Biomedical Research and Therapy 2026, 13(08):8913-8923

Open Access Full Text Article

Original Research

Exploratory Transcriptomic Screening and Clinical Validation of Elevated Circulating HERV-K (HML-2)-Associated *env* Expression in Breast Cancer

Hang Giang Thi Phan[1,*,†], Thanh-loan Tran[1,†], Tuyet-ngoc Thi Tran[2], Phong-son Dinh[3], Ai-nhi Thi Tran[4], Bao-chi Thi Le[5], Xuan-dung Ho[6]

ABSTRACT

Background: Human endogenous retrovirus K (HERV-K, HML-2) is among the most transcriptionally active endogenous retroviral families and has been implicated in breast cancer (BC) pathogenesis. This study aimed to explore HERV-K (HML-2)-associated *env* expression and evaluate its utility as a potential non-invasive circulating biomarker for BC. **Methods:** Public RNA-seq data from breast epithelial transformation models (GSE84275) were screened to identify dysregulated HERV-K (HML-2)-associated *env* signals. An upregulated HERV-K (HML-2)-associated *env* signal overlapping the chromosome 4p16.1 region was prioritized for clinical evaluation. Peripheral family-level HERV-K (HML-2)-associated *env* expression was subsequently quantified by quantitative real-time PCR (qRT-PCR) in peripheral blood leukocytes obtained from 56 BC patients and 45 healthy controls. **Results:** Peripheral HERV-K (HML-2)-associated *env* expression was significantly higher in BC patients than in healthy controls (median: 0.9111 vs. 0.5284, $P = 0.0003$). Receiver operating characteristic (ROC) curve analysis demonstrated moderate discriminatory performance (area under the curve [AUC] = 0.7075, 95% CI: 0.6051–0.8100, $P = 0.0005$). Expression levels were significantly elevated in early-stage disease (stages I–IIA) compared with advanced stages ($P < 0.0001$). Multivariable logistic regression analysis confirmed that elevated *env*-associated expression remained an independent predictor of BC (adjusted OR = 12.66, 95% CI: 3.14–51.02, $P < 0.001$). **Conclusions:** Elevated peripheral HERV-K (HML-2)-associated *env* expression represents a circulating molecular signal prominently associated with early-stage breast cancer, highlighting its potential utility as a complementary non-invasive diagnostic biomarker.

Key words: Endogenous retroviruses, HERV-K (HML-2), *env* gene expression, Breast cancer, Peripheral blood, Biomarkers, Liquid biopsy

[1]*Department of Immunology and Pathophysiology, University of Medicine and Pharmacy, Hue University, Viet Nam*

[2]*Department of Microbiology and Institute of Biomedicine, University of Medicine and Pharmacy, Hue University, Hue, Viet Nam*

[3]*College of Medicine and Pharmacy, Duy Tan University, Da Nang, Viet Nam*

[4]*Medical Laboratory Technology Unit, Medical Technology Department, University of Medicine and Pharmacy, Hue University, Viet Nam*

[5]*Department of Microbiology, University of Medicine and Pharmacy, Hue University, Viet Nam*

[6]*Oncology Department, University of Medicine and Pharmacy, Hue University, Viet Nam*

[†]Those authors contributed equally to the study.

Correspondence

Hang Giang Thi Phan, Department of Immunology and Pathophysiology, University of Medicine and Pharmacy, Hue University, Viet Nam

Email: pthgiang.med@hueuni.edu.vn

INTRODUCTION

Human endogenous retroviruses (HERVs) constitute approximately 8% of the human genome and are increasingly recognized as dynamic, functional components of the human transcriptome [1,2]. Although typically silenced by host epigenetic mechanisms, aberrant HERV activation has been implicated in chronic inflammation, neurodegenerative disorders, and multiple human malignancies [3–7].

Among HERV families, HERV-K (HML-2) retains relatively intact open reading frames and represents one of the most transcriptionally active and evolutionary young retroviral groups in the human genome [8,9]. Elevated HERV-K (HML-2) *env* expression has been associated with epithelial-to-mesenchymal transition (EMT) and the activation of oncogenic signaling pathways, including the extracellular signal-regulated kinase (ERK) cascade [10–12]. In breast cancer (BC), aberrant HERV-K transcription correlates with genomic instability and aggressive tumor phenotypes [13–15]. However, the diagnostic and biomarker relevance of circulating, leukocyte-derived HERV-K transcriptional signals remains insufficiently characterized.

Breast cancer remains a leading cause of cancer-related morbidity and mortality among women worldwide, characterized by marked molecular and clinical heterogeneity [16,17]. Although conventional oncogenic signaling cascades have been extensively studied, the contribution of retroelement-associated transcriptional activation to early epithelial transformation is not yet fully elucidated. Advances in high-throughput RNA sequencing (RNA-seq) now facilitate exploratory characterization of repetitive-element transcription during malignant transformation [18,19]. Integrating transcriptomic screening with clinical validation may therefore clarify the diagnostic potential of peripheral HERV-K (HML-2)-associated *env* expression signals.

Cite this article : Phan THG, Tran TL, Tran TTN, Dinh PS, Tran TAN, Le TBC, Ho XD. **Exploratory Transcriptomic Screening and Clinical Validation of Elevated Circulating HERV-K (HML-2)-Associated *env* Expression in Breast Cancer**. *Biomed. Res. Ther.* 2026; 13(08):8913-8923.

History

- Received: May 22, 2026
- Accepted: Aug 02, 2026
- Published Online: Aug 31, 2026

DOI : 10.15419/bmrat.v13i8.1095

In this study, we screened public RNA-seq data from breast epithelial transformation models (GSE84275) and identified an upregulated HERV-K (HML-2) *env*-associated sequence overlapping the chromosome 4p16.1 locus (chr4:9,123,514–9,133,075). We subsequently quantified peripheral blood leukocyte-derived HERV-K (HML-2)-associated *env* expression at the family level in an independent clinical cohort of 56 BC patients and 45 healthy controls, evaluating its diagnostic performance and association with clinicopathological stages.

MATERIALS AND METHODS

Study Design

This study integrated *in silico* transcriptomic discovery with clinical cohort validation. Public RNA-seq data from dataset GSE84275[20] deposited in the Gene Expression Omnibus (GEO)[21] were analyzed to prioritize dysregulated HERV-K (HML-2)-associated *env* signals in breast epithelial transformation models. Subsequently, family-level *env* expression was quantified via qRT-PCR in peripheral blood samples from BC patients and healthy controls. The clinical diagnostic validation component was conducted and reported in strict accordance with the Standards for Reporting Diagnostic Accuracy (STARD 2015) guidelines (Supplementary File S1).

