enhancement in the barcode analysis tools offered by the database, particularly in the area of generating dynamic barcodes with a single click. In our future endeavors, we will dedicate efforts to augment the capabilities of online analysis and continuously explore new application domains for barcoding technology.

5. Conclusions

This endeavor aimed to create species-specific barcode segments for effective identification of SARS-CoV-2, as well as to test the reliability and specificity of these segments. This is the first paper to construct SDs employing big data, extracting 75 main and subordinate barcode segments from *ORF1ab, S*, *E*, *ORF7a*, and *N* CDSs. The fundamental details

and nucleotide sequences of the aforementioned segments are formally documented and made publicly accessible through combined barcode formation. To facilitate public knowledge of SARS-CoV-2 genomic information, genetic diversity analysis data, and barcode images of SDs can be available through online platforms. In comparison to other barcoding approaches, this article extensively tested and optimized the processes of sequence collection, SDs construction, genetic diversity testing, and barcode segment screening. As a result, a more standardized barcode segment design route is established, serving as a technical resource for a large number of medical professionals. In the future, we will promptly gather additional virus data with real-time updates and create a more extensive and comprehensive virus barcode database in an effort to maximize the support of academics for the creation of improved and more efficient virus detection techniques.

Data availability

All data generated or analyzed during this study are included in this published article. The sequences in training and test sets are available from NCBI and GISAID with accession numbers in Supplementary Table S1 and S2.

Ethics statement

This article does not contain any studies with human or animal subjects performed by any of the authors.

Author contributions

Changqiao You: conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, writing-original draft, writing-review & editing. Shuai Jiang: conceptualization, data curation, formal analysis, methodology, software, validation, visualization, writing – original draft, writing-review & editing. Yunyun Ding: conceptualization, formal analysis, methodology, software, visualization, writing – original draft, writing-review & editing. Shunxing Ye: data curation, methodology, visualization. Xiaoxiao Zou: data curation, investigation, methodology. Hongming Zhang: investigation, methodology, supervision, visualization. Zeqi Li: resources, validation, writing-review & editing. Fenglin Chen: resources, validation. Yongliang Li: funding acquisition, project administration, supervision, validation, writing-original draft. Xingyi Ge: funding acquisition, project administration, resources, supervision, validation, writing-review & editing. Xinhong Guo: conceptualization, funding acquisition, project administration, resources, supervision, validation, writing-original draft, writing-review & editing.

Conflict of interest

The authors have declared no competing interests.

Acknowledgements

This research was supported by grants from (1) Key Research & Development Project of Nanhua Biomedical Co., Ltd. (No. H202191490139), (2) National Natural Science Foundation of China (No. 31872866), (3−4) China Postdoctoral Science Foundation (Nos. 2021M701160 and 2022M721101) and Funds of Hunan university (521119400156).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.virs.2024.01.006.

References

Agosto-Arroyo, E., Coshatt, G.M., Winokur, T.S., Harada, S., Park, S.L., 2017. Alchemy: a web 2.0 real-time quality assurance platform for human immunodeficiency virus, hepatitis C virus, and BK virus quantitation assays. J. Pathol. Inf. 10, 8–18.

Amiral, J., Seghatchian, J., 2022. Autoimmune complications of COVID-19 and potential consequences for long-lasting disease syndromes. Transfus. Apher. Sci. 62, 103625.

Badua, C.L.D.C., Baldo, K.A.T., Medina, P.M.B., 2021. Genomic and proteomic mutation landscapes of SARS-CoV-2. J. Med. Virol. 93, 1702–1721.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., Lipman, D., 2008. The influenza virus resource at the national center for biotechnology information. J. Virol. 82, 596–601.

Blois, S., Goetz, B.M., Bull, J.J., Sullivan, C.S., 2022. Interpreting and de-noising genetically engineered barcodes in a DNA virus. PLoS Comput. Biol. 18, e1010131.

Carvalho, L.P.C., Costa, G.D.S., Pereira-Júnior, A.M., de-Paulo, P.F.M., Silva, G.S., Carioca, A.L.P.M., Rodrigues, B.L., Pessoa, F.A.C., Medeiros, J.F., 2022. DNA barcoding of genus *Culicoides* biting midges (Diptera: Ceratopogonidae) in the Brazilian Amazon. Acta Trop. 235–106619.

