BMC Genomics

RESEARCH Open Access

An audit of the PeptideAtlas database uncovers evidence for repurposed pseudogenes and co-opted retroviral ORFs

Jose Manuel Rodriguez[1,2], Miguel Maquedano[3], Daniel Cerdán-Vélez[3], Andrea Laguillo-Gómez[1], Enrique Calvo[1,2], Federico Abascal[4], Jesús Vázquez[1,2*] and Michael L. Tress[3*]

Abstract

Background The human genome has been the subject of scrutiny for more than two decades, yet new protein coding genes are still being uncovered. Recently ribosome profiling experiments have provided evidence for the translation of thousands of novel open reading frames (ORFs). To determine how many of these novel ORFs have peptide support, we carried out an in-depth investigation of an entire mass spectrometry proteomics database.

Results We analysed the peptides housed in the human build of the PeptideAtlas database and identified reliable evidence for 35 potential coding genes not annotated in the Ensembl/GENCODE reference gene set. Evidence from complementary sources confirmed that 16 were almost certainly coding genes, but we believe that at least 14 are most likely to be undergoing aberrant translation. These 14 genes had reading frames that were not preserved beyond human and their peptides were restricted to cancers or cell lines. Remarkably, three of the sixteen likely coding genes were derived from endogenous retroviral *gag* ORFs and were expressed only in placenta. All three had evidence of purifying selection. Retroviral *env* ORFs (syncytins) with distinct origins are expressed in almost all mammalian placentae and these results suggest that co-opted *gag* ORFs may also play an important role in placental development.

Conclusions Our analysis shows that proteomics data can be used in conjunction with evolutionary evidence to confirm the existence of new coding genes. The evidence suggests that both testis and placenta are the tissues most likely to express still to be identified coding genes, and that there may be other transposon-derived ORF that have been co-opted as coding genes. The strong evidence for the translation of regions under dysregulated conditions has important implications for the annotation of coding genes and in the analysis of cancer and other degenerative diseases.

Keywords Proteomics, Coding genes, Pseudogenes, Endogenous retrovirus, Co-option

*Correspondence:
Jesús Vázquez
jesus.vazquez@cnic.es
Michael L. Tress
mtress@cnio.es
[1]Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid 28029, Spain
[2]CIBER de Enfermedades Cardiovasculares (CIBERCV), Madrid 28029, Spain
[3]Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain
[4]Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK

Background

The human reference genome has been completed with the annotation of heterochromatic regions [1] and the Y chromosome [2] by the T2T consortium. Despite this, the annotation of a final set of human coding genes is still some way from being finished. The principal reason for this is that the main reference databases [3–6] still disagree on which genes code for proteins [7, 8], although the completion of the novel CHM13 human reference, in which preliminary estimations of novel coding genes run from 2 [9] to 300 [4], has added a new layer of complexity. On top of this, there is increasing evidence from ribosome profiling experiments that a surprisingly large number of unannotated open reading frames may undergo translation [10].

The initial drafts of the human reference genome [11, 12] pinned the number of protein coding genes at between 25 and 35,000. Since then, estimates of the number of coding genes have been part of a gradual downward trend [7, 13–17]. The most recent GENCODE release [3] (v49) annotates 19,435 coding genes, although between the RefSeq, UniProtKB and Ensembl/GENCODE reference sets there were still more than 21,800 annotated coding genes in 2024 [18].

Recently, several high-profile large-scale ribosome profiling analyses of the human genome have been published [19, 20] that may reverse this trend. These provide evidence for tens of thousands of novel unannotated open reading frames (ORFs). A consortium has been formed to investigate whether these regions code for proteins [10] following these (and other) studies, with one of the early papers published by members of the consortium suggesting that coding gene numbers might expand by 30% on the back of the more than 7,000 well-supported novel ORFs they find [21].

The paper [21] highlights the discovery of nine novel ORFs that are often cited as evidence for the functional importance of short ORFs as a whole. These are *APELA*, *ASDURF*, *MIURF*, *MRLN*, *MYMX*, *POLGARF*, *TINCR*, and the as yet unnamed uORFs (upstream open reading frames) in *MKKS* and *SLC35A4*. What links these nine examples is not that they are short (three are longer than 100 amino acids and *POLGARF* has 260), but that they are all ancient. Seven can trace their ancestry back to lobe-finned fish at least, while the other two are conserved across all mammals. Although there are undoubtedly further ancient yet to be discovered ORFs (one such example is the upstream overlapping ORF (uoORF) in *GRIN2A* [22]), these 9 conserved ORFs are not representative of the class of novel ORFs as a whole because most novel ORFs detected in ribosome profiling experiments have little or no evidence of cross-species conservation.

One further problem is that although there is a lot of ribosome profiling evidence for novel ORFs, the proteomics evidence is not strong. Even the two large-scale ribosome profiling analyses that detected tens of thousands of novel ORFs failed to find much evidence for their products in standard proteomics experiments. One found peptide evidence for just seven unannotated ORFs [19], and although the other detected 541 peptides for hundreds of novel ORFs in two separate proteomics analyses of heart tissues, only five of these novel peptides (fewer than 1%) were detected in both heart experiments [20]. To determine whether peptides could be detected for these novel ORFs in large-scale experiments in multiple-tissues, we mapped the novel ORF translations from both papers against spectra from five large-scale proteomics experiments [22]. We found substantial evidence for translation upstream of known coding genes, but just two novel coding genes had convincing peptide evidence. The consortium themselves [10] found just 13 peptides for novel ORFs in PeptideAtlas [23], even allowing for one peptide per coding gene and not restricting to tryptic peptides. The extent to which these novel ORFs produce stable proteins is not clear.

If novel ORFs are being translated, as the ribosome profiling experiments suggest, and the proteins are stable, there ought to be peptide evidence for them in proteomics analyses. So, what is happening? One explanation is technical. The smallest of the proteins and those proteins with a special amino acid composition may not be amenable to detection in standard trypsin-based proteomics experiments, although this cannot explain the thousands of undetected ORFs translations on its own. Another possible reason may be that some of the transcripts captured in ribosome profiling experiments are not translated, for example due to control mechanisms at the level of the ribosome [24], or translated in smaller quantities that cannot be detected in proteomics experiments. Even if translated in sufficient quantities, it is also possible that many of these peptides are rapidly degraded [25]. There is certainly evidence for some degradation in the ribosome profiling-based analyses [19, 20], since there are plenty of novel ORF peptides in proteasome-derived human leukocyte antigen class 1 (HLA-I) proteomics experiments [21, 26].

The final possibility is that the novel ORFs are translated in few tissues or under certain conditions. If this was happening, a single proteomics analysis would be unlikely to find much evidence, but an analysis of multiple large-scale proteomics experiments from a large range of tissues ought to turn up substantial support, as long as stringent statistical validation is carried out.

Here, we carry out a manual analysis of the novel ORFs detected in the PeptideAtlas database. PeptideAtlas is a database that maps annotated and predicted proteins to thousands of proteomics experiments. PeptideAtlas interrogates spectra from a large range of proteomics

experiments. One advantage of this analysis is that the protein search database used by PeptideAtlas includes many predicted proteins from large-scale analyses. In particular, it contains almost two thirds of the sequences predicted in the two large-scale ribosome profiling analyses [19, 20]. The combination of these three features (and the possibility of manual analysis) means that PeptideAtlas is a potential source of evidence for the novel ORFs identified in ribosome profiling experiments.

We downloaded the peptides from the PeptideAtlas database with the aim of discovering evidence for the translation of the potential protein coding genes identified in the ribosome profiling analyses and for other genes not yet annotated in reference databases. We found evidence for hundreds of regions that are not annotated as coding in the Ensembl/GENCODE human reference gene set, though most overlapped known coding genes (alternative isoforms, translated upstream regions), There was convincing evidence for the translation to protein of 35 genes not part of the GENCODE reference set, including seven of the novel ORFs predicted by ribosome profiling experiments [19, 20]. Remarkably, three of these potential coding genes derived from endogenous retrovirus (ERV) *gag* ORFs that appear to have been repurposed in their primates hosts to perform biological roles in placenta, a process known as co-option [27].

Methods

Generating a list of novel peptides from PeptideAtlas

We downloaded peptides from the January 2023 build of the human PeptideAtlas repository. Peptides are pre-mapped in PeptideAtlas by the Trans-Proteomic Pipeline (TPP, 28), a suite of locally installed tools. The TPP maps spectra from large and small-scale proteomics experiments to the PeptideAtlas human protein search database [28]. We analysed the peptides from PeptideAtlas because the search database includes proteins from a wide range of sources and because TPP has stringent statistical validation at the peptide and protein level.

The search database (THISP, 29) is made up of sequences from the UniProtKB [6], NextProt [30], and RefSeq [5] databases, as well as likely contaminants, microbes and many non-reference peptides provided by contributors. The version of THISP used in the January 2023 build had 341,040 sequence distinct entries, Sequences from contributors provided the largest number of entries to the database. Contributor sequences are mostly protein sequences culled from large-scale experiments that might be protein coding. More than two thirds (67.8%) of the novel ORFs identified by Chen et al. [19] and van Heesch et al. [20] were annotated in the January 2023 build of the THISP database.

The January 2023 build of PeptideAtlas lists 3,489,945 peptides and these peptides are mapped to 62,245 protein entries. However, all peptides identified by the pipeline, including those that mapped to multiple THISP entries, were mapped to a single THISP entry in the PeptideAtlas file. This means any peptide might map to more than one protein entry, and many do because many of the entries in the THISP database are duplicates or at least have regions of common protein sequences.

We were interested in PeptideAtlas peptides that did not map to the Ensembl/GENCODE reference set [3, 4]. The GENCODE v45 reference set was downloaded from the GENCODE website. Before we remapped the 3,489,945 PeptideAtlas peptides, we excluded all decoy peptides, peptides that mapped to common contaminants and to microbes, and peptides that were not fully tryptic. All these labels were available from the PeptideAtlas peptide file. Peptides that mapped to GENCODE v45 proteins after this step were excluded from the remainder of the analysis (Fig. 1).

After filtering, we were left with a total of 19,266 tryptic peptides that did not map to GENCODE v45 genes. A further filtering step removed 5,910 novel peptides that we knew mapped to UniProtKB immunoglobulins or major histocompatibility proteins. We know that these proteins will differ in amino acid sequence from the GENCODE v45 entries from the reference set. This left us with 13,356 novel peptides (Additional table S1).

Defining strong peptides

Since we have analysed the UniProtKB database as part of another study [18], we know that there are many probable pseudogenes in the THISP search database, and we know that many of the pseudogenes will have peptides that differ from their parent genes by just one or two amino acids. Proteomics experiments do not always correctly distinguish canonical peptides (from known proteins) from predicted peptides that would map to pseudogenes, largely because of possible confusion with single amino acid variations (SAAVs) and post-translational modifications (PTMs) [29]. Since canonical proteins often have many naturally occurring SAAVs and PTMs, it would not be uncommon for a search engine to mistakenly map one of these peptides to a pseudogene. One way to get around these problems would be to just remove any peptide that had one or two SAAV with respect to a GENCODE v45 protein. However, this would mean that *bona fide* novel coding genes with clear peptide support that are close homologues of known coding genes might be lost from the analysis. Another solution would have been to inspect the protein spectrum matches of all these peptides. This is the best solution but is not feasible for thousands of variant peptides.

