International Journal of Advanced Smart Convergence Vol.8 No.3 115-122 (2019)
http://dx.doi.org/10.7236/IJASC.2019.8.3.115

IJASC 19-3-15

Data-processing pipeline and database design for integrated analysis of mycoviruses

Mikyung Je[1], Hyeon Seok Son[2], Hayeon Kim[3,*]

[1]SNU Bioinformatics Institute, Interdisciplinary Graduate Program in Bioinformatics,
College of Natural Science, Seoul National University, Seoul, Korea
[2]Laboratory of Computational Biology & Bioinformatics, Institute of Public Health and
Environment, Graduate School of Public Health, Seoul National University, Seoul, Korea
[3]Department of Biomedical Laboratory Science, Kyungdong University, Wonju, Gangwondo, Korea
e-mail: {[1]jemi57, [2]hss2003}@snu.ac.kr, [3,]hykim1984@kduniv.ac.kr*

Abstract

Recent and ongoing discoveries of mycoviruses with new properties demand the development of an appropriate research infrastructure to analyze their evolution and classification. In particular, the discovery of negative-sense single-stranded mycoviruses is worth noting in genome types in which double-stranded RNA virus and positive-sense single-stranded RNA virus were predominant. In addition, some genomic properties of mycoviruses are more interesting because they have been reported to have similarities with the pathogenic virus family that infects humans and animals. Genetic information on mycoviruses continues to accumulate in public repositories; however, these databases have some difficulty reflecting the latest taxonomic information and obtaining specialized data for mycoviruses. Therefore, in this study, we developed a bioinformatics-based pipeline to efficiently utilize this genetic information. We also designed a schema for data processing and database construction and an algorithm to keep taxonomic information of mycoviruses up to date. The pipeline and database (termed 'mycoVDB') presented in this study are expected to serve as useful foundations for improving the accuracy and efficiency of future research on mycoviruses.

Keywords: *Mycovirus, Database, Bioinformatics, Taxonomic classification*

1. INTRODUCTION

Recent studies for mycoviruses have focused on the discovery of mycoviruses with new properties. Mycoviruses generally have dsRNA and (+)ssRNA as genomes [1]; however, Sclerotinia sclerotimonavirus negative-stranded RNA virus 1 (SsNSRV-1) with (–)ssRNA was reported in 2014 [2]. Polymycoviruses have been reported to exhibit hyper-virulent effects unlike those typically presented by mycoviruses [3,4]. Properties associated with (–)ssRNA viruses, which include a large number of human and animal pathogenic

Manuscript Received: July. 25, 2019 / Revised: Aug. 1, 2019 / Accepted: Aug. 7, 2019
Corresponding Author: hykim1984@kduniv.ac.kr
Tel: +82-33-738-1374, Fax: +82-33-738-0243
Department of Biomedical Laboratory Science, Kyungdong University, Korea