Chaimayo, C., Kaewnaphan, B., Tanlieng, N., Athipanyasilp, N., Sirijatuphat, R., Chayakulkeeree, M., Angkasekwinai, N., Sutthent, R., Puangpunngam, N., Tharmviboonsri, T., Pongraweewan, O., Chuthapisith, S., Sirivatanauksorn, Y., Kantakamalakul, W., Horthongkham, N., 2020. Rapid SARS-CoV-2 antigen detection assay in comparison with real-time RT-PCR assay for laboratory diagnosis of COVID-19 in Thailand. Virol. J. 17, 177.

Chu, H.M., Liu, J.X., Zhang, K., Zheng, C.H., Wang, J., Kong, X.Z., 2022. A binary biclustering algorithm based on the adjacency difference matrix for gene expression data analysis. BMC Bioinf. 23, 381.

CNCB-NGDC Members and Partners, 2023. Database resources of the national genomics data center, China National Center for Bioinformation in 2023. Nucleic Acids Res. 51, D18–D28.

Cohen-Aharonov, L.A., Rebibo-Sabbah, A., Yaacov, A., Granit, R.Z., Strauss, M., Colodner, R., Cheshin, O., Rosenberg, S., Eavri, R., 2022. High throughput SARS-CoV-2 variant analysis using molecular barcodes coupled with next generation sequencing. PLoS One 17, e0253404.

Cosar, B., Karagulleoglu, Z.Y., Unal, S., Ince, A.T., Uncuoglu, D.B., Tuncer, G., Kilinc, B.R., Ozkan, Y.E., Ozkoc, H.C., Demir, I.N., Eker, A., Karagoz, F., Simsek, S.Y., Yasar, B., Pala, M., Demir, A., Atak, I.N., Mendi, A.H., Bengi, V.U., Cengiz-Seval, G., Gunes-Altuntas, E., Kilic, P., Demir-Dora, D., 2022. SARS-CoV-2 mutations and their viral variants. Cytokine Growth Factor Rev. 63, 10–22.

Cotuțiu, V.D., Ionică, A.M., Lefkaditis, M., Cazan, C.D., Hașaș, A.D., Mihalca, A.D., 2022. Thelazia lacrymalis in horses from Romania: epidemiology, morphology and phylogenetic analysis. Parasites Vectors 15, 425.

Daniels, R.S., McCauley, J.W., 2023. The health of influenza surveillance and pandemic preparedness in the wake of the COVID-19 pandemic. J. Gen. Virol. 104, 001822.

Fujito, S., Akyol, T.Y., Mukae, T., Wako, T., Yamashita, K.I., Tsukazaki, H., Hirakawa, H., Tanaka, K., Mine, Y., Sato, S., Shigyo, M., 2021. Construction of a high-density lineage map and graphical representation of the arrangement of transcriptome-based unigene markers on the chromosomes of onion, Allium cepa L. BMC Genom. 22, 481.

Ghoyounchi, R., Ahmadpour, E., Spotin, A., Mahami-Oskouei, M., Rezamand, A., Aminisani, N., Ghojazadeh, M., Berahmat, R., Mikaeili-Galeh, T., 2017. Microsporidiosis in Iran: a systematic review and meta-analysis. Asian Pac. J. Tropical Med. 10, 341–350.

GISAID, 2023. Variant of Concern Reports. https://gisaid.org/lineage-comparison/. (Accessed 11 June 2023).

Gogoi, B., Wann, S.B., Saikia, S.P., 2020. DNA barcodes for delineating *Clerodendrum* species of North East India. Sci. Rep. 10, 13490.

Gong, L., Zhang, D., Ding, X., Huang, J., Guan, W., Qiu, X., Huang, Z., 2021. DNA barcode reference library construction and genetic diversity and structure analysis of *Amomum villosum* Lour. (Zingiberaceae) populations in Guangdong Province. PeerJ 9, e12325.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., Pavé, A., 1980. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 8, r49–r62.

