NAR Genomics and Bioinformatics, 2026, **8**, lqag013
https://doi.org/10.1093/nargab/lqag013
Data Resources

OXFORD

A comprehensive annotation of conserved protein domains in human endogenous retroviruses

Tomàs Montserrat-Ayuso [1,2], **Aurora Pujol** [3,4,5], **Anna Esteve-Codina** [1,2,*]

[1]Centre Nacional d'Anàlisi Genòmica (CNAG), Baldiri Reixac 4, 08028 Barcelona, Spain
[2]Universitat de Barcelona (UB), Barcelona, Spain
[3]Neurometabolic Diseases Laboratory, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain
[4]Centre for Biomedical Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III, Madrid, Spain
[5]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain

*To whom correspondence should be addressed. Email: anna.esteve@cnag.eu

Abstract

Human endogenous retroviruses (HERVs) occupy nearly 8% of the human genome, yet their protein-coding potential remains largely unexplored. Originating from ancestral retroviruses that infected germline cells, HERVs typically follow the canonical proviral structure LTR–gag–pol–env–LTR, where gag, pol, and env encode structural, enzymatic, and envelope proteins. We present a comprehensive resource annotating conserved retroviral domains across 120 000 + ORFs derived from internal HERV regions. Using a reproducible pipeline based on HMMER and InterProScan, we identified over 17 000 domain hits—primarily from pol genes such as reverse transcriptase, RNase H, and protease—and quantified their structural conservation. Hundreds of domains exceed 95% alignment coverage, revealing a surprising abundance of full-length retrovirus-like domains in both young and ancient families. The HERVK (HML-2) subfamily retains the most complete polyprotein architecture, including 13 loci with nearly intact Gag, Pol, and Env, but full-length Pol domains are also found in HERVH, HERVW, and HERVE. Our annotations recover conserved catalytic motifs in Pol and transmembrane features in Env, enabling fine-grained functional interpretation. All results—including BED, FASTA, domain sequences, InterProScan outputs, and transmembrane predictions—are provided as an open resource at Zenodo to support downstream analyses of HERV protein expression, immune modulation, and co-option in health and disease.

Introduction

Endogenous retroviruses (ERVs) are remnants of ancient retroviral infections that became integrated into the germline and are now inherited as part of the host genome [1, 2]. In humans, ERV [human endogenous retroviruse (HERV)] sequences account for ∼8% of the genome and are typically found as fragmented long terminal repeats (LTRs), solo LTRs, and degraded internal regions [1, 3, 4]. Over the course of evolution, most HERVs have been progressively inactivated by accumulated mutations, deletions, and recombination events, resulting in a loss of their functional capacity [1, 5, 6]. Nonetheless, some HERV insertions have retained identifiable sequence features, such as intact open reading frames (ORFs) or conserved *cis*-regulatory motifs [6, 7]. In several cases, these elements have been co-opted by the host genome to contribute to host physiological processes [8]. A well-known example is the syncytin gene family, derived from retroviral envelope (*env*) genes, which is essential for trophoblast fusion and placental morphogenesis in eutherian mammals [9, 10]. These findings suggest that, despite widespread degradation, HERVs may still harbor biologically meaningful sequences, including those encoding functional protein domains.

The canonical retroviral genome encodes three major gene classes: *gag*, which produces capsid and matrix structural proteins; *pol*, encoding key enzymatic machinery including protease, reverse transcriptase (RT), RNase H, and integrase; and *env*, which mediates viral entry through membrane fusion [1, 11]. Although these genes are often fragmented or lost in endogenous retroviruses, several HERV loci retain partial or complete ORFs corresponding to these gene classes [5, 12–14]. The biological relevance of these retained sequences is underscored by examples of co-option, such as the syncytins mentioned above, but also by growing evidence of HERV-derived proteins involved in immune modulation, neurodevelopment, and tumorigenesis [8, 15–19]. Although most HERVs are considered to be usually dormant, they can be reactivated by stimuli such as viral infection, including influenza, HIV, or herpesviruses [15, 20], underscoring their dynamic interplay with the host immune system. This reinforces the importance of identifying conserved HERV ORFs and their associated retroviral domains as potential contributors to cellular processes or pathology.

While much recent research has focused on the transcriptional activity and regulatory roles of HERV-derived sequences—such as their enrichment in enhancer elements, influence on chromatin states, or contribution to long non-coding RNAs [7, 21]—less attention has been paid to their protein-coding capacity, particularly at the domain level. Furthermore, the relevance of protein-domain conservation extends beyond HERVs: recent work has shown that even non-retroviral endogenous viral elements (EVEs) may retain protein domains with possible antiviral functions, underscoring

Received: October 9, 2025. **Revised:** December 2, 2025. **Accepted:** January 20, 2026

the broader relevance of domain-level annotation in paleovirology and host-virus coevolution [22]. Prior studies often focus on specific HERV families or subfamilies (e.g. HERVK or HERVW), or on single co-opted genes, rather than conducting genome-wide surveys of protein-coding domain conservation [8, 9, 23–31]. A representative example is the comprehensive subfamily-specific analysis of HERVW/HERV17 by Grandi *et al.* [28], who generated an exhaustive catalogue of HERVW insertions in the human genome (hg19), defined subgroup-specific phylogenies and LTR-based age estimates, and performed a detailed structural and motif-level characterization of HERVW *env* genes—including comparisons to Syncytin-1. Similarly, Grandi *et al.* [30] delineated the HERVK (HML-10) subfamily, showing that this ancient HERV-K lineage is characterized by recurrent degradation of pol and the selective retention of a small number of Env ORFs.

In parallel, Blomberg and colleagues have provided prototypic structural descriptions for several other lineages, including ERV3/HERV-R and HERVK (HML-6), reconstructing near full-length proviral genomes and documenting lineage-specific patterns of *gag–pol–env* decay [29, 31]. Comprehensive reviews by Grandi and Tramontano have further emphasized that HERV envelope proteins, in particular, can retain fusogenic or immunomodulatory properties and contribute to both physiology and pathology [32]. Collectively, these family-focused studies have established that individual HERV lineages show characteristic "signatures" of gene retention and loss, but by design they do not provide a genome-wide, domain-level map that spans all HERV families.

One close approximation is the analysis by Nakagawa and Takahashi [13], which identified thousands of HERV-derived ORFs with partial or complete retroviral domains using their gEVE database. This pioneering resource was the first large-scale catalogue of viral ORFs across mammalian genomes and remains valuable for exploring EVEs coding potential. The gEVE pipeline, developed in 2016, integrates RepeatMasker [33], RetroTector [34], and HMMER-based motif searches [35], together with extensive BLAT-based refinement, providing a broad inventory of EVE ORFs and associated viral motifs. While highly comprehensive, gEVE was not designed to quantify Hidden Markov Model (HMM)-profile coverage or to resolve subfamily-specific domain retention patterns, which limits its ability to evaluate the degree of structural preservation of retroviral proteins at high resolution. More broadly, there is still no widely adopted community-standard tool for HERV discovery, and most available methods lack a fully shareable and reproducible pipeline [36, 37]. Building upon these foundations, our study provides a reproducible, domain-centric framework that adds quantitative coverage metrics at single-locus resolution, enabling finer-scale interrogation of protein-coding potential and structure across the HERV landscape.

Several resources have contributed to improving the classification, localization, and annotation of HERVs in the genome. The HERVd database was among the first systematic catalogs of HERV families and chromosomal coordinates, based on RepeatMasker [33] and profile HMM-based annotation [38]. The abovementioned gEVE database expanded this framework to multiple vertebrate genomes, integrating domain-level information with genomic context to support comparative analysis [13]. More recently, the Telescope pipeline introduced a high-resolution approach to locus-specific HERV expression quantification in RNA-seq data, including the development of a curated database of HERV loci and family-level annotations that has become foundational for expression-based studies [39]. Additionally, the ERVmap resource provided a locus-specific catalog of over 3000 near full-length HERVs and a dedicated RNA-seq quantification pipeline, enabling high-resolution analysis of HERV expression across cell types and disease states, including autoimmune and cancer contexts [40]. These resources have significantly improved our understanding of HERV transcription and classification, but none have systematically addressed domain-level structural conservation in predicted ORFs.

In this study, we present a genome-wide analysis of the structural conservation of retroviral protein domains within ORFs derived from internal HERV sequences. Throughout this article, "structural conservation" refers to preservation of protein-domain architecture—evaluated as the extent to which ERV-derived sequences align to HMM protein models—rather than conservation of proviral gene structure. Despite extensive mutational decay, we identified hundreds of HERV loci that retain high-confidence domains from the *gag*, *pol*, and *env* gene classes—many with complete or near-complete coverage of HMM profiles. While most studies examine HERVs at the subfamily level, this approach may obscure functional heterogeneity, as only a subset of loci within a given subfamily retain structurally intact domains. By quantifying domain-level conservation at single-locus resolution, we define a set of structurally preserved retroviral proteins that may represent candidates for residual biological activity or host co-option, providing a foundational resource for future functional investigations.

