



## OPEN ACCESS

## EDITED BY

Naveen Kumar,  
ICAR-National Institute of High Security  
Animal Diseases (ICAR-NIHSAD), India

## REVIEWED BY

Rajinder Gupta,  
Technical University Dresden, Germany  
Hongkai Wu,  
The First Affiliated Hospital of Guangzhou  
Medical University, China

## \*CORRESPONDENCE

Rama Shankar  
✉ ramashan@msu.edu  
Bin Chen  
✉ chenbi12@msu.edu

RECEIVED 17 March 2025

ACCEPTED 28 July 2025

PUBLISHED 13 August 2025

## CITATION

Shankar R, Paithankar S, Gupta S and  
Chen B (2025) Detection of viral  
contamination in cell lines using  
ViralCellDetector.  
*Front. Microbiol.* 16:1595180.  
doi: 10.3389/fmicb.2025.1595180

## COPYRIGHT

© 2025 Shankar, Paithankar, Gupta and Chen.  
This is an open-access article distributed  
under the terms of the [Creative Commons  
Attribution License \(CC BY\)](#). The use,  
distribution or reproduction in other forums is  
permitted, provided the original author(s) and  
the copyright owner(s) are credited and that  
the original publication in this journal is cited,  
in accordance with accepted academic  
practice. No use, distribution or reproduction  
is permitted which does not comply with  
these terms.

# Detection of viral contamination in cell lines using ViralCellDetector

Rama Shankar<sup>1\*</sup>, Shreya Paithankar<sup>1</sup>, Suchir Gupta<sup>1</sup> and  
Bin Chen<sup>1,2,3\*</sup>

<sup>1</sup>Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, MI, United States, <sup>2</sup>Department of Pharmacology and Toxicology, College of Human Medicine, Michigan State University, East Lansing, MI, United States, <sup>3</sup>Department of Computer Science and Engineering, College of Engineering, Michigan State University, East Lansing, MI, United States

**Background and aims:** Cell lines are widely used in biomedical research to investigate various biological processes, including gene expression, cancer progression, and drug responses. However, cross-contamination with bacteria, mycoplasma, and viruses remains a persistent challenge. While the detection of bacterial and mycoplasma contamination is relatively straightforward, identifying viral contamination is more difficult. To address this issue, we developed ViralCellDetector, a tool designed to detect viral contamination by mapping RNA-seq data to a comprehensive viral genome library.

**Methods:** ViralCellDetector processes RNA-seq data from any host species by first aligning reads to the host reference genome, followed by mapping the unmapped reads to the NCBI viral genome database. Viral presence is determined using stringent criteria based on the number of mapped reads and viral genome coverage. To further enable the detection of viral contamination from unknown sources, we identified host genes that are differentially expressed during viral infection and used these markers to train a machine learning model for classification.

**Results:** Using ViralCellDetector, we found that approximately 10% (110 samples) of RNA-seq datasets involving MCF7 cells were likely contaminated with viruses. The tool demonstrated high sensitivity in detecting viral sequences. Furthermore, the machine learning model effectively distinguished infected from non-infected samples based on human gene expression profiles, achieving an AUC of 0.91 and an accuracy of 0.93.

**Conclusion:** Our mapping-based approach enables robust detection of viral contamination in RNA-seq data from any host organism, while the marker-based approach accurately identifies viral infections specifically in human cell lines. This capability can help researchers detect and avoid the use of contaminated cell lines, thereby improving the reliability of experimental outcomes.

## KEYWORDS

cell lines, viral contamination, bacterial contamination, differentially expressed genes, RNA-seq data, random forest, machine learning

## Introduction

Cell lines are invaluable tools in biomedical research as they enable scientists to investigate disease mechanisms and develop new treatments. They are widely used to study a wide range of biological processes, including cell signaling, gene expression, and drug metabolism. Additionally, cell lines play a crucial role in assessing the efficacy and toxicity of new drugs prior to vivo testing. These cell lines can be derived from various sources, such as

patient-derived cancer cells, human primary cells, or cells from animals. Immortalized cancer cell lines are often shared across laboratories and research studies. However, contamination with bacteria and mycoplasma is a significant concern (Corral-Vázquez et al., 2017). To mitigate these contamination and ensure experimental integrity, it is essential to follow established best practice guidelines for cell line maintenance and handling (Baust et al., 2017; Reid, 2017).

In addition to bacterial and mycoplasma contamination, viral contamination is a significant concern in cell line-based research, though it is much challenging to detect. Viral contamination can originate from the environment or from the original tissue source used to establish the cell lines. Although human cell lines may potentially harbor latent or active virus, there are limited approaches available for detecting these viruses (Shioda et al., 2018; Cheval et al., 2019; Uphoff et al., 2019; Dolskiy et al., 2020). Many existing approaches are designed to primarily detect only a small subset of well-known pathogenic viruses, such as cytomegalovirus (CMV), Epstein–Barr virus (EBV), human herpesvirus 6 (HHV-6), HHV-7, human polyomavirus BK (BKV), human polyomavirus JC (JCV), human adenovirus (ADV), human parvovirus B19 (B19V), hepatitis B virus (HBV), human T-cell leukemia virus type 1 (HTLV-1), HTLV-2, human immunodeficiency virus 1 (HIV-1), HIV-2, hepatitis A virus (HAV), and hepatitis C virus (HCV). These detection methods, whether PCR-based or reliant on specific viral sequence reads, are tailored for individual viruses, making it challenging to apply them broadly to research scenarios.

Recently, several algorithms have been developed to detect viral integration into the human genome using sequencing data (Kostic et al., 2011; Chen Y. et al., 2013; Li et al., 2013; Naeem et al., 2013; Schelhorn et al., 2013; Wang et al., 2013, 2015; Xia et al., 2019; Selitsky et al., 2020). Among these, VirTect (Xia et al., 2019) is highly advanced and can detect virus integration sites based on whole transcriptome sequencing (RNA-Seq) data. Additionally, an AI-Enabled-Virus-Detect tool (Ghorbani et al., 2024) has been reported to detect viral sequences using a BLAST-based approach. In this method, the authors perform *de novo* assembly of unmapped host reads to identify potential viral contigs. However, this approach presents a limitation, as even a minimal overlap of two reads can lead to contig formation, increasing the likelihood of false-positive viral identifications. Despite these advances, there remains a lack of broadly applicable tools for detecting viral contaminations across various host reference samples. Many existing mapping-based approaches assume that viral sequence reads are abundantly present in sequence data. However, widely adopted polyA-based library preparation protocols primarily enrich for host transcripts with polyadenylated tails, which may limit the detection of viral transcripts. Therefore, there is a critical need for a robust and scalable tool that can detect viral contaminations across diverse sample types using standard RNA-seq data.

