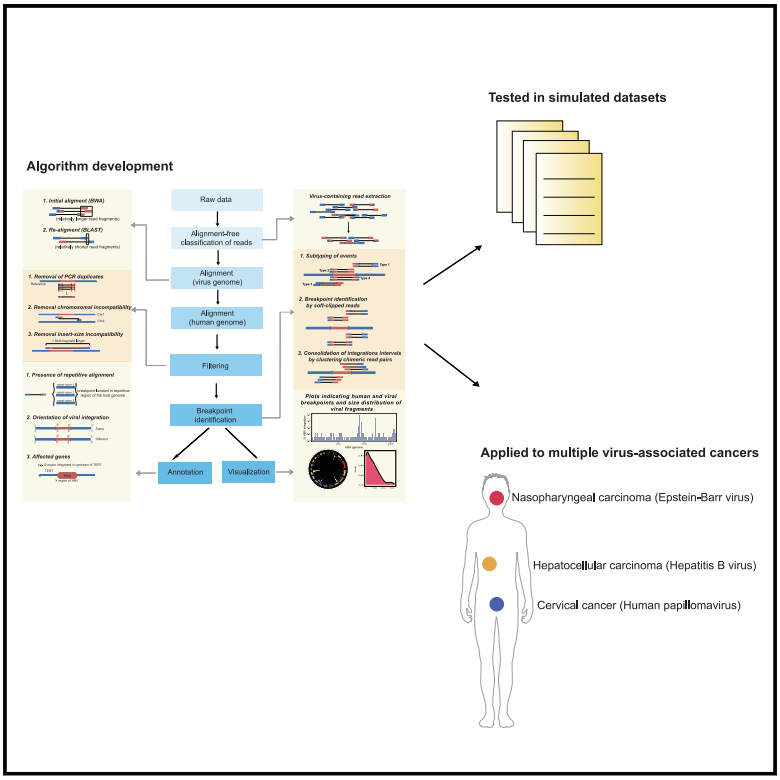
Cell Reports
Methods

Article

AVID enables sensitive and accurate viral integration detection across human cancers

Graphical abstract



Highlights

- AVID detects viral integration sites with high sensitivity and accuracy
- We demonstrate AVID's performance with simulated and experimentally validated datasets
- AVID is applicable to different types of virus-associated human cancers

Authors

Xueying Lyu, Russell Wing-Yeung Mok, Hoi-Ying Chan, ..., Irene Oi-Lin Ng, Loey Lung-Yi Mak, Daniel Wai-Hung Ho

Correspondence

lungyi@hku.hk (L.L.-Y.M.), dwhho@hku.hk (D.W.-H.H.)

In brief

Lyu et al. develop an algorithm called AVID (accurate viral integration detector) for sensitive and accurate detection of viral integration. It is widely applicable to different oncovirus-associated human cancers, supporting the mechanistic investigation of virus-associated tumorigenesis.

Lyu et al., 2025, Cell Reports Methods *5*, 101007
March 24, 2025
https://doi.org/10.1016/j.crmeth.2025.101007

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

AVID enables sensitive and accurate viral integration detection across human cancers

Xueying Lyu,[1,2] Russell Wing-Yeung Mok,[1,2] Hoi-Ying Chan,[1,2] Tina Suoangbaji,[1,2] Qian Li,[1,2] Fanhong Zeng,[1,2] Renwen Long,[1,2] Irene Oi-Lin Ng,[1,2] Loey Lung-Yi Mak,[1,3,*] and Daniel Wai-Hung Ho[1,2,4,*]

[1]State Key Laboratory of Liver Research, The University of Hong Kong, Pokfulam, Hong Kong, China
[2]Department of Pathology, School of Clinical Medicine, The University of Hong Kong, Pokfulam, Hong Kong, China
[3]Department of Medicine, School of Clinical Medicine, The University of Hong Kong, Pokfulam, Hong Kong, China
[4]Lead contact
*Correspondence: lungyi@hku.hk (L.L.-Y.M.), dwhho@hku.hk (D.W.-H.H.)
https://doi.org/10.1016/j.crmeth.2025.101007

MOTIVATION Oncogenic virus infection is strongly associated with increased cancer risk. Some of these viruses are able to integrate into the host genome, resulting in genome stability damage, structure variation, transcription or protein dysfunction, and further carcinogenesis. While some existing tools have been developed to detect viral integration events, no single tool was found to be sufficiently adequate in all settings. Therefore, we are motivated to develop a tool for sensitive and accurate viral integration detection that can hopefully overcome the existing limitations and provide a better solution for the task.

SUMMARY

Oncovirus infection is a key etiological risk factor of human cancers, which triggers virus integration in the host genome. Viral integration can lead to structural variation, gene dysfunction, and genome instability, promoting tumorigenesis. To support the investigation of virus-associated cancer and improve the detection of virus infection, we developed an algorithm called AVID (accurate viral integration detector) for viral integration detection. AVID was built by overcoming the existing detection limitations, enhancing sensitivity and accuracy, and expanding additional functions of viral integration detection. The performance of AVID was estimated in simulated datasets and experimentally validated datasets compared with other tools. To demonstrate its wide applicability, we also tested AVID on viral integration detection in multiple oncovirus-associated human cancers, including hepatocellular carcinoma (HCC), cervical cancer, and nasopharyngeal carcinoma. Taken together, our study developed an improved and applicable tool for viral integration detection and visualization to facilitate further exploration of virus-infected diseases.

INTRODUCTION

Oncovirus infection is a major risk factor for human cancers.[1] Some viruses can integrate into the human genome, leading to structural variation known as viral genomic integration.[2] Viral integration may result in genome instability, disruption of human genes, aberrant human gene expression, and/or expression of chimeric oncogenic proteins, which contribute to consequential tumorigenesis.[3] Infections of common oncoviruses, e.g., hepatitis B virus (HBV), human papillomavirus (HPV), and Epstein-Barr virus (EBV), are known to cause viral integration events, and they are one of the causal factors of liver cancer, cervical cancer and nasopharyngeal carcinoma, and lymphoproliferative diseases, respectively. In particular, hepatocellular carcinoma (HCC) is one of the most lethal cancers and also an extremely difficult-to-treat cancer.[4–7] Among the identified etiological risk factors for HCC, including chronic viral infections (HBV and hepatitis C virus), chronic alcohol consumption, and steatotic liver disease, chronic HBV infection accounts for around 50% of HCC cases worldwide. One of the distinctive features of HBV DNA is that it can integrate into the human genome, which in turn disrupts endogenous tumor suppressors and other regulatory genes or enhances the activity of proto-oncogenes. The imbalance of the overall oncogenic and tumor-suppressive signals may result in enhanced cell survival, proliferation, and reduced apoptosis and lead to HCC development even in non-cirrhotic livers.[8]

Recently, with the emergence of next-generation sequencing (NGS), different NGS-based approaches have been used to provide a more unbiased and comprehensive survey of HBV integration.[9] Due to the huge amount of data generated by NGS, data analyses rely on specific computational tools to derive useful findings. Different tools have emerged to detect viral integration and determine the exact breakpoint position of the human-virus chimera. Existing tools differ in the underlying algorithms,