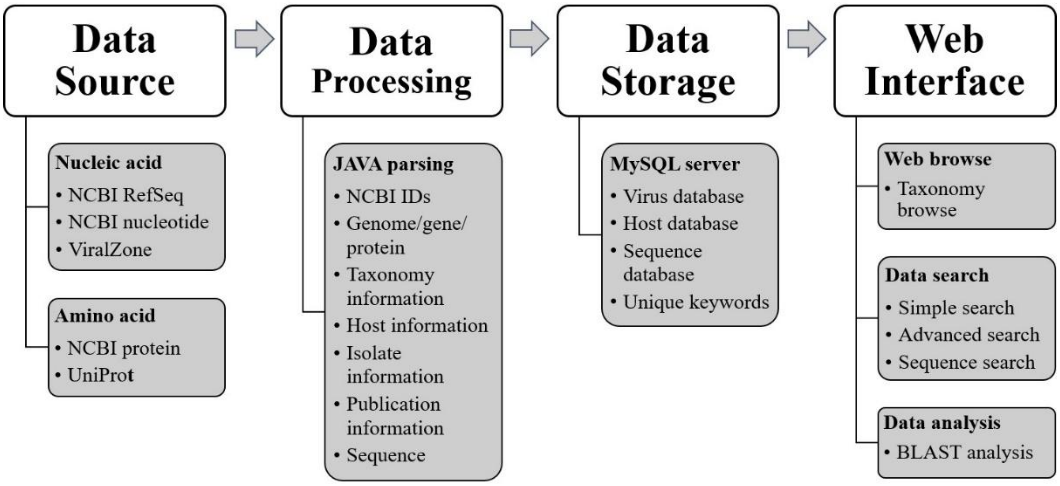


Figure 1. Bioinformatics pipeline for building mycovirus database

The mycovirus database includes taxonomic information, genetic sequence information and analysis tools for performing bioinformatics analysis. viruses, have also been observed in SsNSRV-1 and polymycoviruses [2,5]. These findings have highlighted the need for a new research infrastructure to analyze the evolution and classification of mycoviruses. The National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) is a public repository from which biological sequence data for mycoviruses can be downloaded [6]; alternatives include ViralZone, ViPR, and virusSITE. ViralZone (https://viralzone.expasy.org/) links virus-specific reference sequences to an external database, which in turn is linked to NCBI. RefSeq and UniProt store nucleic acid and protein sequences, respectively [7]. ViPR (https://www.viprbrc.org/) provides information about family Reoviridae mycoviruses [8], whereas virusSITE (http://www.virusite.org/index.php) provides detailed information about selected viruses accompanied by NCBI sequence data links [9]. Databases that focus on mycoviruses include DPVWeb and MVDB. DPVWeb (http://www.dpvweb.net/) contains data provided by GenBank and the European Molecular Biology Laboratory for viruses that infect plants, fungi and protozoa. The DPVWeb search function is designed to retrieve data by searching for 40 types of items; retrieved data can be downloaded in the FASTA format [10]. MVDB (http://mycovirusdb.com/) is a database built specifically for information on mycoviruses; it contains data on 336 mycoviruses, some of which are unclassified, and provides links to access the BLAST, CLUSTALW and MAFFT analytical tools. Data can be retrieved by inputting information such as the NCBI accession number, genome type, taxonomic information, sequence segments and sequence length (bp); sequence data can be downloaded in FASTA, CSV, Excel and PDF formats [11]. The genetic information of mycoviruses continues to accumulate in public repositories; however, only a few such databases are available, and these are insufficient to meet current data utilization needs for mycoviruses. As mycoviruses with new genetic characteristics are discovered, their taxonomic systems may be revised. However, the lack of flexibility among current repositories imposes a limit on their ability to adjust to new discoveries. Therefore, the objective of this study was to design a bioinformatics-based pipeline and database for analysis of mycoviruses genetic information and to confirm the latest taxonomy.

Table 1. Mycovirus taxonomy and NCBI datasets in 'mycoVDB'

Genome	Family	Genus	NCBI datasets	
			Reference	Complete
ssDNA	Genomoviridae	Gemycircularvirus	1	2
	Unclassfied	Unclassfied	1	1
dsRNA	Chrysoviridae	Alphchrysovirus	35	43
		Betachrysovirus	37	45
		Unclassified		12
	Megabirnaviridae	Megabirnavirus	6	8
	Partitiviridae	Alphapartitivirus	20	25
		Betapartitivirus	23	48
		Gammapartitivirus	18	30
		Unclassified	30	92
	Quadriviridae	Quadrivirus	4	9
	Reoviridae	Mycoreovirus	36	40
	Totiviridae	Giardiavirus	1	1
		Totivirus	10	17
		Victorivirus	26	34
		Unclassified	12	14
	Unclassfied	Botybirnavirus	6	11
		Unclassfied	66	75
(+)ssRNA	Alphaflexiviridae	Botrexvirus	1	1
		Sclerodarnavirus	1	1
	Barnaviridae	Barnavirus	0	1
	Deltaflexiviridae	Deltaflexivirus	3	3
	Endornaviridae	Alphaendornavirus	2	4
		Betaendornavirus	6	9
		Unclassified	8	9
	Gammaflexiviridae	Mycoflexivirus	1	1
	Hypoviridae	Hypovirus	11	28
		Unclassified	2	3
	Narnaviridae	Mitovirus	45	49
		Narnavirus	2	6
		Unclassified	3	3
	Unclassified	Unclassified	17	18
(-)ssRNA	Mymonaviridae	Sclerotimonavirus	2	2
	Unclassified	Unclassified	4	4
ssRNA-RT	Metaviridae	Metavirus	1	1
etc.	Satellites	-	2	2
	ssRNA	-	2	2
	Unknown	-	4	4

