Database, 2021, **2021(2021)**, 1–5
DOI: https://doi.org/10.1093/database/baab079
Original article

iCAV: an integrative database of cancer-associated viruses

Bo Liu[1,#], Qingfeng Zhang[2,#], Jingou Wang[1], Shumin Cao[1], Zhiyuan Zhou[1], Ze-Xian Liu[2,*] and Han Cheng[1,*]

[1]School of Life Sciences, Zhengzhou University, Zhengzhou 450001, China
[2]State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

[*]Correspondence may also be addressed to Han Cheng. Tel: +86 0371 6778 0989; Fax: +86 0371 6778 0989; E-mail: chenghan@zzu.edu.cn and Ze-Xian Liu: Tel: +86 20 8734 2522; Fax: +86 20 8734 2522; E-mail: liuzx@sysucc.org.cn
[#]These authors contributed equally to this work.

Citation details: Liu, B., Zhang, Q., Wang, J. *et al.* iCAV: an integrative database of cancer-associated viruses. *Database* (2021) Vol. 2021: article ID baab079; DOI: https://doi.org/10.1093/database/baab079

Abstract

To date, various studies have found that the occurrence of cancer may be related to viral infections. Therefore, it is important to explore the relationship between viruses and diseases. The International Agency for Research on Cancer has defined six types of viruses as Class 1 human carcinogens, including Epstein–Barr virus, hepatitis C virus, hepatitis B virus, human T-cell lymphotropic virus, human herpesvirus 8 and human papillomavirus, while Merkel cell polyomavirus is classified as 'probably carcinogenic to humans' (Group 2A). Therefore, in-depth research on these viruses will help clarify their relationship with diseases, and substantial efforts have been made to sequence their genomes. However, there is no complete database documenting these cancer-associated viruses, and researchers are not able to easily access and retrieve the published genomes. In this study, we developed iCAV, a database that integrates the genomes of cancer-related viruses and the corresponding phenotypes. We collected a total of 18 649 genome sequences from seven human disease-related viruses, and each virus was further classified by the associated disease, sample and country. iCAV is a comprehensive resource of cancer-associated viruses that provides browse and download functions for viral genomes.

Database URL: http://icav.omicsbio.info/

Introduction

Since the early 1900s, various studies have reported the carcinogenic properties of retroviruses (1). In the 1960s, Sir Anthony Epstein, Bert Achong and Yvonne Barr identified the first human tumor virus in a cell culture of samples from pediatric Burkitt's lymphoma patients in Africa; that virus was named the Epstein–Barr virus (EBV) (2). Since then, evidence of the association between cancers and infections with certain viruses has been accumulating, and people have identified several cancer-associated viruses, including EBV, human papillomavirus (HPV), Kaposi's sarcoma-associated herpesvirus (KSHV; also known as human herpesvirus 8, HHV8), hepatitis C virus (HCV), hepatitis B virus (HBV), Merkel cell polyomavirus (MCV) and human T-cell lymphotropic virus (HTLV) (3). Infection with these viruses is the etiology of approximately 15% of all cancer cases worldwide (4). According to the assessment of the International Agency for Research on Cancer (IARC), HBV and HCV are indirect carcinogens that cause cancers by promoting a chronic inflammatory state, while HPV, MCV, EBV, HHV8 and HTLV are direct carcinogens (5).

HPV contains a double-stranded DNA (dsDNA) genome that is approximately 8 kbps in length (6, 7). It causes almost all cervical, anal, genital, head and neck cancers and 30% of oropharyngeal cancers (8). EBV is also a DNA virus that has a dsDNA genome that is 175 kbps in length, and nearly 95% of healthy adults have asymptomatic infections with EBV (9). The effects of EBV infection vary by geographic location, but it mainly causes nasopharyngeal cancer (10), posttransplant lymphoproliferative disorders (PTLDs) (11), Burkitt lymphoma (BL) (12) and Hodgkin lymphoma (13). HBV contains a partial dsDNA genome that is approximately 3.2 kbps in length (14). HCV has a single-stranded RNA genome that is approximately 9.6 kbps in length (15). HBV and HCV can cause hepatitis with variable degrees of damage, which, more seriously, can lead to cirrhosis and hepatocellular carcinoma (16). Moreover, studies have shown that HCV and HBV infections cause pancreatic cancer (17). MCV is a DNA virus that is nearly 5.4 kbps in length (18). MCV often causes a relatively harmless infection that persists lifelong, although it can also cause serious skin cancers and Merkle cell carcinoma (MCC) (19, 20). It has also been reported that the probability of MCC in AIDS patients is 10 times that in ordinary patients (21). HTLV is approximately 9 kb in length (22). The retrovirus human T-cell lymphotropic virus type 1 (HTLV-1) has infected 10–20 million people,