To address these limitations, we developed a ViralCellDetector tool that leveraged the ultrafast STAR aligner (Dobin et al., 2013), followed by the BWA aligner, to incorporate all known viruses from the NCBI virus database. Our tool first maps RNA-seq data to any given host reference genome and transcriptome and subsequently maps unmapped reads to a comprehensive viral genome database. To enhance the specificity, users can exclude viruses not affecting their respective host and applied stringent criteria to accurately identify true viruses. As demonstration of the tool's utility, we analyzed RNA-seq data from more than 1,000 experiments performed on MCF7 cells to identify the potentially infected samples. Additionally, we implemented a machine learning approach

using this dataset to identify host gene expression biomarkers associated with viral contamination. This biomarker-based approach exhibits robust performance and is independent of viral species and library preparation protocols. However, it performs best in samples where viral contamination has induced alterations in host cell biology. We anticipate that the integration of both mapping and biomarker-based approaches will empower researchers to effectively identify potential viral contaminations in their cell lines using standard RNA-seq data, thus ensuring the reliability and reproducibility of their experimental results.

## Methods

The overall workflow of ViralCellDetector is illustrated in Figure 1a. The fastq files from samples were used as input in ViralCellDetector pipeline. The first step of ViralCellDetector is to align the sequencing reads to the reference genome and its corresponding annotation file (e.g., Hg38 transcriptome with ENSEMBL GRCh38.p3 annotation in case of human cells) using STAR aligner (Dobin et al., 2013). Unmapped reads were subsequently mapped to the viral genome database available from NCBI<sup>1</sup> using BWA (Li and Durbin, 2009). The advantage of using BWA over STAR aligner for viral genomes is that it can retain alignments even when only one read from a paired end read maps to the viral genome. To reduce false positives, we calculate the total number of reads mapped to each viral genome. Additionally, we estimate the percentage of genome coverage to more accurately identify true viral presence.

## Dataset

We downloaded the SRA sequencing data from GEO<sup>2</sup>. All the sequencing data were further converted into fastq files using the SRAtoolkit<sup>3</sup>. The fastq files were used as input in the ViralCellDetector to detect the putative viruses.

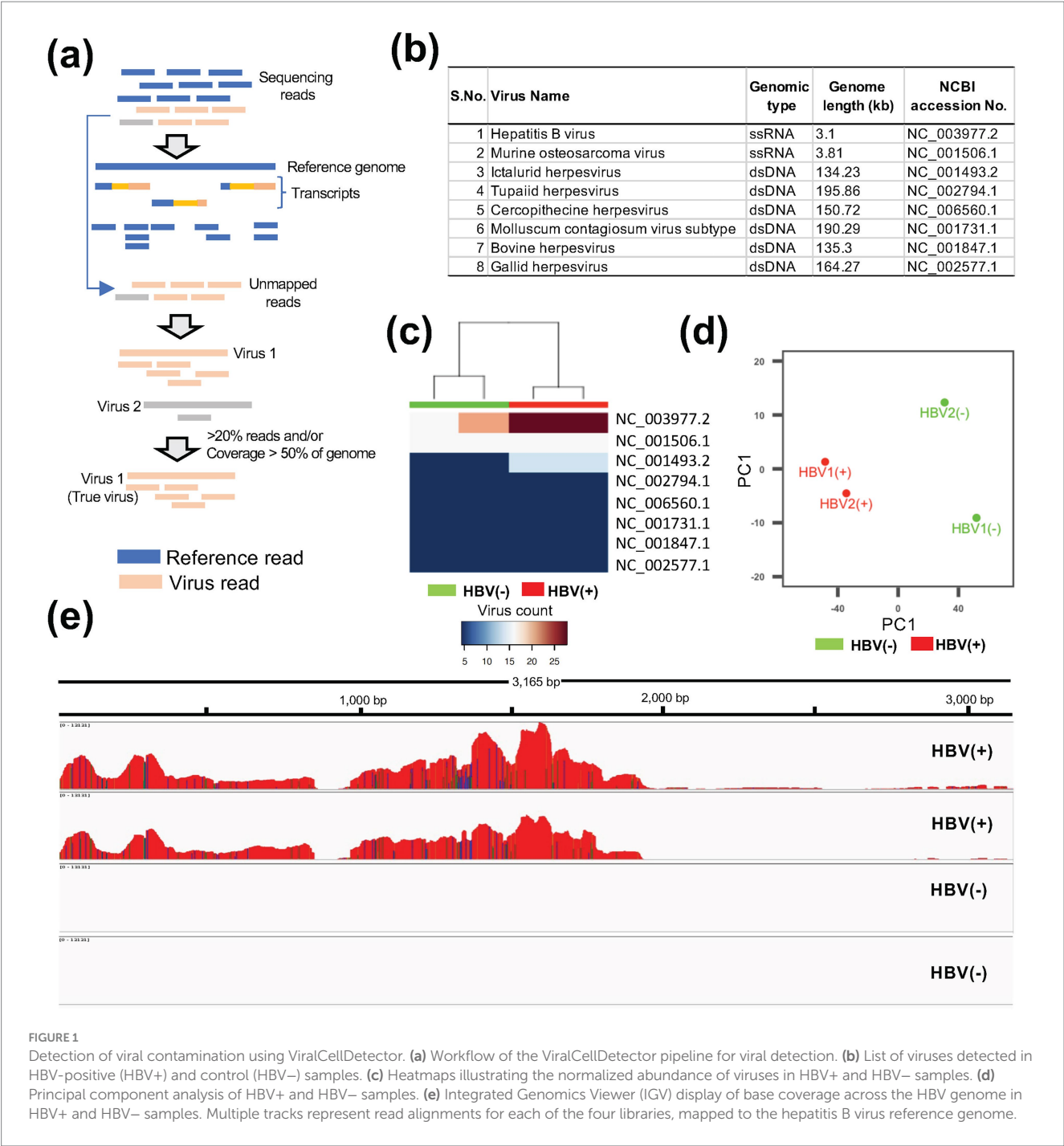
## Data processing

ViralCellDetector is designed to accept raw sequencing data as input. It separates paired end reads into two sets: those mapping to the host reference genome, and those that potentially belonging to viral genomes. Reads aligning to the host genome and transcriptome were first mapped using the ultrafast STAR aligner with default parameters (Dobin et al., 2013). Unmapped reads left after host reference mapping were then aligned to the NCBI viral genome database using the BWA aligner with default parameters (Li and Durbin, 2009). The viral genome sequences can be downloaded from the NCBI virus database (see Footnote 1). After viral sequence download from NCBI, the endogenous retroviruses may be removed before indexing the viral genomes. Once the viral reads were mapped to the viral genome, we applied the following criteria to filter out false positives: (1) at least 20% reads should be mapped to the viral genome (based on our positive sample mapping), (2) the continuous

1 <https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/>

2 <https://www.ncbi.nlm.nih.gov/geo/browse/?view=series>

3 <https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit>



**FIGURE 1** Detection of viral contamination using ViralCellDetector. **(a)** Workflow of the ViralCellDetector pipeline for viral detection. **(b)** List of viruses detected in HBV-positive (HBV+) and control (HBV-) samples. **(c)** Heatmaps illustrating the normalized abundance of viruses in HBV+ and HBV- samples. **(d)** Principal component analysis of HBV+ and HBV- samples. **(e)** Integrated Genomics Viewer (IGV) display of base coverage across the HBV genome in HBV+ and HBV- samples. Multiple tracks represent read alignments for each of the four libraries, mapped to the hepatitis B virus reference genome.

coverage of the viral genome should be more than 50%, and (3) the virus should be known to infect the host.