Transcriptomic Screening Using GEO RNA-seq Data

Normalized RNA-seq data from GSE84275 were retrieved from GEO and evaluated across one non-transformed human mammary epithelial model (HME) and three transformed models (HMLE-Ras, HMLE-Her2, and HCC1954)[20]. HERV-K (HML-2)-associated *env* signals were identified based on repetitive element annotations and genomic coordinates overlapping known HERV-K loci. Relative expression differences between transformed and non-transformed models were evaluated descriptively using normalized expression values. Given the exploratory nature and limited sample size of this public dataset, formal hypothesis testing and multiple-testing corrections were not performed; the dataset was utilized primarily for candidate prioritization. The HERV-K (HML-2)-associated *env* signal overlapping chromosome 4p16.1 (chr4:9,123,514–9,133,075) demonstrated consistent upregulation across transformed models and was selected for clinical validation. Genomic coordinates were mapped to the GRCh38/hg38 reference assembly via the UCSC Genome Browser[22,23].

In Silico Sequence Verification

The candidate HERV-K (HML-2) *env*-associated sequence was aligned against the human reference genome using BLASTn (NCBI) to verify sequence similarity and genomic overlap. Annotated genes situated within ±500 kb of the overlapping genomic region were extracted from the UCSC Table Browser using GENCODE v46 annotations[24]. Given the multicopy nature of HERV-K (HML-2) proviruses, this analysis was intended to confirm family-level sequence identity and genomic context rather than definitive locus-specific transcriptional assignment[25].

Study Participants

Female patients with primary breast cancer were prospectively recruited from Hue University of Medicine and Pharmacy Hospital between January 2025 and January 2026. Diagnoses were histopathologically confirmed, and anatomical tumor staging was classified according to the American Joint Committee on Cancer (AJCC) TNM staging manual (8th edition)[26]. Early-stage disease was defined as stages I to IIA, and advanced-stage disease was defined as stages IIB to IV[27].

Inclusion criteria were: (1) female sex aged ≥30 years; (2) histopathologically confirmed primary breast carcinoma; and (3) no prior history of systemic chemotherapy, radiotherapy, or endocrine therapy before blood collection. Exclusion criteria comprised: (1) previous history of other malignancies; (2) autoimmune or systemic inflammatory disorders; (3) acute infectious episodes within two weeks prior to sampling; and (4) pregnancy or lactation. Age-matched healthy controls were recruited from women undergoing routine annual health check-ups at the same institution during the corresponding timeframe. Control individuals exhibited no history of malignancy, autoimmune disease, or active infection. All participants provided written informed consent. The protocol was approved by the Institutional Ethics Committee of Hue University of Medicine and Pharmacy, Hue University (Approval No. H2025/426) and conducted in accordance with the Declaration of Helsinki (1975).

Clinical Data Collection and Blood Sampling

Baseline demographic characteristics and laboratory parameters, including complete blood counts and fasting blood glucose levels, were collected from electronic medical records. In BC patients, serum

concentrations of cancer antigen 15-3 (CA15-3), carcinoembryonic antigen (CEA), and C-reactive protein (CRP) were determined through routine hospital laboratory assays. Fasting peripheral venous blood (4 mL) was collected in the morning: 2 mL in EDTA tubes for leukocyte isolation and RNA extraction, and 2 mL in non-anticoagulated tubes for biochemical testing. Samples were processed within 2 hours of venipuncture. Leukocytes were isolated using red blood cell (RBC) lysis buffer (Solarbio, Beijing, China) and stored at −80 ℃ until analysis.

RNA Isolation and Quantitative Real-Time PCR

Total RNA was extracted from isolated leukocytes using the AxyPrep Total RNA Miniprep Kit (Axygen, China). RNA concentration and optical purity (A260/A280 ratio) were assessed spectrophotometrically. Reverse transcription was performed using the HiScript® III RT SuperMix kit (Vazyme, China), which incorporates an initial genomic DNA elimination step. Quantitative real-time PCR (qRT-PCR) was executed using SYBR Green chemistry on a StepOne™ Real-Time PCR System (Applied Biosystems, Foster City, CA, USA). Reactions were set up in a 20 μL volume containing 2× SYBR Green Master Mix, 0.4 μM forward and reverse primers, cDNA template, and nuclease-free water. All experimental procedures adhered strictly to the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines.

Target gene specificity was confirmed by post-amplification melting curve analysis, demonstrating single, discrete dissociation peaks (Supplementary Figure S1). Reactions were run in technical duplicates with no-template controls (NTC) and no-reverse-transcriptase (no-RT) controls in each batch. Relative HERV-K (HML-2)-associated *env* expression was calculated using the comparative $2^{-\Delta\Delta Ct}$ method, with glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) serving as the internal reference gene[28]. A single pooled cDNA calibrator derived from five BC patients was included across all runs to ensure batch-to-batch comparability: ΔCt = Ct(*env*) − Ct(*GAPDH*), and ΔΔCt = ΔCt(sample) − ΔCt(calibrator). Validated primers targeting conserved HERV-K (HML-2) *env* family sequences were utilized[29] (amplicon size: 166 bp; Table 1).

Statistical Analysis

Statistical analyses were carried out using SPSS 25.0 (IBM Corp., Armonk, NY, USA) and GraphPad Prism 9.0 (GraphPad Software, San Diego, CA, USA). Continuous variables were evaluated for distributional normality using the Shapiro–Wilk test and are presented as mean ± standard deviation (SD) for normally distributed data or median (interquartile range, IQR: 25th–75th percentiles) for non-normally distributed data. The sample size was exploratory and not predetermined.

Differences between two groups were assessed using the independent Student's *t*-test or the Mann–Whitney *U* test, as appropriate. Differences among three or more groups were evaluated using the Kruskal–Wallis test followed by Dunn's *post hoc* multiple comparisons test. Binary logistic regression analysis was conducted to identify factors independently associated with breast cancer. Multicollinearity among independent variables was assessed using variance inflation factor (VIF) and tolerance metrics. Receiver operating characteristic (ROC) curve analysis was performed to determine diagnostic accuracy, reporting the area under the curve (AUC) and 95% confidence intervals (CI). All statistical tests were two-tailed, and $P < 0.05$ was defined as statistically significant.

RESULTS

Transcriptomic Screening Identifies Upregulation of HERV-K (HML-2) *env* in Transformed Breast Epithelial Models

Exploratory screening of normalized RNA-seq profiles from GSE84275 demonstrated marked transcriptional elevation of an HERV-K (HML-2) *env*-associated signal overlapping the genomic interval chr4:9,123,514–9,133,075 in transformed breast epithelial cell lines (HMLE-Ras, HMLE-Her2, and HCC1954) relative to non-transformed mammary epithelial cells (HME). These transcriptomic findings served as a hypothesis-generating basis for clinical evaluation in patient-derived peripheral blood samples.