Received 14 July 2021; Revised 12 November 2021; Accepted 7 December 2021

although most of them are asymptomatic (23). Some infected patients develop highly aggressive malignancies, such as adult T-cell leukemia/lymphoma and HTLV-1-associated myelopathy/tropical spastic palsy (24, 25). KSHV, also known as HHV8, is a DNA virus that often causes Kaposi's sarcoma, which is a type of skin cancer (26). The entire genome of HHV8 is 14 kbps in length (27).

As research has progressed, the importance of viruses in the etiology of various cancers has become increasingly clear, and there are already several resources that collect and host relevant information on tumor viruses. For example, the NCBI Nucleotide database contains a large number of viral nucleotide sequences submitted by researchers (28). In addition, ViPR, which is a pathogenic virus database and analytical resource, contains more information about these viruses, including their sequences, genes, proteins, immune epitopes and so on, and provides some basic analytical tools, such as those for sequence alignment, phylogenetic inference and BLAST comparisons (29). However, there is no resource that is focused on cancer-associated viruses, and it is still difficult for researchers to obtain the reference genomes. Considering that the number of cancers caused by viral infections has increased dramatically, a complete database that could support research on the relationship between these viruses and diseases is urgently needed. Here, we introduce iCAV, which is an integrative database of cancer-associated viruses, with reference genomes and the related metadata for seven types of cancer-associated viruses. To ensure convenient usage of the database, all viruses are grouped by sample country and disease, and researchers can utilize the browsing functions to obtain the results of interest.

Materials and Methods

Data collection and processing

We searched for the nucleotide sequences of all seven viruses uploaded to the NCBI nucleotide database as of October 2020 using several carefully chosen keywords and then downloaded them (Table 1). To obtain the complete genome sequences, we first filtered the results by the range of genome length, which was defined as the known approximate length, and removed the sequences that only contain a portion of a genome (Table 1). Then, we extracted the relevant information for each virus, including the GenBank ID, definition, strain name, isolate name, geographic origin, sample type, and related phenotype. We also accessed the original study in which the virus sequences were published by searching for the PMID. For those genomes without PMIDs listed, we tried to obtain the relevant studies based on their reference titles. And then we searched the titles in PubMed database. At last, we took about 700 articles. All retrieved articles were carefully curated, the missing information, including countries, samples and phenotypes, were extracted from the original studies. The sample types were categorized by their source, such as plasma, serum, biopsy and cell line. We also carefully determined the country of the samples where the viruses were isolated. With regard to the phenotypes, we classified them into the specific disease or a healthy phenotype. Individuals who did not have any specific disease were defined as healthy (Figure 1). At last, detailed information, such as countries, samples and phenotypes, retrieved from NCBI nucleotide database and NCBI PubMed database were integrated, while the genomic sequences were also provided in iCAV (30).

Table 1. The keywords and length range of each virus

Virus	Keywords	Length
HTLV	Human T-cell lymphotropic virus OR HTLV OR Human T-lymphotropic virus	8000–10 000
HBV	Hepatitis B virus OR HBV	3000–3300
HCV	Hepacivirus C OR HCV	8900–10 000
MCV	Merkel cell polyomavirus OR MCV OR MCPyV	5000–5500
HPV	Human papillomavirus OR HPV	7000–10 000
HHV-8	Human gammaherpesvirus 8 KSHV Kaposi's sarcoma- associated herpesvirus	130 000–140 000
EBV	Human herpesvirus 4 OR Human gamma- herpesvirus 4 OR EBV OR Epstein–Barr virus	160 000–180 000

Construction of the website

The data we collected are stored in a MySQL database. The website was built using HTML, JavaScript and PHP, and several open source front-end libraries, such as jQuery and Bootstrap, were used to further modify the website (31). Then, the website was hosted on an Apache server. In addition, to ensure a smooth user experience, we tested the iCAV site on a variety of browsers, such as Google Chrome and Internet Explorer.