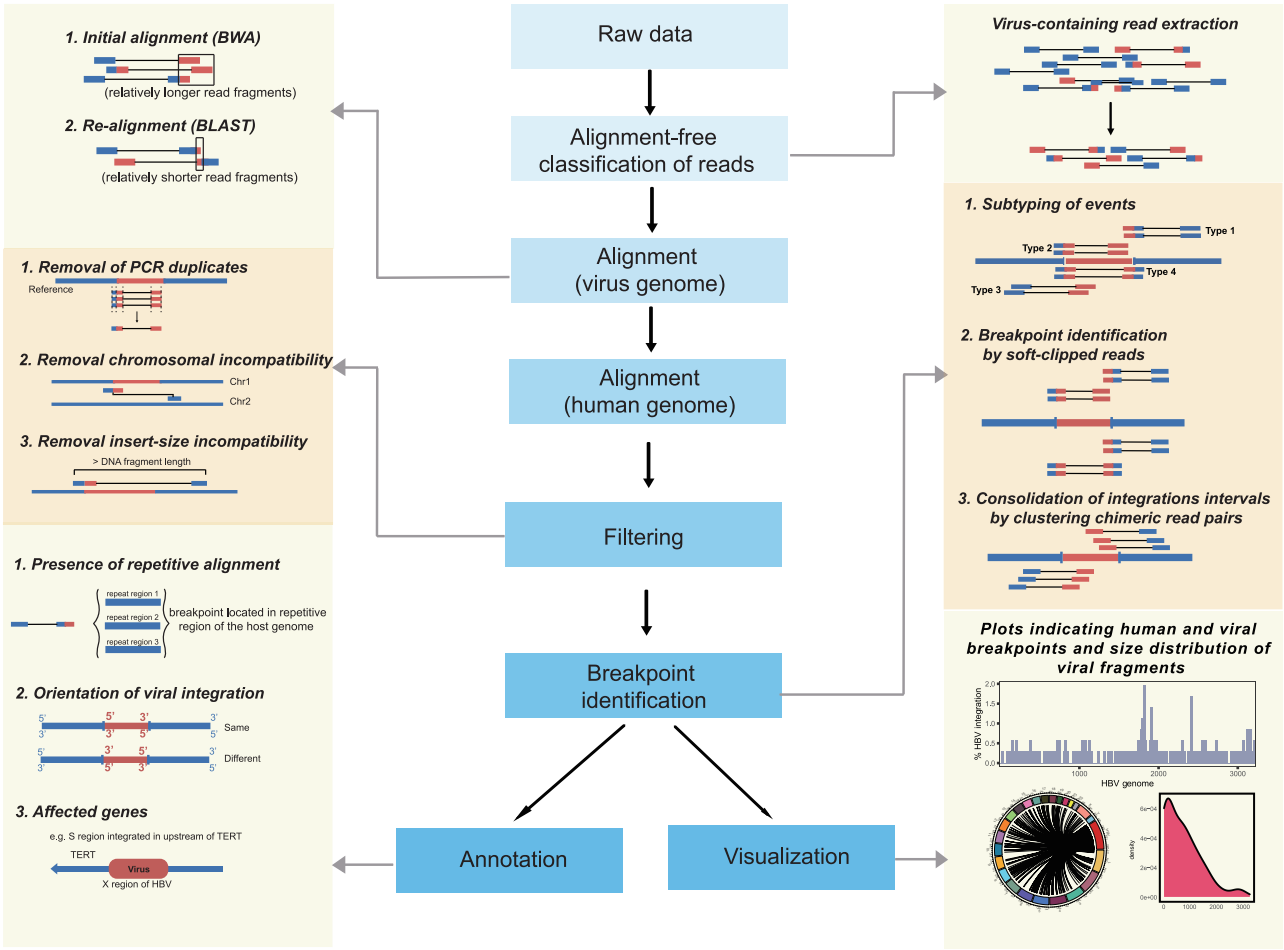


Figure 1. Development of AVID algorithm

The middle image is a flowchart, and the two side images are strategies for improving viral integration detection and additional functions.

pinpointing different unique aspects/features of viral integration. A technical overview of the various viral integration detection tools regarding their underlying algorithms and properties was reported.[3] In general, major steps of viral integration detection include (1) read alignment and extraction of chimeric pairs and/or soft-clipped reads, (2) quality control of aligned and extracted reads, and (3) integration of candidate discovery and determination of integration breakpoints.[3] Notably, some tools (e.g., Exogene,[10] BATVI,[11] and SurVirus[12]) have specific strategies for more efficient alignment, better extraction of informative reads, better handling of repetitive sequences, and/or rescue of unmapped reads. Nevertheless, no single tool was found to be sufficiently accurate in all settings.[13] Hence, there is a strong need to develop better computational tools for studying viral integration, including HBV.

We previously developed the viral integration detection algorithm called Virus-Clip.[14] There are several key limitations regarding the issues of single-read consideration (single reads are evaluated separately no matter if they are generated by single-end or pair-end sequencing), no information on the direction/orientation of integration, no removal of PCR duplicates, high false positive rates, and low sensitivity toward short chimeric fragments (soft-clipped portions are too short to be successfully aligned). Pinpointing the shortcomings of Virus-Clip as mentioned above, we have adopted specific measures to develop a completely new and enhanced algorithm (namely AVID [accurate viral integration detector]). In our study, comprehensive performance evaluation was undertaken using different benchmark datasets. We generated simulation data under different scenarios, including data at different sequencing coverages, with different numbers of viral integration sites, with viral integrations of different viral insert sizes, and with multiple subclones of malignant cells. This could provide empirical evidence to suggest the best analytical tools in different scenarios. To further evaluate the validity of our algorithm, we also assess the tools using patient-derived and experimentally validated datasets.

RESULTS

Development of new AVID algorithm for sensitive and accurate viral integration detection and visualization

We noted the various limitations of our previously developed Virus-Clip algorithm,[3,14] namely the issues of single-read

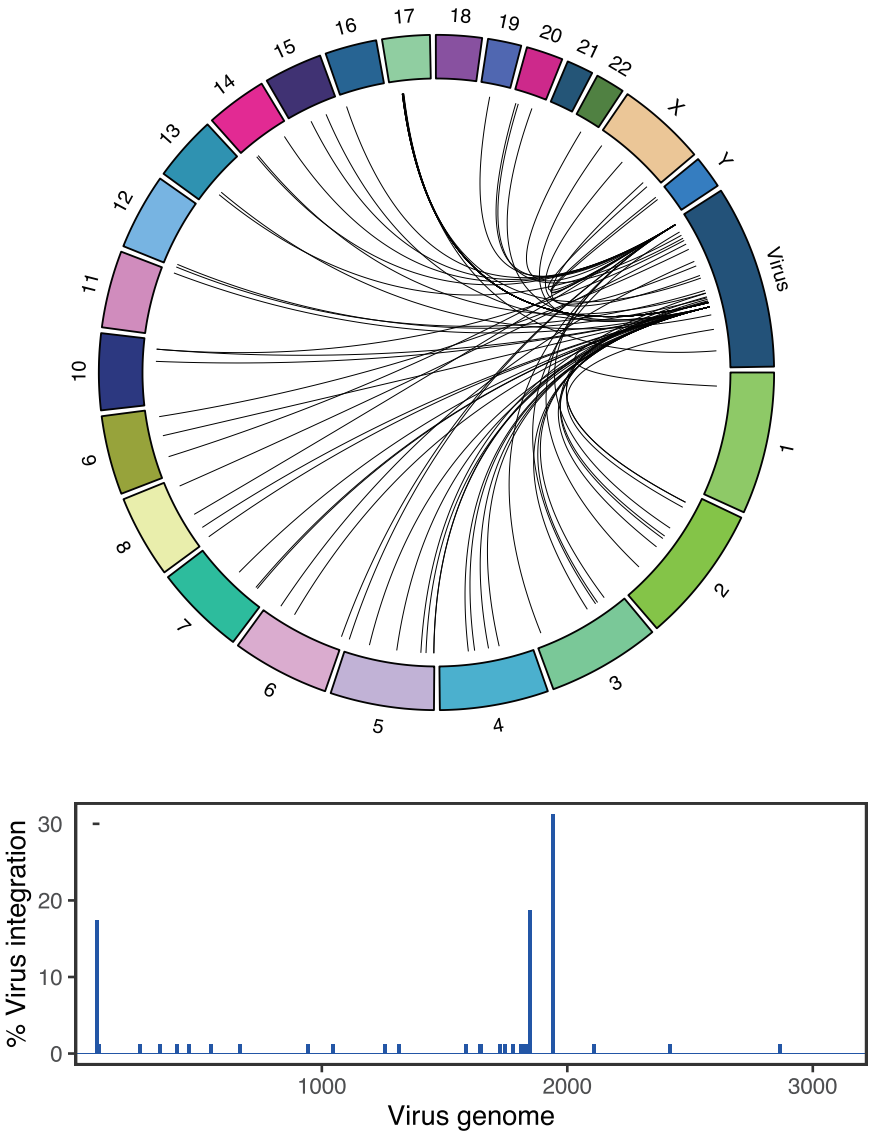


Figure 2. Example of visualization plots

For the Circos plot on the top, each curved line in the center represents a viral integration event, and the two ends indicate the corresponding breakpoint positions in human and viral genomes. The histogram on the bottom summarizes the frequency distribution of viral genomic breakpoints.

