Review

Not peer-reviewed version

Navigating the Landscape: A Comprehensive Review of Current Virus Databases

Muriel Ritsch *, Noriko A. Cassman *, Shahram Saghaei, Manja Marz *

Posted Date: 12 July 2023

doi: 10.20944/preprints202307.0824.v1

Keywords: databases; viruses; genomees; sequences; metadata; FAIR evaluation

Review

Navigating the Landscape: A Comprehensive Review of Current Virus Databases

Muriel Ritsch [1,2,*], Noriko A. Cassman [1,2,*], Shahram Saghaei [1,2] and Manja Marz [1,2,3,4,*]

[1] RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, 07743 Jena, Germany
[2] European Virus Bioinformatics Center, 07743 Jena, Germany; evbc@uni-jena.de
[3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany
[4] FLI Leibniz Institute for Age Research, Jena, Germany
* Correspondence: manja@uni-jena.de (M.M.), anne.muriel.christin.ritsch@uni-jena.de (M.R.); noriko.cassman@uni-jena.de (N.A.C.)

Abstract: Viruses are abundant and diverse entities that have important roles in public health, ecology, and agriculture. Identification and surveillance of viruses rely on understanding their genome organization, sequences, and replication strategy. Despite technological advancements in sequencing methods, our current understanding of virus diversity remains incomplete, highlighting the need to explore undiscovered viruses. Virus databases play a crucial role in providing access to sequences, annotations and other metadata, and analysis tools for studying viruses. However, there has not been a comprehensive review of virus databases in the last five years. This study aimed to fill this gap by identifying 24 active virus databases and included an extensive evaluation of their content, functionality and compliance with the FAIR principles. In this study, we thoroughly assessed the search capabilities of five database catalogs, which serve as comprehensive repositories housing a diverse array of databases and offering essential metadata. Moreover, we conducted a comprehensive review of different types of errors, encompassing taxonomy, names, missing information, sequences, sequence orientation, and chimeric sequences, with the intention of empowering users to effectively tackle these challenges. We expect this review to aid users in selecting suitable virus databases and other resources, and to help databases in error management and improve their adherence to the FAIR principles. The databases listed here represent current knowledge of viruses and will help aid users find databases of interest based on content, functionality, and scope. Use of virus databases is integral to gain new insights into the biology, evolution, and transmission of viruses, and develop new strategies to manage virus outbreaks and preserve global health.

Keywords: viruses; databases; genomes; sequences; metadata; FAIR evaluation

1. Introduction

Viruses, the most abundant and diverse biological entities in the biosphere, play important roles in public health, ecology, and agriculture [1,2]. Identification, characterization, and surveillance of viruses rely heavily on knowledge of their various characteristics, including genome organization, genomic sequences, and replication strategy [3–5]. Until about the 1990's, virus genomes were mainly derived from isolates, came from medically and agriculturally important hosts and were annotated with morphological and medically-related metadata [6]. With the advancements in and lower costs of next-generation sequencing technologies, -omics projects have gained popularity across various fields, leading to the continuous deposition of metagenome-derived viral sequences into sequencing repositories along with computationally-derived metadata. The analysis of uncultivated virus genomes derived from metagenomic sequences has accelerated the rate of discovery of new viruses [7,8]. However, despite these technological strides, the current understanding of virus diversity remains incomplete [9,10]; for example, regarding viruses with zoonotic potential, one estimate suggests that only 1 % of these virus species have been discovered to date [11]. This discrepancy highlights the

 doi:10.20944/preprints202307.0824.v1

need to further explore and uncover the vast array of undiscovered viruses through their genomic sequences.

Web-facing databases provide structured and indexed storage of information which can be easily accessed through an internet browser. In the context of virus research, virus databases represent central hubs of information connecting virus genomic sequences and associated metadata such as viral and host taxonomy, host range, transmission modes, genomic structures, and gene annotations. Researchers use these databases to gain insights into viral genetic diversity and evolutionary relationships and to study new viruses [12,13]. Virus databases play a pivotal role in virus discovery, surveillance, and monitoring, allowing for the comparison of newly identified viruses with known ones [14–16]. Moreover, they contribute to epidemiological studies by integrating diverse information such as clinical characteristics, virus and host genetic variation, and sampling and geographical location. Overall, virus databases are vital resources for comprehensive analyses, comparative studies, and generating innovative strategies in virus detection, prevention, and control.

The presence of multiple virus databases can be attributed to variations in specialization, data types, and aims. Some databases are created for a specific research purpose or to support a specific virus research area, e.g. virus ecology or epidemiology; alternately, some databases focus on specific viruses or cover a wide range of viruses. Further, databases have different goals, ranging from information dissemination to providing web-based tools for data analysis. Database longevity refers to the ability of a database to remain functional and accessible over a long period of time, even as technology and software evolve [17]. Ensuring longevity for virus sequence databases includes regular maintenance and updates, usage of standardized data formats, creation of backups and archives, implementation of open data policies, collaboration, funding, and trust and usage by the community [18,19]. The current varied landscape of virus databases largely reflects the informational needs and funding of different types of virus research.

Many virus databases feature extra curation of a subset of virus sequences from the long-term sequence repositories (e.g. `NCBI Genbank`) as well as additional computed data sets based on these sequences. As such, virus databases contain vital metadata such as taxonomic details of hosts and viruses, location data, publications, years, processing information, morphology, and more. These metadata play a crucial role in studying virus outbreaks, tracking infection pathways, and analyzing viral distribution patterns. Another important type of data found in virus databases are gene annotations, providing valuable information about genetic elements, protein sequences, gene functions, and structural features. This annotation data allows researchers to analyze the functions of viral genes and proteins, gaining insights into their impact on virus biology. These datasets facilitate comparative analysis of viruses and virus communities. On the other hand, incomplete or inaccurate metadata in sequence entries can adversely impact downstream analysis [20,21]. Virus databases also offer analysis tools and software, such as sequence comparisons, phylogenetic analyses, protein structure prediction, and identification of evolutionary selection pressures. A comprehensive comparison of the content based on the number of sequences and species in each database has not yet been performed.

A challenge encountered by all databases – not limited to virus-specific ones – is the handling of errors. Errors are almost unavoidable and can significantly impact subsequent analyses. Many questions arise when considering error management, such as whether users should be allowed to upload their own data, which can lead to a more complete data set but also carries the potential for more errors. Another important aspect is how quickly and effectively errors can be addressed and supported. One solution to address this issue is to provide a curated subset of data, allowing users to decide which data set to utilize. However, this presents difficulties as users may not be aware of the cascading nature of existing errors. As far as we know, there is currently no comprehensive review systematically describing different types of errors to provide users with an overview of potential pitfalls, so users can better address them.

To the best of our knowledge, the last two comprehensive reviews of virus databases were conducted by Sharma *et al.* in 2015 and Mcleod *et al.* in 2017 [22,23]. Sharma et al. reviewed 60

 doi:10.20944/preprints202307.0824.v1

tools and 50 databases, but 39 of the databases are no longer accessible or have not been updated recently, see Table S1. In the Mcleod *et al.* review, they listed 149 entries, including 48 sequence databases; however, only 11 sequence databases met our criteria of being up-to-date for our study, see Table S2. Comparing the two reviews, we found 8 databases that were considered up-to-date in both. Surprisingly, despite the rapid advancements and numerous changes in the field, there has been a notable absence of comprehensive reviews on virus databases in the past 5 years.

In addition to database review articles, several catalogs of databases have been developed that aggregate databases in a clear and well-structured manner. Five such catalogs of databases are (1) `re3data.org`, (2) `FAIRsharing` [24], (3) `The Database Commons` [25], (4) `ELEXIR bio.tools` [26], and (5) `NAR database list` [27]. These catalogs may serve as valuable resources for users who are uncertain about which database to utilize, by providing them with a convenient starting point. In these database lists, users can find database descriptions and searchable metadata, as well as static properties such as the year of establishment, links, and country information. There is currently a lack of comparisons of catalogs of databases evaluating their use for searching virus databases.

When using a virus database, users have diverse requirements beyond accessing relevant content. The usability of the database is crucial, with easy navigation and efficient information retrieval being key factors. Features like keyword search and phrase suggestions enable efficient entry retrieval. Presenting results in meaningful formats, such as tables, enhances usability, particularly when multiple matching entries are involved. Accessibility is important, providing options like one-click downloads and availability through various channels (website, FTP, or API), considering the need for user login. Establishing links to other databases enables comprehensive and integrated research, while integration with workbenches and availability of associated tools empower users to work with their data and utilize additional resources for analysis. Easy sharing of URLs promotes collaboration and efficient communication. Trust and transparency are fostered by the availability of source code, enabling users to comprehend underlying processes. To the best of our knowledge, there are currently no existing reviews that attempt to describe the functionality of databases. This lack of information hinders users in making informed decisions about which database is the most suitable for their needs.

The FAIR data principles encompass the Findability, Accessibility, Interoperability, and Reusability of research data objects [28]. The principles emphasize making these objects FAIR (whether database entries or databases) by following the criteria within each principle described at the FAIR principles website. The FAIR principles can be applied to evaluate three key entities: data (or any digital object), metadata (information about the digital object), and infrastructure. Adhering to the FAIR principles is meant to enhance machine-driven discovery and utilization of data in parallel to facilitating human collaboration and data reuse in scientific research. Evaluating virus databases based on FAIR principles would reveal their potential to advance scientific knowledge and enable data-driven discoveries. As far as our knowledge extends, this is the initial evaluation of virus databases in terms of FAIRness.

In this study, we provide a comprehensive overview of the most up-to-date virus databases, covering five key aspects: (1) content, (2) functionality, (3) FAIRness, (4) comparison of the catalogs of databases, and (5) an analysis of potential errors within the databases. In terms of content, we compare the key features already present in other reviews, such as name, link, last update, short description, and citations. Additionally, we delve deeper into the content of the databases by evaluating the number of sequences and species available, suggesting use cases for each database. Regarding functionality, we compare features such as general usage, presence of a workbench, tool availability, database complexity, download options and ease of use, and the need for user login. Additionally, we analyze whether keyword search functionality is available, whether there are phrase suggestions, the presentation format of the output, the presence of links to other databases, the ability to share URLs, and the availability of source code. Furthermore, we evaluate the FAIRness (Findable, Accessible, Interoperable, and Reusable) of the databases, as it is of paramount importance for their usability and long-term value. In our review, we include an up-to-date assessment of the webpages hosting virus databases, specifically focusing on (1) `re3data.org`, (2) `FAIRsharing` [24], (3) `The Database`

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

Commons [25], (4) ELEXIR bio.tools [26], and (5) NAR database list [27]. Additionally, we provide an overview of various types of errors that can occur in databases, providing users with suggestions to better address and mitigate these issues during deposition of user data to the sequencing repositories.

2. Exploring the distinctions amongst diverse databases

Here we present a comprehensive overview of 24 virus databases which are currently active and contain virus-centric data (e.g. virus sequences), see Table 1 and Table S3.

We compiled the list of databases for comparisons by including all up-to-date databases mentioned in the previous review papers [22,23], along with active databases obtained from the database catalogs (re3data.org, FAIRsharing, The Database Commons, ELEXIR bio.tools, and NAR database list).

We have only examined up-to-date databases, meaning databases with a website or data update from 2022 onwards. Due to the COVID-19 pandemic, numerous coronavirus-related databases have been developed and established. In total, we identified 49 such coronavirus databases, of which we selected and included three important ones in Table 1 based on our assessment. All other coronavirus databases identified here are listed in the Table S4. For further reference please see a relatively recent review of COVID-19 resources [29]. Additionally, we excluded several databases based on various criteria. Some were considered too specific, focusing on a single virus in a particular region or providing exclusively epidemiological data. Others we classified as tools, networks, or tables rather than databases.

