*Nucleic Acids Research*, 2025, gkaf1106
https://doi.org/10.1093/nar/gkaf1106
**Database issue**

# ViMIC 2.0: an updated database of human disease-related viral mutations, integration sites, and multi-omics data

**Chenjun Huang[1,†], Honglian Huang[2,†], Min Ding[3,†], Jiawen Zhu[1], Xin Qin[1], Zeyuan Zhang[1], Xiaoyang Zhao[4], Ziyi Wei[2], Min Wang[5], M. James C. Crabbe[6,7,8], Xiaoyan Zhang[2,*], Ying Wang[1,*]**

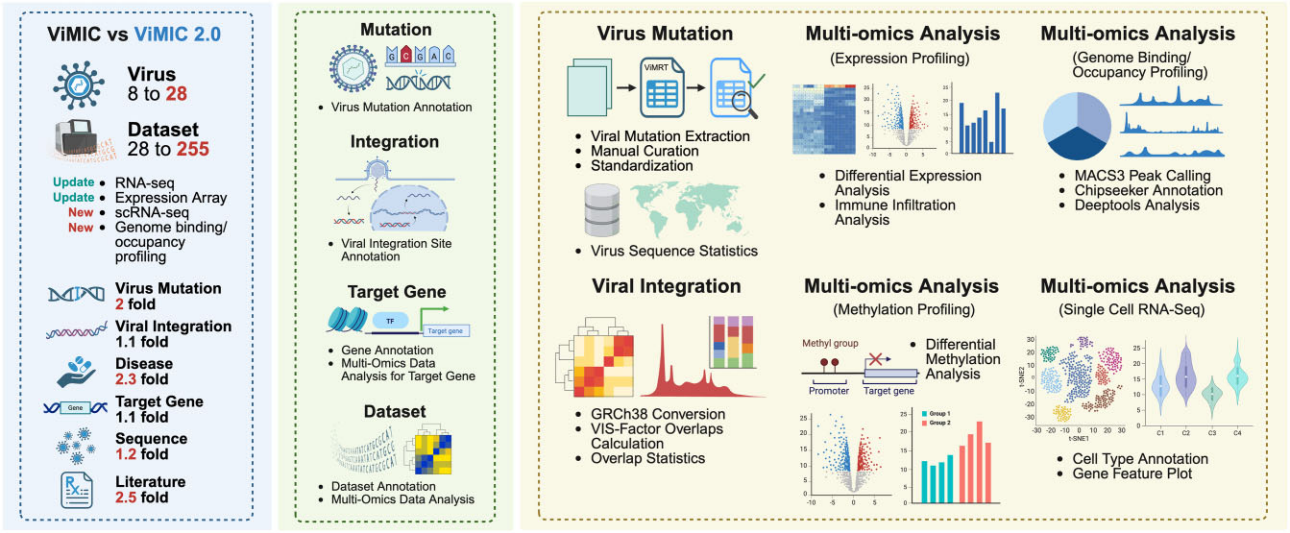
[1]Department of Clinical Laboratory Medicine Center, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai 200437, China
[2]School of Life Sciences and Technology, Tongji University, Shanghai 200092, China
[3]Department of Interventional Oncology, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200127, China
[4]The First Dispatched Outpatient Department, 905th Hospital of PLA Navy, Shanghai 200052, China
[5]Department of Laboratory Medicine, Eastern Hepatobiliary Surgery Hospital, Shanghai 200438, China
[6]Wolfson College, Oxford University, Oxford OX12JD, United Kingdom
[7]Institute of Biomedical and Environmental Science & Technology, University of Bedfordshire, Luton LU13JU, United Kingdom
[8]School of Life Sciences, Shanxi University, Taiyuan 030006, China

*To whom correspondence should be addressed. Email: nadger_wang@139.com
Correspondence may also be addressed to Xiaoyan Zhang. Email: xyzhang@tongji.edu.cn
†The first three authors should be regarded as Joint First Authors.

## Abstract

ViMIC 2.0 is an updated database that provides comprehensively curated data on virus mutations (VMs), viral integration sites (VISs), and multi-omics datasets related to human diseases. Leveraging expanding public data, ViMIC 2.0 significantly enhanced data scale, diversity, and analytical capabilities compared to the previous version. In terms of data volume, the number of virus types has increased from 8 to 28, VM entries have grown from 31 712 to 64 168, virus-related diseases expanded from 77 to 177, literature rose from 2539 to 6433, and omics datasets have substantially increased from 28 sets of single expression profile data to 255 sets of multi-omics data. In addition, ViMIC 2.0 has updated 9409 VISs, 173 048 sequences, newly incorporated sequencing types such as single-cell transcriptomic sequencing (scRNA-seq), and genome binding/occupancy profiling. Regarding the visualization module, ViMIC 2.0 now provides results of differential gene expression analysis for bulk RNA-seq or array, cell type annotation and gene feature plot for scRNA-seq data, and differential methylation analysis for methylation profiling, as well as peak annotation for ChIP-seq/ChIP-on-chip/ATAC-seq data. In summary, ViMIC 2.0 serves as a user-friendly, up-to-date, and well-maintained resource for the virology research community. ViMIC 2.0 is freely accessible at http://www.biomedinfo.cn/ViMIC2.0/index.php.

## Graphical abstract




**Received:** September 10, 2025. **Revised:** October 1, 2025. **Accepted:** October 1, 2025

## Introduction

Viral infection has become one of the most severe challenges in the field of global public health. From the continuously evolving Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1] and recurrent influenza virus (IV) [2] to oncogenic viruses such as human papillomavirus (HPV) [3], hepatitis B virus (HBV) [4], and Epstein–Barr virus (EBV) [5], the outbreaks and epidemics of numerous viral diseases continue to test the response capabilities of the public health systems worldwide. The complexity of viral pathogenic mechanisms further complicates prevention and control efforts. Viruses interact with their hosts through multilayered regulatory mechanisms. Viral proteins or nucleic acid components can target host transcriptional regulatory system. For instance, HPV-16 E6 protein degrades the tumor suppressor p53 and significantly upregulates CIP2A expression. This promotes cell proliferation through G1 checkpoint disruption, thereby mediating tumorigenesis [6]. Integrating viruses such as HBV, HPV, and EBV may exert *cis*-effects by inserting their genomes into human chromosomes [7]. Non-integrating viruses also can broadly and persistently influence the host gene transcription through epigenetic mechanisms, such as DNA methylation and histone modifications (HMs), or by activating specific signaling pathways. For example, after IV infection, genetic variations including mobile element insertions can alter HMs (H3K27ac, H3K4me1) and chromatin accessibility, thereby regulating the expression of immune genes such as TRIM25 [8]. Furthermore, viral mutations may lead to enhancing virulence or drug resistance [9, 10]. Together, these mechanisms enable viruses to effectively evade immune recognition and clearance, establish persistent infection, and ultimately contribute to chronic diseases or even malignant tumors.