Results

Composition of the data in iCAV

In total, 18 649 reference genomes of seven types of viruses were collected, including 9213 HBV genome sequences, 6622 HPV genome sequences, 1366 HCV genome sequences, 1103 EBV genome sequences, 204 HTLV genome sequences, 75 HHV8 genome sequences and 66 MCV genome sequences (Figure 1). Our data were related to 87 phenotypes and 66 sample types, which originated from 143 countries worldwide, while the data types of each virus are presented in Figure 1, Tables 2 and 3. For example, in 9213 individuals infected with HBV, 3652 had chronic hepatitis B (CHB), 2575 were healthy, 462 had hepatitis, 175 had hepatocellular carcinoma (HCC) and 2349 had other diseases (Table 2). The sample types were 5410 serum samples, 119 blood samples, 772 plasma samples, 12 biopsy samples and 2892 other samples (Table 3). In 6622 individuals infected with HPV, 433 were healthy, 5973 had other disease, 196 had cervical cancer (CC), 19 had squamous-cell carcinoma (SCC) and 1 had genital cancer (GC) (Table 2). The sample types were 6372 other samples, 160 biopsies, 78 cervical swabs and 12 serum samples (Table 3).

Usage and presentation in iCAV

All data were classified by virus type, so it is convenient for users to access the corresponding records and relevant information for each type of virus (Figure 2A). We also listed the sample, country and disease on the left, which allows users to further filter the results (Figure 2B). After accessing the viruses

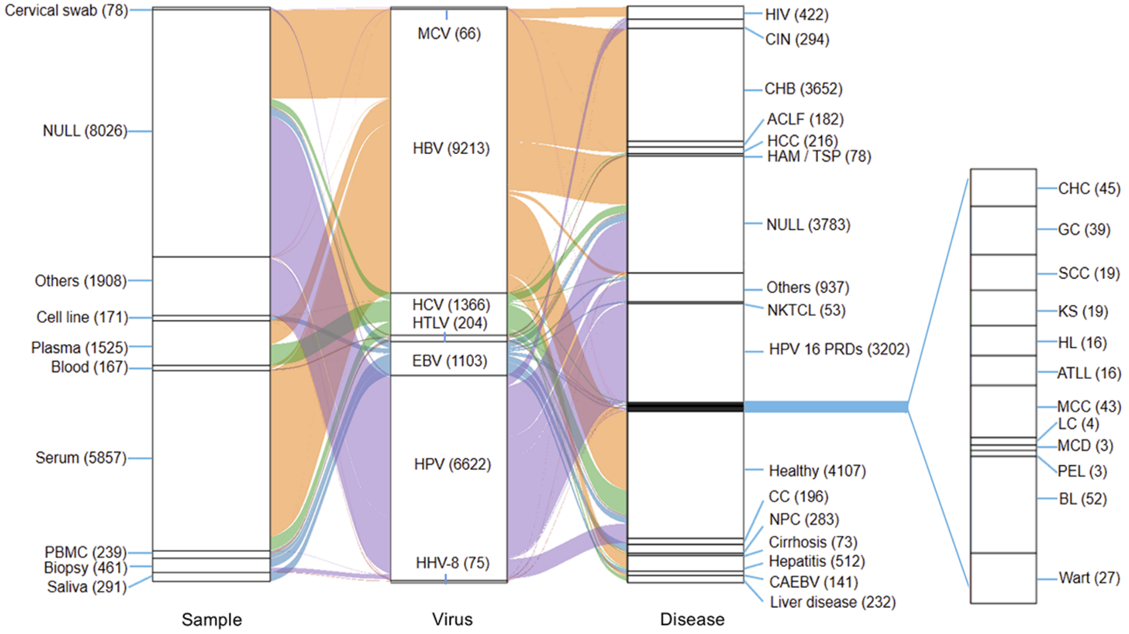


Figure 1. Relationships among samples, viruses and diseases. This figure shows mainly samples and diseases.

