ViMIC: A Database of Human Disease-related Virus Mutations, Integration Sites and Cis-effects

Ying Wang^{1,2,#}, Yuantao Tong^{1,#}, Zeyu Zhang^{1,#}, Rongbin Zheng¹, Danqi Huang¹, Jinxuan Yang¹, Hui Zong¹, Fanglin Tan¹ and Xiaoyan Zhang^{1,*}

1 Research Center for Translational Medicine, Shanghai East Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, 200092, China

2 Department of Laboratory Medicine, Shanghai Eastern Hepatobiliary Surgery Hospital, Shanghai, 200438, China

*To whom correspondence should be addressed. Tel: 86-021-65980233; Fax: 86-021-65981041; Email: xyzhang@tongji.edu.cn

#The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

ABSTRACT

Molecular mechanisms of virus-related diseases involve multiple factors, including viral mutation accumulation and integration of a viral genome into the host DNA. With increasing attention being paid to virus-mediated pathogenesis and the development of many useful technologies to identify virus mutations (VMs) and viral integration sites (VISs), abundant literatures on these topics are available in PubMed. However, knowledge of VMs and VISs is widely scattered in numerous published papers, and the association of VMs with VISs in the viral genome or the functional annotation of VISs still lacks integration and curation. To address these challenges, we built a database of human disease-related **V**irus **M**utations, **I**ntegration sites and **C**is-effects (ViMIC), which specialize in three features: virus mutation sites, viral integration sites and target genes. In total, the ViMIC provides information on 6,461 VMs, 79,089 VISs, and 15,056 viral target genes of 8 viruses in 65 human diseases obtained from literatures. Furthermore, in ViMIC, users are allowed to explore the cis-effects of virus-host interactions by surveying 78 histone modifications, binding of 1,358 transcription regulators, and chromatin accessibility on these VISs. We believe ViMIC will become a valuable resource for the virus research community. The database is available at <http://bmtongji.cn/ViMIC/index.php>.

INTRODUCTION

Viruses are of substantial interest due to the seriousness of their associated diseases in humans. The mechanism of viral pathogenesis is complicated, as viruses interact with the host genome on multiple levels (1,2). For example, studies revealed that a virus may exert cis-effects by integrating its genome into human chromosomes (3-6). Mutations in viral regions or genes may also contribute to an increased risk for disease progression and drug resistance (7-9). Some virus mutations (VMs) may play an important role in promoting the occurrence of diseases by increasing or decreasing the incidence of viral DNA integration into host cells (10-13). Moreover, alterations in epigenetics may also be induced by the viral protein to promote activation or inhibition of target genes (14-17).

At present, millions of virus studies can be found in PubMed. Such studies have generated vast amounts of information relevant to VMs and viral integration sites (VISs) covering different types of viruses and human diseases. Moreover, domain knowledge that can be used to improve the understanding of virus cis-effects, for instance, gene expression and transcription cis-regulatory information, has been deposited in freely accessible databases, such as Cistrome Data Browser (18,19) and the Gene Expression Omnibus (GEO) (20). However, the well-curated information and data are still distributed in separate public domains that lack integration and standardization. Therefore, it is quite challenging for researchers to quickly search, visualize, and re-use such knowledge for their studies of virus-host interactions. Several databases, such as the Viral Integration Site DataBase (VISDB) (21), Dr.VIS (22), the Retrovirus Integration Database (RID) (23) and the ISDB (24) have been developed. However, to our knowledge, these databases focused on virus integration, and none of existing databases or web-servers are designed for systematically evaluating VMs and functionally annotating of VISs. Understanding of viral pathogenic mutations will abet the effective surveillance and early detection for the high-risk virus-infected diseases. The significant viral mutations or regions for affecting integration reactions have also been reported (10,11,13,25). Besides, previous studies indicated that the transcriptional regulators contribute to the control of viral replication and regulate critical cis-elements in a virus (26,27).

Therefore, to overcome the challenges mentioned above, we present a database of human disease-related **V**irus **M**utations, **I**ntegration sites and **C**is-effects (ViMIC). ViMIC is built to fill such gaps from the aspects of virus mutation, integration, and cis-effect of host. The unique of our database over other databases includes: ViMIC is the first database to provide virus mutation information retrieved from a large collection of published literature; ViMIC is the first to integrate VMs with VISs information from the literature; and ViMIC is the first to integrate VIS data with cistrome resources to help researchers who are interested in gene regulation mechanisms to systematically investigate the functional annotation of VISs in the host genome. ViMIC mainly covers three special features, including virus mutation sites, viral integration sites and target genes. First, the database facilitates the search for the VM annotations and the association between VMs and VISs on the viral reference genome. Second, the database can help users determine whether the integration site of viral fragments integrated into the host genome is in a functional element, distinguishing by histone modification, transcription factor binding and chromatin accessibility. Third, the database provides information on host target genes affected by the viral insertion or by the virus gene/protein/region, and on the correlation between gene expression levels and the fraction of infiltrating immune cells in human diseases.

CONSTRUCTION AND CONTENTS

A schematic overview of the ViMIC data collection, text-mining system, data processing and web interface is summarized in **Figure 1**.

Data sources

The major sources of data in ViMIC were derived from abundant published literatures, the Cistrome Data Browser (18,19), VISDB (21) and GEO database (20). Viral annotation information was

retrieved from a large collection of published literature through our text-mining system, which was followed by a manual curation process. The viral integration sites were retrieved from published literature and VISDB. The human transcription regulator binding, histone modification, and chromatin accessibility data were obtained from the Cistrome Data Browser. Moreover, for each target gene, ViMIC collected gene expression datasets of virus-infected diseases from GEO as well as the Cancer Genome Atlas (TCGA). Next, we conducted a correlation analysis between target gene expressions and fractions of infiltrating immune cells estimated by Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORT) (28).