In Silico Verification Confirms Genomic Localization of the Prioritized HERV-K (HML-2) *env* Sequence

BLASTn alignment of the candidate sequence revealed 100% identity (7,836/7,836 bp, 0 gaps, E-value = 0.0) against the human reference provirus HERV-K (HML-2) sequence (JN675026.1; Supplementary Figure S2), corresponding to the HML-2_4p16.1a locus at chromosome 4p16.1 (chr4:9,123,514–9,133,075)[20]. Because HERV-K elements exhibit high sequence paralogy, this *in silico* confirmation established

Table 1: Oligonucleotide primer sequences used for quantitative real-time PCR (qRT-PCR) amplification of the HERV-K (HML-2) *env* family and the *GAPDH* endogenous reference gene. Primers targeting conserved regions of the HERV-K (HML-2) *env* gene yield a specific 166 bp amplicon, while *GAPDH* primers amplify an endogenous reference control across leukocyte cDNA samples.

Target Gene	Primer	Sequence (5′ → 3′)	Amplicon Size
HERV-K (HML-2) *env*	Forward	GCTGTCTCTTCGGAGCTGTT	166 bp
HERV-K (HML-2) *env*	Reverse	CTGAGGCAATTGCAGGAGTT	166 bp
GAPDH	Forward	CAAGGAGTAAGACCCCTGGAC	131 bp
GAPDH	Reverse	TCTACATGGCAACTGTGAGGAG	131 bp

Abbreviations: bp, base pairs; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; HERV-K, human endogenous retrovirus K; HML-2, human MMTV-like 2; qRT-PCR, quantitative reverse transcription polymerase chain reaction.

family-level identity and genomic overlap without inferring exclusive single-locus transcription.

Genomic Context of the Candidate HERV-K (HML-2) Locus

Annotation via the UCSC Genome Browser (GRCh38/hg38) demonstrated that the candidate locus at 4p16.1 possesses a canonical full-length proviral architecture flanked by long terminal repeat (LTR5) sequences (Figure 1), known to harbor potent promoter and enhancer activities[30]. In addition, interrogation of the ±500 kb flanking region identified several pseudogenes and a microRNA gene (*MIR548L2*) (Supplementary Table S1).

Baseline Characteristics of Study Participants

A total of 101 female participants, comprising 56 BC patients and 45 age-matched healthy controls, were enrolled. Mean age did not differ significantly between BC patients and controls (51.50 vs. 47.00 years, $P = 0.1632$), confirming baseline comparability. BC patients exhibited significantly lower red blood cell counts (4.254 ± 0.4212 vs. 4.494 ± 0.2522 × 10^{12}/L, $P = 0.0011$), lower hemoglobin levels (120.9 ± 12.06 vs. 133.6 ± 7.405 g/L, $P < 0.0001$), and a reduced lymphocyte percentage (28.23 ± 8.498% vs. 33.26 ± 6.312%, $P = 0.0013$). Conversely, platelet counts were moderately elevated in BC patients (265.5 vs. 235.0 × 10^{9}/L, $P = 0.0347$). Total leukocyte count and fasting blood glucose did not differ between cohorts (all $P > 0.05$; Table 2, Supplementary Table S2).

HERV-K (HML-2)-Associated *env* Expression is Elevated in Breast Cancer Patients Compared with Healthy Controls

qRT-PCR quantification revealed significantly higher peripheral blood leukocyte HERV-K (HML-2) *env* expression in BC patients than in healthy controls (median [IQR]: 0.9111 [0.5364–1.312] vs. 0.5284 [0.2848–0.7876], $P = 0.0003$; Figure 2). The Hodges–Lehmann median difference between cohorts was 0.3173.

Logistic Regression Analysis of Factors Associated with Breast Cancer

Univariate logistic regression revealed that elevated HERV-K (HML-2) *env* expression (OR = 5.28, 95% CI: 1.90–14.67, $P = 0.001$), lower lymphocyte percentage (OR = 0.92, 95% CI: 0.86–0.97, $P = 0.003$), and lower hemoglobin level (OR = 0.87, 95% CI: 0.82–0.92, $P < 0.001$) were significantly associated with BC. In multivariable logistic regression, elevated *env* expression remained independently associated with BC (adjusted OR = 12.66, 95% CI: 3.14–51.02, $P < 0.001$). Patient age also demonstrated an independent association (adjusted OR = 1.05, 95% CI: 1.00–1.10, $P = 0.044$), while hemoglobin level maintained an inverse association (adjusted OR = 0.85, 95% CI: 0.79–0.91, $P < 0.001$). Lymphocyte percentage was not significant after adjustment ($P = 0.179$; Table 3). Multicollinearity diagnostics confirmed acceptable tolerance (0.832–0.996) and VIF values (1.004–1.203; Supplementary Table S3).

Diagnostic Performance of HERV-K (HML-2)-Associated *env* Expression

ROC curve analysis showed that peripheral HERV-K (HML-2) *env* expression effectively discriminated BC patients from healthy controls, achieving an AUC of 0.7075 (95% CI: 0.6051–0.8100, $P = 0.0005$; Figure 3), reflecting moderate diagnostic discriminatory accuracy.

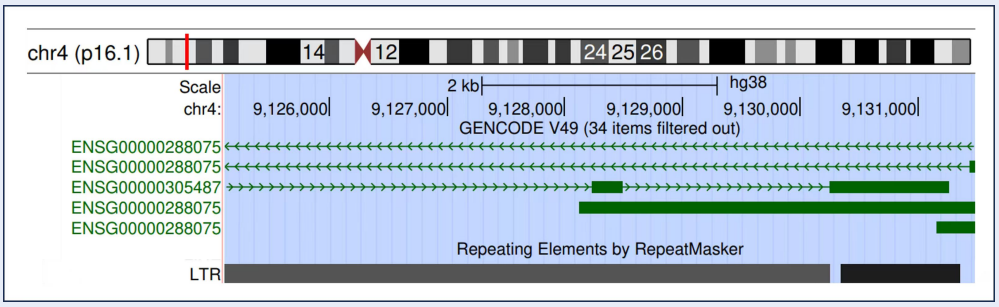

Figure 1: In silico genomic architecture and structural organization of the prioritized HERV-K (HML-2) locus at chromosome 4p16.1. Representative genomic visualization generated using the UCSC Genome Browser (human reference assembly GRCh38/hg38) showing the candidate region (chr4:9,123,514–9,133,075). The chromosomal ideogram (top) indicates the cytogenetic localization at 4p16.1 (red vertical bar). Annotation tracks display the genomic coordinates (scale: 2 kb), overlapping annotated transcripts from GENCODE v49 (green bars and directional chevron arrows indicating transcriptional orientation), and repetitive DNA elements identified by RepeatMasker (bottom). The prominent flanking long terminal repeat (LTR) elements (gray and black bars) delineate the proviral structure corresponding to the HML-2_4p16.1a insertion. *Abbreviations: kb, kilobases; LTR, long terminal repeat; UCSC, University of California, Santa Cruz.*

