Deciphering human endogenous retrovirus expression in colorectal cancers: exploratory analysis regarding prognostic value in liver metastases

Julien Viot,[a,b,*] Romain Loyon,[b] Nawfel Adib,[b] Pierre Laurent-Puig,[c,d] Aurélien de Reyniès,[d] Fabrice André,[e,f] Franck Monnien,[a,b] Thierry André,[g] Magali Svrcek,[h] Anthony Turpin,[i,n] Zohair Selmani,[a,b] Laurent Arnould,[j,k] Laura Guyard,[j,k] Nicolas Gilbert,[l] Anthony Boureux,[l] Olivier Adotevi,[a,b] Angélique Vienot,[a,b] Syrine Abdeljaoued,[b] Dewi Vernerey,[a] Christophe Borg,[a,b] and Daniel Gautheret[m]

[a]Département d'Oncologie Médicale, CHU Besançon, Besançon 25000, France
[b]Université Marie et Louis Pasteur, INSERM, Etablissement Français du Sang Bourgogne Franche-Comté, UMR1098, Interactions Hôte-Greffon-Tumeur/Ingénierie Cellulaire et Génique, Besançon, France
[c]Department of Biology, Institut du Cancer Paris CARPEM, APHP, APHP.Centre-Université Paris Cité, Hôpital Européen G. Pompidou, Paris, France
[d]Centre de Recherche des Cordeliers, Sorbonne Université, INSERM, Université de Paris, EPIGENETEC, Paris 75006, France
[e]Paris-Saclay University, Gustave Roussy, Villejuif, France
[f]Department of Medical Oncology, Gustave Roussy, Villejuif, France
[g]Department of Medical Oncology, Sorbonne University, Saint-Antoine Hospital, AP-HP, Paris, France
[h]Department of Pathology, Saint-Antoine Hospital, AP-HP, Sorbonne Université, Paris, France
[i]Department of Oncology, Lille University Hospital, France
[j]Department of Tumour Biology and Pathology, Georges François Leclerc Cancer Center - UNICANCER, Dijon, France
[k]CCRB Ferdinand Cabanne de Dijon, France
[l]IRMB, INSERM U1183, Hopital Saint-Eloi, Universite de Montpellier, Montpellier, France
[m]Institute for Integrative Biology of the Cell (I2BC), Université Paris-Saclay, CNRS, CEA, Gif-sur-Yvette 91190, France
[n]CNRS UMR9020, INSERM UMR1277, University of Lille, Institut Pasteur, Lille, France

eBioMedicine
2025;116: 105727
Published Online xxx
https://doi.org/10.1016/j.ebiom.2025.105727

Summary

Background Human Endogenous RetroVirus (HERV) expression in tumours reflects epigenetic dysregulation of cancer and an oncogenic factor through promoter/enhancer action on genes. While more than 50% of colorectal cancers develop liver metastases, HERV has not been studied in this context.

Methods We collected 400 RNA-seq samples from over 200 patients with primary and liver metastases, including public data and a novel set of 200 samples.

Findings We observed global stability of HERV expression between liver metastases and primary colorectal cancers, suggesting an early oncogenic footprint. We identified a list of 17 HERV loci for liver metastatic colorectal cancer (lmCRC) characterization; with tumour-specificity validated in single-cell metastatic colorectal cancer data and normal tissue bulk RNA-seq. Eleven loci produced HERV-derived peptides as per tandem mass spectrometry from primary colorectal cancer. Six loci were associated with the risk of relapse after lmCRC surgery. Four, HERVH_Xp22.32a, HERVH_20p11.23b, HERVH_13q33.3, HERVH_13q31.3, had adverse prognostic value (log-rank p-value 0.028, 0.0083, 9e-4, 0.05, respectively) while two, HERVH_Xp22.2c (log-rank p-value 0.032) and HERVH_8q21.3b (in multivariable models) were associated with better prognosis. Moreover, the markers showed a cumulative effect on survival when expressed. Some were associated with decreased cytotoxic immune cells and most of them correlated with cell cycle pathways.

Interpretation These findings provide insights into the lmCRC transcriptome landscape by suggesting prognostic markers and potential therapeutic targets.

Funding This work was supported by funding from institutional grants from Inserm, EFS, University of Bourgogne Franche-Comté, national found "Agence Nationale de la Recherche – ANR-JCJC: Projet HERIC and ANR-22-CE45-0007", and "La ligue contre le cancer".

*Corresponding author. CHU Besançon, Département d'Oncologie Médicale, Besançon 25000, France.
E-mail address: jviot@chu-besancon.fr (J. Viot).

Keywords: Transposable elements; Endogenous retrovirus; Colorectal cancer; Liver metastases; Immunology

Research in context

Evidence before this study

Expression of transposable elements, particularly Human Endogenous Retroviruses (HERVs), has been shown to differentiate between various tissue types. Recent findings have revealed distinct expression patterns of specific HERV loci that differentiate primary colon cancer from adjacent normal tissues. There is also evidence linking HERV expression to immune activation.

Added value of this study

By analysing RNA sequencing (RNA-seq) data from over 400 samples, including normal tissue, primary colorectal cancer, and liver metastases, we identified specific transposable elements and HERVs that characterize liver metastases. HERV expression remained globally consistent across both primary tumours and metastases. Additionally, we identified prognostic markers associated with surgical outcomes in liver metastases from colorectal cancer, which correlated with variations in the immune microenvironment. Notably, the HERVH_8q21.3b locus showed a strong correlation with CALB1 expression.

Implications of all the available evidence

We present a RNA-seq dataset comprising paired samples of primary colorectal cancer and liver metastases. The expression of HERV loci appears to be a significant biological factor in liver metastatic colorectal cancer with potential prognostic implications. Although the interaction between HERV expression and the immune environment is suggested, further in-depth investigation is required.

Introduction

Whereas the transcription of coding genes has been extensively studied in cancer, the expression of repeated/transposable sequences remains a work in progress.[1,2] Metastatic colorectal cancer (mCRC) is a common and fatal disease, although surgical resection of liver metastases combined with chemotherapy is associated with a 5-year survival of 30–60%.[3,4] Liver is the first metastatic site representing 50–80% of colorectal cancer (CRC) metastases.[5,6] Characterization of CRC genomic alterations has led to the clinical use of immune checkpoint inhibitors or RAS/RAF targeting agents in this cancer.[1,2,7] On the other hand, the progress made in characterizing the transcriptome of CRC has yet to be fully exploited. Moreover, the molecular signatures identified from gene expression analysis in primary tumours do not overlap with those of liver metastases.[2] Given this situation, it is necessary to identify transcriptomic biomarkers that could be used to stratify patients with metastatic CRC.

Thus, we used an original approach based on transposable elements (TE) analysis for biomarkers identification of liver metastatic colorectal cancers (lmCRC). Indeed, a large fraction of the non-coding genome is composed by TE, which include retrotransposons divided in three main classes: short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs) and long terminal repeats (LTRs) of which Human Endogenous Retrovirus (HERV) are a part. HERV represents 8% of our genome.[8] By contrast with the previous beliefs that HERV are silent, it is now acknowledged that HERV may be expressed in a cell type-dependent manner in normal tissues.[9,10] The expression of HERV might be modulated by cancer-induced genomic dysregulation. The relationship between HERV expression and oncogenesis is still debated and may be due to epigenetic dysregulation[8,10,11] or acts through their promoter/enhancer activity on oncogenes.[9,12,13]

There are several mechanisms by which HERVs can influence the protein production process: a HERV acting as a long non-coding RNA (lncRNA) capable of modulating protein production from a gene, the direct production of proteins derived from internal HERV sequences (Gag, Pol, Env), and onco-exaptation between a gene and a HERV.[14,15] HERVH elements can serve as promoters for gene expression, for instance, by acting as binding sites for transcription factors in epiblasts, but not in normal tissues.[16,17] While HERVH represents a rare case of endogenous retroviral co-option, the HERVH and its associated long terminal repeat, LTR7, have been implicated in the regulation of human pluripotency and stem-cell self-renewal.[17–19]

Recent investigations reported an expression of HERV at the locus level in primary colorectal cancers.[20] The MER4, HERVL/ERVL, HERVH and HERVK families were the most abundant among re-expressed HERV.[10,21] HERV expression alone could distinguish tumour and normal samples. However, the global HERV expression in primary CRC was lower compared with other cancer types.[10,20] TE re-expression is mainly associated with proximal DNA demethylation and correlates with DNA damage and immune response.[10] However, little is known about the prognostic impact

of HERV expression in mCRC and regulation of HERV is currently not considered as a relevant determinant of mCRC biology.

A possible implication of HERV in oncogenesis might be their role in modulating the microenvironment and the immune contexture, a prognostic value in CRC.[22] Indeed, it has been proposed that HERV re-expression in cancer cells generates a "viral mimicry"[12,23,24] and that HERV-derived peptides might be a possible source of cancer-associated antigens.[25] In addition, several studies pointed out the association of interferon-gamma response to HERV re-expression.[12,26] Indeed, the global HERV load has been associated with enhanced immune cell infiltration in several cancers, including colorectal carcinoma.[27] Moreover, the role of HERV as a modulator of cancer-related immune responses was supported by the correlation between HERV levels and increased survival in patients with urothelial and kidney cancer treated by immune checkpoint inhibitors.[28,29]

Overall, the status of HERV expression in lmCRC and whether it can affect cancer cells, and their immune environment is currently unknown. Thus, our study aims to clarify the level of HERV expression in lmCRC and to determine the possible biological and clinical correlates of such HERV transcription. For this purpose, we generated a new RNA-seq dataset of paired primary and liver metastatic biopsies from 100 patients with lmCRC (BIOMIROX). We performed an integrated analysis of TE and HERV expression in 400 normal, primary, and lmCRC samples. We analysed the stability of HERV expression signatures across primary and metastatic samples. Finally, we identified the prognostic value of HERV expression and analysed their association with clinical variables, immune cell infiltration, and regulatory pathways.

Methods

Tumour samples

Tumour samples from the UHB cohort and part of BIOMIROX were excised by surgeons between 2003 and 2019 from the Department of Digestive Surgery at the University Hospital of Besançon and European Hospital Georges Pompidou, France. The Centre de Ressources Biologiques Ferdinand Cabanne at Dijon University Hospital provided tissue samples from patients with mCRC used to perform RT-qPCR. Primary tumour and liver metastases with paired normal tissue samples were supplied as frozen tissue pieces.

For other tissue selection and preparation, details on library preparation and sequencing please refer to the original article: SRP029880 (GSE50760),[30] SRP060016,[31] SRP095672 (GSE92914),[32] SRP245232 (GSE144259),[33] TCGA COAD,[1] UHB_cohort (GSE207194),[34] METAPRISM (EGAD00001009684).[35]

Biological metadata and mapping information can be found in Supplementary Tables S1 and S2.

Human patients

All patients with metastatic or non-metastatic colorectal cancer who had sufficient biological material for RNA sequencing were eligible. The patients' characteristics are summarised in Supplementary Table S3, including the number of missing values for each collected variable. Sex/gender was not included in the clinical data collection, as the aggregate of anonymized databases contained too many missing data points for the variable to remain relevant. No distinction based on the patient's gender led to exclusion from the study.

Ethics

This is a post hoc study of patients meeting the inclusion criteria of having their primary tumour and metastases preserved in the Biological Resource Center - Tumour Bank Division of the Besançon University Hospital (registration number BB-0033-00024). The project was approved by the scientific board of the biobank (#F2012 CRC HERV). All clinical data associated with patients whose tumour samples were selected were part of the MIROX (NCT00268398), EPITOPES-CRC01 (NCT02838381), and EPITOPES-CRC02 (NCT02817178) trials. These trials were approved by the ethics committees of Lille and Grand-Est II. All patients were enrolled after the signature of informed consent. Clinical variables were retrieved manually from electronic records according to the RGPD legislation.

Annotations

TE and HERV annotations were referenced to RepeatMasker, an annotation database based on RepBase and Dfam.org. Prebuilt annotations from Telescope (https://github.com/mlbendall/telescope_annotation_db/tree/master/builds/HERV_rmsk.hg38.v2) and REdiscoverTE were used to annotate read alignments. TE annotation from REdiscoverTE covers 20 classes, 58 families, and 1497 subfamilies. HERV annotations from Telescope cover 3 families, 229 subfamilies, and 14,968 loci. Bowtie2 v2.3.5 was used with the prebuild GRch38 index proposed on their website. REdiscoverTE gene annotation was based on GENCODE v26. Genomic annotation was performed with R package annotatr v1.20.0 for Hg38 and Telescope hg38 annotation GTF.

Bulk RNA-seq

Globally, all samples were poly-A paired-end RNA-seq between 75 and 150 bases on Illumina sequencer. Tumour RNA was extracted from macro dissected formalin-fixed, paraffin-embedded (FFPE) blocks from primary tumours using the Maxwell RSC RNA FFPE Kit (Promega, France). PolyA-RNA sequencing (RNA-seq) library preparation protocols were performed using 400 ng of template RNA and the QuantSeq 3′mRNA-Seq Kit FWD for Illumina (Lexogen) according to the manufacturer's instructions. Libraries were sequenced on NovaSeq6000 (Illumina).