Our analysis builds upon previous resources such as gEVE and ERVmap [13, 40], shifting the focus from presence/absence annotation to domain-centric conservation with quantitative resolution. To our knowledge, this is the first genome-wide resource to systematically annotate ERV ORFs at the domain level, integrating alignment coverage, conserved motifs, and structural features into a shareable and reproducible dataset.

Materials and methods

Repeat Masker annotation

We annotated HERV insertions across the human genome using RepeatMasker [33] version 4.1.8, employing NCBI/RMBLAST (v2.14.1+) as the search engine. The analysis was conducted on the GRCh38/hg38 reference genome (primary assembly) using the Dfam 3.9 repeat library [41] (root 'dfam39_full.0.h5' and Mammalia 'dfam39_full.7.h5' partitions). RepeatMasker was run with -species human, -s and -no_is options.

Extraction of internal HERV coordinates

To identify internal regions of HERVs, we parsed RepeatMasker .out files for entries labeled with "-int", "_I", "-I" or "-I_MM" in their name and classified as LTR/ERV. These entries were converted into six-column BED format (chromosome, start, end, name, score, strand), where the name included chromosomal position, strand, and divergence from consensus.

Merging of overlapping internal regions

To reduce redundancy and better reflect transcript-like HERV structures, we merged overlapping or closely spaced internal HERV regions (≤200 bp gap) sharing the same subfamily and strand. Regions were grouped by subfamily and strand and collapsed into unified intervals with updated standardized BED names. We selected the 200 bp threshold after noticing that shorter gaps (≤150 bp, as used in our preprint) occasionally fragmented bona fide internal regions. For comparison, we also tested a *continuity rule* as implemented in Telescope [39] which merges fragments from the same subfamily and strand when they align consistently with the model (up to 2.5 kb). Domain counts and conservation rates were highly consistent across all thresholds (150 bp, 200 bp, 2.5 kb; see Supplementary Table S5), supporting the robustness of our conclusions. For clarity and reproducibility, we report results based on the simpler 200 bp rule throughout the manuscript.

Sequence extraction of merged HERVs

We extracted strand-specific DNA sequences of the merged HERV internal regions. Sequences were obtained from the indexed reference genome FASTA (GRCh38) and reverse-complemented if on the negative strand.

ORF prediction

ORFs were predicted from the merged internal HERV sequences using the EMBOSS getorf tool (v6.6.0) [42]. Sequences were provided in FASTA format, and ORFs shorter than 180 nucleotides were discarded using the -minsize 180 parameter. To maximize sensitivity, ORFs were defined from stop codon to stop codon, without requiring a canonical start codon, by using the -find 0 and -methionine N options. This approach, described in Villesen *et al.* (2004) [5] accommodates the nonconventional translation mechanisms commonly used by retroviruses—such as ribosomal frameshifting and termination suppression—particularly at internal *pol* gene. Also, ORF sequences lacking a canonical ATG start codon may still serve as coding exons in spliced retroviral transcripts [13]. ORFs were extracted in the forward strand only (-reverse N), since strand orientation was previously accounted for during sequence preparation.

Protein domain detection

HMM profiles for retroviral protein domains were obtained from the GyDB profile collection and downloaded from the GyDB [43, 44] repository (see the 'Data availability' section for the exact URL). All individual HMM files were concatenated into a single database, and the resulting file was indexed with hmmpress from the HMMER (v3.4) package [35] to generate the binary files required by hmmscan (see Supplemental Methods for exact commands). Protein domains were identified by running hmmscan on the ORFs using this custom GyDB-derived HMM database. Functional class assignment followed a simple rule-based scheme: enzymatic domains (RT, RNaseH, integrase, protease, dUTPase) were classified as Pol; envelope-associated profiles as Env; and capsid/matrix profiles as Gag. Domains annotated as accessory proteins in GyDB were designated as Accessory. Accessory proteins are additional genes beyond the canonical *gag*, *pol*, and *env*. They are nonessential genes often involved in immune evasion, replication efficiency, or modulation of host functions [11]. Profiles not associated with HERVs—such as aphid transmission factors, movement proteins, transactivator/viroplasmin proteins, virion-associated proteins, and chromodomains—were labeled as Other. The full correspondence between HMM model names and functional classes used in this study is available in the Zenodo repository.

Results were parsed and filtered to retain only hits satisfying the following criteria: full-sequence E-value ≤ 1e-5, domain E-value ≤ 1e-5. To avoid redundancy and focus on the most relevant domain instance per locus, we further filtered the results to retain only the best-scoring match per domain type (e.g. Gag, RT, RNaseH, Env) within each ORF.

We calculated domain coverage as the fraction of HMM profile positions aligned by the hit, computed as (hmm_to − hmm_from + 1)/hmm_length. Coverage describes the fraction of the HMM profile aligned to the ORF. According to the GyDB articles [43, 44], HMM profiles were constructed from lineage-specific multiple alignments derived from monophyletic clusters of Ty3/Gypsy and Retroviridae LTR retroelements. Thus, in our study "coverage" reflects the extent to which a sequence within an ORF matches the HMM model for a given retroviral domain, rather than the full length of the corresponding protein domain. Because all loci were evaluated against the same curated HMMs, comparisons of coverage values across domains and subfamilies remain internally consistent, even though the HMMs themselves may not span complete protein-coding regions.

Mapping protein domains back to genomic coordinates

Filtered domain hits were mapped back to their corresponding genomic positions based on ORF-to-internal-sequence offsets and internal-to-genome coordinates. This conversion took into account strand orientation and translated amino acid coordinates into nucleotide positions. The mapping step produced BED-format annotations including the domain type, gene class, alignment score, and domain conservation level.

Extraction of domain sequences

Protein domain coordinates from filtered HMMER output were used to extract precise amino acid subsequences corresponding to each domain. Domain sequences were retrieved from the original ORF FASTA based on alignment start and end positions, generating a FASTA file of domain-specific sequences for alignment and downstream analyses.

Domain and motif annotation with InterProScan and Phobius

To complement HMMER-based domain detection, we further annotated the HMMER-identified domain sequences for conserved motifs and additional structural features using InterProScan [45, 46] (v5.75). All retroviral domain calls originate strictly from HMMER searches using GyDB HMM profiles. InterProScan was applied only to the amino acid sequences of those HMMER-identified domains to annotate motifs and structural features; it was not used to detect additional domains or to classify ORFs lacking HMMER hits.

Protein domain sequences were analyzed using the InterProScan standalone tool with default parameters, enabling all member databases. The XML output was parsed to extract conserved sites predicted by the Conserved Domain Database [47] (CDD), particularly focusing on catalytic residues and

functional motifs. A custom Python script was used to extract and summarize the domain descriptions and conserved residues per protein. During testing, we noted that InterProScan occasionally omitted CDD-annotated active sites when sequences were submitted in multi-FASTA batches, even though the underlying domain hits and e-values were identical to those obtained when processing single sequences. To ensure that all catalytic residues of high-coverage RNase H domains (HMMER coverage > 95%) were correctly captured for the analyses reported in the Results, these sequences were re-analysed using InterProScan in single-sequence mode. This approach consistently returned the full set of CDD conserved-site annotations.

To investigate membrane-associated features of Env proteins, we filtered the predicted ORFs to retain only sequences belonging to *env* genes. These sequences were then analyzed using Phobius [48] with default settings to predict transmembrane domains and signal peptides. The output was parsed to identify Env proteins with predicted transmembrane regions, which may indicate preserved envelope protein topology.

The flow chart of the complete pipeline is shown in Fig. 1, and it is available as a series of standalone scripts (see below and "Code availability"). All domain annotations, including BED files, FASTA sequences, and InterProScan outputs, were generated using this pipeline. These datasets are available at Zenodo (repository: "Domain-Level Annotations and Conservation Scores for Human Endogenous Retroviruses", DOI: https://doi.org/10.5281/zenodo.17662456), and the Python scripts used to generate them are available for reproducibility at Zenodo: https://doi.org/10.5281/zenodo.18326381.

Application to T2T-CHM13 and Mouse Genomes

The same workflow described above was applied without modification to the T2T-CHM13 human assembly [49] and the GRCm39 mouse reference genome. RepeatMasker annotations were generated using species-appropriate parameters and Dfam libraries. For the T2T-CHM13 genome, we additionally used the -cutoff 255 parameter to reproduce the behavior of the NCBI RepeatMasker output available for this assembly. All subsequent steps (internal region extraction, merging, ORF prediction, HMMER-based domain annotation, motif analysis, and BED/FASTA output generation) were performed identically to the GRCh38 analysis. The resulting annotations for both assemblies are available at Zenodo (see the 'Data availability' section).