Data retrieval and text mining

We used the Entrez Programming Utilities (E-Utilities) tool to build a set of full names and abbreviations of virus keywords into a fixed URL syntax to search and retrieve virus-related articles from PubMed. A Python script was run to filter the articles by matching keywords (e.g., mutation) as regular expressions. Full-text articles from PMC and abstracts from PubMed (BioC-XML format) with annotations of biomedical concepts in the Bioconcepts2pubtatorcentral file were downloaded from PubTatorCentral (29), and this file was then used to extract biomedical entities such as genes, mutations, diseases and chemicals. We also developed a Python script to automate the entity recognition task for those articles without PubTator annotation. The procedure includes literature downloading, conversion of PDF documents to plain text files (PDFMiner library), and noise filtering and information extraction for sentences containing mutations. Mutation entries were automatically extracted from PubMed and PubTator annotation file (30), followed by a manual curation to ensure the data quality of our database.

VIS quality control in VM–VIS association module

Some of the collected VIS information may be incomplete, which leads to no way to map the VM–VIS association, so we excluded those VISs that did not provide breakpoint locations and the viral reference genome. For each VIS, we define a level index to evaluate the quality of VIS positioning on the viral reference genome.

1. Level 1: The breakpoint locations in the viral reference genome and the sequence of the integrated fragment are both provided. We further used the multiple sequence alignment tool CLUSTALW (31) to check whether the information of this VIS was correct. In indel cases, all quality levels of VISs are Level 1.
2. Level 2: The precise start and end breakpoint locations in the viral reference genome are provided, while the sequence of the viral integration fragment is missing.
3. Level 3: Only one precise breakpoint location in the viral reference genome is provided, and the sequence of the viral integration fragment is missing. In this level, we can only use this breakpoint position in the virus reference genome to infer whether this specific breakpoint is a well-studied VM.

VM quality control in VM–VIS association module

For each VM and VIS, we retrieved its genotype/subtype and reference sequence of the virus

from the literature. As the collected reference genome information between VMs and VISs may not be matched exactly, we also defined a level index to evaluate the quality of the VM positioning on the viral reference genome for each VM.

1. Level 1: The reference genomes of VMs and VISs are exactly matched.
2. Level 2: The virus genotype/subtype of VMs and VISs are exactly matched. However, the reference genomes of VMs and VISs are not matched, but the genomic-length of two reference sequences is consistent.
3. Level 3: The virus genotype/subtype of VMs and VISs are not matched, but the genomic-length of two reference sequences is consistent. For example, HBV mutations, such as K130M, was frequent in both genotype C and B patients (32-34).
4. Level 4: The reference genome of VMs in the literature is missing. In these cases, we only annotated those VMs whose mutant base/amino acid in the wild-type sequence is consistent with the base/amino acid in the same position of the given reference sequence.

All VMs (Level 1-4) which mutant base/amino acid in the wild-type sequence (e.g., G in the G1896A or K in the K130M) must be consistent with the base/amino acid in the same position of the given reference sequence (e.g., the 130th amino acid in the X region is K in the ref. NC_003977).

VIS-Cistrome Data (Factor) Overlapping Analysis

To construct features for exploring the functional integration of a VIS (VIS-factors), we calculated the overlaps between the VIS and cistromes. The concept of "cistrome" has been widely used in the gene regulation field (18,19,35). As these studies state, a cistrome refers to "the set of cis-acting targets of a trans-acting factor on a genome-wide scale, also known as the in vivo genome-wide location of transcription factor binding-sites or histone modifications." Therefore, the host gene expression may be influenced if the virus integrates into the Cistrome. Cistrome DB collects human and mouse ChIP-seq, DNase-seq, and ATAC-seq data from publicly available resources. In our current study, we applied Cistrome DB data as cistrome factor information to integrate with data from a virus integration site and assist users to explore the cis-effect of VIS. The procedure included four steps.

First, we derived the viral integration sites in the human genome from VISDB (76,264 curated VISs) and from the PubMed literature (2825 VISs, 11 articles) that reported disease-related viral integration events in the human genome. The collected VISs were processed into BED format, which was then converted into human reference genome coordinates in GRCh38 by the LiftOver if a previous genome version was used. Second, we downloaded the ChIP-seq peaks of human transcription factor and histone modifications as well as the DNase-seq and ATAC-seq peaks from Cistrome Data Browser. We then compressed this information into bgzipped (.gz) files (a broad peak file using all peaks and a narrow peak file using 5-fold-enriched peaks). Third, a GIGGLE (genome search engine) (36) index was built using the compressed peak files, and we used GIGGLE to identify and search the significance of shared genomic loci (overlap) between the VIS and three genome interval files (histone modification site, transcription factor binding site and chromatin accessible region). Finally, we incorporated the GSMID and biological resource information from Cistrome Data Browser into the VIS-cistrome entry (overlap > 0). If a VIS overlapped with cistrome factor-related

GSM samples, it would indicate that the virus integration site is a potentially functional DNA element that can influence gene expression in the human.

Database contents

ViMIC is characterized by a comprehensive curation of VMs, functional annotations of VISs in the host genome and target genes. For each collected virus, ViMIC provides three features, as shown in **Figure 1**.

Feature I – Virus Mutation Site

The “Virus Mutation Site” feature provides two modules for users: the VM annotation module and the VM-VIS association module. The VM annotation module includes mutation site, mutation level, virus gene/protein/region, virus-associated disease, genotype/subtype, literature evidence, etc. We designed a description column to indicate more detailed information for each mutation site using the following tags: (i) D: drugs mentioned in the relevant literature; (ii) G: genotypes/subtypes mentioned in the relevant literature; (iii) R: reference genomes mentioned in the relevant literature. In the current version of VM-VIS association module, VISs of 5 viruses including hepatitis B virus (HBV), human papillomavirus virus (HPV), Epstein-Barr virus (EBV), adenovirus associated virus type 2 (AAV2) and Merkel cell polyomavirus (MCV), with precise breakpoint coordinates and the reference genome in the integration literature, were used to map the association between the VMs and VISs.