TE and gene expression were quantified (Supplementary Fig. S1) by REdiscoverTE[10] using the adapted version of the original dataset available at https://github.com/ucsffrancislab/ REdiscoverTE and Salmon v1.6.0[36](Supplementary Tables S4 and S5). HERV locus expression was counted by Telescope[37] v1.0.3 after alignment with bowtie2 v2.3.5.1 in very sensitive mode (optional parameters: -k 100 -p 16 –very-sensitive-local –score-min L, 0,1.6) (Supplementary Tables S6 and S7). Note that this method assigns multi-mapping reads to individual TE families using an expectation maximization algorithm.

Batch effect reduction was performed using Combat-seq (R package sva v3.42.0) on raw counts, adjusted for study and sample types. Multidimensional reduction was used to confirm Batch reduction (Supplementary Figs. S2 and S3). Count normalization for genes and HERV was performed as counts per million (CPM) based on the library size of the genes as previously proposed by Kong et al. using REdiscoverTE.

3' RNA-seq counts need a correction factor of three, probably due to sequencing depth. The quality of the correlation between HERV mean count on paired-end and single-end after correction for TE and HERV loci was 0.8 $p < 2.2e\text{-}16$ and 0.94 $p < 2.2e\text{-}16$, respectively (Supplementary Fig. S4 and S5 and Table S8).

GTEx

We quantified ERVs in 1137 RNA-seq normal tissue samples from GTEx[38] (Supplementary Table S9) using a fast reference-free query software, Reindeer.[39] We selected 541 HERV loci that were differentially expressed between normal and tumoral colorectal tissues (Supplementary Table S10). The DNA sequences of the selected HERVH loci were extracted from the UCSC genome browser based on genomic positions. Tables were normalized to reflect RNA-seq CPM counts.

Quality control was performed using the SRP029880 dataset of 54 colon samples, using Reindeer and either Kallisto[40] for genes or Telescope for HERV. This showed good correlation and comparable value ranges (CC:0.97 p-value < 2.2e-16 and CC:0.94 p-value 2.2e-16, respectively) (Supplementary Fig. S6).

Single-cell RNA-seq

Single-cell RNA-seq atlas from six liver metastatic colorectal cancer (GSE178318)[41] was retrieved as FASTQ files and gene count matrix. ERV counting was performed with STAR-solo v2.7.10a and the following command: STAR –genomeDir refs/STAR_sc –readFilesIn ${base}_2.fastq.gz ${base}_1.fastq.gz –readFilesCommand zcat –soloType CB_UMI_Simple –soloCBwhitelist refs/737K-august-2016.txt –soloBarcodeReadLength 0 –outFileNamePrefix output_HERVc/${sample}/ –soloFeatures Gene GeneFull Velocyto –soloMultiMappers EM –outFilterScoreMin 30 –soloCBmatchWLtype 1 MM_multi_Nbase_pseudocounts –soloUMIfiltering MultiGeneUMI_CR –soloUMIdedup 1 MM_CR –clipAdapterType CellRanger4 –sjdbGTFfile refs/telescope_rmsk_hg38_locus.gtf –outFilterMultimapNmax 200 –winAnchorMultimapNmax 200 –soloCellFilter EmptyDrops_CR –limitOutSJcollapsed 10000000 –outSAMtype BAM Unsorted. Please note that this version of STAR-solo takes care of multimapping from the maximum expectation algorithm on reads.

Analysis was conducted with Seurat v4.2.0 according to manual references. In brief, the raw gene counts matrix and raw ERV matrix were concatenated in a new Seurat object based on cell ID. 44579 cells from six samples were used. Quality checks are supplementary (Supplementary Fig. S7). Then we normalized and clustered on genes. We annotated manually the clusters based on highly expressed genes from the original publication. Then plot HERV expression on UMAP-represented cells. HERV mean expression by sample type and cell cluster can be found in Supplementary Tables S11 and S12.

To identify cancer-specific HERV we performed for each locus two by two rank-sum Wilcoxon test with cancer cells and other cell subtypes. Only significant values with p-value ≤ 0.05 were retained (Supplementary Table S13). Loci with at least one cell subtype with more expression than cancer cells were discarded for the rest of the Bulk RNA-seq analysis.

MS/MS spectrometry

We performed HERV-derived peptides using PePQuery.[42] The previously described annotation from Telescope based on RepeatMasker was enriched for DNA sequences retrieved by positions using the UCSC API. More than 50,000 sequences were processed locally with a custom Python script using pepquery v2.0.2; TCGA colon proteome and gencode human protein database as reference: "os.system ("java -jar pepquery-2.0.2.jar -b CPTAC_TCGA_Colon_Cancer_Proteome_PDC000111 -db gencode:human -hc -o pepquery_python_db/{}/ -i {} -t DNA -cpu 12".format (row.index_HERV, row.sequence))".

The TCGA colon[1] proteome is composed of 90 patients with 95 biospecimens, 64 colon adenocarcinoma samples, and 31 rectal adenocarcinomas, representing 1425 MS/MS spectra for liquid chromatography-tandem mass spectrometry (LC-MS/MS) global proteomic profiling.

For our analysis, we used the peptides considered as confident by PePQuery.

Elastic Net predictions

Among the five independent datasets, the Elastic Net predictor was trained on the SRP029880 dataset, excluding normal liver samples, as this dataset included samples from the same patients, thereby minimizing interindividual variability. Calculations were performed using the R package glmnet v4.1.8. The input matrix

consisted of raw counts for HERV loci obtained from Telescope. On this training dataset, we employed k-fold cross-validation (k = 10) to identify the optimal model by minimizing the misclassification error. The best-performing model from this process was then tested on the remaining four independent datasets to ensure robust external validation, with all normal liver samples excluded.

Differential expression analysis

Differential expression analysis (DE) was performed using DEseq2 v1.34.0 and edgeR v3.36.0 according to the reference manuals. Briefly, we used paired-end RNA-seq samples, then we excluded normal liver samples to focus on colorectal cell lineage and filtered low-represented locus by conserving locus with at least five reads in 5% of samples to speed up the calculation. Designed on studies plus sample type or class. To select a subset of highly expressed, strongly differential elements, we subsampled results on adjusted p-value (Benjamini & Hochberg) < 0.001, log2fold-change <−1 or >1, and base mean > 10. DE has been performed on full cohort and independent datasets if possible. Results from DEseq2 and edgeR were merged.

Pathways enrichment

Spearman's correlation coefficient between protein-coding genes (annotation retrieved from Ensembl) and HERV has been performed on a CPM-normalized concatenated matrix. We retrieved human HALLMARK signatures from Msigdb (https://www.gsea-msigdb.org/gsea/msigdb) with R package v1.2.0. Gene enrichment analysis was performed with fsgea R package v1.20.0, with ranking defined as Spearman's correlation score between HERV locus expression and gene expressions for the global dataset and hallmark collection. If the adjusted p-value (Benjamini & Hochberg) was <0.05, the pathway was associated with HERV locus.

Survival analysis

Clinical data for overall and progression-free survival after liver metastasis resection were available for 137 patients from the UHB and BIOMIROX cohorts, with follow-up exceeding five years.

For the survival analyses, progression-free survival (PFS) was defined as the time from hepatic metastasis resection (origin/start) to either disease progression, as determined by RECIST v1 criteria, or death, whichever occurred first. Patients who were alive without progression were censored at the time of their last radiological evaluation showing no progression. Overall survival (OS) was defined as the time from hepatic metastasis resection to death from any cause. Patients who were alive were censored at the date of their last follow-up.

HERV mRNA was arbitrarily considered expressed if CPM was above 1 for all HERV candidates, except for HERVH_13q31.3 (threshold set at 0.5 CPM). Prognostic value was assessed using log-rank, lasso Cox, stepwise Cox, and multivariable Cox models.

We assessed the assumptions underlying Cox regression. To minimise prediction error, the tuning (penalty) parameter in the LASSO regression was selected through k-fold cross-validation, with the C-index as the criterion. The optimal lambda was determined based on the highest C-index. The stepwise Cox regression was bidirectional, with selection based on AIC. All other parameters were set to defaults, notably with 'alpha = 0.15' to include a greater number of potentially relevant variables in the 'StepwiseCox' function from the StepReg R package.

There was no missing data for the genomic expression of HERV loci, and there was no missing data for age. However, the side of the tumour was unknown for 20 patients, accounting for 14.5% of the cohort. Additionally, 73 patients had missing data for RAS mutation status, representing 52.9% of the total. Finally, there were one (0.7%) patient who had missing age at surgery and two missing data points for chemotherapy status, which corresponds to 1.4% of the patients.

The multivariable Cox regression model was built using HERV loci identified through Cox Lasso regression, along with clinically relevant variables based on existing literature and clinical expertise. These variables included age, tumour localization within the colon, chemotherapy treatment prior to metastasectomy, and RAS mutation status. Microsatellite instability and RAF mutations were excluded due to insufficient case numbers, as most samples were MSS and BRAF wild-type.

Due to missing data, particularly for RAS status, we tested models both with and without RAS status, using multiple imputation of missing data via the mice R package with the 'pmm' (predictive mean matching) method and 1000 imputations. There was no significant difference between the models excluding RAS or those utilizing multiple imputation. However, we chose to retain RAS status in the model given its established clinical relevance.

Immune deconvolution

Immune deconvolution was calculated using the immunedeconv R package v2.1 with mcp-counter[43] methods according to the manual. Subpopulation scores were drawn according to HERVH locus sample expression using the same method as for survival analysis, arbitrarily considered expressed if HERVH locus >1 CPM.

Nucleic acids extraction

Total RNA and genomic DNA nucleic acids from normal and tumour tissue samples from patients with

mCRC were extracted using the Qiagen AllPrep DNA/RNA Mini kit. Tissues were first lysed in Lysing Matrix H tubes (MP Biomedicals) using either FastPrep-24 5G (MP Biomedicals, 30 s cycle) or TissueLyser II (Qiagen, 2 cycles of 2 min lysis at 20 Hz followed by 2 cycles of 2 min lysis at 25 Hz). Supplier protocols were optimized with Turbo DNAse treatment (Invitrogen). The RW1 wash step was modified. After an initial wash with 350 µL of RW1 buffer, column membranes were incubated for 15 min at room temperature with 50 µL of a mix containing Turbo DNAse (Invitrogen, 4 IU), 10 × Turbo DNAse Buffer (Invitrogen) and RNase free water, before RNA was re-precipitated in RLT + Dithiothreitol buffer (Sigma, 40 mM) and 80% ethanol. The protocol was repeated as described by the supplier after a second wash with 350 µL of RW1 buffer. Nucleic acid concentrations were determined using NanoDrop. For long-term storage, RNA was stored at –80 °C or –150 °C and genomic DNA at –20 °C.

Quantitative real-time polymerase chain reaction (q-RT PCR)

q-RT PCR were performed to assess HERV expression level on 59 mCRC patients' tissues (Tumoral and Healthy adjacent tissues). Primer sequences can be found in Supplementary Table S14. Total RNA isolated using Qiagen AllPrep DNA/RNA Mini kit, was reverse transcript with PrimeSCript RT Reagent Kit with gDNA Eraser (Perfect Real Time) (Takara Bio). q-RT PCR was performed using gene-specific SybrGreen probes and TB Green Premix (Takara Bio), following the manufacturer's instructions. Gene expression was normalized to 18s mRNA. Samples were realized in duplicate. Relative expression for the mRNA transcripts was calculated using the 2–ΔΔCt method. Figures & Statistics.

Statistics

The primary objective was to identify differential HERV expression between primary CRC and liver metastases. No formal statistical sample size was determined for this study due to its exploratory nature and the lack of available data in the literature regarding the primary objective.

Statistical comparisons and figures were made using R version 4.1.2. Most of the figures were done with ggpubr packages v0.6.0 and ggplot v3.4.4. Kaplan–Meier curves were plotted using ggpubr packages v0.4.0 and survival v3.2. LogRank p-value was provided on graphs. Heatmap was performed by pheatmap v1.0.12 and ComplexHeatmap v2.10.0. Circular plot with Circlize v0.4.15 and fmsb v0.7.5 R package. VolcanoPlot was drawn by EnhancedVolcano v1.12.0. Multidimensional reduction plot was performed with t-SNE and UMAP through M3C R package v1.16.0.

Distribution between two groups was performed with non-parametric Wilcoxon test. Fisher's exact test for count data was used for clinical variable testing with patient expressing or not HERVH locus. Age was tested with Kruskal–Wallis rank sum test. p-values were considered significant when <0.05 or <0.001 and detailed in subsequent analysis. The 95% confidence intervals are provided to characterize the data distribution.

Role of funders

This work was supported by funding from institutional grants from INSERM, EFS, University of Bourgogne Franche-Comté, national found "Agence Nationale de la Recherche", and "La ligue contre le cancer". The Funders had no role in study design, data collection, data analyses, interpretation, or report writing.