Table 2: Baseline demographic, hematological, and biochemical characteristics of healthy control subjects and breast cancer patients. Continuous variables were tested for normality using the Shapiro–Wilk test. Normally distributed variables are expressed as mean ± standard deviation (SD) and compared using the independent-samples Student's *t*-test (*t*). Non-normally distributed variables are presented as median (interquartile range, IQR: 25th–75th percentiles) and compared using the Mann–Whitney *U* test (*U*). All tests were two-tailed, with **P* < 0.05 indicating statistical significance.

Clinical Parameters	Healthy Controls (*n* = 45)	BC Patients (*n* = 56)	Test Statistic (*t* / *U*)	*P*-value
Age (years)	47.00 (39.50 – 62.00)	51.50 (44.00 – 59.75)	*U* = 1056	0.1632
Red blood cells ($\times 10^{12}$/L)	4.494 ± 0.2522	4.254 ± 0.4212	*t* = 3.368	0.0011*
Hemoglobin (g/L)	133.6 ± 7.405	120.9 ± 12.06	*t* = 6.161	<0.0001*
White blood cells ($\times 10^{9}$/L)	6.850 (5.970 – 7.780)	6.445 (5.550 – 7.748)	*U* = 1075	0.2078
Neutrophils (%)	56.00 (51.02 – 60.75)	59.70 (54.98 – 65.40)	*U* = 923	0.0209*
Lymphocytes (%)	33.26 ± 6.312	28.23 ± 8.498	*t* = 3.301	0.0013*
Platelets ($\times 10^{9}$/L)	235.0 (209.0 – 302.5)	265.5 (219.8 – 315.0)	*U* = 1017	0.0347*
Fasting blood glucose (mmol/L)	5.095 (4.803 – 5.495)	5.155 (4.868 – 5.963)	*U* = 552.5	0.3047

Abbreviations: BC, breast cancer; GLU, fasting blood glucose; Hb, hemoglobin; IQR, interquartile range; LYM, lymphocytes; NEU, neutrophils; PLT, platelets; RBC, red blood cells; SD, standard deviation; WBC, white blood cells.

Association Between HERV-K (HML-2)-Associated *env* Expression and Disease Stage

Stratification of BC patients by disease stage (early-stage [I–IIA], n = 22; advanced-stage [IIB–IV], n = 34; healthy controls, n = 45) demonstrated significant stage-dependent heterogeneity (Kruskal–Wallis H = 21.54, P < 0.0001; Figure 4). Dunn's *post hoc* analysis indicated that *env* expression was significantly higher in early-stage BC patients than in healthy controls (adjusted P < 0.0001) and advanced-stage patients (adjusted P = 0.0166). In contrast, advanced-stage patients did not differ significantly from healthy controls (adjusted P = 0.1481). Median expression was highest in early-stage BC (1.087 [0.8586–1.448]), intermediate in advanced-stage BC (0.7196 [0.4563–1.104]), and lowest in controls (0.5284 [0.2848–0.7876]). Conventional markers (CA15-3, CEA, and CRP) did not differ significantly between stages (all P > 0.05; Supplementary Table S4). Clinicopathological characteristics and subtype distributions are summarized in Supplementary Table S5.

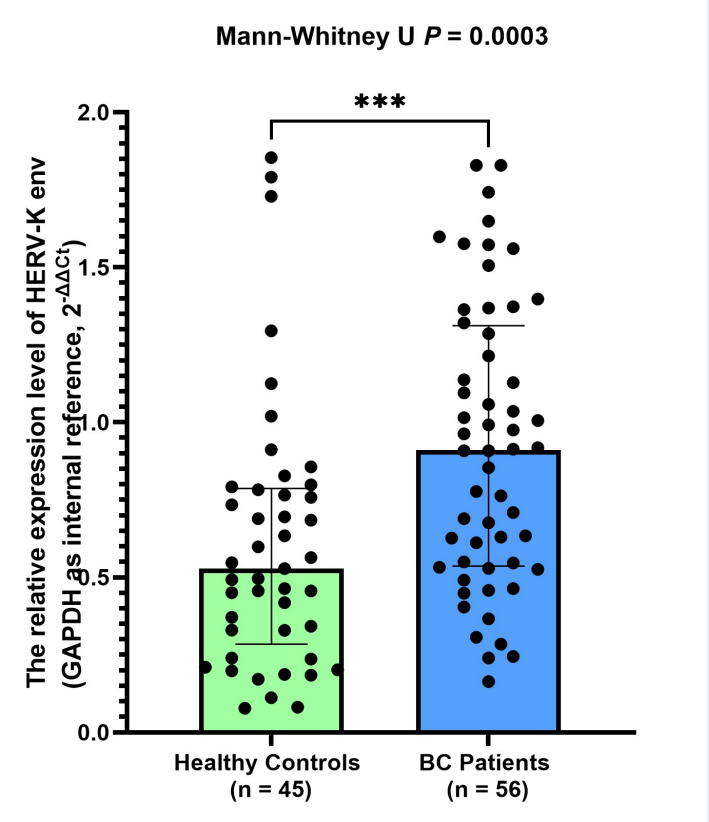

Figure 2: Peripheral blood leukocyte HERV-K (HML-2) *env* expression in breast cancer patients compared with healthy controls. Relative expression levels of family-level HERV-K (HML-2) *env* mRNA were quantified in peripheral blood leukocytes from healthy control subjects ($n = 45$, green column) and patients with histologically confirmed primary breast cancer ($n = 56$, blue column) using SYBR Green quantitative real-time PCR (qRT-PCR). Target gene expression was normalized to the endogenous reference gene *GAPDH* and calculated using the comparative $2^{-\Delta\Delta Ct}$ method relative to a pooled breast cancer calibrator sample. Column heights represent median values; error bars denote the interquartile range (IQR; 25th–75th percentiles). Individual data points represent single participants. Statistical comparison between groups was performed using the two-tailed Mann–Whitney *U* test ($U = 738.5$, $P = 0.0003$). ***$P < 0.001$. *Abbreviations: BC, breast cancer; Ct, cycle threshold; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; HERV-K, human endogenous retrovirus K; HML-2, human MMTV-like 2; qRT-PCR, quantitative reverse transcription polymerase chain reaction.*

Table 3: Univariate and multivariable binary logistic regression analyses identifying independent factors associated with breast cancer. Univariate logistic regression was performed for individual demographic, hematological, and molecular parameters. Multivariable logistic regression was executed adjusting simultaneously for relative HERV-K (HML-2) *env* expression, patient age, lymphocyte percentage, and hemoglobin concentration. Values represent unadjusted odds ratios (OR) and adjusted odds ratios (aOR) with corresponding 95% confidence intervals (CI) and two-tailed *P*-values (*$P < 0.05$ denotes statistical significance).