To reduce the number of variant peptides that we had to manually curate we used the number of times each peptide has been observed across the experiments

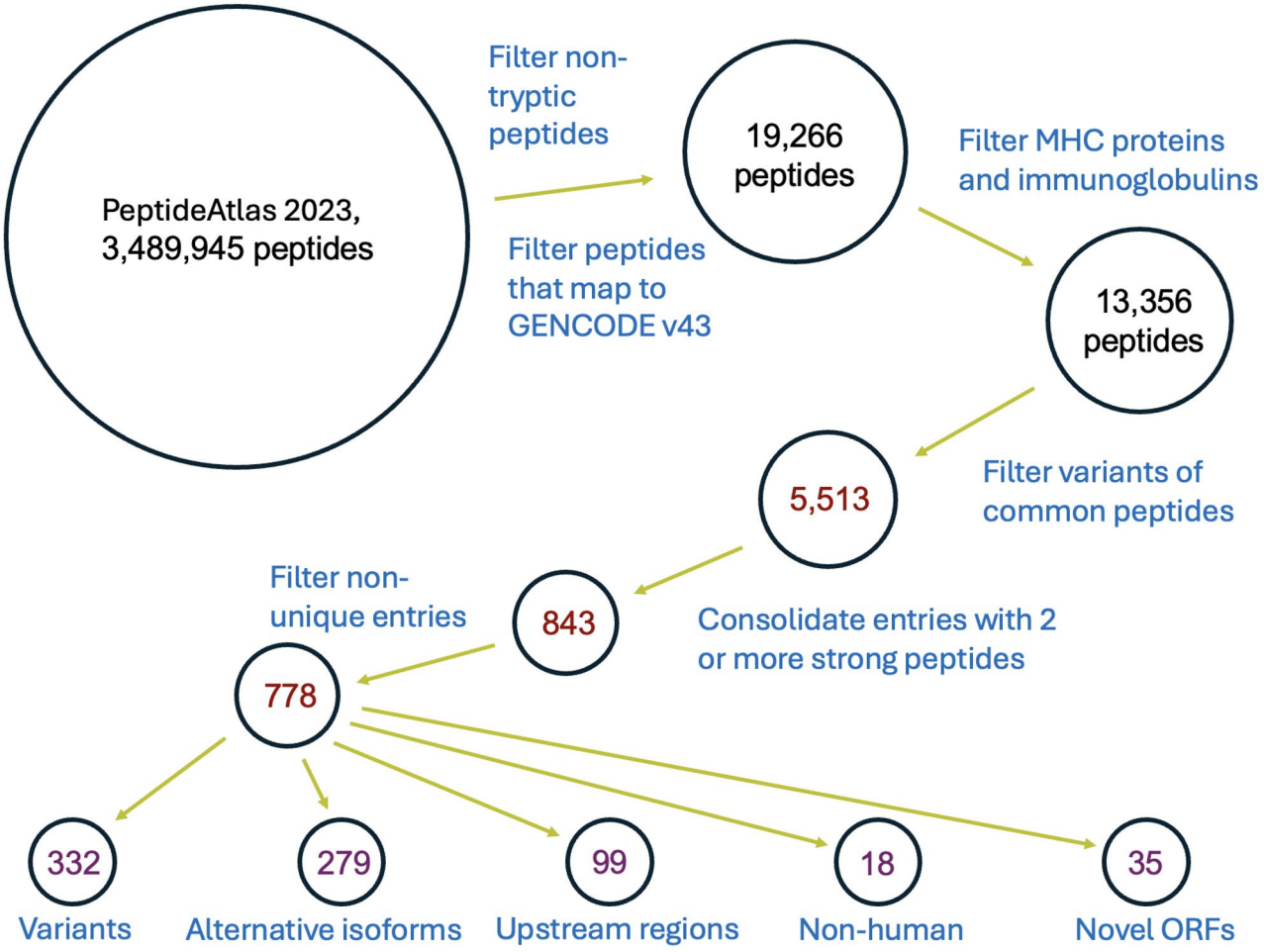


Fig. 1 PeptideAtlas Data Curation Workflow. The workflow of the analysis of PeptideAtlas peptides that did not map to proteins in the Ensembl/GENCODE reference set and the THISP entries that they mapped to. The five largest classes of "missing" proteins that were identified are shown in the last row

interrogated in PeptideAtlas, included in the downloaded list of PeptideAtlas peptides. The more observations a peptide has, the more the peptide is expressed.

We generated a score for all PeptideAtlas peptides based on the number of each peptide. For each PeptideAtlas peptide we summed all the observations of all GENCODE v45 peptides that differed by one or two amino acids. These highly similar known peptides were termed double amino acid variant (DAAV) peptides. The observation score for each PeptideAtlas peptide was then the number of its own observations divided by the sum of its own observations and those of the DAAV peptides.

PeptideAtlas peptides similar to peptides from GENCODE v45 had observation scores below one, while peptides with no similar DAAV peptides had observation scores of one. The peptides with the lowest observation scores are quite likely to be either amino acid variants or post-translational modifications of the peptides in the GENCODE v45 set. Peptides with observation scores between 0.5 and one have similar GENCODE v45 peptides but there are more observations for the novel peptide than for the GENCODE v45 peptides.

For our analysis we regarded peptides with observation scores above 0.9 as "strong" peptides. These strong peptides are referred to in the paper as strong discriminating peptides (SDPs) and were those that were used to search for unannotated genes and coding regions. Out of 13,356 PeptideAtlas peptides that did not map to the GENCODE v45 reference set, 8,170 had DAAVs in GENCODE v45. Of these 327 had an observation score greater than 0.9 for a final total of 5,513 SDPs (Additional table S1). The distribution of observation scores that had DAAV peptides in GENCODE v45 is shown in Additional figure S1.

Manual curation of the entries supported by strong discriminating peptides

We found 5,513 SDPs in PeptideAtlas that did not map to GENCODE v45 proteins, mapping to 3,774 distinct entries in total. Although most entries were supported by

a single SDP, 843 entries had two or more mapping SDPs (Fig. 1).

We manually curated the novel peptides for the 843 entries that had two or more SDPs that did not map to the GENCODE v45 annotation. Manual curation attempted to ascertain where the peptide mapped, whether the peptide was most likely to be a variant of a known protein sequence (despite our filtering for amino acid variants some variants still got through the filters), whether it was evidence for a novel splice isoform or a novel ORF in an untranslated region of a gene, or whether the peptide supported a possible new coding gene that was not yet annotated in Ensembl/GENCODE.

Since peptides from the same gene often map to more than one entry in the THISP database, we also combined entries where the peptide data supported the same protein product. The peptides for the LINE-1 ORF1 protein mapped to nine distinct THISP entries, for example, so we combined the evidence into a single entry during the manual curation process. After combination the number of PeptideAtlas entries with at least two strong novel peptides was 778.

During the curation, we found that many of the apparent strong SDPs were highly similar to GENCODE v45 peptides and were possibly single amino acid variants or post-translational modifications of these peptides. These SDPs were different because of a change to a lysine or arginine residue that produced very different peptides after theoretical trypsin cleavage. There were also SDPs that had insertions or deletions of one or more amino acids in low complexity regions that were also most easily explained as amino acid variants or PTMs of canonical human peptides from known coding genes. Entries with these SDPs were usually classified as likely variants of Ensmbl/GENCODE proteins.

Calculation of dN/dS rates

We generated codon alignments based on alignments of simian sequences from the Cactus [31] 447-way and 470-way mammalian alignments [32] and edited them where necessary with Jalview [33] Species were removed from the alignment if their codon sequence included stop codons or frame shifts. The neighbouring *env* coding genes *ERVV-1* and *ERVV-2* on chromosome 19 required substantial editing.

From curated codon alignments we reconstructed the phylogenetic tree using Phyml v3.3.20250429 [34]. We estimated dN/dS ratios (the ratios of non-synonymous to synonymous substitutions) using Paml/Codeml [35, 36] with model 0 (one, shared dN/dS parameter). The significance was estimated with a likelihood ratio test (LRT) between the free-dN/dS model and a null model with dN/dS fixed at 1 (neutral evolution). The corresponding p-values were adjusted for multiple hypothesis testing with the Benjamini and Hochberg false discovery rate method [37].

Results

We identified 13,356 novel tryptic peptides in PeptideAtlas, peptides that did not map to protein sequences present in the Ensembl/GENCODE reference gene set. Of these 13,356 peptides, 5,513 were strong discriminating peptides (SDPs, those that were substantially different from Ensembl/GENCODE peptides, see methods section). Focusing on PeptideAtlas entries that had high support (see Methods), we carried out an in-depth review of all THISP database entries [29] that were supported by at least two distinct SDPs. We considered each of these 778 novel entries on its merits.

The largest proportion (332) had SDPs that were most easily explained as variants or post-translational modifications of known peptides from highly expressed proteins (see methods section). These variant peptides are almost always mapped erroneously to known pseudogenes. We also identified novel alternative isoforms (as many as 279 entries), translation from 5' untranslated regions (99 entries) [38, 39], potential contaminants (18) and non-reference genes (Fig. 1, Additional figures S2 and S3). These entries, along with the ten genes with the highest number of SDPs, are analysed in more detail in Additional file 1. The ten most detected genes include retroviral genes, non-reference genes and likely contaminants, as well as novel ORFs.

Peptide evidence for coding genes novel to Ensembl/GENCODE

Recent analyses of large-scale ribosome profiling analyses have suggested that there are many new protein-coding genes in the human genome, but to date there has been little confirmatory support for novel ORFs in standard proteomics experiments. Here we investigate whether the protein and peptide data from the 2,416 proteomics experiments in the 2023 build of the PeptideAtlas database support the translation of unannotated protein coding genes.

Peptide support from a large-scale source such as PeptideAtlas may validate the translation and stability of gene products, but manual intervention is required to guarantee that a novel peptide is not a false positive identification. Even if the novel peptides are clearly supported by the spectra, further work is required to confirm whether the gene product is more likely to be a biologically relevant protein or only produced under the dysregulated conditions that are typical of many cancer cells [40, 41].

Genes with peptide support annotated in the GENCODE v49 reference set

We identified 35 possible coding genes that were not annotated in the GENCODE v45 gene set with the 2023 version of the PeptideAtlas (Additional Table S1). The data was passed on to the GENCODE annotators and ten of the 35 genes with PeptideAtlas support (Table 1) have been annotated as coding as of the GENCODE v49 (Ensembl 115) release. In addition, GENCODE has annotated 11 paralogues of these genes as coding, so a total of 21 coding genes have been added to the Ensembl/GENCODE reference set after our analysis.

Seven of the ten genes annotated as coding by Ensembl/GENCODE in version 49 of GENCODE arose by gene duplication. These are *C5orf60* (now *SPATA31J*), *CFAP144P1*, *ENSG00000293661*, *MLS3B*, *MYH16*, *TSPY26P* and *ZNF840P*. Four of these duplications *(C5orf60*, *CFAP144P1*, *MLS3B* and *MYH16*) are testis expressed and four *(C5orf60*, *MLS3B*, *MYH16* and *TSPY26P*) are predicted to produce proteins that are N-terminally truncated with respect to the parent gene. This is part of the reason that these genes were annotated as pseudogenes rather than as coding genes.

ZNF840P has the most convincing evidence (Fig. 2A). It is a simian duplication that has evolved considerably. All PeptideAtlas peptides for *ZNF840P* were detected in oocytes, which fits with the RNAseq data from GTEx [42]. *CFAP144P1* is highly similar to parent gene *CFAP144*, but there is much more evidence for the peptides that map to *CFAP144P1* than for the peptides that map to *CFAP144*. All PeptideAtlas peptides that map to *CFAP144P1* were found in experiments carried out on testis or sperm and this is also corroborated by RNAseq data from GTEx. Finally, all the masses of the fragment ions in peptide spectrum matches (PSMs) for *CFAP144P1* match perfectly with the candidate peptides identified in the protein databases (Fig. 2B). *C5orf60*, *MLS3B*, and *TSPY26P* are discussed in more detail in Additional file 1.