Feature II – Viral Integration Site

The second feature is the VIS-cistrome factor overlap calculation, which we specifically designed to explore whether the VIS is involved in a functional region in the host genome. Therefore, for each VIS, if there are overlaps between the VIS and the genomic peaks of the cistrome factors (sum of overlaps > 0), the detailed results are presented, including the overlap number and annotation of each cistrome sample (GSMID, cell line, cell type, tissue type and factor name). Furthermore, we summarized all the VIS-cistrome factor overlapped records on human chromosomes using heatmap and histogram plots, and the top-ranked cistrome factors ordered by the overlap number in each chromosome using stacked bar plots. Thus, users can view the overlap distribution between the VIS and cistrome factors in the human genome.

Feature III – Target gene

This feature provides information for the target genes affected by viral insertion or by the virus gene/protein/region. We have systematically annotated the target gene information, including official gene symbols, official full names, gene aliases, transcripts, gene types, gene location information, gene functions, gene IDs, drug information, etc. For each gene, if there were available gene expression data for patients with the virus-related disease, we further analysed the relationship between its expression and immune cell infiltration and represented it by heatmap and lollipop plots.

In the current version (as of August 31, 2020), the database contains 8 viruses, including HBV, HPV, EBV, AAV2, MCV, human immunodeficiency virus (HIV), human T-cell lymphotropic virus type-1 (HTLV1) and xenotropic murine leukaemia virus-related virus (XMRV), as well as 6,461 VMs, 79,089 VISs (71,595 VISs, sum overlaps > 0), 11,637,065 overlapping entries between VISs and Cistromes, 65 virus-associated diseases (25 public gene expression datasets performing immune

analysis), 15,056 target genes and 56 datasets with clinical information. The detailed statistics for ViMIC data are listed in **Table 1**.

All of the annotation and analysis results are deposited and managed in a MySQL relational database on a Linux server. Every VM and VIS was assigned unique identifiers, namely, a VVID and DVID, respectively, to perform a cross table query in the SQL database. More detailed information on the VMs, VISs, and target genes is crosslinked to the related resources including PubMed, VISDB, ENCODE, NCBI GenBank (37), Wikipedia, DrugBank (38), NCBI Gene database (39), Human Protein Atlas (HPA), HUGO Gene Nomenclature Committee (HGNC) (40), Vertebrate Genome Annotation (VEGA) (41), Online Mendelian Inheritance in Man (OMIM) (42), STRING (43), Rfam (44), miRbase (45), lncRNASNP2 (46), GEO, TCGA, EMBL (47), International Cancer Genome Consortium (ICGC) and European Genome-phenome Archive (EGA) (48).

Database Utility

Data query

ViMIC allows users to search data of interest in three ways. First, a quick search is available on the homepage. Users can input a keyword of interest including a virus name, disease, histone mark, transcription regulator, or virus gene/protein/region. Second, we designed a search engine at the top-right corner of each table in the web interface, where users can enter any keywords of interest. Third, an advanced search is also available on the home page for sophisticated searches by selecting a virus name of interest, and searching by human chromosome, disease, virus gene/protein/region or target gene. Each search produces tables of matched entries. Users can then view the detailed information by clicking on the hyperlink or the “View” button in the table.

Web interface and data visualization

A web interface was designed to provide user-friendly access to browse, search, analyse, visualize and download data. The database provides an interface for each virus with a summary and links to three ViMIC features to help users quickly check their hypothesis regarding VMs, VIS-cistrome factor overlaps or viral target genes.

First, the “Virus Mutation Site” feature provides the curated mutation information for 8 types of viruses. On the homepage of selected virus, ViMIC presents a catalogue of links to virus genes/proteins/regions and diseases, and a virus genome plot of each viral reference genome is provided for exploring associations between VMs and genome features. Users can check the location of a particular VM in the selected virus genome together with other well-studied mutations labelled with position, virus gene/protein/region, mutation sites and literature count. Moreover, users can click the “View VM-VIS Association” button to browse the association between reported VMs and reported VISs. The detailed virus mutation page is composed of mutation annotation and evidence sentences from the literatures.

Second, the “Viral Integration Site” feature presents analysis results, including the number of VISs and three cistrome factor-overlapped records. On the selected virus integration page, ViMIC provides the sum of VIS-cistrome factor counts and three VIS-cistrome factors overlap counts in each human chromosome, as well as two browse data menus for users to view overlapped entries by

chromosomes or factors. Users can browse VIS-cistrome factor overlapped records data by clicking the “View” button in the table as well as by selecting a chromosome or cistrome factor in the browse data menu. Moreover, ViMIC designs a search module to filter out overlap data according to options of user's interest like disease, biological source (cell line, cell type and tissue type), etc. In addition, ViMIC provides a heatmap of three VIS-cistrome factors overlap counts on each chromosome and two stacked barplots of top-ranked cistrome factors to help users have a better understanding of the distribution of VIS-cistrome factor overlap in the human genome.

Third, the “Target Gene” feature provides the curated target genes affected by viral genome insertion or by virus gene/region/protein. Users can browse the detailed gene information, related VISs/VMs and the correlation analysis results between gene expression levels and the proportions of 22 types of infiltrating immune cells in virus associated human diseases.

ViMIC also developed virus, disease, dataset, upload, download and help modules at the homepage, which allows users to be quickly redirected to a particular feature of users' interest. In addition, users can browse datasets for which clinical characteristics and virus sequencing data are available in the literature in our dataset module. To make the data presentation more intuitive, the statistical plots and tables are also shown on the home page of each module.

Case exploration

Visualization of the virus mutation site

In this article, we take HBV as an example. Users can click the green button with VM number in HBV, as shown in **Figure 2A**. ViMIC will then return a table containing the manually curated information for the HBV mutation sites. Users can quickly enter the keyword in the search box on the top-right of the table, and click the ‘View’ button to explore the detail information of searched mutation, including mutation level, virus gene/protein/region, genotype and evidences mentioned in literatures.