2. METHODS

The database for mycoviruses constructed in this study includes taxonomic information, genetic sequence information and tools for performing bioinformatics analyses (Figure 1). To build such a database, it is necessary to collect data from available public data repositories. Most mycovirus-related data are available from NCBI (Table 1), and additional data can be collected from ViralZone, ViPR and UniProt. We performed a parsing process to extract NCBI/UniProt identification numbers (IDs); genome/gene/protein names; information on taxonomy, host, isolate and publication; and sequence data using JavaScript. This parsing process was important because the efficiency of the database search function depends on the accuracy of data extraction and purification. Extracted data were stored using the MySQL server, and virus,

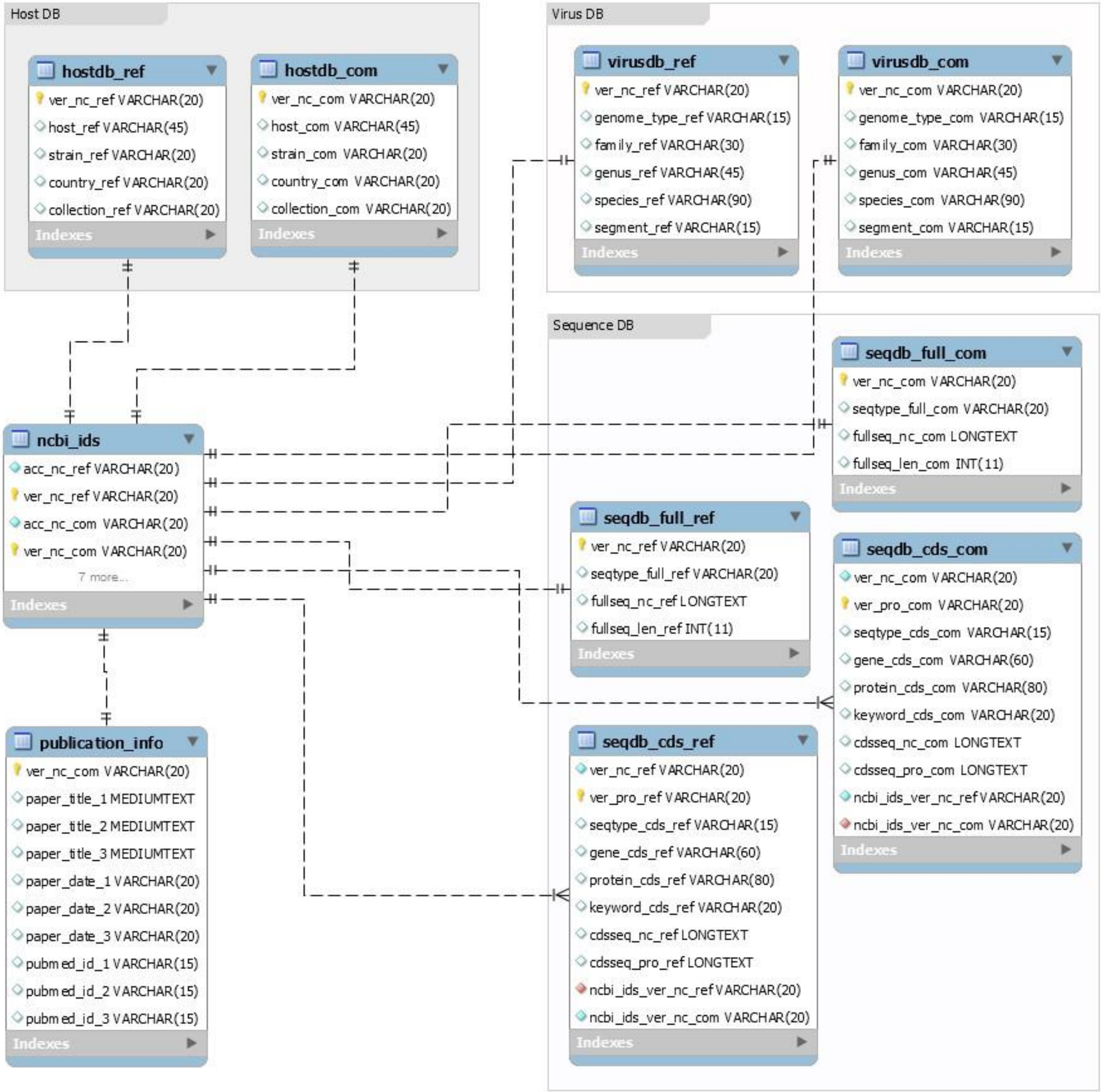
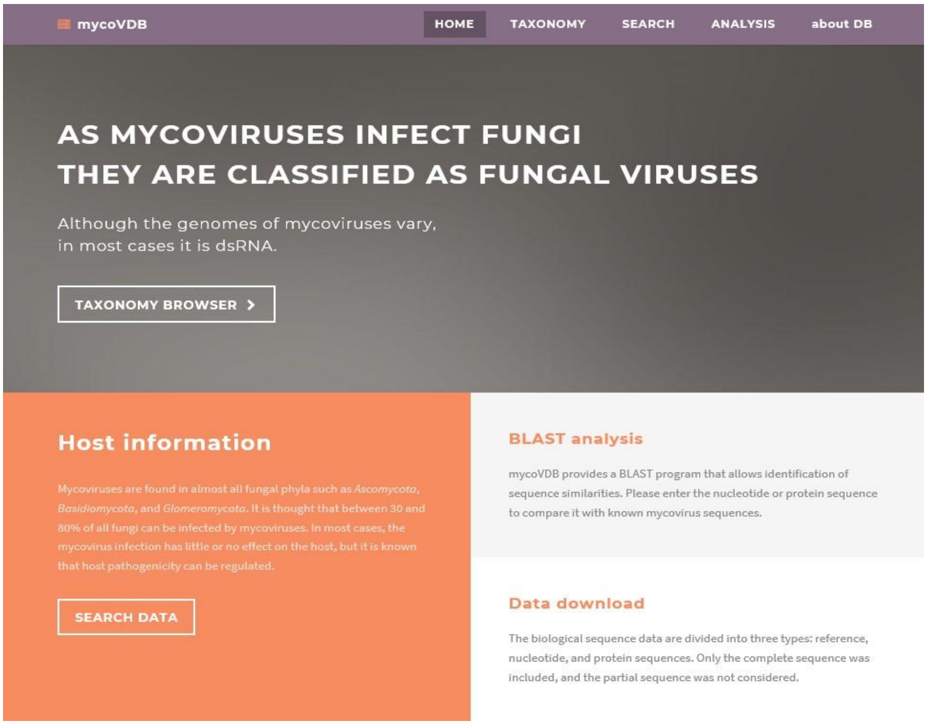


Figure 2. Database scheme for building 'mycoVDB'

This figure shows the database schema for building a mycovirus database. 'mycoVDB' comprises virus database, host database and sequence database; each type of database is linked to the other types.



AS MYCOVIRUSES INFECT FUNGI
THEY ARE CLASSIFIED AS FUNGAL VIRUSES

Although the genomes of mycoviruses vary, in most cases it is dsRNA.

Host information

Mycoviruses are found in almost all fungal phyla such as *Ascomycota*, *Basidiomycota*, and *Glomeromycota*. It is thought that between 30 and 80% of all fungi can be infected by mycoviruses. In most cases, the mycovirus infection has little or no effect on the host, but it is known that host pathogenicity can be regulated.

BLAST analysis

mycoVDB provides a BLAST program that allows identification of sequence similarities. Please enter the nucleotide or protein sequence to compare it with known mycovirus sequences.