Table 1 The 10 PeptideAtlas analysis genes annotated as coding in GENCODE v49 (G49)

THISP entry	UniProtKB Gene	UniProtKB	SDPs	G49 gene name	Coding?
A6NFR6	*C5orf60*	Proteome	16	*SPATA31J*	Yes
Q6ZVS7	*CFAP144P1*	Proteome	6	*CFAP144P1*	Yes
P61550	*ERVS71-1*	Proteome	2	*ENSG00000293570*	Yes
Q8N319	*LINC03040*	Proteome	2	*LINC03040*	No
P0C860	*MSL3P1*	Proteome	9	*MSL3B*	Yes
Q9H6N6	*MYH16*	Proteome	11	*MYH16*	No
A6NNC1	*POM121L1P*	Proteome	3	*ENSG00000288349**	Unplaced
Q9H489	*TSPY26P*	Proteome	2	*ENSG00000293164*	Yes
A6NDX5	*ZNF840P*	Proteome	12	*ZNF840P*	Yes
Q3ZM62	-	TrEMBL	3	*ENSG00000293661**	Yes

Asterisks indicate where GENCODE have annotated multiple paralogues in addition to the gene identified in PeptideAtlas. In total 21 coding genes have been added. The "Coding?" column contains our evaluation of whether these genes are likely to code for a functional gene product and is justified in the main text of the paper

Almost a thousand UniProtKB reference proteome entries do not have equivalent coding genes in GENCODE v45 [18]. Although we have shown that most of the regions that these UniProtKB entries map to are unlikely to code for proteins [18], some will, and it is important to identify these entries. Nine of the ten newly annotated Ensembl/GENCODE coding genes had equivalent UniProtKB reference proteins. The only newly annotated Ensembl/GENCODE coding gene that was not in the UniProtKB proteome is *ENSG00000293661*, which corresponds to TrEMBL entry Q3ZM62.

ENSG00000293661 was clearly a coding gene since there were 3 SDPs in the 2023 PeptideAtlas build and there are 5 SDPs in the most recent build that map to THISP database entry Q3ZM62. They cover the sequence from residues 2 to 73 because the UniProtKB sequence of Q3ZM62 is incorrect after residue 80). *ENSG00000293661* is on chromosome X just downstream of and antisense to *ETDA*, another single exon gene (Fig. 3A). Q3ZM62 and the *ETDA* protein have 40% identity.

ETDA has three close paralogues on chromosome X, *ETDB*, *ETDC* and an unnamed pseudogene which we have called *ETDD* in this paper (Fig. 3A). Each of these three paralogues also has an antisense homologue of *ENSG00000293661*, *SMIM10L2B-AS1*, *ENSG00000293662* and *ENSG00000293663* respectively. These three novel paralogues of *ENSG00000293661* have also now been annotated as coding by GENCODE.

The original duplication that produced both the ETD family gene and the ENSG00000293661 family gene occurred in the mammalian clade because there are orthologues of both gene families throughout placental mammals. Within the primate lineage, segmental duplications generated another three copies of this antisense gene pair. The most recent segmental duplication occurred after the split with chimpanzees and generated the *ETDA*/*ENSG00000293661* and *ETDB*/*SMIM10L2B-AS1* (Fig. 3A) gene pairs. Since this duplication is so recent, the *ETDA* and *ETDB* proteins are identical and *ENSG00000293661* and *SMIM10L2B-AS1* are almost indistinguishable (Fig. 3B), though one of the PeptideAtlas peptides clearly favours *ENSG00000293661* rather than *SMIM10L2B-AS1* (Fig. 3C).

It is unlikely that all of these paralogues are coding. *ETDA* and *ETDB* proteins are indistinguishable, but at least one is coding. *ETDC* genes in humans, chimpanzees and gorillas have a variant upstream of the translation initiation site that would weaken the Kozak sequence, and gorilla *ETDC* has a frame shift, so this gene may be

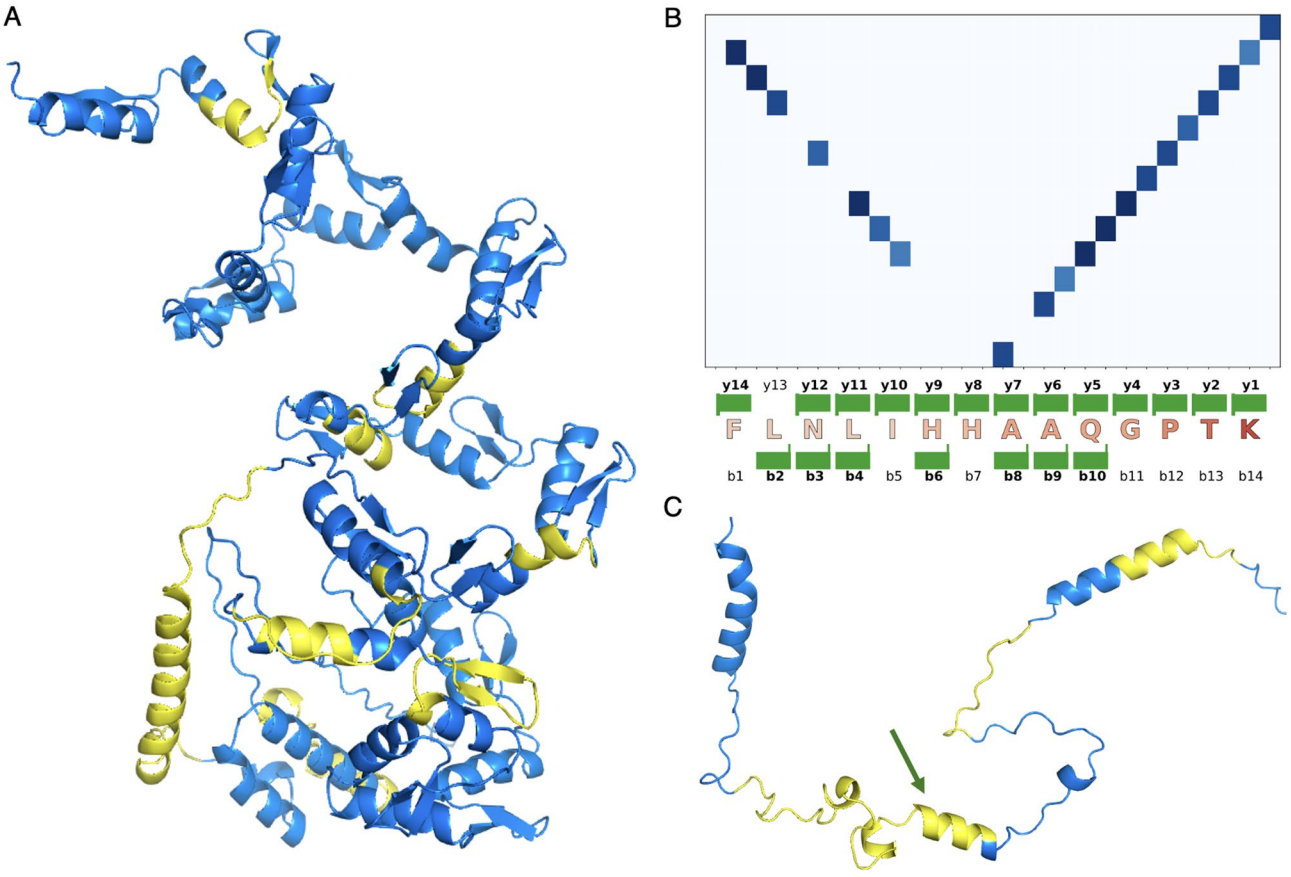


Fig. 2 Peptides and PSMs for coding genes *ZNF840P* and *CFAP144P1*. **A** The AlphaFold [43] structure of macaque ZNF840P with the peptides detected in PeptideAtlas marked in yellow on the structure. **B** A section of the VSeq [44] analysis of a PSM supportive of new coding gene *CFAP144P1*, showing just the intensity of the fragments matched in the b-series (above left) and y-series (above right) as squares, and the peptide sequence that is supported by the fragments (below) as green rectangles. The full output of VSeq for this PSM can be seen in Additional figure S4. **C** The AlphaFold structure of human CFAP144P1 with the peptides detected in PeptideAtlas marked in yellow on the structure, the arrow shows the position of the peptide identified by VSeq

a pseudogene. *ETDD* is annotated as a pseudogene in hominins because it has a premature stop codon.

Of the four newly annotated coding genes, *ENSG00000293661* appears to be coding because one of the PeptideAtlas peptides clearly favours its protein over that of the almost identical *SMIM10L2B-AS1*. Neither *SMIM10L2B-AS1* nor *ENSG00000293662* nor *ENSG00000293663* have any supporting peptides in the most recent version of PeptideAtlas. All old world monkey orthologues of *ENSG00000293663* have frame shifts, while in orangutans and some gibbon species *ENSG00000293662* orthologues have lost their start codons. This suggests that both these genes may be pseudogenes in human too. One further clue comes from the tissue expression of these 8 genes. RNAseq data from GTEx shows that *ENSG00000293661*, *SMIM10L2B-AS1* and *ETDB* all have strong testis specific expression, while the other 5 genes (*ENSG00000293662*, *ENSG00000293663*, *ETDA*, *ETDC*, and the pseudogene *ETDD*) do not.

MYH16 is almost certainly aberrantly expressed

The evidence *for MYH16* translation is conflicting. *MYH16* is coding in many vertebrates and clearly conserved in primates. Outside of humans it is specifically expressed in jaw muscles [45]. However, human *MYH16* has been predicted to be a unitary pseudogene [46]. The gene has multiple radical amino acid changes with respect to mammalian *MYH16*, has lost the 5' splice signal from exon 28 and, most importantly, there is a frameshift in exon 15. RefSeq annotates a coding sequence for *MYH16* because "peptide support [suggests] that a shorter protein containing only a myosin tail domain is expressed in human". The transcript would initiate translation from an ATG in exon 19 and would skip exon 28, avoiding both high impact mutations in human *MYH16*.

We found 11 SDPs in cancer cell experiments and in heart and muscle tissues in PeptideAtlas that support translation. This would be strong peptide support, but deeper analysis of the PeptideAtlas peptides shows that *MYH16* translation is only supported in cancer cell lines. None of the PSMs from heart and muscle experiments

