Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Database article

dbEBV: A database of Epstein-Barr virus variants and their correlations with human health

Ruoqi Xie [a,1], Bijin Cao [b,1], Ze Wu [c,1], Yi Ouyang [a], Hui Chen [d], Weiwei Zhai [e], Ze-Xian Liu [a,*], Miao Xu [a,*], Guanghui Guo [f,*]

[a] State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, PR China
[b] School of Life Sciences, Zhengzhou University, Zhengzhou, China
[c] Shenzhen Longgang District Central Blood Station, Shenzhen 518172, China
[d] Human Genetics, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672, Singapore
[e] Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[f] Clinical Laboratory, The Third People's Hospital of Longgang District, Shenzhen 518115, China

ARTICLE INFO

Keywords:
EBV
Genetic variant
Disease
Global distribution
Phylogenetic analysis

ABSTRACT

Since Epstein-Barr virus (EBV) was discovered in 1964, it has been reported to be associated with various malignancies as well as benign diseases, and the pathogenicity of EBV has been widely studied. Several databases have been established to provide comprehensive information on the virus and its relation to diseases and introduce convenient analysis tools. Although they have greatly facilitated the analysis of EBV at the genome, gene, protein, or epitope level, they did not provide enough insight into the genomic variants of EBV, which have been suggested as relevant to diseases by multiple studies. Here, we introduce dbEBV, a comprehensive database of EBV genomic variation landscape, which contains 942 EBV genomes with 109,893 variants from different tissues or cell lines in 24 countries. The database enables the visualization of information with varying global frequencies and their relationship with the human health of each variant. It also supports phylogenetic analysis at the genome or gene level in subgroups of different characteristics. Information of interest can easily be reached with functions such as searching, browsing, and filtering. In conclusion, dbEBV is a convenient resource for exploring EBV genomic variants, freely available at http://dbebv.omicsbio.info.

1. Introduction

EBV was discovered by Anthony Epstein's team from a Burkitt lymphoma (BL) biopsy In 1964 [1]. As the first human virus to be implicated in cancers, it has been reported to be associated with nasopharyngeal carcinoma (NPC), gastric carcinoma (GC), several kinds of lymphomas, and also benign diseases such as infectious mononucleosis (IM) [2–6]. Despite this, more than 90% of the world's adult population are infected by the virus and asymptomatic life-long carriers of it [7].

Since the first EBV genome sequence B95–8 was published in 1984 [8], many other EBV strains have been sequenced successively, such as AG876, GD1, GD2, Akata, M81, and so on [9–15], which are divided into EBV type 1 and 2 according to the diverged alleles of EBV-determined nuclear antigen (EBNA) 2 and 3 [16–18]. EBV type 1 is predominantly distributed in most parts of the world and has greater transforming potential with a genome like B95–8. In contrast, EBV type 2 is equally prevalent to EBV type 1 in Central Africa, which resembles AG876 [19,20].

The heterogeneity of these EBV strains has been further explored, and multiple variants contributing to the pathogenicity of the virus have been clarified. EBNA2 of EBV type 2 is substantially different from that of EBV type 1 in both length and single nucleotide polymorphism, producing proteins with less B cell transformation potential [21,22]. A single amino acid difference in EBNA2 between EBV type 1 and 2 has been found to be related to B lymphoblastoid cell line growth maintenance [23]. Deletion of EBNA2 is commonly detected in several Burkitt lymphoma-derived cell lines, along with the expression of EBNA3 proteins and increased resistance to apoptosis [17,18]. Besides, truncated

* Corresponding authors.
E-mail addresses: liuzx@sysucc.org.cn (Z.-X. Liu), xumiao@sysucc.org.cn (M. Xu), guanghuig2017@163.com (G. Guo).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2024.04.043
Received 19 December 2023; Received in revised form 16 April 2024; Accepted 16 April 2024
Available online 17 April 2024

EBNA3B has been detected in B cell lines from EBV-positive human lymphoma samples with similar characteristics to tumors-derived lines induced by EBNA3B-knocked-out EBV, suggesting that defective EBNA3B may promote lymphomagenesis [24]. Another crucial gene involved in the pathogenicity of EBV is latent membrane protein (LMP) 1, and a natural variant CAO-LMP1 derived from NPC induces impaired LMP1-mediated upregulation of CD40 and CD54, which is attributed to sequences outside the CTAR-2 domain rather than a common 30-bp deletion within its C-terminal region [25]. Variation of miRNA regions also affects the development of EBV-associated diseases. Deletion of *Bam*HI-rightward transcript (BART) regions is observed in some kinds of lymphomas and accelerated tumor growth [26–28], while the high expression level of BART miRNA in NPC and EBVaGC suggests its potential role in carcinogenesis as well [26].

Although multiple studies have provided insight into critical genomic variants of EBV, most of them focused on only a few variant loci of their interest, and it is difficult for others to get the full variant profiles.