Although technological advances have enabled the generation of viral data at an unprecedented scale and speed, major challenges remain. The vast amounts of viral data are scattered across disparate platforms and databases, lacking standardized organization and systematic analysis. The information fragmentation severely hinders a deeper understanding of viral pathogenesis and constrains the development of effective prevention and treatment strategies. Therefore, it is particularly important to establish a comprehensive database capable of integrating multi-source data and covering diverse virus types, to help researchers to decipher the complex regulatory networks of virus–host interactions, identify key pathogenic factors and biomarkers, and provide critical data support for vaccine design, drug development, and public precision health.

Currently, international virus databases include databases focusing on the integration site information of viruses in the host genome (e.g. VISDB [11]); specialized databases for individual virus species, such as 2019nCoVR [12] and HBVdb [13]); and resources primarily built on the transcriptomic data, such as HVIDB [14]. Some databases (e.g. dbGSRV [15]) have become inaccessible, reflecting the lack of timeliness in updating and maintenance, which limits the sustainability of data resources. Other databases like ViralZone [16] provide systematic virus knowledge, high-quality visual resources, and a broad overview of diverse viral biological processes. ViPR provides genomic and proteomic resources and analysis utilities [17]. VirusMentha focuses on the virus–host protein interactions [18]. However, these databases have limitations in integrating the latest multi-omics data on viral mutations and host interactions, and the timeliness of content updates cannot meet the needs of cutting-edge research. To address the challenges mentioned above, we previously developed ViMIC, a database of human disease-related virus mutations (VMs), viral integration sites (VISs) and *cis*-effects [19]. To our knowledge, ViMIC remains the sole database designed for comprehensive annotation of virus mutation, integration, and *cis*-effects of hosts. Since this version, massive amounts of data (literature and sequencing data) have accumulated in public domains and an updated version became imperative.

To meet these needs, we updated the database to ViMIC 2.0, which represents a comprehensive upgrade from ViMIC, with significant expansions in viral coverage (3.5-fold increase), mutation data (2-fold), multi-omics datasets (9.1-fold), and associated literature (2.5-fold). ViMIC 2.0 now encompasses 28 human pathogenic viruses, including 7 integrating viruses already curated in ViMIC and 21 newly incorporated viruses of high public-health relevance. The virus type was selected based on five criteria: (i) sustained transmission or multi-country outbreaks; (ii) significant clinical severity or population burden, reflected by hospitalization rates, case-fatality ratios, or risk of serious sequelae; (iii) documented carcinogenic potential; (iv) documented immune escape or rapid antigenic evolution with implications for vaccine effectiveness or antiviral susceptibility; and (v) broad geographic distribution or high reactivation-associated morbidity in the general population. The database incorporates 64 168 VMs, 115 033 VISs, 17 604 target genes, 177 virus-related diseases, and 1.28 million viral sequences. Furthermore, ViMIC 2.0 integrates 255 multi-omics datasets related to viral infections from public repositories. The scope of omics data has also been extended from transcriptomics alone to multi-levels. The detailed comparison between the contents of ViMIC 2.0 and ViMIC is provided in Table 1.

## Improved expansion and new features

### Data expansion

ViMIC 2.0 significantly expands virus coverage by incorporating 21 human pathogenic viruses with substantial public health importance, including IV, SARS-CoV-2, varicella-zoster virus, herpes simplex virus types 1/2 (HSV-1/2), yellow fever virus, human cytomegalovirus, Kaposi's sarcoma-associated herpesvirus, hepatitis A/C/D/E viruses, dengue virus, Japanese encephalitis virus, tick-borne encephalitis virus, Crimean–Congo hemorrhagic fever virus, chikungunya virus, Zika virus, Ebola virus, MARV virus, and rabies virus. In the updated version, Adeno-associated virus 2 has been removed due to its predominant use as a tool virus and the limited evidence for its pathogenicity, with only one disease-associated publication.

The literature was updated on 10 June 2025 and involved a re-curation of all articles included in ViMIC. Using combined keywords (e.g. human papillomavirus [tiab] AND mutation [tiab]) in PubMed, nearly 9000 additional mutation-related publications were initially identified. Ten new integration studies were included for subsequent overlap calculations using keywords (e.g. human papillomavirus [tiab] AND integration [tiab]) and manual curation. Viral sequence data were batch-acquired from NCBI Nucleotide database, incorporating 173 048 new sequences. At multi-omics level, we assembled 255 datasets related to virus infection from

**Table 1.** Comparison of the data included in ViMIC 2.0 and ViMIC

| Feature | ViMIC | ViMIC 2.0 | Fold increase |
|---|---|---|---|
| **Virus** | 8 | 28 | 3.5 |
| **VM** | 31 712 | 64 168 | 2.0 |
| **Viral integration** | 105 624 | 115 033 | 1.1 |
| **Virus related disease** | 77 | 177 | 2.3 |
| **Target gene** | 16 310 | 17 604 | 1.1 |
| **Literature** | 2539 | 6433 | 2.5 |
| **Virus sequence** | 1 110 015 | 1 283 063 | 1.2 |
| **Omics data** | 28 | 255 | 9.1 |
| RNA-seq/array | Yes | Yes | |
| scRNA-seq | No | Yes | New |
| Methylation profiling | No | Yes | New |
| Chip-seq | No | Yes | New |
| ChIP-on-chip | No | Yes | New |
| ATAC-seq | No | Yes | New |
| Variation | No | Yes | New |
| **Omics data analysis result** | | | |
| Immune infiltration analysis | Yes | Yes | |
| Differential expression analysis | No | Yes | New |
| Single cell type annotation | No | Yes | New |
| Single cell gene expression analysis | No | Yes | New |
| Differential methylation analysis | No | Yes | New |
| Epigenetic chipseeker annotation | No | Yes | New |
| Epigenetic deeptools analysis | No | Yes | New |

public repositories. The collection spans transcriptomic profiles such as RNA-seq, microarray, and single-cell transcriptomic sequencing (scRNA-seq), and epigenomic profiles such as Chromatin Immunoprecipitation followed by sequencing (ChIP-seq), ChIP-on-chip, Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq), and DNA methylation profiling, thereby expanding both the breadth and the depth of coverage. All public omics datasets underwent standardized manual curation: verification of complete metadata (virus type, experimental platform, and group design), confirmation of infection status or case/control labeling, assessment of data availability (raw and/or processed files), and review of available data processing summaries. Datasets with ambiguous viral labels, incomplete metadata, or inadequate quality were excluded.

## Data processing and standardization

For the initially screened mutation-related literature, this study employed the ViMRT tool [20] to automatically extract viral mutation information from each publication, including entities such as viral mutations, viral genes, diseases, and relevant support sentences. ViMRT is an automated tool developed by our team based on natural language processing (NLP) technology, which is specifically designed to identify and standardize VM-related information in biomedical literature.