Results

TE differentiate colon, primary CRC, and liver metastases

It is already known that the expression of TE is differentially regulated in tumour and healthy tissues.[10] The liver microenvironment harbours a specific immune context promoting immune tolerance within liver metastases.[22] The specific regulation of TE in the context of lmCRC has never been thoroughly studied. Therefore, we decided to investigate if a difference exists between TE expression in primary tumours and liver metastases.

To carry out these analyses with appropriate statistical power, a meta-basis was set up including 436 RNA-seq samples from primary CRC (n = 191), liver metastases (n = 205), and normal adjacent tissues (n = 40) (Fig. 1a, Supplementary Fig. S1 and Table S1). Among these, 160 samples were paired with primary tumours and liver metastases from the same patients and over 400 samples had corresponding biological data. Clinical information was available for 208 out of 240 patients, including primary sidedness, microsatellite status, RAS/RAF mutations, exposure to chemotherapy, and survival (Supplementary Tables S2 and S3). To achieve more accurate calculations of TE and HERV loci expression, we utilized the recent algorithms REdiscoverTE and Telescope that integrate expectation maximization to assign ambiguous multi-mapped reads. Unlike most previous analyses of CRC, our goal was to characterize locus-specific expression to uncover potential positional associations with phenotype. Efficient batch-effect reduction and strong correlation (Pearson's correlation score above 0.9) between paired-end RNA-seq and single-end RNA-seq expression were observed at subfamily and locus level (Supplementary Figs. S2–S4).

TE overall expression may be characteristic of the tissue type.[10] The distribution of TE family and subfamily expression in tumoral tissue (Supplementary Tables S4 and S5) was described using the normal colon tissue as a reference. We first observed that the TE burden, which is the sum of TE expression per sample, was higher in tumours than in normal tissues (Wilcoxon

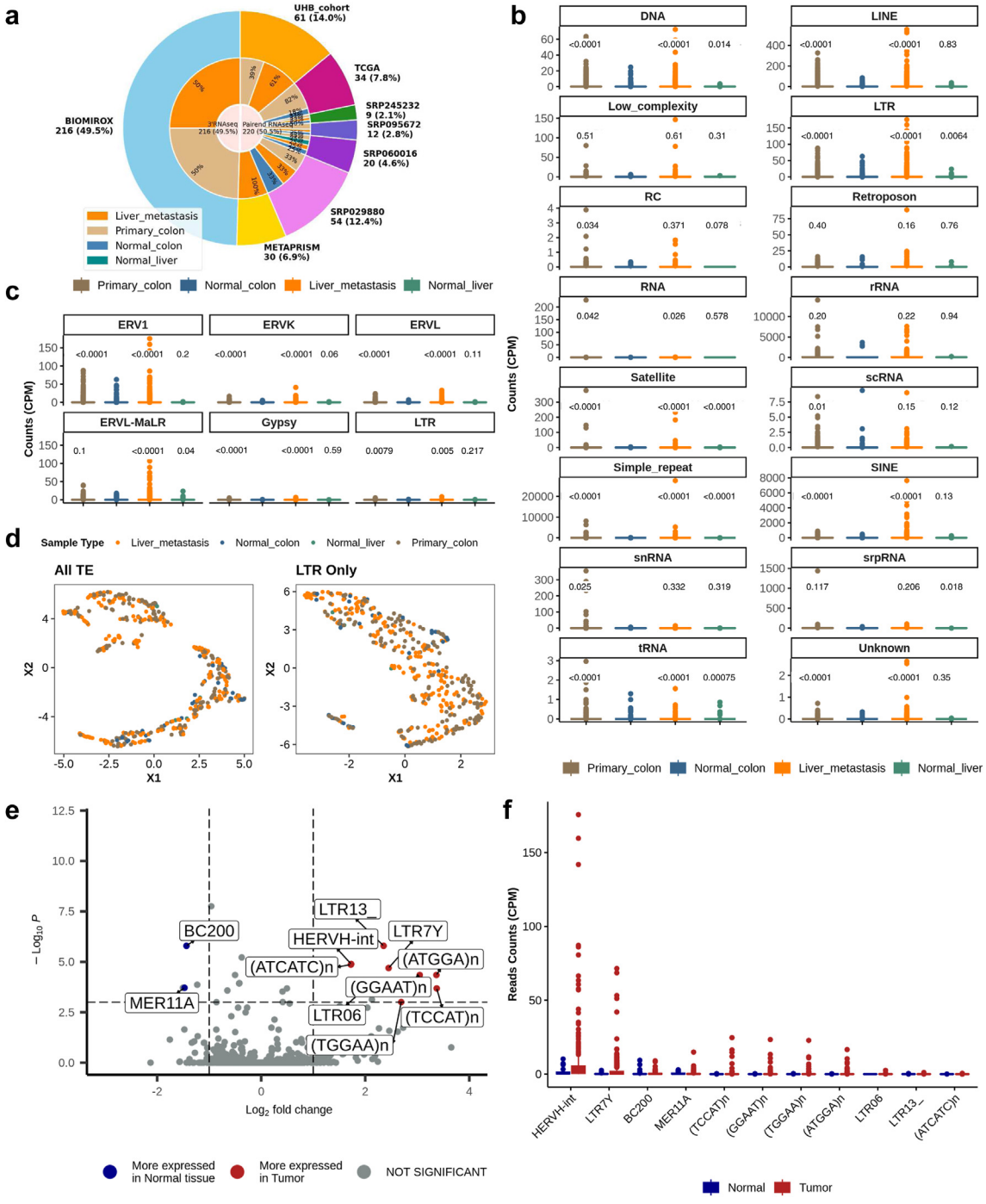

Fig. 1: TE differentiate colon, cancer, and liver metastasis. a) Chart showing datasets included, along with their sample type and the RNA-seq technology used to sequence the samples. The number of samples is shown below the cohort name or accession number, if publicly available, and the percentage of the total cohort is shown in parentheses. Half of the global cohort is paired-end RNA-seq and half is 3' single-end RNA-seq. b) Distribution of TE family expression across sample types. RC: rolling circle. c) Distribution of subfamily expression within the HERV/LTR family. d) UMAP based on TE expression for all samples in the meta-cohort. Samples are coloured according to sample type. The left panel uses the entire TE expression matrix, while the right panel uses LTR family only. e) Volcano plot for differential expression between normal and tumour tissues. Left blue labels are under-expressed in tumours compared to normal tissues. Right red labels are considered over-expressed in tumours compared to normal tissues. f) Distribution of expression of previously identified targets in the present cohort. For both differential expression analyses, all paired-end normal and tumour or primary and metastatic colon samples were normalized and analysed by DEseq2 on raw counts with batch effect reduction in the design. TE subfamilies are considered significant if log2foldchange >1, base mean > 10, and p-value < 0.001. For all plots, counts are normalized to counts per million (CPM) based on the gene size library. The p-values representing the significance levels of the Wilcoxon rank-sum test are displayed above each category, with normal colon serving as the reference.

test, p-value = 0.0048). TE expression burden was equivalent in primary tumours and liver metastases (Wilcoxon test, p-value = 0.32) (Supplementary Fig. S8). Both primary tumours and liver metastases showed significant enrichment for DNA, LINE, LTR, RNA, Satellite, Simple Repeat, SINE, and tRNA (Fig. 1b, Supplementary Table S15). Rolling Circle (RC) and snRNA appeared upregulated in primary tumours (Wilcoxon test p-value 0.034 and 0.025, respectively), but with no significance after adjusted p-value. A high heterogeneity was observed for expression levels of TE families and subfamilies (Supplementary Figs. S9 and S10). Then, we decided to address the heterogeneity of LTR subfamilies among CRCs. Within the LTR subfamilies, also known as ERV, ERV1 was the most abundant, followed by ERV2 (including ERVK) and ERV3 (including ERVL). All ERV families exhibited a significant overexpression in primary CRC and lmCRC samples compared to normal samples (Fig. 1c). Of note, this set of analysis points out a selective upregulation of ERVL-MaLR families in liver metastases but not in primary tumours (Wilcoxon test adjusted p-value $5.4.10^{-5}$). However, this was only driven by one liver metastatic sample overexpressing MLT1D and THE1B subfamilies. Thus, most TE are inactive in healthy colon tissue but can be highly expressed in a specific subset of colorectal carcinomas.

Sample clustering using TE expression identified two groups. Samples were intermixed across sample types, but normal tissues appeared to be enriched in one of the groups (Fig. 1d). Therefore, we investigated the differential expression of TE subfamilies between normal and tumour samples. We observed an increase in HERVH-int and LTR7Y (HERVH-int associated LTR), LTR13_, and LTR06 families (log2FoldChange 1.7, 2.4, 2.35 and 1.99 respectively), as well as simple repeats, and a decrease in MER11A and BC200 families (log2FoldChange 1.4 and 1.47 respectively) (Fig. 1e and Supplementary Table S16). Except for simple repeats and BC200, which is an Alu subfamily, most of the differential expression is driven by the LTR (i.e., ERV) family. The most common ERV family expressed in CRC is HERVH (Fig. 1f). We then analysed the difference in TE expression between primary tumours and liver metastases, focussing on previously identified TE that showed differences between normal and tumoral tissues. We observed a small difference in LTR13_, which was slightly more expressed in primary CRC (log2FoldChange 1.4, p-value $3.87.10^{-6}$), but had extremely low expression in normalized CPM counts (Supplementary Fig. S11). This suggests that there is no significant difference between primary and liver metastases at the TE subfamily level.

Thus, TE expression was heterogeneous between samples, but it was significantly higher in tumour samples, regardless of the primary or metastatic localization, compared to normal tissue. The LTR transposable family, mainly HERVH and LTR7Y, distinguished normal colon from CRC and appeared to be conserved between primary and liver metastases.

Locus-specific HERV expression in colorectal carcinoma

HERV families represent similar transposable nucleotide sequences throughout the genome. However, a specific interplay might occur at a specific locus where restoration of HERV expression modulates adjacent gene expression regulation. Therefore, it is necessary to interpret the impact of each HERV according to its integration site. Notably, the characterization of HERV expression according to the locus where the retrovirus sequence is integrated into the genome has only been performed on 24 primary CRC samples.[20] Therefore, the next set of analyses was dedicated to the description of HERV expression in CRC according to the locus of integration (Supplementary Table S6 and S7).

First, we searched for the general distribution of HERV loci across lmCRC. We ranked 14,968 annotated HERV loci by their mean expression. Only 5% of the annotated loci were expressed in lmCRC and expression levels were globally low (median of mean expression per loci 0.8 [1st Quartile 0.5 – 3rd Quartile 1.6] CPM). Six loci were expressed (count greater than 1 CPM) in more than 90% of samples, while 121 loci were expressed in more than 50% and 637 loci in more than 10%. Around twenty loci exhibit an average expression above 10 CPM, with approximately half of them being from the HERVH family (Fig. 2a). The most highly expressed locus in lmCRC was HERVH_Xp22.32a. We identified a high degree of heterogeneity between samples and loci, with some consistently expressed and others modulated in tumours (Supplementary Fig. S12).

To gain insight into the selective distribution of HERV according to the integration sites, we investigated overall HERV expression per sample. As previously described, looking at the sum of HERV expression per sample, termed HERV burden, there was a significant increase in HERV expression in tumours compared to normal tissue (rank-sum Wilcoxon test, p-value = $1.4.10^{-11}$ and $1.8.10^{-11}$ for primaries and metastases, respectively). However, no statistically significant differences were observed between primary CRC and liver metastases (Wilcoxon test, p-value = 0.37) (Fig. 2b, Supplementary Fig. S13). This difference in HERV burden between normal and tumour samples was not consistent across all chromosomes: no statistically significant differences were observed in chromosomal HERV burden between normal and tumour samples on chromosomes Y, 15, 17, 21, and 22, while there was a strong overexpression of HERV in tumours on chromosomes X, 7, and 13 (Fig. 2c and Supplementary Fig. S14). This suggests the presence of