The following non-COVID databases were found in our search of databases and were not included in our Table 1 based on the criteria for rejection listed above: (1) Disease Monitoring Dashboard, (2) RID, (3) Virus-CKB, (4) VirHostNet 3.0, (5) HIV Drug Inter-actions, (6) HEP Drug Interactions, (7) HIV and COVID-19 Registry in Europe, (8) United States Swine Pathogen Database, (9) Global.health, and (10) WestNile.ca.gov.

The database comparison includes the database name, a hyperlink to the database, and an assessment of whether a login is required. The content section specifies the number of sequences and species available. Additionally, the functionality is described, considering aspects such as user-friendliness, the availability of a workbench, the ability to work with personal data, tool availability, subjective complexity, the type of downloads offered, and the possibility of accessing all the data. Last, a short description of each database is provided along with the corresponding references. In Table S3, an additional table can be found that further investigates the following aspects: the presence of keyword search functionality, the availability of phrase suggestion, the presence of cross-links to other databases, the ability to share the URL, whether the webpage is generated via an API, the quality of the result table as output, and the availability of the source code. To facilitate a comprehensive analysis of the database catalogs, we have included a comparison to determine whether each respective database is included or not. For a detailed examination of the catalogs see Section 2.4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

Table 1. Overview and comparison of recently updated (2022 or later) virus-related databases. A comprehensive description of sequence counts especially in the metagenomic context (vOTUs) can be found in Section 2.3. To obtain more detailed information regarding the download process, please see Table S3. 🔒: To access the complete range of features and download data, user authentication is required; #seq-n – number of nucleotide sequences; #seq-p – number of protein sequences; #spec – number of species; – – not known/ no access; use – subjective impression of usability workb. – workbench; own d. – own data: the possibility to work with personal data; compl. – complexity of data base, ranging from simple (🔋) to extensive (🔋); tools – availability of tools and a quantitative ranking: ✔✔– a lot of tools; ✔– available; ✓– only few tools; ✘– no tools available; F – Findability; A – Accessibility; I – Interoperability; R – Reusability (see Table S5 and Figure S1); Down – downloadable via Web (W), FTP (F), and API (A); Click – by one click no data (no), one dataset (one), selected data (sel), or all data can be download (all); re3da – `re3data`; Fshare – `FAIRsharing.org`; DBcom– `Database Commons`; elexir–`ELEXIR bio.tools`; NAR – `NAR Database list`; for VVR: just one of the seven VVR resources is listed;

Name	#seq-n	#seq-p	#spec	Functionality					F	A	I	R	Down	Click	listed in					Cite
				use	workb.	own d.	tools	compl							re3da	Fshare	DBcom	elexir	NAR	
knowledge databases																				
ICTV	0	0	11,273	☺	✘	✘	✘	🔋	😐	😐	😐	😐	W	all	✘	✘	✔	✔	✘	[30,31]
ViralZone	0	0	7	😃	✘	✘	✘	🔋	😐	☺	😐	😃	–	–	✔	✔	✔	✔	✔	[32]
VIPERdb	0	0	211,450	☺	✘	✘	✔	🔋	😃	😃	☹	☺	W	one	✔	✔	✔	✔	✔	[33,34]
Virus-Host DB	0	0	15,178	☺	✘	✘	✔	🔋	😃	😃	☹	😃	F	all	✘	✘	✘	✘	✘	[35]
genomic sequences databases																				
BV-BRC	295,306,161	480,376,932	24,824	😃	✔	✔	✔✔	🔋	😐	😐	☹	😐	WFA	sel	✔	✔	✔	✔	✔	[36]
NCBI Virus	11,345,662	52,734,161	52,414	😃	✘	✓	✓	🔋	☺	😃	😐	☺	WA	all	✘	✔	✘	✘	✘	[37]
NCBI Viral Genomes	261,488	0	52,414	☺	✘	✓	✓	🔋	–	–	–	–	WFA	one	✘	✔	✔	✘	✔	[38]
RVDB	8,454,473	0	11,273	😃	✘	✓	✓	🔋	☹	😐	☹	☹	W	all	✘	✘	✘	✘	✘	[39]
VOGDB	10,327	0	10,327	😐	✘	✘	✘	🔋	☺	😃	☹	😐	WF	all	✘	✘	✘	✔	✘	
Virxicon	599,538	0	11,273	😃	✘	✘	✘	🔋	☹	😐	☹	☹	WA	sel	✘	✘	✔	✔	✘	[40]
ZOVER	64,289	64,289	–	☺	✘	✓	✓	🔋	☹	☹	☹	☹	W	sel	✘	✘	✔	✔	✔	[41–43]
other -omics databases																				
IMG/VR 🔒	15,677,623	0	8,606,551	😐	✔	✔	✔✔	🔋	😃	😃	😐	😐	WA	all	✘	✔	✔	✔	✔	[44,45]
MVIP	0	0	77	😃	✘	✘	✓	🔋	😐	😃	☹	😐	W	one	✘	✘	✘	✔	✔	[46]
Viral Host Range DB	0	0	>279	☺	✘	✔	✔	🔋	😐	😃	☹	😐	–	–	✘	✔	✔	✔	✘	[47]
virus-specific databases																				
GISAID 🔒	15,320,801	0	3	☺	✘	✔	✔	🔋	😐	😐	☹	😐	W	sel	✘	✔	✘	✘	✘	[48–50]
COVID-19 Data Portal	17,317,438	5,340	52	😐	✘	✘	✘	🔋	😃	😃	😐	😃	WFA	all	✔	✔	✔	✔	✔	[51]
COVDB	0	0	1	☺	✘	✔	✓	🔋	☹	☹	☹	😐	–	–	✔	✔	✘	✘	✘	[52]
LANL HIV Database	1,119,274	74,726	2	☺	✘	✔	✔✔	🔋	😐	☹	😐	😐	W	sel	✘	✘	✔	✔	✔	[53,54]
EuResist 🔒	269,205	0	2	😐	–	–	–	–	–	–	–	–	–	–	✘	✘	✘	✘	✘	[55]
HIV Drug Resistance DB	0	0	1	☺	✘	✔	✓	🔋	☹	☹	😐	☺	W	one	✘	✔	✘	✘	✔	[56]
HBVdb	132,609	116,318	1	☺	✘	✔	✔	🔋	😐	😐	☹	😐	W	one	✘	✘	✔	✘	✔	[57]
PaVE	7,228	3,985	1	😃	✘	✔	✓	🔋	☺	😃	☹	☺	WA	sel	✘	✘	✔	✘	✔	[58]
NCBI VVR	932,938	1,254,767	7	☺	✘	✔	✓	🔋	☺	😃	😐	☺	WA	sel	✔	✘	✓	✘	✘	[37]
PSD	4,876	0	1	😃	✘	✔	✔	🔋	☹	😃	😐	☹	W	all	✘	✘	✘	✘	✘	[59]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

2.1. Knowledge databases

These databases play a crucial role in research, facilitating knowledge transfer and providing foundational information. It is important to note that these sources do not directly contain sequences, but provide external links to the sequences, instead serving as repositories of centralized knowledge on viruses.

One of the fundamental tasks in virology is establishing a robust taxonomy to facilitate effective comparison and study of viruses. The International Committee on Taxonomy of Viruses (ICTV) is the global organization tasked by the International Union of Microbiological Societies (IUMS) with developing, refining and maintaining the virus taxonomy down to the level of species [30,31,60]. As of May 2023, the virus taxonomy curated by the ICTV (version: MSL38) comprises 264 families, 2,818 genera, and 11,273 species. The taxonomy is periodically updated, with revisions released at least once and up to twice a year. New entries are proposed by the scientific community and reviewed by expert subcommittees within the ICTV. The categorization of virus groups is based on various characteristics, such as genetic material, genome organization, replication strategy, and host range. Notably, the ICTV has recently embraced the inclusion of virus groups based solely on sequence information, departing from the traditional reliance on virus morphology. Users can download the entire taxonomy as an Excel sheet or browse through it, available at the visual browser. In our opinion, the website may profit from a clearer structure. Helpful are the provided "How-to Videos" as a valuable resource to assist users in effectively navigating the taxonomy.

The ViralZone is a powerful and up-to-date online encyclopedic database that provides summarized expert knowledge on various aspects of viruses, including genomic structure, virus replication cycle, host range, virus taxonomy, and molecular biology [32]. Widely embraced by the research community, it has become a prevalent and trusted resource for obtaining information about specific viruses, serving as a key starting point for addressing novel research questions. In total, it houses detailed descriptions of over 128 families, 567 genera, and 7 virus species (e.g. Influenza A virus and SARS coronavirus 2). Every entry within the database comprises a fact sheet that presents visual representations of the virion and genome, alongside comprehensive details concerning gene expression and replication mechanisms. While the ViralZone itself does not contain sequences in bulk download form, it does provide links to protein and nucleotide sequences of reference sequences within the fact sheet. It is a structured, user-friendly and well-connected website where users and especially virologists can quickly find the information for which they are looking.

The VIPERdb, a specialized database dedicated to icosahedral virus capsid structures, offers a wealth of information derived from both structural and computational analyses [33,34]. The database provides diverse visualizations on various levels, multiple sequence alignments, relevant publications, and useful tools such as anomaly analysis and contact finder. Each of the 1,332 structures are linked

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

to their respective protein sequences on PubMed. The search functionality allows users to explore structures based on taxonomic classifications or specific criteria. One limitation of VIPERdb is its focus on icosahedral virus capsid structures. However, there are ongoing efforts to expand its scope and some helical structures are already included. With its latest release, VIPERdb has introduced a new standalone database on its website, namely Virus World. This comprehensive database encompasses information on 181,476 viruses belonging to 158 families. Virus World also provides the capsid protein sequences for these viruses in some instances. In our opinion, the search function of VIPERdb could be improved as users must have prior knowledge of their virus of interest before using it. Please note that there was an additional database known as VIPR, which has recently been incorporated into the BV-BRC (Bacterial and Viral Bioinformatics Resource Center) as described below.

The Virus-Host DB is a comprehensive and manually curated database that links viruses and hosts using pairs of NCBI taxonomy Ids [35]. It includes viruses with complete genomes stored in NCBI/RefSeq and GenBank, with the accession numbers listed in EBI Genomes. Host information is collected from various sources, including RefSeq, GenBank, UniProt, and ViralZone, and supplemented with additional data obtained from literature surveys. The database offers comprehensive information on 15,179 virus species, encompassing scientific names, lineages, Baltimore groups, RefSeq sequences, database links, and details of 3,791 hosts, enabling users to investigate interactions from both virus and host viewpoints. The database is well interconnected, making it user-friendly and valuable for obtaining an overview of interactions. The database contains a limited amount of information, as it focuses on providing specific linkages rather than comprehensive data.

2.2. Databases containing virus sequences

Genomic, transcriptomic, and proteomic virus sequences serve as a foundational element for a wide range of virus bioinformatics analyses. For example, phylogenetic analysis typically starts with multiple sequence alignment of a collection of sequences. For host sequences we refer to additional -omics databases, see below. Sequence databases serve as a critical starting point for examining genetic variations and functional components. When working with all or a significant portion of sequences from a database for further analyses, e.g. in virus ecology, it is important to be aware of the imbalance in virus representation. In other words, the composition of viruses sequences within a database does not reflect the natural occurrence of viruses.