Building upon the automated extraction process, we have optimized the structure of the mutation annotation table in ViMIC. The mutation annotation in ViMIC 2.0 covers the following 14 fields: PMID, Mutation, Mutation Level, Mutation Type, Disease, Gene/Protein/Region, Genotype/Subtype, Virus Reference, Sequence, Targetgene, Immune, Treatment, Clinical Information, and Location. Terms are standardized based on the output from ViMRT. Disease terms were extracted based on the research purpose and main content of the literature. For studies mainly based on *in vitro* cell line experiments, the disease name was marked as Cell line. The standardization of disease terms mainly refers to the NCBI MeSH database. The "Gene/Protein/Region" field retains the original descriptions from literature, with an additional column providing the corresponding encoding genes or proteins based on the NCBI Gene database. When a standard term for a disease or gene/protein is unavailable in these databases, a commonly accepted term is adopted to ensure consistency. The "Genotype/Subtype" field records virus genotype or subtype as reported in the original study. In addition to the "Viral Reference" field, ViMIC 2.0 introduces a "Sequence" information field. The "Viral Reference" field is defined as the reference sequence used for mutation alignment and analysis, or the original strain of virus used for site-directed mutagenesis. The "Sequence" field refers to the accession numbers (e.g. from GenBank) of sequencing data deposited in public databases. The "Targetgene" field represents host target genes involved in the study, with original terms standardized to Gene Symbols. It should be noted that target genes are primarily derived from studies on viral mutation or integration, while those from several virus-related studies not involving these two themes are currently excluded in this version. The "Immune" tag indicates whether the study contains immune-related research (e.g. mutation-associated immune escape). The "Treatment" field documents certain treatment methods, such as antiviral drugs. The "Clinical Information" field indicates whether the study includes clinical data, such as basic demographic details. The "Location" field annotates geographic region/country of data origin. All extracted results underwent manual curations.

For the viral sequence data update in the "Mutation" module, information such as GenBank ID, viral gene, genotype, and geographical origin was extracted from the NCBI Nucleotide database for subsequent annotation and statistical analysis. Moreover, the "Viral integration sites" help users to identify shared genomic loci and explore the potential impact of viral integration on host regulatory elements. For the virus integration data update, the genomic coordinates of all VISs were uniformly converted to the GRCh38 reference genome using the UCSC LiftOver tool. Subsequently, the GIGGLE tool [21] was employed to analyze the overlap between VISs and three categories of functional genomic regions from the

Cistrome database [22], defined as Cistrome factors including HM sites, transcription factor binding sites, and chromatin accessible regions.

### Omics data analysis

ViMIC 2.0 features a fully upgraded "Dataset" module, designed to provide researchers with virus-related multi-omics data sourced from public databases. The current release integrates three categories of omics data, including transcriptome, epigenome, and genome, covering the following experimental types: the expression profiling by array/ high-throughput sequencing, the methylation profiling by high-throughput sequencing /array/genome tiling array, the genome binding/occupancy profiling by genome tiling array, and the genome variation profiling by high-throughput sequencing. Since genomic datasets do not provide publicly accessible raw data, only the basic dataset information is listed in ViMIC 2.0.

At the transcriptome level, for bulk RNA-seq or expression microarray data, the limma package [23] in R was used for differential analysis based on sample groups (e.g. virus-infected group versus control). Differentially expressed genes (DEGs) were identified using the following criteria: $|\log_2$ fold change (FC)$| > 1$ and adjusted False Discovery Rate (FDR) $< 0.05$. Immune cell infiltration was inferred from bulk transcriptomic data using xCell [24]. We selected the online xCell tool by uploading normalized gene expression matrix to its web platform to estimate enrichment scores for 64 immune and stromal cell types. This analysis was conducted on virus-infected datasets. The combined metrics of scores and $p$-values were used to evaluate the significance of inferred cell-type enrichment in individual samples. The results are presented through raincloud plots and heatmaps.

For scRNA-seq dataset, a standard analytical pipeline was implemented. The Seurat package was utilized for processing scRNA-seq data and performing cell clustering [25]. A Seurat object was generated using the CreateSeuratObject function, followed by data normalization with the NormalizeData method. The FindVariableFeatures function was employed to identify the top 2000 highly variable genes. To address batch effects, data integration was performed using the "anchors" strategy [26]. The datasets were scaled using the ScaleData function, and principal component analysis was conducted. Data visualization was performed using UMAP algorithms. Marker genes for each cluster were identified with the FindAllMarkers function, utilizing the Wilcoxon rank-sum test. Cell types were manually annotated by integrating literature-based annotations with the top 10 marker genes identified for each cluster.

At the epigenomic level, DNA methylation analysis was performed using normalized $\beta$-value matrices. Differential analysis was conducted with the limma package, and CpG sites meeting $|\Delta\beta| > 0.2$ and FDR $< 0.05$ were identified as the significant differentially methylated positions (DMPs). For ChIP-seq/ChIP-on-chip/ATAC-seq datasets, a similar processing pipeline was applied. Raw sequencing reads were first subjected to quality control using FastQC, followed by alignment to the GRCh38 reference genome with Bowtie2 [27]. Subsequent filtering steps involved removing mitochondrial DNA (chrM), unplaced contigs (chrUn), and randomly assigned fragments. The alignmentSieve tool deepTools2 [28] was used for advanced filtering, involving deduplication, application of a minimum mapping quality threshold (Q25), exclusion of low-quality alignments, and masking of blacklisted regions, yielding a high-quality processed BAM file. Peak calling was performed using MACS3 on the filtered BAM files, and significantly enriched peaks were annotated genomically with the ChIPseeker package [29] to determine their genomic contexts, such as promoters, gene bodies, and distal regulatory regions, as well as their associated genes.

## Database construction and improved user interface

ViMIC 2.0 was built on the LNMP schema (Linux + Nginx + MySQL + PHP), leveraging MySQL for data storage and management, PHP for server-side logic processing, and Nginx for high-performance web service delivery and request distribution, collectively ensuring system stability and dynamic interactivity. Data visualization was conducted by multiple JavaScript libraries, including jQuery (v1.7) and ECharts (v5.5). ViMIC 2.0 database is freely accessible at http://www.biomedinfo.cn/ViMIC2.0/index.php or an alternative access address http://www.bmtongji.cn/ViMIC2.0/index.php. The ViMIC version remains available and can be accessed via the "Our Databases" panel on the ViMIC 2.0 homepage. The overall architecture of the database is illustrated in Fig. 1.