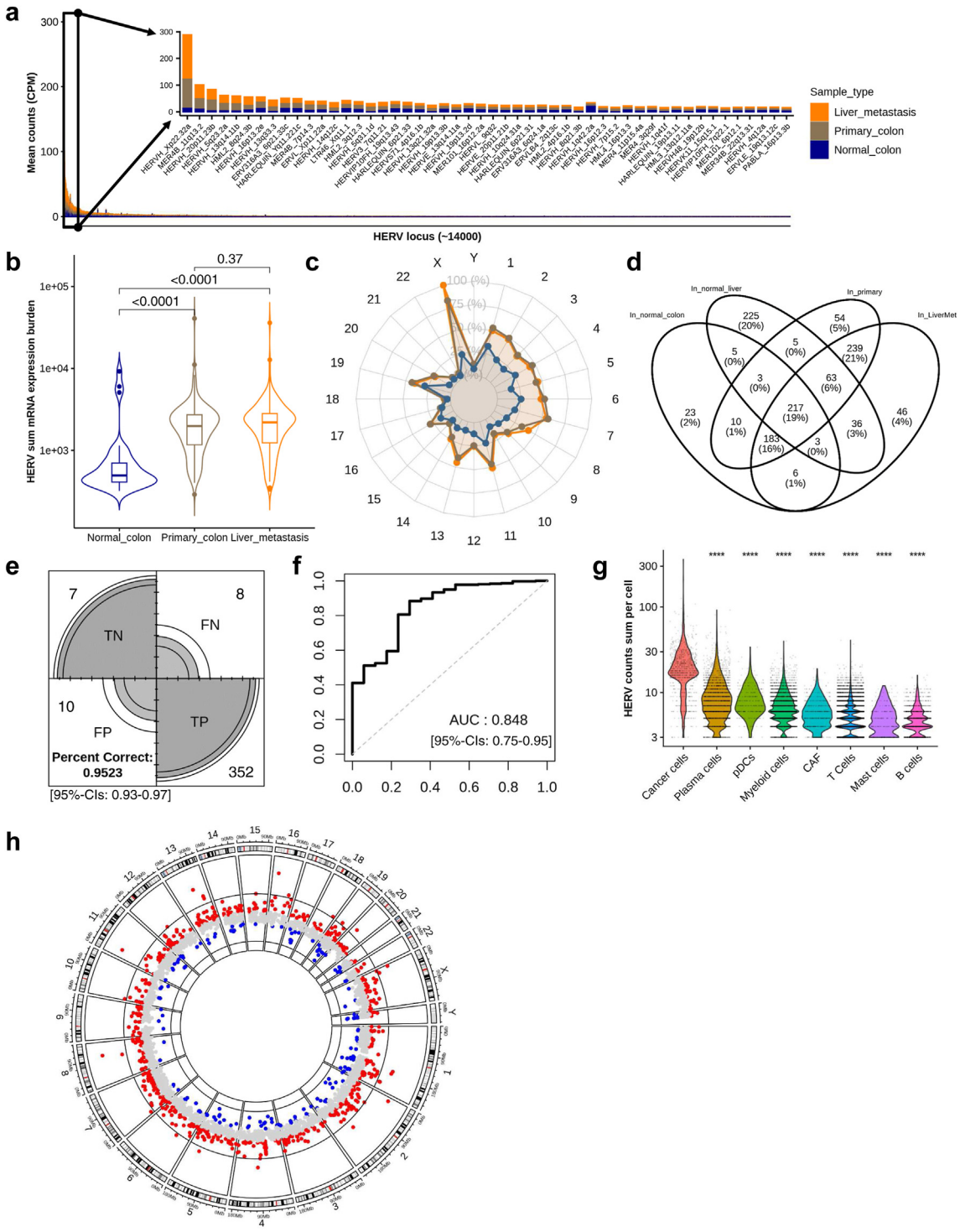

Fig. 2: Locus specific HERV expression in colorectal carcinoma. a) Expression distribution of the 14,968 HERV loci across sample types. Counts are normalized to counts per million (CPM) based on the gene size library. Focus on the fifty most highly expressed HERVH loci. b) Distribution of HERV total expression (considered as HERV burden) per sample according to sample type. Observations outside the interquartile range are considered outliers and are marked as individual points in the plot. The p-values representing the significance levels of the Wilcoxon rank-sum test are displayed above each category. c) HERV total expression per chromosome, for each sample type. The expression range is from 0 to 300 CPM. d) Venn diagram of expressed HERV loci in samples. Only loci expressed above 1 CPM in >10% of tumours are shown. e) Confusion matrix

hotspot HERV loci. Here too, the chromosomal HERV burden distribution is similar in paired primary CRC and liver metastases.

Consequently, we explored HERV loci recurrent expression in tissues. Out of the 14,968 annotated loci, we found that 1118 (7.5%) were expressed (>1 CPM) in more than 10% of the samples (Fig. 2d). Among these, 253 (22%) were exclusively expressed in normal tissues and 339 (30%) were exclusively expressed in tumours but not in normal tissues. All of the 35 HERV loci expressed in 80% of tumour samples were also identified at lower levels in normal colon (Supplementary Fig. S15). Thus, a minority of HERV in lmCRC are selectively transcribed during CRC oncogenesis. HERV burden in CRC is mainly the result of overexpression of sequences transcribed at low levels in normal tissues.

Given the above observation, we assessed whether differential HERV expression could discriminate between healthy and tumoral colonic tissue. Unsupervised clustering discriminated normal colon and CRC tissue with HERV loci in a different way than gene expression, suggesting that gene and HERV expression may be complementary for tumour sample classification (Supplementary Fig. S16). Then, we used supervised machine learning classifiers based on raw HERV counts. A normal/tumour classifier with 7750 parameters was trained on the SRP029880 dataset (18 patients with paired primary CRC/Liver metastases/Adjacent normal colon tissues). When tested on all other datasets, it achieved an accuracy of 0.95 [95%-CIs: 0.93–0.97] (Fig. 2e) and AUROC of 0.848 [95%-CIs: 0.75–0.95] (Fig. 2e). Altogether, these analyses point out HERV as a relevant determinant to discriminate against mCRC.

Metastatic spreading is commonly associated with the acquisition of tumour genomic heterogeneity. Consequently, we assessed if the expression of HERV differs between primary and liver metastatic CRC. Based on the list of dysregulated normal/tumour loci identified previously and excluding HERV loci expressed in normal liver samples, only nine targets were revealed as differentially expressed between primary CRCs and liver metastases (Supplementary Figs. S17 and S18). PABLA_19q13.32 locus was the most overexpressed in liver metastases (log2FoldChange –1.38, adjusted p-value 8.10^{-4}). Conversely, HML3_17q21.32 and ERVLE_19q13.31a were preferentially expressed in primary CRC (log2FoldChange 4.64 and 3.91, respectively) (Supplementary Fig. S19 and Table S17). However, those HERV exhibited either low expression in tumour samples, basal expression in normal tissue, or low reproducibility across the independent dataset tested. Therefore, these analyses failed to delineate a specific HERV-related signature discriminating between liver metastases and primary CRC.

HERV locus efficiently discriminate lmCRC from normal tissues

Having demonstrated that HERV might be re-expressed in CRC samples and remains stable across metastatic spreading, we sought to identify the list of HERV candidates potentially linked to the disease activity. In a first step, we investigated the cell types that produce the HERV RNA within the CRC microenvironment. To elucidate which cell subsets transcribed HERV mRNA, a single-cell RNA-seq dataset of colorectal liver metastases[41] was analysed (Supplementary Tables S11 and S12). The HERV-related mRNA load of the samples was indeed carried by cancer cells (Fig. 2g, Supplementary Figs. S7 and S20). We identified 510 specific loci (Supplementary Table S13) producing HERV mRNA in cells belonging to the tumour microenvironment. It seems that HERV loci, transcribed by non-cancer cells, are mainly located in T lymphocytes.

This analysis allowed the selection of a list of HERV loci selectively expressed in CRC cells. Over 155 loci that exhibited differential expression (log2FoldChange above or under 1, baseMean above 10, p-value adjusted $<10^{-3}$) between normal and tumour tissues, 111 loci were found to be overexpressed, while 44 were underexpressed in tumours (Table 1 and SupplementaryFigs. S21–S24 and Table S18). Finally, HERV loci characterizing lmCRC were not randomly expressed on the genome (Fig. 2h). Overexpressed loci (log2FoldChange above four) were present in greater numbers on chromosomes 1, 13, and X.

Thus, ERV locus expression efficiently discriminates colorectal cancers from normal tissues. We have demonstrated the global stability of HERV locus regulation across primary colorectal cancer and liver metastases, suggesting that HERV expression is an early event in tumour carcinogenesis. A list of HERV loci candidates is proposed to investigate how HERV re-expression contributes to the biological and clinical features of lmCRC.

of the combined datasets (all but the training one) used to test the Elastic Net model trained to categorize normal and tumour tissue on SRP029880. TN: True Negative, FN: False Negative, FP: False Positive, TP: True Positive. "Percent correct" represents the accuracy of the model. f) ROC curve of the previous model. g) Distributions of log-transformed total HERV counts per cell for each tumour component from a single-cell RNA-seq atlas of liver metastatic colorectal cancer. 28,266 cells were classified based on gene expression into T cells (n = 18,262), Plasma cells (n = 2454), Myeloid cells (n = 4082), B cells (n = 1615), Cancer cells (n = 902), pDCs (n = 289), Cancer-associated Fibroblast (CAF) (n = 443), and Mast cells (n = 219). h) Positional representation of differentially expressed HERVs. Each dot represents the log2 fold-change of normal versus tumour expression of a single HERV locus, with values ranging from 4 to –4. Coloured dots are for significantly differential loci (blue: under-expressed, red: overexpressed), based on DEseq2 criteria: padj < 0.001 and abs (log2(Fold Change) > 1) and mean CPM >10. For all plots, counts are normalized to counts per million (CPM) based on the number of reads aligned on genes.

HERV_locus	baseMean	log2FoldChange	padj
HERVH_8q22.2	309.44	8.05	<0.0001
HERVL_16p12.3b	43.58	7.76	<0.0001
HARLEQUIN_2p24.3	48.62	7.26	<0.0001
HERVH_13q21.32a	300.44	7.01	<0.0001
HERVH_7p15.2	52.34	6.43	<0.0001
HERVH_Xp22.32a	8480.37	6.04	<0.0001
HERVL_16p12.3a	17.18	6.00	<0.0001
HERVH_1q25.2	118.73	5.82	<0.0001
MER4B_1q31.3	14.64	5.82	<0.0001
HERVH_4q24b	159.63	5.45	<0.0001
MER61_10q26.12	14.65	−2.07	0.000779
HERVH_4p15.2a	27.68	−2.29	<0.0001
HERVH_14q21.2c	10.37	−2.41	<0.0001
ERVLB4_11q23.3a	110.21	−2.74	<0.0001
HERVL_17q21.32	12.71	−2.74	<0.0001
HERVH_11q24.1d	17.30	−2.76	0.00089
HERVIP10FH_4p13	14.81	−3.05	<0.0001
HML3_17q21.32	353.34	−3.10	<0.0001
HERVL74_2q11.2	30.42	−3.32	<0.0001
ERVLE_19q13.31a	77.05	−3.52	<0.0001

Table 1: **Top 10 up and down regulated in differential expression normal colon versus tumours.**

High HERV burden mCRC display a low innate immune infiltration

In the next investigations, we aimed to characterize liver metastatic patients' biological and clinical features based on the ability of mCRC to reactivate HERV loci transcription. For this purpose, we clustered patients with liver metastases using the K-means method based on previously proposed normal/tumour differential HERV loci, assuming their stability across tumour differentiation. We observed two groups with no batch effects associated with one on the other (Supplementary Fig. S25), distinguished by high (114 patients, mean number of HERV loci expressed = 27) or low (91 patients, mean HERV loci = 9) HERV expression.

No significant correlations were observed between HERV expression and the primary cancer sidedness, RAS/RAF mutations, microsatellite instability, age, and pre-treatment by chemotherapy (Supplementary Tables S19 and S20). Additionally, overall survival after liver metastasis surgery did not different between the two groups (log-rank p-value = 0.64) (Supplementary Fig. S26). The transcriptomic profile of mCRC was also investigated using DEseq2 in patients with a low or high number of overexpressed HERV loci. 487 genes were differentially expressed, but no significant biological pathways could recapitulate a significant pattern of the two groups. By contrast, immune deconvolution of cellular subpopulation with MCP counter suggested a decrease in innate immune cells such as macrophage/monocyte, myeloid dendritic cell, and neutrophil when HERV expression was enhanced (Wilcoxon p-value 2.10^{-4}, $6.5.10^{-5}$, 3.10^{-6}, respectively) (Supplementary Fig. S27).

Thus, the HERV expression burden does not seem to be associated with clinical variables nor biological processes. A weak association was suggested between immune cell infiltration and the number of overexpressed HERV loci. This implies that while HERV expression may not directly correlate with the primary clinical and molecular characteristics of mCRC, it could play a role in modulating the immune landscape within the tumour microenvironment. Further studies are necessary to elucidate the mechanisms by which HERV reactivation influences immune cell dynamics and to determine its potential as a therapeutic target or biomarker in metastatic colorectal cancer.

Identification of HERV subset candidates potentially involved in mCRC

Since the analysis of the global HERV burden did not identify a significant impact of HERV on lmCRC behaviour, we next sought to clarify if individual HERV loci expression might be correlated with the clinical or biological disease characteristics. Only 25% of the HERV sequences from the full cohort analysis were expressed in at least two single datasets (Supplementary Fig. S28). Then, we further investigated HERV expression at the single locus level, to identify potential regulatory elements or colorectal biomarkers in these HERV loci. We selected HERV candidates exhibiting high RNA expression levels, very high differential expression between normal and tumour tissues, and validation in at least two independent cohorts. This resulted in a list of 17 ERV-derived RNA associated with lmCRC (Table 2, Fig. 3a and b, Supplementary Fig. S29). The 17 selected HERV loci were not generally coexpressed, except for HERVH_20p11.23b and HERVH_Xp22.32a (Pearson correlation score of 0.51).