Variables	Univariate OR	95% CI	*P*-value	Adjusted OR	95% CI	*P*-value
HERV-K (HML-2) *env* expression	5.28	1.90 – 14.67	0.001*	12.66	3.14 – 51.02	<0.001*
Age (years)	1.02	0.98 – 1.06	0.256	1.05	1.00 – 1.10	0.044*
Lymphocytes (%)	0.92	0.86 – 0.97	0.003*	0.95	0.88 – 1.02	0.179
Hemoglobin (g/L)	0.87	0.82 – 0.92	<0.001*	0.85	0.79 – 0.91	<0.001*

Abbreviations: aOR, adjusted odds ratio; CI, confidence interval; HERV-K, human endogenous retrovirus K; HML-2, human MMTV-like 2; OR, odds ratio.

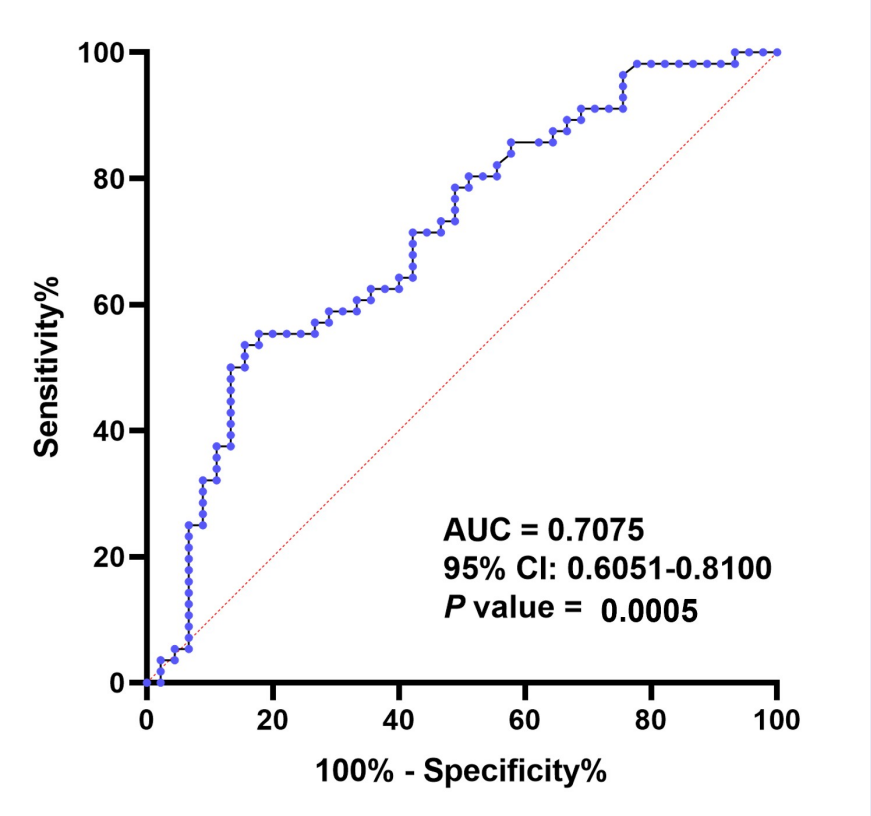

Figure 3: **Receiver operating characteristic (ROC) curve evaluating the discriminatory performance of circulating HERV-K (HML-2) *env* expression.** The ROC curve (blue solid line with circular markers) illustrates sensitivity versus 100% – specificity across various expression cut-off values for differentiating breast cancer patients (n = 56) from age-matched healthy control subjects (n = 45). The diagonal dashed red line represents the reference line of no discrimination (AUC = 0.50). The area under the receiver operating characteristic curve (AUC) is 0.7075 (95% confidence interval [CI]: 0.6051–0.8100; P = 0.0005), indicating moderate diagnostic discriminatory capacity for circulating family-level HERV-K (HML-2) *env* expression. *Abbreviations: AUC, area under the curve; BC, breast cancer; CI, confidence interval; HERV-K, human endogenous retrovirus K; HML-2, human MMTV-like 2; ROC, receiver operating characteristic.*

DISCUSSION

This study combined exploratory transcriptomic screening with clinical validation to evaluate circulating, leukocyte-derived HERV-K (HML-2) *env* expression in breast cancer. Transcriptomic analysis of cell models identified an upregulated *env* signal overlapping chromosome 4p16.1. Subsequent clinical qRT-PCR quantification confirmed significantly elevated peripheral *env* expression in BC patients relative to healthy controls (P = 0.0003), predominantly driven by early-stage disease (stages I–IIA, P < 0.0001). ROC analysis demonstrated moderate diagnostic discrimination (AUC = 0.7075), supporting the potential of circulating retroelement transcription as a complementary biomarker.

These findings align with prior reports documenting HERV-K (HML-2) transcriptional derepression in breast malignancies[31,32]. However, unlike earlier studies that evaluated HERV-K at the global family level without genomic contextualization, we annotated the candidate locus to chromosome 4p16.1 (HML-2_4p16.1a)[20]. Because short-read sequencing and conserved primer assays cannot uniquely resolve multicopy retroviral loci[25], our quantification reflects family-level *env* expression contextualized by the 4p16.1 prioritization.

