Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Gene Length Bias in Human Papillomavirus Integration Sites: A Statistical Analysis Using KNIME

Sahil Khandekar

sahil04mail@gmail.com

Urja Bait

Short Report

Keywords: HPV integration, gene length bias, viral oncogenesis, genomic hotspots, statistical genomics, HPV16, HPV18

Posted Date: March 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8941085/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Integration of high-risk Human Papillomavirus (HPV) into the human genome represents a critical event in carcinogenesis. While recurrent integration hotspots are well documented, the structural determinants governing host gene susceptibility remain incompletely defined. Here, we evaluate whether human gene length influences HPV integration probability using VISDB integration datasets for HPV16 and HPV18 merged with Ensembl BioMart annotations. Statistical analysis performed in KNIME reveals that HPV-integrated genes exhibit significantly greater length than non-integrated genes (Two-sided Wilcoxon-Mann-Whitney, p = 0.0001). Distributional analysis demonstrates a consistent rightward shift in log-transformed gene length among integrated genes. Genomic validation using the UCSC Genome Browser confirms integration within established oncogenic regions, including 8q24 near MYC. These findings identify gene length as a measurable structural bias contributing to HPV integration susceptibility, while underscoring the role of genomic context in hotspot formation.

1 Introduction

High-risk HPV types, particularly HPV16 and HPV18, promote malignant transformation through integration into host genomic DNA. Integration events can disrupt tumor suppressor genes, amplify oncogenic loci, and alter regulatory architecture. Recurrent hotspots such as *PVT1*, *FHIT*, *MACROD2*, and loci within 8q24 have been consistently reported [1-3]. However, whether intrinsic structural properties of host genes influence integration frequency remains unclear.

We hypothesized that longer genes, by occupying larger genomic intervals, present increased opportunity for viral integration.

2 Materials and Methods

2.1 Data Acquisition

HPV16 and HPV18 integration site data were obtained from VISDB [4]. Human gene annotations were retrieved from Ensembl BioMart (GRCh38) [5]. Integration coordinates were mapped to corresponding host genes using genomic overlap criteria.

2.2 Computational Workflow

All preprocessing and statistical analyses were performed using the KNIME Analytics Platform [7].

VISDB Data Processing

HPV16 and HPV18 datasets were imported separately using CSV Reader nodes and subsequently concatenated. The following preprocessing steps were applied:

- Column filtering to retain only relevant gene-level variables
- Removal of rows with missing gene annotations
- Cell splitting of multi-gene integration entries
- Ungrouping to separate duplicated gene entries
- Removal of duplicate rows
- Elimination of unnecessary identifiers (e.g., Ensembl ID where redundant)

For hotspot identification, gene counts were aggregated using GroupBy nodes, followed by rule-based filtering and ranking using Sorter and Top K Row Filter nodes.

Ensembl BioMart Processing

Gene annotation data from Ensembl BioMart were:

- Filtered to retain essential genomic coordinates
- Cleaned to remove unnecessary chromosome entries
- Gene length calculated from genomic start and end positions
- Log transformation applied to gene length values to correct skewness

Data Integration

Processed VISDB and BioMart datasets were merged using genomic overlap criteria through Joiner nodes. A rule-based column was generated to classify genes into:

- Integrated
- Non-integrated

Statistical Validation

Group comparison was performed using the Wilcoxon–Mann–Whitney test node. Distributional visualization included:

- Box plot
- Kernel density estimation

2.3 Statistical Analysis

Gene length distributions between integrated and non-integrated groups were compared using the Wilcoxon–Mann–Whitney test. Statistical significance was defined at $p < 0.05$.

3 Results

3.1 Gene Length Distribution Bias

Integrated genes demonstrated higher median log-transformed lengths relative to non-integrated genes (Fig. 3). Distributional differences are further illustrated in the density plot (Fig. 4). Statistical testing confirmed a significant difference between groups ($p = 0.0001$), supporting a gene length bias in HPV integration.

A total of 10063 integrated genes and 13199 non-integrated genes were included in the analysis.

3.2 Hotspot Context

Genomic inspection using the UCSC Genome Browser [6] revealed that several integration hotspots reside within established oncogenic regions, including 8q24 proximal to *MYC*. Importantly, certain shorter genes within highly active regulatory domains also exhibited integration events, suggesting that chromatin state and regulatory architecture influence integration beyond structural length bias. The computational workflow used for hotspot identification is shown in Fig. 1 and Fig. 2.