ViMIC 2.0 offers a user-friendly interface and diverse data query options. Users can easily navigate and utilize the database through the following steps: the homepage features three core modules: "Virus Mutation Site," "Viral Integration Site," and "Target Gene," allowing users to click and explore detailed information on specific virus. The right side of the page displays icons of 28 viruses; clicking on any icon provides an overview of the corresponding virus. Alternatively, users can also access the same information through the "Virus" module at the top of the page. The homepage search box facilitates rapid retrieval of virus genes/regions/proteins, diseases, and Cistrome factors, with all matching records presented on the results page. Users can also use "Advanced Search" to perform multi-condition combination queries.

In the "Virus Mutation Site" module, ViMIC 2.0 has enhanced information presentation. The "Mutation" page provides three drop-down menus, including virus gene/protein, disease, and description, along with shortcut buttons for frequently mutated regions. The table below lists detailed information for each mutation entry, including PMID of the source literature, viral gene region, encoded gene/protein, associated diseases, and description tags. The "Sequence" page features improved visualization, displaying the distribution of sequences across different geographical regions on a global map to help users intuitively understand the worldwide prevalence of viruses. Clicking the view button on the "Mutation" page allows users to access detailed annotation information for a specific mutation entry, organized into four sections: the basic characteristics of mutations, the functional impact and mechanisms, the clinical and epidemiological correlations, and the literature information.

ViMIC 2.0 significantly enhances the "Target Gene" module, which displays host target gene information mentioned in viral mutations or integration events researches. The module is organized into three sections: (i) Gene Information provides official gene symbols, full names, aliases, transcripts, gene types, genomic locations, functional descriptions, gene IDs, and links to external databases; (ii) Target Gene Related to VMs/VISs displays a table of associated viral mu-

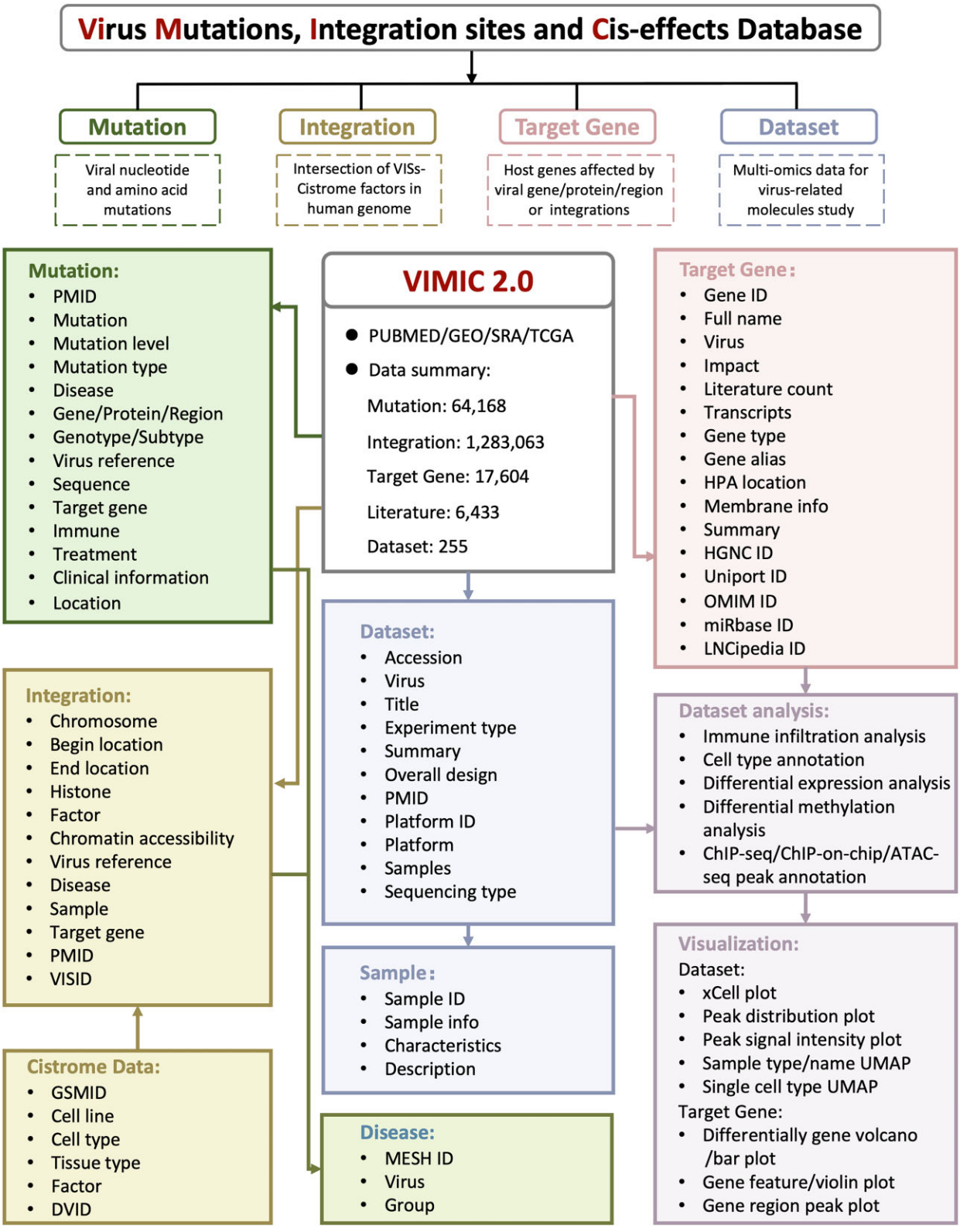



**Figure 1.** ViMIC2.0 database structure diagram. This diagram shows the content overview and data architecture of ViMIC2.0. ViMIC2.0 consists of four core modules (mutation, integration, target gene, and dataset), with additional modules for data analysis/visualization and disease. Each module contains sub-information tables. Each color represents a category of modules, and arrows of the same color indicate the table structure of all the information contained in that module.

tation and integration events; (iii) Target Gene Related to Omics Data integrates multi-omics data and visualization charts, including volcano/bar plots of DEGs, single-cell sample/sample type/cell type annotation UMAP and feature/violin plots of gene expression, volcano/bar plots of DMPs, as well as binding peak distribution maps from ChIP-seq/ChIP-on-chip/ATAC-seq data within the selected gene region. These data facilitate the multi-dimensional analysis of target gene expression and epigenetic regulation related to viral infection.

ViMIC 2.0 also introduces a newly developed "Dataset" module, which systematically presents public datasets across multi-levels for viruses. In this updated version, we mainly process transcriptome and epigenome data. For transcriptomic data, immune cell infiltration analysis using the xCell method was performed, with results visualized through raincloud plots and heatmaps. The scRNA-seq data are displayed via UMAP visualizations illustrating samples, types, and cell subpopulations. For each sample in ChIP-seq/ChIP-on-chip/ATAC-seq, pie charts, heatmaps of regions near transcription start sites, and line graphs of signal intensity are provided to illustrate enrichment patterns of protein binding or epigenetic modifications. All analytical results are available for download to facilitate further research and in-depth exploration.