### RNA-seq analysis

All the sequencing reads were mapped on Hg38 transcriptome using the ENSEMBL GRCh38.p3 annotation with the STAR aligner (Dobin et al., 2013). Gene counts obtained from the alignment were used for identification of differentially expressed (DE) genes. The edgeR package (Robinson et al., 2010) was employed to quantify DE genes based on following criteria:  $\log_2$  fold change  $\geq 1$  or  $\leq -1$  and an adjusted  $p$ -value (False Discover Rate)  $\leq 0.01$ . DE genes were identified between the two

groups (control and infection) in two independent datasets. Gene ontology enrichment and KEGG pathways analyses were performed using the enrichR (Chen E. Y. et al., 2013; Kuleshov et al., 2016). DE genes involved in viral infection-related pathways were further utilized as features in machine learning model. Data visualization was conducted using the ggplot2 package in R. All the analyses were performed on R (version 4.2.1).

### Random forest for feature selection and prediction

Random forest (RF) is a robust and widely used machine learning algorithm based on bagging techniques, known for its consistent

performance across various classification tasks (Subash et al., 2022; Wadood et al., 2022). It consists of an ensemble of independent decision trees, where the final prediction is determined by aggregating the outputs from all trees (Montes et al., 2021). Key hyperparameters influencing model performance, namely 'ntree' (number of trees), 'mtry' (number of variables tried at each split), and 'nodesize' (minimum size of terminal nodes); can be tuned by the user (Oukawa et al., 2022). In our study, we set 'ntree' to 600 and 'mtry' to 8 for feature selection and classification. The dataset was randomly partitioned into 80% training and 20% testing subsets. We employed a recursive feature elimination (RFE) to select informative features. To evaluate the stability and generalizability of the model, we implemented 10-fold cross-validation. Features that achieved an accuracy of  $\geq 0.9$  across cross-validation folds were retained for final model training and testing. Model performance on the test dataset was assessed using accuracy, area under the curve (AUC), sensitivity, and specificity. Additionally, to address class imbalance, we applied downsampling of the majority class and repeated 10-fold cross-validation to ensure the robustness and reliability of the classifier.

## Results

### Validation of the mapping-based viral contamination detection approach

The workflow of the pipeline is provided in Figure 1a. To validate the tool, we analyzed publicly available RNA-seq datasets containing known viral infection/contamination. One dataset included Hepatitis B virus (HBV) infected samples (GSE65485), comprising two control and two infected samples. The second dataset contained SARS-CoV-2 infected cell lines (GSE187420), consisting of three control and three infected samples. In the HBV dataset, we identified a total of eight viruses in the list (Figure 1b). However, upon examining the expression patterns in the control and infected samples, the HBV-infected samples exhibited the highest abundance of HBV reads compared to control samples (Figure 1c). Furthermore, a clear transcriptional distinction between control and infected samples was observed (Figure 1d). Visualization of read coverage across viral genomes revealed that only the HBV genome achieved more than 90% coverage (Figure 1e), confirming the presence

of HBV infection. Similarly, in the SARS-CoV-2 dataset (GSE187420) (Xing et al., 2022), infected samples displayed a high proportion of unmapped reads when aligned to the human reference transcriptome (Supplementary Table S1), along with a substantial number of reads mapping to the SARS-CoV-2 genome (Supplementary Figure S1). These findings support the effectiveness of our pipeline in accurately detecting viral contaminations in RNA-seq data derived from cell lines.

### Comparison with other tools in simulated data

To further evaluate the applicability of the ViralCellDetector tool, we generated six simulated paired-end FASTQ files set by combining varying proportions of viral reads (10, 20, 30, 40, and 50%) from different viruses with host-derived FASTQ reads. We then applied the ViralCellDetector pipeline to these datasets. Additionally, we compared its performance to another method, referred to as the "AI-enabled tool" (Ghorbani et al., 2024). Notably, this tool utilizes *de novo* assembly of unmapped reads from the host genome, followed by virus detection using the Basic Local Alignment Search Tool (BLAST). Despite being labeled as AI-based, the method relies solely on BLAST and does not implement any artificial intelligence algorithms.

We assessed the performance of ViralCellDetector and the AI-enabled tool by computing sensitivity and false positive rate (FPR) across all simulated datasets. We also recorded the time required by each method to process each sample. ViralCellDetector consistently outperformed the AI-enabled tool in both sensitivity and FPR (Figures 2a,b). Moreover, it completed analyses significantly faster, delivering a putative virus list in much less time (Figure 2c), highlighting the efficiency of the ViralCellDetector tool.

### Detection of viral contamination from unlabeled dataset

To detect viral contaminations in unlabeled (without known contamination) RNA-seq data, we collected MCF7 cell line datasets from multiple studies (SRP065220, SRP142602, and SRP163132). The

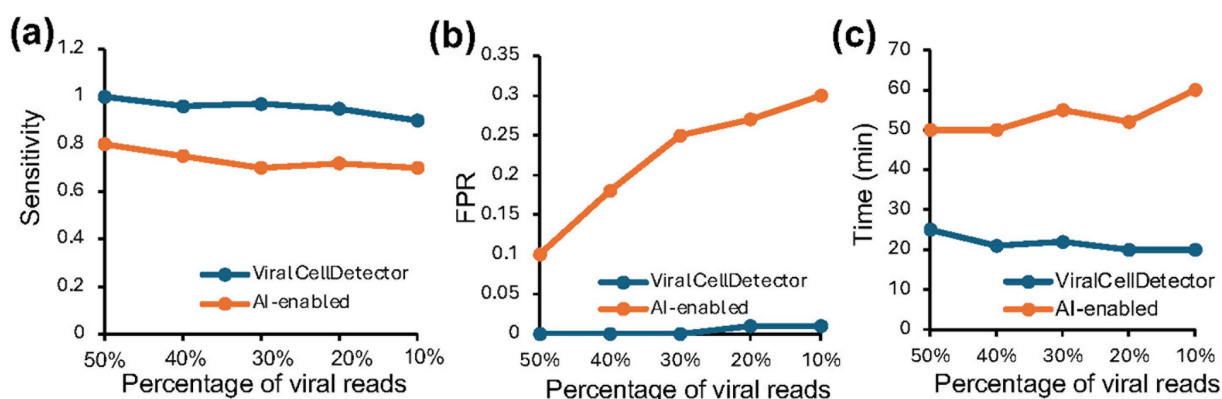


FIGURE 2  
Comparison of ViralCellDetector with other tools on simulated data containing varying proportions of viral reads mixed with host sequences. (a,b) Comparison of sensitivity and false positive rate (FPR) between ViralCellDetector and other AI-enabled tools reported to detect viruses in samples. (c) Time taken to process the simulated data.