The Bacterial and Viral Bioinformatics Resource Cen-ter (BV-BRC) is a recently merged platform that integrates various NIAID-funded pathogen-related resources, including the Virus Pathogen Resource (ViPR), the Influenza Research Database (IRD) and PATRIC (the Bacterial Bioinformatics Database and Analysis Resource). With diverse computational tools, the BV-BRC empowers researchers to analyze and interpret large genetic datasets originating from NCBI GenBank and Refseq (see below) as well as specific projects. Users have the ability to search, browse, download, and analyze a multitude of data types, including metadata, taxonomy, genomes, features (ORFs), proteins, protein structures, domains and motifs, epitopes, and experimental data. It also offers a private workbench for the secure analysis and storage of private datasets. The BV-BRC encompasses integrated datasets from mainly pathogenic bacteria, archaea, viruses, and eukaryotes, allowing users to search, browse, download, and analyze various data types such as metadata, taxonomy,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

genomes, features, proteins, and experimental data. Tools and services are categorized into genomics, phylogenomics, protein analysis, metagenomics, transcriptomics, and utilities. The BV-BRC provides access to 295,306,161 virus sequences including 9,763,946 genomes/segments, representing 106 virus families, 1,946 genera, and 24,824 species. Additionally, there are 480,376,932 protein sequences. The BV-BRC has categorized plasmids under the virus category, resulting in the inclusion of plasmid sequences within the overall sequence count. Note that the number of species is higher than that of the official ICTV number because the BV-BRC includes unclassified taxa. Despite the complexity of the platform, efforts have been made to maintain user-friendliness and visual accessibility. Workshops and training opportunities are provided regularly to enhance user proficiency in utilizing the database effectively.

The recently established NCBI Virus interface is a consolidation of various NCBI resources: NCBI Viral Genomes (a former version of NCBI Virus), NCBI Nucleotide (selected for taxonomic classification to viruses), including Refseq, Genbank, Virus Variation Resource, and the old resource NCBI Retroviruses [37]. Virus genome sequences are submitted by users into the public sequence repositories which are part of the International Nucleotide Sequence Database Collaboration (INSDC). The INSDC collaboration is composed of three organizations: the National Center for Biotechnology Information (NCBI) GenBank, the EMBL-EBI European Nucleotide Archive (ENA) and the DNA DataBank of Japan (DDBJ). These three repositories contain the same data, ensuring data consistency across platforms. The sequences from these repositories are frequently utilized as a starting point by other databases, which then apply diverse analyses or visualizations to further explore the data.

NCBI Virus is regularly updated, with the core component being the GenBank and Refseq sequences and well-curated metadata, and additional features being new analysis or visualization functionalities. As of June 2023, 11,345,662 virus nucleotide and 52,734,161 virus protein sequences are accessible through NCBI Virus. These sequences are linked to the NCBI Nucleotide database, providing extensive metadata such as organism, host, taxonomy, publication, organization (e.g. ORF or domains), as well as the corresponding nucleotide or protein sequences. The number of species in the NCBI Virus database, which is 52,414, surpasses the official ICTV count due to the inclusion of unclassified taxa. However, it is important to note that within the broader NCBI GenBank database, there are instances of erroneous sequences that can potentially contribute to false-positive results in analyses (see Section 3 for more details). NCBI offers the Reference Sequence (RefSeq) collection as a comprehensive, integrated, and well-annotated dataset containing diverse data types, including 19,975 nucleotide sequences and 710,847 protein sequences of viruses. These high-quality sequences are extensively utilized by the scientific community.

The search interface is user-friendly and the results are filterable by a wide range of curated metadata, such as taxonomy, length, completeness, host, submitter, genome molecule type, and date. Users can perform a sequence blast or keyword search, with example searches such as "all viruses" or "bacteriophages". Based on our experience, datasets containing up to 100,000 sequences can be readily downloaded, offering users a range of options to select from. Additionally, users can conveniently create their own custom FASTA headers. Several tools are available to perform alignments or phylogenetic analyses with selected sequences.Over the years, NCBI Virus resources have evolved, offering enhanced functionality. Compared to the older NCBI Viral Genomes database [38], NCBI Virus is more organized, functional, and visually appealing.The Virus Variation Resource (VVR) covers seven viruses (Influenza Virus, Dengue Virus, Zika Virus, Rotavirus, West Nile Virus, MERS coronavirus, and Ebolavirus). Only the sub-database for Influenza Virus provides extra functionality, such as an annotation tool. NCBI SARS-CoV-2 Resources is another specific virus database for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

COVID-19 only. In summary, NCBI Virus serves as the go-to resource when working with NCBI virus-related data, offering a visually appealing and user-friendly interface to virus sequence data and metadata.

RVDB

The Reference Viral Database (RVDB) comprises a comprehensive collection of nucleotide sequences, encompassing viral, virus-related, and virus-like sequences (excluding bacterial viruses) [39]. The database provides two versions: an unclustered and a clustered version based on sequence similarity. Researchers can conveniently download all sequences, although it should be noted that the large file size (approximately 20 GB) may result in longer download times. The RVDB, a curated subset of GenBank, is preferred by researchers for bioinformatics analyses due to its comprehensive sequence coverage and ongoing curation efforts. The database is designed with simplicity in mind, offering user-friendly functionalities. Additionally, the database provides a BLAST tool for performing sequence searches, further enhancing its usability for various research needs.

VOGDB
Virus Orthologous Groups

The Virus Orthologous Groups Database (VOGDB) is a regularly updated database that is based on RefSeq virus genomes, providing a comprehensive representation of viral lineages in Virus Orthologous Groups (VOG) for comparative virus (meta-)genomics. The VOGDB contains 10,327 sequences from 10,327 species, grouped into 30,218 virus-specific VOGs, allowing for multiple assignments of the same sequence to different VOGs, reflecting the small functional parts of the genome typically represented by a VOG. While VOGDB currently supports searching for VOGs and provides taxonomic information, direct downloading of a single VOG is not available. Instead, users can access fileshare platforms or choose from 11 different (compressed) file formats for their downloads, which may be slightly disorganized. Surprisingly, there have been no publications published on VOGDB to date.

Virxicon lexicon of viral sequences

The Virxicon is a centralized knowledge base gathering information about viruses and their associated sequences [40]. Virxicon is a database that maintains the ICTV virus taxonomy, incorporating virus sequences from the NCBI Viral Genomes database and GenBank, and annotating them based on the Baltimore classification system. The database comprises a total of 599,538 sequences (the website statistics were not retrievable for the numbers of families, genera or species represented). In their research paper, the authors compare Virxicon with other databases, such as ViralZone, NCBI Virus, and ViPR (now BV-BRC), aiming to combine the strengths of these databases. Virxicon facilitates the bulk download of virus sequences with searchable, well-curated metadata, namely Baltimore class, molecular types, and topological resources. However, it is our impression that the database does not provide unique functionality compared to other virus databases. NCBI Virus and BV-BRC provide a larger number of sequences and more extensive functionality related to sequences, including

 doi:10.20944/preprints202307.0824.v1

search and tools, while ViralZone serves as a more comprehensive lexicon including virus sequence download and curated simple metadata. The Virxicon website offers an intuitive and user-friendly interface, providing search and easy access to information.

ZOVER: the database of zoonotic and vector-borne viruses

The ZOVER, a comprehensive database of zoonotic and vector-borne viruses, aims to integrate virological, ecological, and epidemiological information to enhance understanding of animal-associated viruses and their significant impact on human and animal health [41–43]. ZOVER is a valuable resource, offering a curated subset of NCBI GenBank data and manually collected from published literature focused on four specific hosts: bats, rodents, mosquitoes, and ticks. ZOVER was merged from the Database of Bat-associated Viruses and Database of Rodent-associated Viruses. The ZOVER database includes 64,289 sequences, combining both protein and nucleotide sequences, making it challenging to differentiate them individually. ZOVER presents data in a well-organized, visualized and user-friendly manner, providing a comprehensive and visually appealing platform for accessing information. ZOVER offers a valuable tool for researchers in the field, as it provides curated and easily accessible data, enhances data visualization, and offers a user-friendly interface for efficient exploration and analysis. Users can easily navigate the database using taxonomy-based searches or various search options, including sequence-based, text-based or region-based.

2.3. Omics databases

The emergence of databases dedicated to -omics data and analyses represents a remarkable advancement in the field of virology. These specialized resources go beyond traditional databases, providing a next-level platform for researchers to delve into the vast realm of -omics data sets and unlock hidden viral treasures. By focusing on -omics data, which encompass various '-omics' disciplines such as genomics, metagenomics, transcriptomics, and proteomics, these databases offer a comprehensive view of the viral world at a molecular level. These databases serve as central hubs for storing, organizing, and analyzing -omics data sets, enabling researchers to explore uncharted territories and uncover previously unknown virus species.

The online platform The Integrated Microbial Genomes/Virus (IMG/VR) provides their own geNomad analysis workflow, with which IMG/VR systematically identifies viral sequences from user-contributed and publicly available datasets, providing researchers with a comprehensive collection of 15,677,623 Uncultivated Viral Genomes (UViGs) [44,45] with different levels of confidence for download and analysis. The resource incorporates data from the metagenomic and metatranscriptomic JGI database IMG/M, RefSeq database, and three specific virus databases, while enhancing the annotation process with genome quality estimation, up-to-date taxonomic classification, and microbial host taxonomy prediction. IMG/VR offers users a comprehensive platform with abundant information, analysis tools, and links to sub-databases, including detailed meta-information and statistics for each virus. However, for us its navigation and accessibility pose challenges, necessitating substantial time investment for users to become acquainted with its features. The database contains an impressive collection of 15,677,623 putative viral sequences, categorized into viral genomes and Single-scaffold UViGs, organized into viral operational taxonomic units (vOTUs) across various viral families, genera, and species. Despite its importance in the study of Uncultivated Viruses (UViGs), users should be mindful of IMG/VR's complexity and lack of user-friendliness, requiring a login for full functionality and demanding considerable effort to effectively explore and utilize all available features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

The Multi-omics Portal of Virus Infection (MVIP) collects and analyzes virus infection-related high-throughput sequencing data, integrating comprehensive meta-information [46]. It enables -omics data analysis and visualization, presenting a summary table of samples for specific tissues and viruses. Users can access detailed datasets, including differential expression, pathway enrichment, and alternative splicing, which are downloadable. MVIP provides external resource links and allows user submissions for broader analyses and database enhancement. MVIP offers valuable information and analysis for specific biosamples, driving advancements in the understanding of virus infection, and provides users with the opportunity to suggest biosamples for integration. Currently, MVIP boasts a dataset comprising approximately 6,586 sequencing samples derived from 77 distinct viruses, such as SARS-CoV-2, SARS-CoV, DENV, ZIKV, and IAV, across 33 host species, including *Homo sapiens* and *Mus musculus*. MVIP is a visually appealing database that provides comprehensive -omics data analysis capabilities, serving as both a resource for analyzing existing data and a knowledge base for researchers conducting their own sequencing projects, offering insights into the availability of suitable datasets for specific research questions, despite occasional short loading delays.

The Viral Host RangeDatabase (VHRdb) is a unique resource that consolidates experimental data on the range of hosts a virus can infect [47]. Despite the wealth of host-range experiments conducted in laboratories, this valuable data is often inaccessible and underutilized. The VHRdb is an online platform that centralizes experimental data on viral host ranges, allowing users to browse, upload, analyze, and visualize results. Currently, it contains 17,170 interactions between 776 viruses and 2,041 hosts from 20 datasets. Among the 776 viruses in the VHRdb, 303 are linked to the NCBI, representing 279 species from 25 families. The comprehensive overview of virus-host interactions is presented in a visually appealing table, categorizing the relationships into "No infection," "Intermediate," and "Infection." The VHRdb provides extensive and helpful documentation, including Quick Start guides, to assist users in navigating and utilizing the database effectively. However, a limitation of the VHRdb is its relatively limited representation of viruses, which are not be evenly distributed across various viral families, relying heavily on available studies that may not provide comprehensive coverage. To mitigate this, users can upload their own data for public access or private use. Despite these limitations, the existing studies are visually presented excellently, allowing for straightforward interpretation and analysis.