Data download

The biological sequence data are divided into three types: reference, nucleotide, and protein sequences. Only the complete sequence was included, and the partial sequence was not considered.

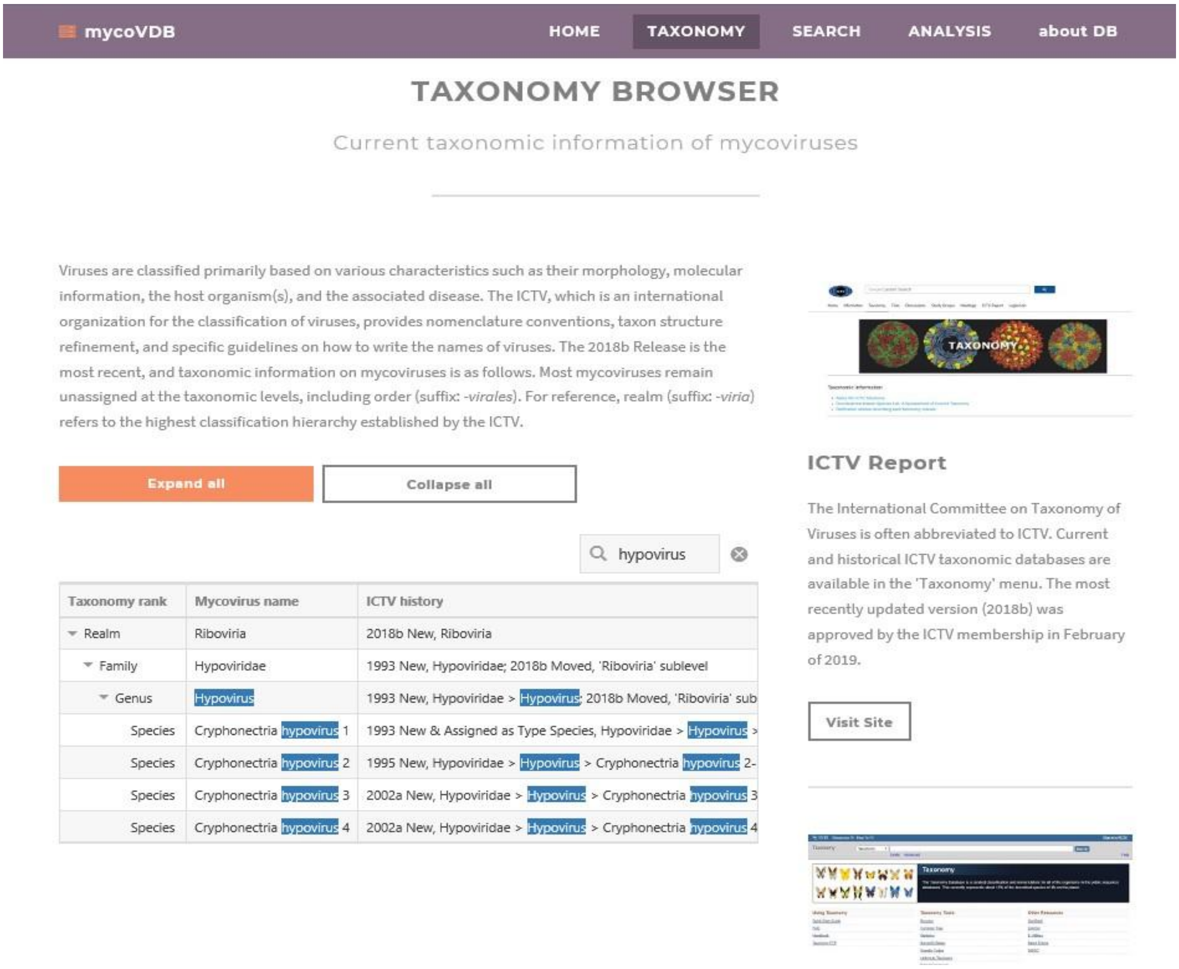
Figure 3. Web page in the 'mycoVDB'

'mycoVDB' is a MySQL-based database for mycoviruses; its main functions are mycovirus-related taxonomic information search, genetic information search and data analysis.
host and sequence databases were constructed using the collected information. The database was built using a web-based system using Hypertext Markup Language (HTML) and JavaScript. Also, the database supports various types of searches. To increase the number of search options, we designed the database and organized the data to be searched using genome type, mycovirus taxonomic information, viral genes/proteins, isolate, host and publication information. The database was constructed to include complete sequence data, allowing the user to download retrieved sequence files in the FASTA format.