Table 2. The phenotypes of each virus

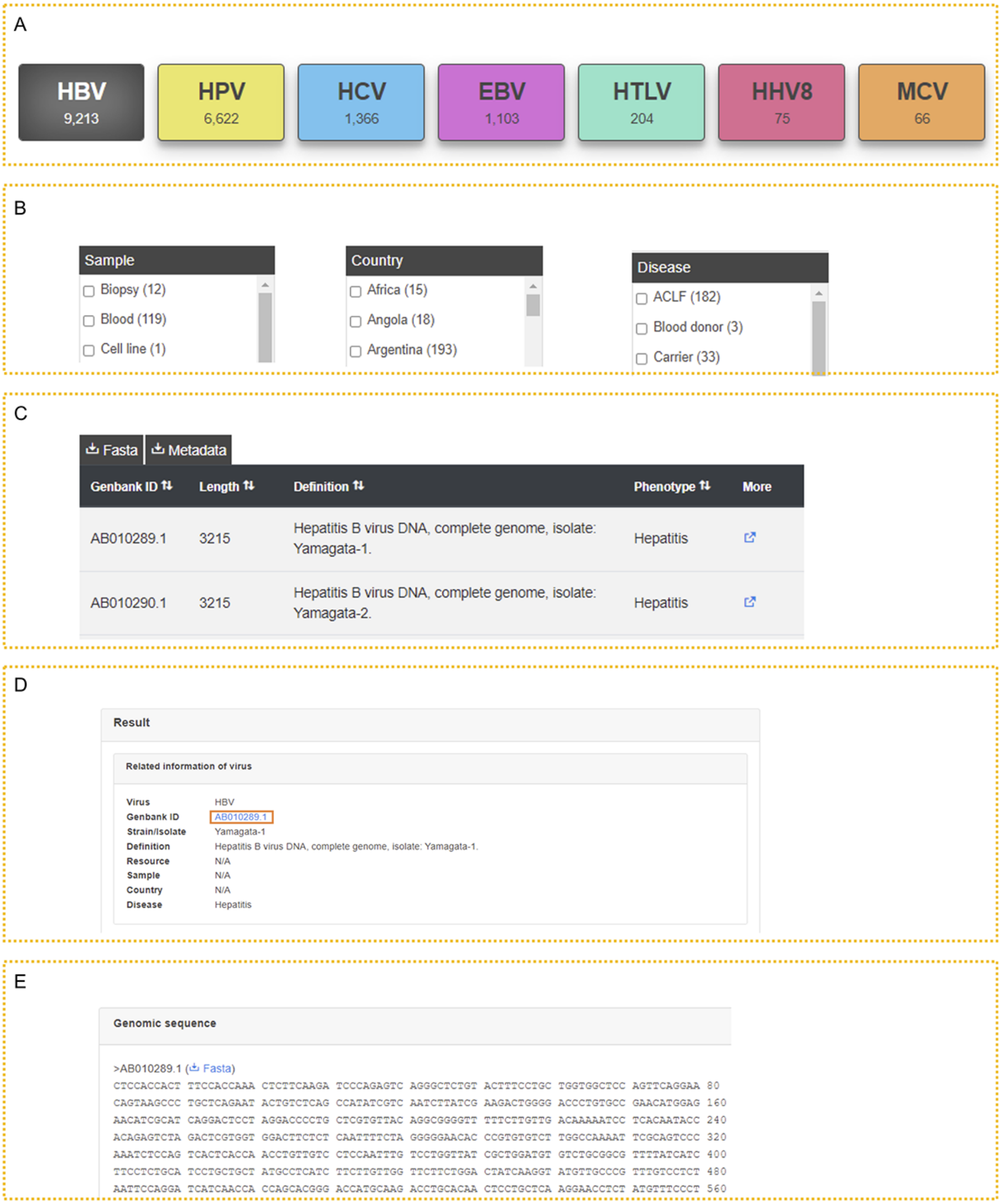
Virus	Phenotypes	Count
HBV	CHB	3652
	Health	2575
	Hepatitis	462
	HCC	175
	Other diseases	2349
HPV	CC	196
	SCC	19
	GC	1
	Health	433
	Other diseases	5973
HCV	Hepatitis	50
	Chronic hepatitis C (CHC)	45
	HCC	41
	Health	794
	Other diseases	436
EBV	Nasopharyngeal cancer	283
	Natural-killer/T cell lymphoma (NKTCL)	53
	GC	26
	Hodgkin lymphoma	16
	Lung cancer (LC)	4
	Health	213
HTLV	HTLV-1-associated myelopathy/tropical spastic palsy	78
	Health	59
	Adult T-cell leukemia-lymphoma (ATLL)	16
	Other diseases	51
HHV8	Kaposi's sarcoma (KS)	19
	Primary effusion lymphoma (PEL)	3
	Multicentric castleman disease (MCD)	3
	Other diseases	50
MCV	MCC	28
	Other diseases	5
	Health	33