Guan, Q., Sadykov, M., Mfarrej, S., Hala, S., Naeem, R., Nugmanova, R., Al-Omari, A., Salih, S., Al-Mutair, A., Carr, M.J., Hall, W.W., Arold, S.T., Pain, A., 2020. A genetic barcode of SARS-CoV-2 for monitoring global distribution of different clades during the COVID-19 pandemic. Int. J. Infect. Dis. 100, 216–223.

Guo, Y.Y., Huang, L.Q., Liu, Z.J., Wang, X.Q., 2016. Promise and challenge of DNA barcoding in Venus slipper (*Paphiopedilum*). PLoS One 11, e0146880.

Guruprasad, L., 2021. Human coronavirus spike protein-host receptor recognition. Prog. Biophys. Mol. Biol. 161, 39–53.

Hall, B.G., 2013. Building phylogenetic trees from molecular data with MEGA. Mol. Biol. Evol. 30, 1229–1235.

Hebert, P.D., Cywinska, A., Ball, S.L., deWaard, J.R., 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–321.

Hu, B., Guo, H., Zhou, P., Shi, Z.L., 2021. Characteristics of SARS-CoV-2 and COVID-19. Nat. Rev. Microbiol. 19, 141–154.

Jiang, S., Chen, F., Qin, P., Xie, H., Peng, G., Li, Y., Guo, X., 2022. The specific DNA barcodes based on chloroplast genes for species identification of *Theaceae* plants. Physiol. Mol. Biol. Plants 28, 837–848.

Katoh, K., Rozewicki, J., Yamada, K.D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings Bioinf. 20, 1160–1166.