4 Discussion

Our findings demonstrate that human gene length is significantly associated with HPV integration susceptibility. The most direct interpretation is probabilistic: longer genes span greater genomic territory and therefore present increased opportunity for viral insertion events.

However, gene length alone does not fully explain hotspot localization. Integration within shorter genes positioned in transcriptionally active or enhancer-rich regions indicates that epigenetic and chromatin features likely refine integration targeting. These observations support a composite model in which structural exposure establishes baseline susceptibility, while regulatory context determines hotspot specificity.

5 Conclusion

Human gene length is positively associated with HPV integration probability. Integrated genes are significantly longer than non-integrated genes. Although gene length contributes to susceptibility, integration hotspot formation likely arises from interaction between structural and regulatory genomic features.

6 Limitations

This analysis evaluated a single genomic feature (gene length) and Integration frequency was not normalized by gene density or chromosomal length distribution.

7 Future Directions

Future investigations should integrate chromatin accessibility, histone modification profiles, gene expression data, replication timing, and three-dimensional genome organization to develop predictive models of HPV integration susceptibility.

Abbreviations

HPV: Human Papillomavirus; VISDB: Viral Integration Site Database.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

HPV integration data were obtained from VISDB (http://bioinfo.ahu.edu.cn/VISDB). Gene annotation data were retrieved from Ensembl BioMart (https://www.ensembl.org/biomart). Genomic validation was performed using the UCSC Genome Browser (https://genome.ucsc.edu).

Competing Interests

The authors declare no competing interests.

Funding

This research received no external funding.

Authors' contributions

Sahil Khandekar conceptualized the study, performed data analysis, and drafted the manuscript. Urja Bait contributed to data interpretation and manuscript revision. Both authors approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Bodelon C, Untereiner ME, Machiela MJ, Vinokurova S, Wentzensen N. Genomic characterization of viral integration sites in HPV-related cancers. *Int J Cancer.* 2016;139(9):2001–2011.
2. Hu Z, Zhu D, Wang W, et al. Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet.* 2015;47(2):158–163.
3. Akagi K, Li J, Broutian TR, et al. Genome-wide analysis of HPV integration in human cancers reveals recurrent, focal genomic instability. *Genome Res.* 2014;24(2):185–199.
4. Cao L, et al. VISDB: a manually curated database of viral integration sites in the human genome. *Nucleic Acids Res.* 2020;48(D1):D633–D641.
5. Kinsella RJ, Kähäri A, Haider S, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford).* 2011;2011:bar030.
6. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996–1006.
7. Berthold MR, Cebron N, Dill F, et al. KNIME – the Konstanz Information Miner: version 2.0 and beyond. *SIGKDD Explor.* 2009;11(1):26–31.