3. RESULTS

3.1 mycoVDB

We designed a database schema for the construction of a MySQL-based database that stores collected data, as shown in the bioinformatics pipeline for integrated analysis of mycoviruses (Figures 1, 2). The constructed database was named 'mycoVDB' because it contains only mycovirus-related data. 'mycoVDB' consists of three databases (host, viral and sequence databases) for searching of mycovirus-related taxonomic and genetic information (Figure 3). The host database contains data on the host species, strain, country and collection date. The database contains taxonomic information (family, genus and species) related to the virus, including genome and segment data. The host and viral databases are each divided into two tables, one containing reference sequence information and the other complete sequence information. The sequence database contains information on the full and coding DNA sequence regions in mycoviruses and is organized into reference and complete sequence tables, respectively. Each database is associated with an NCBI ID and



TAXONOMY BROWSER

Current taxonomic information of mycoviruses

Viruses are classified primarily based on various characteristics such as their morphology, molecular information, the host organism(s), and the associated disease. The ICTV, which is an international organization for the classification of viruses, provides nomenclature conventions, taxon structure refinement, and specific guidelines on how to write the names of viruses. The 2018b Release is the most recent, and taxonomic information on mycoviruses is as follows. Most mycoviruses remain unassigned at the taxonomic levels, including order (suffix: *-virales*). For reference, realm (suffix: *-viria*) refers to the highest classification hierarchy established by the ICTV.

Taxonomy rank	Mycovirus name	ICTV history
Realm	Riboviria	2018b New, Riboviria
Family	Hypoviridae	1993 New, Hypoviridae; 2018b Moved, 'Riboviria' sublevel
Genus	Hypovirus	1993 New, Hypoviridae > Hypovirus; 2018b Moved, 'Riboviria' sub
Species	Cryphonectria hypovirus 1	1993 New & Assigned as Type Species, Hypoviridae > Hypovirus >
Species	Cryphonectria hypovirus 2	1995 New, Hypoviridae > Hypovirus > Cryphonectria hypovirus 2-
Species	Cryphonectria hypovirus 3	2002a New, Hypoviridae > Hypovirus > Cryphonectria hypovirus 3
Species	Cryphonectria hypovirus 4	2002a New, Hypoviridae > Hypovirus > Cryphonectria hypovirus 4

ICTV Report

The International Committee on Taxonomy of Viruses is often abbreviated to ICTV. Current and historical ICTV taxonomic databases are available in the 'Taxonomy' menu. The most recently updated version (2018b) was approved by the ICTV membership in February of 2019.

Figure 4. Taxonomy Browser

The 'Taxonomy Browser' provides taxonomic information for more than 100 species of mycoviruses. contains an additional table providing publication information associated with the GenBank accession ID. Thus, 'mycoVDB' is useful for systematic and detailed mycovirus searches.

3.2 Taxonomy browser

Unlike mycovirus-related databases, 'mycoVDB' contains a taxonomy browser that provides the latest phylogenetic information of mycoviruses (Figure 4). Mycoviruses can be searched by inputting a taxonomic classification system or virus name. When a viral species is selected, the user can review the complete history of the changes in phylogenetic classification recorded by the International Committee on Taxonomy of Viruses (ICTV, https://talk.ictvonline.org/) for the selected virus. Mycoviruses in the database include not only fungi, but also a variety of species that can infect other hosts, such as plants. With the discovery of mycoviruses with various genome types, it is important to understand all recent revisions to their taxonomic information. Therefore, we tried to provide taxonomic classification and nomenclature for mycoviruses based on ICTV reports through 'mycoVDB', and as a result, 'Taxonomy Browser' has the latest taxonomic data for mycoviruses revised by 2018. Therefore, we tried to provide taxonomic classification and nomenclature for mycoviruses based on ICTV reports through 'mycoVDB', and as a result, 'Taxonomy Browser' shows the latest taxonomic data for mycoviruses revised by 2018. Information on mycoviruses that have been reported to infect fungi but have not yet been officially classified is provided in the data search section, with reference to the related literature.