Additionally, the "Viral Integration Site" module has been expanded with 9409 new VISs that overlap with Cistrome factors (overlap count >0), detailed overlap statistics and sample annotations, including GSMID, cell line, cell type, tissue type, and factor name are displayed. It also provides heatmaps and histograms to comprehensively visualize genome-wide overlap distributions between VISs and the functional regions. The "Disease" module consolidates mutation, integration, and literature information for various virus-related diseases and enables cross-linking with the mutation and integration modules. Finally, the "Download" module offers multiple resources for downloading VM, integration, sequence, omics, and gene data, along with detailed information such as file type, size, and access links.

## Case exploration

This study uses HPV as an example to demonstrate, through two cases, how to utilize the ViMIC 2.0 to retrieve information for subsequent in-depth research.

### Case 1: Multi-omics data reveals the potential role of NOVA1 in HPV-driven diseases

Users can explore human genes of interest through the "Target Gene" module. Taking the Neuro-oncological ventral antigen 1 (NOVA1) gene as an example (Fig. 2), the Description column shows that NOVA1 has been detected in HPV-related multi-omics data, marked by "E" for expression profiling (array/RNA-seq), "S" for scRNA-seq, "C" for genome binding/occupancy profiling, and "M" for methylation profiling. The "F" label denotes that ViMIC 2.0 provides visualized differential analysis charts for this gene [panel (a)]. Clicking "View" navigates users to the gene details page, which is divided into four sections: The first section displays basic gene information [panel (b)]. The second section notes reported viral integration events in or near NOVA1 in the context of HPV infection, as documented in prior studies [30, 31] [panel (c)]. The third section summarizes omics datasets containing NOVA1 [panel (d)]. The fourth section collectively presents analytical visualizations from these datasets. A label on the left of dataset ID means figures are shown below. In expression profiling, GSE140662 reveals NOVA1 downregulation in condyloma acuminatum [panel (e)]. Methylation analysis from GSE169622 [panel (f)] and TCHA-CESC indicates hypermethylation at NOVA1 promoter region in cervical cancer (CC). GSE143026 [panel (g)] includes ChIP-seq and ATAC-seq data, with ViMIC 2.0 displaying binding peak profiles of NOVA1 across samples. The scRNA dataset GSE197461 [panel (h)] shows NOVA1 expression predominantly in fibroblasts within the CC microenvironment. NOVA1 is a neuron-specific RNA splicing factor. Previous studies have shown its high expression in T cells and fibroblast/stromal cells within tertiary lymphoid structures formed during benign immune inflammation, while it is generally downregulated in tumor cells and other components in tumor microenvironment [32]. Another study also identifies NOVA1 as a potential key gene in cervical lesion progression and carcinogenesis. HPV16 or HPV18 infection can alter NOVA1 expression in human vaginal and foreskin keratinocytes, with both viral E6 and E7 suppressing NOVA1 expression [33]. These findings align with our transcript-level data. ViMIC 2.0 further enhances this understanding by offering a comprehensive multi-omics perspective, supporting deeper investigation into NOVA1 as a potential biomarker in HPV-mediated carcinogenesis and disease progression.

### Case 2: Multi-module interactive analysis reveals the potential mechanism of ZNF671 in cervical carcinogenesis

Users can explore viral mutations of interest through the "Mutation" module. Taking the D32E mutation in HPV16 E6 protein as an example (Fig. 3), D32E has been reported in 10 HPV-related studies. Our prior study demonstrated that in a C-33A CC cell line stably expressing the D32E mutation, the ZNF671 expression was significantly downregulated compared to wild-type C-33A cells ($p < .001$) [34]. Subsequently, by searching ZNF671 in the "Target Gene" module, users can find that ZNF671 has been detected in both expression and methylation sequencing data within HPV-related studies. Paired TCGA–CESC expression and methylation data showed that, compared to HPV-negative samples, ZNF671 is downregulated in HPV-positive CC samples, accompanied by pronounced promoter hypermethylation. Similarly, GSE169622 also indicated hypermethylated CpG sites of ZNF671 in the CC group, relative to the normal group. In our previous study, we evaluated the methylation status of ZNF671 in 41 CC patients and 29 non-cancer patients, all infected with HPV-16. Methylation analysis revealed that ZNF671 exhibited a sensitivity of 93% in diagnosing CC, with a positivity rate of only 28% in non-cancer patients [34]. Notably, ZNF671 is also a transcriptional repressor. A homepage search for the Cistrome factor ZNF671 revealed partial overlap between its binding sites and genomic regions of integration fragments from six viruses. In HPV-related data, VISs and Cistrome factors (especially on chromosome 2) exhibited significant enrichment. Interestingly, 38 out of 39 VISs overlapping with ZNF671 binding regions target the LINC00486 gene in the host genome. A study by Zeng *et al.*, which examined nearly a thousand HPV integration sites in tumor tissues, identified LINC00486 as a high-frequency integration gene [35].