To further validate the cancer-specific profile of the selected HERVs, we measured their expression in 1137 normal samples from GTEx (Supplementary Fig. S30 and Tables S9 and S10), an RNA-seq dataset of normal tissues from non-cancer patients. ERV expression measures using a reference-free counting method[25] confirmed that the proposed targets were generally not expressed in normal liver and colon tissues but could be expressed sporadically in other tissues (Fig. 3c, Supplementary Figs. S30 and S31).

We performed RT-PCR to confirm the presence of five HERV candidates displaying high levels of mRNA expression in RNA-seq (HERVH_Xp22.2c, HERVH_4q24b, HERVH_13q33.3, HERVH_13q31.3a, and HERVH_Xp22.32a) in lmCRC. RT-PCRs were conducted on 59 samples of liver metastases, associated primary CRC, and adjacent normal colon/liver tissues in an independent cohort of patients with lmCRC. This revealed increased expression of the selected HERV in

Locus	Start	End	HERV Family	LTR	qRT-PCR validation	GTEx expression	Associated Peptide	Prognostic value	CNA association	TME modulation	Cell cycle correlation
Xp22.32a	4,540,474	4,546,320	HERVH	LTR7Y	yes	Not in colon	GGCSLVGRGGSHK	yes:poor	no	yes	yes
20p11.23b	19,752,049	19,756,776	HERVH	LTR7B	no	Not in Colon	none	yes:poor	yes	no	yes
13q33.3	109,265,090	109,271,116	HERVH	LTR7Y/ LTR7u2	yes	Not in Colon	CFSGLLQAGLGAAWEPRVGEIK	yes:poor	no	yes	yes
5q31.1d	136,540,943	136,545,393	HERVH	LTR7Y/ LTR7u2	no	Never expressed	ICTLSTK	no	no	yes	yes
13q21.32a	66,141,332	66,147,036	HERVH	LTR7	no	Never expressed	none	no	no	yes	no
8q21.3b	90,090,224	90,095,868	HERVH	LTR7Y/ LTR7u2	no	Never expressed	none	yes:good	no	no	yes
7p15.2	26,024,200	26,029,809	HERVH	LTR7Y	no	Rarely expressed in colon	LYIPYGPPSSRK	no	no	yes	yes
1q41f	221,965,924	221,971,633	HERVH	LTR7Y/ LTR7u2	no	Never expressed	CFSGLLQAGLGAAWEPRVGEIK	no	no	partially	yes
Xp22.2c	16,179,202	16,184,434	HERVH	LTR7/ LTR7d1	yes	Never expressed	NTRALPADR	yes:good	no	partially	yes
2p24.3	13,525,956	13,533,609	HARLEQUIN	LTR2	no	Never expressed	none	no	no	no	yes
1q25.2	178,040,241	178,045,930	HERVH	LTR7Y/ LTR7B	no	Not in Colon	none	no	no	partially	yes
8q22.2	99,943,695	99,949,609	HERVH	LTR7/ LTR7u2	no	Not in Colon	none	no	no	partially	yes
13q31.3a	90,839,564	90,845,144	HERVH	LTR7Y/ LTR7u2	yes	Never expressed	CFSGLLQAGLGAAWEPRVGEIK	yes:poor	no	yes	yes
4q24b	103,553,770	103,559,474	HERVH	LTR7Y/ LTR7u2	yes	Not in colon	ESAKGHGVGPFYR	no	no	yes	yes
1q41g	221,973,078	221,978,957	HERVH	LTR7/ LTR7up1	no	Never expressed	CFSGLLQAGLGAAWEPRVGEIK GASTLNPFSFTLTGK	no	no	yes	no
13q31.3b	90,848,573	90,854,278	HERVH	LTR7/ LTR7up1	no	Never expressed	NTRALPADR	yes:poor	no	yes	yes
19q12	28,606,436	28,615,273	HERVE	LTR2	no	Not in colon	VVEMGETQSKPTPLGTMLK	no	no	no	no

Table 2: **Selection of mCRC-associated HERV elements.**

80% of the liver metastases (mean relative expression 19 [range for individual locus 7–37]) and 70% of the primary CRC (mean relative expression 9 [range for individual locus 5–15]) (Fig. 3d, Supplementary Table S21). HERVs were only detected in 25% of normal colon and liver tissues but at very low levels (mean relative expression 1.3) compared to the lmCRC. Moreover, the five HERV candidates were expressed at similar levels between primary tumours and liver metastases. This provides an independent validation of the presence of HERV-derived RNA in liver metastases of patients with CRC.

Expressed HERV can be translated into functional, immunogenic proteins. To investigate this, we analysed tumour MS/MS spectra. DNA sequences of all (~50,000) HERV elements were retrieved from RepeatMasker coordinates and processed by PePQuery[42] against 90 individual primary colorectal cancer spectra from TCGA.[1] We identified about 2400 peptides overall (Supplementary Table S22). Among the 155 HERV loci identified in the normal/tumour differential analysis, fifty peptides were identified, and these were shared by a maximum of four patients. Peptides were identified from eleven of the seventeen selected HERV CRC markers (Supplementary Fig. S32).

HERV loci displaying a prognostic value after resection of lmCRC

We next sought to assess the clinical and biological characteristics of HERV-expressing mCRC. We observed no significant differences in the primary cancer sidedness, RAS/RAF mutations, microsatellite instability, and pre-treatment by chemotherapy, according to the expression of the selected HERV loci (Supplementary Table S23). Survival after surgical resection of liver metastasis was analysed in 137 patients from the BIOMIROX and UHB cohorts. We discretized HERV expression as either expressed or silent and applied a Lasso-Cox model to determine how HERV expression predicts survival. Six significant markers were identified (Fig. 3e). Four of them, HERVH_Xp22.32a, HERVH_20p11.23b, HERVH_13q33.3, and HERVH_13q31.3, were correlated with poor prognostic when expressed (log-rank p-value 0.028,

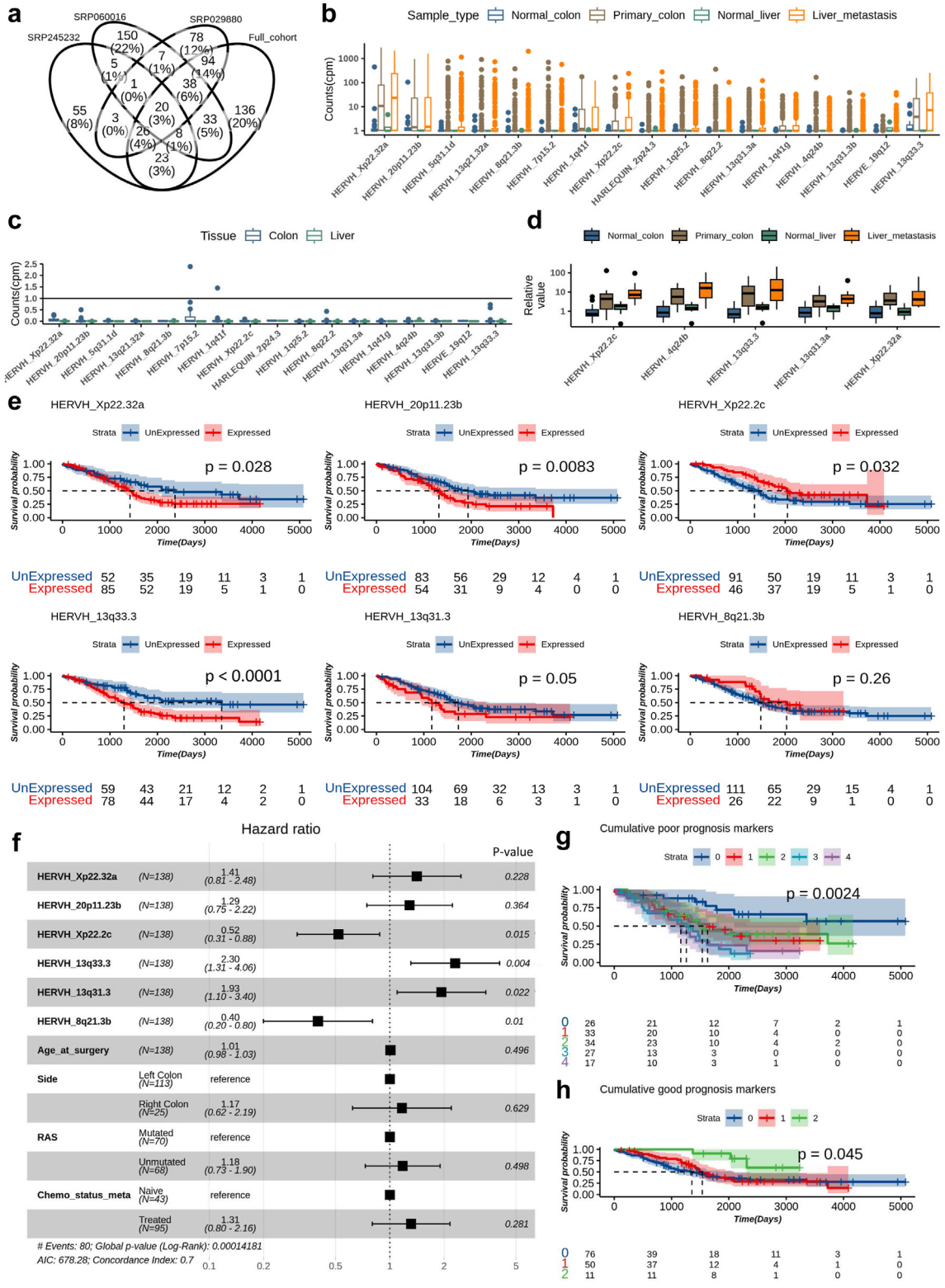

Fig. 3: ERV-derived RNAs have prognostic value after surgery for liver metastases. a) Venn diagram for HERV locus significantly different from independent and full meta-cohort differential expression analysis of normal versus tumour tissues. Locus is considered significant if log2 (fold change) > 1, mean CPM >10, and p-value < 0.001. b) Count distribution of proposed colorectal cancer biomarker HERV elements. Selected

$8.3.10^{-3}$, 9.10^{-4}, 0.05, respectively and Hazard Ratio of 1.7, 1.8, 2.2, 1.55 respectively). By contrast, HERVH_Xp22.2c expression was associated with a better prognosis (log-rank p-value 0.032 and Hazard Ratio of 0.6). All multivariable models tested confirmed HERVH_8p21.3b as a good prognostic marker. Only HERVH_13q33.3 showed a significant association with progression-free survival (log-rank p-value 0.016 and Hazard Ratio of 1.6). Furthermore, we built a multivariable prognostic model including clinical variables and the HERV prognostic markers (Fig. 3f). The model showed a global p-value of 0.00014 and a C-index of 0.7 [95% CIs 0.64–0.75]. HERVH_Xp22.2c and HERVH_8q21.3b exhibited good prognostic value, whereas HERVH_13q33.3 and HERVH_13q31.3 were associated with poor prognostic value. HERVH_Xp22.32a, HERVH_20p11.23b, and the available clinical markers (age, tumour side, RAS mutation status, and chemotherapy treatment) did not demonstrate independent prognostic value. Next, we assessed if the proposed prognosis markers were independent or displayed a cumulative impact on survival. Poor prognostic markers had an additive contribution to low survival, with likely three sub-groups, none, 1 or 2, 3 or 4 poor prognostic markers expressed (Global log-rank p-value = 0.0024, Hazard Ratio respectively 2.6 and 4.4) (Fig. 3g). The same observation was made with the two good prognostic markers, when they are conjointly expressed, patients seem to have better survival after liver metastasis surgery (Global log-rank p-value = 0.045, Hazard Ratio of 0.25) (Fig. 3h).

Correlation of candidate HERV expression and mCRC pathways

To explore the biological pathways associated with the expression of CRC-specific HERVs, we performed gene set enrichment analysis on the genes co-expressed with those elements (Supplementary Fig. S33). All HERV markers were associated with cell cycle signatures, particularly MYC targets, E2F targets, and G2M checkpoint (Fig. 4a, Supplementary Table S24). Most of them were associated with DNA repair, MTORC1 signalling, mitotic spindle, and TNFα signalling via NFκB, regardless of their positive or negative prognostic value (Fig. 4a, Supplementary Fig. S34). HERVH_20p11.23b were associated with a larger number of activated pathways, 23 out of 26. Additionally, three out of four poor prognosis markers, HERVH_13q33.3, HERVH_20p11.23b, and HERVH_Xp22.32a, were significantly associated with stemness (Supplementary Table S25).

A cytotoxic immune environment is known to have prognostic value in colorectal cancer.[22] We performed immune deconvolution using the mcp-counter algorithm.[43] Except for HERVH_20p11.23b, all poor prognostic markers showed lower T cell counts (Fig. 4b). HERVH_Xp22.32a, HERVH_Xp22.2c, and HERVH_13q33.3 showed significantly lower NK cell counts (p-value $1.2.10^{-3}$, $5.2.10^{-5}$, $2.5.10^{-3}$, respectively) (Fig. 4c). B cells, myeloid dendritic cells, and neutrophils were decreased for all markers except HERVH_20p11.23b and HERVH_8q21.3b. The monocyte expression did not significantly affect the HERV score (Supplementary Fig. S35).

