International Journal of Environmental and Biological Sciences
ISSN: 3054-3924
https://doi.org/10.63939/pxh43951
Copyright © The Author(s)
Vol. 1 , No 4 , October 2026

RESEARCH Open Access

The Role of Bioinformatics in Understanding Cancer Viruses: From Genome Sequencing to Antiviral Discovery

Zainab Abd Al_hassan Zwain[1]. Email: zainaba.zwain@student.uokufa.edu.iq

Sahar Mohammed Jawad[2]. Email: saharm.alkarawy@uokufa.edu.iq

Emad Kareem Alsabari[3]. Email: emadk.alsabari@uokufa.edu.iq

Alaa Mohammed Khaleel[4]. Email: alaam.jabri@uokufa.edu.iq

Dhefaf Hameed Al-Mudhafar[5]. Email: dhifaf.almudhaffer@uokufa.edu.iq

Hasanain Ali Shubbar[6]. Email: hasanain.a.shubbar@jmu.edu.iq

Sara salih hassan[7]. Email: sarah.s.bidany@jmu.edu.iq

[6,7]Jabir Ibn Hayyan University for Medical and Pharmaceutical Sciences, Faculty of Medical Science, Academic Center for Epidemiological Diseases , Najaf, Iraq

Dr. Samer A. MH. Al-Hilali[8]. Email: sameram.hussein@uokufa.edu.iq

Eman Hassani Shabat[9]. Email: emanh.alsalami@uokufa.edu.iq

Tabarek Hadi Al-Naqeeb[10]. Email: tabarakh.alnaqib@uokufa.edu.iq

Liqaa Hadi Naser[11]. Email: liqaah.alzurfi@uokufa.edu.iq

Ayat Raad jawad[12]. Email: ayatraadjawad98@gmail.com

Saif Jabbar Yasir[13]. Email: saif.alshehmani@uokufa.edu.iq

Middle Euphrates Cancer Research Unit, College of Medicine,

University of Kufa, Iraq. research.med@uokufa.edu.iq

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

Abstract

Background: Virus-associated cancers are caused by several oncogenic mechanisms such as direct oncoprotein expression, chronic inflammation and viral genome integration, and are responsible for a substantial part of the global cancer burden. High-throughput sequencing has revolutionized tumor virology research, but studying viral-host interactions remains computationally and methodologically challenging because host genomes are repetitive, sequencing error rates are high, and samples are contaminated. **Objectives:** This review aims to integrate existing computational and sequencing approaches in cancer virology and to emphasize the importance of bioinformatics in bridging the gap between the raw genomic data and the identification of viral biomarkers and anti-viral targets. **Methodological and Technological Framework:** We assess the joint performance of state-of-the-art bioinformatic pipelines (e.g. VIRUS Breakend, nf-VIF, Exogene), target-enrichment capture methods and short- and long-read sequencing systems (e.g. PacBio HiFi and Oxford Nanopore). We are focusing on the integration of passenger event drivers, intrahost viral diversity assessment, and microbial detection resolution for low-biomass tumor samples. **Key Points:** Long read sequencing has revolutionized the mapping of integration sites, resolving full-length virus-host junctions and complex structural rearrangements. Integration of multi-omic data including transcriptomics, chromatin conformation data and variant analysis is also key to understand how viral insertions disrupt host regulatory neighborhoods and drive oncogene expression, e.g. E6/E7 in HPV-associated cancers. But reproducibility needs orthogonal validation, contamination filtering, and strict quality control. **Conclusion:** Advanced bioinformatic frameworks are the answer for mapping viral integration landscapes and tracking intrahost evolution. These computational insights will need to be prospectively validated in clinical cohorts with standardized reporting to translate them into actionable prognostic markers and standard treatments.

Keywords: Viral integration; Tumor microbiome; Cancer virology; Long-read sequencing; Bioinformatics; Next-generation sequencing (NGS); Antiviral discovery.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

1. Introduction and Scope

Virus-associated cancers are an important component of the global cancer burden. The mechanisms by which the oncogenic viruses that cause them, including human papillomavirus (HPV), hepatitis B virus (HBV), and hepatitis C virus (HCV), contribute to tumorigenesis, differ greatly. Some of these viruses integrate into the host genome, while others do so infrequently [1]. The placement of these agents within the larger category of virus-associated malignancies indicates that viral carcinogenesis can proceed via multiple pathways including direct oncoprotein expression, chronic inflammation and host immune evasion. This has implications for the design of computational approaches to capture viral and host determinants of disease [2][3].