Table 3. The samples of each virus

Virus	Samples	Count
HBV	Serum	5410
	Blood	119
	Plasma	772
	Biopsy	12
	Other samples	2892
HPV	Biopsy	160
	Other samples	6372
	Cervical swab	78
	Serum	12
HCV	Plasma	677
	Serum	434
	Cell lines	10
	Other samples	245
EBV	Saliva	291
	Other samples	371
	Biopsy	284
	Plasma	1
	Cell line	156
HTLV	Peripheral blood mononuclear cell (PBMC)	106
	Other samples	52
	Blood	45
	Plasma	1
HHV8	Other samples	56
	Biopsy	16
	Cell line	3
MCV	Other samples	54
	PBMC	5
	Biopsy	5
	Cell line	1
	Serum	1

of interest, users can download the genome sequences in FASTA format and the metadata (Figure 2C). Detailed information on the virus is provided if they click the 'More' link, including the GenBank ID, strain name, isolate name, definition, resource, sample, country and disease (Figure 2C–D).

The complete genome sequence is also displayed (Figure 2E), and users can download it separately. Moreover, users can obtain all the data for their further analysis in the download page. We guarantee that we will not record any information about our visitors, including IP, private information and browsing histories.



Sample
□ Biopsy (12)
□ Blood (119)
□ Cell line (1)

Country
□ Africa (15)
□ Angola (18)
□ Argentina (193)

Disease
□ ACLF (182)
□ Blood donor (3)
□ Carrier (33)

Fasta | Metadata

Genbank ID ⇅	Length ⇅	Definition ⇅	Phenotype ⇅	More
AB010289.1	3215	Hepatitis B virus DNA, complete genome, isolate: Yamagata-1.	Hepatitis	↗
AB010290.1	3215	Hepatitis B virus DNA, complete genome, isolate: Yamagata-2.	Hepatitis	↗

Result

Related information of virus

Virus	HBV
Genbank ID	AB010289.1
Strain/Isolate	Yamagata-1
Definition	Hepatitis B virus DNA, complete genome, isolate: Yamagata-1.
Resource	N/A
Sample	N/A
Country	N/A
Disease	Hepatitis

Genomic sequence

```
>AB010289.1 (Fasta)
CTCCACCACT TTCCACCAAA CTCTTCAAGA TCCCAGAGTC AGGGCTCTGT ACTTTCCTGC TGGTGGCTCC AGTTCAGGAA 80
CAGTAAGCCC TGCTCAGAAT ACTGTCTCAG CCATATCGTC AATCTTATCG AAGACTGGGG ACCCTGTGCC GAACATGGAG 160
AACATCGCAT CAGGACTCCT AGGACCCCTG CTCGTGTTAC AGGCGGGGTT TTTCTTGTTG ACAAAAATCC TCACAATACC 240
ACAGAGTCTA GACTCGTGGT GGACTTCTCT CAATTTTCTA GGGGGAACAC CCGTGTGTCT TGGCCAAAAT TCGCAGTCCC 320
AAATCTCCAG TCACTCACCA ACCTGTTGTC CTCCAATTTG TCCTGGTTAT CGCTGGATGT GTCTGCGGCG TTTTATCATC 400
TTCCTCTGCA TCCTGCTGCT ATGCCTCATC TTCTTGTTGG TTCTTCTGGA CTATCAAGGT ATGTTGCCCG TTTGTCCTCT 480
AATTCCAGGA TCATCAACCA CCAGCACGGG ACCATGCAAG ACCTGCACAA CTCCTGCTCA AGGAACCTCT ATGTTTCCCT 560
```

Figure 2. The usage of iCAV. (A–E) Five steps to use iCAV.