Genomic annotation revealed that the candidate locus is flanked by LTR5 elements containing promoter and enhancer motifs susceptible to epigenetic derepression during oncogenic transformation[33–35]. Oncogenic activation driven by Ras or HER2 signaling triggers global chromatin remodeling, DNA hy-

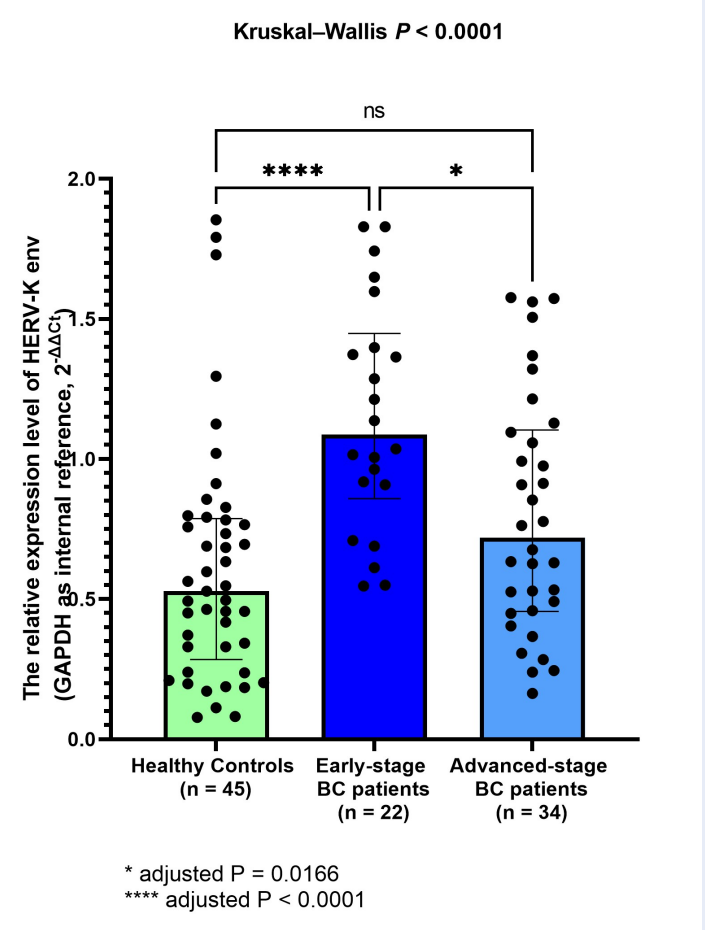

Figure 4: Stage-stratified expression of peripheral HERV-K (HML-2) *env* in breast cancer patients and healthy controls. Relative HERV-K (HML-2) *env* expression levels in peripheral blood leukocytes from healthy controls ($n = 45$, light green), early-stage breast cancer patients (AJCC Stages I–IIA, $n = 22$, dark blue), and advanced-stage breast cancer patients (AJCC Stages IIB–IV, $n = 34$, light blue). Expression values were determined by qRT-PCR using *GAPDH* as an endogenous reference and calculated via the $2^{-\Delta\Delta Ct}$ method. Column heights represent median values; error bars denote the interquartile range (IQR). Individual data points represent single subjects. Statistical significance among the three groups was evaluated by the non-parametric Kruskal–Wallis test ($H = 21.54$, $P < 0.0001$), followed by Dunn's *post hoc* multiple comparisons test. Adjusted *P*-values: healthy controls vs. early-stage BC, $P < 0.0001$ (****); early-stage BC vs. advanced-stage BC, $P = 0.0166$ (*); healthy controls vs. advanced-stage BC, $P = 0.1481$ (ns, not statistically significant). *Abbreviations: AJCC, American Joint Committee on Cancer; BC, breast cancer; Ct, cycle threshold; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; HERV-K, human endogenous retrovirus K; HML-2, human MMTV-like 2; ns, not significant; qRT-PCR, quantitative reverse transcription polymerase chain reaction; TNM, tumor, node, metastasis.*

pomethylation, and retroelement reactivation[36–38]. Thus, elevated peripheral *env* expression likely mirrors broad systemic epigenetic dysregulation rather than isolated transcription from a single proviral locus.

A central finding is the non-linear, stage-dependent expression pattern: *env* expression peaked in early-stage BC and declined in advanced disease. During early oncogenesis, acute epigenetic instability and initial anti-tumor immune activation may transiently derepress retroelements across circulating leukocytes[39,40]. In advanced disease, progressive immune exhaustion, tumor-induced immunosuppression, and altered leukocyte subpopulations may attenuate this response. Importantly, because non-malignant inflammatory disease controls were not evaluated, this signal should be viewed as an indirect systemic reflection of tumor-associated immune-epigenetic remodeling rather than a cancer-specific transcript.

Multivariable logistic regression demonstrated that *env* expression remained independently associated with BC after controlling for hematological parameters (adjusted OR = 12.66). However, the wide confidence interval (3.14–51.02) reflects estimation uncertainty inherent to exploratory sample sizes, necessitating validation in larger cohorts. Furthermore, while conventional serum markers (CA15-3, CEA, CRP) failed to differentiate early- from advanced-stage disease in our cohort, circulating *env* expression showed distinct stage sensitivity, highlighting its promise within multi-analyte liquid biopsy panels[41].

Several limitations should be noted. First, qRT-PCR primers captured family-level *env* transcripts; long-read sequencing or targeted locus capture is required for definitive locus-specific attribution. Second, single reference gene normalization (*GAPDH*) was employed; validating multiple housekeeping genes will enhance quantitative precision[28]. Third, the cross-sectional cohort was modest, though post hoc power analysis indicated adequate statistical power (86.1%) for stage comparisons (Supplementary Table S6). Finally, mechanistic investigations into causal pathways were beyond the current scope.

CONCLUSION

In conclusion, this study demonstrates that circulating leukocyte-derived HERV-K (HML-2)-associated *env* expression is significantly increased in breast cancer patients, particularly in early-stage disease. While representing family-level retroelement transcriptional activation, this circulating signal holds promise as a complementary, non-invasive biomarker reflecting tumor-associated epigenetic and immune alterations. Expanded multicenter trials and locus-resolved sequencing are warranted to validate these findings and facilitate clinical translation.

ABBREVIATIONS

AJCC: American Joint Committee on Cancer; aOR: Adjusted odds ratio; AUC: Area under the receiver operating characteristic curve; BC: Breast cancer; BLASTn: Basic Local Alignment Search Tool (nucleotide); bp: Base pairs; CA15-3: Cancer antigen 15-3; cDNA: Complementary DNA; CEA: Carcinoembryonic antigen; CI: Confidence interval; CRP: C-reactive protein; Ct: Cycle threshold; EDTA: Ethylenediaminetetraacetic acid; EMT: Epithelial-to-mesenchymal transition; ER: Estrogen receptor; ERK: Extracellular signal-regulated kinase; GAPDH: Glyceraldehyde 3-phosphate dehydrogenase; GEO: Gene Expression Omnibus; GLU: Fasting blood glucose; Hb: Hemoglobin; HER2: Human epidermal growth factor receptor 2; HERV: Human endogenous retrovirus; HERV-K: Human endogenous retrovirus K; HML-2: Human MMTV-like 2; IQR: Interquartile range; LTR: Long terminal repeat; LYM: Lymphocytes; MIQE: Minimum Information for Publication of Quantitative Real-Time PCR Experiments; NEU: Neutrophils; OR: Odds ratio; PLT: Platelets; PR: Progesterone receptor; qRT-PCR: Quantitative reverse transcription polymerase chain reaction; RBC: Red blood cells; RNA-seq: RNA sequencing; ROC: Receiver operating characteristic; SD: Standard deviation; STARD: Standards for Reporting Diagnostic Accuracy; TNBC: Triple-negative breast cancer; TNM: Tumor, Node, Metastasis; UCSC: University of California, Santa Cruz; VIF: Variance inflation factor; WBC: White blood cells.