A reliable first step is the detection of the viral genome in tumor cells, as the oncogenic mechanisms of these viruses are only partially known. This becomes especially difficult when the viral family is highly divergent from known reference sequences [4]. Computational pathogen identification pipelines bypass this limitation through aligning unmapped sequencing reads from tumor genomes and transcriptomes against curated viral databases. This method enables the identification of divergent viral sequences that would be overlooked by conventional alignment. When applied to the scale of thousands of tumors–normal pairs, this approach has revealed viral associations in several cancer types, including those sequences that were not previously associated with those malignancies [5].

Such pan-cancer surveys depend on harmonized data resources, as large public cohorts like TCGA provide the RNA-seq and clinical annotations needed to correlate viral presence with host gene expression changes across tumor types [6]. However, these integrated analyses also expose the limits of current resources, since existing multi-omics efforts remain constrained by narrow data scope and the difficulty of linking viral signals to specific host signaling pathways [7]. Addressing this gap requires bioinformatic frameworks that integrate long-read sequencing—which resolves viral insertions and chimeric virus–host transcripts—with transcriptome-wide expression profiling to link viral presence to dysregulated host pathways ([8][9].

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

Only seven viruses are currently known to cause human cancers: human gammaherpesvirus 4 (EBV), human herpesvirus 8 (HHV-8), human papillomavirus (HPV), human T-cell lymphotropic virus type 1 (HTLV-1), hepatitis B virus (HBV), hepatitis C virus (HCV) and Merkel cell polyomavirus (MCPyV) [10]. These frameworks should therefore be considered. Collectively, these agents account for a growing fraction of cancers worldwide, and their genomes provide the starting point for computational workflows ([11][12]. Bioinformatic pipelines processing these genomes must first recover viral sequences reliably from tumor RNA-seq data. Recent untargeted assembly tools like VirMAP match or exceed established classifiers such as Kraken while offering finer taxonomic resolution [13].

Applied to TCGA RNA-seq data, this assembly-based strategy allows viral expression to be tested for association with overall survival, revealing that exogenous cancer viruses are linked to better outcomes while human endogenous viruses are linked to worse outcomes [14]. Because such associations are drawn from bulk RNA-seq, they reflect correlations within tumor tissue rather than established causal roles, and validating them requires orthogonal experimental confirmation of viral presence and activity in matched samples.

Current understanding of virus-associated carcinogenesis suggests that while detection of viral sequences in tumor tissues is an important first step, it is not sufficient to establish a direct oncogenic link. The biological relevance of viral identification is dependent on the interaction of tumor microenvironment, transcriptional activity, host genomic aberrations and viral genetic components. In this sense, bioinformatics should be seen not only as a method of viral sequence detection, but also as an analytical framework that can establish a link between the presence of viruses and molecular mechanisms of malignant transformation. Combining viral genomic data with host transcriptomic and clinical data could help to differentiate between incidental viral detection and biologically meaningful viral involvement. However, causal interpretations should not be made until the computational associations have been validated both molecularly and clinically and in an independent fashion.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

The use of bioinformatics in cancer virology has become an important tool for systematically identifying and characterizing oncogenic viruses in a range of cancers. Its main scientific contribution is the integration of molecular information of the host and the virus to explore the mechanisms underlying the development of virus-associated tumors. Future advances will require standardized computational methods, comprehensive multi-omics data and experimentally validated correlations between virus activity and cancer progression.

2. Cancer Virus Genomics and Integration Dynamics

Viral genomes associated with cancer are typically compact double-stranded DNA or single-stranded RNA molecules encoding a limited set of oncoproteins, and their genomic architecture determines whether they persist as episomes or integrate into host chromosomes [15]. Integration is not an all-or-nothing event; for HBV, HPV16, HPV18, and adeno-associated virus-2, viral insertion has been linked to local variations in genomic copy number that drive cancer through regional gene amplification [16].