Identification and analysis of conserved protein domains in HERV ORFs

Amino acid sequences were loaded in R using the Biostrings package and characterized by length and chromosomal distribution. Domains were classified as conserved if coverage was ≥40%, and as highly conserved if coverage exceeded 95%. ORFs were grouped by domain class, and the number of ORFs harboring conserved domains was summarized. Domain coverage distributions were visualized using violin and beeswarm plots, highlighting thresholds used for conservation categorization (40% and 95%).

Subfamily-level analyses focused on four major HERV subfamilies (HERVK/HML-2, HERVH, HERV17, and HERVE). Within each subfamily, the coverage scores of eight key domain types (Gag, RT, RNaseH, DUT, INT, AP, Env, and accessory) were assessed. For each ORF and domain class, only the best-scoring domain match was retained.

To evaluate co-occurrence of domain types within individual loci, we grouped ORFs by locus and determined the presence of Gag, Pol, and Env domains. Loci containing all three gene classes (i.e. "trios") with conserved domains were flagged as potentially retaining multi-domain retroviral architecture. A stricter criterion requiring > 80% coverage for each domain was also applied to define high-confidence candidate loci. For these loci, domain-specific coverage and subfamily information were compiled into summary tables for further inspection.

All statistical analyses and plots were performed in R (v4.3.3) using the dplyr [50], ggplot2 [51], ggbeeswarm, Biostrings [52], and openxlsx [53] packages.

Results

Genome-wide landscape of HERV ORFs and conserved retroviral domains

To explore the functional potential of endogenous retroviral sequences, we systematically annotated ORFs across the internal HERV sequences found by RepeatMasker [33] and assessed the presence of conserved retroviral domains (see full annotation at Zenodo; repository: "Domain-Level Annotations and Conservation Scores for Human Endogenous Retroviruses", DOI: https://doi.org/10.5281/zenodo.17662456). A total of 128 427 ORFs were identified and analyzed, with lengths ranging from 60 to 1415 nucleotides (median = 78 nt; mean = 92.22 nt). ORFs were distributed across all autosomes and sex chromosomes, with chromosomes 4 and X harboring the highest numbers (10 418 and 10 064 ORFs, respectively), while the fewest were observed on chromosome 22 (995 ORFs). When normalized by chromosome size, distinct patterns emerged: the Y chromosome showed the highest ORF density, with one ORF every ∼10.2 kb, followed by chromosomes 19 and X. In contrast, chromosomes 16, 20, and 22 had the lowest densities (one ORF every 46–52 kb). As previously reported [3, 54, 55], the accumulation of HERVs on the Y chromosome likely reflects reduced recombination, relaxed purifying selection, and a permissive environment for repetitive element retention. These observations point to a nonrandom genomic distribution of HERV-ORFs, potentially shaped by chromosome-specific differences in evolutionary constraint, integration bias, and retrotransposon dynamics.

Despite the short length of most ORFs, a total of 17 209 sequences still had identifiable retroviral-like domains. We defined conserved domains as those covering ≥40% of the corresponding HMM profile—a threshold selected to balance sensitivity and specificity, allowing detection of partially preserved domains while minimizing inclusion of highly degraded fragments likely to be biologically uninformative. Applying this criterion, we identified 6589 domains with moderate to high conservation, each spanning hallmark gene classes including *gag*, *pol*, *env*, and accessory (additional genes beyond the canonical *gag*, *pol*, and *env*). The majority of hits corresponded to Pol-related domains ($n = 5969$), which include key enzymatic components such as RT, integrase, RNase H, and aspartic protease (AP). Fewer ORFs carried Env domains ($n = 353$), followed by Gag ($n = 216$), and accessory proteins ($n = 51$; Fig. 2A). This distribution underscores the prevalence of conserved Pol-derived fragments in the genomic remnants of HERVs.

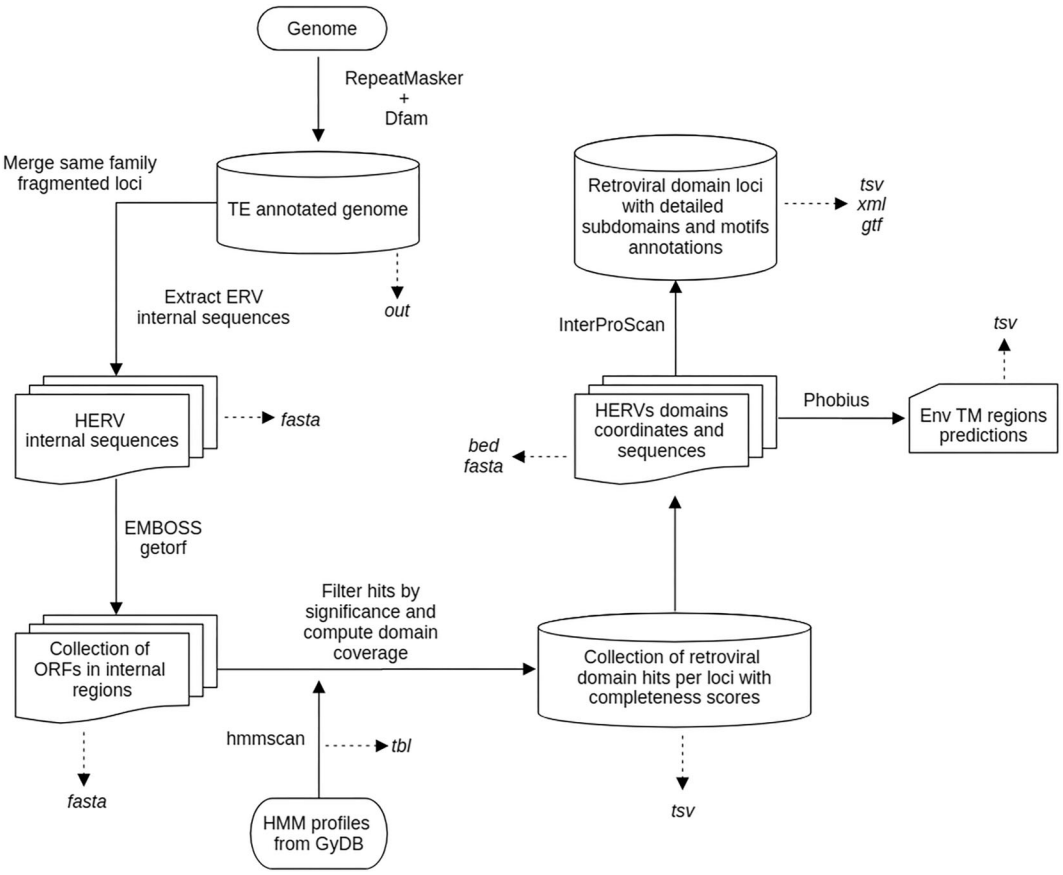

Figure 1. The annotation pipeline begins with the output file from RepeatMasker. While the main output is a TSV file containing domain coordinates and completeness scores, several intermediate and supporting files are also generated, including FASTA, TSV, BED, and TBL files, as well as InterProScan and Phobius output files providing additional information for each detected domain.

The presence of these domains across a diverse set of chromosomal locations suggests widespread distribution and long-term retention of retroviral elements with partially conserved coding capacity. These findings support the notion that while many HERVs are degenerated, a subset retains traces of protein-coding potential, particularly within the *pol* gene family.

Distribution and characteristics of conserved domains in HERV-derived ORFs

To further characterize the quality of the annotated HERV-derived domains, we evaluated their structural conservation by analyzing alignment coverage against reference HMM profile. As mentioned above, many domains exhibited high coverage values: many surpassed the 0.4 coverage threshold, with numerous domains reaching coverage values above 0.8 or even complete alignment (coverage = 1.0; Fig. 2B). These high-coverage matches suggest that several endogenous retroviral elements retain extensive homology to their exogenous retroviral counterparts.

Several representative examples illustrate the degree of conservation observed. One Env domain from a HERVK locus (chr5: 156 658 763–156 665 917) showed 99.5% coverage over 356 amino acids, with an exceptionally high HMMER score (829.4) and a highly significant e-value (4.9e–251), strongly supporting its structural integrity. To further evaluate its functional conservation, we inspected the results from InterProScan and Phobius. The analysis revealed a complete GP41-like region, including conserved HR1–HR2 heptad repeats, a predicted transmembrane helix, and cytoplasmic and noncytoplasmic domains (Supplementary Fig. S1). These structural features are essential for membrane fusion in retroviral envelope proteins, suggesting that this Env domain may retain key aspects of its ancestral function.