Significant efforts have been made to integrate a rapidly increasing number of EBV genome sequences and related information and comprehensively characterize the virus. GenBank incorporates DNA sequences from all available public sources, with most EBV genomic sequences uploaded to this database [29]. ViPR is a more professional database focusing on viruses, providing additional information, including genes, proteins, and immune epitopes, and fundamental analysis tools such as sequence alignment, phylogenetic inference, and BLAST comparison [30]. EBVdb is explicitly built for EBV-associated T cell immunology and vaccinology, with 2622 curated EBV antigenic proteins, 610 verified T cell epitopes, 26 verified human leukocyte antigen (HLA) ligands, and several computational tools for analysis, which facilitates data mining for EBV [31].

These databases have greatly facilitated the systematic analysis of EBV at the genome, gene, protein, or epitope level. However, they all failed to depict the geographical and pathological distribution of variants all across the EBV genome and among different EBV strains.

In this study, we developed dbEBV (http://dbebv.omicsbio.info), a comprehensive database of EBV genomic variation landscape, which contains 942 EBV genomes with 109,893 variants from different tissues or cell lines in 24 countries, as well as the global frequency and relationship with human health of each variant. The dbEBV also supports phylogenetic analysis at the genome or gene level in subgroups of different characteristics. The statistic results and variant profiles are visualized, supporting functions like searching, browsing, and filtering.

2. Materials and methods

2.1. Public EBV genome collecting

Until May 2019, we had carefully selected 559 public EBV genomes from GenBank with detailed information, including location, sample type, NPC incidence, EBV type, and host phenotype. The genomes were downloaded in FASTA format, and loci within repeat regions or represented with letters other than A, T, C, and G were considered poor quality and not included in subsequent variant calculating. Related information, including GenBank accession, strain or isolate name, genome definition, and genome reference name, was also extracted from GenBank. Moreover, each gene's coding sequence (CDS) regions in the reference genome B95–8 (NC_007605.1) were collected for further analysis.

2.2. Self-generated EBV genome sequencing

Three hundred eighty-three strains were isolated from the recruited participants, and 270 were included in a study we carried out before [15]. Our private cohort mainly involves patients in South China with EBV-related cancers, including NPC, BL, HL, NKTCL, and GC, as well as healthy subjects. The whole genome sequencing (WGS) was performed using the Illumina HiSeq 2000 platform. Then, the raw reads were mapped to the reference genome B95–8 through Burrows-Wheeler Aligner (BWA, version 0.7.5a) [32,33] after pre-processing and quality control. The variants were called following the GATK best practice workflows (version 3.2–2) after base and variant recalibration [33], and those with low depth ($<10\times$) were filtered out. Then, the variants were annotated with gene, variant function, and locus coverage. Besides, the related information was collected, including geographical origin (location), pathological origin (sample type), NPC incidence, EBV type, and host phenotype.

2.3. Multiple sequences alignment and variant calling

The FASTA sequences of self-generated EBV genomes were created with custom scripts, and loci with missing genotype information, being heterozygous, with the allele frequency of 40–60%, within repeat regions, or represented with a letter other than A, T, C, and G were considered poor quality and not included in subsequent analysis. The results were subsequently combined with the 559 public EBV genomes, and then rapid multiple sequence alignment was performed using MAFFT software tools (version 7) [34]. Considering most of the EBV genomes are assembled in accordance with B95–8, we defined the variants based on the difference between each aligned EBV genome and the reference genome B95–8, and the variant ID was determined as described in Fig. 1.

2.4. Detecting disease-associated variants

After carefully detecting the variant profiles of each EBV, the number and the proportion of variants in different phenotypes were made statistics, and Fisher's exact tests were performed to distinguish disease-associated variants between disease and healthy phenotype. FDR was calculated to reduce false positive results, with an acceptable cut-off value of 0.05. Moreover, we evaluated the OR of each variant to determine whether it was a risk or a protective factor. Considering the limits on the number of EBVs in each phenotype, we removed the diseases that contained less than 5 EBVs, and thus, LCL, LC, DLBCL, and nasopharyngitis were not taken into analysis.

In addition, the location, the sample type, the dataset of each EBV strain, and the combination of these factors were used to establish generalized linear models (GLM) to reduce the potential impacts of covariants at the association tests.

2.5. Distinguishing EBV types

EBV types 1 and 2 differ mainly in EBNA2 and EBNA3s [16,18]. B95–8 is a typical strain of EBV type 1, while AG875 (DQ279927.1) represents EBV type 2. Thus, to determine the EBV type of the strains we collected, we compared each to the strains B95–8 and AG876, respectively, and counted the nucleotide differences of CDS in ENBA2 (36216 to 37679) and ENBA3s (ENBA3A, 79955 to 80293; ENBA3B, 80382 to 82877; ENBA3A, 86517 to 89135). EBV strains containing an acceptable number of variants in EBNA2 (≤ 100) and EBNA3s (≤ 500) were defined as the corresponding EBV types.