The HERVH 8q21.3b integration site is associated with CALB1 expression

HERV re-expression at specific target integration sites might interfere with gene expression regulation and influence cancer-related biology. We assessed the genomic locations and annotations of regulatory elements identified in previous experiments. HERVH_8q21.3b, HERVH_13q33.3, and HERVH_20p11.23b are associated with exons and promoters, while others are intergenic (Fig. 5a). HERVH_13q33.3 and HERVH_20p11.23b are associated with coding sequences specifically, MYO16 and RIN2 respectively.

HERVH_20p11.23b is located near the promoter of RIN2, a gene involved in the MAP kinase pathway, but there was no correlation between HERVH_20p11.23b

elements have very high differential expression (log2FC > 4 and absolute mean difference >2) between normal and tumour tissues, or positive differential expression confirmed in at least two independent cohorts and strong absolute expression difference (>10 CPM), or positive differential expression with identified protein expression. Observations outside the interquartile range are considered outliers and are marked as individual points in the plot. c) Expression of the above HERV loci in GTEx normal colon (blue, 48 samples) and liver (green, 14 samples). HERV Expression was quantified using the reference-free software Reindeer on 1137 samples from the GTEx database using HERV DNA sequences extracted from DFAM.org. d) Distribution of the relative RT-PCR quantification of 5 selected HERV loci in 59 independent samples. Normal colon (n = 16), normal liver (n = 10), primary colon (n = 18), and liver metastasis (n = 15). e) Kaplan–Meier curves for HERV loci associated with overall survival after surgical resection of liver metastasis in univariate survival analysis and multivariable lasso cox model. f) Multivariable Cox model for overall survival after surgical resection of liver metastases incorporating previous HERV prognostic markers and available clinical variables. Multiples imputations of missing data. Right: p-value, middle: hazard ratio and 95% confidence interval. g) Kaplan–Meier curve showing the cumulative prognostic value of poor prognostic markers. The number of each category represents the sum of the poor prognostic markers presented in E expressed for each individual. h) Kaplan–Meier curve showing the cumulative prognostic value of good prognostic markers. The number of each category represents the sum of good prognostic markers shown in E expressed for each individual. In e, g and h, The shaded area refers to the 95% confidence band. For all plots, counts are normalized to counts per million (CPM) based on the gene size library. Normal colon samples are blue, normal liver samples are dark green, primary colorectal samples are brown, and liver metastasis samples are orange. For all Kaplan–Meier curves, the p-value of the log-rank is shown on the graph.

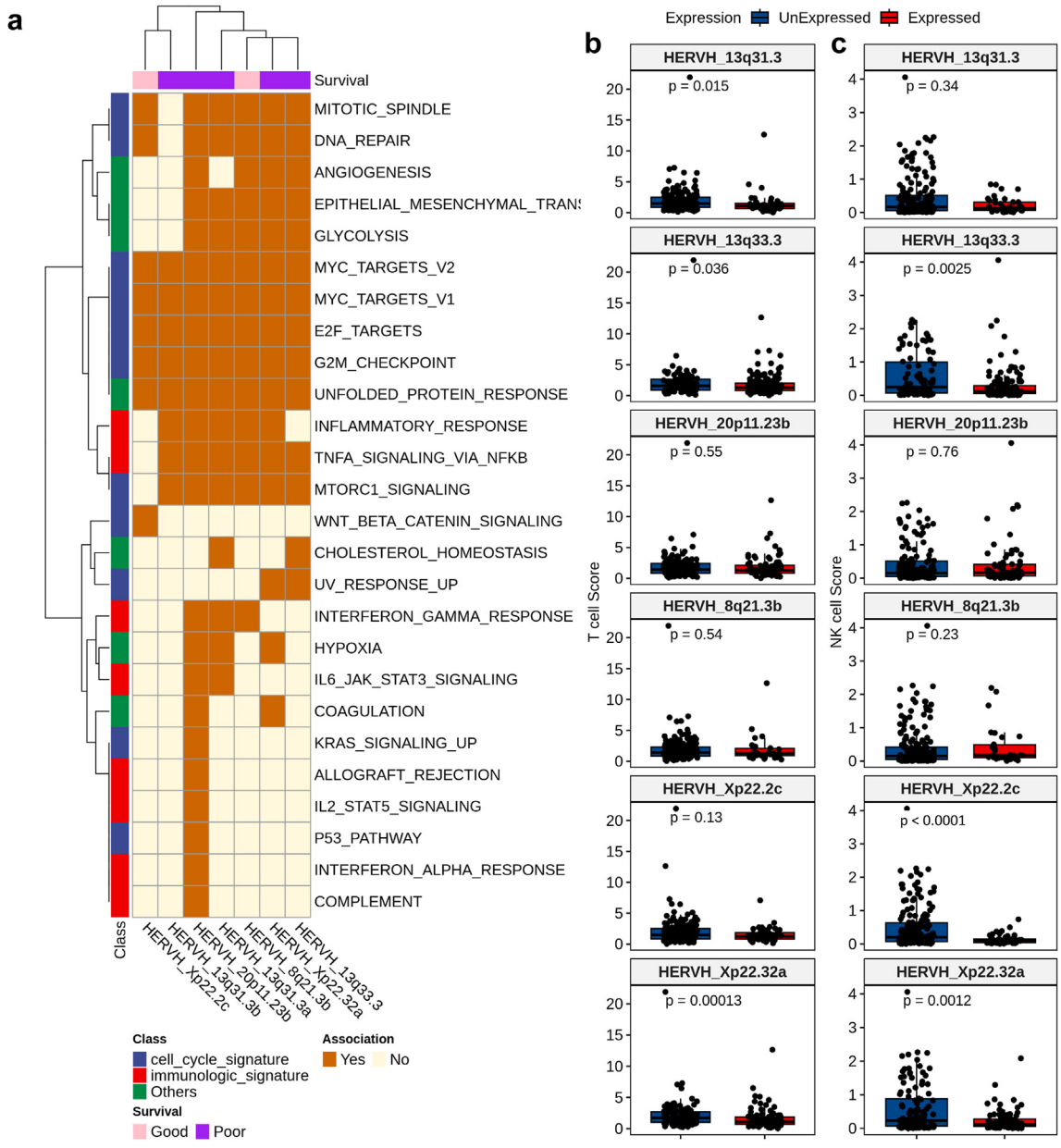

Fig. 4: ERV biomarkers of metastatic colorectal cancer are associated with the activation of cell cycle pathways and a decrease in the cytotoxic immune environment. a) Association of MsigDB Hallmark pathways with HERV loci considered as prognostic markers. The gene set enrichment score was calculated by ranking Spearman's correlation score between the HERVH locus and gene expression. Enrichments with p-value < 0.05 were considered as significant (coloured in heatmap). Annotations are provided for prognosis (columns) and pathway class (rows). b) T cell score from immune deconvolution with MCP-counter for target loci. Samples are categorized according to whether the locus is expressed (red) or not (blue). Locus is considered expressed if over 1 CPM in the sample. p-values are from Wilcoxon tests. c) Same as b with NK cell score.

and the related gene expression (Spearman correlation coefficient of 0.042 [95% CIs –0.053 to 0.13], p-value = 0.39) (Fig. 5b). HERVH_8q21.3b correlated strongly with CALB1 expression (Spearman correlation coefficient of 0.92, 95% CIs 0.90–0.93, p-value < 0.0001) (Fig. 5c). Tumour purity did not affect the observed correlation. Notably, HERVH_8q21.3b was identified as an alternative promoter of the CALB1 gene (Fig. 5d), a member of the calcium-binding protein family. However, despite the good prognostic value of HERVH_8q21.3b, CALB1 was not associated with survival after liver metastasis resection (log-rank p-

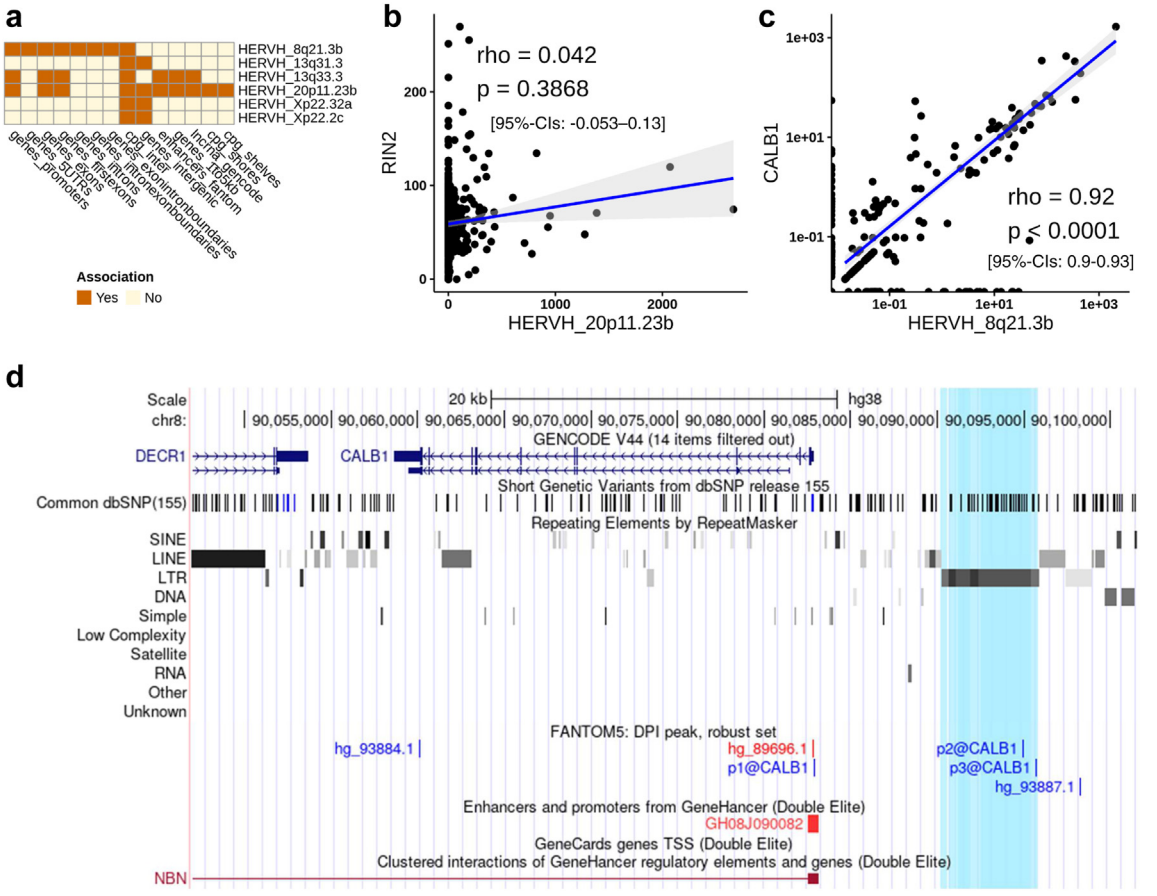

Fig. 5: The good prognostic marker HERVH_8q21.3b presents co-expression with the CALB1 gene. a) Heatmap showing the association between prognostic HERV loci and genomic annotations. Annotations (columns) represent genomic features at the locus positions. b) Correlation between expression of HERVH_20p11.23b and gene RIN2 (CPM counts). R: Pearson's correlation coefficient. c) Correlation between expression of HERVH_8q21.3b and gene CALB1 (CPM counts). R: Pearson's correlation coefficient. In b and c, the shaded area refers to the 95% confidence band. d) UCSC Genome Browser view of CALB1 and HERVH_8q21.3b. Tracks shown from top to bottom: genomic position, CALB1 transcripts, dbSNP, RepeatMasker annotation, FANTOM5 and GeneHancer (enhancers). Position of HERVH_8q21.3b is highlighted in blue.

value = 0.45) (Supplementary Fig. S36), suggesting an independent effect of the HERV locus on prognosis.

Discussion

We aimed to assess the pattern of HERV expression in primary and liver metastatic CRC and to investigate the biological and clinical correlates of transposable element modulation. First, we found that HERV can accurately classify tumoral and normal colon tissues. HERV elements discriminating metastatic from primary tumours are host-organ related rather than tumour-specific, consistent with global stability of HERV expression across primary CRC and metastases. Second, we identified 17 cancer-specific HERV-derived RNAs associated with lmCRC, confirmed not to be expressed in normal tissues. Among those, we extracted six HERV loci of prognostic value following liver metastases resection in patients with lmCRC (HERVH_Xp22.32a, HERVH_20p11.23b, HERVH_13q33.3, HERVH_13q31.3, HERVH_Xp22.2c, and HERVH_8q21.3b) and characterized their association with the cell cycle pathway and immune microenvironment.