consideration (single reads are evaluated separately no matter if they are generated by single-end or paired-end sequencing), no consideration on the direction/orientation of integration, no removal of PCR duplicates, high false positive rates, and low sensitivity toward short chimeric fragments (soft-clipped portions are too short to be successfully aligned). Addressing these shortcomings, we developed a completely new and enhanced algorithm (namely AVID). First, AVID takes advantage of paired-end reads to better detect viral integration. It utilizes both a soft-clipped read (a chimeric read that simultaneously has both human and virus portions) and a chimeric read pair (pair-end reads with one of them fully mapped to human and the other one fully mapped to virus) to achieve more sensitive identification of viral integration events. While soft-clipped reads can provide information on the exact integration breakpoint positions, the chimeric read pairs only indicate the presence of integration events and their approximate genomic intervals. It also determines and subtypes the viral integration events according to the characteristics of the events. Second, it removes duplicated reads to reduce over-estimation of supporting read counts due to PCR duplicates and, hence, reduces false positive detections. Third, with the information derived from paired-end reads, i.e., a pair of reads, it determines the chromosome compatibility and viral insert size to filter the events and effectively enhances accuracy by further reducing false positives. Fourth, it reports the orientation of viral integration and the involvement of repetitive region alignment, helping users to confine the reported events. Fifth, AVID will perform re-alignment of soft-clipped reads in the case of low read coverage and short chimeric fragments to improve the chimeric fragment confirmation. Finally, it will perform clustering to collate supporting reads with common integration breakpoints or narrow down the potential human and virus breakpoints using chimeric read pairs. This will help to confine the breakpoint positions. The workflow and key features of the AVID algorithm are summarized in Figure 1. Apart from the aforementioned analytical functions, AVID can generate relevant visualization plots to indicate the landscapes of human and viral breakpoints (Figure 2). They provide a visual summarization of the viral integration events and their frequency distribution.

Evaluation of the AVID algorithm using simulated datasets

The complexity of viral integration detection hinges upon various factors, including the data coverage, number of integration events, length of the inserted viral fragments, and presence of intra-tumoral subclones of tumor cells. We simulated whole-genome sequencing data with designated viral integration events inserted at predefined values. The simulated events pinpointed a range of values for the aforementioned factors, and we subjected them to viral integration detection by AVID and other algorithms (Virus-Clip, SurVirus, BATVI, and Exogene).[3,10–12,14] Notably, AVID outperformed the others in different scenarios, as exemplified by the superior precision, recall, and F1 score achieved (Figures 3A and 3B; Tables S1, S2, S3, and S4). Besides, AVID can better detect orientation-specific virus

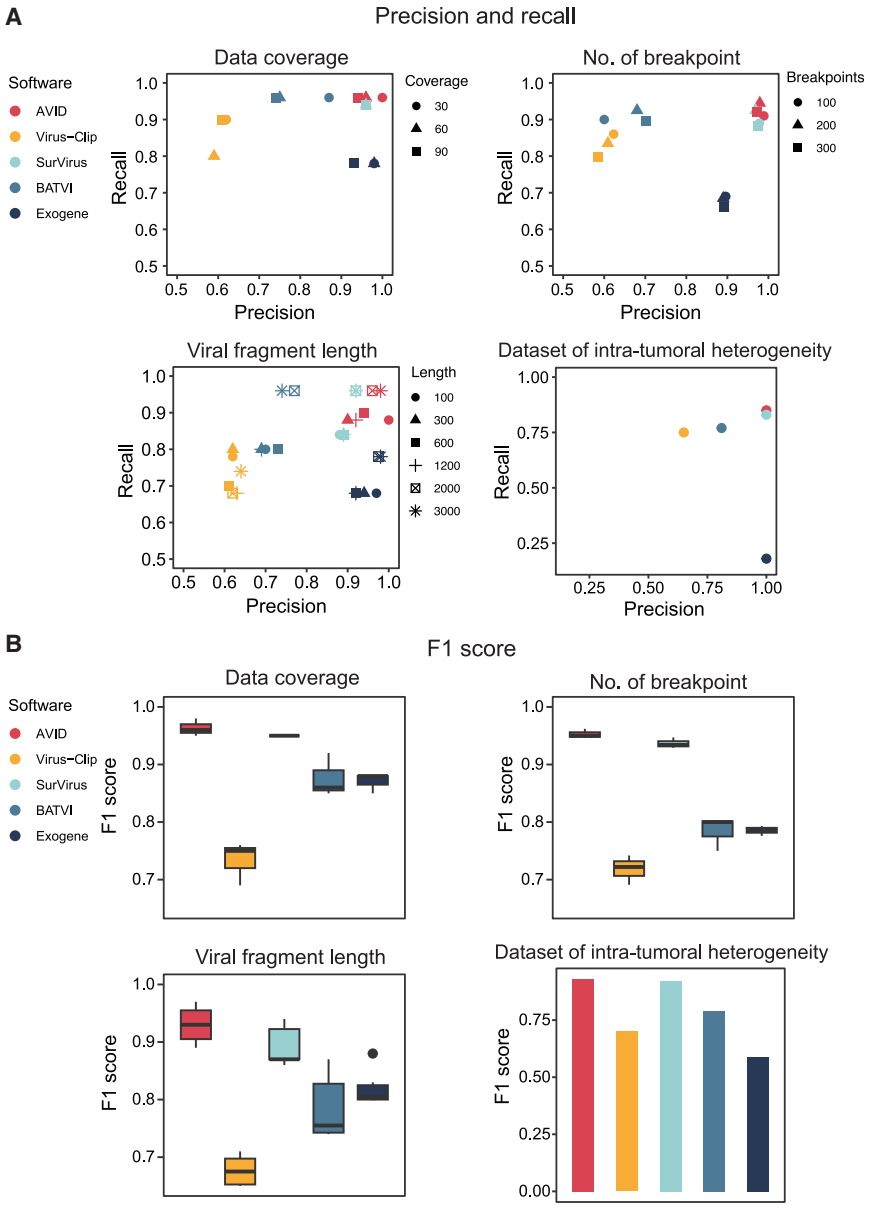


Figure 3. Performance of AVID in simulated datasets

In simulated datasets with the complexity of data coverage, the number of breakpoints, viral fragment length, and intra-tumoral heterogeneity, the performance of AVID was evaluated by precision, recall (A), and F1 score (B) and compared with the other existing tools. Error bars: mean ± SD. SD, standard deviation.

See also Tables S1, S2, S3, and S4.

integration as compared to SurVirus and BATVI (Exogene and Virus-Clip could not report the orientation of viral integration) (Figure 4; Tables S1, S2, S3, and S4). Moreover, upon comparing the computational memory usage and execution time using simulated datasets, the AVID algorithm had relatively low memory usage and shorter execution times, indicating its practicability (Figure 5; Tables S1, S2, S3, and S4).

Evaluation of the AVID algorithm using experimentally validated datasets

We additionally tested the algorithms using experimentally validated datasets. In the data by Sung et al. (EGA: ERP00196),[15] they reported 23 HBV integration events detected in HCC and confirmed by Sanger sequencing, whereas in the data by Hu et al. (SRA: SRP048813),[16] 11 confirmed HPV integration events in cervical cancer were documented. Viral integration detection tools were executed on these publicly available datasets. The AVID algorithm was able to detect most of the viral integration events while at the same time requiring relatively low memory usage and short execution times (Figures 6A and 6B; Table S5). Taken together, the various evaluation outcomes justified the use of the AVID algorithm in different aspects of our subsequent investigation in this study.

Application of AVID in viral integration detection of multiple oncoviruses

To confirm the broad applicability of AVID, it was applied to four different cohorts,[15–18] including the viral integration detection of HBV in HCC (EGA: ERP00196; SRA: SRP068532), EBV in nasopharyngeal carcinoma (SRA: SRP035573), and HPV in cervical cancer (SRA: SRP048813) (Figure 7). Generally, over 90% of patients were detected with viral integration in virus-enriched/captured sequencing data (Zhao et al. cohort and Hu et al. cohort) or whole-genome sequencing data (Sung et al. cohort). We could also detect EBV integration in about 55% (34 out of 61 cases) of patients with nasopharyngeal carcinoma using exome sequencing data (Lin et al. cohort), which only target the exonic regions, i.e., integration events at promoter or intronic regions may not be detected and probably cannot provide full coverage of the genome. Since whole-genome sequencing or virus-enriched/captured sequencing could provide a more comprehensive and systematic survey of the viral integration landscape, our findings indicate that the choice of sequencing platform may impact the efficiency of viral integration detection. Notably, in the two cohorts of HBV-associated HCC (Zhao et al. cohort and Sung et al. cohort), AVID was able to achieve detection rates of 92.5% (detected in 394 out of 426 cases) and 92.6% (detected in 75 out of 81 cases), respectively, which was more sensitive than the statistics reported in their original publications (76.9% and 91.4%, respectively)[15,17] (Figures 7A and 7B). Consistently,