To browse the association of VMs and VISs, users can click “View VM-VIS Association” on the homepage of HBV. ViMIC will show the level 1's VIS data and matched reported VMs by default. As shown in the example of HBV reference genome NC_003977.1, users can view the distribution of mutations harbored in VISs compared with the NC_003977.1 sequence. The sequence map allows users to see all level 1's VISs and click DVID to query the detailed information of each VIS, view mutations that have been reported in this range, and enter the VM annotation module by clicking the icon of the VM location. On the other hand, ViMIC designed buttons for users to browse VIS data of other levels (e.g., level 2), if they are interested in which mutations are reported in the range of integration. Additionally, ViMIC set a series of dropdown menu items above the table for users to choose different reference sequence of virus, virus gene/protein/region, and VM browsing options (**Figure 2A**).

Visualization of the viral integration site overlapped with cistrome factors

If users wish to acquire a detailed factor list for one VIS entry, they can follow the step in **Figure 2B**. We take the entry of “1003124” on chromosome 4 in HBV as an example. After users click the “chr4” and search the entry of “1003124”, ViMIC shows the overlaps result of three cistrome factors. By clicking the green button with number at “TFBS” (Transcription factor binding site) column, ViMIC shows the ranking of all transcriptional regulators that have overlaps with this VIS. For instance, users

can view cistrome samples by clicking the 'View' button of the transcription factor HNF1A (hepatocyte nuclear factor 1 homeobox A), and can find that the sample ENCSR800QIT_1 shows an overlap between "1003124" and HNF1A at the position on chromosome 4 (coordinate range: 99028288-99028424, hg38). ViMIC provides information about this VIS integrating into the host genome which may be a functional element interacting with HNF1A. Moreover, users can also select "chr 4" in the integration menu of HBV to inspect the overlap distribution of three factors and virus integration fragments on human chromosome 4. In addition, users can select the factor browse menu to view all overlapped entries between curated HBV VISs and HNF1A in the human genome (**Figure 2B**).

Visualization of target gene information and immune analysis result

As METAP1 is a target gene of "1003124" (49), ViMIC presents the immune analysis provided by CIBERSORT between METAP1 gene expression and 22 types of infiltrating immune cells (**Figure 2C**). On the detailed information page of METAP1, ViMIC shows the heatmap and correlation analysis results for HBV infected diseases as well as providing the METAP1 gene name, ID, full name, alias, transcripts, type, HPA location and protein classification information, and several gene ID links to other resources. For instance, the METAP1 gene expression analysis results show the highest negative correlation with macrophage M2 in HBV+ liver cirrhosis samples.

DISCUSSION AND FUTURE DIRECTIONS

ViMIC is developed as a comprehensive database of human disease-related viruses from the perspective of VMs, VISs and virus cis-effects, and is a well-maintained resource that will be of great value to different users.

ViMIC collected virus mutation information retrieved from a large collection of published literature. This feature will provide researchers who are interested in VMs with annotations of the reported pathogenic mutations in 8 viruses related to human diseases easily. ViMIC integrated VMs and VISs information from the literatures and it integrates VIS data with cistrome resources. Previous studies pointed out that HBx can enhance virus replication leading to a higher frequency of HBV-DNA integration events into the human genome (11). The HBx sequence easily to integrate into hepatocytes with multiple point mutations, plays an important role in HBV infection and liver cancer development (10). This feature may help spur the further investigation of these two aspects of complicated mechanisms of viral pathogenesis. ViMIC also integrate VIS data with cistrome resources. The viral sequence inserted into the host genome may be a functional element interacting with histone modification, transcriptional regulation and cis-regulatory elements. For example, researcher have found that the transcriptional factor HNF1A contributes to the control of HBV replication (26). The liver-specific activity of one of the critical cis-elements, Enhancer II (nt. 1636-1741, partially contained in VIS 1003124), is regulated by multiple liver-enriched transcription factors, including HNF1A (27). Our work will contribute to and help researchers who are interested in gene regulation mechanisms to systematically investigate the functional annotation of VISs in the host genome. Furthermore, Jungnam Joo et al. (50) indicated that the viral target genes related to immune response may be associated with the viral integration patterns to avoid the immune surveillance system in the body and affect the response to treatment. ViMIC provides three aspects of viral target gene information for users to browse the gene information, VISs insertion information, and the

correlation between the expression level of viral target genes and infiltrating immune cell fractions. Additionally, advanced users can conveniently download the annotated virus data and conduct further analysis by themselves.

In the future, ViMIC will be maintained and updated continuously, since VMs, VISs and cistrome data are expected to accumulate. The utility of ViMIC will be improved in several aspects. For example, we will further classify and annotate drug resistance sites and immune escape sites. The sequence around the viral integration site will be downloaded from public databases to perform in-depth analysis to investigate the functional role of virus-host interactions. The upload module will be optimized for users to upload potential VISs of interest in batches and perform online analysis. In addition, we will also integrate other immune deconvolution tools such as TIMER (51) and xCell (52), to optimize the analysis of the association between gene expression and immune cell infiltration in the viral target gene module. In summary, we hope that ViMIC can give the biomedical community a better comprehensive understanding of virus-related research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

DATA AVAILABILITY

ViMIC is available at <http://bmtongji.cn/ViMIC/index.php>. The E-Utilities is a set of eight server-side programs that provide a stable interface into the Entrez query and database system available in NCBI (<https://www.ncbi.nlm.nih.gov/books/NBK25500/>). 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>. PubTator is a web service that can automatically annotate biomedical entries including genes and mutations, from the literature, which is available at <https://www.ncbi.nlm.nih.gov/research/pubtator/>. CIBERSORT uses the gene expression data to estimate the abundances of member cell types from a mixed cell population, and it can be freely accessed at <https://cibersort.stanford.edu/>.

ACKNOWLEDGEMENTS

We are grateful to Dr. Shipeng Chen for reading the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China [81972914, 81573023], the Fundamental Research Funds for the Central Universities [22120200014], Shanghai “Rising Stars of Medical Talent” Youth Development Program [2019-72], and the National Key R&D Program of China [2016YFC1303200].