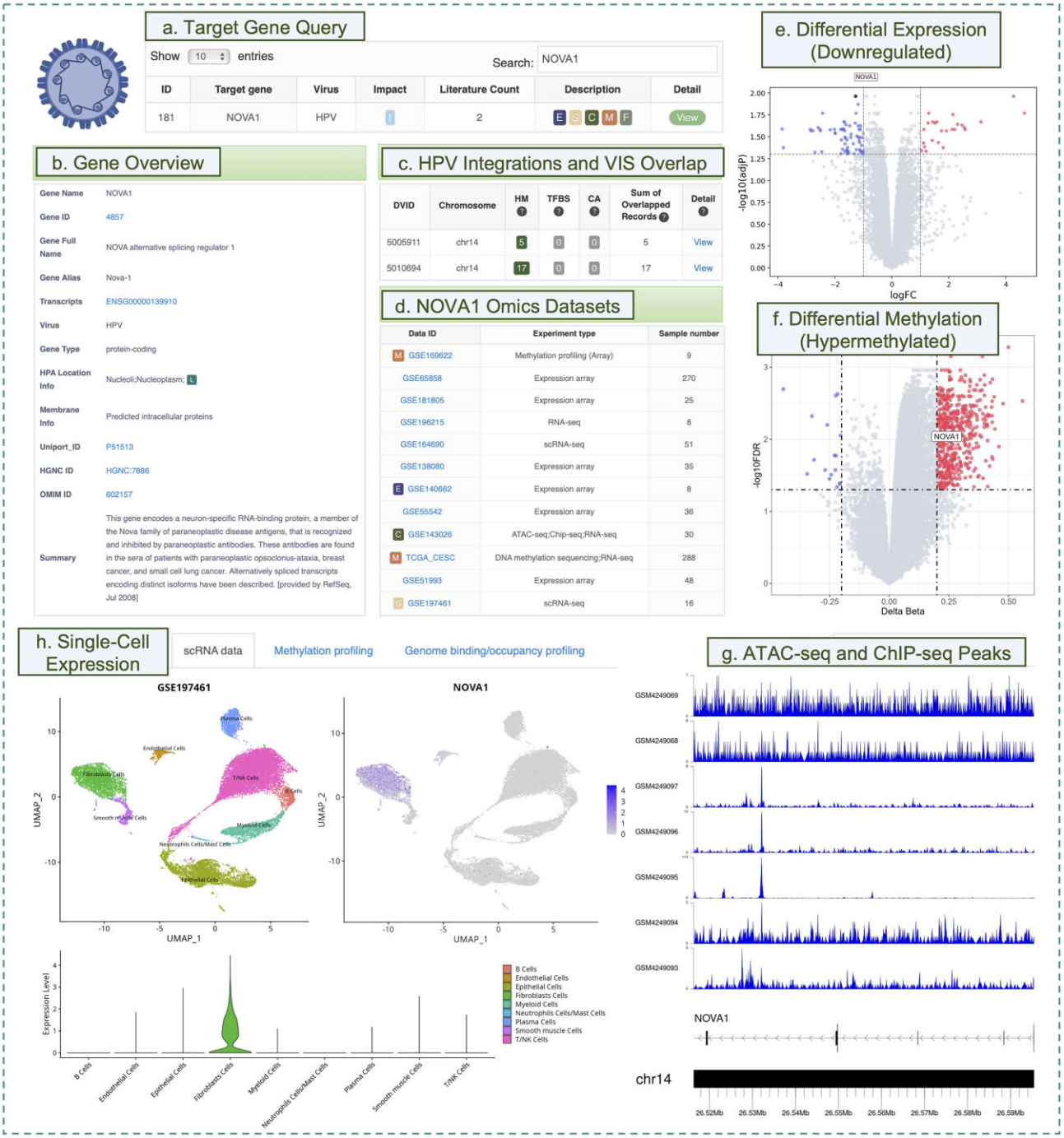



**Figure 2.** In the case of NOVA1, ViMIC2.0 demonstrates multi-omics exploration of host genes involved in viral pathogenesis. (**A**) Users can search for target genes (e.g. NOVA1) using the "Target Gene" module. The Description column indicates that NOVA1 is included in multiple HPV-related omics data. (**B**) The details page displays basic annotation information of NOVA1. (**C**) A table of viral integration studies mentioning NOVA1 and the associated VIS-factor overlap calculation results are provided. (**D**) A table list of omics datasets containing NOVA1. (**E**) Volcano plot of differential expression analysis, showing significant downregulation of the NOVA1. (**F**) Volcano plot of differential methylation analysis, showing hypermethylation of NOVA1. (**G**) Peak plots of NOVA1 detected in ATAC-seq and ChIP-seq data. (**H**) UMAP of cell type annotation, NOVA1 feature plot, and violin plots of NOVA1 expression in various cell types in scRNA-seq data. The icon used is licensed from BioRender: Wang, Y. (2025) https://BioRender.com/gjnudmr.

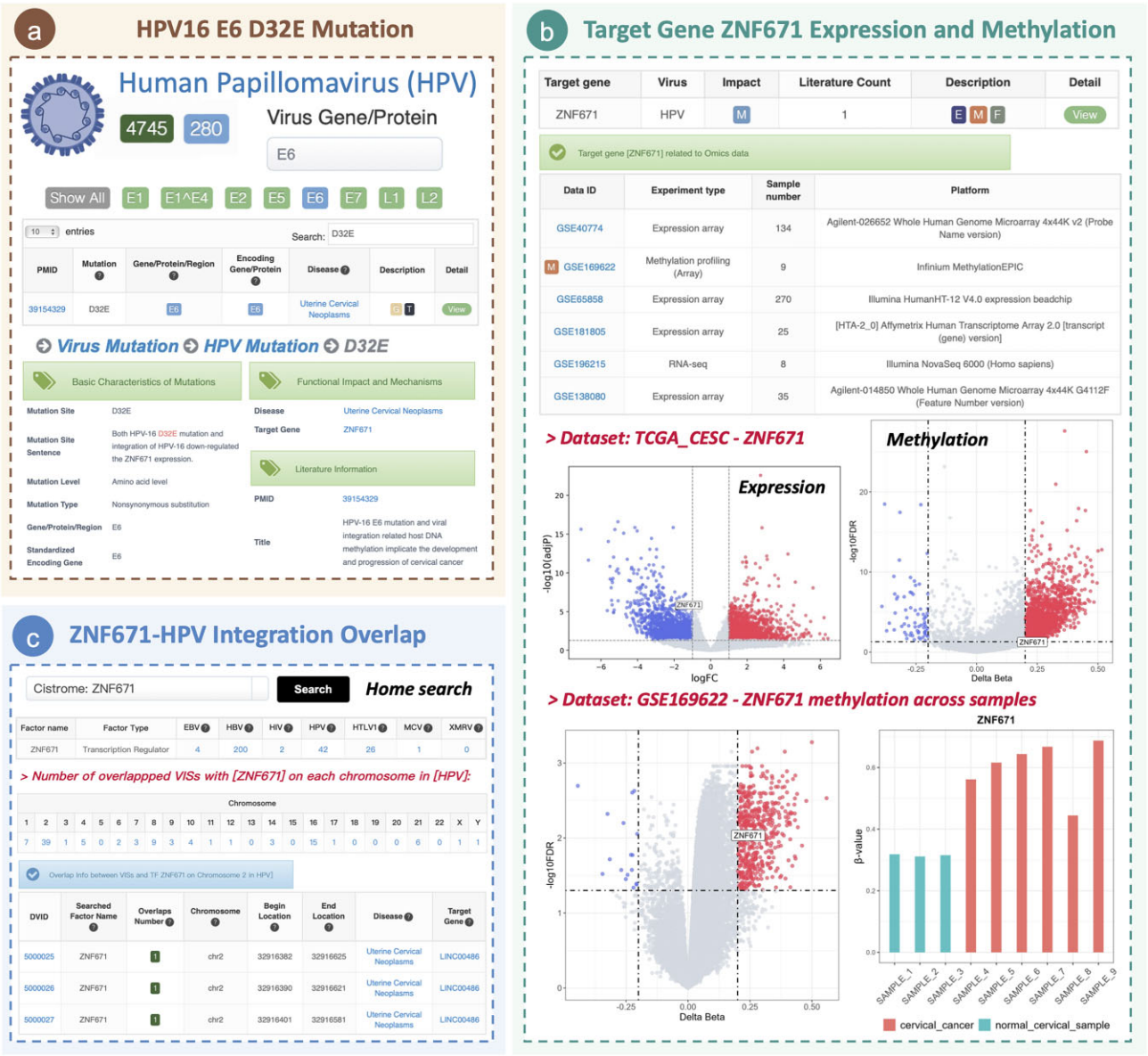



**Figure 3.** In the case of ZNF671, integrated multi-module analysis in ViMIC 2.0 reveals its potential role in HPV-mediated diseases. (**A**) VM: The search results for the D32E mutation in HPV. (**B**) Target Gene: Link to the "Target Gene" module for ZNF671 returns multi-omics evidence, including volcano and bar plots of differential ZNF671 expression and methylation. (**C**) Viral Integration: Searching for ZNF671 on the homepage displays the overlaps between ZNF671 and VIS in various viruses. Click the number under HPV to view the overlaps for each chromosome. Further click the number under a specific chromosome (e.g. chromosome 2) to view overlap information between VISs and ZNF671 on chromosome 2 in HPV. The icon used is licensed from BioRender: Wang, Y. (2025) https://BioRender.com/zmer4nv.