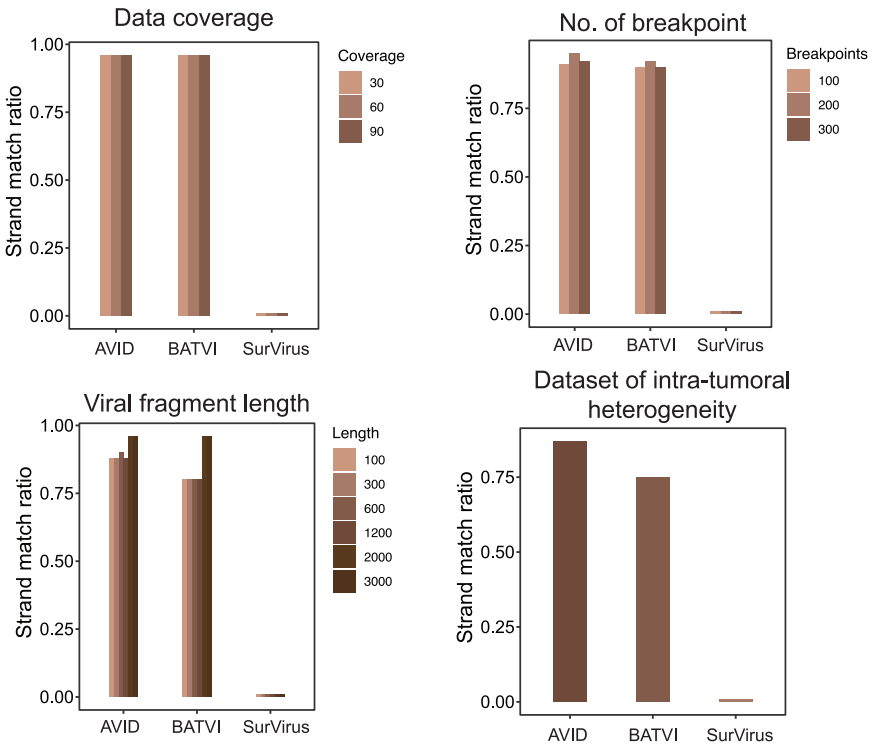


Figure 4. Performance of AVID in orientation-specific viral integration detection in simulated datasets compared with the other tools

Error bars: mean ± SD. SD, standard deviation.

telomerase reverse transcriptase (*TERT*) and *KMT2B* were the most frequent genes detected with HBV integration in both HBV-associated HCC cohorts, suggesting AVID could identify relevant and concordant results in biologically similar cohorts. Other well-known HBV-integrated genes, including *CCNE1*, *CCNA2*, and *ARID1B*, were also detected with events. Moreover, in the nasopharyngeal carcinoma cohort, we found EBV integration events in key genes for carcinogenesis (*KMT2D*, *SMG1*, *TP53*, and *APOBEC3A*), further indicating the relevance of the findings reported by AVID (Figure 7C). In the cohort of HPV-associated cervical cancer, the virus integration detection rate was 99.3% (detected in 134 out of 135 cases), which also clearly outperformed the detection of 76.3% in their original study[16] (Figure 7D). Apart from the frequent HPV integration events affecting genes reported in the original study (*FHIT*, *LRP1B*, *KLF5*, *DLG2*, and *SEMA3D*), we additionally detected a substantial amount of events affecting genes encoding for long intergenic non-protein-coding RNAs and microRNAs, as well as kinases (*RYK* and *SDK1*), substantiating the potentially important and oncogenic roles in viral integration event-driven cervical carcinogenesis.[19,20]

DISCUSSION

Oncoviruses integrated into the human genome can lead to tumorigenesis. To support the investigation of viral integration landscapes and the mechanistic delineation in different oncovirus-associated human cancers, we developed our AVID algorithm for sensitive and accurate viral integration detection. By pinpointing the limitations of our previously developed Virus-Clip algorithm, we have adopted specific measures to develop a completely new and enhanced AVID algorithm. Notably, we fixed different issues by improving sensitivity and accuracy and further extended various functionalities.

To comprehensively evaluate the performance of AVID, we generated simulated datasets with special focuses on the aspects of the complexity of data coverage, the number of integrations, and viral fragment length. They resulted in analyses with variable complexities from different perspectives. Notably, as the clonality of virus integration was associated with tumorigenesis,[21,22] to mimic the high heterogeneity and complexity of an HCC tumor, i.e., the presence of different subclones of malignant cells having their specific sets of molecular alterations, we simulated subclones of malignant cells with different aforementioned parameters and combined them into a single hybrid dataset. We believe this likely recapitulates the genuine biological condition that exists in HCC tumors. Encouragingly, AVID was able to outperform the other testing counterparts using all simulated datasets that pinpoint different underlying determining factors. More importantly, since evaluating different types of sequencing data and disease models of virus-associated human cancers may help to eliminate undesirable artifacts and bias, we applied AVID in different experimentally validated datasets (DNA sequencing data on HCC and RNA sequencing data on cervical cancer). These validation datasets further confirmed the superior capability of AVID in viral integration detection.

In our previous reports,[23,24] we adopted a target-panel sequencing approach to survey HBV integration in our patients' HBV-associated HCCs (n = 95). HBV integration at the human *TERT* gene promoter was frequent (35.8%, n = 34/95) in HCC tumors and was associated with increased *TERT* mRNA expression and more aggressive tumor behavior. The transcription activity of *TERT* was modulated by the orientation of HBV integration, suggesting its importance in leading to the functional consequence of viral-integrated human genes. Moreover, mitochondrial DNA is also a target of HBV integration, which may contribute to HCC development.[25,26] To this end, we built AVID with specific functionality to allow orientation-specific detection of viral integration. Moreover, AVID also provides a useful visualization capability to summarize the viral integration landscape in a graphical manner. Apart from HBV-associated HCC, AVID

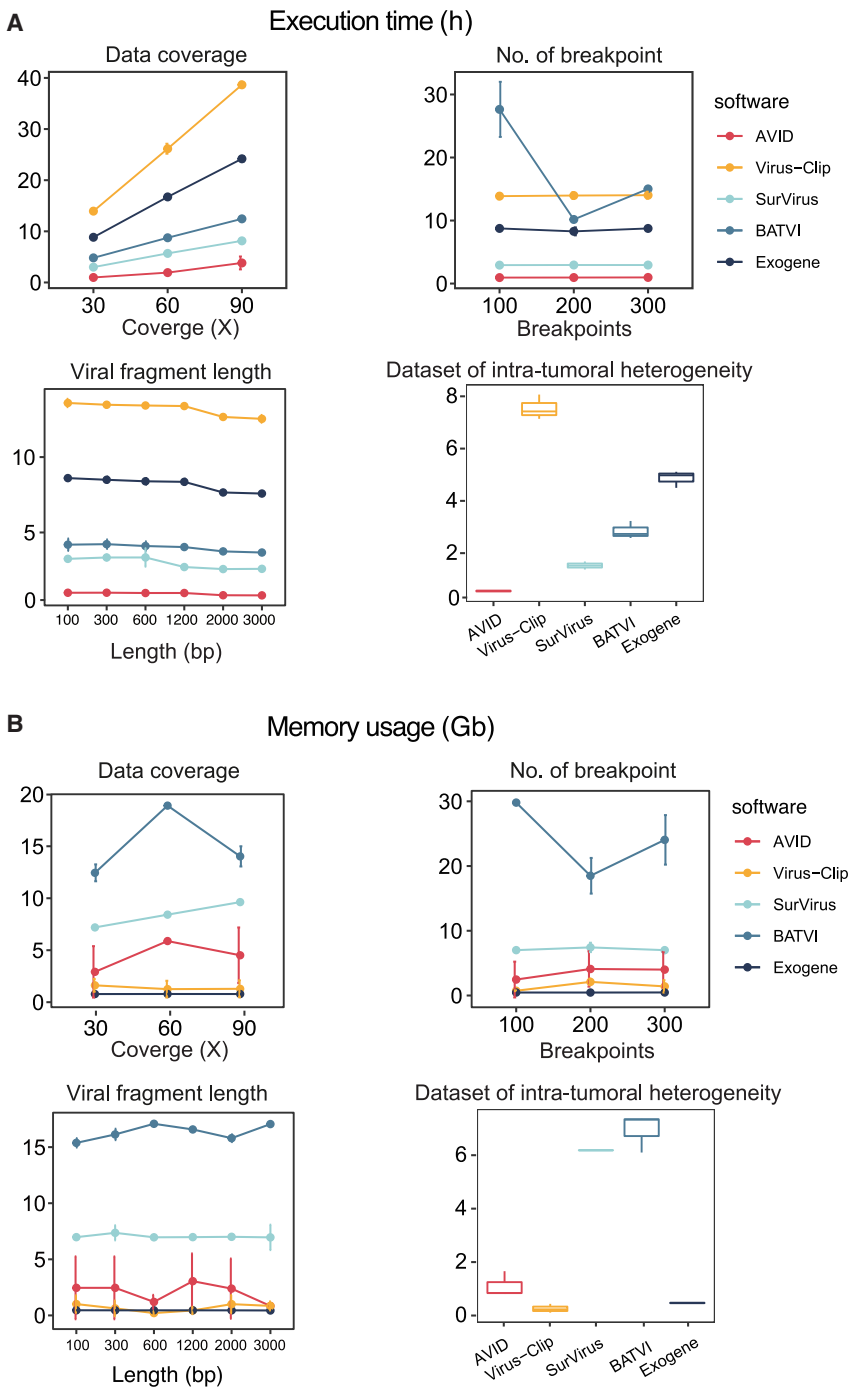


Figure 5. Performance of AVID in computational efficiency

Evaluation of AVID's computational efficiency in terms of execution time (A) and memory usage (B) and compared with other tools in simulated datasets. Error bars: mean ± SD. SD, standard deviation.