2.3.1. Specific databases

In addition to comprehensive databases, there are numerous virus-specific databases available. If one is working on a specific virus, it is worthwhile to explore specialized databases dedicated to that particular species. Here is a brief list of potential databases that primarily focus on a single virus species. One exception is the NCBI VVR database, which specifically addresses seven different viruses, as mentioned earlier. For coronaviruses, we rely on the following databases: (1) GISAID (2) COVID-19 Data Portal and (3) Stanford Coronavirus Antiviral & Resistance Database (COVDB) [48–52]. For HIV, we have the following databases listed in Table 1: (1) LANL HIV Database, (2) EuResist, (3) HIV Drug Resistance DB, and (4) PSD [53–56,59]. Due to our inability to access EuResit by the time of submission, the investigation could not be conducted as extensively as with other databases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

For Hepatitis B virus, the dedicated database is HBVdb [57]. And for papillomaviruses, we have the PapillomaVirus Episteme (PaVE) database [58].

2.3.2. Non-viral specific databases

In addition to virus-specific databases, there exist numerous databases that are also important for virus research in the broader fields of biology and genomics. We would suggest also for interested users to keep an eye on initiatives such as the Global Core Biodata Resources which seek to identify invaluable, long-term resources for the life sciences.

UniProt [61] provides a vast collection of protein sequences and functional insights, including those from viral sources, enabling researchers to unravel the molecular mechanisms and biological functions of viruses. The Rfam database [62], widely recognized and utilized, encompasses RNA families with detailed sequence alignments, secondary structures, and covariance models, while the Pfam, now InterPro database serves as an extensively employed resource, offering multiple sequence alignments and hidden Markov models for protein families [63]. The NCBI houses additional non-virus-specific databases, such as the Gene Expression Omnibus, which serves as an international public repository for high-throughput functional genomic data sets, or Sequence Read Archive(SRA), a valuable resource that provides access to biological sequence data, fostering reproducibility and enabling new discoveries through data set comparisons within the research community [64]. Of note is the new NCBI datasets browser (currently in beta version), which provides easy searchable access to different NCBI databases and NCBI Taxonomy via fact sheets. Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive biological database that represents molecular networks and pathways, facilitates analysis of genomic data, and integrates drug labels and disease databases, making it one of the most widely used resources in the field [65–67]. The miRBase is the central repository for microRNA (miRNA) [68]. It enables users to search and browse entries representing hairpin and mature miRNA sequences. Entries can be retrieved by various criteria, and both sequence and annotation data are available for download. The database currently includes 320 precursors and 510 mature miRNAs related to viruses.

2.3.3. Other databases

Additionally, there exist virus-related online platforms that link together pre-existing tools, databases, and datasets. These websites serve as valuable resources for researchers and practitioners seeking to leverage existing resources and foster collaboration within the scientific community. By linking together disparate resources, these platforms contribute to the dissemination and accessibility of scientific information, promoting efficient utilization of available resources for further research and innovation. One example is the European Virus Bioinformatics Center (EVBC) website, on which a total of 275 entries are linked, sorted by software type (such as database, command-line tool, or similar), virus family, or functionality [69,70]. Another example is iVirus.us, which provides a platform to access 27 tools and 21 datasets [71,72].

2.3.4. FAIR evaluation

Many virus databases aim to support the (re)use of virus data and enable processing using machine-learning methods. Both goals can be facilitated by adopting the FAIR principles. We therefore included an evaluation of FAIR properties in our database overview using FAIR principles checklist. Where a virus database had a table featuring one virus per row, the entries were evaluated as research objects (please refer to the data sources of the FAIR evaluation of the databases at the Table S5). Where available, a virus sequence was considered to be "data". Note that some databases in the list were therefore excluded from the FAIR evaluation because of a lack of a comparable research object. The databases that did not have comparable research objects were NCBI Viral Genomes (due to being a central website linking to different resources) and the EuResist database (to which we did not have access by the time of submission). The FAIR scores are based on presence/absence for each of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

checklist criteria as has been done previously in the context of data deposition of nuclear magnetic resonance data [73]. The scores are out of four for the subcriteria in Findability, Accessibility and Reusability, and are out of three for those of Interoperability. A more complete description of the FAIR Principles checklist can be found in the Figure S1.

In general the FAIR scores of the content of the active databases reviewed here (summarized in Table 1 and the full table available in Table S5) ranged from less FAIR for the smaller or older databases and more FAIR for the larger and newer databases. An important component of the Findability score is the assignment of a database-given global and persistent identifier; while the large platforms such as BV-BRC and IMG/VR featured this, the smaller databases such as HBVdb often used an external id, e.g. the NCBI Accession ID or TaxIDs. This might be due to the differing aims of the virus databases as some are focused on data reuse and machine-readability while others may have simpler goals such as cross-linking available knowledge. Accessibility for the databases was generally positive owing to web-accessible links and straightforward download options (see also Table S3). Further, the overall low score for Interoperability reflects the lack of standards for all virus metadata; while there exist clear ontologies, e.g. for clinical data (as for the HIV drug resistance DB) or for pathogenic virus metadata (see the Genomic Standards Consortium (GSC)) this is not yet the case for metadata for all viruses. This is currently a target for various groups such as the GSC (which are responsible for the Minimum Information about Sequencing standards which are used by the INSDC repositories), the Gene Ontology consortium, the Genomes Online Database (GOLD) which complements the IMG databases of the Joint Genome Institute JGI and other efforts such as Bernasconi *et. al.* with the Viral Conceptual Model [74,75]. This shows that community-wide metadata standards are poised to improve interoperability in the near future. Last, the Reusability of many of the virus databases would benefit from the inclusion of formal licenses describing the reuse of their data (see Choose A License). Overall, this FAIR evaluation was a first for virus databases and highlighted several areas for improvement.

2.4. Catalogs of databases

To assist users in selecting appropriate databases, scholarly journals and other entities have established catalogs that employ various criteria for indexing databases based on different criteria to improve their findability and accessibility. Here we describe five catalogs of databases: (1) re3data.org, (2) FAIRsharing [24], (3) The Database Commons [25], (4) ELEXIR bio.tools [26], and (5) NAR database list [27], see Figure 1. We analyzed a range of entries, narrowing down to virus-specific databases, categorizing them based on their up-to-date status and relevance to COVID-19, while excluding non-virus databases that didn't meet the criteria.

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

The re3data.org website is a web-based registry that facilitates data discovery, access, and sharing for researchers. Its comprehensive metadata on data repositories allows researchers to identify repositories that align with their specific data management needs. The platform has a particular focus on the FAIR principles. There are 3,125 entries on this platform, of which 2,181 are databases or scientific and statistical data formats in terms of content types. Among them, there are 186 virus-related entries identified using the search term "virology," of which only 24 are virus-specific, and 17 are considered up-to-date. Nine of these databases have been extensively described in our curated Table 1, seven are dedicated to coronavirus research (see Table S4), and one database, namely WestNile.ca.gov, was excluded due to its narrow focus.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

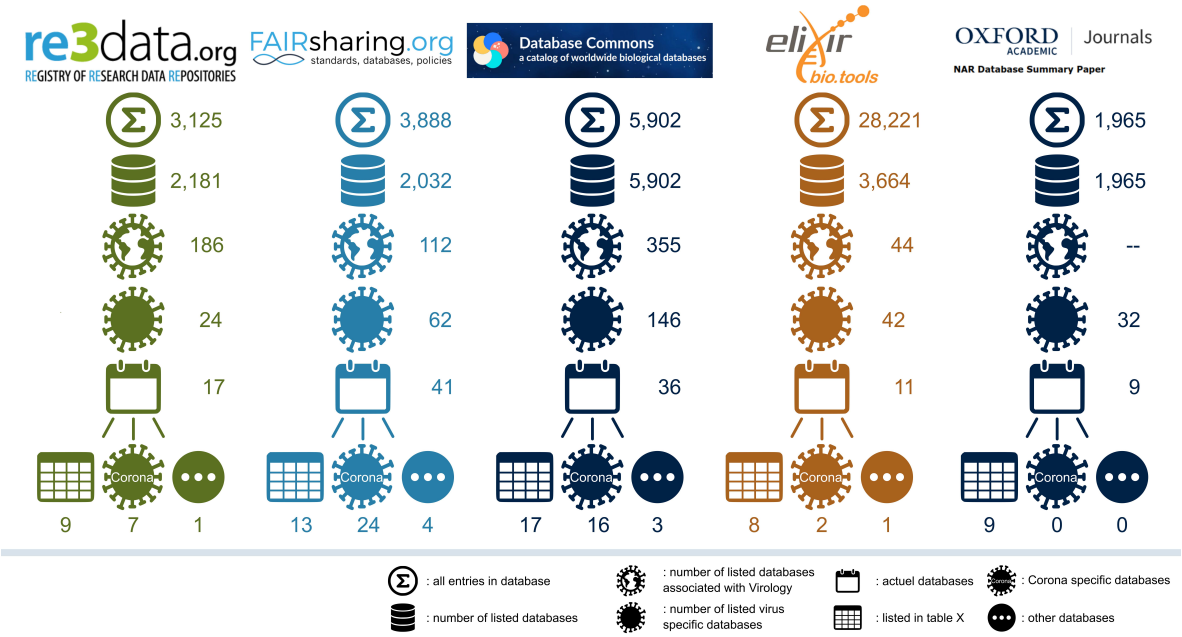

Figure 1. Comparison of five catalogs content: (1) re3data.org, (2) FAIRsharing [24], (3) The Database Commons [25], (4) ELEXIR bio.tools [26], and (5) NAR database list [27]. Among all up-to-date virus-specific databases, they were categorized into three groups: inclusion in our curated Table 1, exclusive focus on coronavirus (see Table S4), or non-corona-related databases mentioned in the text, see Section 2.4.

The online platform FAIRsharing, is designed to enhance the visibility of scientific data standards, databases, and policies for the scientific community. The platform includes a registry of data standards, databases, policies, collections, and organizations that details each resource, such as its scope, history, and adoption status. In total, 3,888 entries are listed in the registry, of which 2,032 are repositories or knowledge-bases. Among these, 112 are virus-related (identified using the keyword "virology"). However, only 62 of these resources are virus-specific, and only 41 are up-to-date. These 41 resources can be further classified into 13 listed in our curated Table 1, 24 related to coronavirus data (see Table S4), and 4 we excluded: (1) HIV Drug Interactions, (2) HEP Drug Interactions, (3) Global.health, and (4) HIV and COVID-19 Registry in Europe. These databases were excluded due to their restricted focus, such as focusing only on drug interactions of a particular virus or containing primarily epidemiological data, which did not align with our definition of a comprehensive virus database. Additionally, one of the databases resembled more of a network than a traditional database.

The Database Commons is a curated catalog of biological databases that organize databases based on data type, species, and subject matter. It provides detailed metadata for each database, including name, URL, description, hosting institution, and contact information. Within the Database Commons, there are currently 5,902 entries listed. Among them, 355 databases fall under the "Data Object" virus category. Of these, 146 are virus-specific, and 36 are considered up-to-date. These 36 databases can be further categorized as 17 listed in our curated Table 1, 16 coronavirus databases (see Table S4), and

3 other databases (Disease Monitoring Dashboard, RID, and Virus-CKB). The additional databases were excluded due to their specific nature, such as being more tool-oriented or containing limited data with only two tables rather than meeting the criteria of a comprehensive virus database.