Integration into recurrent insertion hotspots disrupts host genes, and further sculpts integration-driven clonal expansion. Thus, genome-scale profiling of integration sites is critical for understanding how these viruses drive tumor growth. Integration events at both viral and host loci are frequently accompanied by structural rearrangements. By sequencing HPV-positive cervical tumors with long reads, these rearrangements were found to be structurally complex, recurrent, and associated with heterologous viral configurations that interfere with the allele-specific expression of neighboring human genes [17].

Furthermore, viral integration generates chimeric transcripts and copy-number alterations persisting across tumor clones, complicating efforts to attribute a given breakpoint to a single clonal integration event [18]. Long-read sequencing addresses this ambiguity by spanning the full length of the integrated viral genome and its flanking host sequence, resolving both the virus–host junction and the insertion site within a single molecule—a feat ligation-mediated short-read PCR often fails to achieve [19][20].

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

Integration of a virus into the host genome may be an important molecular event in virus-associated carcinogenesis especially if it affects genomic stability, gene regulation or cellular signaling pathways. However, the biological consequences of integration are not homogeneous and the presence of viral DNA in a host chromosome does not imply that it is oncogenic. The functional relevance of integration is defined by the genomic location, structural configuration, transcriptional effects, and association with tumor clonal expansion. Thus, functional analysis of neighboring genes and regulatory elements should be performed in parallel with the identification of recurrent integration sites. In this context, long-read sequencing offers unique advantages, enabling the detection of complex integration structures that may not be identified using traditional short-read approaches. Integration of cancer-associated viruses is variable and can affect both the organization of the host genome and gene expression. Detailed characterization of viral integration sites, structural rearrangements and associated transcriptional changes is needed to differentiate incidental integrations from potentially oncogenic events. High-resolution genome sequencing of the viral genome provides a more robust structure for functional analysis of the role of viral integration in tumor initiation and progression.

3. Genome Sequencing Technologies: Short-Read vs. Long-Read Paradigms

Whole-genome sequencing of HPV-positive tumors is often performed using Illumina and other short-read platforms. The short read lengths of these platforms, however, present a challenge to resolving long tandem repeats and translocated sequences at integration sites [21]. Long-read platforms (PacBio HiFi and Oxford Nanopore) overcome this limitation by generating reads that span the full length of the integrated viral genome and the flanking host sequences [22][23]. This capability is important for validation of insertion structures like the HPV16 insertion in C5orf67, and to resolve distinct integration patterns at single-molecule resolution [24,25]. However, nanopore reads are more error-prone and therefore require consensus polishing tools (e.g., Nanopolish) and orthogonal validation (PCR or targeted capture sequencing) to confirm true integration events and rule out sequencing artifacts [26].

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

The sequencing technology employed can greatly influence the accuracy and completeness of the characterization of viral genomes. Short read sequencing still has a place because of its accuracy, appropriateness for large scale genomic studies, and the presence of established data analysis pipelines. But the short read length can restrict the reconstruction of complex viral integrations, repetitive genomic regions, and large structural rearrangements. Long-read technologies allow for more information by allowing longer genomic fragments and more complete virus-host junctions to be analyzed. However, one must consider the differences in error profiles between platforms, sequencing costs, sample quality requirements, and computational demands. Therefore, hybrid sequencing approaches could be a feasible compromise between sequence accuracy and structural resolution particularly when considering clinically relevant viral integration events. Long-read and short-read sequencing methods are complementary to each other in cancer virology. Short-read methods provide accurate and scalable sequence analyzes. Long-read platforms improve characterization of complex viral integrations and genomic rearrangements. By applying these technologies, together with rigorous bioinformatic analysis and independent validation, a thorough comprehension of the viral genomic architecture in cancer can be achieved.

3.1. Liquid Biopsy and Targeted Enrichment

Extending long-read frameworks to circulating cell-free DNA (cfDNA) enables non-invasive monitoring of HPV-associated cancers and capture of virus–host chimeric DNA from plasma ([27][28]. Enrichment approaches like HPV DNA capture-seq increase the proportion of viral reads by 1,000 to 10,000 times, facilitating recovery of comprehensive viral sequences from limited clinical specimens [29]. This, combined with fragment-size profiling [30], enables the distinction of cancer-associated integration signals from active or concurrent infections.