successfully demonstrated its wide applicability to detect viral integration of other oncoviruses and the identification of relevant and enhanced viral integration landscapes in multiple human cancers. Its versatility and adaptability also justified its usefulness in performing a mechanistic investigation of viral-integration-driven carcinogenesis of human cancers.

In summary, our study developed a viral integration detection algorithm that we believe will be applicable to a wide range of viral-integration-induced human cancers. Our results suggest that the tool can delineate a more comprehensive viral integration landscape for better mechanistic exploration, biomarker discovery, and patient stratification of human cancers.[27,28] Notably, given that viral integration can be successfully detected in plasma cell-free DNA of viral-associated human cancers,[29–31] the development of more sensitive and accurate detection of viral integration will be very much in need under such drastic conditions to distill the minority of event-containing reads among the vast majority of background reads without the viral integration event.

Limitations of the study

Given that viral integration events may involve complex structural alterations that link multiple human chromosomes and viral genomic fragments, e.g., heterocateny in HPV integration,[32] an ordinary short-read sequencing approach (including our simulated datasets) may not be able to fully capture and recapitulate the genuine viral integration landscape and the underlying complicated structural organization of genomic components. Hence, in future studies of viral integration and the corresponding development of computational detection algorithms, it will be more advisable to take advantage of long-read sequencing strategies that utilize the relatively longer length of sequencing reads to better reconstruct the genomic architecture and draw mechanistic insights into how the various genomically integrated viral components lead to carcinogenesis consequences. Although AVID outperformed the other tools in this study, the small size of experimentally validated cohorts limits the justification for its absolute superiority. When we applied the comparison in the full cohort (n= 135) of the Hu et al. study, which only experimentally validated a subset of all detected events (we believe there could be false positive detection), we observed the variability of HPV integration detection among tools as indicated by no single events being

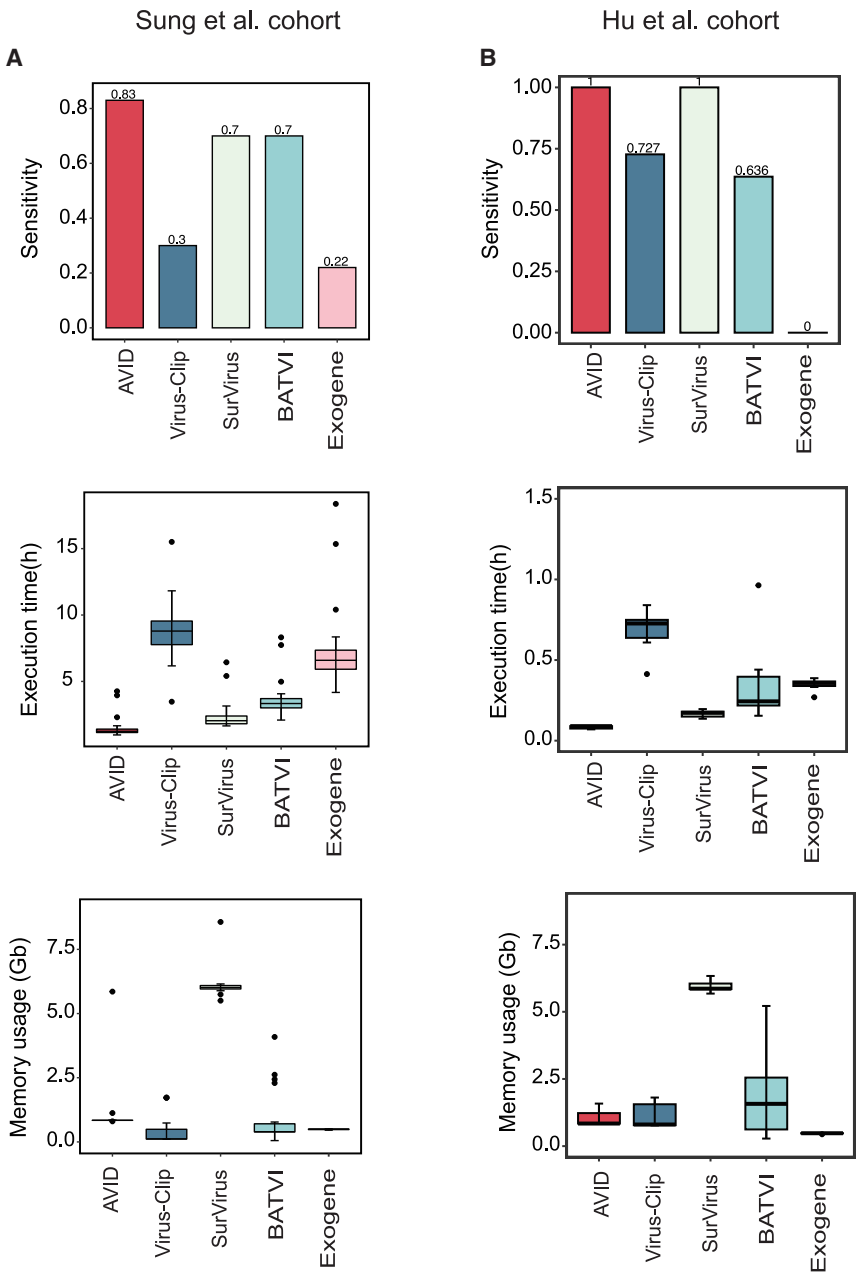


Figure 6. Performance of AVID in experimentally validated datasets

Performance evaluation of AVID in terms of sensitivity, execution time, and memory usage using experimentally validated HBV-associated HCC (A) and HPV-associated cervical cancer (B) cohorts, respectively. Error bars: mean ± SD. SD, standard deviation.

See also Table S5.

consistently detected by all tools (Table S6). This variability probably emphasizes the edge of each tool, which has been developed with specific focuses, and results in unique advantages and disadvantages.