A comprehensive registry of bioinformatics resources is established through a community-driven curation effort supported by ELIXIR, a (ELIXIR). ELIXIR bio.tools serves as the dedicated registry within this infrastructure, ensuring the sustainable upkeep of the curated information [26]. Collaborative curation, tailored to local needs and facilitated by a network of partners, enables the continuous development and accessibility of this valuable resource. In total, there are over 28,211 resources listed in the registry, including various tools. Among them are 3,664 databases, and a search using the keyword "Virology" identified 44 databases in this category. Out of these, 42 databases are virus-specific, with 11 being up-to-date. Eight of these virus-specific databases are included in our curated Table 1. Additionally, we have identified two up-to-date coronavirus-specific databases and one particular database, namely the U[illegible]hogen Database, which we excluded.

NAR Database Summary Paper

To our knowledge, the Nucleic Acids Research Journal Database Summary Issue NAR is the oldest known list of databases. Published annually, it provides descriptions of new and updated databases that contain nucleic acid and protein sequences and structures [27]. The NAR provides the links for these databases at the Molecular Biology Database Collection. These databases are categorized into genomics, transcriptomics, proteomics, metabolomics, and structural biology. Presently, it includes a total of 1965 databases. Each database is described in detail, including its scope, content, features, relevant citations, and links to access the resource. The most recent issue from January 2023 lists 32 databases in the "virus genome database" category. Among them, 9 are considered up-to-date and included in Table 1.

In conclusion, despite the availability of database catalogs that assist researchers in finding relevant resources, there are still challenges and limitations to address. These catalogs lack virus-specific content and often do not reflect the current status or usability of the databases. Furthermore, there is a need for better metadata standardization and information on the reliability and quality of the databases. Although these catalogs serve as a starting point, they may not provide comprehensive and detailed information for researchers to make informed decisions about utilizing the databases effectively.

3. Evaluation of errors in the NCBI and BV-BRC

Despite diligent curation, databases like the NCBI Nucleotide database often harbor errors, including those arising from user-generated data. These errors extend beyond user mistakes and stem from the need to adapt databases to rapidly evolving scientific fields like virus taxonomy. Viromics poses challenges, such as the absence of a universal viral gene, the facilitation of horizontal gene transfer, and the need for specific data models and standards. Addressing these challenges requires specialized protocols for RNA and DNA viruses and mitigating experimental biases related to enrichment methods [76]. As user-friendly pipelines for comparative genomics become more prevalent, the quality of viral sequences from databases used as input becomes crucial for reliable and accurate bioinformatic analyses. Incorrect input data can undermine the validity of downstream results, regardless of the pipeline employed. Therefore, it is essential to critically scrutinize and validate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

the outputs obtained from these pipelines, mainly when errors exist in the utilized databases. Taking NCBI and BV-BRC as an illustrative example, it is important to acknowledge the various types of errors that can occur in databases, namely (1) Taxonomy errors, (2) Naming and labeling errors, (3) Missing information, (4) Sequences errors, (5) Wrong orientation, and (6) Chimeric sequences.

3.1. Taxonomy errors

As outlined above, the ICTV serves as the authoritative source for virus taxonomy. The NCBI Taxonomy [77] encompasses most of the ICTV entries, supplemented by a high number of additional taxa, see Figure 2. ICTV defines only approximately one-fifth of the species mentioned in NCBI. In rare cases, inconsistencies may occur, but they are typically resolved in subsequent updates. The higher number of species in the NCBI Virus database compared to the official ICTV count is primarily attributed to the inclusion of unclassified and unverified taxa. TaxIDs are commonly used for taxonomic groups in databases like NCBI and UniProt, but not in ICTV. The NCBI Virus database houses around 53,000 virus species, with nearly 1.5 % having more than one TaxID assigned (e.g., Lomovskaya virus with 6 TaxIDs), see Figure 2. Multiple TaxIDs in NCBI for certain viruses result from ongoing refinement of virus classification, including species mergers and new TaxID assignments. An additional challenge arises in cases where assigning two TaxIDs to a specific sequence is needed, as seen in studies involving integration sites where a sequence represents both the virus and the host [78]. Implementing a mechanism to accommodate such dual assignments would better reflect the intricacies of these scenarios.

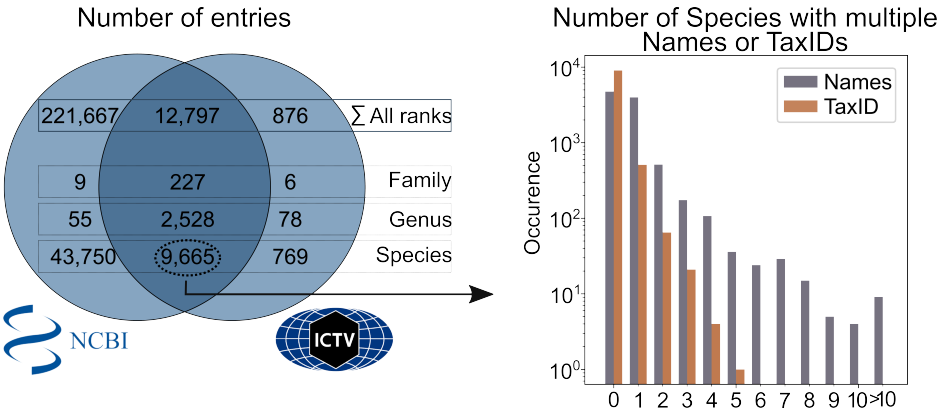

Figure 2. Here, we comprehensively compare ICTV and NCBI Taxonomy for viruses and highlight the prevalent existence of multiple names and TaxIDs within the species' taxonomic units. **Left:** The upper part showed the number of taxonomic virus entries in the respective database for individual ranks. Only virus species that are currently in ICTV and NCBI Taxonomy were listed on the right. **Right:** Among these 8,163 species, the investigation examined the occurrence of multiple names and TaxIDs. The y-axis is presented in a logarithmic scale. Entries with a value of 0 indicate no presence of additional names or TaxIDs.

3.2. Naming and labeling errors

Within the realm of virus labeling, an ongoing systematic error can be observed, where users label sequences as "complete genomes" despite the partial or complete absence of untranslated region (UTR) sequences [79]. For example, the Hepatitis C virus sequence (EU255989.1), even though the UTR sequences are missing. The widespread occurrence of this systematic error is exemplified by the Hepatitis C virus in the BV-BRC, where a search for complete genomes yielded 3,676 results. Upon closer examination, 69 specific sequences did not meet the expected length of 8.500 bp, indicating that these are incomplete genomes.

The length of the virus genomes alone is not sufficient to demonstrate the extent of the problem. To assess the prevalence, a blastn (version: 2.5.0, E-value 10^{-4}) search was conducted on the 3,676

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

listed complete Hepatitis C virus sequences using representative 5'UTR sequences, revealing that only approximately 80 % of the sequences contain the conserved 5'UTR. Inconsistency in the usage of completeness terminology across databases is another issue. When examining all sequences labeled as complete in the NCBI Virus database, it was discovered that 1.95 % of them are classified differently in terms of completeness in the BV-BRC database.

Incorrect metadata can extend beyond labeling, as exemplified by the case of sequence AJ000888.1, initially labeled as Hepatitis C virus but actually belonging to a human sequence. As a result, it is incorrectly identified as of viral origin despite being present in the human genome.

Another source of errors besides the labeling of the virus is the official name, which is used by many databases. Approximately 15 % of virus species described in NCBI Taxonomy have multiple names, with the virus species *Gallid alphaherpesvirus 2* having an additional 16 names. In Table S6 lists different types of submission and naming errors. One cause could be the lack of a convention for naming virus species for a long time, which allows multiple names to be submitted [80]. With every sequence upload of a user potential new errors may occur, e.g. by filling the organism field incorrectly.

The naming of fasta headers, can also contribute to confusion. Sequences can contain cryptic names (e.g., JB021961.1-Sequence 12 from Patent EP2495325, A12995.1-Fragment Ba131, or E04227.1-cDNA to satellite RNA). This can cause difficulties for forward analysis. A more user-friendly alternative is implemented in the NCBI Virus, where the user can build the name of the sequence from predefined properties independently (e.g., Accession, Species, Length of sequence, and further).

3.3. Missing information

To maximize the utility of genomic sequences for various purposes, it is essential to collect metadata on the properties of the pathogen and make them available in organized, clear, and consistent formats. Several studies have focused on identifying the necessary minimum of metadata [21,81]. In this specific case, the variation among publications and their focus on different types of data make it challenging to reach a clear consensus. Nonetheless, we propose four metadata groups to consider: collection attributes (Year, Source, Country, Host), database crosslinks (GenBank Accession, Taxon Linage ID, Publication), species variations (Lineage, Strain, Subtype), and sequence information (Segment Information, Genome Quality, and Genome Completeness Status).

Almost a quarter (28.57 %) of the 9,763,946 genomes/segments described in the BV-BRC, are missing more than half of the described metadata above. Astonishingly, for only 68.12 % of the sequences a lineage information is provided (see Table S7 for detailed findings). The annotation of features, such as genes, UTRs, and CDSs, plays also a role and can be erroneous or missing. This is exemplified by the case of sequence CS179664.1, a Hepacivirus hominis sequence, where the feature annotations are found to be entirely missing.

To counteract missing or incorrect metadata, several guides and tools are available to help explain how to submit data to the INSDC repositories (see Submitting Data to ENA and NCBI GenBank Submissions, the NCBI Submission Help Page and the NCBI Virus Submission Help Page [82–84]. Relevant information such as sample collection, sequencing methodology, and bioinformatic procedures should be noted [85–88]. Assembly strategies have been shown to significantly affect resulting sequences, emphasizing the importance of including this information [85–88]. Guidelines are available to help ensure verification, addressing the issue of missing or incorrect metadata [82].

3.4. Sequence errors

Likewise, nucleotide or protein sequences in NCBI and other databases can contain errors, see Figure 3.

In addition to the challenge of differentiating sequencing errors from the impact of rapid viral evolution, in some cases sequences lack even any meaningful sequence information. A striking illustration of this is observed in the sequence of 1PFI_C-Chain C, PF1 VIRUS STRUCTURE, which is composed of a solitary N nucleotide.

 doi:10.20944/preprints202307.0824.v1

Furthermore, errors in (reference) sequences can have a cascading effect on subsequent sequences e.g., during reference-based assemblies, potentially propagating inaccuracies throughout computational downstream analyses. These errors, combined with methodological challenges in genome assemblies, further emphasize the need for scrutiny and validation of genomic data to ensure the reliability of research findings [85,88,89]. Alternately, this serves as an extra caution for users to perform a quality control check when conducting bioinformatics analyses using existing data sets.

3.5. Sequence orientation error

Viruses have diverse genomes and are classified into different Baltimore classes based on their genome. The `NCBI` default for user-uploaded virus sequences is the positive sense strand, which is counter intuitive for e.g. viruses in Baltimore class 5 (negative single-stranded viruses).

Upload of sequences in wrong orientation, as seen in examples like the *Orthomarburgvirus marburgense* sequence KU059750.1, leads to misinterpretation and further complications. This issue is exacerbated by the fact that many tools do not verify the correct orientation, resulting in nonsensical output that can misguide subsequent analyses. This issue extends to the annotation of functional units, including proteins, which are also provided in reverse order. To identify incorrectly uploaded sequences in terms of orientation, the ORF density was used, which refers to the number of open reading frames (ORFs) found on a particular strand. Out of the 3,676 Hepatitis C virus genomes labeled as complete in the `BV-BRC`, a total of 205 sequences exhibited discrepancies in ORF density, indicating potential errors in their orientation. This highlights a systematic problem in the database.