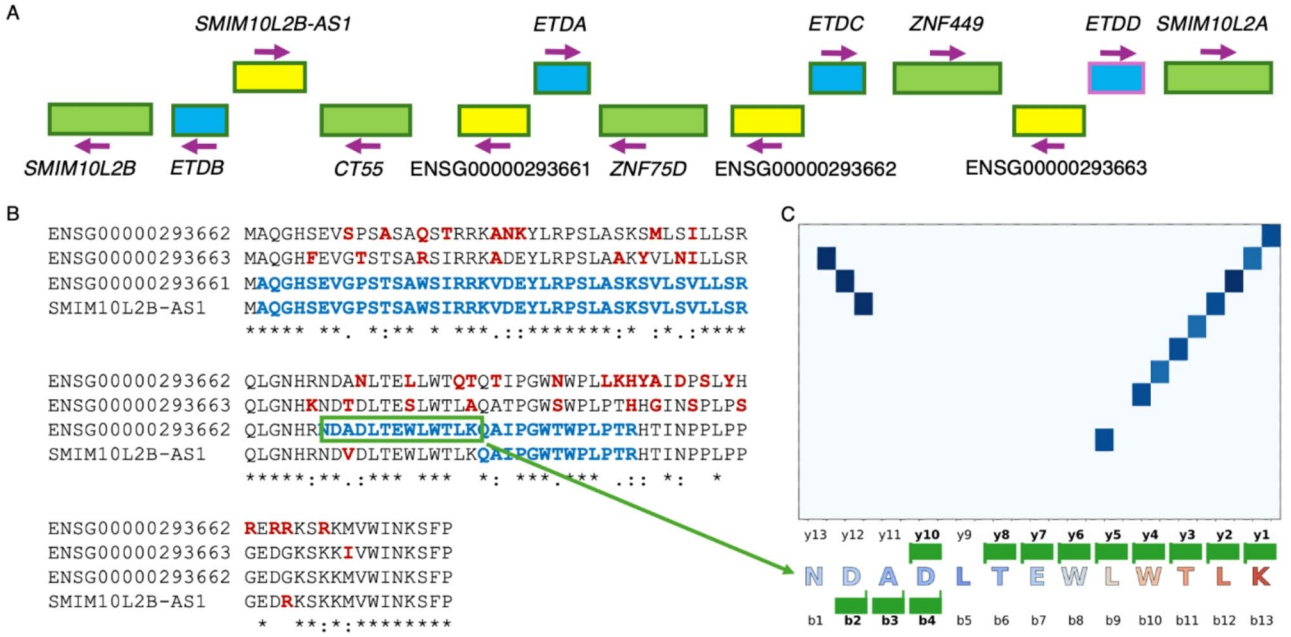


Fig. 3 The eight ETD-like genes on chromosome X. **A** A schematic representation of the human ETD family and the related novel paralogues on chromosome X. ETD genes are shown as blue boxes, paralogous antisense ETD-like genes are in yellow, other intervening genes are marked in green. Gene sense is marked with arrows. Annotated pseudogenes have a pink border. **B** An alignment between the four novel antisense paralogues. Peptides detected for these genes are in bold and coloured in blue, amino acid differences from the *ENSG00000293661* product are shown in bold and red. **C** A section of the VSeq analysis of a PSM supportive of *ENSG00000293661*, showing just the intensity of the fragments matched in the b-series (above left) and y-series (above right) as squares, and the peptide sequence that is supported by the fragments (below) as green rectangles. This peptide is the one that distinguishes *ENSG00000293661* from *ENSG00000293662*. The full output of VSeq for this PSM can be seen in Additional figure S5

in PeptideAtlas unequivocally support the *MYH16* peptides, instead the peptides found in these tissues are likely to be from other myosin genes (Fig. 4).

The original experiments in which multiple peptides were identified for *MYH16* used an epidermoid carcinoma cell line [47]. *MYH16* has been found to be aberrantly upregulated in lung adenocarcinoma and other cancers [48] and its expression correlates with worse overall survival rates. These results suggest that *MYH16* may not be a protein coding gene after all but instead is translated only in dysregulated conditions.

ERVS71-1, *LINC03040* and a gaggle of *POM121L1P* paralogues

ERVS71-1 is an almost complete *env* gene from a HERV-T (human endogenous retrovirus-T) provirus that lacks just the five C-terminal amino acids [49]. The two SDPs we found (there are four in the most recent version of PeptideAtlas) are supported by spectra in which all fragment ion masses match perfectly (Fig. 4C). It has been shown that the virus was inserted into the germline at least 13 million years ago [50]. There is support for a possible functional role in contributing to the protection against HERV-T retrovirus in ancestral hominids; experimental evidence suggests that *ERVS71-1* might interfere with cell surface levels of *SLC16A1*, the receptor responsible for the cellular access of ancient HERV-T viruses [50]. Analysis of the provirus found that it is highly unlikely that the reading frame (626 amino acids) would have been largely maintained under neutral evolution [50].

Unfortunately, *ERVS71-1* has no direct orthologue in any other species. While there is a conserved HERV-T *env* ORF in orangutan [50] it is clearly orthologous to one of the other multiple *env* pseudogenes in the human genome rather than *ERVS71-1*. The same is true of the related HERV-T *env* pseudogenes in chimpanzee and gorilla. If there are no orthologues of *ERVS71-1* in chimpanzee or gorilla, *ERVS71-1* may be a recent duplication that has undergone considerable changes. This does not rule out a functional role for *ERVS71-1*, but it does cast doubt on its role in protection against HERV-T retrovirus. In addition, while *ERVS71-1* transcript evidence is almost exclusively in thyroid tissues [42], PeptideAtlas peptides are detected mostly in stem cells and not in healthy thyroid. The expression in stem cells, plus the maintaining of the reading frame of almost the entire *env* ORF in humans means that this gene might still be coding but it is not a clear cut case.

LINC03040, also known as *C6orf223*, was originally protein-coding, but the gene was reclassified as non-coding because it had little or no evolutionary support. It has now been re-annotated as coding. Meanwhile, it was used to define a Pfam domain [51], but the Pfam domain

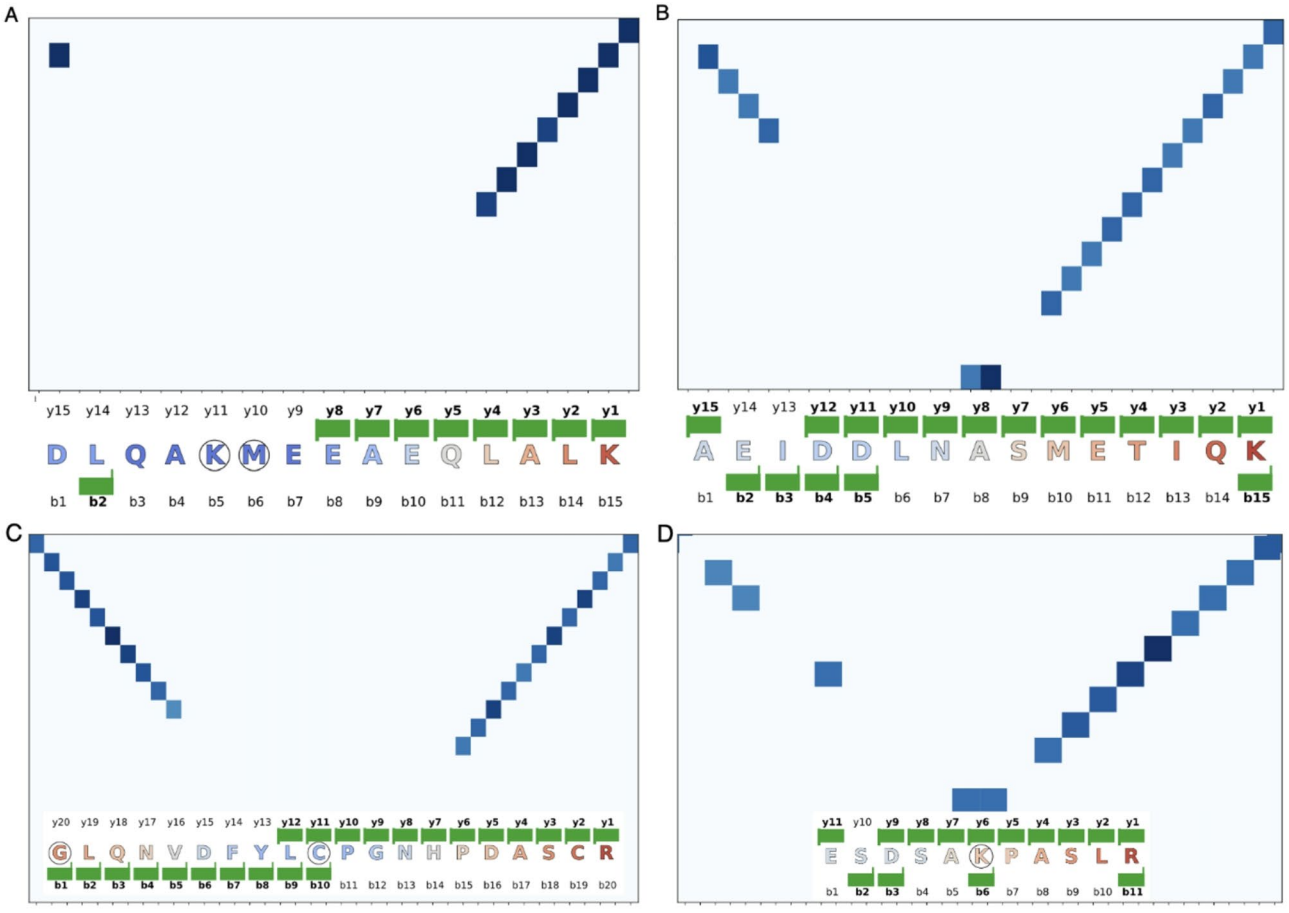


Fig. 4 PeptideAtlas PSM for *MYH16*, *ERVS71-1*, and *LINC03040*. **A** A section of the VSeq analysis of a PSM that is not supportive of MYH16 translation in heart tissues. It shows the intensity of the fragments matched in y-series only (above right) as squares, and the peptide sequence that is supported by the fragments (below) as green rectangles. The circles indicate the position of post-translational modifications. This supported C-terminal section of this peptide is also in seven known myosin proteins, while the MYH16-specific region (the N-terminal section) has no support. The full output of VSeq for this PSM can be seen in Additional figure S6. **B** A section of the VSeq analysis of a PSM supportive of MYH16 translation in U2OS cells showing just the intensity of the fragments matched in the b-series (above left) and y-series (above right) as squares, and the peptide sequence that is supported by the fragments (below) as green rectangles. The full output of VSeq for this PSM can be seen in Additional figure S7. **C** A section of the VSeq analysis of a PSM supportive of ERVS71-1 translation in U2OS cells showing just the intensity of the fragments matched in the b-series (above left) and y-series (above right) as squares, and the peptide sequence that is supported by the fragments (below) as green rectangles. The circles indicate the position of post-translational modifications. The full output of VSeq for this PSM can be seen in Additional figure S8. **D** A section of the VSeq analysis of a PSM supportive of LINC03040 translation in ubiquitinated Jurkat cells showing just the intensity of the fragments matched in the b-series (above left) and y-series (above right) as squares, and the peptide sequence that is supported by the fragments (below) as green rectangles. The circles indicate the position of post-translational modifications. The full output of VSeq for this PSM can be seen in Additional figure S9

is based on a single [primate] sequence and not on any conservation signal. *LINC03040* has multiple distinguishing peptides in PeptideAtlas, so it is clearly translated (Fig. 4D) though almost all *LINC03040* peptides were detected solely in HLA-I proteomics experiments [52]. Even in the most recent version of PeptideAtlas (2025) only four peptides with 12 observations are fully tryptic. Eleven observations are found in cancer cell lines and/or HLA-1 experiments (Fig. 4D) and although one PSM was observed in blood cells, the spectrum does not support the peptide.

LINC03040 is upregulated in colon cancer [53, 54] and there is considerable evidence for transcript expression in TCGA colon cancer samples [55] and none in normal tissue. High levels of *LINC03040* expression are linked directly to much poorer survival rates [54]. It was recently shown that the hypermethylation of *LINC03040* plays an important role in colon carcinoma [56]. Experimental evidence and the lack of conservation points to *LINC03040* being a cancer antigen, so the translation of *LINC03040* is most likely to be aberrant.

Finally, there are three SDPs in PeptideAtlas for UniProtKB entry A6NNC1, which is tagged as gene *POM121L1P*. *POM121L1P* is a bizarre case. There are multiple POM121L1P-like ORFs in the genome on chromosome 5, and on chromosome 6 and 22. None of the

possible ORFs have any conservation support. The UniProtKB entry does not map to any of the products of the various potential *POM121L1P* ORFs in the human genome. The few *POM121L1P* references that exist are all for the chromosome 22 pseudogene.

All the potential *POM121L1P* proteins are made up of three or four long repeats of a sequence derived from the parent gene, *POM121*. It is not clear how many unbroken *POM121L1P* ORFs exist, since it is not clear what constitutes an unbroken ORF - all *POM121L1P* ORFs have multiple large indels. Although most *POM121L1P* peptides are detected in cancer cell lines, some are found in sperm, so at least one of the *POM121L1P* genes may have gained a functional role. GENCODE has annotated nine of the *POM121L1P* ORFs on chromosome 5 as coding genes based on the PeptideAtlas peptide evidence from a single THISP entry. Just two of these nine novel genes have all the SDPs detected for this gene.