2.6. Evolutionary tree building

According to the CDS region information of B95–8, segments of each gene were obtained from the results of multiple sequence alignment. After masking the repetitive regions and removing the poor coverage regions, we used Randomized Accelerated Maximum Likelihood (RAxML version 8) with a general time reversible (GTR) model to infer the maximum likelihood of the phylogenetic relationship [35]. After building the evolutionary trees at the whole genome level or a single gene, the subtrees were generated by deleting unnecessary branches and

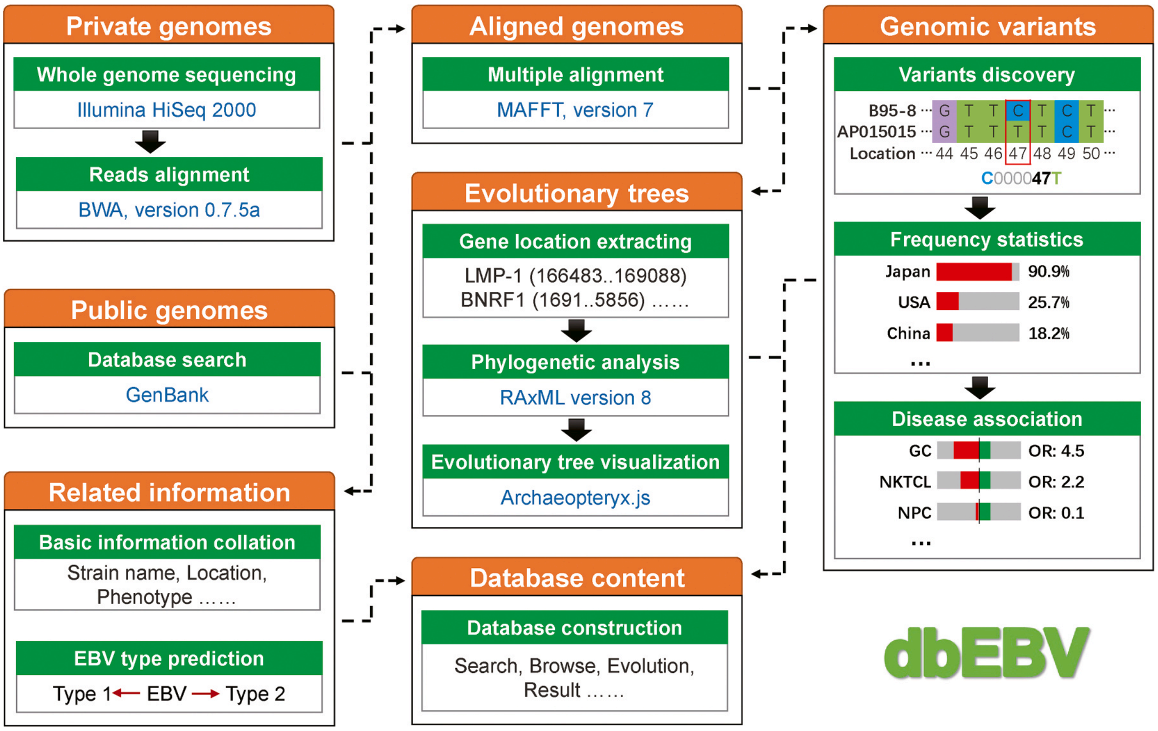

Fig. 1. Construction process of dbEBV.

nodes.

2.7. Implementation of a web server

The data of dbEBV were stored in a MySQL database. In the front end, the website's framework was built with HTML, CSS, and Bootstrap, while the interactive functions were implemented using JavaScript and JQuery. The interactive map of the global distribution of variant frequency was created with Leaflet, and the evolutionary trees were presented with Archaeopteryx.js. The visualization of statistical results and variant profiles were displayed by Echarts. In the back-end, PHP was used to receive and process the input from the front-end. To ensure its stability, we also tested the dbEBV website on various web browsers, such as Mozilla Firefox, Google Chrome, and Internet Explorer.