Another example comes from a Pol RNaseH domain encoded by a HERVH locus (chr14: 53 129 175–53 135 122), which showed full coverage over 148 amino acids, with a high HMMER score (153.7) and a significant e-value (6.6e–47). InterProScan annotation identified a complete RNaseH_HI_RT_Bel domain, matching multiple conserved RNase H family signatures, including RNASE_H_1 (PFAM), RNaseH_domain (PROSITE), and RNaseH_sf (CATH-Gene3D). The region exhibited consistent hits across multiple databases, including PFAM, SUPERFAMILY, and CDD, supporting its classification as a structurally conserved RNase H-like fold. The conserved DEDD residues—characteristic of active RNase H enzymes [56]—were also detected, reinforcing the potential enzymatic activity of these domains. (Supplementary Fig. S2). Although catalytic activity was not directly tested, the presence of conserved residues typically associated with RNaseH function supports potential enzymatic functionality, possibly contributing to nucleic acid metabolism or retroelement regulation.

As a final example, we highlight a Pol protease domain from a HERV-E locus (chr1: 20 154 322–20 160 102). Despite its shorter length (76 amino acids), this domain showed complete

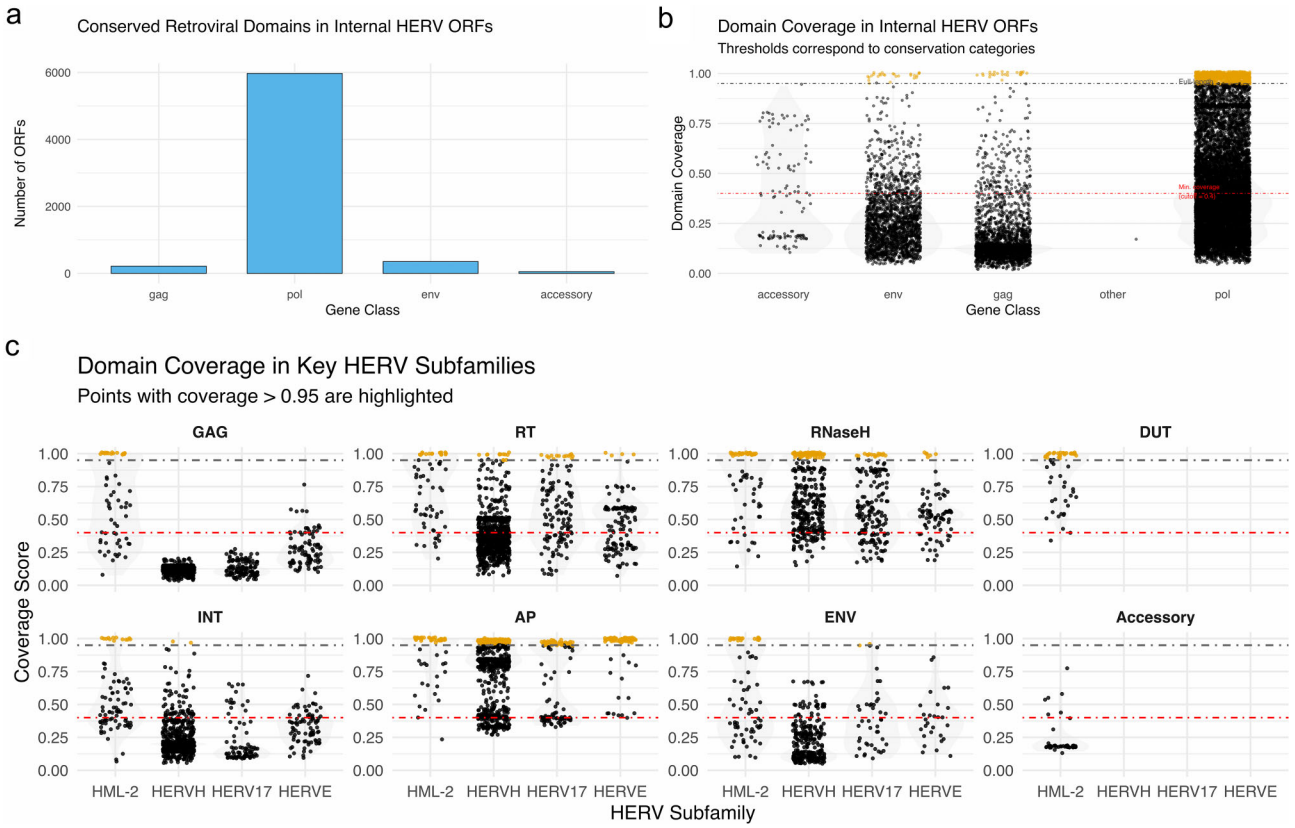

Figure 2. (**A**) Bar plot showing the number of annotated ORFs containing conserved domains (≥ 40% HMM profile coverage) associated with each major retroviral gene class (*gag, pol, env,* and accessory). (**B**) Violin plots showing the distribution of domain coverage values for annotated HERV ORFs grouped by gene class (accessory, *env, gag,* and *pol*). Each point represents an individual domain hit passing quality thresholds (e-value < 1e–5). The red dashed line marks the 40% minimum domain coverage. Higher density of points near full coverage (1.0) is observed especially in the *pol, gag,* and *env* classes, indicating a substantial number of highly conserved domains. Yellow points highlight full-length alignments (coverage > 0.95), reflecting potentially intact or structurally well-preserved retroviral sequences. (**C**) Data points represent both total identified domains (black) and highly conserved domains (coverage > 0.95; highlighted). HERVK/HML-2 and HERVH exhibit the most extensive domain retention, while other subfamilies show gene-specific or restricted conservation. HML-2 refers to the HERVK subfamily.

coverage of the HMM profile, with a strong HMMER score (67.5) and a significant e-value (1.4e–20). InterProScan identified a complete ASP_PROT_RETROV domain, supported by multiple overlapping annotations across PFAM (RVP – Retroviral Aspartyl Protease), PROSITE (Peptidase_A2_cat), SUPERFAMILY (Acid Proteases), and CATH-Gene3D (Peptidase_aspartic_dom_sf; Supplementary Fig. S3). These consistent annotations indicate that the domain adopts the conserved structure of pepsin-like retroviral aspartyl proteases, suggesting structural retention in this ancient HERV lineage. Interestingly, the annotated gene ENSG00000227066 in Gencode [57] (v48), described as an uncharacterized long noncoding RNA (lncRNA), neatly overlaps this HERV, specifically the RNase H (full-length), and integrase (degraded) domains.

Beyond individual examples, we performed a systematic analysis of 1015 full-length domains (HMMER coverage > 95%), encompassing RNaseH, protease, integrase, RT, dUTPase, Env, Gag, and accessory proteins. Among these, 359 domain instances contained conserved catalytic or structural residues as annotated by InterProScan when searching in the CDD. RNaseH domains accounted for the majority of these cases, frequently retaining residues essential for enzymatic activity (e.g. conserved D, E, D, D) and RNA/DNA hybrid binding sites (166 instances retaining at least the latter, 60 retaining both). Similarly, 79 dUTPase and 72 protease domains showed well-conserved active or catalytic site residues. For Env domains, five full-length instances—belonging to the HERVS71, HERVW, and HERVE-a subfamilies—contained sequences that matched known immunosuppressive regions in the CDD. Complementary analysis with Phobius further revealed that full-length ENV domains frequently encode C-terminal transmembrane helices, reinforcing their structural completeness and potential for functionality.

These results highlight that, while many HERV loci are fragmented and degraded, a notable fraction retain high levels of sequence conservation in specific protein domains. This supports the potential relevance of these elements in host biology and evolutionary dynamics.

Subfamily-level patterns of domain retention

To investigate the evolutionary preservation of endogenous retroviral proteins, we examined the distribution of conserved domains across HERV subfamilies (Supplementary Table S1 and Fig. 2C). While some subfamilies showed widespread retention of the canonical retroviral genes *gag*, *pol*, and *env*, others exhibited more restricted or partial conservation.

The HERVK family displayed the most diverse and abundant repertoire of conserved domains, with multiple clades

(e.g. HERVK/HML-2, HERVK9, HERVK11, HERVK14, HERVK22) retaining domains from all major gene classes: *gag*, *pol*, *env*, and in some cases, accessory genes. For example, HERVK9 alone contributed over 100 hits each for AP, RT, and RNase H domains. Even under stringent filtering (coverage >0.95, Supplementary Table S2), HERVK/HML-2 retained a notable number of full-length domains—including 33 AP, 28 RNaseH, 27 dUTPase (DUT), 17 RT, 13 integrase, 17 Gag, and 19 Env—indicating preservation of near-complete polyproteins in several loci.