RESOURCE AVAILABILITY

Lead contact

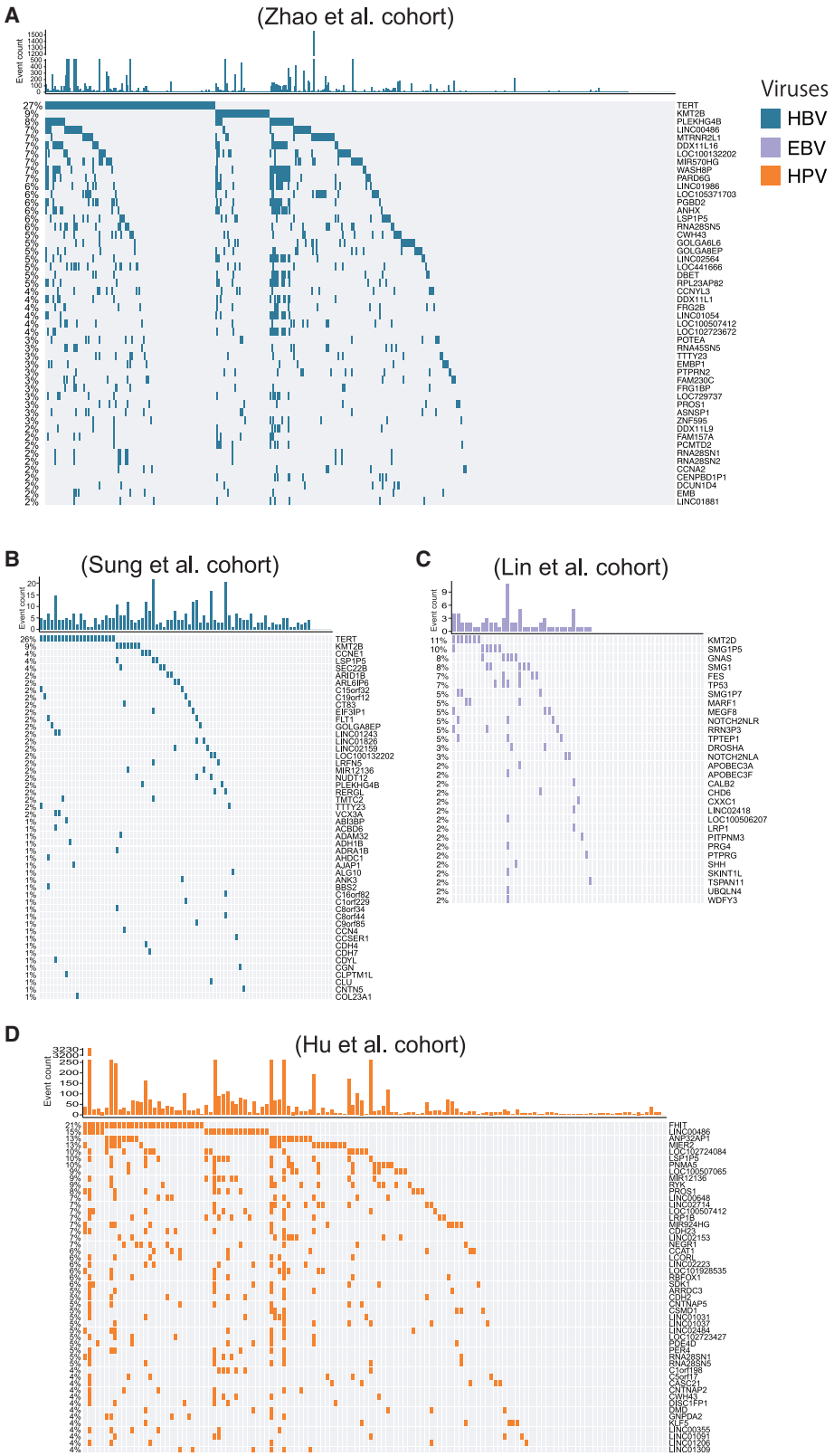
Further information and requests for resources should be directed to the lead contact, Daniel Wai-Hung Ho (dwhho@hku.hk).

Material availability

This study did not generate new unique reagents.

Data and code availability

- This study utilized publicly available data. Accession numbers are listed in the key resources table. Analysis results are provided within the paper materials and also at https://github.com/xueyinglyu/AVID.
- Code details are as follows:
  Project name: AVID.
  Project homepage: https://github.com/xueyinglyu/AVID.
  Archival DOI: https://doi.org/10.5281/zenodo.14915154.
  Operating system(s): platform independent.
  Programming language: Python, Perl, and R.
  Other requirements: Python 3.8 or higher.
  License: GNU GPL.
- Any further information needed to re-analyze the data reported in this paper is available from the lead contact upon request.



(legend on next page)

ACKNOWLEDGMENTS

The study was supported by the Hong Kong Research Grants Council Theme-based Research Scheme (T12-704/16-R and T12-716/22-R), an Innovation and Technology Commission grant to the State Key Laboratory of Liver Research (ITC PD/17-9), the Health and Medical Research Fund (10212956 and 07182546), the RGC General Research Fund (17100021 and 17117019), the National Natural Science Foundation of China (81872222), and the University Development Fund of The University of Hong Kong. I.O.-L.N. is the Loke Yew Professor in Pathology.

AUTHOR CONTRIBUTIONS

D.W.-H.H. and I.O.-L.N. were involved in the study concept and design. X.L., R.W.-Y.M., Q.L., H.-Y.C., T.S., F.Z., and R.L. were involved in software development and data analysis. X.L. was involved in drafting the manuscript. X.L., L.L.-Y.M., and D.W.-H.H. were involved in data acquisition and critical revisions of the manuscript.

DECLARATION OF INTERESTS

The authors have disclosed that they have no significant relationships with, or financial interest in, any commercial companies mentioned in this article.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Datasets
  - Workflow of AVID algorithm
  - Alignment-free classification of virus-containing reads
  - Staged read alignment strategy
  - Filtering of non-compatible read pairs
  - Removal of duplicated reads
  - Identification and classification of viral integration events
  - Analysis output, auxiliary information and visualization of integration events
  - Data simulation
  - Criteria for viral integration detection comparison
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2025.101007.

Received: July 11, 2024
Revised: November 25, 2024
Accepted: February 25, 2025
Published: March 24, 2025

REFERENCES

1. Muller-Coan, B.G., Caetano, B.F.R., Pagano, J.S., and Elgui de Oliveira, D. (2018). Cancer Progression Goes Viral: The Role of Oncoviruses in Aggressiveness of Malignancies. Trends Cancer *4*, 485–498. https://doi.org/10.1016/j.trecan.2018.04.006.
2. Zapatka, M., Borozan, I., Brewer, D.S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sültmann, H., and Moch, H.; PCAWG Pathogens (2020). The landscape of viral associations in human cancers. Nat. Genet. *52*, 320–330. https://doi.org/10.1038/s41588-019-0558-9.
3. Ho, D.W., Lyu, X., and Ng, I.O. (2021). Viral integration detection strategies and a technical update on Virus-Clip. Biocell *45*, 1495-–1500.
4. Llovet, J.M., Kelley, R.K., Villanueva, A., Singal, A.G., Pikarsky, E., Roayaie, S., Lencioni, R., Koike, K., Zucman-Rossi, J., and Finn, R.S. (2021). Hepatocellular carcinoma. Nat. Rev. Dis. Primers *7*, 6. https://doi.org/10.1038/s41572-020-00240-3.
5. Fonseca, L.G., and Carrilho, F.J. (2023). Current landscape and future directions for systemic treatments of hepatocellular carcinoma. Hepatoma Res. *9*, 27. https://doi.org/10.20517/2394-5079.2023.63.
6. Wu, Y.C.J., Wakil, A., Salomon, F., and Pyrsopoulos, N. (2023). Issue on combined locoregional and systemic treatment for hepatocellular carcinoma. Hepatoma Res. *9*, 6. https://doi.org/10.20517/2394-5079.2022.37.
7. Kung, J.W.C., and Ng, K.K.C. (2022). Role of locoregional therapies in the management of patients with hepatocellular carcinoma. Hepatoma Res. *8*, 17. https://doi.org/10.20517/2394-5079.2021.138.
8. Ho, D.W.H., Lo, R.C.L., Chan, L.K., and Ng, I.O.L. (2016). Molecular Pathogenesis of Hepatocellular Carcinoma. Liver Cancer *5*, 290–302. https://doi.org/10.1159/000449340.
9. Zhao, K., Liu, A., and Xia, Y. (2020). Insights into Hepatitis B Virus DNA Integration-55 Years after Virus Discovery. Innovation *1*, 100034. https://doi.org/10.1016/j.xinn.2020.100034.
10. Stephens, Z., O'Brien, D., Dehankar, M., Roberts, L.R., Iyer, R.K., and Kocher, J.-P. (2021). Exogene: A performant workflow for detecting viral integrations from paired-end next-generation sequencing data. PLoS One *16*, e0250915. https://doi.org/10.1371/journal.pone.0250915.
11. Tennakoon, C., and Sung, W.K. (2017). BATVI: Fast, sensitive and accurate detection of virus integrations. BMC Bioinf. *18*, 71. https://doi.org/10.1186/s12859-017-1470-x.
12. Rajaby, R., Zhou, Y., Meng, Y., Zeng, X., Li, G., Wu, P., and Sung, W.K. (2021). SurVirus: a repeat-aware virus integration caller. Nucleic Acids Res. *49*, e33. https://doi.org/10.1093/nar/gkaa1237.
13. Chen, X., Kost, J., and Li, D. (2019). Comprehensive comparative analysis of methods and software for identifying viral integrations. Brief. Bioinform. *20*, 2088–2097. https://doi.org/10.1093/bib/bby070.
14. Ho, D.W.H., Sze, K.M.F., and Ng, I.O.L. (2015). Virus-Clip: a fast and memory-efficient viral integration site detection tool at single-base resolution with annotation capability. Oncotarget *6*, 20959–20963. https://doi.org/10.18632/oncotarget.4187.
15. Sung, W.K., Zheng, H., Li, S., Chen, R., Liu, X., Li, Y., Lee, N.P., Lee, W.H., Ariyaratne, P.N., Tennakoon, C., et al. (2012). Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat. Genet. *44*, 765–769. https://doi.org/10.1038/ng.2295.
16. Hu, Z., Zhu, D., Wang, W., Li, W., Jia, W., Zeng, X., Ding, W., Yu, L., Wang, X., Wang, L., et al. (2015). Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nat. Genet. *47*, 158–163. https://doi.org/10.1038/ng.3178.
17. 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

Figure 7. Top genes detected with viral integration in different cohorts
(A) Top 50 genes identified with HBV integration in Zhao et al. cohort.
(B) Top 50 genes identified with HBV integration in Sung et al. cohort.
(C) All genes detected with EBV integration in Lin et al. cohort.
(D) Top 50 genes identified with HPV integration in Hu et al. cohort.
See also Table S6.

preference of HBV integration in hepatocellular carcinoma. Nat. Commun. *7*, 12992. https://doi.org/10.1038/ncomms12992.