In summary, ViMIC 2.0's interactive analysis across its VM, integration, and target gene modules reveals the complexity of HPV-driven carcinogenesis, suggesting that HPV may modulate ZNF671 expression by altering its methylation status. Moreover, the overlap between ZNF671 binding sites and VISs, along with VIS-mediated regulation of LINC00486, offers critical insights into their potential synergistic roles in carcinogenesis, paving the way for further mechanistic validation.

## Conclusions and future extensions

We present ViMIC 2.0 as an updated version of ViMIC, which not only expands the data volume but also introduces new features. ViMIC 2.0 has made significant improvements across several aspects. First, ViMIC 2.0 incorporates 21 additional viruses, increasing the number of virus types by 3.5-fold, doubling the number of mutations, and nearly tripling the number of covered publications. Furthermore, ViMIC 2.0 extends omics data coverage from single transcriptome to multiple levels such as expression profiling, methylation profiling, single-cell transcriptome, and the genome binding/occupancy profiling, enabling users to examine virus-related host gene alterations from a more comprehensive perspective and thereby facilitating in-depth mechanistic studies. This upgrade is expected to attract substantial interest from researchers in virus field, significantly expanding its user base beyond the original version.

To ensure the sustainability as the virology literature and multi-omics datasets continue to expand, ViMIC 2.0 adopts a hybrid update strategy. First, new data are incorporated on a semiannual schedule through expert curation supported by ViMRT-assisted text mining. In the next phase, we will introduce NLP and large language models to optimize literature classification, virus entities extraction, and automatic standardize metadata generation, while maintaining human curation to ensure accuracy. Scheduled pipelines will continue to collect viral sequences and multi-omics datasets from major public repositories. By combining automation with expert validation, ViMIC2.0 will provide timely updates and maintain scalability over the long term.

A limitation of ViMIC 2.0 is that it reflects the underlying imbalance of the literature, in which well-studied viruses such as HIV-1, HBV, and HPV are heavily represented, whereas emerging or less-studied pathogens remain comparatively underrepresented. This bias arises from the uneven availability of published data and cannot be fully eliminated. To mitigate its impact, ViMIC 2.0 applies a rigorous standardization procedure, including term normalization and remapping of all integration sites to the GRCh38 human reference genome. Furthermore, mutation and integration records are structured according to a consistent schema and are manually verified against the primary literature. These measures provide a standardized framework that reduces heterogeneity across sources.

Looking ahead, we plan to enhance the fine-grained classification of mutation-related literature. For instance, categorizing studies into the cell line research, crystal structure analysis, and vaccine development. We also intend to incorporate AI tools (such as chatbots or retrieval-augmented generation technology) in future versions to enable intelligent generation and interaction with viral knowledge, further improving user engagement and experience. To support sustainable database updates and reduce the burden of manual curation, we will further develop advanced methods for intelligent mining of virological literature. These efforts will facilitate the long-term maintenance and updating of the ViMIC series database. With these new datasets and functionalities, ViMIC 2.0 and its future iterations are poised to greatly benefit a wide range of biomedical users, particularly those in virology research.

## Acknowledgements

We gratefully acknowledge the use of BioRender.com for creating the schematic diagrams and biological illustrations used in this publication. The graphic abstract is licensed from BioRender: Wang, Y. (2025) https://BioRender.com/0huqmt1.

*Author contributions*: Chenjun Huang (Data curation [equal], Formal analysis [equal], Validation [equal], Visualization [equal], Writing—original draft [equal]), Honglian Huang (Data curation [equal], Formal analysis [equal], Visualization [equal], Writing—original draft [equal]), Min Ding (Data curation [equal], Formal analysis [equal], Visualization [equal]), Jiawen Zhu (Data curation [equal]), Xin Qin (Data curation [equal]), Zeyuan Zhang (Data curation [equal]), Xiaoyang Zhao (Data curation [equal]), Ziyi Wei (Data curation [equal]), Min Wang (Data curation [equal]), M. James C. Crabbe (Writing—review & editing [equal]), Xiaoyan Zhang (Funding acquisition [equal], Supervision [equal], Writing—review & editing [equal]), and Ying Wang (Conceptualization [equal], Funding acquisition [equal], Methodology [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal])

## Conflict of interest

The authors disclose no conflicts of interest.

## Funding

This work was supported by the National Natural Science Foundation of China (32370694, 81972914) and Shanghai Public Health Research Project of Shanghai Municipal Health Commission (2024GKM25).

## Data availability

ViMIC2.0 is accessible at http://www.biomedinfo.cn/ViMIC2.0/index.php or an alternative access address http://www.bmtongji.cn/ViMIC2.0/index.php. The ViMIC database is available at http://www.biomedinfo.cn/ViMIC/index.php or http://www.bmtongji.cn/ViMIC/index.php. ViMRT is available at http://bmtongji.cn:1225/mutation/index. GIGGLE can search and rank the overlapping genomic loci between query genomic locations such as VISs and genome interval files and it is available as a GitHub repository at https://github.com/ryanlayer/giggle. LiftOver is a web service that provides genome coordinate transformation between different human genome reference assemblies and can be freely accessed at https://genome.ucsc.edu/cgi-bin/hgLiftOver. xCell is a webtool that performs cell type enrichment analysis from gene expression data for 64 immune and stroma cell types (https://comphealth.ucsf.edu/app/xcell).