CONFLICT OF INTEREST

The authors disclose no conflicts of interest.

REFERENCES

1. Herrington, C.S., Coates, P.J. and Duprex, W.P. (2015) Viruses and disease: emerging concepts for prevention, diagnosis and treatment. *J Pathol*, **235**, 149-152.
2. Chen, Y., Williams, V., Filippova, M., Filippov, V. and Duerksen-Hughes, P. (2014) Viral carcinogenesis: factors inducing DNA damage and virus integration. *Cancers (Basel)*, **6**, 2155-2186.
3. Peng, R.J., Han, B.W., Cai, Q.Q., Zuo, X.Y., Xia, T., Chen, J.R., Feng, L.N., Lim, J.Q., Chen, S.W., Zeng, M.S. *et al.* (2019) Genomic and transcriptomic landscapes of Epstein-Barr virus in extranodal natural killer T-cell lymphoma. *Leukemia*, **33**, 1451-1462.
4. Nault, J.C., Datta, S., Imbeaud, S., Franconi, A., Mallet, M., Couchy, G., Letouze, E., Pilati, C., Verret, B., Blanc, J.F. *et al.* (2015) Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat Genet*, **47**, 1187-1193.
5. Zhao, L.H., Liu, X., Yan, H.X., Li, W.Y., Zeng, X., Yang, Y., Zhao, J., Liu, S.P., Zhuang, X.H., Lin, C. *et al.* (2016) Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nat Commun*, **7**, 12992.
6. Kataoka, K., Nagata, Y., Kitahara, A., Shiraishi, Y., Shimamura, T., Yasunaga, J., Totoki, Y., Chiba, K., Sato-Otsubo, A., Nagae, G. *et al.* (2015) Integrated molecular analysis of adult T cell leukemia/lymphoma. *Nat Genet*, **47**, 1304-1315.
7. Kanda, T., Yajima, M. and Ikuta, K. (2019) Epstein-Barr virus strain variation and cancer. *Cancer Sci*, **110**, 1132-1139.
8. Liu, L., Liu, Y., Chen, R., Li, X., Luo, D., Zhao, Y., Li, Q., Huang, B., Wang, F.S., Liu, X. *et al.* (2019) Prevalence of the entecavir-resistance-inducing mutation rtA186T in a large cohort of Chinese hepatitis B virus patients. *Antiviral Res*, **164**, 131-138.
9. An, P., Xu, J., Yu, Y. and Winkler, C.A. (2018) Host and Viral Genetic Variation in HBV-Related Hepatocellular Carcinoma. *Front Genet*, **9**, 261.
10. Zhang, X. and Ding, H.G. (2015) Key role of hepatitis B virus mutation in chronic hepatitis B development to hepatocellular carcinoma. *World J Hepatol*, **7**, 1282-1286.
11. Mohamadkhani, A., Montazeri, G. and Poustchi, H. (2011) The importance of hepatitis B virus genome diversity in Basal core promoter region. *Middle East J Dig Dis*, **3**, 13-19.
12. Tanaka, Y., Mukaide, M., Orito, E., Yuen, M.F., Ito, K., Kurbanov, F., Suguchi, F., Asahina, Y., Izumi, N., Kato, M. *et al.* (2006) Specific mutations in enhancer II/core promoter of hepatitis B virus subgenotypes C1/C2 increase the risk of hepatocellular carcinoma. *J Hepatol*, **45**, 646-653.
13. Mesplede, T., Quashie, P.K., Osman, N., Han, Y., Singhroy, D.N., Lie, Y., Petropoulos, C.J., Huang, W. and Wainberg, M.A. (2013) Viral fitness cost prevents HIV-1 from evading dolutegravir drug pressure. *Retrovirology*, **10**, 22.
14. Kvaratskhelia, M., Sharma, A., Larue, R.C., Serrao, E. and Engelman, A. (2014) Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Res*, **42**, 10209-10225.
15. Hatano, T., Sano, D., Takahashi, H., Hyakusoku, H., Isono, Y., Shimada, S., Sawakuma, K., Takada, K., Oikawa, R., Watanabe, Y. *et al.* (2017) Identification of human papillomavirus (HPV) 16 DNA integration and the ensuing patterns of methylation in HPV-associated head and neck squamous cell carcinoma cell lines. *Int J Cancer*, **140**, 1571-1580.
16. Gao, W., Jia, Z., Tian, Y., Yang, P., Sun, H., Wang, C., Ding, Y., Zhang, M., Zhang, Y., Yang, D. *et al.* (2020) HBx Protein Contributes to Liver Carcinogenesis by H3K4me3 Modification Through Stabilizing WD Repeat Domain 5 Protein. *Hepatology*, **71**, 1678-1695.
17. Li, Y., Fu, Y., Hu, X., Sun, L., Tang, D., Li, N., Peng, F. and Fan, X.G. (2019) The HBx-CTTN interaction promotes cell proliferation and migration of hepatocellular carcinoma via CREB1. *Cell Death Dis*, **10**, 405.
18. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.H., Brown, M., Zhang, X., Meyer, C.A. *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res*, **47**, D729-D735.
19. Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res*, **45**, D658-D662.
20. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*, **41**, D991-995.