18. Lin, D.C., Meng, X., Hazawa, M., Nagata, Y., Varela, A.M., Xu, L., Sato, Y., Liu, L.Z., Ding, L.W., Sharma, A., et al. (2014). The genomic landscape of nasopharyngeal carcinoma. Nat. Genet. *46*, 866–871. https://doi.org/10.1038/ng.3006.
19. Sartorius, K., Makarova, J., Sartorius, B., An, P., Winkler, C., Chuturgoon, A., and Kramvis, A. (2019). The Regulatory Role of MicroRNA in Hepatitis-B Virus-Associated Hepatocellular Carcinoma (HBV-HCC) Pathogenesis. Cells *8*, 1504. https://doi.org/10.3390/cells8121504.
20. Li, H.C., Yang, C.H., and Lo, S.Y. (2022). Long noncoding RNAs in hepatitis B virus replication and oncogenesis. World J. Gastroenterol. *28*, 2823–2842. https://doi.org/10.3748/wjg.v28.i25.2823.
21. Péneau, C., Imbeaud, S., La Bella, T., Hirsch, T.Z., Caruso, S., Calderaro, J., Paradis, V., Blanc, J.F., Letouze, E., Nault, J.C., et al. (2022). Hepatitis B virus integrations promote local and distant oncogenic driver alterations in hepatocellular carcinoma. Gut *71*, 616–626. https://doi.org/10.1136/gutjnl-2020-323153.
22. Jiang, Z., Jhunjhunwala, S., Liu, J., Haverty, P.M., Kennemer, M.I., Guan, Y., Lee, W., Carnevali, P., Stinson, J., Johnson, S., et al. (2012). The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. Genome Res. *22*, 593–601. https://doi.org/10.1101/gr.133926.111.
23. Sze, K.M.F., Ho, D.W.H., Chiu, Y.T., Tsui, Y.M., Chan, L.K., Lee, J.M.F., Chok, K.S.H., Chan, A.C.Y., Tang, C.N., Tang, V.W.L., et al. (2021). Hepatitis B Virus-Telomerase Reverse Transcriptase Promoter Integration Harnesses Host ELF4, Resulting in Telomerase Reverse Transcriptase Gene Transcription in Hepatocellular Carcinoma. Hepatology *73*, 23–40. https://doi.org/10.1002/hep.31231.
24. Lyu, X., Sze, K.M.F., Lee, J.M.F., Husain, A., Tian, L., Imbeaud, S., Zucman-Rossi, J., Ng, I.O.L., and Ho, D.W.H. (2024). Disparity landscapes of viral-induced structural variations in HCC: Mechanistic characterization and functional implications. Hepatology. https://doi.org/10.1097/HEP.0000000000001087.
25. Giosa, D., Lombardo, D., Musolino, C., Chines, V., Raffa, G., Casuscelli di Tocco, F., D'Aliberti, D., Caminiti, G., Saitta, C., Alibrandi, A., et al. (2023). Mitochondrial DNA is a target of HBV integration. Commun. Biol. *6*, 684. https://doi.org/10.1038/s42003-023-05017-4.
26. Lyu, X.Y., Tsui, Y.M., Tam, I.K.K., Li, P.M., Cheung, G.C.H., Lee, J.M.F., Ng, I.O.L., and Ho, D.W.H. (2024). Resolution of Optimal Mitochondrial and Nuclear DNA Enrichment in Target-Panel Sequencing and Physiological Mitochondrial DNA Copy Number Estimation in Liver Cancer and Non-Liver Cancer Subjects. Cancers *16*, 3012. https://doi.org/10.3390/cancers16173012.
27. Ahn, J.C., Lee, Y.T., Agopian, V.G., Zhu, Y., You, S., Tseng, H.R., and Yang, J.D. (2022). Hepatocellular carcinoma surveillance: current practice and future directions. Hepatoma Res. *8*, 10. https://doi.org/10.20517/2394-5079.2021.131.
28. Zhou, K., and Terrault, N. (2022). Promise and pitfalls of new viral biomarkers for hepatocellular carcinoma risk prediction in patients with chronic hepatitis B. Hepatoma Res. *8*, 15. https://doi.org/10.20517/2394-5079.2022.06.
29. Cheung, T.T., Wai-Hung Ho, D., Lyu, S.X., Zhang, Q., Tsui, Y.M., Ching-Yun Yu, T., Man-Fong Sze, K., Man-Fong Lee, J., Lau, V.W.H., Yin-Lun Chu, E., et al. (2024). Multimodal Integrative Genomics and Pathology Analyses in Neoadjuvant Nivolumab Treatment for Intermediate and Locally Advanced Hepatocellular Carcinoma. Liver Cancer *13*, 70–88. https://doi.org/10.1159/000531176.
30. Lyu, X., Tsui, Y.M., Ho, D.W.H., and Ng, I.O.L. (2022). Liquid Biopsy Using Cell-Free or Circulating Tumor DNA in the Management of Hepatocellular Carcinoma. Cell. Mol. Gastroenterol. Hepatol. *13*, 1611–1624. https://doi.org/10.1016/j.jcmgh.2022.02.008.
31. Parida, P., Baburaj, G., Rao, M., Lewis, S., and Damerla, R.R. (2024). Circulating cell-free DNA as a diagnostic and prognostic marker for cervical cancer. Int. J. Gynecol. Cancer *34*, 307–316. https://doi.org/10.1136/ijgc-2023-004873.
32. Akagi, K., Symer, D.E., Mahmoud, M., Jiang, B., Goodwin, S., Wangsa, D., Li, Z., Xiao, W., Dunn, J.D., Ried, T., et al. (2023). Intratumoral Heterogeneity and Clonal Evolution Induced by HPV Integration. Cancer Discov. *13*, 910–927. https://doi.org/10.1158/2159-8290.Cd-22-0900.
33. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
34. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.
35. van Buuren, N., Ramirez, R., Soulette, C., Suri, V., Han, D., May, L., Turner, S., Parvangada, P.C., Martin, R., Chan, H.L.Y., et al. (2022). Targeted long-read sequencing reveals clonally expanded HBV-associated chromosomal translocations in patients with chronic hepatitis B. JHEP Rep. *4*, 100449. https://doi.org/10.1016/j.jhepr.2022.100449.
36. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164. https://doi.org/10.1093/nar/gkq603.
37. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. Bioinformatics *30*, 2811–2812. https://doi.org/10.1093/bioinformatics/btu393.
38. Wilkinson, L. (2011). ggplot2: Elegant Graphics for Data Analysis by H. WICKHAM. Biometrics *67*, 678–679. https://doi.org/10.2307/41242513.
39. Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., Chen, Y., Mu, D., Zhang, H., Li, N., et al. (2012). pIRS: Profile-based Illumina pair-end reads simulator. Bioinformatics *28*, 1533–1535. https://doi.org/10.1093/bioinformatics/bts187.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Sung et al. cohort	Sung et al.[15]	https://www.ncbi.nlm.nih.gov/sra?term=ERP001196
Hu et al. cohort	Hu et al.[16]	https://www.ncbi.nlm.nih.gov/sra?term=SRP048813
Zhao et al. cohort	Zhao et al.[17]	https://www.ncbi.nlm.nih.gov/sra?term=SRP068532
Lin et al. cohort	Lin et al.[18]	https://www.ncbi.nlm.nih.gov/sra?term=SRP035573
Software and algorithms		
AVID	This paper	https://github.com/xueyinglyu/AVID and https://10.5281/zenodo.14915154
BATVI	Tennakoon and Sung[11]	https://quay.io/repository/biocontainers/batvi?tab=tags&tag=latest
Exogene	Stephens et al.[10]	https://github.com/zstephens/exogene
SurVirus	Rajaby et al.[12]	https://github.com/kensung-lab/SurVirus
Virus-Clip	Ho et al.[14]	https://github.com/dwhho/Virus-Clip

METHOD DETAILS

Datasets

Lin et al. cohort of EBV-associated NPC was downloaded from Sequencing Read Archive (SRA) (accession ID: SRP035573). Sung et al. cohort of HBV-associated HCC was downloaded from European Genome phenome Archive (accession: ERP00196). Zhao et al. cohort of HBV-associated HCC was downloaded from SRA (accession: SRP068532). Hu et al. cohort of HPV-associated cervical cancer was downloaded from SRA (accession: SRP048813 and SRP048861). The code of AVID is provided in GitHub (https://github.com/xueyinglyu/AVID).