The 'Taxonomy Browser' has 'realm' as the top level of phylogenetic information on mycoviruses; most mycoviruses in the database are identified at the family, genus and species levels. When there are other taxonomic levels of phylogenetic information, the corresponding data can be additionally displayed. The 'Taxonomy Browser' in ‘mycoVDB’ contains taxonomic information of genus Rhizidiovirus (unassigned family, 1 species) in dsDNA viruses, and genus Gemycircularvirus (family Genomoviridae, 1 species) in ssDNA viruses. It also contains taxonomic information of dsRNA viruses in the following genera: Alphachrysovirus (12 species) and Betachrysovirus (8 species) in family Chrysoviridae, Megabirnavirus (1 species) in family Megabirnaviridae, Alphapartitivirus (10 species), Betapartitivirus (10 species), Gammapartitivirus (8 species) and an unassigned genus (3 species) in family Partitiviridae, Quadrivirus (1 species) in family Quadriviridae, Mycoreovirus (3 species) in family Reoviridae, genus Totivirus (7 species) and Victorivirus (14 species) in family Totiviridae, and genus Botybirnavirus (unassigned family, 1 species). Among (+)ssRNA viruses, it includes taxonomic information in the following genera: Botrexvirus (1 species) and Sclerodarnavirus (1 species) in family Alphaflexiviridae, Barnavirus (1 species) in family Barnaviridae, Deltaflexivirus (2 species) in family Deltaflexiviridae, Alphaendornavirus (3species) and Betaendornavirus (5 species) in family Endornaviridae, Mycoflexivirus (1 species) in family Gammaflexiviridae, Hypovirus (four species) in family Hypoviridae, and Mitovirus (5 species) and Narnavirus (2 species) in family Narnaviridae. And it includes taxonomic information of genus Sclerotimonavirus (family Mymonaviridae, 3 species) in (+)ssRNA viruses, and genus Metavirus (family Metaviridae, 5 species), genus Hemivirus and Pseudovirus (family Pseudoviridae, 6 species) in ssRNA-RT viruses.

4. DISCUSSION

In this study, we constructed the mycovirus database ‘mycoVDB’ based on a newly developed bioinformatics pipeline and database schema. ‘mycoVDB’ integrates genetic information on mycoviruses, facilitating data searching and collection; its search function allows users to retrieve taxonomic classification information that is unavailable in other mycovirus databases. We designed a taxonomy browser for ‘mycoVDB’ because clear evolutionary relationships determined by phylogenetic analyses are not available in other databases for all mycoviruses with genetic information. Some mycoviruses have not been classified at the genus or family level, and defined taxonomic classifications for mycoviruses have changed frequently over time. Therefore, we constructed ‘mycoVDB’ to improve the current understanding of mycovirus-related taxonomy and evolution by providing the latest taxonomic classification statuses and complete revision history. Recently discovered mycoviruses often have different characteristics from previously discovered mycoviruses; related studies have shown that mycovirus infection may weaken or enhance virulence among hosts such as fungi. For these reasons, more integrated and in-depth analyses of mycoviruses are required. In this study, we created a searchable, state-of-the-art database containing genetic sequence information to support such future bioinformatics analyses of mycoviruses by organizing data accumulated from the latest research on mycoviruses.