The 35 potential coding genes that were identified with PeptideAtlas SDPs peptides can be divided into three groups (Fig. 5A), those that are most likely to be produced from *bona fide* protein coding genes (16 genes), those that appear to be translated only in the dysregulated conditions typical of cancer cells (aberrant translations, 14) and those that we cannot be mapped to the genome for one reason or other (unplaced ORFs, 5). Of the 10 genes already annotated as coding in GENCODE v49 based on PeptideAtlas data (Table 1), we believe that seven are coding genes, that two (*MYH16* and *LINC03040*) are likely to produced only by aberrant translation and that one (*POM121L1P1*) cannot be correctly located in the genome.

Classifying the remaining 25 potential coding genes

As well as the seven coding genes that have been added to the GENCODE v49 reference set, we believe that another nine predicted genes with peptide support are likely to be coding (Table 2). These genes are *ANKRD26P1*, *EP400P1*, *ERVFRD-2*, *GPRIN2-1*, *LIPT2-AS1*, *LNCPRESS1*, *LOC107986768*, *PLAC4* and *WASHC1*. Six of these genes, *ERVFRD-2*, *GPRIN2-1*, *LIPT2-AS1*, *LNCPRESS1*, *PLAC4* and *WASHC1*, were not previously part of any reference proteome. All 9 genes are either repurposed transposon-derived genes (*ERVFRD-2*, *LNCPRESS1*, *LIPT2-AS1*, *LOC107986768*, *PLAC4*) or were generated by gene duplication. *ANKRD26P1* and *EP400P1 are* truncated versions of their parent genes.

Genes *WASHC1* and *GPRIN2-1* are certainly protein coding genes but can only be annotated in the CHM13 assembly [9]. Beyond these two genes, one of the clearest examples is *LIPT2-AS1*, currently a lncRNA (long non-coding RNA) in RefSeq. The first half of the protein would have 40% identity to the *JRK* gene, which came from a Tigger transposon domestication event [57]. The *LIPT2-AS1* protein does not have the C-terminal DDE endonuclease domain of the *JRK* protein. The 470 mammal alignments from the UCSC browser [32] show clear protein coding conservation right across primates for *LIPT2-AS1* (see Additional figure S11), and UniProtKB annotates highly similar proteins for rabbit, naked mole rat, beaver, green anole and bamboo shark. However, these genes do have a DDE endonuclease domain and given the retroviral origin of *LIPT2-AS1*, it may be that co-option has occurred more than once.

We found 3 SDPs for *LIPT2-AS1* in the 2023 build of PeptideAtlas (Fig. 6B), and there are five in the 2024 build. Most of the PSM are from cancer cell line experiments, but there are PSM in frontal cortex tissues and this is further supported by the transcriptomics support for *LIPT2-AS1* in brain tissues. The gene model predicted by the UniProtKB protein Q8TB74 is truncated because we find another ATG 51 codons upstream that would complete a CENPB N-terminal DNA-binding domain.

PeptideAtlas supports the co-option of gag ORFs in placenta

Remarkably, 3 of the 9 coding genes not yet annotated by GENCODE are derived from endogenous retrovirus (ERV) *gag* genes and the PeptideAtlas peptides are mostly or entirely found in placental tissue. The longest of the three *gag* ORFs, known variously as *ERVFRD-2* or *gagV1*, is part of a cluster of ERV-derived genes in a zinc finger gene rich region of chromosome 19 (Fig. 6A). Despite the name, it is not related to *ERVFRD-1*. The retrovirus that gave rise to *ERVFRD-2* was incorporated into the ancestor of the simian primates at least 43 million years ago [58] and is flanked by a predicted upstream coding ORF, pre-gagV1 [58], and two copies of the *env* ORF, *ERVV-1* and *ERVV-2* (Fig. 6A). Monkeys have a second, almost identical, *gag* ORF (*gagV3*) [59] between *ERVV-1* and *ERVV-2*, though this gene has been lost in apes.

Both *ERVV-1* and *ERVV-2* are expressed in placenta, but the similarity between the two genes mean that it is difficult to know which is functional. Evidence for gene conversion between the two only complicates matters [59]. Only *ERVV-2* has distinguishing peptides in PeptideAtlas, and conservation evidence suggests that *ERVV-1* may have pseudogenised (the chimpanzee ORF has an early stop codon and the human ORF is shortened by a frameshift that leads to a different premature stop). Despite this, there is supporting evidence for a role in placenta for *ERVV-1* [60]. Both are highly expressed in placenta [61].

We calculated dN/dS ratios for the four predicted ERV ORFs. Three, *ERVFRD-2* (0.385), *ERVV-1* (0.384) and *ERVV-2* (0.352), all had dN/dS ratios significantly lower than one with all p-values below 2×10^{-7}(see Additional table S3). Clearly, all three have been under purifying

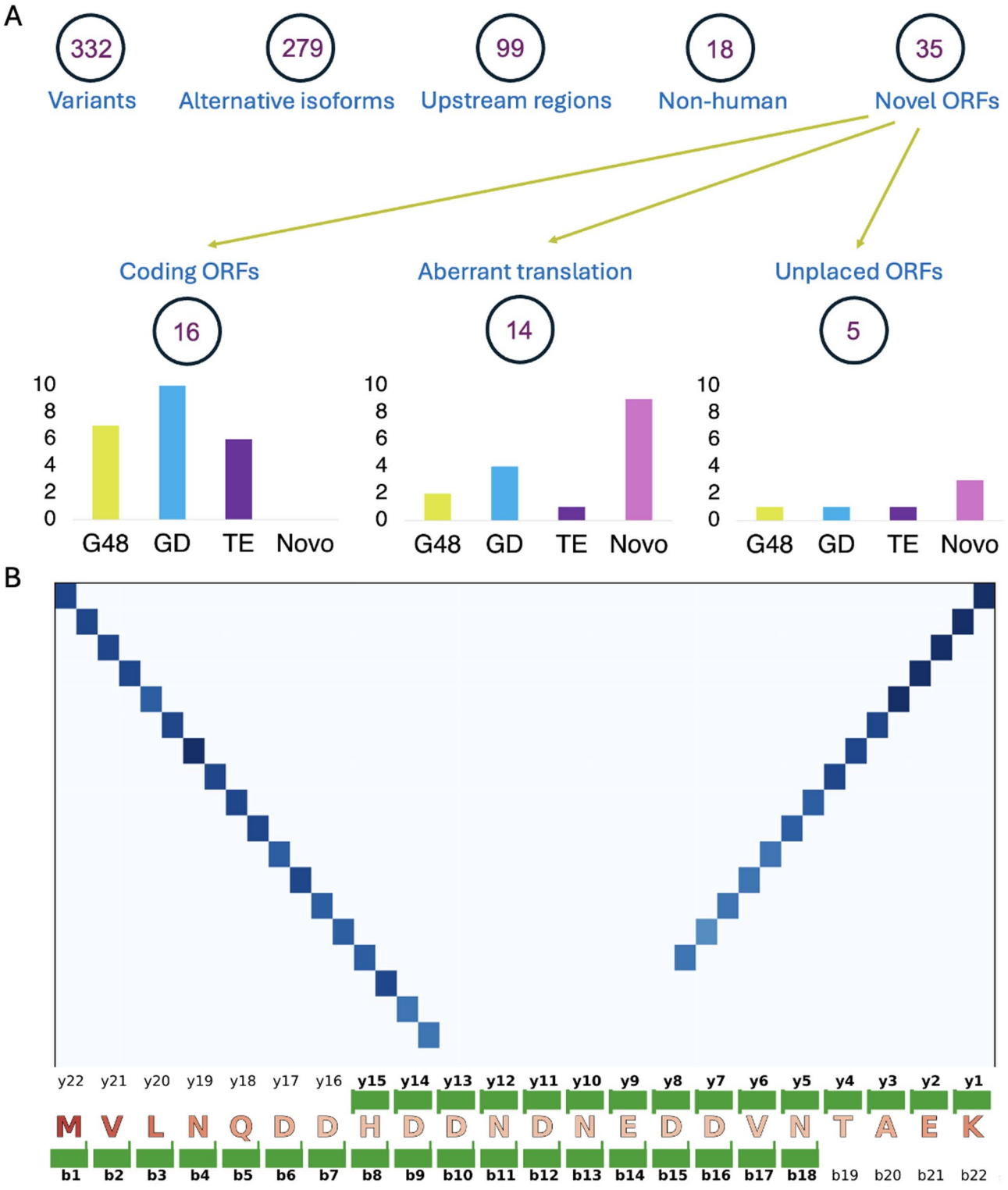


Fig. 5 The novel coding ORFs in PeptideAtlas and a supporting PSM for *LIPT2-AS1*. **A** Overview of the analysis of the 35 potential unannotated coding ORFs that were detected in the PeptideAtlas database. For the genes in each of the three sections we show how many are already annotated as coding by GENCODE (G49), how many arose by gene duplication (GD), how many were co-opted from transposable elements ORFs (TE) and how many appear to have evolved de novo (Novo). **B** A section of the VSeq analysis of a PSM supportive of LIPT2-AS1 translation showing the intensity of the fragments matched in the b-series (above left) and y-series (above right) as squares, and the peptide sequence that is supported by the fragments (below) as green rectangles. The full output of VSeq for this PSM can be seen in Additional figure S10

Table 2 Predicted possible coding genes from the PeptideAtlas analysis

Gene	SDPs	UniProtKB	Entry	Derived from
ANKRD26P1	9	Proteome	Q6NSI1	Duplication
EP400P1	6	Proteome	Q6ZTU2	Duplication
ERVFRD-2	3	TrEMBL	Q6ZRZ8	Retrovirus
GPRIN2-1	3	-	-	Duplication
LIPT2-AS1	3	TrEMBL	Q8TB74	Transposon
LNCPRESS1	2	-	-	Transposon
LOC107986768	4	-	-	Retrovirus
PLAC4	3	TrEMBL	Q8NAM4	Retrovirus
WASHC1	10	-	-	Duplication

The 9 genes supported by PeptideAtlas that are not yet in the Ensembl/GENCODE reference set and that we believe are likely to be protein coding

selection. For *ERVFRD-2*, the low dN/dS ratio and the placenta-specific expression confirms that the ORF has been co-opted to a placental-specific role. If the *ERVFRD-2* product is not post-processed, it would produce a protein with four structural domains (Fig. 6B).

ERVFRD-2 has been claimed as the oldest intact co-opted *gag* gene in the human genome [58], although we believe that another of the three placental *gag* genes should have at least a share of that title. The *gag* ORF-derived *PLAC4* gene on chromosome 7 is much shorter and would produce a protein with only the N-terminal matrix domain (Fig. 6B), but it is just as conserved. *PLAC4* cDNA was first discovered in placenta tissues more than 30 years ago [62]. As with *ERVFRD-2*, *PLAC4* was incorporated into the ancestor of simians and is clearly under purifying selection across simians with a dN/dS of 0.272 and a q-value of 6.39×10^{-7} (see Additional table S3 and Additional figure S12).

The third *gag* ORF, RefSeq predicted ORF *LOC107986768*, is a more recent evolutionary innovation and is only entirely conserved in apes. It is on the opposite strand from the *SCIN* coding gene on chromosome 7. In the 2024 build of PeptideAtlas there are 11 peptides for this gene (5 strong and 6 non-tryptic); all peptides are restricted to placenta. Despite the recent innovation, evidence from the UCSC 447-way vertebrate alignments [32], which include more ape species than the 470-way vertebrate alignments, indicates that *LOC107986768* is also likely to be under purifying selection (dN/dS = 0.61, *p*-value = 0.031).