21. Tang, D., Li, B., Xu, T., Hu, R., Tan, D., Song, X., Jia, P. and Zhao, Z. (2020) VISDB: a manually curated database of viral integration sites in the human genome. *Nucleic Acids Res*, **48**, D633-D641.
22. Yang, X., Li, M., Liu, Q., Zhang, Y., Qian, J., Wan, X., Wang, A., Zhang, H., Zhu, C., Lu, X. *et al.* (2015) Dr.VIS v2.0: an updated database of human disease-related viral integration sites in the era of high-throughput deep sequencing. *Nucleic Acids Res*, **43**, D887-892.
23. Shao, W., Shan, J., Kearney, M.F., Wu, X., Maldarelli, F., Mellors, J.W., Luke, B., Coffin, J.M. and Hughes, S.H. (2016) Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. *Retrovirology*, **13**, 47.
24. Sibley, T.R., Silberman, E.J. and Mullins, J.I. (2019) ISDB: a database toolkit for storing and analyzing viral integration site data. *Bioinformatics*, **35**, 1073-1075.
25. Li, S., Shen, H., Liu, Z., Li, N., Yang, S., Zhang, K. and Li, J. (2017) Integration of human papillomavirus 16 in esophageal carcinoma samples. *Infect Agent Cancer*, **12**, 53.
26. Quasdorff, M., Hosel, M., Odenthal, M., Zedler, U., Bohne, F., Gripon, P., Dienes, H.P., Drebbler, U., Stippel, D., Goesser, T. *et al.* (2008) A concerted action of HNF4alpha and HNF1alpha links hepatitis B virus replication to hepatocyte differentiation. *Cell Microbiol*, **10**, 1478-1490.
27. Cai, Y.N., Zhou, Q., Kong, Y.Y., Li, M., Viollet, B., Xie, Y.H. and Wang, Y. (2003) LRH-1/hB1F and HNF1 synergistically up-regulate hepatitis B virus gene transcription and DNA replication. *Cell Res*, **13**, 451-458.
28. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M. and Alizadeh, A.A. (2015) Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*, **12**, 453-457.
29. Wei, C.H., Allot, A., Leaman, R. and Lu, Z. (2019) PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res*, **47**, W587-W593.
30. Wei, C.H., Kao, H.Y. and Lu, Z. (2013) PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res*, **41**, W518-522.
31. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673-4680.
32. Lin, X., Xu, X., Huang, Q.L., Liu, Y.Q., Zheng, D.L., Chen, W.N. and Lin, J.Y. (2005) Biological impacts of "hot-spot" mutations of hepatitis B virus X proteins are genotype B and C differentiated. *World J Gastroenterol*, **11**, 4703-4708.
33. Leon, B., Taylor, L., Vargas, M., Luftig, R.B., Albertazzi, F., Herrero, L. and Visona, K. (2005) HBx M130K and V131I (T-A) mutations in HBV genotype F during a follow-up study in chronic carriers. *Viral J*, **2**, 60.
34. Kim, H.J., Park, J.H., Jee, Y., Lee, S.A., Kim, H., Song, B.C., Yang, S., Lee, M., Yoon, J.H., Kim, Y.J. *et al.* (2008) Hepatitis B virus X mutations occurring naturally associated with clinical severity of liver disease among Korean patients with chronic genotype C infection. *J Med Virol*, **80**, 1337-1343.
35. Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y. *et al.* (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol*, **12**, R83.
36. Layer, R.M., Pedersen, B.S., DiSera, T., Marth, G.T., Gertz, J. and Quinlan, A.R. (2018) GIGGLE: a search engine for large-scale integrated genome analysis. *Nat Methods*, **15**, 123-126.
37. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2017) GenBank. *Nucleic Acids Res*, **45**, D37-D42.
38. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*, **46**, D1074-D1082.
39. Brown, G.R., Hem, V., Katz, K.S., Ovetsky, M., Wallin, C., Ermolaeva, O., Tolstoy, I., Tatusova, T., Pruitt, K.D., Maglott, D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*, **43**, D36-42.
40. Bruford, E.A., Lush, M.J., Wright, M.W., Sneddon, T.P., Povey, S. and Birney, E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res*, **36**, D445-448.
41. Wilming, L.G., Gilbert, J.G., Howe, K., Trevanion, S., Hubbard, T. and Harrow, J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res*, **36**, D753-760.

42. Amberger, J.S., Bocchini, C.A., Scott, A.F. and Hamosh, A. (2019) OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*, **47**, D1038-D1043.
43. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*, **47**, D607-D613.
44. Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res*, **43**, D130-137.
45. Kozomara, A., Birgaoanu, M. and Griffiths-Jones, S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res*, **47**, D155-D162.
46. Miao, Y.R., Liu, W., Zhang, Q. and Guo, A.Y. (2018) lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res*, **46**, D276-D280.
47. Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res*, **47**, D745-D751.
48. Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R. *et al.* (2015) The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet*, **47**, 692-695.
49. Lau, C.C., Sun, T., Ching, A.K., He, M., Li, J.W., Wong, A.M., Co, N.N., Chan, A.W., Li, P.S., Lung, R.W. *et al.* (2014) Viral-human chimeric transcript predisposes risk to liver cancer development and progression. *Cancer Cell*, **25**, 335-349.
50. Joo, J., Omae, Y., Hitomi, Y., Park, B., Shin, H.J., Yoon, K.A., Sawai, H., Tsuiji, M., Hayashi, T., Kong, S.Y. *et al.* (2019) The association of integration patterns of human papilloma virus and single nucleotide polymorphisms on immune- or DNA repair-related genes in cervical cancer patients. *Sci Rep*, **9**, 13132.
51. Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., Li, B. and Liu, X.S. (2020) TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res*, **48**, W509-W514.
52. Aran, D., Hu, Z. and Butte, A.J. (2017) xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol*, **18**, 220.

TABLE AND FIGURES LEGENDS

Table 1. Statistics of ViMIC data

Data Type	Count	Description
Virus	8	Including 5 DNA viruses and 3 RNA viruses.
Virus Mutation (VM)	6,461	Curated mutation information entries.
Histone Mark	78	Histone Mark types overlapped with VISs in ViMIC.
Transcription Regulator	1,358	Transcription regulator types overlapped with VISs in ViMIC.
Viral Integration Site (VIS) (Sum Overlaps >0)	71,595	VISs that overlaps with cistrome data.
VIS-Histone Modification (HM) Overlaps >0	66,322	VISs that overlaps with cistrome HM data.
VIS-Transcription Factor Binding Site (TFBS) Overlaps >0	53,719	VISs that overlaps with cistrome TFBS data.
VIS-Chromatin Accessibility (CA) Overlaps >0	29,172	VISs that overlaps with cistrome CA data.
Virus Related Disease	65	Disease types which associated with the 8 ViMIC viruses.
Target Gene	15,056	Curated target genes affected by viral genome insertion or virus gene/protein/region regulation.
Literature	2,153	Count of the literature related to virus mutation and integration.
Clinical Annotation and Data	56	Clinical characteristics and virus sequencing data which are available in literatures.