Kirtipal, N., Bharadwaj, S., Kang, S.G., 2020. From SARS to SARS-CoV-2, insights on structure, pathogenicity and immunity aspects of pandemic human coronaviruses. Infect. Genet. Evol. 85, 104502.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. Mega X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35, 1547–1549.
Lago, S.G., Tomasik, J., van Rees, G.F., Ramsey, J.M., Haenisch, F., Cooper, J.D., Broek, J.A., Suarez-Pinilla, P., Ruland, T., Auyeug, B., Mikova, O., Kabacs, N., Arolt, V., Baron-Cohen, S., Crespo-Facorro, B., Bahn, S., 2020. Exploring the neuropsychiatric spectrum using high-content functional analysis of single-cell signaling networks. Mol. Psychiatr. 25, 2355–2372.
Lam, T.T., Jia, N., Zhang, Y.W., Shum, M.H., Jiang, J.F., Zhu, H.C., Tong, Y.G., Shi, Y.X., Ni, X.B., Liao, Y.S., Li, W.J., Jiang, B.G., Wei, W., Yuan, T.T., Zheng, K., Cui, X.M., Li, J., Pei, G.Q., Qiang, X., Cheung, W.Y., Li, L.F., Sun, F.F., Qin, S., Huang, J.C., Leung, G.M., Holmes, E.C., Hu, Y.L., Guan, Y., Cao, W.C., 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature 583, 282–285.
Langat, S.K., Eyase, F., Bulimo, W., Lutomiah, J., Oyola, S.O., Imbuga, M., Sang, R., 2021. Profiling of RNA viruses in biting midges (*Ceratopogonidae*) and related Diptera from Kenya using metagenomics and metabarcoding analysis. mSphere 6, e0055121.
Letunic, I., Bork, P., 2021. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49, W293–W296.
Li, H., Xiao, W., Tong, T., Li, Y., Zhang, M., Lin, X., Zou, X., Wu, Q., Guo, X., 2021. The specific DNA barcodes based on chloroplast genes for species identification of *Orchidaceae* plants. Sci. Rep. 11, 1424.
Linhart, C., Shamir, R., 2005. The degenerate primer design problem: theory and applications. J. Comput. Biol. 12, 431–456.
Mahima, K., Sunil-Kumar, K.N., Rakhesh, K.V., Rajeswaran, P.S., Sharma, A., Sathishkumar, R., 2022. Advancements and future prospective of DNA barcodes in the herbal drug industry. Front. Pharmacol. 13, 947512.
Markov, P.V., Ghafari, M., Beer, M., Lythgoe, K., Simmonds, P., Stilianakis, N.I., Katzourakis, A., 2023. The evolution of SARS-CoV-2. Nat. Rev. Microbiol. 21, 361–379.
Meacham, F., Boffelli, D., Dhahbi, J., Martin, D.I., Singer, M., Pachter, L., 2011. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinf. 12, 451.
Meng, X., Zou, S., Li, D., He, J., Fang, L., Wang, H., Yan, X., Duan, D., Gao, L., 2022. Nanozyme-strip for rapid and ultrasensitive nucleic acid detection of SARS-CoV-2. Biosens. Bioelectron. 217, 114739.
Minervina, A.A., Pogorelyy, M.V., Kirk, A.M., Crawford, J.C., Allen, E.K., Chou, C.H., Mettelman, R.C., Allison, K.J., Lin, C.Y., Brice, D.C., Zhu, X., Vegesana, K., Wu, G., Trivedi, S., Kottapalli, P., Darnell, D., McNeely, S., Olsen, S.R., Schultz-Cherry, S., McGargill, M.A., Wolf, J., Thomas, P.G., 2022. SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8+ T cells. Nat. Immunol. 23, 781–790.
Nimavat, N., Singh, S., Fichadiya, N., Sharma, P., Patel, N., Kumar, M., Chauhan, G., Pandit, N., 2021. Online medical education in India - different challenges and probable solutions in the age of COVID-19. Adv. Med. Educ. Pract. 12, 237–243.
Peng, C., He, M., Cutrona, S.L., Kiefe, C.I., Liu, F., Wang, Z., 2020. Theme trends and knowledge structure on mobile health apps: bibliometric analysis. JMIR Mhealth Uhealth 8, e18212.
Peng, Q., Xie, Y., Kuai, L., Wang, H., Qi, J., Gao, G.F., Shi, Y., 2023. Structure of monkeypox virus DNA polymerase holoenzyme. Science 379, 100–105.
Pickett, B.E., Sadat, E.L., Zhang, Y., Noronha, J.M., Squires, R.B., Hunt, V., Liu, M., Kumar, S., Zaremba, S., Gu, Z., Zhou, L., Larson, C.N., Dietrich, J., Klem, E.B., Scheuermann, R.H., 2012. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res. 40, D593–D598.
Raphael, C.E., Alkhouli, M., Maor, E., Panaich, S.S., Alli, O., Coylewright, M., Reeder, G.S., Sandhu, G., Holmes, D.R., Nishimura, R., Malouf, J., Cabalka, A., Eleid, M.F., Rihal, C.S., 2017. Building blocks of structural intervention: a novel modular paradigm for procedural training. Circ. Cardiovasc. Interv. 10, e005686.
Rodríguez-Hernández, C., Sanz-Moreno, L., 2020. Inmunidad frente a SARS-CoV-2: caminando hacia la vacunación [Immunity against SARS-CoV-2: walking to the vaccination]. Rev. Española Quimioter. 33, 392–398.
Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sánchez-Gracia, A., 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol. Biol. Evol. 34, 3299–3302.