Curiously, all peptides detected for *LOC107986768* map between residues 140 and 270, which corresponds to the second of the two protein domains of the *gag* protein (the capsid N-terminal domain) and part of the long linker (Fig. 6B). The region with peptide support is also undisturbed across old world monkeys.

Peptide support for aberrant translation

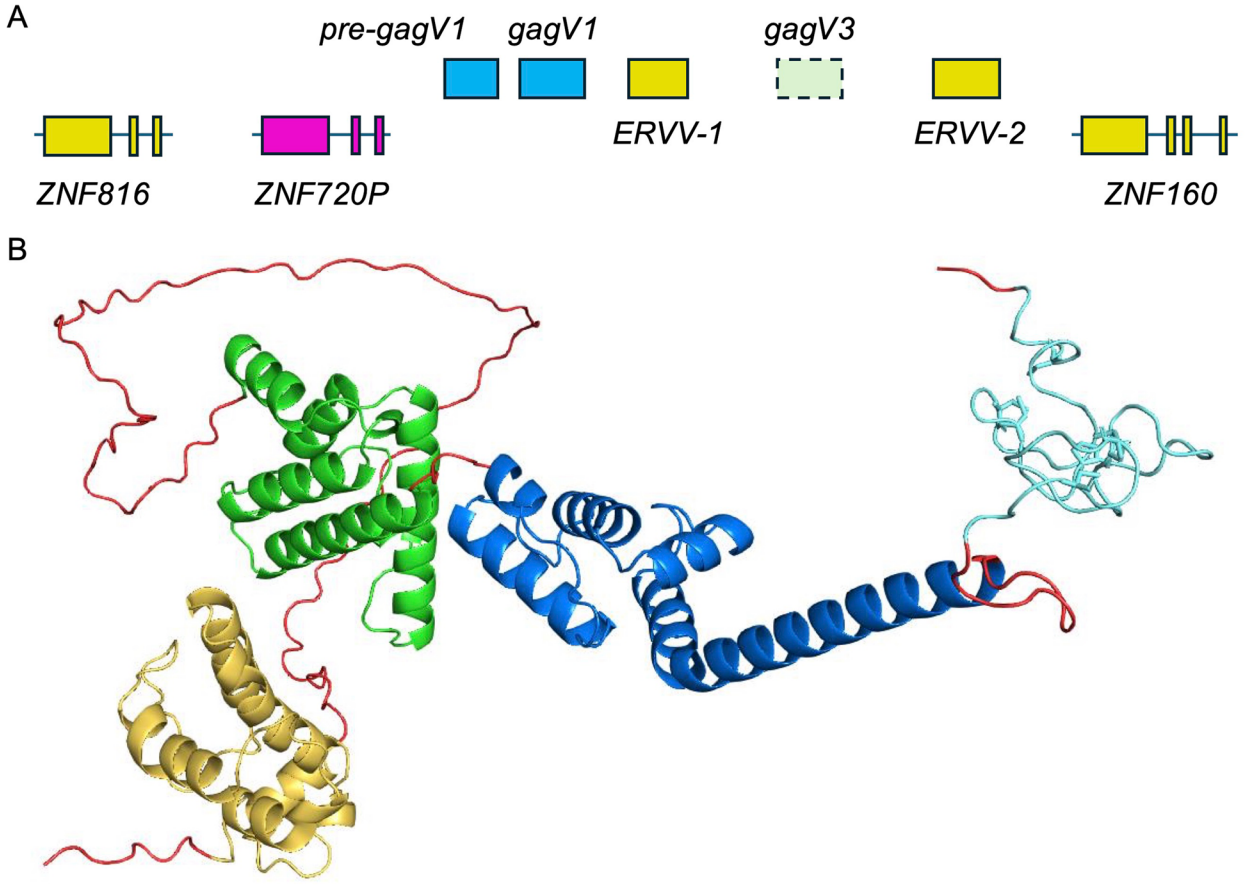
Twelve of the 25 potential novel coding genes had conspicuously poor cross-species conservation. Eleven did not have any evidence of cross species conservation at all, while orthologous regions of the reading frame of the *NECTIN3-AS1* ORF are only unbroken in chimpanzees. Despite this, many had substantial peptide support for their translation. *LINC00839* had 10 supporting SDPs in PeptideAtlas, *ZNF252P* and *LOC124900476* both had 9, *DOCK8-AS1* had 8, and *TXLNGY* had 5. These 12 THISP database entries are listed in Additional table S4.

Non-conserved, *de novo* ORFs such as these almost certainly require further corroboration beyond the peptide support if they are to be annotated as coding genes. This support could include experimental evidence [50], tissue-specific expression or evidence of selection from human genetic variants. However, when we analysed the supporting evidence for these 12 potential genes, we found that the experimental evidence and tissue expression suggested that most were cancer related.

Two examples are *LINC00839* and *ZNF252P*. The reading frame of the *LINC00839* gene is not conserved outside of humans because primate species have a whole spectrum of different premature stop codons. There are two predicted alternatively spliced isoforms annotated for this gene in UniProtKB Trembl and both have peptide support. One of the two isoforms has a long hydrophobic tail, so would be expected to be degraded via the *BAG6* pathway under normal conditions [25], while the other would be a NMD candidate.

Although GTEx shows that *LINC00839* transcripts are expressed mostly in nervous tissues, the peptides for both isoforms are detected almost exclusively in a single sumoylation experiment carried out on the U2OS cancer cell line [63]. Protein sumoylation plays a role in degradation [64]. Peptides are also detected in an ovarian cancer cell line. Abnormal expression of *LINC00839* has been found in a range of distinct tumors and cell lines and increased levels of *LINC00839* have been directly related to adverse clinical outcomes [65]. Although it will never be possible to prove that *LINC00839* is only ever translated in cancer cells, the *LINC00839* isoforms appear most likely to be the result of aberrant translation.

The PeptideAtlas peptides that map exclusively to *ZNF252P* isoforms come from the same U2OS sumoylation experiments as *LINC00839* [62]. Like *LINC00839*, *ZNF252P* also has two isoforms, both of which have peptides. Although there are signs of coding conservation of this gene in more distant primates, *ZNF252P* appears to be a pseudogene among great apes because of premature stops and frameshifts. Human, chimpanzee and gorilla transcripts are truncated. *ZNF252P* is upregulated in a variety of cancer types, and its overexpression improves cancer cell viability, increases proliferation and boosts tumorigenicity in vivo via a positive feedback loop with *MYC* [66]. Again, given



C

```
LOC107986768   MGASQSTPPKTTPLGCLLRNRKALGLRSGIRSKRLIFYCNTAWPQYQLDNGSQWPENGTFDFNILRDLDNFCHRNEKWSEIPYIQAFFIL
ERVFRD-2       MGNQSSV-PQDSPLGCILRNWDKFDPQ-ALKRKRLVFLCNTVWPKYELEGQEAWPVGGSLNVNTILQLDVYCRQQRKWSEVPYVQTFMIL
PLAC4          MGTTRSI-PRDSPLGCLLSYWSKFKF--GLKKKKLILYCNTIWVKYKLDDEYIWPKHGSIYYSDILRLGLFCKREGKWEEIPYVQAFMAL
               **.  *  *:.:****:***   :  : .:: *:*:: ***.**:*:*:. . **  *::: * :  ** :*::::.****:**:*:*:**

LOC107986768   ---RNHPSLCHSCSTFQILLTSSKPDSSPVTPPTDPADDSSSFDPVDFLPPRQHHDEPPPHHDPPPYVPAPALPLSPTLSNQPTSDSESS
ERVFRD-2       ---RENPDFCKGCKIDPALLAILSRPLQRPQPGG--------FNDFLVNPPQP---PLPETKEKEQAPPAPSS-LYRTLSLH---GSAST
PLAC4          YWITLLPGIRMPCIRDPHLAAPHRKFISLPRVSSVPQS-------------------------------------------------
                  *. *.:*:.*     **.          * .        *:     **:    * ** ::     ***.  *. ***     .* *:

LOC107986768   LPPPLTRSRAQCAQQPAPLLPLREVAGVEGIVHVHVPFSFYDLLQIEERLGSFSSDPDTYIKEFKYLTQSYELTWHDLYFILSSTLLPEE
ERVFRD-2       YTRPSGPR--------SGICLLPVVSRPVGPVQVQVPFSMQDLSQVKEGLGKFSENPGKFLEGFRKLTLTFELTWKDVAILLGQTLSLEE
                  * .          . :  *  *.   * *. . ***: ** *::* ** **.:*. *:: *: ** ::****.*: .:*. **  **

LOC107986768   KERVWLAAQAYADDLHQQDPTKPVGAAAVPWEEPSWEYQPTDPGQVYCSHMITCLIAGLNKAAHRL-----------------------
ERVFRD-2       RQTIWEAARQCGDELHLADANYPVGATAVPLQDPNWDYDTP-AGICARNHMLLCLIEGMKRSQVKPVNYNKLATIDQGPHENPTAFLERL
               :: :* **:  .*:**. * . ****:*** ::*.*:*.    *    .**: *** *. :.  :

ERVFRD-2       QETLIKHTNLDPGSPEGQLVLKDHFLTQAAPDIRRKLRMLALGTRAPMSEILKFASSVFYNRDQDERDRAERKEKQKEERQAQLLAALQV

ERVFRD-2       HQPPPGCPKDTFPGNCYQCGKPGHWKANCPYGPRGEKPCTACPLCRKLRYWKENCPESQKGP
```

Fig. 6 Placenta-expressed HERV-V gag genes. **A** A schematic representation of the section of chromosome 19 around the two inserted HERV-V virus. Annotated coding genes (including the two HERV-V env genes, *ERVV1* and *ERVV2*) are shown in yellow, pseudogenes in pink. Predicted retroviral genes in blue, position of lost great ape HERV-V *gag* gene in green. **B** The AlphaFold model structure of the product of the *ERVFRD-2/gagV1 gag* gene. The colours correspond to the regions in the alignment. **C** An alignment between the predicted products of the three placenta-expressed gag genes, *LOC107986768*, *ERVFRD-2* and *PLAC4*. The colours correspond to the structure in section B, the pink residues show the conserved N-terminal myristoylation motif. Detected peptides for *LOC107986768* are underlined

the results, the peptide evidence for *ZNF252P* is likely to be the result of aberrant translation.

Unplaced UniProtKB entries with peptide support

Finally, four entities with peptides in PeptideAtlas defy classification because like entry A6NNC1 (*POM121L1P*) none are placeable in the genome (Additional table S5). Q8NHH4 (tagged as hepatocellular carcinoma-associated antigen 25a, HCA25a) is supposedly a cancer antigen. Only the central region identified by peptides maps to the genome, but it is in multiple locations (14 regions in total) across the genome. This region of the HCA25a sequence is highly similar to the theoretical protein sequences of oesophageal carcinoma antigens [68]. In all likelihood, whichever ORF is producing these peptides is undergoing aberrant translation, though there are peptides in normal lung and liver tissues.

The Q9UF83 entry is supported by 20 SDPs. The predicted protein sequence of Q9UF83 is made up almost entirely of degenerate nine amino acid repeats. All 20 peptides were detected in cancer cell lines. However, there is no region in the human genome that could produce the Q9UF83 protein. The most similar protein would come from a single exon in the heterochromatin-rich p arm of chromosome 13 and would have no more than 93% identity to Q9UF83. This single exon ORF is annotated as coding in a patch (chr13_KN538372v1_fix, ENSG00000283864) though not in the main reference set. UniProtKB entry B4DYQ5 is distantly related to Q9UF83 and like Q9UF83 it does not correspond to any genomic sequence. The nearest equivalent is *FAM230D*, one of the Q9UF83-related lncRNA on chromosome 22 annotated by RefSeq. Even here much of the predicted protein sequence of B4DYQ5 would be missing. *FAM230D* is expressed in testis and peptides that map to B4DYQ5 are found in testis and cancer cell lines.