5. CONCLUSION

The development of molecular biology techniques has rapidly increased the amount of genetic data available on mycoviruses. This has facilitated more research on mycoviruses than in the past and improved our understanding of mycovirus evolution and interactions with fungal hosts. Mycoviruses have not yet been reported to have strong infectivity or severe pathogenicity in humans; however, they have genetic similarities to pathogenic zoonotic viruses, and the role of mycoviruses in controlling host virulence has been partially identified. Therefore, we anticipate a greater breadth and depth of research on evolutionary patterns, survival strategies and host ranges for mycoviruses, based on recently accumulated biological information. Bioinformatics methods can facilitate searches for similar sequences, sequence alignments and conserved motifs, as well as the construction of phylogenetic trees and evolutionary models. The first step toward producing significant results using these bioinformatics methods is to collect genetic information on the mycoviruses under study. For this purpose, we designed a bioinformatics pipeline for data collection and processing and a schema for constructing a new database, 'mycoVDB', which improves the convenience and efficiency of mycovirus research. Because hosts of mycoviruses can infect protozoa and plants, research using 'mycoVBD' may be useful for understanding a wide range of infections in addition to mycoviruses.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT and MOE) (No. 2016R1C1B2015511 and No. 2017R1D1A1B03033413).

REFERENCES

[1] M.N. Pearson, R.E. Beever, B. Boine, and K. Arthur, "Mycoviruses of filamentous fungi and their relevance to plant pathology," *Mol. Molecular plant pathology*, Vol. 10, No. 1, pp. 115-128, Jan 2009.

[2] L. Liu, J. Xie, J. Cheng, Y. Fu, G. Li, X. Yi, and D. Jiang, "Fungal negative-stranded RNA virus that is related to bornaviruses and nyaviruses,' *in Proc. of the National Academy of Sciences of the United States of America*, Vol. 111, No. 33, pp. 12205-12210, Aug 2014.

[3] S. Özkan and R.H. Coutts, "Aspergillus fumigatus mycovirus causes mild hypervirulent effect on pathogenicity when tested on Galleria mellonella," *Fungal genetics and biology*, Vol. 76, No. 76, pp. 20-26, Mar 2015.

[4] I. Kotta-Loizou and R.H. Coutts, "Studies on the virome of the entomopathogenic fungus Beauveria bassiana reveal novel dsRNA elements and mild hypervirulence," *PLoS pathogens*. Vol. 13, No. 1, pp. e1006183, Jan 2017.

[5] L. Kanhayuwa, I. Kotta-Loizou, S. Özkan, A.P. Gunning, and R.H. Coutts, "A novel mycovirus from Aspergillus fumigatus contains four unique dsRNAs as its genome and is infectious as dsRNA," *in Proc. of the National Academy of Sciences of the United States of America*, Vol. 112, No. 29, pp. 9100-9105, Jul 2015.

[6] NCBI Resource Coordinators, "Database Resources of the National Center for Biotechnology Information," *Nucleic acids research*, Vol. 45, No. D1, pp. D12-D17, Jan 2017.

[7] C. Hulo, E. de Castro, P. Masson, L. Bougueleret, A. Bairoch, I. Xenarios, and P. Le Mercier, "ViralZone: a knowledge resource to understand virus diversity," *Nucleic acids research*, Vol. 39, No. Database issue, pp. D576-D582, Jan 2011.

[8] B.E. Pickett, E.L. Sadat, Y. Zhang, J. Noronha, R.B Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu, L. Zhou, C.N. Larxon, J. Dietrich, E.B. Klem, and R.H. Scheuermann, "ViPR: an open bioinformatics database and analysis resource for virology research," *Nucleic acids research*, Vol. 40, No. Database issue, pp. D593-D598, Jan 2012.

[9] M. Stano, G. Beke, and L. Klucar, "viruSITE-integrated database for viral genomics," *Database: the journal of biological databases and curation*, Vol. 2016, pii. baw162, Dec 2016.

[10] M.J. Adams and J.F. Antoniw, "DPVweb: a comprehensive database of plant and fungal virus genes and genomes," *Nucleic acids research*, Vol. 34, No. Database issue, pp. D382-D385, Jan 2016.

[11] W. Shamsi, A. Jamal, N. Virk, and M.F. Bhatti, "The mycovirus database an e-bank for mycoviral genomes," *International Journal of Latest Trends in Engineering and Technology*, Vol. 12, No. 6, pp. 7-11, Nov 2018.