Discussion

These seven types of viruses, including EBV, HCV, HBV, HTLV, HHV8 and HPV, have been defined as Group 1 human carcinogens by the IARC, and MCV is classified as Group 2A (32). Among all cases of cancer caused by viral infections, the vast majority (>85%) occur in developing countries (33). Therefore, there is a growing urgency to study the relationship between viruses and cancers.

iCAV is the first database focused on cancer-associated viruses, and it provides users with the genome sequences and related phenotypes for seven human tumor viruses. By carefully collecting and integrating the metadata, iCAV can provide detailed information about the relationships between viruses and diseases that is easy for users to access. Our website provides a simple and straightforward interface for users to browse for the viruses of interest, and the results are clearly displayed and can be downloaded. In conclusion, iCAV is a convenient resource for researchers studying the relationships between virus genomic sequences and diseases and can improve research efficiency.

In this study, we collected the full and nearly full genome sequences of seven types of human cancer-associated viruses. The NCBI nucleotide database provides detailed sequence information, but the phenotypic information is missing. Therefore, we developed the iCAV database, which provides detailed phenotypic information to help us sort through large amounts of data to find the relevant information. Our goal is to integrate the available information about viruses and phenotypes. In the future, iCAV will be regularly maintained and updated every 2 years through surveying the lasted virus data of complete genome to provide more detailed and comprehensive information. We anticipate that the iCAV database will facilitate subsequent analyses by other researchers.

Funding

Natural Science Foundation of China (81972239 and 91953123 to Z.-X.L., 31601067 to H.C.), Fostering Fund of Fundamental Research for Young Teachers of Zhengzhou University (JC21343016 to H.C.), Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2017ZT07S096 to Z.-X.L.), Guangdong Esophageal Cancer Institute Science and Technology Program (M201911 to Z.-X.L.) and Tip-Top Scientific and Technical Innovative Youth Talents of Guangdong special support program (2019TQ05Y351 to Z.-X.L.).

Conflict of interest

The authors declare no conflict of interest.