Liquid biopsy is a promising advance in molecular cancer virology, as it can potentially provide information on tumor-associated molecular alterations without the need for repeated invasive tissue sampling based on circulating viral DNA and virus–host chimeric fragments. Targeted enrichment mechanisms can be used to improve analytical

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin



sensitivity when viral nucleic acids are present in low concentrations. However, detection of circulating viral DNA does not necessarily indicate the presence of an active malignancy, since viral sequences can also be present as a consequence of a nonmalignant infection or other biological sources. Therefore, the interpretation of liquid biopsy results should include a detailed evaluation of the clinical context, patterns of viral integration, characteristics of fragments and molecular specificity. In longitudinal studies the reliability of circulating viral biomarkers to reflect tumor burden, treatment response or disease recurrence is of particular importance.

The use of liquid biopsy together with targeted viral sequencing has the potential to enable minimally invasive molecular monitoring of virus-associated malignancies. Viral–host junctions identified in tumors can improve the specificity of biomarkers. Enrichment strategies can improve the retrieval of viral sequences from limited clinical specimens. These techniques can only be implemented in routine cancer diagnosis or surveillance after they have been subjected to standardized detection thresholds, analytical validation, and prospective clinical investigation.

3.2. Quality Control, Decontamination, and Artifact Reduction

Cross-sample carryover (e.g., WaGa DNA contamination in Merkel cell carcinoma cell lines) can generate spurious integration calls that persist through analysis (Czech-Sioli et al., 2020). Because low-biomass tumor samples are susceptible to spurious detections, microbial profiling must be benchmarked against known contamination patterns using matched normal tissues and extraction blanks [31][32]. Standardized reporting of filtering thresholds, duplicate removal, and validation criteria ensures discrepancies are traced to methodology rather than biology [33][5].

The accuracy of computational viral detection largely depends on the correctness of the sequencing and analysis procedures, as well as the quality of the biological specimens. In samples with low levels of microbial or viral nucleic acid, spurious viral signals can arise from contaminating nucleic acid introduced during sample collection, nucleic acid extraction, library preparation, or sequencing. Such artifacts may bias estimates of virus prevalence and lead to unwarranted interpretations of virus–cancer relationships.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

Therefore, quality control should be considered an integral part of the study design rather than a post-computational correction. To separate true biological signals from technical artifacts, contamination-aware analytical approaches, matched normal samples, proper negative controls and independent molecular validation are required. Quality control is essential for reproducibility in cancer virology. Contamination control, transparent bioinformatic filtering, and orthogonal validation all reduce the risk of false-positive viral detections and mis-assigned integrations. Standardized quality-control procedures need to be implemented to improve the clinical utility, reliability and reproducibility of viral genomic studies.

4. Variant Analysis and Intrahost Diversity

Variant analysis in cancer-associated viruses quantifies intrahost viral diversity. Low-frequency single-nucleotide variants (SNVs) reveal mutational processes acting on the viral genome during persistence and tumor progression [34]. In HPV16, amplicon-based long-read workflows show that a single variant occurs approximately every 31 nucleotides, enriched in E5, L2, and E2 regions while E7 remains highly conserved [35]. Tracking these variants over time helps determine whether dominant lineages are selected during progression from precancerous lesions to invasive carcinoma, provided sequencing depth is sufficient [9][36].

The genetic diversity of the ntrahost virus is of major importance in the evolutionary processes that may be involved in the development of cancer and persistent infection. Identification of viral single-nucleotide variants and their distribution across functional genomic regions may help to characterize viral population structure and identify candidate variants associated with biological adaptation. However, the presence of sequence variation alone is not enough to assign malignant transformation or treatment resistance to a specific mutation. These interpretations need to take into account the accuracy of sequencing, the frequency of variants, longitudinal consistency and functional consequences. It may be possible to identify whether particular viral lineage or mutation is associated with clinically relevant tumor features by combining viral variant analysis with host genomic and transcriptomic data. Variant analysis is important to study the molecular

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

diversity and evolution of cancer-related viruses. High resolution sequencing can identify viral genetic changes that may be relevant for disease progression, clonal selection and persistence. However, functional validation and longitudinal studies are required to confirm the clinical relevance of these variants before they can be considered as reliable prognostic or therapeutic biomarkers.