Workflow of AVID algorithm

AVID algorithm allows the identification of virus-host integration junctions at single-base resolution (Figure 1). Initially, raw sequencing reads in FASTQ format were imported as input. Initial alignment-free identification of virus-containing reads was performed to confine the pool of candidates that potentially harbor soft-clipped sequencing reads (with partial human and partial viral chimeric fragments) or chimeric read pairs (one read of human origin and another one of viral origin). These candidates were initially aligned to virus genome followed by alignment with human genome. Next, optional user-defined quality-control filtering was applied to the data. Undesirable PCR duplicates may be removed by using mapping positions. False positive artifacts may also be reduced by checking strand compatibility and viral insert size. Moreover, short chimeric fragments were subgrouped, and event clustering was performed to collate related reads for more accurate breakpoint estimation.

Alignment-free classification of virus-containing reads

Viral reads were first extracted from raw sequencing files to narrow down the proportion of reads for efficient downstream analysis. We followed the strategy of gapped k-mer hashing table[12] with modifications. Briefly, in order to more rapidly identifying sequencing reads containing viral DNA fragments, instead of examining the full length of sequencing reads, we only generated k-mers on the 8bp of the two ends of sequencing reads. Reads containing viral DNA sequence at either end were preserved for subsequent alignment.

Staged read alignment strategy

Based on the previously classified virus-containing reads, they were initially aligned to the virus reference genome by BWA (v0.7.17) with default parameters.[33] For the unmapped reads or soft-clipped portion of reads, they were subjected to alignment to the hg38 human genome by BWA. On the other hand, to boost up the sensitivity of detection, undetermined portions of reads with a length <26bp that could not fulfill the minimum input length requirement of BWA, they were re-aligned by BLASTN (v2.5.0) to the virus reference genome.[34] Default parameters and an identity threshold of >80% were utilized to define successful mapping to virus genome. Moreover, to avoid false positive detection causing by overlapping regions between virus and human genomes, we discarded reads that mapped to both genomes.

Filtering of non-compatible read pairs

As PCR amplification during library preparation can generate PCR-mediated recombination upon hybridization, it is possible to falsely introduce viral integration at different chromosomes with varied fragment sizes. Removal of chromosomal and insert-size incompatibility can reduce technical artifacts and false positive discoveries.

To reduce the false positive integration detection, AVID algorithm removed the non-compatible reads: (1) paired-end reads mapping to different chromosomes of the host genome; (2) paired-end reads mapping to the host genome separated by a distance over 1.5-fold of the DNA fragment size; (3) the sum of human and viral portions of a soft-clipped read exceeded the full length of sequencing read by 10%.

However, given the possibility of viral integration events flanked by DNA from different chromosomes[35] or complex structural rearrangements involving more than one chromosome,[32] we implement this checking as an optional filter instead of a mandatory step.

Removal of duplicated reads

Duplicated reads due to the amplification process can cause the overestimation of virus integration, leading to potential false positive outcomes. AVID algorithm removed duplicates by filtering out paired-end reads with the same mapping positions at both the host and virus genomes, instead of considering the position in either one of them. For the identified PCR duplicates, only one pair of reads was retained for downstream analysis.

Identification and classification of viral integration events

AVID algorithm can take advantage of pair-end reads to have better detection of viral integration. It utilized both soft-clipped read (a chimeric read that simultaneously has both human and virus portions) and chimeric read pair (pair-end reads with one of them fully mapped to human and the other one fully mapped to virus) to have more sensitive identification of viral integration events. It also subtyped the viral integration events: Type 1: one soft-clipped and one fully human-mapped reads; Type 2: one soft-clipped and one fully virus-mapped reads; Type 3: one fully human-mapped and one fully virus-mapped reads; and Type 4: two soft-clipped reads. Through the initial identification of hallmark events (soft-clipped reads and/or chimeric read pairs) and the subsequent subtyping of the events, this can streamline the analytic process and achieve improved efficiency. Breakpoints defined by soft-clipped reads were collated with an interval of 10bp and the mean location within the cluster was assigned as the breakpoints of the host and the virus genome. When the virus DNA inserted into the host genome and produced two host-to-virus junctions, the length of the inserted viral DNA could be inferred by the distance between the left and right viral breakpoints.

Chimeric read pairs classified as Type 3 events, i.e., one read fully mapped to host and another one fully mapped to virus, were used for coarse integration region identification but without precise breakpoints. They were clustered within a distance less than the fragment size of the DNA library and their intervals in the host and viral genome were indicative of the putative regions that harboring the integration breakpoints. Moreover, both soft-clipped reads and chimeric read pairs were also contributing to the supportive read count evidence for the event.

Analysis output, auxiliary information and visualization of integration events

Output of AVID algorithm reported the information of viral and human breakpoint positions, orientation of integration, supporting read count, status of repetitive alignment, and affected viral and human genes. AVID algorithm summarized integration events with soft-clipped reads having secondary alignment by BWA. Caution was taken for events with repetitive alignment. It also reported the orientation of viral integration e.g., positive orientation defined as positive strand of virus inserted into positive strand of the host genome, whereas negative orientation was the reverse. Host and viral genes affected with the integration events were determined by the breakpoint and/or integration interval. AVID algorithm utilized ANNOVAR (2020-06-07) for gene and genic position annotation.[36] It also utilized R packages circlize and ggplot2 to provide visualization functions for viral integration analysis such as generating plots for summarizing host and viral breakpoint positions and the size distribution of viral fragments.[37,38]

Data simulation

Software package pIRS (v2.0.2) was used to simulate whole-genome sequencing data.[39] We have simulated the data under different scenarios, which included data at different sequencing coverage (30X, 60X and 90X), different numbers of viral integration sites (100, 200 and 300), and viral integrations of different insert sizes of viral fragments (100, 300, 600, 1200, 2000 and 3000bp). More importantly, to mimic the high heterogeneity and complexity of HCC tumor i.e., the presence of different subclones of malignant cells having their specific sets of molecular alterations, we simulated 3 individual subclones of malignant cells with different aforementioned parameters (each with 20 unique integration events and were generated at coverage of 3X, 5X and 8X respectively) and combined them into a single hybrid dataset. We believe this likely recapitulates the genuine biological condition that exist in HCC tumors.

Criteria for viral integration detection comparison

Since not all tools reported integration breakpoint position at the virus genome and the orientation of integration, we compared the tools using breakpoint positions reported at the human genome for comparison. For AVID algorithm, events with supportive evidence (soft-clipped reads and/or chimeric read pairs) count of at least 3 (at least one soft-clipped read among them) were considered as candidate events in experimentally validated datasets. For simulated datasets, a more stringent cutoff of at least 3 soft-clipped reads was applied. Other tools were executed using default parameters and events with read count of at least 3 were retained for analyses. HBV integration events detected within an interval of 5bp from the designated breakpoint positions (both simulated and experimentally validated data) were considered correct detection. We calculated the precision, recall, and F1-score statistics. For the evaluation of execution time and memory usage, data analysis was performed in triplicate runs under identical computational conditions and the

corresponding mean values were used for comparison. For HPV integration comparison in the full cohort (N = 135) of Hu et al. study, we applied a uniform cut-off of at least three supporting soft-clipped reads for each tool. Since Exogene did not detect any HPV integration, we excluded it from the comparison. Samples that failed in execution were excluded.

QUANTIFICATION AND STATISTICAL ANALYSIS

The performance of AVID was evaluated by precision, recall and F1 score, as indicated in the figure legends. Error bar definitions are provided in the figure legends.