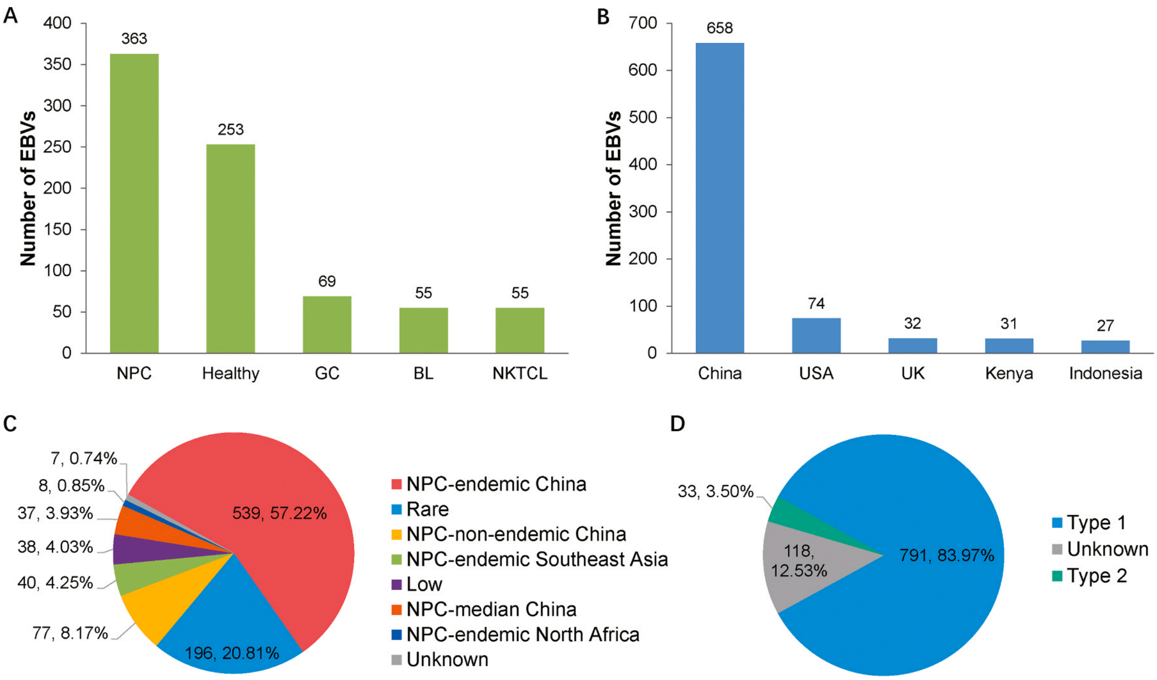

Fig. 2. Statistics of EBV characteristics. (A) The number of EBV strains in the top 5 phenotypes. (B) The number of EBV strains in the top 5 locations. (C) The statistics of NPC incidence. (D) The statistics of EBV type.

3. Results

3.1. Summary of dbEBV data

The database was built as described in Fig. 1. As an integration of public EBV genomes from GenBank and self-generated EBV sequencing data, dbEBV finally hosted 942 EBV genomes, covering 12 EBV-associated diseases and two lymphoblast cell lines, as well as healthy subjects. NPC, healthy subjects, GC, BL, and natural killer (NK) or T cell lymphoma (NKTCL) were the top five host phenotypes, accounting for 363, 253, 69, 55, and 55 EBV genomes, respectively (Fig. 2A). Others included Infectious mononucleosis (IM), post-transplant lymphoproliferative disease (PTLD), Hodgkin lymphoma (HL), chronic-active EBV infection (CAEBV), lung carcinoma (LC), nasopharyngitis and diffuse large B-cell lymphoma (DLBCL), as well as non-cancer spontaneous lymphoblast cell lines (sLCL) and non-cancer lymphoblast cell lines (LCL).

Although our data involved EBV genomes from 24 countries worldwide, most were obtained from China (658 of 942). The other four major countries from which we collected EBV genomes were the USA, UK, Kenya, and Indonesia, accounting for 74, 32, 31, and 27 EBV genomes, respectively (Fig. 2B). Considering NPC, the most closely associated malignancy of EBV, is rare in most regions of the world while relatively common in Southeast Asia and North Africa, especially Southern China [36], we further annotated whether they were from the NPC-endemic areas. As for the 658 EBV genomes from China, 539 were from NPC-endemic regions (57.22%), 37 were from the NPC-median areas (3.93%), and 77 EBV genomes were from NPC-non-endemic regions (8.17%) (Fig. 2C). Other EBV genomes were distributed in the NPC-endemic areas in southeast Asia (40, 4.25%), NPC-endemic regions in north Africa (8, 0.85%), and regions with rare or low NPC incidence out of China (196, 20.81%; 38, 4.03%) (Fig. 2C).

Among all EBV genomes, 791 EBV genomes were classified as type 1 (83.97%), while only 33 EBV genomes were of type 2 (3.50%), and the remaining 118 EBVs failed to be divided into either of the two types (12.53%) (Fig. 2D).