In contrast, HERVH showed a clear Pol-centric conservation pattern. It contributed the highest number of AP (n = 700), RNase H (n = 309), and RT (n = 316) domains across all subfamilies (≥40% coverage), but lacked any conserved GAG domains and retained only 38 ENV hits. Under the high-confidence threshold, it preserved 185 full-length AP, 76 RNase H, 13 RT, and 2 integrase domains, highlighting a selective retention of enzymatic functions and suggesting pressure against structural gene conservation.

The HERVW lineage, primarily represented by HERV17 [3], exhibited moderate but consistent conservation of Pol, and Env domains. It retained dozens of hits across Pol domains, and under the >0.95 coverage threshold, 51 AP, 17 RNase H, 14 RT, and 2 ENV domains remained—supporting a degree of functional maintenance within this subfamily.

Several other lineages showed more specialized conservation patterns. HERVE and HERV9, for instance, contributed primarily Pol-derived domains (especially proteases), while Env-enriched conservation was mostly restricted to lineages like HERVIP10FH or Harlequin. Notably, HERVE—considered evolutionarily ancient—retained 69 full-length AP domains, 6 RNase H, and 4 RT domains, suggesting unexpected structural integrity.

Overall, the conservation of full-length domains in specific subfamilies, including both evolutionarily young (e.g. HERVK/HML-2) and ancient (e.g. HERVE) lineages, highlights the selective retention of replication-related proteins. This pattern of domain conservation raises the possibility that some HERVs may preserve structural or functional potential long after their integration.

Co-occurrence of conserved retroviral domains reveals residual proviral architecture

The presence of all three domain types—Gag, Pol, and Env—is essential for the formation of retroviral particles, as they respectively encode structural components, enzymatic machinery for reverse transcription and integration, and the envelope proteins required for cell entry.

To assess the potential for residual functionality among HERV loci, we examined the co-occurrence of conserved domains within each HERV insertion. We first identified loci that contain at least one Gag, one Pol, and one Env domain, without applying a stringent conservation threshold (only 40% of domain coverage). Such configurations are a prerequisite for viral-like particle formation. Under this relaxed criterion, 44 loci were found to carry a combination of structural and enzymatic domains, suggesting the preservation of core retroviral architecture at a broad scale (Supplementary Table S3).

To prioritize high-confidence candidates, we applied a stricter filter requiring all three domain types (Gag, Pol, and Env) to have at least one domain of each with a minimum alignment coverage of 0.8. This refinement yielded 13 loci, all belonging to the HERVK/HML-2 subfamily, that retained a full complement of core retroviral domains with high sequence conservation (Table 1 and Supplementary Table S4).

These 13 loci typically encode up to seven distinct domains, encompassing AP, dUTPase (DUT), RT, RNase H, integrase (INT), Env, and Gag. Most domains exhibited near-complete alignment (coverage > 0.95), with several loci achieving full-length coverage across all seven domains. All 13 loci retained both 5′ and 3′ LTRs, consistent with full-length proviral structure. For example, the insertion at chr1: 155 627 723–155 634 872 shows perfect conservation (coverage = 0.995–1.0) for every domain. Similar complete or near-complete configurations were observed on chromosomes 3, 5, 7, 8, 11, and 12 (Table 1 and Supplementary Table S3). Interestingly enough, several of these HERVK loci are located within introns of human genes (e.g. *SGCD*, *MEI4*, *DEFB107B*), often in the antisense orientation, suggesting possible regulatory interactions.

These loci represent the best candidates for retained protein-coding potential among endogenous retroviruses and capacity to form infectious viral-like particles. Their intact domain architecture may reflect recent integration, selective constraint, or co-option into host regulatory or structural functions. The functional and structural potential of these elements motivates further studies into their transcriptional activity, epigenetic regulation, and potential immunogenicity, and, notable, two loci (e.g. chr1:155 627 723–155 634 872) encode Env in the same ORF as several Pol domains, an unusual configuration that may reflect evolutionary fusion events.

Application of the annotation pipeline to additional genomes (T2T-CHM13 and mouse GRCm39)

To assess the generalizability of our workflow, we applied the full annotation pipeline to two additional reference genomes: the telomere-to-telomere human assembly T2T-CHM13 and the mouse GRCm39 assembly. In T2T-CHM13, we identified 17 946 retroviral-like domain hits, whereas in mouse GRCm39 the workflow recovered 47 840 internal domain hits.

A comprehensive comparison between T2T-CHM13 and GRCh38, as well as a full exploration of the mouse ERV landscape, lies beyond the scope of the present study and would merit a dedicated analysis. Nonetheless, these results demonstrate that the current implementation of the pipeline is readily applicable across distinct RepeatMasker annotations and ERV repertoires. While the workflow ran without modification for both human assemblies and for mouse, extending it to more distantly related genomes may require minor adjustments to capture species-specific nomenclature used by RepeatMasker for internal ERV regions (see the ‘Materials and methods’ section). The complete BED, FASTA, and domain-level annotation datasets for both assemblies are publicly available at Zenodo (see the ‘Data availability’ section).

Discussion

Our systematic analysis reveals that thousands of ORFs embedded within HERV loci encode recognizable retroviral protein domains, despite the extensive genomic erosion these elements typically undergo. Many of these ORFs exhibit high domain coverage and strong alignment scores, indicating the retention of structural features beyond what would be expected

Table 1. Top HERVK/HML-2 loci with co-occurring Gag, Pol, and Env domains (coverage >0.8 in at least one domain)

Locus	Subfamily	AP	DUT	ENV	GAG	INT	ORFX	RNaseH	RT	Domain count
chr11:101696031–101703471_+	HML-2	1	1	0.995	1	1	0.182	1	1	8
chr12:58328516–58335944_-	HML-2	0.989	1	0.995	1	1	0.182	1	0.85	8
chr1:155627723–155634872_-	HML-2	1	1	0.995	1	1	0.182	1	1	8
chr1:160691754–160698959_+	HML-2	0.989	1	0.995	0.8	0.597		1	1	7
chr1:75378054–75382401_+	HML-2		1	0.995	1	0.13				4
chr22:18939648–18946795_+	HML-2	0.822	1	0.995	1	1	0.182	1	0.59	8
chr3:185563605–185570759_-	HML-2	1	1	0.995	1	1	0.182	1	1	8
chr5:156658763–156665917_-	HML-2	1	1	0.995	1	1	0.182	1	1	8
chr5:30487615–30495147_-	HML-2	1	1	1	0.834	0.68	0.182	0.801	1	8
chr6:77717994–77725406_-	HML-2	1	1	0.995	1	0.763		1	0.547	7
chr7:4583483–4590929_-	HML-2	1	1	0.995	1	1	0.182	1	1	8
chr7:4591987–4599421_-	HML-2	1	0933	0.995	0.816	1	0.182	1	1	8
chr8:7498932–7506377_-	HML-2	1	1	0.995	1	1	0.182	1	1	8

Each locus encodes seven distinct retroviral domains with high alignment coverage, except one locus, encoding 3.

from random degradation. Notably, we observe unexpectedly high conservation within Pol-derived domains, including RT, RNase H, integrase, and AP. Several RNase H and protease domains align fully or nearly fully to their respective HMM profiles and retain catalytic motifs such as the DEDD residues in RNase H domains, a hallmark of enzymatic activity [56]. While prior studies have described isolated ORFs or co-opted genes for *env* or *gag* [5, 8, 12] our genome-wide approach extends this view by systematically quantifying domain-level conservation across all major HERV subfamilies, including Pol domains. These findings suggest that HERV-derived *pol* genes may represent a broader and more structurally intact source of retroviral legacy than previously recognized.

The conservation of protein domains is not uniformly distributed across HERV lineages but instead reflects subfamily-specific patterns. Among these, the youngest HERVK subfamily (HML-2), as expected [3], retains multiple loci with near-complete retroviral architecture—including co-occurring Gag, Pol, and Env domains—in configurations reminiscent of intact retroviruses. This is consistent with reports of active HERVK expression and translation in early embryonic tissues and cancer [21, 58–60]. In contrast, HERVH rarely retains intact Gag and Env ORFs, yet exhibits striking conservation of Pol-encoded enzymatic domains, especially RNase H and protease. This pattern aligns with previous findings indicating extensive recombination and truncation in HERVH loci, which have been maintained at high copy number due to regulatory or transcriptomic utility [7, 61]. Our results refine this view by showing that, despite structural erosion, a subset of HERVH elements preserves enzymatic domain integrity—suggesting selective maintenance of Pol-related functions (specially RNase H and AP activity).