Rozewicki, J., Li, S., Amada, K.M., Standley, D.M., Katoh, K., 2019. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res. 47, W5–W10.
Saini, S.K., Hersby, D.S., Tamhane, T., Povlsen, H.R., Amaya-Hernandez, S.P., Nielsen, M., Gang, A.O., Hadrup, S.R., 2021. SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial CD8+ T cell activation in COVID-19 patients. Sci. Immunol. 6, eabf7550.
Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I., 2020. NCBI Taxonomy: a Comprehensive Update on Curation, Resources and Tools. Database, Oxford, 2020, baaa062.
Selingerova, I., Valik, D., Gescheidtova, L., Sramek, V., Cermakova, Z., Zdrazilova-Dubska, L., 2021. Interpretive discrepancies caused by target values inter-batch variations in chemiluminescence immunoassay for SARS-CoV-2 IgM/IgG by MAGLUMI. J. Med. Virol. 93, 1805–1809.
Shariat, S.F., Lotan, Y., Vickers, A., Karakiewicz, P.I., Schmitz-Dräger, B.J., Goebell, P.J., Malats, N., 2010. Statistical consideration for clinical biomarker research in bladder cancer. Urol. Oncol. 28, 389–400.
Sheth, B.P., Thaker, V.S., 2017. DNA barcoding and traditional taxonomy: an integrated approach for biodiversity conservation. Genome 60, 618–628.
Shu, Y., McCauley, J., 2017. GISAID: global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 22, 30494.
Singhal, T., 2022. The emergence of omicron: challenging times are Here again. Indian J. Pediatr. 89, 490–496.
Swanson, S.J., Conant, L.L., Humphries, C.J., LeDoux, M., Raghavan, M., Mueller, W.M., Allen, L., Gross, W.L., Anderson, C.T., Carlson, C.E., Busch, R.M., Lowe, M., Tivarus, M.E., Drane, D.L., Loring, D.W., Jacobs, M., Morgan, V.L., Szaflarski, J., Bonilha, L., Bookheimer, S., Grabowski, T., Phatak, V., Vannest, J., 2020. Changes in description naming for common and proper nouns after left anterior temporal lobectomy. Epilepsy Behav. 106, 106912.
Sylla, M., Bosio, C., Urdaneta-Marquez, L., Ndiaye, M., 2009. Gene flow, subspecies composition, and dengue virus-2 susceptibility among Aedes aegypti collections in Senegal. PLoS Neglected Trop. Dis. 3, e408.
Tamura, K., Stecher, G., Kumar, S., 2021. MEGA11: molecular evolutionary genetics analysis version 11. Mol. Biol. Evol. 38, 3022–3027.
Tan, M., Xia, J., Luo, H., Meng, G., Zhu, Z., 2023. Applying the digital data and the bioinformatics tools in SARS-CoV-2 research. Comput. Struct. Biotechnol. J. 21, 4697–4705.
Tiecco, G., Storti, S., Arsuffi, S., Degli-Antoni, M., Focà, E., Castelli, F., Quiros-Roldan, E., 2022. Omicron BA.2 lineage, the "stealth" variant: is it truly a silent epidemic? a literature review. Int. J. Mol. Sci. 23, 7315.
Timilsina, U., Umthong, S., Ivey, E.B., Waxman, B., Stavrou, S., 2022. SARS-CoV-2 ORF7a potently inhibits the antiviral effect of the host factor SERINC5. Nat. Commun. 13, 2935.
Trávníček, P., Čertner, M., Ponert, J., Chumová, Z., Jersáková, J., Suda, J., 2019. Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. New Phytol. 224, 1642–1656.
Ullah, A., Mabood, N., Maqbool, M., Khan, L., Ullah, M., 2021. Cytidine deamination-induced perpetual immunity to SAR-CoV-2 infection is a potential new therapeutic target. Int. J. Med. Sci. 18, 3788–3793.
Wang, L., Møhlenberg, M., Wang, P., Zhou, H., 2023. Immune evasion of neutralizing antibodies by SARS-CoV-2 Omicron. Cytokine Growth Factor Rev. 70, 13–25.
Welch, J.J., Eyre-Walker, A., Waxman, D., 2008. Divergence and polymorphism under the nearly neutral theory of molecular evolution. J. Mol. Evol. 67, 418–426.
Westhaus, A., Cabanes-Creus, M., Rybicki, A., Baltazar, G., Navarro, R.G., Zhu, E., Drouyer, M., Knight, M., Albu, R.F., Ng, B.H., Kalajdzic, P., Kwiatek, M., Hsu, K., Santilli, G., Gold, W., Kramer, B., Gonzalez-Cordero, A., Thrasher, A.J., Alexander, I.E., Lisowski, L., 2020. High-throughput in vitro, ex vivo, and in vivo screen of adeno-associated virus vectors based on physical and functional transduction. Hum. Gene Ther. 31, 575–589.
WHO, 2023. Tracking SARS-CoV-2 Variants. https://www.who.int/activities/tracking-SARS-CoV-2-variants. (Accessed 11 June 2023).
Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C., Zhang, Y.Z., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269.
Zhang, J., Zheng, X., Wang, H., Jiang, H., Dong, H., Huang, L., 2020. Novel Sulfolobus fuselloviruses with extensive genomic variations. J. Virol. 94, e01624 e01619.
Zhou, H., Ji, J., Chen, X., Bi, Y., Li, J., Wang, Q., Hu, T., Song, H., Zhao, R., Chen, Y., Cui, M., Zhang, Y., Hughes, A.C., Holmes, E.C., Shi, W., 2021. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell 184, 4380–4391.