3.2. Variant characteristics and phylogenetic trees

All EBV genome sequences were aligned to the reference genome B95–8, and 238,339 variants were detected. The variant loci were further incorporated according to the principle of complementary base pairing. Six types of single nucleotide variants (SNVs) and two types of deletions were counted, with 12,215, 5714, 5736, 3684, 3558, 2401, 44,235 and 32,350 loci in C>T, C>A, T > C, C>G, T > G, T > A, C> - and T > - mutations, respectively (Fig. 3A). The number of SNVs in each EBV genome ranges from 30 to 3681, while the percentage of each SNV shows no significant difference (Fig. 3B). We found that most of the SNVs are C>T transition, accounting for over 30% of all SNVs. Additionally, the functions of disease-associated variants were analyzed, and the proportion of noncoding or synonymous variants is small (Fig. S1). Disease-associated variants were defined by Fisher's exact test, showing the specific genomic background of each disease. Interestingly, EBV genomes from NPC possess more variants with protective tendencies, while others contain more variants with a high disease risk (Fig. 3C). Further analyses considering co-variants were performed with GLMs, and the tendency still persists (Fig. S2). Disease-associated variants at the gene level were also analyzed using Fisher's exact test, with EBNA genes possessing the most variants in different diseases (Fig. S3). We then explored the evolutionary relationship of the EBV genomes through phylogenetic analysis at the whole genome and single gene level. In the evolutionary tree based on whole genomes, the EBV genomes are divided into two branches, which is consistent with the EBV types we predicted before (Fig. 3D).

3.3. Website function and result presentation

The dbEBV mainly provides three functions for users to explore the database and access information of interest, including search, browse, and evolutionary tree building. Users can query a specific EBV strain on the home page by entering its sample ID or strain name. Moreover, they can get the visualized distribution of each variant by entering its variant ID or find all variants of a specific locus by entering its chromosome location (Fig. 4A).

The result page of each EBV strain is accessed by searching the strain on the home page, which contains related information, classified information, and variant profile of the strain (Fig. 4B). The related information involves sample ID, strain name, GenBank accession, GenBank definition, and GenBank reference. The classified information consists of five characteristics, including location, phenotype, NPC incidence, EBV type, and sample type. The variant profile part presents mutation information, including single nucleotide variants and deletions.

After querying on the home page, users can further click one locus of the genome to see the global distribution of the particular variant

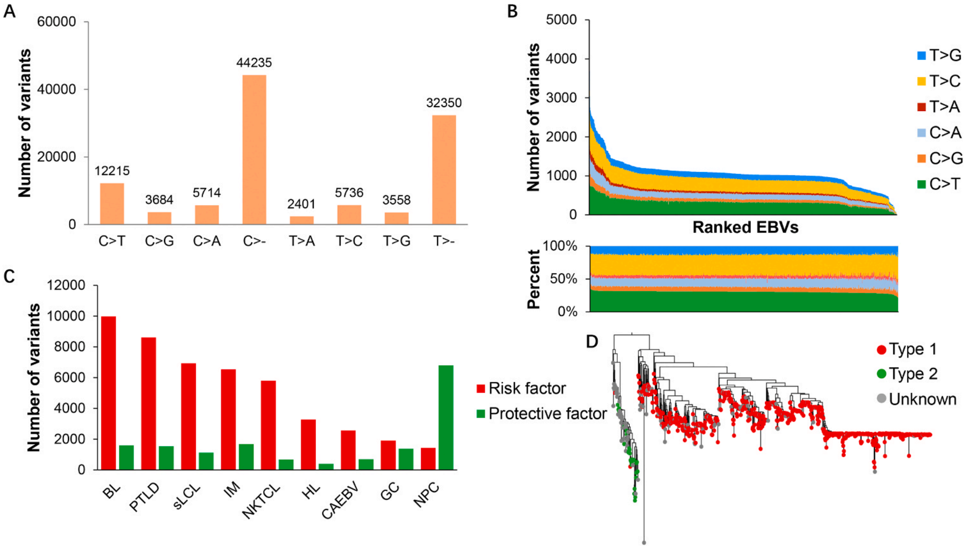

Fig. 3. Statistics of genomic variants. (A) The total number of variants at single nucleotide level (B) The number and proportion of variants in EBV strains. (C) The number of risk and protective variants in EBV-associated diseases. (D) The evolutionary tree of 942 EBV strains at whole genome level.