FIGURE

Figure 1. A schematic overview of the ViMIC database. ViMIC collects publicly available virus mutations (VMs), viral integration sites (VISs), cistrome factors and target genes information from multiple public resources, including PubMed, Cistrome Data Browser, VISDB, GEO, etc. ViMIC uses text mining to extract the virus-related bioconcepts from published citations, followed by a manual curation to ensure annotation accuracy. The data processing layer is responsible for specific tasks, such as establishing annotation standards stored in a MySQL relationship database, performing VM-VIS mapping to the reference genome, calculating overlaps between VIS and cistrome factors, analysing the correlation between gene expression and infiltrating immune cell fractions and generating statistical plots. We designed the ViMIC interface with three main features: virus mutation sites, viral integration sites and target genes. ViMIC provides three ways to query and explore data: by keyword search, quick search engine or an advanced search menu. In addition, ViMIC includes other modules, labelled as Virus, Disease, Dataset, Download and Help that allow users to better explore the collected data.

Figure 2. Screenshot depicting an example exploration of the ViMIC database. (A) Assuming the user is interested in HBV, he or she can access a page on ViMIC containing brief mutation information on HBV by clicking HBV VM Count. The user can input a keyword through the quick search engine and enter the detailed mutation information page to view the mutation site. Users can click the “View VM-VIS Association” button to browse the association between VMs and VISs. By selecting a reference genome (e.g., NC_003977.1), ViMIC will return a result showing the distribution of mutations harbored in VISs or the reported VMs in the range of viral integration fragments based on different levels of VIS data. (B) Assuming the user is interested in acquiring a detailed factor list for VIS entry “1003124” on chromosome 4 in HBV, ViMIC will show the overlaps result of three cistrome factors after the users clicks the “chr4” and search the entry of “1003124”. Assuming the user is interested in the “24” overlap number within the “1003124” and “TFBS” entry, ViMIC will then return the ranking of all transcriptional regulators that have overlaps with this VIS. By clicking the ‘View’ button of HNF1A, ViMIC will generate a HNF1A-related table including the GSMID/ENCODE ID, cell line, cell type, tissue type and factor name. On the HBV VIS-cistrome factor homepage, the user can select the chromosome menu to view the overlap distribution of three factors and virus integration fragments on a specific chromosome (e.g., chr4). By selecting the factor menu, the user can explore more overlap statistics and distribution information for a specific transcription factor (e.g., HNF1A). (C) ViMIC provides the “Target Gene” feature for the curated target genes affected by viral genome insertion or by virus gene/protein/region. Assuming the user is interested in the VIS 1003124 target gene METAP1 in HBV, ViMIC will show METAP1 gene information, inserted VISs reported in the literature, and the correlation between gene expression levels and the fraction of infiltrating immune cells in HBV related human disease on the detailed gene information page.

1

Data Source and Collection

Virus Mutation Site

- ✓ Mutation Site
- ✓ Virus Gene/Protein/Region
- ✓ Disease
- ✓ Genotype/Subtype
- ✓ Reference Genome
- ✓ Literature

Target Gene

- ✓ Gene Name Info
- ✓ Gene-related Virus
- ✓ Gene-related VIS
- ✓ Location Info
- ✓ Immune-related Analysis Results
- ✓ Other Resources' IDs

Viral Integration Site

- ✓ Histone Modification Overlaps
- ✓ Transcription Factor Binding Site Overlaps
- ✓ Chromatin Accessibility Overlaps
- ✓ Chromosome
- ✓ Location (GRCh38)
- ✓ Disease
- ✓ Viral Reference Genome
- ✓ GSM/ENCODE Sample
- ✓ Target Gene
- ✓ VISDB ID
- ✓ Literature
- ✓ Cell Line
- ✓ Cell/Tissue Type
- ✓ Factor Name

3

Data Processing

Mutation Feature

- Establish mutation annotation standards
- VM quality control (genotype, reference)
- VIS quality control (breakpoint location, sequence, reference)
- Mapping VM and VIS to reference genome

VIS-Factor Feature

- Establish VIS-factor annotation standards
- GRCh38 conversion
- VIS and factor overlaps calculation
- Sample mapping

Target Gene Feature

- Establish gene annotation standards
- Link to other databases
- Immune related analysis

2

Text-mining System

4

Web Interface and Visualization

Virus Mutation Site

ID	VVID	Mutation	Gene/Protein/Region	Literature Count	Description	Detail
1	VVHBV0001	K130M		17		View
2	VVHBV0002	K130N		1		View
3	VVHBV0003	I127L		1		View
4	VVHBV0004	I127M		1		View
5	VVHBV0005	I127N		3		View

DVID	Start_End	Breakpoint	Orient	Viral Mutation	VVID	Gene/Protein/Region	VM level
1003124	1671-1811		+	A1762T	VVHBV0018	BCP	Level 2
1003124	1671-1811		+	G1764A	VVHBV0118	BCP	Level 2

Quick search

Please input a virus, disease, histone mark, or transcription regulator

Search

Advanced search

Please select one virus name in the first step:

HBV

Please select chromosomes of interest:

Please select disease of interest:

Please select virus gene/protein/region of interest:

Select keywords of interest

Please select target gene of interest:

Search

Browse by other modules

- Virus
- Disease
- Dataset

Download

- Mutation
- VIS-Factor
- Target gene

Help

- Document
- Video

A

Virus Mutation Site
Reported mutations in 8 viruses related to human diseases

Viral Integration Site
Intersection of viral integration site and Cistrome factors in the human genome

Target Gene
Correlation between gene expression and fraction of infiltrating immune cells