UniProtKB entry V9H0H3 is the product of a human-specific HERV-K (HML2) virus. It appears to be translated entirely or almost entirely in cancer cell lines. However, there are multiple copies of HERV-K viruses in the human genome, and it is not clear which of the copies PeptideAtlas detects peptides for. LINE-1 ORF1 could conceivably be added to the list of unplaced ORFs. LINE-1 ORF1 is present in multiple copies in the genome. It is also the unannotated ORF with most peptide evidence in PeptideAtlas (see Additional file 1), and again almost all peptide-spectrum matches are from cancer cells or cell lines.

Peptide evidence for novel ribosome profile supported coding genes

Finally, we asked whether scaling up the proteomics search space to the 2,416 proteomics experiments had turned up more support for the novel ORFs from the large-scale ribosome profiling experiments [19, 20]. There were 20,337 unique entries in the two large-scale ribosome profiling experiments we analysed, though only a tenth of these entries fit our description of novel ORFs.

Most other entries were associated with known coding genes, for example, alternative transcripts, upstream translation initiation (uORFs, uoORFs and 5' coding extensions), downstream translation initiation (downstream ATGs or non-canonical start codons). Entries associated with known coding genes were not analysed in depth in this analysis, but we did find substantial peptide evidence for unannotated alternative isoforms and for translation from upstream start sites.

We broke down the 20,337 ribosome profile supported ORFs into 4 groups based on the definition in the two analyses: upstream translation initiation, downstream translation initiation, alternative transcripts, and potential novel coding genes. The first three groups of novel ORFs all overlapped coding or UTR exons of known coding genes. Almost 40% of the ribosome profiling novel ORFs had a start codon upstream of known coding genes, 21% were alternative transcripts of known genes and just 9.3% did not overlap known coding genes (Additional figure S13).

Our analysis would only have been able to detect peptides for 940 of the 1,981 potential novel coding genes because just 989 of the translations were in the PeptideAtlas THISP database and 49 of them were already annotated as coding in GENCODE v45. We found multiple peptides in PeptideAtlas for seven of these 940 THISP entries (0.74%). *MSL3P1*, *TSPY26P*, *TXLNGY*, *ZNF252P* and *LNCPRESS1* were detected by Chen et al. [19], while *LIPT2-AS1 and ARRDC1-AS1* were reported by Van Heesch et al. [20].

Another 7 potential novel coding genes from the Chen and Van Heesch analyses had the support of just one SDP. Curiously, for six of these genes all PeptideAtlas peptides came from HLA-I experiments. Observations for the seventh (*LINC00526*) were only found in cancer tissue.

Discussion

We found strong peptide evidence for 35 potential new coding genes in PeptideAtlas. We believe that 16 of these are *bona fide* coding genes, but that the peptides we found were the result of aberrant translation in at least 14 cases.

Six of the 16 coding genes supported by PeptideAtlas data appear to be transposon-derived and three are co-opted retroviral *gag* ORFs that appear to have gained important roles in placental tissues. We have shown that *PLAC4* and *ERVFRD-2* were inserted into the genome in the ancestor of simians and are clearly evolving under negative selection (dN/dS of 0.272 and 0.385, and q-values of 6.39×10^{-7} and 3.55×10^{-9} respectively), while the

third ERV *gag* ORF, *LOC107986768*, is conserved among apes and also appears to be under selective pressure.

There are well-known cases of retroviral genes in human placenta in the literature, but as yet only for ERV-derived *env* genes such as *ERVW-1* (Syncytin 1) and *ERV-FRD-1* (Syncytin 2) which mediate membrane fusion during the formation of placental syncytiotrophoblasts [67]. There are at least eight primate-derived *env* genes thought to play roles in syncytiotrophoblast formation [60], including *ERVV-1* and *ERVV-2* that are adjacent to *ERVFRD-2* on chromosome 19 [58]. Retroviral *gag* genes have also been shown to have placental roles in other species [68]– [69], and both *PLAC4* and *ERVFRD-2* can trace their origins back to simians and therefore predate the most well studied placental *env* proteins, Syncytin 1 and Syncytin 2. Both the mRNA and protein of *PLAC4* have also been shown to be located in the syncytiotrophoblast of placental villi [62, 70, 71].

There were also peptides for a more recently co-opted *env* ORF, *ERVS71-1*, which has been shown to have a role in preventing HERV-T infections [50]. The cases of transposon-derived coding genes reported here expand the known repertoire of functional co-option and suggest the need for further functional studies into the conservation of retroviral genes in the human genome.

While almost all of the 16 likely coding genes were under selection pressure, none of 14 regions that appear to be producing peptides from aberrant translation had conservation support. Twelve were not conserved beyond humans and the other two were not conserved beyond great apes. In addition, these ORFs have peptides detected only in HLA proteomics experiments (for example, *LINC03040*), only in cancer cell experiments (*ARRDC1-AS1*, *PMS2CL*), or only in degradation-related sumoylation experiments (*ZNF252P*). Ten are already described as cancer relevant genes, either known to be cancer markers (*DOCK8-AS1* [72], *NECTIN3-AS1* [73]), or to exert oncogenic effects (*MHENCR* [74]). *MNX1-AS1* exerts such wide effects on a range of cancers that it has its own review [75], and there are almost 30 publications relating *LINC00839* with a spectrum of human disorders, and in particular with promoting malignancy [65].

The examples of aberrant translation of ORFs that we show here present a whole new level of complexity to the annotation of protein coding genes. These regions produce peptides, so they cannot be ignored, but they have no experimental evidence of a function, nor the level of conservation evidence that could be used to infer functional importance. If these genes produce proteins, yet are not functionally important, how should they be annotated? Should they be added to the reference gene set as coding genes for the sake of completeness, as is currently the case with genes *HMHB1* [76] and *MYEOV* [77], or should they be labelled separately so that they cannot be mistaken for genes that produce biologically relevant proteins?

The remaining 5 proteins with multiple SDPs cannot be mapped to genomic coordinates. Peptides for these entries map to multiple regions in the human genome, so it is impossible to know which ORFs are being translated. For example, there are multiple distinct *POM121* pseudogene ORFs in the human genome, scattered mostly across chromosomes 5 and 22. LINE1 ORF1 could be added to this list (see Additional file 1).

Although there is considerable evidence for potential novel coding genes in large-scale ribosome profiling experiments, very few have been verified in proteomics analyses [19, 20, 22, 23]. We hypothesized that scaling up the proteomics search space to the 2,416 proteomics experiments deposited in the PeptideAtlas database, we might see substantially more proteomics evidence. The novel coding genes that we chose to search for were predicted in two large-scale ribosome profiling analyses [19, 20]. However, PeptideAtlas peptides only validated 7 of these potential novel coding genes.

The depth and breadth of the PeptideAtlas proteomics experiments (the range of tissues and conditions that these experiments cover) ought to have been sufficient to detect the translation of most ribosome profiling novel ORFs. Since PeptideAtlas had so little evidence for novel genes that have been shown to be undergoing translation by ribosome profiling analyses, it suggests that the novel isoforms are either produced in quantities that are too low to be detected in proteomics experiments or translated and quickly degraded.

None of the 35 genes that we validated with PeptideAtlas peptides and that were not annotated in Ensembl/GENCODE can be regarded as completely novel discoveries since PeptideAtlas entries have to have been already discovered to be included in the PeptideAtlas THISP search database. However, 32 were not annotated as coding in the RefSeq reference set either and 17 were not part of the UniProtKB human reference proteome. Out of the 16 genes we believe to be coding, seven were not previously part of any reference coding gene set.

Conclusions

We have carried out a comprehensive interrogation of the human build of PeptideAtlas to search for evidence of translation of open reading frames that are not annotated as coding in the Ensembl/GENCODE reference set. As well as 35 possible new coding genes, we also found peptide evidence for more than 250 novel alternative isoforms and almost 100 novel translated upstream regions. These coding regions are in the process of being analysed by GENCODE curators and so far, 10 of the possible

coding genes have been annotated as coding, along with 11 paralogues.

We carried out a manual curation of these 35 potential novel coding genes, including analysis of the PSMs, of the cross-species conservation, of the tissue expression patterns of the transcripts and the isoforms, and of published experimental evidence supporting function or disease-related expression. We believe that 16 are coding genes with functional roles in the cell.

Most of the 16 likely coding genes appear to be recent evolutionary innovations yet still appear to be under purifying selection. The identification of three co-opted *gag*-ORF derived genes in placenta suggests that the importance of ERV-derived genes in certain tissues may have been overlooked until now.

We believe that at least 14 of the 35 ORFs are probably producing peptides from aberrant translation. The evidence for translation of non-coding regions under dysregulated conditions will require annotators to redefine what constitutes a coding region and has important implications for the study of cancers and other degenerative diseases.

Finally, a quarter of the new coding genes that we validated were testis-expressed primate gene duplications and the same number were retrovirus-derived and detected in placenta or stem cells. While it remains to be seen how many novel ribosome profiling ORFs will turn out to be *bona fide* coding genes, our results do provide clues as to where best to look for these genes [78].

Abbreviations

DAAV	Peptides that had two or fewer differences from canonical peptides in GENCODE v45 (Double Amino Acid Variants)
dN/dS	Ratio of non-synonymous to synonymous substitutions
ERV	Endogenous retrovirus
G49	GENCODE v49
HERV	Human endogenous retrovirus
HLA	I-human leukocyte antigen class 1
LncRNA	Long non-coding ribonucleic acid
ORF	Open reading frame
PSM	Peptide-spectrum match
PTM post	Translational modification
SAAV	Single amino acid variant peptide
SDP	Strong discriminating peptides (see methods section for description)
TPP	Trans-Proteomic Pipeline
uORF	Upstream open reading frame
uoORF	Upstream overlapping open reading frame

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-12238-w.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements

This research was also supported by the National Human Genome Research Institute of the National Institutes of Health [grant U41 HG007234].

Authors' contributions

J.M.R and M.L.T. conceived of and designed the analysis. D.C.V. J.M.R., and M.L.T. were involved in the acquisition of the PeptideAtlas data. M.M. and M.L.T. carried out the manual curation of the PeptideAtlas data. F.A. and M.L.T. carried out the dN/dS analysis. M.L.T. carried out the analysis of coding potential. J.M.R., F.A. and M.L.T. wrote the original draft. A.L.G., E.C.A. and J.M.R. carried out the Vseq analysis for figures 2-5 and the additional material. F.A., D.C.V., J.M.R., M.L.T. and J.C.V. reviewed and edited the paper. All authors reviewed the manuscript.

Funding

This study was supported by competitive grants PID2021-122348NB-I00 and PID2024-155650NB-I00 funded by MICIU/AEI/ 10.13039/501100011033 and by "ERDF/EU", PLEC2022-009298, PLEC2022-009235 and EQC2021-007053-P funded by MICIU/AEI/10.13039/501100011033 and by "European Union NextGenerationEU/ PRTR", andS2022/BMD-7333-CM (INMUNOVAR-CM) funded by Comunidad de Madrid. The project leading to these results has received funding from "la Caixa" Foundation under the project code LCF/PR/HR22/52420019. The CNIC is supported by the Instituto de Salud Carlos III (ISCIII), the Ministerio de Ciencia, Innovación Y Universidades (MICIU) and the Pro CNIC Foundation), and is a Severo Ochoa Center of Excellence (grant CEX2020-001041-S funded by MICIU/AEI/10.13039/501100011033).