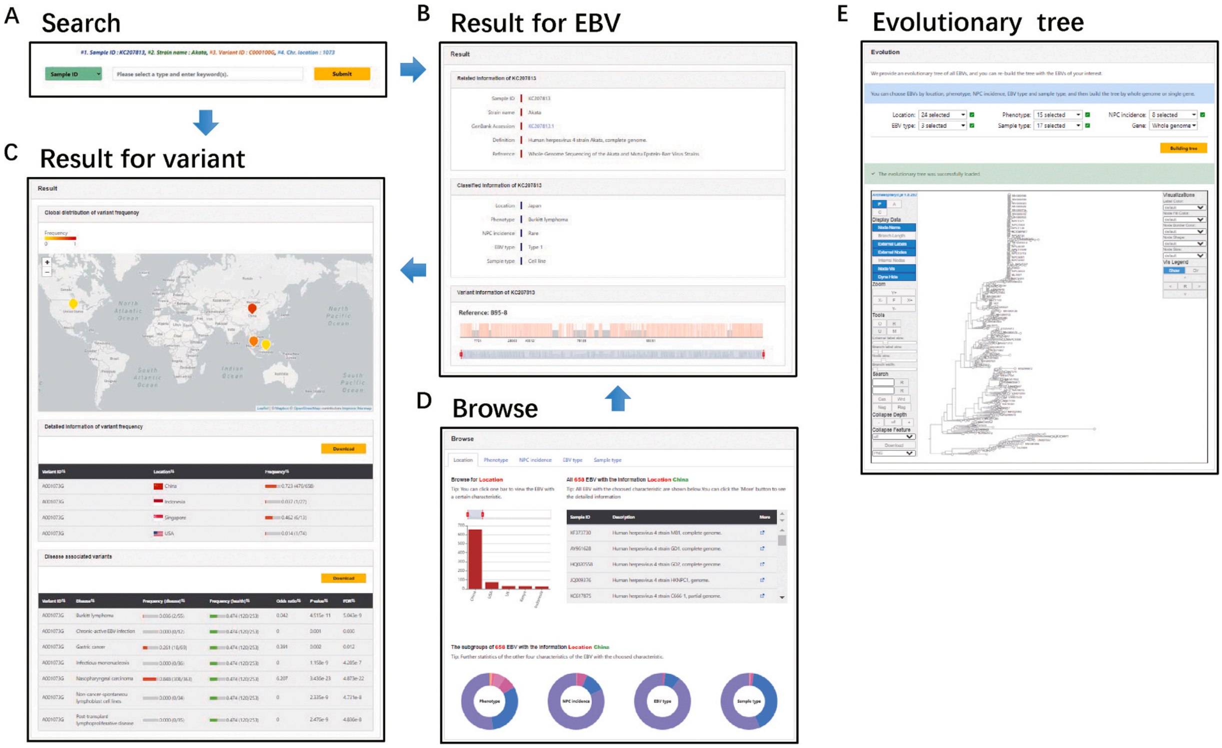

Fig. 4. Database presentation of dbEBV. (A) Search function of dbEBV. (B) Query result of an EBV strain. (C) Query result of a variant. (D) Browse function of dbEBV. (E) Evolutionary tree building function of dbEBV.

visualized along with its frequencies, and the detailed data are listed in the table below. (Fig. 4C). The association with diseases of the variant is further displayed at the bottom, with a drop-down selection box for co-variants selecting and sortable columns including variant ID, disease, frequency in disease, frequency in health, odds ratio (OR), p-value, false discovery rate (FDR), gene, variant function, and locus coverage. This page can also be reached from the home page using the variant ID mentioned above.

The browse page is divided into five subpages according to the five characteristics on the result page. On each subpage, the distribution of one characteristic of the EBV strains is displayed in a bar chart (Fig. 4D). Users can click one bar to see the matched EBV strains in a table beside the bar chart. They can also view detailed information on a particular strain on the result page mentioned above by clicking the 'More' button on the right side of the table. In addition, the distribution of the other four characteristics among the chosen EBV strains is displayed in the pie charts at the bottom of the subpage.

The evolution page allows users to build evolutionary trees with EBV strains filtered by the five characteristics on the result page (Fig. 4E). Additionally, users can make the tree at the whole-genome level, which is the default setting, or at the level of a particular gene. There are two control panels for modification of the tree. The left one allows users to change the style, labels, size, and direction of the image, as well as scaling and collapsing. Users can also search for particular nodes by entering a specific characteristic. The right panel facilitates visualization by providing options to annotate a particular characteristic with label color, node color, or node size. Users can also download the evolutionary tree they create in several file formats.

4. Discussion

The relationship between EBV and various diseases has been demonstrated for decades [2–6], and this relationship varies among different EBV strains [9–15], which may be partially explained by the genomic variants. Multiple studies have clarified the pathogenic mechanism of some variants by conducting *in vitro* or *in vivo* experiments [17,18,21–27]. However, there are also variants reported frequently appearing in certain diseases while not contributing a lot to pathogenesis, and the high frequency might reflect the incidence of the variants in the geographical location studied [25]. Interestingly, the C > T transition is the most abundant variant type in the EBV genome, which is consistent with the variant distribution in the human genome. One reason may be that transitions cause more minor changes in the shape of a DNA backbone, thus having less influence on gene expression [37], leading to smaller selection pressure. Another reason may be that methylated pyrimidine is more prone to deamination and becomes thymine, thus causing CG suppression [38], which helps viruses escape from the immune system of their hosts and survive. Besides, samples obtained from different tissues or with other approaches may also contain different variants [39]. Therefore, it's necessary to consider characteristics such as geographical locations and sample type and expand the sample size when exploring the influence of genomic variants on diseases.

As a comprehensive database of the genomic variation profile of EBV, dbEBV has carefully collected up to 942 available EBV genomes involving all diseases correlated with the virus as well as 253 of them derived from healthy populations as controls, and 109,893 variants have been detected at single nucleotide level with statistical test information. Besides, each variant's geographical origin and sample type have also been integrated. Furthermore, the database supports phylogenetic profiling, indicating the variants' mutual relation. With these efforts, dbEBV has dramatically facilitated the exploration of the potential significance of the variants in diseases, with disruption factors taken into consideration.