Several family-specific studies provide important context for interpreting our domain-level annotations. For HERVW/HERV17, the detailed analysis by Grandi *et al.* [28] highlighted the characteristic retention of Env—including Syncytin-1—and documented recurrent structural deletions elsewhere in the provirus; our results are consistent with this model, as we also recover four Env domains with coverage >0.9, while further indicating that Pol-derived domains are more widely preserved in this lineage than previously appreciated. A similar agreement is seen for HERVK (HML-10), where earlier work reported selective Env retention alongside extensive loss of Pol [30]; although exact locus correspondence differs between hg19 and GRCh38, likely due to assembly and annotation differences, the overall structural profile matches what we observe (some recognizable Env domains and few Pol domains). For the ERV3-like group, prior studies emphasized a single intact *env* at 7q11 and widespread degradation across other loci [31], and our map recapitulates this pattern while also revealing isolated Pol subdomains with unexpectedly high conservation. Likewise, the ancient HERVK (HML-6) lineage—first characterized by Medstrand *et al.* [29] as defective but retaining key retroviral motifs—exhibits the expected asymmetry between domains: Env is consistently eroded, whereas multiple loci retain well-preserved Pol subdomains (several exceeding 0.8 HMM profile coverage), reinforcing the notion that even deeply decayed families can maintain enzymatic remnants. Finally, the family-level trends we detect broadly agree with the earlier genome-wide survey by Vargiu *et al.* [14] (such as the relative preservation of HML-2 and the well-known Env retention in HERVW Syncytin-1), despite methodological and assembly differences that limit direct locus-level comparisons. Together, these consistencies indicate that our domain-centric framework integrates well with previous lineage-focused analyses while extending them by providing a unified, quantitative perspective on structural conservation across the full HERV landscape.

Although most HERV loci are fragmented and nonfunctional due to accumulated mutations, our analysis reveals that retroviral protein domains exhibit variable levels of structural conservation across loci. In many cases, no single locus retains all three canonical domains in full. However, the presence of complementary domain preservation at distinct genomic sites raises the possibility of functional trans-complementation, whereby transcripts or proteins encoded by separate HERV loci could, in theory, assemble into a partially functional retroviral-like complex. This modular conservation suggests a distributed potential for retroelement activity, particularly if such loci are co-expressed in the same cellular context. While trans-complementation has been proposed as a mechanism to increase retroelement copy number [3], it is generally considered rare in the context of HERVs [62]. Nevertheless, under conditions of widespread HERV reactivation—such as those observed in certain diseases [21]—this mechanism cannot be ruled out as a possible contributor to the formation of viral-like particles.

Beyond the potential for trans-complementation, the domain-level preservation also opens the door to alternative functional scenarios. In particular, the widespread conservation of enzymatic domains—especially RNase H and AP—in ancient subfamilies [6] such as HERVH and HERVE raises the possibility of residual activity or structural co-option. Given the established role of retroviral proteases in polyprotein maturation [63], their persistence may reflect functional retention or repurposing in host cellular contexts. Such conservation of retroviral enzymatic functions echoes known cases of co-option, such as the syncytins, which not only mediate placental fusion but also possess immunosuppressive properties [9, 64]. Similarly, the preservation of full-length Env domains in several HERVK loci supports a potential antiviral role through receptor interference, a mechanism initially proposed in koalas and bats whereby Env blocks viral entry by binding host receptors [15, 56]. In this context, it is conceivable that Pol-derived enzymatic domains such as RNase H and AP may also contribute to antiviral defense through dominant-negative interference or residual catalytic activity. For instance, HERV-derived RNase H might degrade viral RNA–DNA hybrids during reverse transcription or modulate immune sensing by processing nucleic acid hybrids that accumulate in the cytoplasm. In a similar manner, residual protease activity could interfere with the processing of viral polyproteins. These actions would mirror established antiviral strategies such as receptor interference by HERV Env proteins, and would conceptually parallel restriction factors such as TRIM5α and tetherin [65, 66], which block replication by binding and disrupting key viral processes. Although direct experimental validation remains lacking, the conservation of these catalytic folds across ancient HERV lineages supports the notion that HERV proteins may persist not merely as evolutionary remnants, but as active participants in host-virus interactions.

More generally, our findings also support a broader evolutionary perspective: that HERVs, long dismissed as genomic "junk," may have been retained in part due to their structural and regulatory utility [15, 67–71]. Rather than being purely parasitic, these sequences can act as white sheets of genomic material—regions with the capacity to incorporate, rearrange, and express novel DNA. In this view, HERVs offer evolutionary scaffolds from which new genes, transcripts, or regulatory networks may emerge. The retention of coding domains in even ancient subfamilies suggests that this material may not only be tolerated by the genome, but in some cases, positively selected as a source of innovation.

To further assess the generalizability of our framework, we applied the same pipeline to the human T2T-CHM13 assembly and the mouse GRCm39 genome. Both analyses ran without modification and recovered extensive sets of retroviral-like domains—17 946 in T2T-CHM13 and 47 840 in mouse—demonstrating that the domain-centric workflow is portable across species and compatible with distinct RepeatMasker annotations and ERV repertoires. While additional genomes may require minor adjustments to accommodate species-specific naming conventions for internal ERV regions, the successful application to two evolutionarily distant mammals highlights the potential of this approach for broader cross-species investigations into retroviral protein-domain conservation and the evolutionary trajectories of endogenous retroviruses.

Nevertheless, several important limitations must be considered. Before interpreting these patterns, it is important to acknowledge a methodological limitation inherent in the construction of our input dataset. Internal HERV regions were defined using RepeatMasker and associated libraries, which in turn reflect early HERV classification efforts that used sequence similarity in the *pol* gene as the primary criterion [14]. As a consequence, loci that retain Pol-like sequence are more likely to be detected, whereas internal regions in which *gag* or *env* have degraded—genes that contain few generic motifs and can therefore be missed—may be underrepresented in the annotated set. This upstream bias could contribute to the predominance of Pol-derived domains in our results. While our proposed interpretation—selective retention of retroviral enzymatic functions—is consistent with known evolutionary dynamics insofar as the HMM-based domain profiles reveal conserved enzymatic domain structures, we cannot fully disentangle biological signal from this detection bias.

Additionally, HERV loci do not necessarily use canonical 5′ LTR-driven transcription. Host-derived promoters, cryptic splice sites, and exonization of retroelement fragments—as exemplified in ACE2 and BRD9 [72, 73]—can generate unexpected transcripts. Because HERV "gene models" remain largely undescribed, the genomic ORFs we annotate may not fully capture the expressed repertoire of HERV-derived transcripts.

Our analysis relies solely on computational predictions of protein domains within candidate ORFs and lacks direct experimental evidence of expression or activity. We did not assess whether the identified ORFs are transcribed, translated, or post-translationally modified *in vivo*. Additionally, no functional assays were conducted to validate enzymatic activity or fusogenic potential. Therefore, while the structural conservation we observe provides strong evidence for potential functionality, empirical validation—including transcriptomics, proteomics, and targeted assays—is essential to fully assess biological relevance.

To address these open questions, future work should integrate domain annotations with bulk and single-cell transcriptomic data to identify actively expressed HERV loci. Ribosome profiling and mass spectrometry can further clarify which ORFs are translated and potentially functional. Moreover, structure prediction tools such as AlphaFold [74] may help assess whether conserved domains adopt native-like folds compatible with enzymatic or structural roles. Compar-

ative analyses across primates and other mammals could reveal lineage-specific patterns of domain retention and identify cases of convergent co-option. Finally, functional studies should explore whether these HERV-derived domains participate in host immunity, neurodevelopment, placentation, or other biological pathways where HERVs are increasingly implicated [15–18, 56]. HERVs can modulate innate immunity by activating cytosolic sensors (e.g. RIG-I, MDA5) or TLR pathways such as TLR3, TLR9, and TLR4, particularly via Env proteins like HERV-W ENV and by RNA–DNA hybrids [15]. Env proteins have been shown to induce inflammatory cytokines including IL-1β and TNF-α, and may contribute to antiviral defense or pathology, as proposed in multiple sclerosis [15, 75]. Together, these efforts will help determine whether the sequences we have identified are evolutionary relics—or underrecognized components of human molecular biology.

In conclusion, our study provides the first genome-wide map of structurally conserved HERV protein domains in humans and identifies hundreds of loci with potential for residual function or evolutionary co-option. These findings suggest that HERV-derived sequences may retain structural integrity that enables functional activity, with potential implications for both physiology and pathophysiology. By pinpointing specific candidates that preserve retroviral enzymatic or structural architecture, we offer a foundation for future functional and evolutionary studies into the enduring impact of endogenous retroviruses on human biology.