View VM-VIS Association

Select Viral Reference Genome: NC_003977.1

Custom filter:
Virus gene/protein/region: all
VM browsing options:
☐ Nucleotide level ☐ Amino acid level
Submit

Selected Reference Genome: NC_003977.1

Select one integration quality control:
☒ level 1 ☐ level 2 ☐ level 3

Chronic HBV (CHB) infection
HBV infection
Hepatocellular carcinoma
HBV-HIV coinfection
Occult HBV infection
Click here to see all HBV associated human diseases

Hepatitis B Virus (HBV) 1559

Show 10 entries Search: Input the search keyword

ID	VVID	Mutation	Gene/Protein/Region	Literature Count	Description	Detail
1	VVHBV0001	K130M	X	17	D G R	View
2	VVHBV0002	K130N	X	1	R	View
3	VVHBV0003	I127L	X	1		View

Start_End Breakpoint Orient VIS Count VM Count Gene/Protein/Region detail

1_1092 + 1 452 8 view

VVID	Mutation	Gene/Protein/Region	Nucleotide Start_End	VM level
VVHBV0183	G102C	S	458-460	Level 3
VVHBV0184	G102R	S	458-460	Level 4
VVHBV0185	Q101H	S	455-457	Level 4

DVID	Start_End Breakpoint	Orient	Viral Mutation	VVID	Gene/Protein/Region	VM level
1003124	1671-1811	+	A1762T	VVHBV0018	BCP	Level 2
1003124	1671-1811	+	G1764A	VVHBV0118	BCP	Level 2
1003124	1671-1811	+	T1753C	VVHBV0099	BCP	Level 2

Range: 94-459
Reference Genome: NC_003977.1

View sequence map at level 1

94-193 TCCGACTACTGCCTCACCCATATCGTCAATCTCTCGAGGACTGGGGACCCCTGCACCGAACATGGAGAGCAC
1001042 TCCGACTACTGCCTCACCCATATCGTCAATCTCTCGAGGACTGGGGACCCCTGCACCGAACATGGAGAGCAC
1006394

B

Viral Integration Site
Intersection of viral integration site and Cistrome factors in the human genome

Hepatitis B Virus (HBV) 17585

Chromosome	VIS Count	Sum VIS-Factor	Histone Modification	Transcription Factor Binding Site	Chromatin Accessibility	Detail
chr1	1525	462352	182064	215651	64637	View
chr2	1801	559345	256150	218456	84739	View

VIS Count and VIS-Cistrome Factors Overlap in [chr4]

Show 10 entries Search: Input the search keyword

Rank of Overlapped Records	DVID	Location	HM	TFBS	CA	Sum of Overlapped Records	Detail
1	1003124	99028288-99028424	13	24	127	164	View

VIS [DVID: 1003124] Location [chr4: 99028288-99028424] (GRCh38)

Show 10 entries Search:

Factor Rank on this VIS	Factor Name	Overlapped Records	Details
1	HNF1A	3	View
2	TFAP2A	3	View
3	YAP1	3	View

> Overlapped records detail between VIS [DVID: 1003124] and transcription factor binding site [HNF1A]:

Show 10 entries Search: liver

GSMID or ENCODE	Cell line	Cell type	Tissue type	Factor
ENCSTR000QIT_1	HepG2	Epithelium	Liver	HNF1A [Search on ENCODE]

First Previous 1 Next Last

Chromosome:
☒ Chromosome option
chr1
chr2
chr3
chr4
chr5

Factor:
Cistrome Factor option

Show 10 entries Search: 1003124

TFBS rank of VIS on chr4	Location	DVID	Transcription Factor Binding Site	Detail
296	99028288-99028424	1003124	24	View

Chromosome:
Chromosome option

Cistrome Factor:
☒ Cistrome Factor option
Histone Modification
Transcription Factor Binding Site
Chromatin Accessibility

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Show 10 entries Search: HNF1A

Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y	
HNF1A	14	54	5	22	8	2	4	5	20	21	4	7	2	1	7	50	7	3	7	1			1	3	7

DVID	Searched Factor Name	Overlaps Number	Chromosome	Begin Location	End Location	Disease
1000023	HNF1A	5	chr4	49657675	49657676	Hepatocellular carcinoma
1000021	HNF1A	5	chr4	49657548	49657549	Hepatocellular carcinoma
1000020	HNF1A	5	chr4	49657558	49657559	Hepatocellular carcinoma

C

Target Gene
Correlation between gene expression and fraction of infiltrating immune cells

HBV 76 4979 5033

Target gene	Virus	Gene Function	Literature Count	Description	Immune	Detail
METAP1	HBV		2	D L T	Immune	View

Gene Information

Gene Name: METAP1
Gene ID: 23173
Gene Full Name: Methionyl aminopeptidase 1
Gene Alias: KIAA0094, MAP1A, MetAP1A
Transcripts: ENSG00000164024
Virus: HBV
Gene Type: Protein coding

HPA Location Info: Cytosol;Nucleoplasm;Plasma membrane; L
HPA Protein Class Info: Enzymes, Plasma proteins, Predicted intracellular proteins
HGNC ID: HGNC:15789
VEGA ID: OTTHUMG00000161231
OMIM ID: 610151
String ID: 9606.ENSPP00000296411
Drugbank ID: BE0004117

> Dataset: GSE114783

Lollipop of METAP1 expression and immune infiltrating cell level correlation result in HBV+ Liver cirrhosis samples

Macrophages M2
Plasma cells
Mast cells activated
NK cells resting
T cells CD8
T cells regulatory Tregs
NK cells activated
B cells memory
Plasma cells
B cells naive
Macrophages M0
Mast cells resting
Mast cells activated
Macrophages M2
T cells follicular helper
Macrophages M1
Eosinophils
Dendritic cells resting
T cells CD4 naive
T cells CD4 memory activated
Dendritic cells activated
T cells gamma delta
Neutrophils

Pearson
-0.50
-0.25
0.00
0.25
0.50

Pvalue
0.75
0.50
0.25