4.1. Oncoprotein Regulation and Chimeric Transcripts

Integration often leads to the loss of E2-mediated repression (since E2 normally binds the viral long control region to repress E6 and E7 transcription). This removal allows unchecked, integration-dependent expression of E6 and E7 oncoproteins, a key driver of cervical carcinogenesis [37]. The resulting virus–host chimeric transcripts are stabilized by downstream host polyadenylation signals, making them accessible to siRNA-based targeting [38]. However, such strategies require selective suppression to avoid off-target consequences near disrupted host genes like INPP4B and MAPK10

In HPV-associated cancers, the relationship between viral integration and oncogene expression is of particular importance, as disruption of viral regulatory regions may result in altered expression of E6 and E7 oncoproteins. However, regulation of viral oncogenes is subject to a variety of molecular mechanisms and their expression should not be assumed to be universally dependent on integration. Virus–host chimeric transcripts give further evidence for transcriptionally active integration events and may also reveal interactions between viral sequences and host regulatory machinery. Their potential medical relevance is of particular interest due to the possibility that specific viral transcripts may serve as molecular targets. However, the development of transcript-directed therapies requires accurate characterization of transcript structure, expression specificity, and potential effects on adjacent host genes.

Control of viral oncoproteins and production of virus-host chimeric transcripts are two prominent molecular features of some virus-associated cancers. Their characterization may lead to a better understanding of viral oncogenic activity and identify possible therapeutic targets. In future work, greater emphasis should be placed on functional validation of transcriptionally active integration events and on determination of the

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

specificity, efficacy and safety of strategies aimed at targeting expression of viral oncogenes.

5. Bioinformatics Approaches and Pipeline Integration

Read cleanup, variant calling and alignment are the first steps in computational workflows. When we shift the start positions, the references are more capable to recover reads across the ends of circular genomes such as HPV [39]. Aligners such as Winnowmap2, Mapquik and minimap2 [40] can handle error-prone long reads well. Integration detection tools (ViR, VIRUSBreakend, nf-VIF and CLC Microbial Genomics workflows) report broken genes and identify viral-host breakpoints ([41][42][43]. Long-read sequencing, targeted capture and chromatin conformation data clarifies the functional relevance of disrupted genes in their regulatory neighborhoods [44][17]. Ultimately, translation of these tools to clinical practice will require validation of integration markers against longitudinal treatment responses that account for intrahost viral diversity and clonal selection under therapeutic pressure [45][3][46].

The power of bioinformatics in cancer virology is determined by the availability, biological relevance, reproducibility and interpretability of the output of computational tools. In particular, the performance of individual pipelines in identifying viral sequences, characterizing structural rearrangements, and detecting virus-host integration sites may differ when assessing complex or low abundance genomic signals. The choice of a computational workflow should therefore be driven by the research objective, the sequencing technology, the specimen characteristics and the expected analytical limitations. The integration of genomic, transcriptomic and chromatin interaction data with complementary computational approaches allows for a more complete understanding of the role of viruses in carcinogenesis. However, it is important to differentiate between computational predictions and experimentally validated biological mechanisms, especially when proposing diagnostic biomarkers or therapeutic targets.

Integrated bioinformatic pipelines provide a systematic approach to the study of viral genomic diversity, integration events and host molecular responses in cancer. Transparent reporting, independent validation and standardized analytical procedures are

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

used to maximize their scientific value. Integration of different types of molecular and sequencing data could improve identification of biologically relevant changes in viruses and development of candidate applications to diagnostics, prognostics, and therapeutics.

6. Scientific consensus

The literature review shows that bioinformatics has significantly improved the analysis capacity of cancer virology, specifically in the areas of virus-host molecular interaction, integration site characterization, and viral genome identification. However, the scientific value of these advances depends on separating biological causality and clinical significance from computational detection. The detection of viral nucleic acid in malignant tissues, although informative, is insufficient for the establishment of an oncogenic mechanism. To explain this more convincingly, it is important to integrate information on genomic structure, the activity of viral transcription, regulation of host genes and functional molecular evidence.

The combination of short-read and long-read sequencing is particularly beneficial methodically, as the two technologies avoid separate analytical limits. However, these advances in sequencing resolution need to be complemented with equally rigorous standards of computational reproducibility, contamination control, and experimental verification. These safeguards may enhance analytical sensitivity but may also result in increased uncertainty rather than more robust biological inferences. The study also illustrates the difference between being able to identify potential molecular targets and being able to demonstrate therapeutic efficacy.