Acknowledgements

We would like to thank our colleague Jessica Gómez-Garrido, from the Genome Assembly and Annotation team at CNAG, for her valuable advice during the development of the annotation pipeline. We acknowledge the use of artificial intelligence tools (ChatGPT, OpenAI) to assist in code drafting and language editing. Finally, we thank the reviewers for their constructive feedback, which helped improve this manuscript.

Author contributions: T.M.-A. and A.E.-C. conceived the project. T.M.-A. performed the bioinformatic analyses. T.M.-A. and A.E.-C. wrote the manuscript. A.E.-C. supervised the project. A.P. contributed to scientific discussion and provided supervision during the revision stage.

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

This publication and all its results are supported by the AGAUR-FI predoctoral grant program (2025 FI-1 00642) Joan Oró, from the Secretariat for Universities and Research of the Department of Research and Universities of the Government of Catalonia, and by the European Social Fund Plus. Institutional support to CNAG was from the Spanish Ministry of Science, Innovation and Universities, Fondo de Investigaciones Sanitarias cofunded with ERDF funds (PI19/01772), the 2014–2020 Smart Growth Operating Program, and the Generalitat de Catalunya through the Departament de Recerca i Universitats and Departament de Salut.

Data availability

The human genome assembly GRCh38 (primary assembly) can be downloaded from the GENCODE website: https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_49/GRCh38.primary_assembly.genome.fa.gz.

The telomere-to-telomere human assembly T2T-CHM13v2.0 can be downloaded from NCBI: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/914/755/GCF_009914755.1_T2T-CHM13v2.0/GCF_009914755.1_T2T-CHM13v2.0_genomic.fna.gz.

The mouse genome assembly GRCm39 (primary assembly) can be downloaded from the GENCODE website: https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M38/GRCm39.primary_assembly.genome.fa.gz.

The Dfam database partitions used for RepeatMasker were downloaded from: https://www.dfam.org/releases/current/families/FamDB/dfam39_full.0.h5.gz (root), and https://www.dfam.org/releases/current/families/FamDB/dfam39_full.7.h5.gz (Mammalia):

The HMM profiles used for domain annotation were obtained from the GyDB database:

https://gydb.org/extensions/Collection/collection/db/GyDB_collection.zip. A combined HMM profile database ready to use with the pipeline is also provided in the Zenodo repository (see below).

The full HERV domain annotation datasets—including the RepeatMasker output files used for this analysis, BED and FASTA files, and InterProScan outputs—are available at Zenodo under the following DOIs: https://doi.org/10.5281/zenodo.17662456 (GRCh38), https://doi.org/10.5281/zenodo.17752356 (T2T-CHM13), and https://doi.org/10.5281/zenodo.17752474 (Mouse GRCm39).

The complete pipeline used to identify open reading frames, annotate retroviral domains, and process the results is available as a collection of Python, Bash, and R scripts at: https://github.com/funcgen/herv-domain-map.git and https://doi.org/10.5281/zenodo.18326381.

References

1. Griffiths DJ. Endogenous retroviruses in the human genome sequence. *Genome Biol* 2001;2:reviews1017.1. https://doi.org/10.1186/gb-2001-2-6-reviews1017
2. Stoye JP. Endogenous retroviruses: still active after all these years? *Curr Biol* 2001;11:R914–6. https://doi.org/10.1016/S0960-9822(01)00553-X
3. Bannert N, Kurth R. The evolutionary dynamics of human endogenous retroviral families. *Annu Rev Genom Hum Genet* 2006;7:149–73. https://doi.org/10.1146/annurev.genom.7.080505.115700
4. Lander ES, Linton LM, Birren B *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921.
5. Villesen P, Aagaard L, Wiuf C *et al.* Identification of endogenous retroviral reading frames in the human genome. *Retrovirology* 2004;1:32. https://doi.org/10.1186/1742-4690-1-32
6. Ueda MT, Kryukov K, Mitsuhashi S *et al.* Comprehensive genomic analysis reveals dynamic evolution of endogenous retroviruses that code for retroviral-like protein domains. *Mobile DNA* 2020;11:29. https://doi.org/10.1186/s13100-020-00224-w
7. Ito J, Sugimoto R, Nakaoka H *et al.* Systematic identification and characterization of regulatory elements derived from human

endogenous retroviruses. *PLoS Genet* 2017;13:e1006883. https://doi.org/10.1371/journal.pgen.1006883
8. Wang J, Han G-Z. Frequent retroviral gene co-option during the evolution of vertebrates. *Mol Biol Evol* 2020;37:3232–42. https://doi.org/10.1093/molbev/msaa180
9. Mi S, Lee X, Li X *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 2000;403:785–9. https://doi.org/10.1038/35001608
10. Cornelis G, Vernochet C, Carradec Q *et al.* Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc Natl Acad Sci USA* 2015;112:E487–496. https://doi.org/10.1073/pnas.1417000112
11. Coffin JM, Hughes SH, Varmus HE (eds.) *Retroviruses*. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press, 1997.
12. de Parseval N, Lazar V, Casella J-F *et al.* Survey of human genes of retroviral origin: identification and transcriptome of the genes with coding capacity for complete envelope proteins. *J Virol* 2003;77:10414–22. https://doi.org/10.1128/JVI.77.19.10414-10422.2003
13. Nakagawa S, Takahashi MU. gEVE: a genome-based endogenous viral element database provides comprehensive viral protein-coding sequences in mammalian genomes. *Database* 2016;2016:baw087. https://doi.org/10.1093/database/baw087
14. Vargiu L, Rodriguez-Tomé P, Sperber GO *et al.* Classification and characterization of human endogenous retroviruses; mosaic forms are common. *Retrovirology* 2016;13:7. https://doi.org/10.1186/s12977-015-0232-y
15. Srinivasachar Badarinarayan S, Sauter D. Switching sides: how endogenous retroviruses protect us from viral infections. *J Virol* 2021;95:e02299–20. https://doi.org/10.1128/JVI.02299-20
16. Duarte RRR, Nixon DF, Powell TR. Ancient viral DNA in the human genome linked to neurodegenerative diseases. *Brain Behav Immun* 2025;123:765–70. https://doi.org/10.1016/j.bbi.2024.10.020
17. Duarte RRR, Pain O, Bendall ML *et al.* Integrating human endogenous retroviruses into transcriptome-wide association studies highlights novel risk factors for major psychiatric conditions. *Nat Commun* 2024;15:3803. https://doi.org/10.1038/s41467-024-48153-z
18. Küry P, Nath A, Créange A *et al.* Human endogenous retroviruses in neurological diseases. *Trends Mol Med* 2018;24:379–94.
19. 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;11:936. https://doi.org/10.3390/biomedicines11030936
20. Evans EF, Saraph A, Tokuyama M. Transactivation of human endogenous retroviruses by viruses. *Viruses* 2024;16:1649. https://doi.org/10.3390/v16111649
21. Wang J, Lu X, Zhang W *et al.* Endogenous retroviruses in development and health. *Trends Microbiol* 2024;32:342–54. https://doi.org/10.1016/j.tim.2023.09.006
22. Belyi VA, Levine AJ, Skalka AM. Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog* 2010;6:e1001030. https://doi.org/10.1371/journal.ppat.1001030
23. Ono R, Nakamura K, Inoue K *et al.* Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality. *Nat Genet* 2006;38:101–6. https://doi.org/10.1038/ng1699
24. Matsui T, Miyamoto K, Kubo A *et al.* SASPase regulates stratum corneum hydration through profilaggrin-to-filaggrin processing. *EMBO Mol Med* 2011;3:320–33. https://doi.org/10.1002/emmm.201100140
25. Nakaya Y, Koshi K, Nakagawa S *et al.* Fematrin-1 is involved in fetomaternal cell-to-cell fusion in bovinae placenta and has contributed to diversity of ruminant placentation. *J Virol* 2013;87:10563–72. https://doi.org/10.1128/JVI.01398-13
26. Pastuzyn ED, Day CE, Kearns RB *et al.* The neuronal gene arc encodes a repurposed retrotransposon gag protein that mediates intercellular RNA transfer. *Cell* 2018;172:275–288.e18. https://doi.org/10.1016/j.cell.2017.12.024
27. Ashley J, Cordy B, Lucia D *et al.* Retrovirus-like gag protein Arc1 binds RNA and traffics across synaptic boutons. *Cell* 2018;172:262–274.e11. https://doi.org/10.1016/j.cell.2017.12.022
28. Grandi N, Cadeddu M, Blomberg J *et al.* Contribution of type W human endogenous retroviruses to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes. *Retrovirology* 2016;13:67. https://doi.org/10.1186/s12977-016-0301-x
29. Medstrand P, Mager DL, Yin H *et al.* Structure and genomic organization of a novel human endogenous retrovirus family: HERV-K (HML-6). *J Gen Virol* 1997;78:1731–44. https://doi.org/10.1099/0022-1317-78-7-1731
30. Grandi N, Cadeddu M, Pisano MP *et al.* Identification of a novel HERV-K(HML10): comprehensive characterization and comparative analysis in non-human primates provide insights about HML10 proviruses structure and diffusion. *Mobile DNA* 2017;8:15. https://doi.org/10.1186/s13100-017-0099-7
31. Andersson A-C, Yun Z, Sperber GO *et al.* ERV3 and related sequences in humans: structure and RNA expression. *J Virol* 2005;79:9270–84. https://doi.org/10.1128/JVI.79.14.9270-9284.2005
32. Grandi N, Tramontano E. HERV envelope proteins: physiological role and pathogenic potential in cancer and autoimmunity. *Front. Microbiol.* 2018;9:462. https://doi.org/10.3389/fmicb.2018.00462
33. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 2009;25: Chapter 4, 4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25
34. Sperber GO, Airola T, Jern P *et al.* Automated recognition of retroviral sequences in genomic data—RetroTector©. *Nucleic Acids Res* 2007;35:4964–76. https://doi.org/10.1093/nar/gkm515
35. Potter SC, Luciani A, Eddy SR *et al.* HMMER web server: 2018 update. *Nucleic Acids Res* 2018;46:W200–4. https://doi.org/10.1093/nar/gky448
36. Ritsch M, Brait N, Harvey E *et al.* Endogenous viral elements: insights into data availability and accessibility. *Virus Evol* 2024;10:veae099. https://doi.org/10.1093/ve/veae099
37. Goubert C, Craig RJ, Bilat AF *et al.* A beginner's guide to manual curation of transposable elements. *Mobile DNA* 2022;13:7. https://doi.org/10.1186/s13100-021-00259-7
38. Pačes J, Pavlíček A, Pačes V. HERVd: database of human endogenous retroviruses. *Nucleic Acids Res* 2002;30:205–6.
39. Bendall ML, Mulder Md, Iñiguez LP *et al.* Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput Biol* 2019;15:e1006453. https://doi.org/10.1371/journal.pcbi.1006453
40. Tokuyama M, Kong Y, Song E *et al.* ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. *Proc Natl Acad Sci USA* 2018;115:12565–72. https://doi.org/10.1073/pnas.1814589115
41. Storer J, Hubley R, Rosen J *et al.* The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* 2021;12:2. https://doi.org/10.1186/s13100-020-00230-y
42. Rice P, Longden I, Bleasby A. EMBOSS: the european molecular biology open software suite. *Trends Genet* 2000;16:276–7. https://doi.org/10.1016/S0168-9525(00)02024-2
43. Llorens C, Futami R, Covelli L *et al.* The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res* 2011;39:D70–4. https://doi.org/10.1093/nar/gkq1061
44. Lloréns C, Futami R, Bezemer D *et al.* The Gypsy Database (GyDB) of mobile genetic elements. *Nucleic Acids Res* 2008;36:D38–46.
45. Blum M, Chang H-Y, Chuguransky S *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 2021;49:D344–54. https://doi.org/10.1093/nar/gkaa977