3.6. Chimeric sequences

Another type of error is the presence of chimeric sequences, where a portion of the sequence is derived from the virus while another part originates from a non-viral source (e.g. the host). It is important to note that such chimeric sequences can arise due to biological factors or process errors, such as assembly mistakes. In extreme cases, this leads to thousands of hits being found in the host genome with a viral sequence, although the originating sequence part is not viral.

One example highlighting this issue is observed in the last 280 nucleotides of a specific Zika virus sequence KY766069. A comparative alignment between this sequence and the two other `RefSeq` Zika virus sequences clearly demonstrates that the 3' end of KY766069 does not originate from Zika virus, as illustrated in Figure S2.

The Zika virus sequence (KY766069) exhibits a fragment of the AluSx repetitive element in its 3' end, resulting in over 200,000 false positive hits in a `blastn` search (version: 2.5.0, E-value $< 10^{-4}$) [90] on the human genome (hg38.p13). The inclusion of a human partial sequence within the Zika virus sequence is likely a result of a sequencing or assembly error. This highlights the impact that a single chimeric sequence within a dataset can have on subsequent results or outputs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 July 2023 doi:10.20944/preprints202307.0824.v1

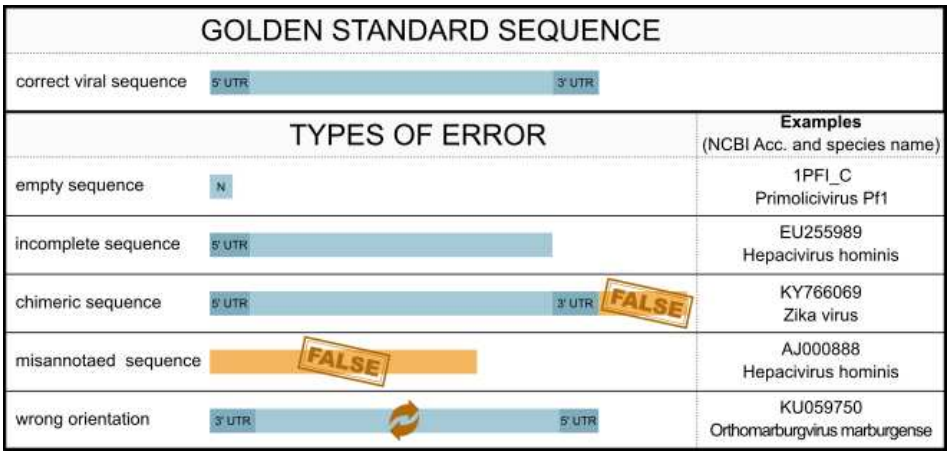

Figure 3. Various sequence based errors. An artificial, correctly labeled viral sequence serves as an example of a golden standard sequence.

In summary, it is crucial to acknowledge the presence of errors in databases like the `NCBI Nucleotide` database to ensure reliable and accurate downstream analyses. Errors can arise from various factors, including taxonomy inconsistencies, naming and labeling errors, missing information, sequence errors, wrong orientation, and chimeric sequences. These errors can undermine the validity of downstream results and highlight the need for critical scrutiny and validation of data. Addressing these errors is essential to maintain the integrity and reliability of virus databases for scientific research.

4. Outlook and Conclusion

Here, we provided a comprehensive assessment of active virus databases, defining these as any database last updated in 2022 or later which contained virus-related research data. Our list of 24 databases was compiled through an exhaustive search of active virus databases from two previous reviews of virus databases as well as five catalogs. For the first time, our review includes a thorough evaluation of database usability and content – including the number of species and sequences contained therein – as well as a FAIRness comparison. We hope this overview will help guide prospective users of these databases. We refrain from suggesting a particular database for use because this highly depends on an individual researcher's needs. For the knowledge databases that cover a broad spectrum of viruses, we provide a detailed overview and offer suggestions for their potential applications.

The content of the virus databases varied widely depending on the scope, number of data types and tools, and number of virus species. Here we presented in detail four knowledge databases, seven sequences databases, three -omics databases and ten databases focusing on specific viruses. The databases featured a range of -omics datasets integrated and in combination with sampling, host, collection, environmental and other metadata. In terms of virus species represented, several focused only on one or a few species (our so-called "specific virus databases") while on the other end of the range, the large databases featured upwards of 1 million species. The virus databases also varied in terms of usability, with larger databases sometimes presenting a more complex user experience in which it takes longer to orientate oneself, although some smaller databases could also improve their usability. In general, the databases we listed here were good in terms of Findability and Accessibility, but we found that community-wide metadata standard development and explicit listing of formal data usage licenses would improve Interoperability and Reusability.

We found that researchers might encounter challenges when using existing catalogs of databases for virus-related databases due to limited virus-specific content and searchable metadata. To address this, efforts are ongoing to curate virus-specific database lists (as in the virus subsection of the `NAR database list`) and to improve domain-specific metadata (which is a goal of `re3data`, `FAIRsharing`, the `Database Commons` and `ELIXIR bio.tools`).

In part due to FAIRification, we expect that these database catalogs will only improve. We suggest that these catalogs can serve as a useful starting point for any researcher, which can complement curated collections, review articles and word of mouth within specific disciplines specifically for searching a particular database, for example, a particular virus or a certain data type. While finding the right database for one's research needs may be challenging, using these existing resources is crucial for modern virus research especially due to the high volume of multi-dimensional data available on viruses. We suggest that the landscape of virus databases be regularly evaluated, and for core virus databases to be included in at least university-level virus-related courses when relevant (e.g. `NCBI Virus`).

Recent sequencing advancements and the growing interest in viruses within the field of virus bioinformatics led to rapid changes in the virus database landscape. We illustrated this rapid development by referencing the Sharma *et al.* 2015 and Mcleod and Upton 2017 reviews of virus databases and tools [22,23]) in which only 22 % and 23 % of the virus databases listed in either review are still active.

Ensuring longevity for virus sequence databases includes FAIR principle-like criteria in addition to regular maintenance and updates, creation of backups and archives, collaboration, funding, and trust and usage by the community [18,19]. For databases, improving FAIRness could help to foster these content-related aspects and contribute to longevity. Moreover, funding plays a large role in the staying power of a database, which in turn is influenced by the database's ease of use and content quality. Ensuring data and metadata quality is a key concern, and efforts are being made by teams at the `INSDC` repositories and other institutions to curate existing metadata. We underline the need for clear data submission guidelines and the inclusion of curated sequences together with regular updating and removal of outdated or redundant sequences. Database content errors can arise from various factors, including taxonomy inconsistencies, naming and labeling errors, missing information, sequence errors, wrong orientation, and chimeric sequences, which can be attributed to either user input or inherent database discrepancies. These errors can undermine the validity of downstream results and highlight the need for critical scrutiny and validation of data. Addressing these errors is essential to maintain the integrity and reliability of virus databases for scientific research.

In conclusion, the databases listed here represent current knowledge of viruses and the current review will help aid future users find databases of interest based on content, functionality, and scope. Use of virus database is integral to gain new insights into the biology, evolution, and transmission of viruses, and develop new strategies to manage virus outbreaks and preserve global health.

Author Contributions: Conceptualization, M.R. N.A.C., and M.M.; methodology, M.R. and N.A.C.; investigation, M.R. N.A.C. and S.S.; writing—original draft preparation, M.R., and N.A.C.; writing—review and editing, all; visualization, M.R. and S.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the DFG grant "NFDI4Microbiota" number NFDI 28/1; the TMWWDG grant "DigLeben" number 5575/10-9 TMWBDG; the EU Horizon 2020 grant "VIROINF" number 955974; the DFG grant "Aquadiva" number CRC 1076; and the DFG grant "Balance of the Microverse" number EXC 2051.

Data Availability Statement: Data is contained within the article or supplementary material. The data presented in this study are available in the Supplementary Material.

Acknowledgments: The authors gratefully acknowledge Dr. Franziska Hufsky for help with proofreading and formatting the manuscript. We would like to thank Sandra Triebel and Dr. Kevin Lamkiewicz for providing examples that were incorporated into the error Sec.3. We used ChatGPT (May 24 Version) to help write starting material for the introduction.

Conflicts of Interest: N.A.C., S.S. and M.M. are in the beginning stages of developing a virus sequence database as part of the NFDI4Microbiota consortium.

References

1. Hendrix, R.W.; Smith, M.C.M.; Burns, R.N.; Ford, M.E.; Hatfull, G.F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proceedings of the National Academy of Sciences* **1999**, *96*, 2192–2197. Publisher: Proceedings of the National Academy of Sciences, doi:10.1073/pnas.96.5.2192.