References

1. Gaglia,M.M. and Munger,K. (2018) More than just oncogenes: mechanisms of tumorigenesis by human viruses. *Curr. Opin. Virol.*, **32**, 48–59.
2. Epstein,M.A., Achong,B.G. and Barr,Y.M. (1964) Virus particles in cultured lymphoblasts from Burkitt's lymphoma. *Lancet*, **1**, 702–703.
3. Moore,P.S. and Chang,Y. (2010) Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat. Rev. Cancer*, **10**, 878–889.
4. Martin,D. and Gutkind,J.S. (2008) Human tumor-associated viruses and new insights into the molecular mechanisms of cancer. *Oncogene*, **27**, S31–S42.
5. Chen,C.J., Hsu,W.L., Yang,H.I. *et al.* (2014) Epidemiology of virus infection and human cancer. *Recent Results Cancer Res.*, **193**, 11–32.
6. Graham,S.V. (2017) The human papillomavirus replication cycle, and its links to cancer progression: a comprehensive review. *Clin. Sci.*, **131**, 2201–2221.
7. Scheurer,M.E., Tortolero-Luna,G. and Adler-Storthz,K. (2005) Human papillomavirus infection: biology, epidemiology, and prevention. *Int. J. Gynecol. Cancer*, **15**, 727–746.
8. Szymonowicz,K.A. and Chen,J. (2020) Biological and clinical aspects of HPV-related cancers. *Cancer Biol. Med.*, **17**, 864–878.
9. Kanda,T., Yajima,M. and Ikuta,K. (2019) Epstein-Barr virus strain variation and cancer. *Cancer Sci.*, **110**, 1132–1139.
10. Teow,S.Y., Yap,H.Y. and Peh,S.C. (2017) Epstein-barr virus as a promising immunotherapeutic target for nasopharyngeal carcinoma treatment. *J. Pathog.*, **2017**,7349268.
11. Dharnidharka,V.R., Webster,A.C., Martinez,O.M. *et al.* (2016) Post-transplant lymphoproliferative disorders. *Nat. Rev. Dis. Primers*, **2**, 15088.
12. Rochford,R. and Moormann,A.M. (2015) Burkitt's lymphoma. *Curr. Top. Microbiol. Immunol.*, **390**, 267–285.
13. Farrell,K. and Jarrett,R.F. (2011) The molecular pathogenesis of Hodgkin lymphoma. *Histopathology*, **58**, 15–25.
14. Herrscher,C., Roingeard,P. and Blanchard,E. (2020) Hepatitis B virus entry into cells. *Cells*, **9**, 1486.
15. Chen,M., Ma,Y., Chen,H. *et al.* (2019) Complete genome sequencing and evolutionary analysis of HCV subtype 6xg from IDUs in Yunnan, China. *PLoS One*, **14**, e0217010.
16. Dandri,M. and Locarnini,S. (2012) New insight in the pathobiology of hepatitis B virus infection. *Gut*, **61**, i6–i17.
17. Fiorino,S., Cuppini,A., Castellani,G. *et al.* (2013) HBV- and HCV-related infections and risk of pancreatic cancer. *JOP J. Pancreas*, **14**, 603–609.
18. Shuda,M., Arora,R., Kwun,H.J. *et al.* (2009) Human Merkel cell polyomavirus infection I. MCV T antigen expression in Merkel cell carcinoma, lymphoid tissues and lymphoid tumors. *Int. J. Cancer*, **125**, 1243–1249.
19. DeCaprio,J.A. (2017) Merkel cell polyomavirus and Merkel cell carcinoma. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **372**, 1732.
20. Buck,C.B., Van Doorslaer,K., Peretti,A. *et al.* (2016) The ancient evolutionary history of polyomaviruses. *PLoS Pathog.*, **12**, e1005574.
21. Engels,E.A., Frisch,M., Goedert,J.J. *et al.* (2002) Merkel cell carcinoma and HIV infection. *Lancet*, **359**, 497–498.
22. Martinez,M.P., Al-Saleem,J. and Green,P.L. (2019) Comparative virology of HTLV-1 and HTLV-2. *Retrovirology*, **16**, 21.
23. Gessain,A. and Cassar,O. (2012) Epidemiological Aspects and World Distribution of HTLV-1 Infection. *Front Microbiol.*, **3**, 388.
24. Taylor,G.P. and Matsuoka,M. (2005) Natural history of adult T-cell leukemia/lymphoma and approaches to therapy. *Oncogene*, **24**, 6047–6057.
25. Matsuoka,M. and Jeang,K.T. (2007) Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation. *Nat. Rev. Cancer*, **7**, 270–280.
26. Chang,Y., Cesarman,E., Pessin,M.S. *et al.* (1994) Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science*, **266**, 1865–1869.
27. Olp,L.N., Jeanniard,A., Marimo,C. *et al.* (2015) Whole-genome sequencing of Kaposi's sarcoma-associated herpesvirus from Zambian Kaposi's sarcoma biopsy specimens reveals unique viral diversity. *J. Virol.*, **89**, 12299–12308.
28. Sayers,E.W., Cavanaugh,M., Clark,K. *et al.* (2019) GenBank. *Nucleic Acids Res.*, **47**, D94–D99.
29. Pickett,B.E., Sadat,E.L., Zhang,Y. *et al.* (2012) ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.*, **40**, D593–D598.
30. Ullah,S., Lin,S., Xu,Y. *et al.* (2016) dbPAF: an integrative database of protein phosphorylation in animals and fungi. *Sci. Rep.*, **6**, 23534.
31. Ullah,S., Ullah,A., Rahman,W. *et al.* (2021) An innovative user-friendly platform for Covid-19 pandemic databases and resources. *Comput. Methods Programs Biomed. Update*, **1**, 100031.
32. Hatano,Y., Ideta,T., Hirata,A. *et al.* (2021) Virus-driven carcinogenesis. *Cancers (Basel)*, **13**, 2625.
33. Schiller,J.T. and Lowy,D.R. (2014) Virus infection and human cancer: an overview. *Recent Results Cancer Res.*, **193**, 1–10.