46. Jones P, Binns D, Chang H-Y *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;30:1236–40. https://doi.org/10.1093/bioinformatics/btu031
47. Wang J, Chitsaz F, Derbyshire MK *et al.* The conserved domain database in 2023. *Nucleic Acids Res* 2023;51:D384–8. https://doi.org/10.1093/nar/gkac1096
48. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 2004;338:1027–36. https://doi.org/10.1016/j.jmb.2004.03.016
49. Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;376:44–53. https://doi.org/10.1126/science.abj6987
50. Wickham H, François R, Henry L *et al.* dplyr: a grammar of data manipulation. *R package* 2025. https://cran.r-project.org/package=dplyr [last accessed 21 January 2026].
51. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.
52. Pagès H, Aboyoun P, Gentleman R *et al.* Biostrings: efficient manipulation of biological strings. *Bioconductor package* 2025. https://bioconductor.org/packages/Biostrings (21 January 2026, date last accessed).
53. Schauberger P, Walker A. openxlsx: read, write and edit xlsx files. *R package*. 2025. https://cran.r-project.org/package=openxlsx (21 January 2026, date last accessed).
54. Flockerzi A, Burkhardt S, Schempp W *et al.* Human endogenous retrovirus HERV-K14 families: status, variants, evolution, and mobilization of other cellular sequences. *J Virol* 2005;79:2941–9. https://doi.org/10.1128/JVI.79.5.2941-2949.2005
55. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 2003;423:825–37. https://doi.org/10.1038/nature01722
56. Moelling K, Broecker F, Russo G *et al.* RNase H as gene modifier, driver of evolution and antiviral defense. *Front Microbiol* 2017;8:1745. https://doi.org/10.3389/fmicb.2017.01745
57. Mudge JM, Carbonell-Sala S, Diekhans M *et al.* GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Res* 2025;53:D966–75. https://doi.org/10.1093/nar/gkae1078
58. Li Z, Sheng T, Wan X *et al.* Expression of HERV-K correlates with status of MEK-ERK and p16INK4A-CDK4 pathways in melanoma cells. *Cancer Invest* 2010;28:1031–7. https://doi.org/10.3109/07357907.2010.512604
59. Zhao J, Rycaj K, Geng S *et al.* Expression of human endogenous retrovirus type K envelope protein is a novel candidate prognostic marker for human breast cancer. *Genes Cancer* 2011;2:914–22. https://doi.org/10.1177/1947601911431841
60. Rycaj K, Plummer JB, Yin B *et al.* Cytotoxicity of human endogenous retrovirus K-specific T cells toward autologous ovarian cancer cells. *Clin Cancer Res* 2015;21:471–83. https://doi.org/10.1158/1078-0432.CCR-14-0388
61. Jern P, Sperber GO, Blomberg J. Definition and variation of human endogenous retrovirus H. *Virology* 2004;327:93–110. https://doi.org/10.1016/j.virol.2004.06.023
62. Magiorkinis G, Gifford RJ, Katzourakis A *et al.* Env-less endogenous retroviruses are genomic superspreaders. *Proc Natl Acad Sci USA* 2012;109:7385–90. https://doi.org/10.1073/pnas.1200913109
63. Spall VE, Shanks M, Lomonossoff GP. Polyprotein processing as a strategy for gene expression in RNA viruses. *Seminars Virol* 1997;8:15–23. https://doi.org/10.1006/smvy.1997.0102
64. Muir A, Lever A, Moffett A. Expression and functions of human endogenous retroviruses in the placenta: an update. *Placenta* 2004;25:S16–25. https://doi.org/10.1016/j.placenta.2004.01.012
65. Stremlau M, Owens CM, Perron MJ *et al.* The cytoplasmic body component TRIM5α restricts HIV-1 infection in Old World monkeys. *Nature* 2004;427:848–53. https://doi.org/10.1038/nature02343
66. Neil SJD, Zang T, Bieniasz PD. Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature* 2008;451:425–30. https://doi.org/10.1038/nature06553
67. Smit AF. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* 1999;9:657–63. https://doi.org/10.1016/S0959-437X(99)00031-3
68. Garcia-Perez JL, Widmann TJ, Adams IR. The impact of transposable elements on mammalian development. *Development* 2016;143:4101–14. https://doi.org/10.1242/dev.132639
69. Platt RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. *Chromosome Res* 2018;26:25–43. https://doi.org/10.1007/s10577-017-9570-z
70. Nishihara H. Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. *Genes Genet Syst* 2020;94:269–81. https://doi.org/10.1266/ggs.19-00029
71. Johnson WE. Origins and evolutionary consequences of ancient endogenous retroviruses. *Nat Rev Micro* 2019;17:355–70. https://doi.org/10.1038/s41579-019-0189-2
72. Ng KW, Attig J, Bolland W *et al.* Tissue-specific and interferon-inducible expression of non-functional ACE2 through endogenous retroelement co-option. *Nat Genet* 2020;52:1294–302. https://doi.org/10.1038/s41588-020-00732-8
73. Inoue D, Chew G-L, Liu B *et al.* Spliceosomal disruption of the non-canonical BAF complex in cancer. *Nature* 2019;574:432–6. https://doi.org/10.1038/s41586-019-1646-9
74. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2
75. Christensen T. Association of human endogenous retroviruses with multiple sclerosis and possible interactions with herpes viruses. *Rev Med Virol* 2005;15:179–211. https://doi.org/10.1002/rmv.465

Received: October 9, 2025. **Revised:** December 2, 2025. **Accepted:** January 20, 2026