## References

1. Hu B, Guo H, Zhou P *et al.* Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol* 2021;19:141–54. https://doi.org/10.1038/s41579-020-00459-7
2. Javanian M, Barary M, Ghebrehewet S *et al.* A brief review of influenza virus infection. *J Med Virol* 2021;93:4638–46. https://doi.org/10.1002/jmv.26990
3. Molina MA, Steenbergen RDM, Pumpe A *et al.* HPV integration and cervical cancer: a failed evolutionary viral trait. *Trends Mol Med* 2024;30:890–902. https://doi.org/10.1016/j.molmed.2024.05.009
4. Jia L, Gao Y, He Y *et al.* HBV induced hepatocellular carcinoma and related potential immunotherapy. *Pharmacol Res* 2020;159:104992. https://doi.org/10.1016/j.phrs.2020.104992
5. Bos J, Groen-van Schooten TS, Brugman CP *et al.* The tumor immune composition of mismatch repair deficient and Epstein–Barr virus-positive gastric cancer: a systematic review. *Cancer Treat Rev* 2024;127:102737. https://doi.org/10.1016/j.ctrv.2024.102737
6. Tian Y, Chen H, Qiao L *et al.* CIP2A facilitates the G1/S cell cycle transition via B-Myb in human papillomavirus 16 oncoprotein E6-expressing cells. *J Cell Mol Med* 2018;22:4150–60. https://doi.org/10.1111/jcmm.13693
7. Tian R, Wang Y, Li W *et al.* Genome-wide virus-integration analysis reveals a common insertional mechanism of HPV, HBV and EBV. *Clin Transl Med* 2022;12:e971. https://doi.org/10.1002/ctm2.971
8. Groza C, Chen X, Pacis A *et al.* Genome graphs detect human polymorphisms in active epigenomic state during influenza infection. *Cell Genom* 2023;3:100294. https://doi.org/10.1016/j.xgen.2023.100294

9. De Francesco MA, Gargiulo F, Dello Iaco F *et al.* Immune escape and drug resistance mutations in Patients with hepatitis B virus infection: clinical and epidemiological implications. *Life* 2025;15:672.
10. Briganti L, Annamalai AS, Bester SM *et al.* Structural and mechanistic bases for resistance of the M66I capsid variant to lenacapavir. *mBio* 2025;16:e0361324. https://doi.org/10.1128/mbio.03613-24
11. Tang D, Li B, Xu T *et al.* VISDB: a manually curated database of viral integration sites in the human genome. *Nucleic Acids Res* 2020;48:D633–41. https://doi.org/10.1093/nar/gkz867
12. Lu G and Moriyama EN. 2019nCoVR-A comprehensive genomic resource for SARS-CoV-2 variant surveillance. *Innovation* 2021;2:100150.
13. Hayer J, Jadeau F, Deleage G *et al.* HBVdb: a knowledge database for hepatitis B virus. *Nucleic Acids Res* 2013;41:D566–70. https://doi.org/10.1093/nar/gks1022
14. Yang X, Lian X, Fu C *et al.* HVIDB: a comprehensive database for human–virus protein–protein interactions. *Brief Bioinform* 2021;22:832–44. https://doi.org/10.1093/bib/bbaa425
15. Li P, Zhang Y, Shen W *et al.* dbGSRV: a manually curated database of genetic susceptibility to respiratory virus. *PLoS One* 2022;17:e0262373. https://doi.org/10.1371/journal.pone.0262373
16. De Castro E, Hulo C, Masson P *et al.* ViralZone 2024 provides higher-resolution images and advanced virus-specific resources. *Nucleic Acids Res* 2024;52:D817–21. https://doi.org/10.1093/nar/gkad946
17. Pickett BE, Sadat EL, Zhang Y *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* 2012;40:D593–8. https://doi.org/10.1093/nar/gkr859
18. Calderone A, Licata L, Cesareni G. VirusMentha: a new resource for virus–host protein interactions. *Nucleic Acids Res* 2015;43:D588–92. https://doi.org/10.1093/nar/gku830
19. Wang Y, Tong Y, Zhang Z *et al.* ViMIC: a database of human disease-related virus mutations, integration sites and *cis*-effects. *Nucleic Acids Res* 2022;50:D918–27.https://doi.org/10.1093/nar/gkab779
20. Tong Y, Tan F, Huang H *et al.* ViMRT: a text-mining tool and search engine for automated virus mutation recognition. *Bioinformatics* 2023;39:btac721.
21. Layer RM, Pedersen BS, DiSera T *et al.* GIGGLE: a search engine for large-scale integrated genome analysis. *Nat Methods* 2018;15:123–6. https://doi.org/10.1038/nmeth.4556
22. Zheng R, Wan C, Mei S *et al.* Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* 2019;47:D729–35. https://doi.org/10.1093/nar/gky1094
23. Ritchie ME, Phipson B, Wu D *et al. limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47. https://doi.org/10.1093/nar/gkv007
24. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* 2017;18:220. https://doi.org/10.1186/s13059-017-1349-1
25. Hao Y, Stuart T, Kowalski MH *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 2024;42:293–304. https://doi.org/10.1038/s41587-023-01767-y
26. Stuart T, Butler A, Hoffman P *et al.* Comprehensive integration of single-cell data. *Cell* 2019;177:1888–902.e21. https://doi.org/10.1016/j.cell.2019.05.031
27. Langmead B, Wilks C, Antonescu V *et al.* Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* 2019;35:421–32. https://doi.org/10.1093/bioinformatics/bty648
28. Ramirez F, Ryan DP, Gruning B *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;44:W160–5. https://doi.org/10.1093/nar/gkw257
29. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 2015;31:2382–3. https://doi.org/10.1093/bioinformatics/btv145
30. Yang W, Liu Y, Dong R *et al.* Accurate detection of HPV integration sites in cervical cancer samples using the nanopore MinION sequencer without error correction. *Front Genet* 2020;11:660. https://doi.org/10.3389/fgene.2020.00660
31. Cao C, Hong P, Huang X *et al.* HPV-CCDC106 integration alters local chromosome architecture and hijacks an enhancer by three-dimensional genome structure remodeling in cervical cancer. *J Genet Genomics* 2020;47:437–50. https://doi.org/10.1016/j.jgg.2020.05.006
32. Kim EK, Cho YA, Seo MK *et al.* NOVA1 induction by inflammation and NOVA1 suppression by epigenetic regulation in head and neck squamous cell carcinoma. *Sci Rep* 2019;9:11231. https://doi.org/10.1038/s41598-019-47755-8
33. Xu J, Liu H, Yang Y *et al.* Genome-wide profiling of cervical RNA-binding proteins identifies human papillomavirus regulation of RNASEH2A expression by viral E7 and E2F1. *mBio* 2019;10:e02687–18. https://doi.org/10.1128/mBio.02687-18
34. Huang C, Xiao X, Ai W *et al.* HPV-16 E6 mutation and viral integration related host DNA methylation implicate the development and progression of cervical cancer. *Infect Dis* 2025;57:66–80. https://doi.org/10.1080/23744235.2024.2391538
35. Zeng X, Wang Y, Liu B *et al.* Multi-omics data reveals novel impacts of human papillomavirus integration on the epigenomic and transcriptomic signatures of cervical tumorigenesis. *J Med Virol* 2023;95:e28789. https://doi.org/10.1002/jmv.28789


**Received:** September 10, 2025. **Revised:** October 1, 2025. **Accepted:** October 1, 2025