We assembled a cohort of over 400 RNA-seq samples from both primary and metastatic colorectal cancer (CRC) across eight independent datasets. This cohort includes more than 200 RNA-seq samples from liver metastases, with over 100 previously unpublished metastasis/primary CRC pairs. Previous research on primary CRC has largely focused on transposable element (TE) expression at the family level[10,44,45] or on single HERV loci using RT-PCR.[46] In contrast, prior studies examining locus-level ERV expression were limited to sample sizes of 24, 55, and 114.[20,27,47]

This study provides large-scale validation of locus-level analyses in primary colorectal cancer and suggest

potential clinical implications of HERV expression in more advanced tumours compared to those previously described. Additionally, we performed an extensive data normalization process, incorporating both single-end and paired-end sequencing data across the different cohorts, demonstrating the utility of cost-effective 3' single-end RNA-seq for ERV element discovery.

The study has several limitations. First, unmeasured confounding arises from the inclusion of parameters that are either unknown or unavailable for analysis. The statistical design could not be fully implemented due to the limited and retrospectively acquired data. Notably, the required sample size for the proposed statistical analysis was not determined a priori. However, the narrow width of the 95% confidence intervals suggests sufficient statistical power, indirectly supporting the adequacy of the sample size. Additionally, missing values, such as RAS mutations, necessitated the use of statistical methods to retain clinically relevant variables. Second, the samples were not specifically collected for this analysis, rendering RNA-seq methods suboptimal in some cases.

The use of hazard ratios (HR) for causal inference is not straightforward, even in the absence of unmeasured confounding.[48] It is important to consider the inherent selection bias in hazard ratios, particularly in this study. Therefore, the presented HR values are exploratory and should not be interpreted as definitive.

However, we identified two results with could have a clinical significance. The first is based on the stability of the expression of different HERV loci between primary hepatic CRCs, suggesting HERVs as biomarkers for metastatic CRCs and as targets for therapeutic strategies aimed at overcoming tumour heterogeneity. The second finding is that HERVs are independent markers associated with patient survival after hepatic metastasis resections. This suggests a complementary role in tumour biology alongside the well-established prognostic factors.

The characterization of TE expression in cancer is still ongoing. Analysis of a subset of TCGA colon tumours identified differentially expressed ERV loci across cancer types.[20] Comparing the top 10 up- and downregulated ERV loci from this study and ours, we could identify a common upregulated loci (HERVH_Xp22.32a) and 5 common downregulated loci (ERVLE_19q13.31a, HERVL74_2q11.2, HML3_17q 21.32, HERVL_17q21.32, ERVLB4_11q23.3a). However, this previous study only included 24 colon samples, thus we ascribe this difference to our larger cohort and selection of tumour cell-specific loci based on single-cell data. Furthermore, three of our CRC-specific candidates were common with tumour-specific ERV previously reported in Head and Neck cancer samples (HERVH_18q22.3a, HERVH_8q21.3b or HERVH_4q24b)[49] and two candidates (HERVH_Xp22.32 and HERVH_13q33.3) in esophageal adenocarcinoma.[50]

Interestingly, all upregulated HERV markers identified in our CRC datasets are part of the HERV-H family. It has been suggested that the integrity of HERV-H LTRs is essential for the regulation of HERV-H transcription levels.[17,51,52] Through a process resembling co-option, certain HERV-H loci beneficial to the host may have retained conserved sequences, which could have restricted their diversification. These loci, initially tolerated, may eventually have been co-opted to contribute to stem cell identity.[17,19,52,53] Sun et al. proposed a mechanism for LTR7/HERV-H re-expression: YTHDC2 recruits the DNA 5-methylcytosine (5 mC)-demethylase, TET1, to remove 5 mC from LTR7/HERV-H and prevent epigenetic silencing.[54] In esophageal and colorectal cell lines, HERVH_Xp22.32, HERVH_13q33.3, and HERVH associated with CALB1 (HERVH_8q21.3b) were linked to KLF5 and SOX9. KLF5 binding to the proviral LTRs, and its loss of activity, resulted in reduced expression of these three proviruses.[50] Similar findings for HERVH associated with CALB1 were observed in lung squamous cell carcinoma, reinforcing the biological association between KLF5 and LTR7Y.[55] In contrast, ARID1A loss, recently implicated in global HERVH activation in COAD,[45] appeared specific to HERVH_1p31.3.[50] However, this study was not designed to identify new processes linked to HERVH family reactivation in CRC, and further analyses of epigenetic mechanisms are needed to confirm the role of HERVs in this context.

Prior studies based on samples derived from primary CRC identified a negative prognostic value of TE[44] and HERV.[47,56] In addition, HERV could discriminate between different groups of survival in Head and Neck carcinoma. Recent studies on hepatocellular carcinoma[57] and lymphoma (Singh B, Locus specific human endogenous retroviruses reveal new lymphoma subtypes, unpublished), extensively characterized the HERV-loci regulated in these diseases and linked HERV expression to lower cancer-related survival. Also, while high expression of the ERV1 family in renal clear cell carcinoma was associated with decreased survival,[58] an elevated HERV expression has been correlated with better survival and immune infiltration of effector T cells in ovarian cancers.[59] Overall, this depicts a complex set of relationships between ERV expression and survival, depending on cancer type.

A possible effect of HERV on cancer progressions might be driven by the immunosuppressive functions of HERV-encoded envelope proteins. Such HERV elements were shown to contribute to feto-maternal tolerance.[60] It has been suggested that HERVH_19p13.12, also known as UCA1, is involved in certain functions of normal tissue, such as human trophoblast development.[61] Gene expression of this locus has also been reported in some cancers.[62,63] In our analyses, this locus was not included in our filters because its expression was just below our cutoff; however, its expression was

higher in colorectal tumours than in normal tissue and could be considered in an extended list. However, the HERVH locus at 21q22.3 was not differentially expressed between normal and tumoral in our analysis and PepQuery research did not identify suppressyn-derived peptide in MS/MS from TCGA's colorectal cancer. The prognostic impact of HERV expression and their interactions with the tumour microenvironment across tumour subtypes and disease stages remain poorly understood. Our findings suggest that HERV expression in lmCRC is associated with signalling pathways in favour of tumour growth and immuno-suppression. Consistent with these observations, Golkaram et al. identified a subset of primary CRC patients with high ERV expression and suppressed immune infiltrate who exhibited lower survival rates.[47] Conversely, some studies propose an immune activation role to HERVs. Indeed, HERV could modulate the immune microenvironment through viral mimicry and Toll-like receptors activation leading to danger signalling pathway activation.[23,24] Moreover, peptides derived from HERV can be recognized by T-cell receptors leading to specific immune response.[25] For instance, ERVH-2 (HERV-H loci on chromosome Xp22.3) has been shown to increase the lysis of CRC cell lines through CD8+ T-cell proliferation in a peptide-specific dependent manner[64]; the same results have been observed in melanoma for HERV-K-MEL derived peptides.[65] In the CT26 colorectal murine tumour model, the combination of a vaccine producing the ERV envelope peptide, and an anti-PD1 checkpoint inhibitor achieved excellent curative efficacy.[66] Here we observed that HERV associated with poorer survival correlates with lower T cell infiltration, while HERV associated with better survival presented no significant correlation with infiltration. This suggests the existence of a third factor influencing survival models, which may be independent of or complementary to the immune system.

Therapeutic vaccination strategies using HERVH-derived epitopes have been proposed for several cancers.[25] Other strategies, including the use of epigenetic modulators to increase the burden of repeat elements to trigger an interferon-gamma response in conjunction with immunotherapy, have shown success in melanoma.[67] Altogether, HERV-derived peptides are promising candidates for tumour-targeted immunotherapy. Our study provides a list of HERV sequences that might be used for further identification of antigen candidates in colorectal carcinoma. Our team is currently investigating the presence of a peripheral and intra-tumoral repertoire of HERV-restricted lymphocytes.

Co-expression of HERVH_8q21.3b and CALB1 has been described in lung squamous cell carcinoma (LUSC).[55] However, while CALB1 upregulation was a prognostic factor in lung carcinoma, we did not observe any impact of its expression in lmCRC clinical outcomes. Attig J. et al.[55] identified a chimeric transcript linking HERVH and CALB1 that is expressed early in tumour carcinogenesis. It was associated with cancer cell growth and reduced pro-tumoral inflammation despite its association with better overall survival in LUSC. CALB1 has also been described to be associated with tumour growth in ovarian cancer cells.[68] In these experiments, CALB1 was shown to bind MDM2 and promote p53-MDM2 interactions. As reported in lung cancer we observed a better survival in mCRC displaying an enhanced HERVH_8q21.3b transcription. HERVH-CALB1 fusion transcripts are expressed in pluripotent epiblasts.[16] Their expression is driven by transcription factor KLF5 instead of KLF4 which usually drives LTR7-HERVH transcription. Interestingly, focal amplification of KLF5 seems to be a genetic mark of CRC.[1] In colorectal cancer, CALB1 has been proposed in molecular signatures as a poor prognostic marker.[69–71] All of these observations highlight a potential interest in CALB1-HERVH regulation in CRC and support further investigation to better characterize the molecular bases of HERVH_8q21.3b and CALB1 interplay.

In conclusion, HERV expression is a biological hallmark of liver metastatic CRC. We demonstrated the prognostic value of some HERV biomarkers in patients who underwent liver metastasis resection. Our study provides significant insights into the interplay between HERVs, gene expression, and the immune context in CRC. The integration of large RNA-seq datasets, including both primary and metastatic CRC samples, allowed for a more comprehensive understanding of HERV expression patterns and their clinical implications.

Contributors

Daniel Gautheret and Christophe Borg planned and supervised the work, Pierre Laurent-Puig, Aurélien de Reyniès, Fabrice André, Anthony Turpin, Olivier Adotevi, Nicolas Gilbert, Anthony Boureux, Thierry André, and Syrine Abdeljaoued supply computational materials for the study, Franck Monnien, Laurent Arnould, Laura Guyard, and Magali Svrcek, supply biological materials for the study, Julien Viot and Daniel Gautheret accessed and verified the underlying data, processed the experimental data, performed the analysis, and designed the figures.

Dewi Vernerey assisted with survival data analysis, Zohair Selmani and Angélique Vienot assisted with Bioinformatics analysis.

Nawfel Adib and Romain Loyon performed biological experiments, Julien Viot wrote the manuscript with support from Daniel Gautheret, Christophe Borg and Romain Loyon. Julien Viot, Daniel Gautheret, Christophe Borg and Romain Loyon, were responsible for the decision to submit the manuscript.

All authors discussed the results, read the manuscript and approved the final version of the manuscript.

Data sharing statement

All data collected for this study are presented within this manuscript and are available upon request from the corresponding authors. After publication, access to sequencing and clinical data may be granted following the approval of a proposal and the signing of a data access agreement.

New experimental data generated in this study from the BIOMIROX dataset (association of MIROX and EPITOPE CRC01 NCT02838381) have been deposited to the European Genome-phenome Archive under the accession number EGAD50000000443.

Previously published Bulk RNA-seq can be retrieved from their original depository at SRP029880 (GSE50760),[30] SRP060016,[31] SRP095672 (GSE92914),[32] SRP245232 (GSE144259),[33] TCGA COAD,[1] UHB_cohort (GSE207194),[34] METAPRISM (EGAD00001009684).[35]

Single-cell RNA-seq used in this study came from GSE178318.[41]

Codes can be found at https://gitlab.com/jviot/liver_metastatic_crc_te_herv.

All programs and versions used are summarised at the end of the notebooks.

Declaration of interests

Pierre Laurent-Puig is chairman of Ile-De-France Canceropole and declare stock option in MethysDx, and Consulting fees from Pierre Fabre, Servier, Blocartis and BMS.

Aurélien De Reynies declare consulting fees from Qlucore as Member of the SAB.

Thierry André reports attending advisory board meetings and receiving consulting fees from, Aptitude health, Bristol Myers Squibb, Gritstone Oncology, Gilead, GlaxoSmithKline, Merck & Co. Inc., Nimbus, Nordic, Seagen, Servier, Pfizer and Takeda. Reports honoraria for lectures, presentations, speakers bureaus, manuscript writing or educational events from Bristol Myers Squibb, Merck & Co. Inc; Merck Serono, Seagen, and Servier. Support for attending meetings and/or travel from Bristol Myers Squibb and, Merck & Co. Inc and Takeda. Participation on a Data Safety Monitoring Board or Advisory Board for Inspirna. President of ARCAD Foundation.

Dewi Vernerey reports consulting fees from OSE Immunotherapeutics, Janssen-Cilag, HalioDx, Pfizer, cellprothera, GERCOR, INCYTE, FSK, INVECTYS, AC Biotech, Veracyte, CURE51, Apmonia Therapeutics.

Christophe Borg reports Grants from Bayer, Boehringer, Roche, Molecular partner, Payement for expert testimony from Molecular partner, support for attending meeting from Takeda and MSD, Participation on a Data Safety Monitoring Board from Sanofi.

Other authors declare no competing interest related to this study.