sequencing data in these studies were generated using a RiboMinus approach, enabling the capture of all expressed mRNAs from humans and non-human sources, including viruses (Zhao et al., 2018; Chen et al., 2020). Specifically, the SRP065220 dataset includes 12 MCF7 samples, SRP142602 includes four samples, and SRP163132 includes 128 samples. The raw sequencing reads were downloaded and first mapped to the human reference genome and transcriptome. For each sample, the proportion of unmapped reads was then calculated. We observed that eight samples exhibited more than 45% unmapped reads (Figure 3a), suggesting potential viral contaminations. This observation is consistent with previous report (Yuan et al., 2021) and our data (Supplementary Table S1), which demonstrate that viral infections/contaminations are often associated with elevated levels of unmapped reads in RNA-seq data. The unmapped reads were then mapped to the viral genome using ViralCellDetector pipeline, resulting in the detection of a total of 22 viruses (Supplementary Figure S2). After applying stringent criteria, we identified six viruses: three dsDNA viruses (Woodchuck hepatitis virus, BeAN 58,058 virus, and *Eptesicus fuscus* gamma herpes virus) and three ssRNA viruses (Human immunodeficiency virus 1, Encephalomyocarditis virus and Human endogenous retrovirus) (Figure 3b).

To further investigate the presence of these viruses in MCF7 cells, we collected RNA-seq data from over 1,021 MCF7 samples. Raw reads were first mapped to the human reference transcriptome. The unmapped reads of each sample were then analyzed to assess the expression levels of six identified viruses. Across all the samples, we observed considerable variation in the expression levels of these viruses (Supplementary Figure S2). To identify the experiments with exceptionally high viral count, we ranked the viral expression values and applied an upper quartile threshold to select samples with markedly elevated viral counts (Supplementary Figure S3). Using this approach, we identified approximately 110 samples (~10% of total) that exhibited significantly higher viral expression counts compared to the remaining samples (Figures 3c,d), suggesting the likely presence of one or more of these viruses. However, we noted that around 50% of these samples originated from studies (e.g., GSE106694, GSE103520, GSE100099, GSE63189, and GSE67295) in which viruses were intentionally introduced through transfection as part of the experimental design. Therefore, not all detected viral signals necessarily indicate unintended contamination. Furthermore, the presence of viral sequences does not always imply a biological impact on host gene expression. In many cases, the detected viral reads simply reflect the viral load in the samples, without indicating any downstream functional consequences in host cells.

## Identify DE genes involved in viral contamination in different cell lines

Accurately quantifying viral loads in RNA-seq data generated using polyA enrichment protocols is challenging. Therefore, we focused on identifying host-derived biomarkers that reflect the biological impact of viral infection. We analyzed data from two independent studies (GSE198398 and GSE187420) involving two different cell lines (Vero 6 and Calu-3). In the GSE198398 dataset, Vero E6 cells were infected with HCoV-OC43 (OC43), while in GSE187420 dataset, Calu-3 cells were infected with SARS-CoV-2. Differential expression analysis identified 953 and 4,101 DE genes in the two datasets, respectively (Figures 4a,b).

These genes were further subjected to KEGG pathway enrichment analysis, which revealed enrichment of pathways associated with viral infections, including Influenza A, Herpes simplex virus 1, Hepatitis B, Hepatitis C, and Epstein-Barr virus (Figures 4c,d). From this analysis, we selected a list of 300 DE genes that were consistently active in infected samples compared to controls. These genes were then used as input features for feature selection and classification using a Random Forest (RF) model.

## Biomarker discovery for viral contamination in cell lines

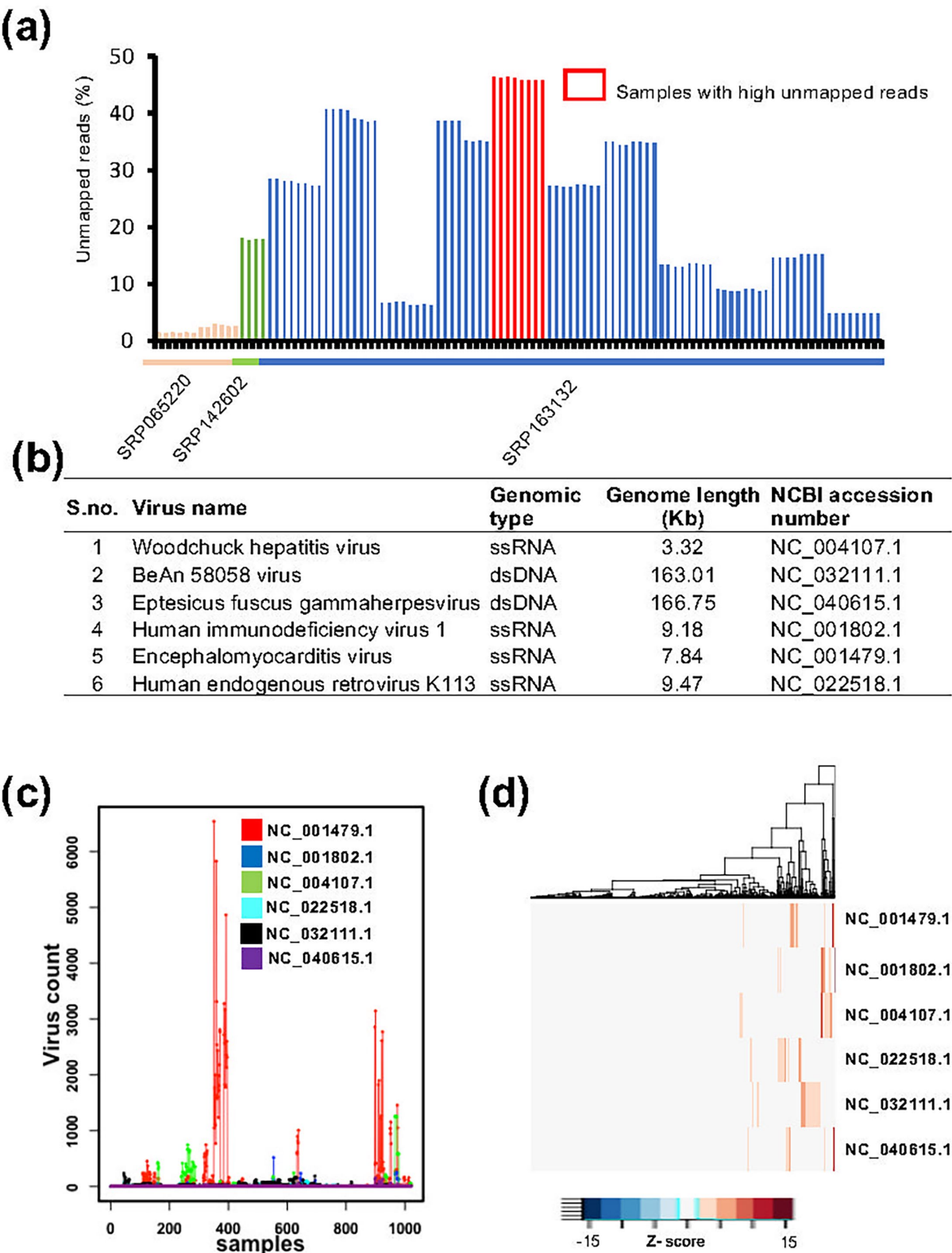
After identifying the infected samples in MCF7 cell lines, we categorized them into infected (110 samples) and control (911 samples) groups. Subsequently, we collected the normalized expression matrix of all samples (1,021 samples) from ARCHS4<sup>4</sup>. To establish a gene expression-based biomarker for viral contamination, we applied a RF machine learning approach for feature selection and prediction. The DE genes identified from two independent viral infection datasets, which were enriched in viral infection-related pathways and consistently activity in infected samples, were used as candidate features. Feature selection was performed using the training dataset based on feature importance values generated by the RF model. Top 12 features (*JAK2*, *ZNF614*, *ZNF613*, *PPP2R2A*, *ZNF595*, *GADD45B*, *ZNF433*, *ITGA5*, *ZNF627*, *NFKBIA*, *PIK3R3*, and *ZNF333*) were selected, each demonstrating classification accuracy greater than 0.92 and a kappa value exceeding 0.44 (Figures 5a,b).