Figures

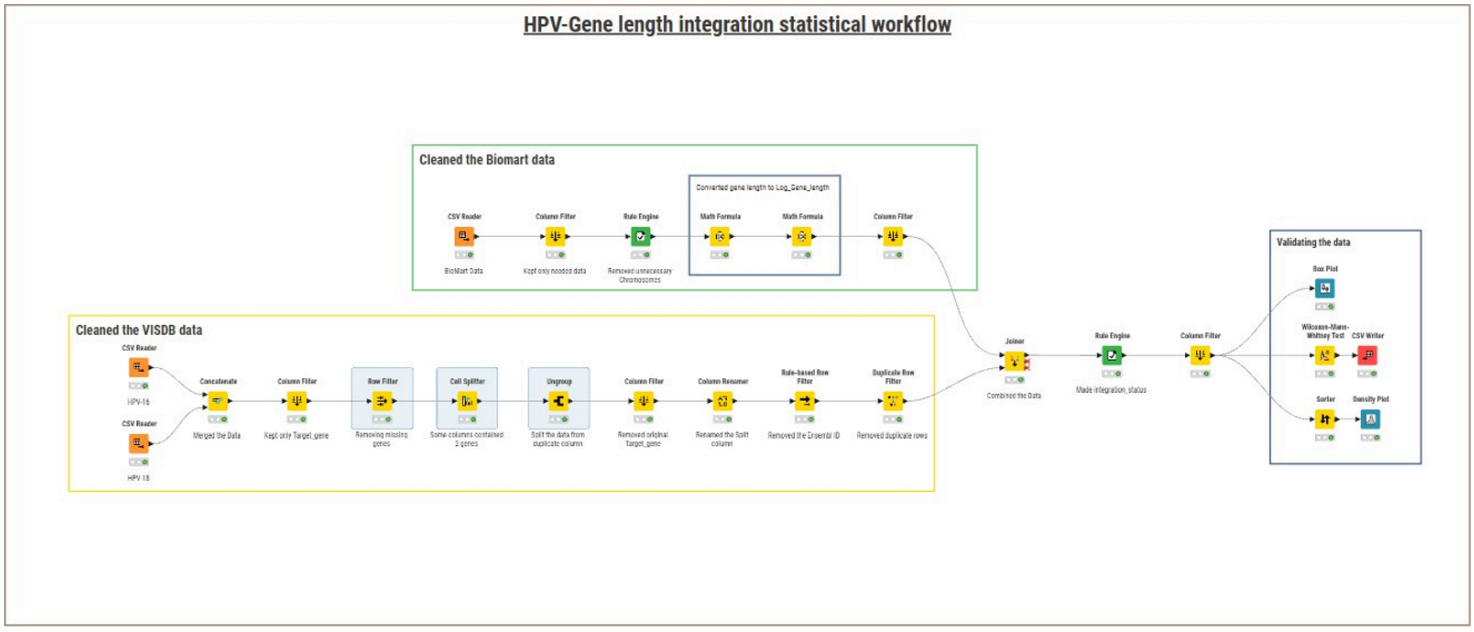


Figure 1

KNIME workflow for preprocessing VISDB and BioMart datasets, data integration, gene classification, and statistical testing.