However, the database remains to be improved in some aspects. Although massive EBV genomes have been collected, their distribution among various characteristics is unbalanced, possibly leading to biased variant profiling. Thus, continual update of newly sequenced EBV genomes is needed. In addition, the genomic variation of EBV may also be affected by immune selection, suggesting that information such as HLA types of hosts and T-cell epitope should be integrated into the database. Overall, despite the improvement to be achieved, dbEBV is still a convenient and comprehensive resource for users to explore the genomic variation landscape.

5. Conclusion

The dbEBV depicts variant profiles of a large number of EBV strains integrated with comprehensive phenotypic information, which are visualized for convenient browsing and filtering. It also supports phylogenetic analysis with adjustable output formats. In conclusion, the database serves as a convenient resource for exploring EBV genomic variants.

Funding

This work was supported by National Key Research and Development Program of China (2022YFC2305400 to M.X.), National Natural Science Foundation of China (82122050 to M.X.), Young Talents Program of Sun Yat-sen University Cancer Center (YTP-SYSUCC-0029 to Z.-X.L.).

CRediT authorship contribution statement

Guanghui Guo: Conceptualization, Supervision, Writing – review & editing. **Ruoqi Xie:** Data curation, Formal analysis, Methodology, Visualization, Writing – original draft. **Yi Ouyang:** Data curation, Formal analysis, Methodology, Visualization, Writing – original draft. **Bijin Cao:** Data curation, Formal analysis, Methodology, Visualization, Writing – original draft. **Ze Wu:** Formal analysis, Methodology, Resources. **Hui Chen:** Formal analysis, Methodology, Resources. **Weiwei Zhai:** Formal analysis, Methodology, Resources. **Ze-Xian Liu:** Conceptualization, Supervision, Writing – review & editing. **Miao Xu:** Conceptualization, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The dataset is available on the Download page of dbEBV (http://dbebv.omicsbio.info).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.043.

References

[1] Epstein MA, Achong BG, Barr YM. Virus particles in cultured lymphoblasts from Burkitt's lymphoma. Lancet 1964;1(7335):702–3.

[2] Griffin BE. Relation of Burkitt's tumor-associated herpes-type virus to infectious mononucleosis. Rev Med Virol 1998;8(2):61–6.

[3] Fukayama M, Hayashi Y, Iwasaki Y, Chong J, Ooba T, et al. Epstein-Barr virus-associated gastric carcinoma and Epstein-Barr virus infection of the stomach. Lab Invest 1994;71(1):73–81.

[4] Hausen H, Schulte-Holthausen H, Klein G, Henle W, Henle G, et al. EBV DNA in biopsies of Burkitt tumours and anaplastic carcinomas of the nasopharynx. Nature 1970;228(5276):1056–8.

[5] Weiss LM, Strickler JG, Warnke RA, Purtilo DT, Sklar J. Epstein-Barr viral DNA in tissues of Hodgkin's disease. Am J Pathol 1987;129(1):86–91.

[6] Jones JF, Shurin S, Abramowsky C, Tubbs RR, Sciotto CG, et al. T-cell lymphomas containing Epstein-Barr viral DNA in patients with chronic Epstein-Barr virus infections. N Engl J Med 1988;318(12):733–41.

[7] Young LS, Yap LF, Murray PG. Epstein-Barr virus: more than 50 years old and still providing surprises. Nat Rev Cancer 2016;16(12):789–802.

[8] Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, et al. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 1984;310(5974): 207–11.

[9] Kanda T, Yajima M, Ahsan N, Tanaka M, Takada K. Production of high-titer Epstein-Barr virus recombinants derived from Akata cells by using a bacterial artificial chromosome system. J Virol 2004;78(13):7004–15.

[10] Zeng MS, Li DJ, Liu QL, Song LB, Li MZ, et al. Genomic sequence analysis of Epstein-Barr virus strain GD1 from a nasopharyngeal carcinoma patient. J Virol 2005;79(24):15323–30.

[11] Dolan A, Addison C, Gatherer D, Davison AJ, McGeoch DJ. The genome of Epstein-Barr virus type 2 strain AG876. Virology 2006;350(1):164–70.

[12] Liu P, Fang X, Feng Z, Guo YM, Peng RJ, et al. Direct sequencing and characterization of a clinical isolate of Epstein-Barr virus from nasopharyngeal carcinoma tissue by using next-generation sequencing technology. J Virol 2011;85 (21):11291–9.

[13] Tsai MH, Raykova A, Klinke O, Bernhardt K, Gärtner K, et al. Spontaneous lytic replication and epitheliotropism define an Epstein-Barr virus strain found in carcinomas. Cell Rep 2013;5(2):458–70.

[14] Kanda T, Yajima M, Ikuta K. Epstein-Barr virus strain variation and cancer. Cancer Sci 2019;110(4):1132–9.