Integration sites of viruses, oncogenic transcripts and regulatory changes may be opportunities for targeted intervention but the translation of these into clinically useful antiviral or anticancer therapies requires appropriate experimental design, preclinical studies and clinical trials. Bioinformatics is generally considered a bridge science between observations of virus genomes and testable hypotheses of cancer development. Future contributions will be underpinned by the successful integration of computational predictions, experimental evidence and clinically meaningful outcomes.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

7. Conclusion:

This review highlights the role of bioinformatics in improving our knowledge of cancer associated viruses, particularly in the areas of genome sequencing, viral integration, intrahost variant characterization, and virus-host molecular interaction. The literature available reports that computational methods have proven to be very useful in identifying and characterizing structurally viral genetic elements in malignant tissues, offering valuable insights into possible mechanisms of virus-associated carcinogenesis. Long-read sequencing platforms, including PacBio HiFi and Oxford Nanopore, have improved the capacity to resolve complex viral integration structures and genomic rearrangements that can be difficult to characterize using short-read sequencing alone. However, complete viral genomic characterization requires a combination of both sequencing approaches to allow the complementary strengths to be supported.

Moreover, the integration of genomic, transcriptomic and regulatory data may help to distinguish biologically relevant viral changes from incidental findings. These approaches are particularly useful for the study of expression of viral oncogenes, virus-host chimeric transcripts, and molecular changes that may be relevant for tumor progression. However, challenges remain in computational standardization, contamination and sequencing artifacts, and functional interpretation of viral genomic findings. The limitations highlight the need for rigorous quality control procedures, transparent analytical reporting and independent experimental validation. The clinical application of bioinformatic knowledge is an area of active research. Validation studies that are well designed can demonstrate the diagnostic, prognostic and therapeutic utility of computational analyzes, but they can also identify potential biomarkers and molecular therapeutic targets.

Bioinformatics provides a powerful platform to explore the molecular complexity of virus-associated cancers, but its greatest clinical impact will be realized by integrating functional molecular validation, high-resolution sequencing, and prospective clinical evidence. In the future, the integration may help develop more accurate diagnostics, reliable molecular biomarkers and targeted treatments for virus-associated cancers.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

References

1. Gu, Z. (2024). Oncogenic mechanisms of human tumor viruses. *Annual Review of Pathology*, 19, 145-170.

2. Xiao, Q. *et al.* (2025) "Viral oncogenesis in cancer: from mechanisms to therapeutics," *Signal Transduction and Targeted Therapy*, 10(1), p. 151. doi:10.1038/s41392-025-02197-9.

3. Li, S. *et al.* (2026) "Data from HPV Integration in Head and Neck Cancer: Downstream Splicing Events and Expression Ratios Linked with Poor Outcomes." doi:10.1158/1078-0432.c.8289922.

4. Consul, S., Robertson, A., & Vikalo, H. (2025). Detecting divergent viral sequences in tumor transcriptomes. *Computational and Structural Biotechnology Journal*, 23, 400-410.

5. Zapatka, M., et al. (2020). The landscape of viral associations in human cancers. *Nature*, 578(7793), 133-139.

6. Tang, G., Cho, M. and Wang, X. (2021) "OncoDB: an interactive online database for analysis of gene expression and viral infection in cancer," *Nucleic Acids Research*, 50. doi:10.1093/nar/gkab970.

7. Huang, X., et al. (2025). Multi-omics limitations in linking viral signals to host signaling pathways. *Trends in Genetics*, 41(2), 150-162.

8. Elbasir, A. *et al.* (2023) "A deep learning approach reveals unexplored landscape of viral expression in cancer," *Nature Communications*, 14(1), p. 785. doi:10.1038/s41467-023-36336-z.

9. Ye, R. *et al.* (2023) "Viral oncogenes, viruses, and cancer: a third-generation sequencing perspective on viral integration into the human genome," *Frontiers in Oncology*, 13, p. 1333812. doi:10.3389/fonc.2023.1333812.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

10. Nam, S., et al. (2025). Recognized causal agents of human cancers. *Lancet Oncology*, 26(1), 85-98.

11. Rodriguez, R. *et al.* (2020) "Tissue-associated microbial detection in cancer using human sequencing data," *BMC Bioinformatics*, 21, p. 523. doi:10.1186/s12859-020-03831-9.