Data availability

Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 14 August 2025 / Accepted: 16 October 2025
Published online: 21 November 2025

References

1. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376:44–53.
2. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, et al. The complete sequence of a human Y chromosome. Nature. 2023;621:344–54.
3. Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Res. 2023;51:D942–9.
4. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res. 2023;51:D933–41.
5. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National center for biotechnology information in 2023. Nucleic Acids Res. 2023;51:D29–38.
6. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51:D523-31.
7. Abascal F, Juan D, Jungreis I, Kellis M, Martinez L, Rigau M, et al. Loose ends: almost one in five human genes still have unresolved coding status. Nucleic Acids Res. 2018;46:7070–84.
8. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang YC, et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 2018;19:208.
9. Cerdán-Vélez D, Tress ML. The T2T-CHM13 reference assembly uncovers essential WASH1 and GPRIN2 paralogues. Bioinformatics Adv. 2024;4:vbae029.
10. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022;40:994–9.
11. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.

12. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291:1304–51.
13. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, et al. The ensembl genome database project. Nucleic Acids Res. 2022;30:38–41.
14. Southan C. Has the yo-yo stopped? An assessment of human protein-coding gene number. Proteomics. 2004;4:1712–26.
15. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, et al. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci USA. 2007;104:19428–33.
16. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. PLoS Biol. 2009;7:e1000112.
17. Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet. 2014;23:5866–78.
18. Maquedano M, Cerdán-Vélez D, Tress ML. The state of the human coding gene catalogues. Database (Oxford). 2025;2025:baaf045.
19. Chen J, Brunner AD, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020;367:1140–6.
20. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The translational landscape of the human heart. Cell. 2019;178:242–e6029.
21. Prensner JR, Abelin JG, Kok LW, Clauser KR, Mudge JM, Ruiz-Orera J, et al. What can ribo-seq, immunopeptidomics, and proteomics tell us about the noncanonical proteome? Mol Cell Proteomics. 2023;22:100631.
22. Rodriguez JM, Abascal F, Cerdán-Vélez D, Martínez Gómez L, Vázquez J. Tress ML evidence for widespread translation of 5′ untranslated regions. Nucleic Acids Res. 2024;52:8112–26.
23. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, et al. The PeptideAtlas project. Nucleic Acids Res. 2006;34:D655–8.
24. Inada T, Beckmann R. Mechanisms of translation-coupled quality control. J Mol Biol. 2024;436:168496.
25. Kesner JS, Chen Z, Shi P, Aparicio AO, Murphy MR, Guo Y, et al. Noncoding translation mitigation. Nature. 2023;617:395–402.
26. Yewdell J, Antón LC, Bacik I, Schubert U, Snyder HL, Bennink JR. Generating MHC class i ligands from viral gene products. Immunol Rev. 1999;172:97–108.
27. Wang J, Han GZ. Frequent retroviral gene co-option during the evolution of vertebrates. Mol Biol Evol. 2020;37:3232–42.
28. Deutsch EW, Mendoza L, Shteynberg DD, Hoopmann MR, Sun Z, Eng J, et al. Trans-proteomic pipeline: robust mass spectrometry-based proteomics data analysis suite. J Proteome Res. 2023;22:615–24.
29. Deutsch EW, Sun Z, Campbell DS, Binz PA, Farrah T, Shteynberg DD, et al. Tiered human integrated sequence search databases for shotgun proteomics. J Proteome Res. 2016;15:4091–100.
30. Zahn-Zabal M, Michel PA, Gateau A, Nikitin F, Schaeffer M, Audot E, et al. The nextprot knowledgebase in 2020: data, tools and usability improvements. Nucleic Acids Res. 2020;48:D328–34.
31. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020;587:246–51.
32. Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, et al. The UCSC genome browser database: 2025 update. Nucleic Acids Res. 2025;53:D1243–9.
33. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. Bioinformatics. 2004;20:426–7.
34. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.
35. Yang Z, Nielsen R. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 1998;46:409–18.
36. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289–300.
38. Fedorova AD, Kiniry SJ, Andreev DE, Mudge JM, Baranov PV. Thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. Nat Commun. 2022;13:7910.
39. Rodriguez JM, Pozo F, Cerdán-Vélez D, Di Domenico T, Vázquez J, Tress ML. APPRIS: selecting functionally important isoforms. Nucleic Acids Res. 2022;50:D54-9.
40. Bradley RK, Anczuków O. RNA splicing dysregulation and the hallmarks of cancer. Nat Rev Cancer. 2023;23:135–55.
41. Mahé M, Rios-Fuller T, Katsara O, Schneider RJ. Non-canonical mRNA translation initiation in cell stress and cancer. NAR Cancer. 2024;6:zcae026.
42. GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318–30.
43. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022;50:D439–44.
44. Cogliati S, Calvo E, Loureiro M, Guaras AM, Nieto-Arellano R, Garcia-Poyatos C, et al. Mechanism of super-assembly of respiratory complexes III and IV. Nature. 2016;539:579–82.
45. Lee LA, Karabina A, Broadwell LJ, Leinwand LA. The ancient sarcomeric myosins found in specialized muscles. Skelet Muscle. 2019;9:7.
46. Stedman HH, Kozyak BW, Nelson A, Thesier DM, Su LT, Low DW, et al. Myosin gene mutation correlates with anatomical changes in the human lineage. Nature. 2004;428:415–8.
47. Branca RM, Orre LM, Johansson HJ, Granholm V, Huss M, Pérez-Bercoff Å, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nat Methods. 2014;11:59–62.
48. Zhang L, Liu J, Wang H, Xu Z, Wang Y, Chen Y, et al. MYH16 upregulation is associated with lung adenocarcinoma aggressiveness and immune infiltration. J Biochem Mol Toxicol. 2023;37:e23490.
49. de Parseval N, Lazar V, Casella JF, Benit L, Heidmann T. 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.
50. Blanco-Melo D, Gifford RJ, Bieniasz PD. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. Elife. 2017;6:e22519.
51. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, et al. The Pfam protein families database: embracing AI/ML. Nucleic Acids Res. 2025;53:D523–34.
52. Ely ZA, Kulstad ZJ, Gunaydin G, Addepalli S, Verzani EK, Casarrubios M, et al. Pancreatic cancer-restricted cryptic antigens are targets for T cell recognition. Science. 2025;388:eadk3487.
53. Xu Y, Zhao J, Dai X, Xie Y, Dong M. High expression of CDH3 predicts a good prognosis for colon adenocarcinoma patients. Exp Ther Med. 2019;18:841–7.
54. Xiong Z, Li W, Luo X, Lin Y, Huang W, Zhang S. Seven bacterial response-related genes are biomarkers for colon cancer. BMC Bioinformatics. 2023;24:103.
55. Kahles A, Lehmann KV, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, et al. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. Cancer Cell. 2018;34:211–e2246.
56. Bai H, Yan DS, Chen YL, Li QZ, Qi YC. Potential biomarkers: the hypomethylation of cg18949415 and cg22193385 sites in colon adenocarcinoma. Comp Biol Med. 2024;2024(169):107884.
57. Waldron R, Rodriguez MLAB, Williams JM, Ning Z, Ahmed A, Lindsay A, et al. JRK binds satellite III DNA and is necessary for the heat shock response. Cell Biol Int. 2024;48:1212–22.
58. Boso G, Fleck K, Carley S, Liu Q, Buckler-White A, Kozak CA. The oldest co-opted gag gene of a human endogenous retrovirus shows placenta-specific expression and is upregulated in diffuse large B-cell lymphomas. Mol Biol Evol. 2021;38:5453–71.
59. Kjeldbjerg AL, Villesen P, Aagaard L, Pedersen FS. Gene conversion and purifying selection of a placenta-specific ERV-V envelope gene during Simian evolution. BMC Evol Biol. 2008;8:266.
60. Roberts RM, Ezashi T, Schulz LC, Sugimoto J, Schust DJ, Khan T, et al. Syncytins expressed in human placental trophoblast. Placenta. 2021;113:8–14.
61. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015;347:1260419.
62. Kido S, Sakuragi N, Bronner MP, Sayegh R, Berger R, Patterson D, Strauss JF 3rd, et al. D21S418E identifies a cAMP-regulated gene located on chromosome 21q22.3 that is expressed in placental syncytiotrophoblast and choriocarcinoma cells. Genomics. 1993;17:256–9.
63. Hendriks IA, Lyon D, Young C, Jensen LJ, Vertegaal AC, Nielsen ML. Site-specific mapping of the human SUMO proteome reveals co-modification with phosphorylation. Nat Struct Mol Biol. 2017;24:325–36.
64. Gutierrez-Morton E, Wang Y. The role of sumoylation in biomolecular condensate dynamics and protein localization. Cell Insight. 2024;3:100199.

65. Hu Y, Hu Y, Lu X, Luo H, Chen Z. LINC00839 in human disorders: insights into its regulatory roles and clinical impact, with a special focus on cancer. J Cancer. 2024;15:2179–92.
66. Qu T, Zhang C, Lu X, Dai J, He X, Li W, et al. 8q24 derived ZNF252P promotes tumorigenesis by driving phase separation to activate c-Myc mediated feedback loop. Nat Commun. 2025;16:1986.
67. Kagaya A, Shimada H, Shiratori T, Kuboshima M, Nakashima-Fujita K, Yasuraoka M, et al. Identification of a novel SEREX antigen family, ECSA, in esophageal squamous cell carcinoma. Proteome Sci. 2011;9:31.
68. Naruse M, Ono R, Irie M, Nakamura K, Furuse T, Hino T, et al. Sirh7/Ldoc1 knockout mice exhibit placental P4 overproduction and delayed parturition. Development. 2014;141:4763–71.
69. Dini P, Carossino M, Balasuriya UBR, El-Sheikh Ali H, Loux SC, et al. Paternally expressed retrotransposon Gag-like 1 gene, RTL1, is one of the crucial elements for placental angiogenesis in horses†. Biol Reprod. 2021;104:1386–99.
70. Tsui NB, Wong CS, Chow KC, Lo ES, Cheng YK. Investigation of biological factors influencing the placental mRNA profile in maternal plasma. Prenat Diagn. 2014;34:251–8.
71. Tuohey L, Macintire K, Ye L, Palmer K, Skubisz M, Tong S, et al. PLAC4 is upregulated in severe early onset preeclampsia and upregulated with syncytialisation but not hypoxia. Placenta. 2013;34:256–60.
72. Shu X, Zhang Z, Yao ZY, Xing XL. Identification of five ferroptosis-related LncRNAs as novel prognosis and diagnosis signatures for renal cancer. Front Mol Biosci. 2022;8:763697.
73. Liu H, Chen C, Liu L, Wang Z. A four-lncrna risk signature for prognostic prediction of osteosarcoma. Front Genet. 2023;13:1081478.
74. Chen X, Dong H, Liu S, Yu L, Yan D, Yao X, et al. Long noncoding RNA MHENCR promotes melanoma progression via regulating miR-425/489-mediated PI3K-Akt pathway. Am J Transl Res. 2017;9:90–102.
75. Li T, Zhou S, Yang Y, Xu Y, Gong X, Cheng Y, et al. LncRNA MNX1-AS1: a novel oncogenic propellant in cancers. Biomed Pharmacother. 2022;149:112801.
76. Dolstra H, Fredrix H, Maas F, Coulie PG, Brasseur F, Mensink E, et al. A human minor histocompatibility antigen specific for B cell acute lymphoblastic leukemia. J Exp Med. 1999;189:301–8.
77. Tang R, Ji J, Ding J, Huang J, Gong B, Zhang X, et al. Overexpression of MYEOV predicting poor prognosis in patients with pancreatic ductal adenocarcinoma. Cell Cycle. 2020;19:1602–10.
78. Dopkins N, Singh B, Michael S, Zhang P, Marston JL, Fei T, et al. Ribosomal profiling of human endogenous retroviruses in healthy tissues. BMC Genomics. 2024;25:5.

Publisher's Note