To gain deeper biological insights into these features, we performed protein-protein interaction analysis, gene ontology (GO) enrichment, and KEGG pathway analysis. Among the selected genes, four (*PIK3R3*, *NFKBIA*, *JAK2*, and *GADD45B*) demonstrated direct interaction with each other (Figure 5c). However, all 12 genes were found to be involved in pathways related to viral infection, cancer, and cell signaling (Figure 5d). These features were subsequently used for predictive modeling. Using RF classification based on these selected features we successfully distinguished infected from control samples, achieving an area under the curve (AUC) of 0.91, an accuracy of 0.93, a sensitivity of 0.99, and a specificity of 0.60 (Figure 5e). In addition, we applied downsampling of the majority class along with 10-fold cross-validation to evaluate the robustness of the model. The model demonstrated a mean accuracy of 0.87 and an average area under the precision-recall curve (AUPRC) of 0.75. These results indicate that the selected features are highly effective for classifying infected and non-infected samples across different cell lines with very high sensitivity. The relatively lower specificity suggests that the biological impact of viral contamination may not uniform across all infected cells.

## Discussion

Cell lines are invaluable tools in biomedical research, providing powerful systems to investigate disease mechanisms and support the development of new therapeutic strategies. However, the use of cell

<sup>4</sup> <https://maayanlab.cloud/archs4/data.html>



**FIGURE 3**  
Detection of viral contamination using ViralCellDetector on MCF7 cell line data. **(a)** Percentage of unmapped reads across samples from three independent studies. The highest percentage of unmapped reads (45%) was observed in eight samples from SRP163132 study. These samples were analyzed for viral detection, resulting in the identification of six viruses. **(b)** List of viruses detected in MCF7 samples by ViralCellDetector. **(c)** Expression levels of the six identified viruses across 1,021 MCF experiments, revealing a subset of samples with viral contamination. **(d)** Heatmaps illustrating elevated expression of the six viruses across various MCF7 experiments.