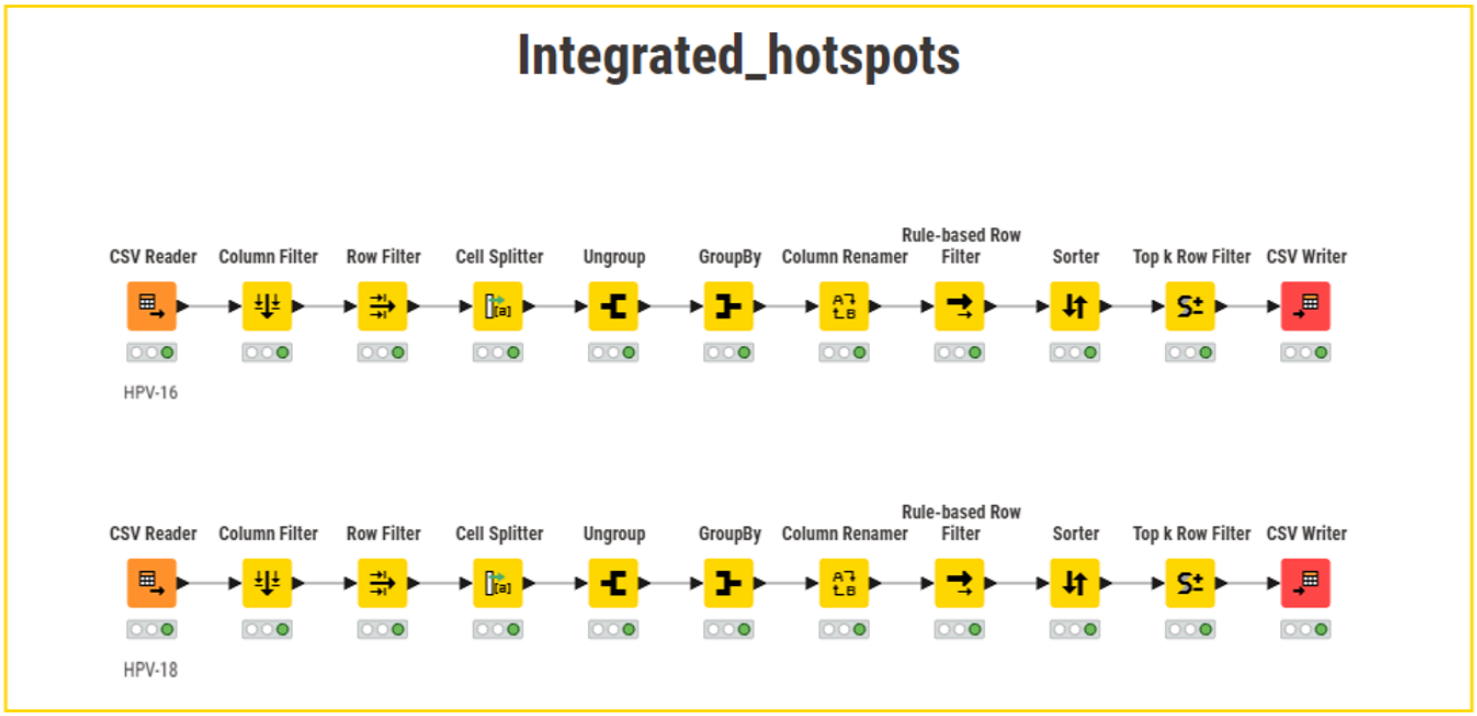


Figure 2

Workflow for identification and ranking of HPV16 and HPV18 integration hotspots using aggregation and top-k filtering.

(This workflow is available on GitHub link:- https://github.com/iamsahill4/HPV-KNIME-Workflow)

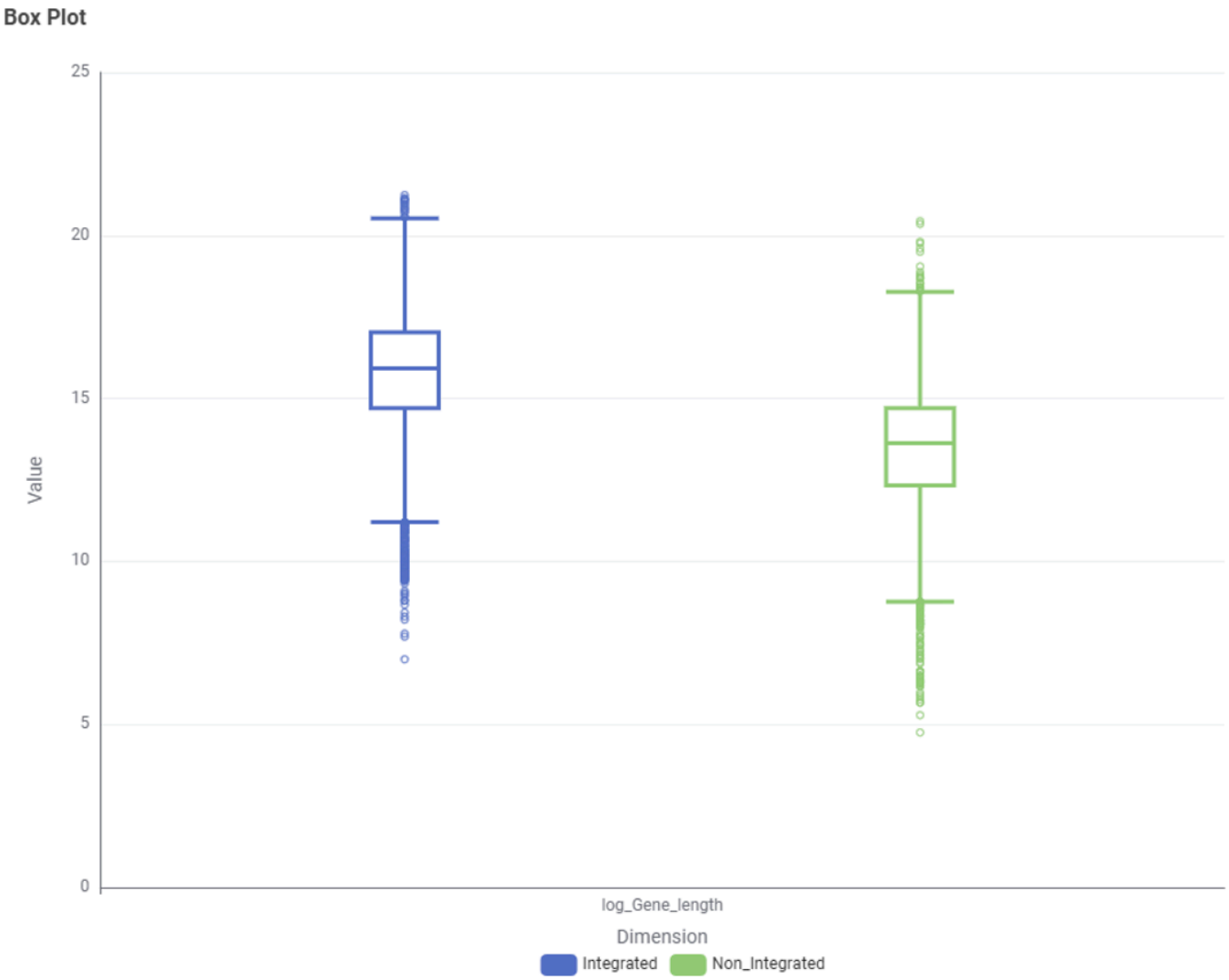


Figure 3

Box plot comparing log-transformed gene lengths between integrated and non-integrated genes.