[15] Xu M, Yao Y, Chen H, Zhang S, Cao SM, et al. Genome sequencing analysis identifies Epstein-Barr virus subtypes associated with high risk of nasopharyngeal carcinoma. Nat Genet 2019;51(7):1131–6.

[16] Sample J, Young L, Martin B, Chatman T, Kieff E, et al. Epstein-Barr virus types 1 and 2 differ in their EBNA-3A, EBNA-3B, and EBNA-3C genes. J Virol 1990;64(9): 4084–92.

[17] Kelly G, Bell A, Rickinson A. Epstein-Barr virus-associated Burkitt lymphomagenesis selects for downregulation of the nuclear antigen EBNA2. Nat Med 2002;8(10):1098–104.

[18] Kelly GL, Milner AE, Tierney RJ, Croom-Carter DS, Altmann M, et al. Epstein-Barr virus nuclear antigen 2 (EBNA2) gene deletion is consistently linked with EBNA3A, -3B, and -3C expression in Burkitt's lymphoma cells and with increased resistance to apoptosis. J Virol 2005;79(16):10709–17.

[19] Young LS, Yao QY, Rooney CM, Sculley TB, Moss DJ, et al. New type B isolates of Epstein-Barr virus from Burkitt's lymphoma and from normal individuals in endemic areas. J Gen Virol 1987;68(Pt 11):2853–62.

[20] Sixbey JW, Shirley P, Chesney PJ, Buntin DM, Resnick L. Detection of a second widespread strain of Epstein-Barr virus. Lancet 1989;2(8666):761–5.

[21] Dambaugh T, Hennessy K, Chamnankit L, Kieff E. U2 region of Epstein-Barr virus DNA may encode Epstein-Barr nuclear antigen 2. Proc Natl Acad Sci USA 1984;81 (23):7632–6.

[22] Palser AL, Grayson NE, White RE, Corton C, Correia S, et al. Genome diversity of Epstein-Barr virus from multiple tumor types and normal infection. J Virol 2015;89 (10):5222–37.

[23] Tzellos S, Correia PB, Karstegl CE, Cancian L, Cano-Flanagan J, et al. A single amino acid in EBNA-2 determines superior B lymphoblastoid cell line growth maintenance by Epstein-Barr virus type 1 EBNA-2. J Virol 2014;88(16):8743–53.

[24] White RE, Rämer PC, Naresh KN, Meixlsperger S, Pinaud L, et al. EBNA3B-deficient EBV promotes B cell lymphomagenesis in humanized mice and is found in human tumors. J Clin Invest 2012;122(4):1487–502.

[25] Johnson RJ, Stack M, Hazlewood SA, Jones M, Blackmore CG, et al. The 30-base-pair deletion in Chinese variants of the Epstein-Barr virus LMP1 gene is not the major effector of functional differences between variant LMP1 genes in human lymphocytes. J Virol 1998;72(5):4038–48.

[26] Qiu J, Cosmopoulos K, Pegtel M, Hopmans E, Murray P, et al. A novel persistence associated EBV miRNA expression profile is disrupted in neoplasia. PLoS Pathog 2011;7(8):e1002193.

[27] Lin X, Tsai MH, Shumilov A, Poirey R, Bannert H, et al. The Epstein-Barr virus BART miRNA cluster of the M81 strain modulates multiple functions in primary B cells. PLoS Pathog 2015;11(12):e1005344.

[28] Okuno Y, Murata T, Sato Y, Muramatsu H, Ito Y, et al. Defective Epstein-Barr virus in chronic active infection and haematological malignancy. Nat Microbiol 2019;4 (3):404–13.

[29] Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J, et al. GenBank. Nucleic Acids Res 2018;46(D1). p. D41-d47.

[30] Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 2012;40:D593–8 (Database issue).

[31] Zhang, G.L., L. Chitkushev, D.B. Keskin, E.L. Reinherz, and V. Bruisic. *EBVdb: a data mining system for knowledge discovery in Epstein-Barr virus with applications in T cell immunology and vaccinology*. in *2015 International Workshop on Artificial Immune Systems (AIS)*. 2015.

[32] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754–60.

[33] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43(5):491–8.

[34] Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30 (14):3059–66.

[35] Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22(21):2688–90.

[36] Chen YP, Chan ATC, Le QT, Blanchard P, Sun Y, et al. Nasopharyngeal carcinoma. Lancet 2019;394(10192):64–80.

[37] Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, et al. Transversions have larger regulatory effects than transitions. BMC Genom 2017;18(1):394.

[38] Takata MA, Goncalves-Carneiro D, Zang TM, Soll SJ, York A, et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 2017;550 (7674):124–7.

[39] Chang CM, Yu KJ, Mbulaiteye SM, Hildesheim A, Bhatia K. The extent of genetic diversity of Epstein-Barr virus and its geographic and disease patterns: a need for reappraisal. Virus Res 2009;143(2):209–21.