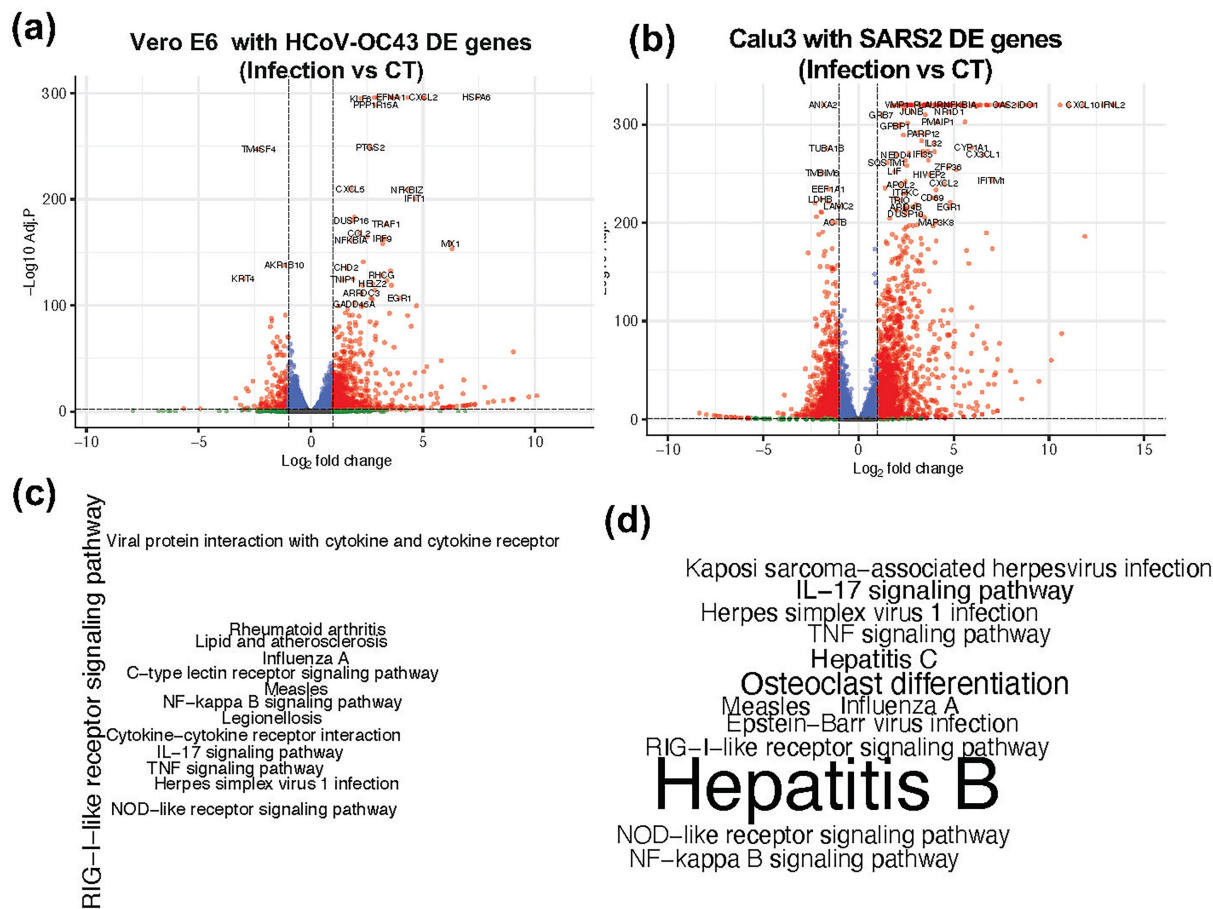


FIGURE 4

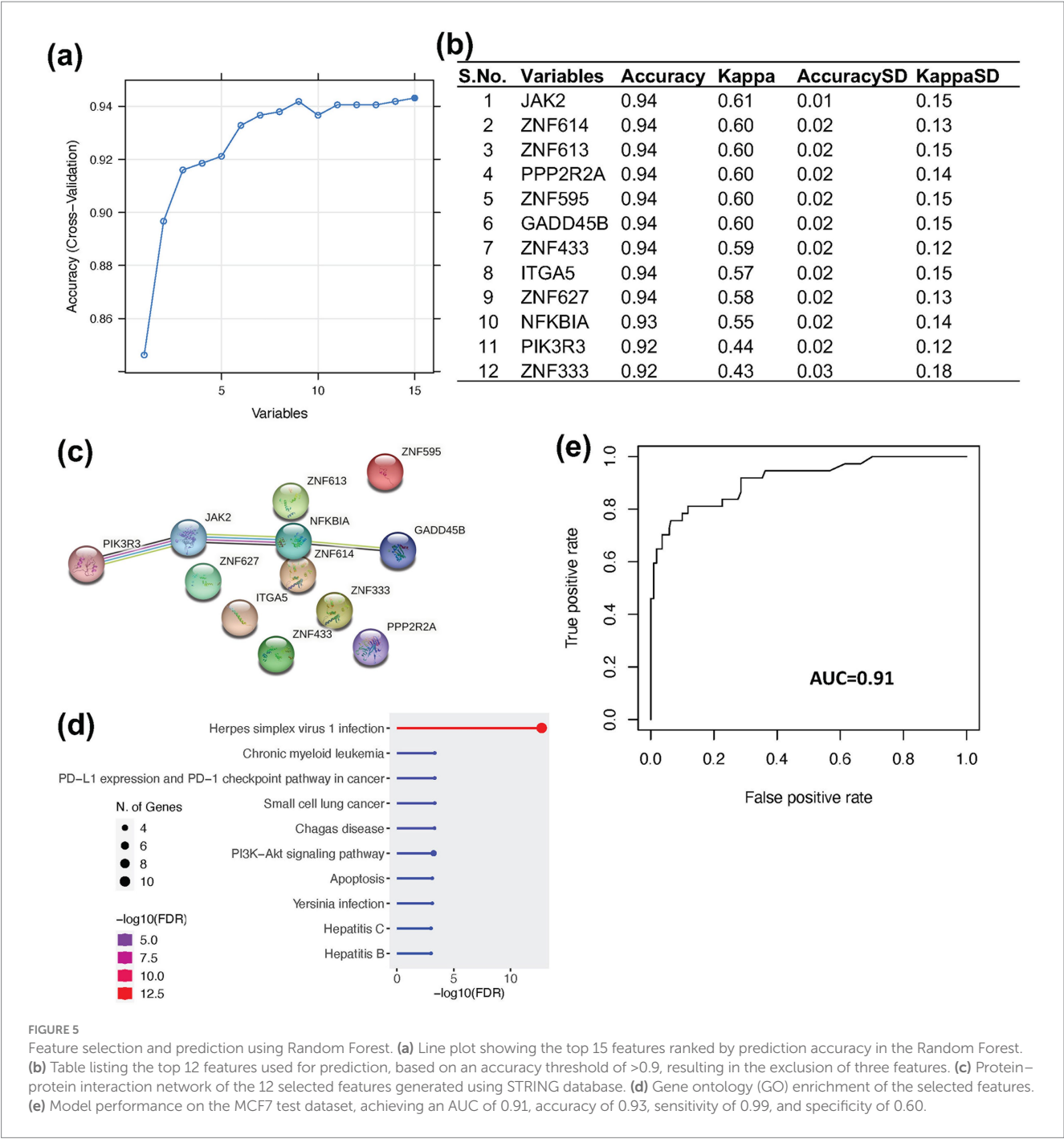
Differentially expressed genes and their KEGG pathway enrichment in two different cell lines. **(a,b)** Volcano plot showing the differentially expressed (DE) genes in Infected vs. Control (CT) samples from **(a)** Vero 6 cells and **(b)** Calu3 cells. **(c,d)** KEGG pathway enrichment analysis of these DE genes in **(c)** Vero 6 cells and **(d)** Calu3 cells, highlighting pathways associated with viral infection. The font size of the pathway labels corresponds to the  $-\log_{10}$  (Adjusted  $p$ -value).

lines is not without challenges, with contamination representing a persistent concern. While bacterial and mycoplasma contamination can often be mitigated through adherence to established laboratory protocols and are relatively straightforward to detect, viral contamination poses a far greater challenge. Viruses are not readily detectable by standard microscopy and often remain undetected without targeted assays.

To address this limitation, we developed ViralCellDetector, a computational pipeline designed to detect viral contaminations directly from RNA-seq data. ViralCellDetector accepts raw sequencing reads as input, systematically filters host-mapped reads, and identifies potential viral sequences by mapping unmapped reads to a comprehensive viral genome database. The pipeline outputs a list of candidate viruses, offering a practical and scalable approach to screen for viral contamination across diverse cell lines and experimental datasets. Additionally, comparison of ViralCellDetector with other tools revealed its better efficacy and faster processing. Building on ViralCellDetector pipeline, we classified a large set of MCF7 cell line experiments into infected and non-infected categories. To further explore the host response to viral contaminations, we identified DE genes from two independent datasets involving two different cell lines

and viral infections. The DE genes associated with viral infection were subsequently used in a machine learning approach to identify a gene expression-based biomarker capable of distinguishing infected and non-infected samples. Through this approach, we identified a set of 12 genes as biomarkers that consistently differentiated infected samples.