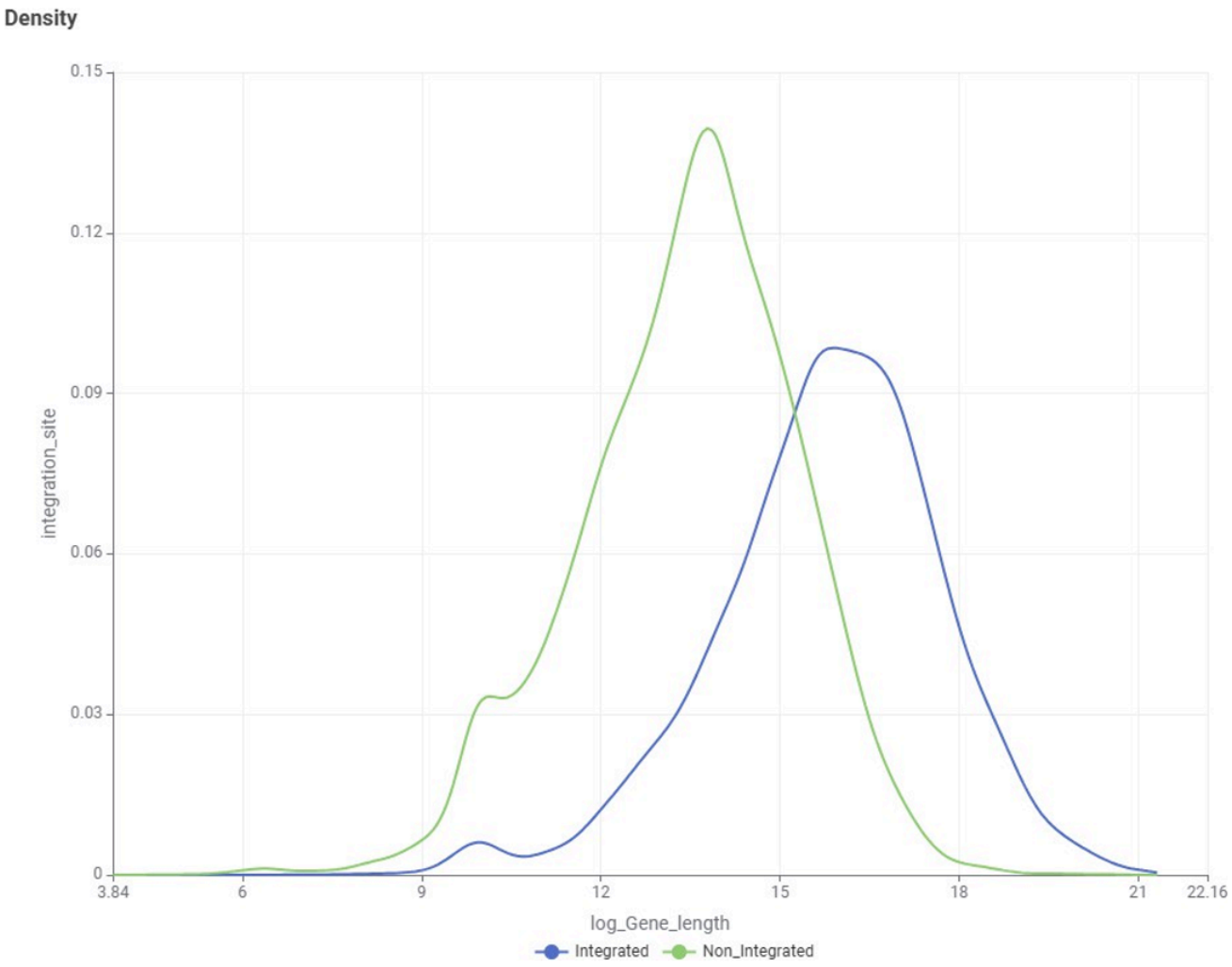


Figure 4

Kernel density plot showing rightward shift in gene length distribution among integrated genes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- UCSCgenevalidation.docx