Acknowledgements

We thank the "CRB Ferdinand Cabanne – Dijon", the GERCOR, Gustave Roussy Institute, and Tumorothèque Franche-Comte for providing biological materials for this study.

This work was supported by funding from institutional grants from Inserm, EFS, University of Bourgogne Franche-Comté, national found "Agence Nationale de la Recherche – ANR-JCJC: Projet HERIC and ANR-22-CE45-0007", and "La ligue contre le cancer".

We thank the "Mésocentre de Franche-Comté" for providing the computational resources for this study.

During the preparation of this work the author(s) used ChatGPT from OpenAI and the Chat from Mistral AI in order to improve readability and language of the work. The author(s) have reviewed and confirmed the validity of the text and take(s) full responsibility for the content of the publication.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2025.105727.

References

1 Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330–337.
2 Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J. Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat Rev Cancer*. 2017;17(2):79–92.
3 Adam R, De Gramont A, Figueras J, et al. The oncosurgery approach to managing liver metastases from colorectal cancer: a multidisciplinary international consensus. *Oncologist*. 2012;17(10):1225–1239.
4 Nordlinger B, Sorbye H, Glimelius B, et al. Perioperative FOLFOX4 chemotherapy and surgery versus surgery alone for resectable liver metastases from colorectal cancer (EORTC 40983): long-term results of a randomised, controlled, phase 3 trial. *Lancet Oncol*. 2013;14(12):1208–1215.
5 Väyrynen V, Wirta EV, Seppälä T, et al. Incidence and management of patients with colorectal cancer and synchronous and metachronous colorectal metastases: a population-based study. *BJS Open*. 2020;4(4):685–692.
6 Chakedis J, Schmidt CR. Surgical treatment of metastatic colorectal cancer. *Surg Oncol Clin N Am*. 2018;27(2):377–399.
7 Ciardiello F, Ciardiello D, Martini G, Napolitano S, Tabernero J, Cervantes A. Clinical management of metastatic colorectal cancer in the era of precision medicine. *CA Cancer J Clin*. 2022;72(4):372–401.
8 Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Microbiol*. 2019;17(6):355–370.
9 Geis FK, Goff SP. Silencing and transcriptional regulation of endogenous retroviruses: an Overview. *Viruses*. 2020;12(8):E884.
10 Kong Y, Rose CM, Cass AA, et al. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat Commun*. 2019;10:5228.
11 Zhao S, Lu J, Pan B, et al. TNRC18 engages H3K9me3 to mediate silencing of endogenous retrotransposons. *Nature*. 2023;623(7987):633–642.
12 Jansz N, Faulkner GJ. Endogenous retroviruses in the origins and treatment of cancer. *Genome Biol*. 2021;22(1):147.
13 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;11(3):936.
14 Zhang M, Liang JQ, Zheng S. Expressional activation and functional roles of human endogenous retroviruses in cancers. *Rev Med Virol*. 2019;29(2):e2025.
15 Rivas SR, Valdez MJM, Govindarajan V, et al. The role of HERV-K in cancer stemness. *Viruses*. 2022;14(9):2019.
16 Singh M, Kondraskhina AM, Hurst LD, Izsvák Z. Staring at the onco-exaptation: the two-faced medley of an ancient retrovirus, HERVH. *J Clin Invest*. 2023;133(14):e172278.
17 Santoni FA, Guerra J, Luban J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*. 2012;9:111.
18 Ohnuki M, Tanabe K, Sutou K, et al. Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc Natl Acad Sci USA*. 2014;111(34):12426–12431.
19 Wang J, Xie G, Singh M, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*. 2014;516(7531):405–409.
20 Steiner MC, Marston JL, Iñiguez LP, et al. Locus-specific characterization of human endogenous retrovirus expression in prostate, breast, and colon cancers. *Cancer Res*. 2021;81(13):3449–3460.
21 Kang Q, Guo X, Li T, et al. Identification of differentially expressed HERV-K(HML-2) loci in colorectal cancer. *Front Microbiol*. 2023;14:1192900.
22 Bruni D, Angell HK, Galon J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat Rev Cancer*. 2020;20(11):662–680.
23 Roulois D, Loo Yau H, Singhania R, et al. DNA-demethylating agents target colorectal cancer cells by inducing viral mimicry by endogenous transcripts. *Cell*. 2015;162(5):961–973.
24 Chiappinelli KB, Strissel PL, Desrichard A, et al. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell*. 2015;162(5):974–986.
25 Attermann AS, Bjerregaard AM, Saini SK, Grønbæk K, Hadrup SR. Human endogenous retroviruses and their implication for immunotherapeutics of cancer. *Ann Oncol*. 2018;29(11):2183–2191.
26 Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*. 2016;351(6277):1083–1087.
27 Topham JT, Titmuss E, Pleasance ED, et al. Endogenous retrovirus transcript levels are associated with immunogenic signatures in multiple metastatic cancer types. *Mol Cancer Ther*. 2020;19(9):1889–1897.
28 Solovyov A, Vabret N, Arora KS, et al. Global cancer transcriptome quantifies repeat element polarization between immunotherapy responsive and T cell suppressive classes. *Cell Rep*. 2018;23(2):512–521.
29 Panda A, de Cubas AA, Stein M, et al. Endogenous retrovirus expression is associated with response to immune checkpoint

blockade in clear cell renal cell carcinoma. *JCI Insight*. 2018;3(16):e121522.
30 Kim SK, Kim SY, Kim JH, et al. A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol Oncol*. 2014;8(8):1653–1666.
31 Lee JR, Kwon CH, Choi Y, et al. Transcriptome analysis of paired primary colorectal carcinoma and liver metastases reveals fusion transcripts and similar gene expression profiles in primary carcinoma and liver metastases. *BMC Cancer*. 2016;16:539.
32 Ma YS, Huang T, Zhong XM, et al. Proteogenomic characterization and comprehensive integrative genomic analysis of human colorectal cancer liver metastasis. *Mol Cancer*. 2018;17(1):139.
33 Ji Q, Zhou L, Sui H, et al. Primary tumors release ITGBL1-rich extracellular vesicles to promote distal metastatic tumor growth through fibroblast-niche formation. *Nat Commun*. 2020;11(1):1211.
34 Viot J, Abdeljaoued S, Vienot A, et al. CD8+ CD226high T cells in liver metastases dictate the prognosis of colorectal cancer patients treated with chemotherapy and radical surgery. *Cell Mol Immunol*. 2023;20(4):365–378.
35 Pradat Y, Viot J, Yurchenko AA, et al. Integrative pan-cancer genomic and transcriptomic analyses of refractory metastatic cancer. *Cancer Discov*. 2023;13(5):1116–1143.
36 Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417–419.
37 Bendall ML, de Mulder M, Iñiguez LP, et al. Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput Biol*. 2019;15(9):e1006453.
38 GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat Genet*. 2013;45(6):580–585.
39 Marchet C, Iqbal Z, Gautheret D, Salson M, Chikhi R. REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. *Bioinformatics*. 2020;36(Suppl_1):i177–i185.
40 Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525–527.
41 Che LH, Liu JW, Huo JP, et al. A single-cell atlas of liver metastases of colorectal cancer reveals reprogramming of the tumor microenvironment in response to preoperative chemotherapy. *Cell Discov*. 2021;7(1):80.
42 Wen B, Wang X, Zhang B. PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Res*. 2019;29(3):485–493.
43 Becht E, Giraldo NA, Lacroix L, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol*. 2016 20;17(1):218.
44 Zhu X, Fang H, Gladysz K, Barbour JA, Wong JWH. Overexpression of transposable elements is associated with immune evasion and poor outcome in colorectal cancer. *Eur J Cancer*. 2021;157:94–107.
45 Yu C, Lei X, Chen F, et al. ARID1A loss derepresses a group of human endogenous retrovirus-H loci to modulate BRD4-dependent transcription. *Nat Commun*. 2022;13(1):3501.
46 Dolci M, Favero C, Toumi W, et al. Human endogenous retroviruses long terminal repeat methylation, transcription, and protein expression in human colon cancer. *Front Oncol*. 2020;10:569015.
47 Golkaram M, Salmans ML, Kaplan S, et al. HERVs establish a distinct molecular subtype in stage II/III colorectal cancer with poor outcome. *NPJ Genom Med*. 2021;6(1):13.
48 Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21(1):13–15.
49 Kolbe AR, Bendall ML, Pearson AT, et al. Human endogenous retrovirus expression is associated with Head and Neck cancer and differential survival. *Viruses*. 2020;12(9):956.
50 Kazachenka A, Loong JH, Attig J, et al. The transcriptional landscape of endogenous retroelements delineates esophageal adenocarcinoma subtypes. *NAR Cancer*. 2023;5(3):zcad040.
51 Gemmell P, Hein J, Katzourakis A. Phylogenetic analysis reveals that ERVs "die young" but HERV-H is unusually conserved. *PLoS Comput Biol*. 2016;12(6):e1004964.
52 Gemmell P, Hein J, Katzourakis A. The exaptation of HERV-H: evolutionary analyses reveal the genomic features of highly transcribed elements. *Front Immunol*. 2019;10:1339.
53 Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. *Mob DNA*. 2016;7:24.
54 Sun T, Xu Y, Xiang Y, Ou J, Soderblom EJ, Diao Y. Crosstalk between RNA m6A and DNA methylation regulates transposable element chromatin activation and cell fate in human pluripotent stem cells. *Nat Genet*. 2023;55(8):1324–1335.
55 Attig J, Pape J, Doglio L, et al. Human endogenous retrovirus onco-exaptation counters cancer cell senescence through calbindin. *J Clin Invest*. 2023;133(14):e164397.
56 Gibb EA, Warren RL, Wilson GW, et al. Activation of an endogenous retrovirus-associated long non-coding RNA in human adenocarcinoma. *Genome Med*. 2015;7(1):22.
57 Chang YS, Hsu MH, Chung CC, et al. Comprehensive analysis and drug modulation of human endogenous retrovirus in hepatocellular carcinomas. *Cancers (Basel)*. 2023;15(14):3664.
58 Zapatka M, Borozan I, Brewer DS, et al. The landscape of viral associations in human cancers. *Nat Genet*. 2020;52(3):320–330.
59 Natoli M, Gallon J, Lu H, et al. Transcriptional analysis of multiple ovarian cancer cohorts reveals prognostic and immunomodulatory consequences of ERV expression. *J Immunother Cancer*. 2021;9(1):e001519.
60 Kassiotis G, Stoye JP. Immune responses to endogenous retroelements: taking the bad with the good. *Nat Rev Immunol*. 2016;16(4):207–219.
61 Kong X, Li R, Chen M, et al. Endogenous retrovirus HERVH-derived lncRNA UCA1 controls human trophoblast development. *Proc Natl Acad Sci USA*. 2024;121(12):e2318176121.
62 Wang F, Li X, Xie X, Zhao L, Chen W. UCA1, a non-protein-coding RNA up-regulated in bladder carcinoma and embryo, influencing cell growth and promoting invasion. *FEBS Lett*. 2008;582(13):1919–1927.
63 Cao WJ, Wu HL, He BS, Zhang YS, Zhang ZY. Analysis of long non-coding RNA expression profiles in gastric cancer. *World J Gastroenterol*. 2013;19(23):3658–3664.
64 Mullins CS, Linnebacher M. Endogenous retrovirus sequences as a novel class of tumor-specific antigens: an example of HERV-H env encoding strong CTL epitopes. *Cancer Immunol Immunother*. 2012;61(7):1093–1100.
65 Schiavetti F, Thonnard J, Colau D, Boon T, Coulie PG. A human endogenous retroviral sequence encoding an antigen recognized on melanoma by cytolytic T lymphocytes. *Cancer Res*. 2002;62(19):5510–5516.
66 Daradoumis J, Ragonnaud E, Skandorff I, et al. An endogenous retrovirus vaccine encoding an envelope with a mutated immunosuppressive domain in combination with anti-PD1 treatment eradicates established tumours in mice. *Viruses*. 2023;15(4):926.
67 Noviello TMR, Di Giacomo AM, Caruso FP, et al. Guadecitabine plus ipilimumab in unresectable melanoma: five-year follow-up and integrated multi-omic analysis in the phase 1b NIBIT-M4 trial. *Nat Commun*. 2023;14(1):5914.
68 Cao LQ, Wang YN, Liang M, Pan MZ. CALB1 enhances the interaction between p53 and MDM2, and inhibits the senescence of ovarian cancer cells. *Mol Med Rep*. 2019;19(6):5097–5104.
69 Yang W, Lu S, Peng L, et al. Integrated analysis of necroptosis-related genes for evaluating immune infiltration and colon cancer prognosis. *Front Immunol*. 2022;13:1085038.
70 Man Y, Xin D, Ji Y, Liu Y, Kou L, Jiang L. Identification and validation of a novel six-gene signature based on mucinous adenocarcinoma-related gene molecular typing in colorectal cancer. *Discov Oncol*. 2024;15(1):63.
71 Xie Y, Li J, Tao Q, et al. Identification of glutamine metabolism-related gene signature to predict colorectal cancer prognosis. *J Cancer*. 2024;15(10):3199–3214.