In our comprehensive analysis of MCF7 cell lines, we detected contamination with three DNA viruses and three RNA viruses. The detected RNA viruses included Encephalomyocarditis virus (EMCV), a member of the Picornaviridae family associated with encephalitis and myocarditis in various mammalian species (Griffin, 2011), and Hepatitis virus, known for its etiological role in hepatocellular carcinoma (Liu, 2020); Among the DNA viruses, we identified *Eptesicus fuscus* gammaherpesvirus (EfGHV), a gammaherpesvirus primarily infecting bat species (Subudhi et al., 2018); BeAn 58,058 virus, a poxvirus isolated from *Oryzomys* rodents (Wanzeller et al., 2017); Human endogenous retrovirus K113 (HERV-K113), which represents an integrated component of the human genome with potential transcriptional activity under certain conditions; and HIV-1, a well-characterized human retrovirus. Importantly, several of these viruses, such as EMCV, HIV-1, and HERV-K113, are frequently used as experimental models in virology and molecular biology



laboratories, which may increase the risk of unintentional carryover into unrelated cell culture systems. However, it is important to note that some of these experiments may have intentionally introduced viral infections as part of their experimental design, and thus not all detected contaminations necessarily reflect unintended viral contamination. Additionally, not all the samples with viral contamination will have impact on host gene expression, and detection of those viruses revealed the virus sequence load in the samples rather any downstream biological effect on the host cells.

Furthermore, the DE analysis in two independent datasets identified a set of genes consistently upregulated in infected samples and associated with viral infection pathways. These genes provided a

valuable foundation for feature selection in our machine learning approach. Using this strategy, we identified a list of 12 distinct genes that serve as potential biomarkers capable of accurately classifying infected and non-infected cell lines. Interestingly, these genes are involved in multiple pathways related to viral infection, cellular signaling, and cancer, highlighting their functional relevance in viral entry, replication, and pathogenesis. Several of these genes have already been reported to support viral contamination in cells (Shen et al., 2014; Liu et al., 2017; Talledo et al., 2012). Their strong association with viral infection and key regulatory pathways likely contributes to their high predictive performance. This biomarker offers a valuable tool for the detection of potential viral contamination



in cell line experiments, enhancing quality control and enabling researchers to identify unrecognized viral contaminations that may otherwise confound experimental outcomes.

One limitation of the present study is that biomarker discovery was performed using the MCF7 cell line, for which labeled datasets with confirmed viral contamination are not publicly available. To address this, we downloaded and processed fastq files to label over 1,000 MCF7 samples. Additionally, we implemented 10-fold cross-validation and a downsampling approach to assess the robustness of our identified biomarkers. Another limitation is the use of stringent criteria in our pipeline for detecting viral contamination specifically, requiring >20% of reads to be mapped with >50% viral genome coverage. While this enhances the reliability of detection, it may lead to the omission of low-level viral nucleic acids, which may nonetheless have minimal biological impact on the host. Moreover, we validated the tool using simulated datasets and found that even with as little as 10% viral reads, ViralCellDetector was able to detect the viruses with high sensitivity, supporting the effectiveness of our approach.

Future research directions should be considered to further strengthen this work. Our exploratory study revealed a surprisingly widespread occurrence of viral contamination in commonly used cell lines. Further investigations are warranted to elucidate the underlying causes of this phenomenon, which may include laboratory handling practices, viral persistence, or cross-contamination. Such insights will be critical for the scientific community to improve experimental rigor and minimize the risk of unintended viral contamination in future research.

In conclusion, we have developed a computational pipeline and validated the reliability of this tool using the simulated data and compared with identified a biomarker panel for the detection of viral contamination in cell lines using readily available RNA-seq data. This approach offers researchers a valuable tool to investigate unexpected experimental outcomes and implement appropriate corrective measures to ensure data integrity and reproducibility. Incorporating such screening strategies should be considered an essential component of best laboratory practice guidelines.

## Data availability statement

The source code of ViralCellDetector is freely available at <https://github.com/Bin-Chen-Lab/ViralCellDetector>. The data used in this study are public datasets. Those data can be accessible from Gene Omnibus Expression (<https://www.ncbi.nlm.nih.gov/geo/browse/?view=series>) or Archs4 (<https://maayanlab.cloud/archs4/data.html>).

## References

- Baust, J. M., Buehring, G. C., Campbell, L., Elmore, E., Harbell, J. W., Nims, R. W., et al. (2017). Best practices in cell culture: an overview. *In Vitro Cell. Dev. Biol. Anim.* 53, 669–672. doi: 10.1007/s11626-017-0177-7
- Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* 14:128. doi: 10.1186/1471-2105-14-128
- Chen, L., Yang, R., Kwan, T., Tang, C., Watt, S., Zhang, Y., et al. (2020). Paired rRNA-depleted and polyA-selected RNA sequencing data and supporting multi-omics data from human T cells. *Sci Data* 7:376. doi: 10.1038/s41597-020-00719-4
- Chen, Y., Yao, H., Thompson, E. J., Tannir, N. M., Weinstein, J. N., and Su, X. (2013). Virusseq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue. *Bioinformatics* 29, 266–267. doi: 10.1093/bioinformatics/bts665
- Cheval, J., Muth, E., Gonzalez, G., Couplier, M., Beurdeley, P., Cruveiller, S., et al. (2019). Adventitious virus detection in cells by high-throughput sequencing of newly synthesized RNAs: unambiguous differentiation of cell infection from carryover of viral nucleic acids. *mSphere* 4, e00298–e00219. doi: 10.1128/mSphere.00298-19
- Corral-Vázquez, C., Aguilar-Quesada, R., Catalina, P., Lucena-Aguilar, G., Ligerio, G., Miranda, B., et al. (2017). Cell lines authentication and mycoplasma detection as

## Author contributions

RS: Data curation, Visualization, Software, Investigation, Conceptualization, Writing – review & editing, Project administration, Writing – original draft, Formal analysis, Supervision. SP: Data curation, Writing – review & editing. SG: Data curation, Writing – review & editing, Validation, Resources. BC: Conceptualization, Supervision, Writing – review & editing.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The research was supported by R01GM134307, K99HD111575, R01GM145700, and the MSU Global Impact Initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of sponsors.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2025.1595180/full#supplementary-material>