2. Mushegian, A. Are there 1031 virus particles on earth, or more, or fewer? *Journal of bacteriology* **2020**, *202*, e00052–20.
3. Grubaugh, N.D.; Ladner, J.T.; Lemey, P.; Pybus, O.G.; Rambaut, A.; Holmes, E.C.; Andersen, K.G. Tracking virus outbreaks in the twenty-first century. *Nature Microbiology* **2019**, *4*, 10–19. doi:10.1038/s41564-018-0296-2.
4. Armstrong, G.L.; MacCannell, D.R.; Taylor, J.; Carleton, H.A.; Neuhaus, E.B.; Bradbury, R.S.; Posey, J.E.; Gwinn, M. Pathogen Genomics in Public Health. *New England Journal of Medicine* **2019**, *381*, 2569–2580, [https://doi.org/10.1056/NEJMsr1813907]. PMID: 31881145, doi:10.1056/NEJMsr1813907.
5. Malmstrom, C.M.; Martin, M.D.; Gagnevin, L. Exploring the emergence and evolution of plant pathogenic microbes using historical and paleontological sources. *Annual Review of Phytopathology* **2022**, *60*, 187–209.
6. Jones, R.A.C.; Boonham, N.; Adams, I.P.; Fox, A. Historical virus isolate collections: An invaluable resource connecting plant virology's pre-sequencing and post-sequencing eras. *Plant Pathology* **2021**, *70*, 235–248. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ppa.13313, doi:10.1111/ppa.13313.
7. Roux, S.; Adriaenssens, E.M.; Dutilh, B.E.; Koonin, E.V.; Kropinski, A.M.; Krupovic, M.; Kuhn, J.H.; Lavigne, R.; Brister, J.R.; Varsani, A.; Amid, C.; Aziz, R.K.; Bordenstein, S.R.; Bork, P.; Breitbart, M.; Cochrane, G.R.; Daly, R.A.; Desnues, C.; Duhaime, M.B.; Emerson, J.B.; Enault, F.; Fuhrman, J.A.; Hingamp, P.; Hugenholtz, P.; Hurwitz, B.L.; Ivanova, N.N.; Labonté, J.M.; Lee, K.B.; Malmstrom, R.R.; Martinez-Garcia, M.; Mizrachi, I.K.; Ogata, H.; Páez-Espino, D.; Petit, M.A.; Putonti, C.; Rattei, T.; Reyes, A.; Rodriguez-Valera, F.; Rosario, K.; Schriml, L.; Schulz, F.; Steward, G.F.; Sullivan, M.B.; Sunagawa, S.; Suttle, C.A.; Temperton, B.; Tringe, S.G.; Thurber, R.V.; Webster, N.S.; Whiteson, K.L.; Wilhelm, S.W.; Wommack, K.E.; Woyke, T.; Wrighton, K.C.; Yilmaz, P.; Yoshida, T.; Young, M.J.; Yutin, N.; Allen, L.Z.; Kyrpides, N.C.; Eloe-Fadrosh, E.A. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology* **2019**, *37*, 29–37. Number: 1 Publisher: Nature Publishing Group, doi:10.1038/nbt.4306.
8. Lauber, C.; Seitz, S. Opportunities and Challenges of Data-Driven Virus Discovery. *Biomolecules* **2022**, *12*. doi:10.3390/biom12081073.
9. Hatano, Y.; Ideta, T.; Hirata, A.; Hatano, K.; Tomita, H.; Okada, H.; Shimizu, M.; Tanaka, T.; Hara, A. Virus-Driven Carcinogenesis. *Cancers* **2021**, *13*, 2625. doi:10.3390/cancers13112625.
10. Carroll, D.; Daszak, P.; Wolfe, N.D.; Gao, G.F.; Morel, C.M.; Morzaria, S.; Pablos-Méndez, A.; Tomori, O.; Mazet, J.A.K. The Global Virome Project. *Science* **2018**, *359*, 872–874. Publisher: American Association for the Advancement of Science, doi:10.1126/science.aap7463.
11. Carroll, D.; Watson, B.; Togami, E.; Daszak, P.; Mazet, J.A.; Chrisman, C.J.; Rubin, E.M.; Wolfe, N.; Morel, C.M.; Gao, G.F.; others. Building a global atlas of zoonotic viruses. *Bulletin of the World Health Organization* **2018**, *96*, 292.
12. Santiago-Rodriguez, T.M.; Hollister, E.B. Unraveling the viral dark matter through viral metagenomics. *Frontiers in Immunology* **2022**, *13*.
13. Liang, Y.; Zheng, K.; McMinn, A.; Wang, M. Expanding diversity and ecological roles of RNA viruses. *Trends in Microbiology* **2023**, *31*, 229–232. doi:10.1016/j.tim.2022.12.004.
14. Edgar, R.C.; Taylor, J.; Lin, V.; Altman, T.; Barbera, P.; Meleshko, D.; Lohr, D.; Novakovsky, G.; Buchfink, B.; Al-Shayeb, B.; Banfield, J.F.; de la Peña, M.; Korobeynikov, A.; Chikhi, R.; Babaian, A. Petabase-scale sequence alignment catalyses viral discovery. *Nature* **2022**, *602*, 142–147. doi:10.1038/s41586-021-04332-2.
15. Rodriguez-Morales, A.J.; Balbin-Ramon, G.J.; Rabaan, A.A.; Sah, R.; Dhama, K.; Paniz-Mondolfi, A.; Pagliano, P.; Esposito, S. Genomic Epidemiology and its importance in the study of the COVID-19 pandemic. *Le Infezioni in Medicina* **2020**, *28*, 139–142.
16. Martin, J.; Klapsa, D.; Wilton, T.; Zambon, M.; Bentley, E.; Bujaki, E.; Fritzsche, M.; Mate, R.; Majumdar, M. Tracking SARS-CoV-2 in Sewage: Evidence of Changes in Virus Variant Predominance during COVID-19 Pandemic. *Viruses* **2020**, *12*, 1144. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute, doi:10.3390/v12101144.
17. Lin, Y.; Qian, Y.; Qi, X.; Shen, B., Databases, Knowledgebases, and Software Tools for Virus Informatics. In *Translational Informatics: Prevention and Treatment of Viral Infections*; Springer Nature Singapore: Singapore, 2022; pp. 1–19. doi:10.1007/978-981-16-8969-7_1.
18. Lin, D.; Crabtree, J.; Dillo, I.; Downs, R.R.; Edmunds, R.; Giaretta, D.; De Giusti, M.; L'Hours, H.; Hugo, W.; Jenkyns, R.; Khodiyar, V.; Martone, M.E.; Mokrane, M.; Navale, V.; Petters, J.; Sierman, B.; Sokolova, D.V.; Stockhause, M.; Westbrook, J. The TRUST Principles for digital repositories. *Scientific Data* **2020**, *7*, 144. Number: 1 Publisher: Nature Publishing Group, doi:10.1038/s41597-020-0486-7.

19. Wren, J.D.; Bateman, A. Databases, data tombs and dust in the wind. *Bioinformatics* **2008**, *24*, 2127–2128. doi:10.1093/bioinformatics/btn464.
20. Orchard, S.; Salwinski, L.; Kerrien, S.; Montecchi-Palazzi, L.; Oesterheld, M.; Stümpflen, V.; Ceol, A.; Chatr-Aryamontri, A.; Armstrong, J.; Woollard, P.; others. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nature biotechnology* **2007**, *25*, 894–898.
21. Roux, S.; Adriaenssens, E.M.; Dutilh, B.E.; Koonin, E.V.; Kropinski, A.M.; Krupovic, M.; Kuhn, J.H.; Lavigne, R.; Brister, J.R.; Varsani, A.; others. Minimum information about an uncultivated virus genome (MIUViG). *Nature biotechnology* **2019**, *37*, 29–37.
22. Sharma, D.; Priyadarshini, P.; Vrati, S. Unraveling the web of viroinformatics: computational tools and databases in virus research. *Journal of virology* **2015**, *89*, 1489–1501.
23. McLeod, K.; Upton, C. Virus Databases. *Reference Module in Biomedical Sciences* **2017**, pp. B978–0–12–801238–3.95728–3. doi:10.1016/B978-0-12-801238-3.95728-3.
24. Sansone, S.A.; McQuilton, P.; Rocca-Serra, P.; Gonzalez-Beltran, A.; Izzo, M.; Lister, A.L.; Thurston, M.; Community, F. FAIRsharing as a community approach to standards, repositories and policies. *Nature biotechnology* **2019**, *37*, 358–367.
25. Ma, L.; Zou, D.; Liu, L.; Shireen, H.; Abbasi, A.A.; Bateman, A.; Xiao, J.; Zhao, W.; Bao, Y.; Zhang, Z. Database Commons: A Catalog of Worldwide Biological Databases. *Genomics, Proteomics & Bioinformatics* **2022**.
26. Ison, J.; Rapacki, K.; Ménager, H.; Kalaš, M.; Rydza, E.; Chmura, P.; Anthon, C.; Beard, N.; Berka, K.; Bolser, D.; others. Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic acids research* **2016**, *44*, D38–D47.
27. Rigden, D.J.; Fernández, X.M. The 2023 Nucleic Acids Research Database Issue and the online molecular biology database collection. *Nucleic Acids Research* **2023**, *51*, D1–D8.
28. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; others. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **2016**, *3*, 1–9.
29. Bernasconi, A.; Canakoglu, A.; Masseroli, M.; Pinoli, P.; Ceri, S. A review on viral data sources and search systems for perspective mitigation of COVID-19. *Briefings in Bioinformatics* **2020**, p. bbaa359. doi:10.1093/bib/bbaa359.
30. Lefkowitz, E.J.; Dempsey, D.M.; Hendrickson, R.C.; Orton, R.J.; Siddell, S.G.; Smith, D.B. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic acids research* **2018**, *46*, D708–D717.
31. Walker, P.J.; Siddell, S.G.; Lefkowitz, E.J.; Mushegian, A.R.; Adriaenssens, E.M.; Dempsey, D.M.; Dutilh, B.E.; Harrach, B.; Harrison, R.L.; Hendrickson, R.C.; others. Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020), 2020.
32. Hulo, C.; De Castro, E.; Masson, P.; Bougueleret, L.; Bairoch, A.; Xenarios, I.; Le Mercier, P. ViralZone: a knowledge resource to understand virus diversity. *Nucleic acids research* **2011**, *39*, D576–D582.
33. Carrillo-Tripp, M.; Shepherd, C.M.; Borelli, I.A.; Venkataraman, S.; Lander, G.; Natarajan, P.; Johnson, J.E.; Brooks III, C.L.; Reddy, V.S. VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic acids research* **2009**, *37*, D436–D442.
34. Montiel-Garcia, D.; Santoyo-Rivera, N.; Ho, P.; Carrillo-Tripp, M.; Iii, C.L.B.; Johnson, J.E.; Reddy, V.S. VIPERdb v3. 0: a structure-based data analytics platform for viral capsids. *Nucleic Acids Research* **2021**, *49*, D809–D816.
35. Mihara, T.; Nishimura, Y.; Shimizu, Y.; Nishiyama, H.; Yoshikawa, G.; Uehara, H.; Hingamp, P.; Goto, S.; Ogata, H. Linking virus genomes with host taxonomy. *Viruses* **2016**, *8*, 66.
36. Olson, R.D.; Assaf, R.; Brettin, T.; Conrad, N.; Cucinell, C.; Davis, J.J.; Dempsey, D.M.; Dickerman, A.; Dietrich, E.M.; Kenyon, R.W.; others. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic acids research* **2023**, *51*, D678–D689.
37. Hatcher, E.L.; Zhdanov, S.A.; Bao, Y.; Blinkova, O.; Nawrocki, E.P.; Ostapchuck, Y.; Schäffer, A.A.; Brister, J.R. Virus Variation Resource–improved response to emergent viral outbreaks. *Nucleic acids research* **2017**, *45*, D482–D490.
38. Brister, J.R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI viral genomes resource. *Nucleic acids research* **2015**, *43*, D571–D577.