12. Zhang, Z. *et al.* (2021) "Uniform genomic data analysis in the NCI Genomic Data Commons," *RISalud Institutional Health Repository of Andalucía*, 12(1), p. 1226. doi:10.1038/s41467-021-21254-9.

13. Seo, Y.D. *et al.* (2023a) "Abstract 641: VirMAP for cancer: Characterization of the intratumoral virome in virally-associated cancers and a resource for investigators," *Cancer Research*, 83, p. 641. doi:10.1158/1538-7445.am2023-641.

14. Elbasir, A. *et al.* (2023) "A deep learning approach reveals unexplored landscape of viral expression in cancer," *Nature Communications*, 14(1), p. 785. doi:10.1038/s41467-023-36336-z.

15. Hu, Y., et al. (2025). Architecture of cancer-associated viral genomes. *Journal of Virology*, 99(3), e00120-25.

16. Giunta, S. (2021) "Decoding human cancer with whole genome sequencing: a review of PCAWG Project studies published in February 2020," *Cancer and Metastasis Reviews*, 40(3), pp. 909–924. doi:10.1007/s10555-021-09969-z.

17. Porter, V., et al. (2024). Structural rearrangements in HPV-positive cervical tumors. *Nature Genetics*, 56(3), 420-432.

18. Zhou, Y. *et al.* (2022) "A comprehensive SARS-CoV-2–human protein–protein interactome reveals COVID-19 pathobiology and potential host therapeutic targets," *Nature Biotechnology*, 41(1), pp. 128–139. doi:10.1038/s41587-022-01474-0.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

19. Artesi, M., et al. (2021). A fast and accurate long-read sequencing workflow for viral detection. *Genome Medicine*, 13(2), 45.

20. Bergmann, J., Pacholewska, A., & Schweiger, S. (2025). Hybrid sequencing strategies for low-copy viral detection. *Nucleic Acids Research*, 53(2), e15.

21. Yang, S. *et al.* (2022) "Whole Genome Assembly of Human Papillomavirus by Nanopore Long-Read Sequencing," *Frontiers in Genetics*, 12, p. 798608. doi:10.3389/fgene.2021.798608.

22. Akagi, K., et al. (2023). Long-read sequencing in viral integration analysis. *Nature Communications*, 14(1), 1120.

23. Elrick, H. *et al.* (2025) "SAVANA: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing," *Nature Methods*, 22(7), pp. 1436–1446. doi:10.1038/s41592-025-02708-0.

24. Cui, X. *et al.* (2023) "Multiple HPV integration mode in the cell lines based on long-reads sequencing," *Frontiers in Microbiology*, 14, p. 1294146. doi:10.3389/fmicb.2023.1294146.

25. Arsdale, A.V. *et al.* (2024) "Structure and transcription of integrated HPV DNA in vulvar carcinomas," *npj Genomic Medicine*, 9(1), p. 35. doi:10.1038/s41525-024-00418-8.

26. Ho, D.W., Lyu, X. and Ng, I.O. (2021) "Viral integration detection strategies and a technical update on Virus-Clip," *Biocell*, 45(6), pp. 1495–1500. doi:10.32604/biocell.2021.017227.

27. Chen, Z. and He, X. (2021) "Application of third-generation sequencing in cancer research," *Medical Review*, 1(2), pp. 150–171. doi:10.1515/mr-2021-0013.

28. Li, C., Yeh, S. and Chen, P. (2022) "Circulating Virus–Host Chimera DNAs in the Clinical Monitoring of Virus-Related Cancers," *Cancers*, 14(10), p. 2531. doi:10.3390/cancers14102531.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

29. Morodomi, S. *et al.* (2025) "Robust HPV -16 Detection Workflow for Formalin-Fixed Cancer Tissue and Its Application for Oral Squamous Cell Carcinoma," *Cancer Medicine*, 14(4). doi:10.1002/cam4.70544.

30. Underhill, H.R. (2021) "Leveraging the Fragment Length of Circulating Tumour DNA to Improve Molecular Profiling of Solid Tumour Malignancies with Next-Generation Sequencing: A Pathway to Advanced Non-invasive Diagnostics in Precision Oncology?," *Molecular Diagnosis & Therapy*, 25(4), pp. 389–408. doi:10.1007/s40291-021-00534-6.