- minimum quality control of cell lines in biobanking. *Cell Tissue Bank*. 18, 271–280. doi: 10.1007/s10561-017-9617-6
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635
- Dolskiy, A. A., Grishchenko, I. V., and Yudkin, D. V. (2020). Cell cultures for virology: usability, advantages, and prospects. *Int. J. Mol. Sci.* 21:7978. doi: 10.3390/ijms21217978
- Ghorbani, A., Rostami, M., and Guzzi, P. H. (2024). AI-enabled pipeline for virus detection, validation, and SNP discovery from next-generation sequencing data. *Front. Genet.* 15:15. doi: 10.3389/fgene.2024.1492752
- Griffin, D. E. (2011). Viral encephalomyelitis. *PLoS Pathog.* 7:e1002004. doi: 10.1371/journal.ppat.1002004
- Kostic, A. D., Ojesina, A. I., Pedamallu, C. S., Jung, J., Verhaak, R. G., Getz, G., et al. (2011). Pathseq: software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* 29, 393–396. doi: 10.1038/nbt.1868
- Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Li, J.-W., Wan, R., Yu, C. S., Co, N. N., Wong, N., and Chan, T. F. (2013). Viralfusionseq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* 29, 649–651. doi: 10.1093/bioinformatics/btt011
- Liu, L. (2020). Clinical features of hepatocellular carcinoma with hepatitis B virus among patients on Nucleos(t) ide analog therapy. *Infect. Agent. Cancer* 15:8. doi: 10.1186/s13027-020-0277-y
- Liu, X., Happel, C., and Ziegelbauer, J. M. (2017). Kaposi's sarcoma-associated herpesvirus micrornas target GADD45B to protect infected cells from cell cycle arrest and apoptosis. *J. Virol.* 91:e02045-16. doi: 10.1128/JVI.02045-16
- Montes, C., Kapelan, Z., and Saldarriaga, J. (2021). Predicting non-deposition sediment transport in sewer pipes using random forest. *Water Res.* 189:116639. doi: 10.1016/j.watres.2020.116639
- Naem, R., Rashid, M., and Pain, A. (2013). Readscan: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics* 29, 391–392. doi: 10.1093/bioinformatics/bts684
- Oukawa, G. Y., Kreci, P., and Targino, A. C. (2022). Fine-scale modeling of the urban heat island: a comparison of multiple linear regression and random forest approaches. *Sci. Total Environ.* 815:152836. doi: 10.1016/j.scitotenv.2021.152836
- Reid, Y. A. (2017). Best practices for naming, receiving, and managing cells in culture. *In Vitro Cell. Dev. Biol. Anim.* 53, 761–774. doi: 10.1007/s11626-017-0199-1
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616
- Schelhorn, S.-E., Fischer, M., Tolosi, L., Altmüller, J., Nürnberg, P., Pfister, H., et al. (2013). Sensitive detection of viral transcripts in human tumor transcriptomes. *PLoS Comput. Biol.* 9:e1003228. doi: 10.1371/journal.pcbi.1003228
- Selitsky, S. R., Marron, D., Hollern, D., Mose, L. E., Hoadley, K. A., Jones, C., et al. (2020). Virus expression detection reveals RNA-sequencing contamination in TCGA. *BMC Genomics* 21:79. doi: 10.1186/s12864-020-6483-6
- Shen, W., Shen, W.-j., Dong, R., Chen, G., and Zheng, S. (2014). MicroRNA-222 modulates liver fibrosis in a murine model of biliary atresia. *Biochem. Biophys. Res. Commun.* 446, 155–159. doi: 10.1016/j.bbrc.2014.02.065
- Shioda, S., Kasai, F., Watanabe, K., Kawakami, K., Ohtani, A., Iemura, M., et al. (2018). Screening for 15 pathogenic viruses in human cell lines registered at the JCRB cell Bank: characterization of *in vitro* human cells by viral infection. *R. Soc. Open Sci.* 5:172472. doi: 10.1098/rsos.172472
- Subash, T., David, A., ReetaJanetSurekha, S., Gayathri, S., Samuelkamaleshkumar, S., Magimairaj, H. P., et al. (2022). Comparing algorithms for assessing upper limb use with inertial measurement units. *Front. Physiol.* 13:1023589. doi: 10.3389/fphys.2022.1023589
- Subudhi, S., Rapin, N., Dorville, N., Hill, J. E., Town, J., Willis, C. K. R., et al. (2018). Isolation, characterization and prevalence of a novel Gammaherpesvirus in *Eptesicus fuscus*, the north American big brown bat. *Virology* 516, 227–238. doi: 10.1016/j.virol.2018.01.024
- Talledo, M., López, G., Huyghe, J. R., Verdonck, K., González, E., Clark, D., et al. (2012). Possible implication of *NFKB1A* and *NKG2D* genes in susceptibility to HTLV-1-associated myelopathy/tropical spastic paraparesis in Peruvian patients infected with HTLV-1. *J. Med. Virol.* 84, 319–326. doi: 10.1002/jmv.22255
- Uphoff, C. C., Pommerenke, C., Denkmann, S. A., and Drexler, H. G. (2019). Screening human cell lines for viral infections applying RNA-Seq data analysis. *PLoS One* 14:e0210404. doi: 10.1371/journal.pone.0210404
- Wadood, A., Ajmal, A., Junaid, M., Rehman, A. U., Uddin, R., Azam, S. S., et al. (2022). Machine learning-based virtual screening for STAT3 anticancer drug target. *Curr. Pharm. Des.* 28, 3023–3032. doi: 10.2174/1381612828666220728120523
- Wang, Q., Jia, P., and Zhao, Z. (2013). Virusfinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS One* 8:e64465. doi: 10.1371/journal.pone.0064465
- Wang, Q., Jia, P., and Zhao, Z. (2015). VERSE: a novel approach to detect virus integration in host genomes through reference genome customization. *Genome Med.* 7:2. doi: 10.1186/s13073-015-0126-6
- Wanzeller, A. L. M., Souza, A. L. P., Azevedo, R. D. S. D. S., Júnior, E. S., Filho, L. C. F., Oliveira, R. S., et al. (2017). Complete genome sequence of the BeAn 58058 virus isolated from *Oryzomys* sp. rodents in the Amazon region of Brazil. *Genome Announc.* 5, e01575–e01516. doi: 10.1128/genomea.01575-16
- Xia, Y., Liu, Y., Deng, M., and Xi, R. (2019). Detecting virus integration sites based on multiple related sequencing data by VirTect. *BMC Med. Genet.* 12:19. doi: 10.1186/s12920-018-0461-8
- Xing, J., Shankar, R., Ko, M., Zhang, K., Zhang, S., Drelich, A., et al. (2022). Deciphering COVID-19 host transcriptomic complexity and variations for therapeutic discovery against new variants. *iScience* 25:105068. doi: 10.1016/j.isci.2022.105068
- Yuan, Z., Fan, X., Zhu, J. J., Fu, T. M., Wu, J., Xu, H., et al. (2021). Presence of complete murine viral genome sequences in patient-derived xenografts. *Nat. Commun.* 12:2031. doi: 10.1038/s41467-021-22200-5
- Zhao, S., Zhang, Y., Gamini, R., Zhang, B., and von Schack, D. (2018). Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.* 8:4781. doi: 10.1038/s41598-018-23226-4