39. Goodacre, N.; Aljanahi, A.; Nandakumar, S.; Mikailov, M.; Khan, A.S. A reference viral database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *MSphere* **2018**, *3*, e00069–18.
40. Kudla, M.; Gutowska, K.; Synak, J.; Weber, M.; Bohnsack, K.S.; Lukasiak, P.; Villmann, T.; Blazewicz, J.; Szachniuk, M. Virxicon: a lexicon of viral sequences. *Bioinformatics* **2020**, *36*, 5507–5513.
41. Chen, L.; Liu, B.; Yang, J.; Jin, Q. DBatVir: the database of bat-associated viruses. *Database* **2014**, *2014*.
42. Chen, L.; Liu, B.; Wu, Z.; Jin, Q.; Yang, J. DRodVir: A resource for exploring the virome diversity in rodents. *Journal of Genetics and Genomics* **2017**, *44*, 259–264.
43. Zhou, S.; Liu, B.; Han, Y.; Wang, Y.; Chen, L.; Wu, Z.; Yang, J. ZOVER: the database of zoonotic and vector-borne viruses. *Nucleic Acids Research* **2022**, *50*, D943–D949.
44. Chen, I.M.A.; Chu, K.; Palaniappan, K.; Ratner, A.; Huang, J.; Huntemann, M.; Hajek, P.; Ritter, S.; Varghese, N.; Seshadri, R.; others. The IMG/M data management and analysis system v. 6.0: new tools and advanced capabilities. *Nucleic acids research* **2021**, *49*, D751–D763.
45. Camargo, A.P.; Nayfach, S.; Chen, I.M.A.; Palaniappan, K.; Ratner, A.; Chu, K.; Ritter, S.J.; Reddy, T.; Mukherjee, S.; Schulz, F.; others. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research* **2023**, *51*, D733–D743.
46. Tang, Z.; Fan, W.; Li, Q.; Wang, D.; Wen, M.; Wang, J.; Li, X.; Zhou, Y. MVIP: multi-omics portal of viral infection. *Nucleic Acids Research* **2022**, *50*, D817–D827.
47. Lamy-Besnier, Q.; Brancotte, B.; Ménager, H.; Debarbieux, L. Viral Host Range database, an online tool for recording, analyzing and disseminating virus–host interactions. *Bioinformatics* **2021**, *37*, 2798.
48. Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **2017**, *22*, 30494.
49. Elbe, S.; Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global challenges* **2017**, *1*, 33–46.
50. Khare, S.; Gurry, C.; Freitas, L.; Schultz, M.B.; Bach, G.; Diallo, A.; Akite, N.; Ho, J.; Lee, R.T.; Yeo, W.; others. GISAID's role in pandemic response. *China CDC Weekly* **2021**, *3*, 1049.
51. Harrison, P.W.; Lopez, R.; Rahman, N.; Allen, S.G.; Aslam, R.; Buso, N.; Cummins, C.; Fathy, Y.; Felix, E.; Glont, M.; others. The COVID-19 Data Portal: accelerating SARS-CoV-2 and COVID-19 research through rapid open access data sharing. *Nucleic acids research* **2021**, *49*, W619–W623.
52. Tzou, P.L.; Tao, K.; Pond, S.L.K.; Shafer, R.W. Coronavirus Resistance Database (CoV-RDB): SARS-CoV-2 susceptibility to monoclonal antibodies, convalescent plasma, and plasma from vaccinated persons. *PLoS One* **2022**, *17*, e0261045.
53. Kuiken, C.; Korber, B.; Shafer, R.W. HIV sequence databases. *AIDS reviews* **2003**, *5*, 52.
54. Kuiken, C.; Yoon, H.; Abfalterer, W.; Gaschen, B.; Lo, C.; Korber, B. Viral genome analysis and knowledge management. In *Data Mining for Systems Biology*; Springer, 2013; pp. 253–261.
55. Shafer, R.W. Rationale and uses of a public HIV drug-resistance database. *The Journal of infectious diseases* **2006**, *194*, S51–S58.
56. Rhee, S.Y.; Gonzales, M.J.; Kantor, R.; Betts, B.J.; Ravela, J.; Shafer, R.W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research* **2003**, *31*, 298–303.
57. Hayer, J.; Jadeau, F.; Deleage, G.; Kay, A.; Zoulim, F.; Combet, C. HBVdb: a knowledge database for Hepatitis B Virus. *Nucleic acids research* **2013**, *41*, D566–D570.
58. Van Doorslaer, K.; Li, Z.; Xirasagar, S.; Maes, P.; Kaminsky, D.; Liou, D.; Sun, Q.; Kaur, R.; Huyen, Y.; McBride, A.A. The Papillomavirus Episteme: a major update to the papillomavirus sequence database. *Nucleic acids research* **2017**, *45*, D499–D506.
59. Shao, W.; Shan, J.; Hu, W.S.; Halvas, E.K.; Mellors, J.W.; Coffin, J.M.; Kearney, M.F. HIV proviral sequence database: a new public database for near full-length HIV proviral sequences and their meta-analyses. *AIDS research and human retroviruses* **2020**, *36*, 1–3.
60. Siddell, S.G.; Smith, D.B.; Adriaenssens, E.; Alfenas-Zerbini, P.; Dutilh, B.E.; Garcia, M.L.; Junglen, S.; Krupovic, M.; Kuhn, J.H.; Lambert, A.J.; Lefkowitz, E.J.; Łobocka, M.; Mushegian, A.R.; Oksanen, H.M.; Robertson, D.L.; Rubino, L.; Sabanadzovic, S.; Simmonds, P.; Suzuki, N.; Van Doorslaer, K.; Vandamme, A.M.; Varsani, A.; Zerbini, F.M. Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV). *The Journal of General Virology* **2023**, *104*, 001840. doi:10.1099/jgv.0.001840.

61. Consortium, U. UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Research* **2023**, *51*, D523–D531.
62. Kalvari, I.; Nawrocki, E.P.; Ontiveros-Palacios, N.; Argasinska, J.; Lamkiewicz, K.; Marz, M.; Griffiths-Jones, S.; Toffano-Nioche, C.; Gautheret, D.; Weinberg, Z.; others. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* **2021**, *49*, D192–D200.
63. Paysan-Lafosse, T.; Blum, M.; Chuguransky, S.; Grego, T.; Pinto, B.L.; Salazar, G.A.; Bileschi, M.L.; Bork, P.; Bridge, A.; Colwell, L.; others. InterPro in 2022. *Nucleic Acids Research* **2023**, *51*, D418–D427.
64. Wheeler, D.L.; Barrett, T.; Benson, D.A.; Bryant, S.H.; Canese, K.; Chetvernin, V.; Church, D.M.; DiCuccio, M.; Edgar, R.; Federhen, S.; others. Database resources of the national center for biotechnology information. *Nucleic acids research* **2007**, *35*, D5–D12.
65. Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **2000**, *28*, 27–30.
66. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Science* **2019**, *28*, 1947–1951.
67. Kanehisa, M.; Furumichi, M.; Sato, Y.; Kawashima, M.; Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research* **2023**, *51*, D587–D592.
68. Kozomara, A.; Birgaoanu, M.; Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic acids research* **2019**, *47*, D155–D162.
69. Ibrahim, B.; McMahon, D.P.; Hufsky, F.; Beer, M.; Deng, L.; Mercier, P.L.; Palmarini, M.; Thiel, V.; Marz, M. A new era of virus bioinformatics. *Virus Research* **2018**, *251*, 86–90. doi:10.1016/j.virusres.2018.05.009.
70. Hufsky, F.; Abecasis, A.; Agudelo-Romero, P.; Bletsa, M.; Brown, K.; Claus, C.; Deinhardt-Emmer, S.; Deng, L.; Friedel, C.C.; Gismondi, M.I.; Kostaki, E.G.; Kühnert, D.; Kulkarni-Kale, U.; Metzner, K.J.; Meyer, I.M.; Miozzi, L.; Nishimura, L.; Paraskevopoulou, S.; Pérez-Cataluña, A.; Rahlff, J.; Thomson, E.; Tumescheit, C.; van der Hoek, L.; Van Espen, L.; Vandamme, A.M.; Zaheri, M.; Zuckerman, N.; Marz, M. Women in the European Virus Bioinformatics Center. *Viruses* **2022**, *14*, 1522. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, doi:10.3390/v14071522.
71. Bolduc, B.; Youens-Clark, K.; Roux, S.; Hurwitz, B.L.; Sullivan, M.B. iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *The ISME journal* **2017**, *11*, 7–14.
72. Bolduc, B.; Zablocki, O.; Guo, J.; Zayed, A.A.; Vik, D.; Dehal, P.; Wood-Charlson, E.M.; Arkin, A.; Merchant, N.; Pett-Ridge, J.; others. iVirus 2.0: Cyberinfrastructure-supported tools and data to power DNA virus ecology. *ISME Communications* **2021**, *1*, 77.
73. Conte, S.I.; Fina, F.; Psalios, M.; Ryal, S.; Lebl, T.; Clements, A. Integration of an Active Research Data System with a Data Repository to Streamline the Research Data Lifecyle: Pure-NOMAD Case Study. *International Journal of Digital Curation* **2017**, *12*, 210–219. Number: 2, doi:10.2218/ijdc.v12i2.570.
74. Field, D.; Sterk, P.; Kottmann, R.; De Smet, J.W.; Amaral-Zettler, L.; Cochrane, G.; Cole, J.R.; Davies, N.; Dawyndt, P.; Garrity, G.M.; Gilbert, J.A.; Glöckner, F.O.; Hirschman, L.; Klenk, H.P.; Knight, R.; Kyrpides, N.; Meyer, F.; Karsch-Mizrachi, I.; Morrison, N.; Robbins, R.; San Gil, I.; Sansone, S.; Schriml, L.; Tatusova, T.; Ussery, D.; Yilmaz, P.; White, O.; Wooley, J.; Caporaso, G. Genomic standards consortium projects. *Standards in Genomic Sciences* **2014**, *9*, 599–601. doi:10.4056/sigs.5559680.
75. Bernasconi, A.; Guizzardi, G.; Pastor, O.; Storey, V.C. Semantic interoperability: ontological unpacking of a viral conceptual model. *BMC Bioinformatics* **2022**, *23*, 491. doi:10.1186/s12859-022-05022-0.
76. García-López, R.; Pérez-Brocal, V.; Moya, A. Beyond cells–The virome in the human holobiont. *Microbial Cell* **2019**, *6*, 373.
77. Schoch, C.L.; Ciufo, S.; Domrachev, M.; Hotton, C.L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; others. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database* **2020**, *2020*.
78. Xu, B.; Chotewutmontri, S.; Wolf, S.; Klos, U.; Schmitz, M.; Dürst, M.; Schwarz, E. Multiplex identification of human papillomavirus 16 DNA integration sites in cervical carcinomas. *PloS one* **2013**, *8*, e66693.
79. Rasekhian, M.; Roohvand, F.; Habtemariam, S.; Marzbany, M.; Kazemimanesh, M. The Role of 3'UTR of RNA Viruses on mRNA Stability and Translation Enhancement. *Mini Reviews in Medicinal Chemistry* **2021**, *21*, 2389–2398. doi:10.2174/1389557521666210217092305.

80. Zerbini, F.M.; Siddell, S.G.; Mushegian, A.R.; Walker, P.J.; Lefkowitz, E.J.; Adriaenssens, E.M.; Alfenas-Zerbini, P.; Dutilh, B.E.; García, M.L.; Junglen, S.; others. Differentiating between viruses and virus species by writing their names correctly. *Archives of virology* **2022**, *167*, 1231–1234.
81. Dugan, V.G.; Emrich, S.J.; Giraldo-Calderón, G.I.; Harb, O.S.; Newman, R.M.; Pickett, B.E.; Schriml, L.M.; Stockwell, T.B.; Stoeckert Jr, C.J.; Sullivan, D.E.; others. Standardized metadata for human pathogen/vector genomic sequences. *PloS one* **2014**, *9*, e99979.
82. Turner, D.; Adriaenssens, E.M.; Tolstoy, I.; Kropinski, A.M. Phage Annotation Guide: Guidelines for Assembly and High-Quality Annotation. *PHAGE* **2021**, *2*, 170–182.
83. Roncoroni, M.; Droesbeke, B.; Eguinoa, I.; De Ruyck, K.; D'Anna, F.; Yusuf, D.; Grüning, B.; Backofen, R.; Coppens, F. A SARS-CoV-2 sequence submission tool for the European Nucleotide Archive. *Bioinformatics* **2021**, *37*, 3983–3985. doi:10.1093/bioinformatics/btab421.
84. Schäffer, A.A.; Hatcher, E.L.; Yankie, L.; Shonkwiler, L.; Brister, J.R.; Karsch-Mizrachi, I.; Nawrocki, E.P. VADR: validation and annotation of virus sequence submissions to GenBank. *BMC bioinformatics* **2020**, *21*, 211. doi:10.1186/s12859-020-3537-3.
85. Arroyo Mühr, L.S.; Lagheden, C.; Hassan, S.S.; Kleppe, S.N.; Hultin, E.; Dillner, J. De novo sequence assembly requires bioinformatic checking of chimeric sequences. *Plos one* **2020**, *15*, e0237455.
86. García-López, R.; Vázquez-Castellanos, J.F.; Moya, A. Fragmentation and coverage variation in viral metagenome assemblies, and their effect in diversity calculations. *Frontiers in bioengineering and biotechnology* **2015**, *3*, 141.
87. Orakov, A.; Fullam, A.; Coelho, L.P.; Khedkar, S.; Szklarczyk, D.; Mende, D.R.; Schmidt, T.S.; Bork, P. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome biology* **2021**, *22*, 1–19.
88. Sutton, T.D.; Clooney, A.G.; Ryan, F.J.; Ross, R.P.; Hill, C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **2019**, *7*, 1–15.
89. Salzberg, S.L.; Phillippy, A.M.; Zimin, A.; Puiu, D.; Magoc, T.; Koren, S.; Treangen, T.J.; Schatz, M.C.; Delcher, A.L.; Roberts, M.; others. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research* **2012**, *22*, 557–567.
90. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **1990**, *215*, 403–410.