31. Ghaddar, B., Blaser, M.J. and De, S. (2026) "Reliable detection of Host-Microbe Signatures in cancer using PRISM," *Cancer Cell*, 44(4), p. 879. doi:10.1016/j.ccell.2026.01.007.

32. Dohlman, A.B. *et al.* (2026) "Biodiversity and biogeography of the multi-kingdom cancer microbiome," *Cell*, 189(12), p. 3817. doi:10.1016/j.cell.2026.04.015.

33. Mühr, L.S.A. *et al.* (2021) "Human Papillomavirus Detection by Whole-Genome Next-Generation Sequencing: Importance of Validation and Quality Assurance Procedures," *Viruses*, 13(7), p. 1323. doi:10.3390/v13071323.

34. Solis-Ponce, L. *et al.* (2025) "Genotype distribution and molecular characterization of HPV in the Peruvian amazon: insights into prevalence, lineage diversity, and viral integration," *Scientific Reports*, 15(1), p. 32535. doi:10.1038/s41598-025-18455-3.

35. Maina, T.K. *et al.* (2025) "Scalable long-read nanopore HPV16 amplicon-based whole-genome sequencing," *Scientific Reports*, 15(1), p. 34892. doi:10.1038/s41598-025-18664-w.

36. Jurasz, H., Pawłowski, T. and Perlejewski, K. (2021) "Contamination Issue in Viral Metagenomics: Problems, Solutions, and Clinical Perspectives," *Frontiers in Microbiology*, 12, p. 745076. doi:10.3389/fmicb.2021.745076.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

37. Song, X. *et al.* (2020) "Characterization of HPV and host genome interactions in primary head and neck cancers," *UNC Libraries* [Preprint]. doi:10.17615/wkmh-9g54.

38. Yu, L. *et al.* (2024) "HPV oncogenes expressed from only one of multiple integrated HPV DNA copies drive clonal cell expansion in cervical cancer," *PubMed Central*, 15(5). doi:10.1128/mbio.00729-24.

39. Sasa, N. *et al.* (2025) "Intratumor heterogeneity of HPV integration in HPV-associated head and neck cancer," *Nature Communications*, 16(1), p. 1052. doi:10.1038/s41467-025-56150-z.

40. Ermini, L. and Driguez, P. (2024) "The Application of Long-Read Sequencing to Cancer," *Cancers*, 16(7), p. 1275. doi:10.3390/cancers16071275.

41. Cameron, D. *et al.* (2021) "VIRUSBreakend: Viral Integration Recognition Using Single Breakends," *Bioinformatics*, 37(19), pp. 3115–3119. doi:10.1093/bioinformatics/btab343.

42. Mainguené, J. *et al.* (2022) "Human papilloma virus integration sites and genomic signatures in head and neck squamous cell carcinoma," *Molecular Oncology*, 16(16), pp. 3001–3016. doi:10.1002/1878-0261.13219.

43. Shen-Gunther, J., Cai, H. and Wang, Y. (2022) "HPV Integration Site Mapping: A Rapid Method of Viral Integration Site (VIS) Analysis and Visualization Using Automated Workflows in CLC Microbial Genomics," *International Journal of Molecular Sciences*, 23(15), p. 8132. doi:10.3390/ijms23158132.

44. Groves, I.J. *et al.* (2021) "Short- and long-range cis interactions between integrated HPV genomes and cellular chromatin dysregulate host gene expression in early cervical carcinogenesis," *PLoS Pathogens*, 17(8). doi:10.1371/journal.ppat.1009875.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin

45. Qin, T. *et al.* (2021) "Molecular Tumor Subtypes of HPV-Positive Head and Neck Cancers: Biological Characteristics and Implications for Clinical Outcomes," *Cancers*, 13(11), p. 2721. doi:10.3390/cancers13112721.

46. Rudroff, T. (2025) "Convergent Mechanisms in Virus-Induced Cancers: A Perspective on Classical Viruses, SARS-CoV-2, and AI-Driven Solutions," *Infectious Disease Reports*, 17(2), p. 33. doi:10.3390/idr17020033.

Creative Commons Attribution-Non-Commercial 4.0 International License

Published by DAC Democratic Arabic Center
DAC Democratic Arabic Center GmbH, Berlin