ACKNOWLEDGMENTS

The authors thank the patients and healthy volunteers who participated in this study, as well as the clinical and laboratory staff at Hue University of Medicine and Pharmacy Hospital for their administrative and technical assistance.

AUTHOR'S CONTRIBUTIONS

All authors contributed to study conceptualization, study design, data acquisition, experimental investigation, statistical analysis, and manuscript preparation. All authors read and approved the final manuscript.

FUNDING

This study was funded by Hue University under grant number DHH2025-04-233.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Public transcriptomic RNA-seq datasets are accessible via the NCBI Gene Expression Omnibus repository under accession number GSE84275.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study protocol was approved by the Institutional Ethics Committee of Hue University of

Medicine and Pharmacy, Hue University (Approval No. H2025/426) and was conducted in accordance with the ethical standards established in the Declaration of Helsinki. All enrolled participants provided written informed consent prior to inclusion.

CONSENT FOR PUBLICATION

Not applicable. No individual personal identifiers or images are included in this manuscript.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

None. Generative AI or AI-assisted technologies were not utilized in the writing or editing of this manuscript.

COMPETING INTERESTS

The authors declare that they have no competing financial or non-financial interests.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb;409(6822):860–921. PMID: 11237011. Available from: https://doi.org/10.1038/35057062.
2. Kitsou K, Lagiou P, Magiorkinis G. Human endogenous retroviruses in cancer: oncogenesis mechanisms and clinical implications. Journal of Medical Virology. 2023 Jan;95(1):e28350. PMID: 36428242. Available from: https://doi.org/10.1002/jmv.28350.
3. Bo M, Carta A, Cipriani C, Cavassa V, Simula ER, Huyen NT, et al. HERVs endophenotype in autism spectrum disorder: human endogenous retroviruses, specific immunoreactivity, and disease association in different family members. Microorganisms. 2024 Dec;13(1):9. PMID: 39858776. Available from: https://doi.org/10.3390/microorganisms13010009.
4. Küry P, Nath A, Créange A, Dolei A, Marche P, Gold J, et al. Human endogenous retroviruses in neurological diseases. Trends in Molecular Medicine. 2018 Apr;24(4):379–394. PMID: 29551251. Available from: https://doi.org/10.1016/j.molmed.2018.02.007.
5. Nelson PN, Roden D, Nevill A, Freimanis GL, Trela M, Ejtehadi HD, et al. Rheumatoid arthritis is associated with IgG antibodies to human endogenous retrovirus gag matrix: a potential pathogenic mechanism of disease? Journal of Rheumatology. 2014 Oct;41(10):1952–1960. PMID: 25225282. Available from: https://doi.org/10.3899/jrheum.130502.
6. Russ E, Mikhalkevich N, Iordanskiy S. Expression of human endogenous retrovirus group K (HERV-K) HML-2 correlates with immune activation of macrophages and type I interferon response. Microbiology Spectrum. 2023 Apr;11(2):e0443822. PMID: 36861980. Available from: https://doi.org/10.1128/spectrum.04438-22.
7. Wu Y, Huang S, Sha Q, Yu J. Emerging and Re-emerging viruses as triggers of human endogenous retrovirus activation: implications for aging and age-related pathologies. Molecular Aspects of Medicine. 2025 Dec;106:101422. PMID: 41207217. Available from: https://doi.org/10.1016/j.mam.2025.101422.
8. Scognamiglio S, Grandi N, Pessiu E, Tramontano E. Identification, comprehensive characterization, and comparative genomics of the HERV-K(HML8) integrations in the human genome. Virus Research. 2023 Jan;323:198976. PMID: 36309315. Available from: https://doi.org/10.1016/j.virusres.2022.198976.
9. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology. 2011 Nov;8(1):90. PMID: 22067224. Available from: https://doi.org/10.1186/1742-4690-8-90.
10. Lemaître C, Tsang J, Bireau C, Heidmann T, Dewannieux M. A human endogenous retrovirus-derived gene that can contribute to oncogenesis by activating the ERK pathway and inducing migration and invasion. PLoS Pathogens. 2017 Jun;13(6):e1006451. PMID: 28651004. Available from: https://doi.org/10.1371/journal.ppat.1006451.
11. Lin Y, Satta RR, Simula ER, Tang S, Molicotti P, Cossu A, et al. The role of human endogenous retroviruses in the initiation and progression of melanoma. Biomedicines. 2025 Jul;13(7):1662. PMID: 40722734. Available from: https://doi.org/10.3390/biomedicines13071662.
12. Zhou F, Li M, Wei Y, Lin K, Lu Y, Shen J, et al. Activation of HERV-K Env protein is essential for tumorigenesis and metastasis of breast cancer cells. Oncotarget. 2016 Dec;7(51):84093–84117. PMID: 27557521. Available from: https://doi.org/10.18632/oncotarget.11455.
13. Shek J, Sun C, Wilson EM, Moadab F, Hastie KM, Rajamanickam RR, et al. Human endogenous retrovirus K (HERV-K) envelope structures in pre- and postfusion by cryo-EM. Science Advances. 2025 Aug;11(35):eady8168. PMID: 40864726. Available from: https://doi.org/10.1126/sciadv.ady8168.
14. Stricker E, Peckham-Gregory EC, Scheurer ME. HERVs and cancer-a comprehensive review of the relationship of human endogenous retroviruses and human cancers. Biomedicines. 2023 Mar;11(3):936. PMID: 36979914. Available from: https://doi.org/10.3390/biomedicines11030936.
15. Záveský L, Jandáková E, Weinberger V, Minář L, Kohoutová M, Slanař O. Human endogenous retroviruses in breast cancer: altered expression pattern implicates divergent roles in carcinogenesis. Oncology. 2024;102(10):858–867. PMID: 38408442. Available from: https://doi.org/10.1159/000538021.
16. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians. 2024;74(3):229–263. PMID: 38572751. Available from: https://doi.org/10.3322/caac.21834.
17. Loibl S, Poortmans P, Morrow M, Denkert C, Curigliano G. Breast cancer. Lancet. 2021 May;397(10286):1750–1769. PMID: 33812473. Available from: https://doi.org/10.1016/S0140-6736(20)32381-3.
18. Liu S, et al. Long-read single-cell sequencing reveals expressions of hypermutation clusters of isoforms in human liver cancer cells. bioRxiv. 2023 Aug;Available from: https://doi.org/10.1101/2023.03.16.532991.
19. Kang Q, Guo X, Li T, Yang C, Han J, Jia L, et al. Identification of differentially expressed HERV-K(HML-2) loci in colorectal cancer. Frontiers in Microbiology. 2023 Jun;14:1192900. PMID: 37342563. Available from: https://doi.org/10.3389/fmicb.2023.1192900.
20. Montesion M, Williams ZH, Subramanian RP, Kuperwasser C, Coffin JM. Promoter expression of HERV-K (HML-2) provirus-derived sequences is related to LTR sequence variation and polymorphic transcription factor binding sites. Retrovirology. 2018 Aug;15(1):57. PMID: 30126415. Available from: https://doi.org/10.1186/s12977-018-0441-2.
21. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research. 2013 Jan;41(Database issue):D991–D995. PMID: 23193258. Available from: https://doi.org/10.1093/nar/gks1193.
22. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome

Research. 2002 Jun;12(6):996–1006. PMID: 12045153. Available from: https://doi.org/10.1101/gr.229102.

23. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Research. 2017 May;27(5):849–864. PMID: 28396521. Available from: https://doi.org/10.1101/gr.213611.116.
24. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. GENCODE 2021. Nucleic Acids Research. 2021 Jan;49(D1):D916–D923. PMID: 33270111. Available from: https://doi.org/10.1093/nar/gkaa1087.
25. Li X, Lu K, Chen X, Tu K, Xie D. capTEs enables locus-specific dissection of transcriptional outputs from reference and non-reference transposable elements. Communications Biology. 2023 Sep;6(1):974. PMID: 37741908. Available from: https://doi.org/10.1038/s42003-023-05349-1.
26. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: A Cancer Journal for Clinicians. 2017 Mar;67(2):93–99. PMID: 28094848. Available from: https://doi.org/10.3322/caac.21388.
27. Cardoso F, Senkus E, Costa A, Papadopoulos E, Aapro M, André F, et al. 4th ESO-ESMO international consensus guidelines for advanced breast cancer (ABC 4). Annals of Oncology. 2018 Aug;29(8):1634–1657. PMID: 30032243. Available from: https://doi.org/10.1093/annonc/mdy192.
28. Dheda K, Huggett JF, Bustin SA, Johnson MA, Rook G, Zumla A. Validation of housekeeping genes for normalizing RNA expression in real-time PCR. Biotechniques. 2004 Jul;37(1):112–114. PMID: 15283208. Available from: https://doi.org/10.2144/04371RR03.
29. Jasemi S, Simula ER, Pantaleo A, Sechi LA. Transcriptional upregulation of HERV-env genes under simulated microgravity. Viruses. 2025 Feb;17(3):306. PMID: 40143237. Available from: https://doi.org/10.3390/v17030306.
30. Fuentes DR, Swigut T, Wysocka J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. eLife. 2018 Aug;7:e35989. PMID: 30070637. Available from: https://doi.org/10.7554/eLife.35989.
31. Gao Y, Yu XF, Chen T. Human endogenous retroviruses in cancer: Expression, regulation and function. Oncology Letters. 2021 Feb;21(2):121. PMID: 33552242. Available from: https://doi.org/10.3892/ol.2020.12382.
32. Wang-Johanning F, Frost AR, Johanning GL, Khazaeli MB, LoBuglio AF, Shaw DR, et al. Expression of human endogenous retrovirus k envelope transcripts in human breast cancer. Clinical Cancer Research. 2001 Jun;7(6):1553–1560. PMID: 11410490.
33. Garcia-Montojo M, Doucet-O'Hare T, Henderson L, Nath A. Human endogenous retrovirus-K (HML-2): a comprehensive review. Critical Reviews in Microbiology. 2018 Nov;44(6):715–738. PMID: 30318978. Available from: https://doi.org/10.1080/1040841X.2018.1501345.
34. Suntsova M, Garazha A, Ivanova A, Kaminsky D, Zhavoronkov A, Buzdin A. Molecular functions of human endogenous retroviruses in health and disease. Cellular and Molecular Life Sciences. 2015 Oct;72(19):3653–3675. PMID: 26082181. Available from: https://doi.org/10.1007/s00018-015-1947-6.
35. Rivas SR, Valdez MJ, Govindarajan V, Seetharam D, Doucet-O'Hare TT, Heiss JD, et al. The role of HERV-K in cancer stemness. Viruses. 2022 Sep;14(9):2019. PMID: 36146825. Available from: https://doi.org/10.3390/v14092019.
36. Chiappinelli KB, Strissel PL, Desrichard A, Li H, Henke C, Akman B, et al. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. Cell. 2015 Aug;162(5):974–986.
37. Ryspayeva D, Seyhan AA, MacDonald WJ, Purcell C, Roady TJ, Ghandali M, et al. Signaling pathway dysregulation in breast cancer. Oncotarget. 2025 Mar;16(1):168–201. PMID: 40080721. Available from: https://doi.org/10.18632/oncotarget.28701.
38. Katoh H, Honda T. Roles of human endogenous retroviruses and endogenous virus-like elements in cancer development and innate immunity. Biomolecules. 2023 Nov;13(12):1706. PMID: 38136578. Available from: https://doi.org/10.3390/biom13121706.
39. Liu M, Jia L, Li H, Liu Y, Han J, Wang X, et al. p53 binding sites in long terminal repeat 5Hs (LTR5Hs) of human endogenous retrovirus K family (HML-2 subgroup) play important roles in the regulation of LTR5Hs transcriptional activity. Microbiology Spectrum. 2022 Aug;10(4):e0048522. PMID: 35867400. Available from: https://doi.org/10.1128/spectrum.00485-22.
40. Salavatiha Z, Soleimani-Jelodar R, Jalilvand S. The role of endogenous retroviruses-K in human cancer. Reviews in Medical Virology. 2020 Nov;30(6):1–13. PMID: 32734655. Available from: https://doi.org/10.1002/rmv.2142.
41. Wang M, Xie F, Lin J, Zhao Y, Zhang Q, Liao Z, et al. Diagnostic and prognostic value of circulating circRNAs in cancer. Frontiers in Medicine (Lausanne). 2021 Mar;8:649383. PMID: 33816529. Available from: https://doi.org/10.3389/fmed.2021.649383.