A deep audit of the PeptideAtlas database uncovers evidence for unannotated coding genes and aberrant translation

Authors: Jose Manuel Rodriguez[1-2], Miguel Maquedano[3], Daniel Cerdan-Velez[3], Enrique Calvo[1-2], Jesús Vazquez[1-2], Michael L. Tress[3]

1. Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), 28029 Madrid, Spain
2. CIBER de Enfermedades Cardiovasculares (CIBERCV), 28029 Madrid, Spain
3. Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), 28029 Madrid, Spain

* Corresponding author: mtress@cnio.es

Abstract

The human genome has been the subject of intense scrutiny by experimental and manual curation projects for more than two decades. Novel coding genes have been proposed from large-scale RNASeq, ribosome profiling and proteomics experiments. Here we carry out an in-depth analysis of an entire proteomics database.

We analysed the proteins, peptides and spectra housed in the human build of the PeptideAtlas proteomics database to identify coding regions that are not yet annotated in the GENCODE reference gene set. We find support for hundreds of missing alternative protein isoforms and unannotated upstream translations, and evidence of cross-contamination from other species.

There was reliable peptide evidence for 34 novel unannotated open reading frames (ORFs) in PeptideAtlas. We find that almost half belong to coding genes that are missing from GENCODE and other reference sets. Most of the remaining ORFs were not conserved beyond human, however, and their peptide confirmation was restricted to cancer cell lines. We show that this is strong evidence for aberrant translation, raising important questions about the extent of aberrant translation and how these ORFs should be annotated in reference genomes.

Introduction

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 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] (v46) annotates 19,411 coding genes, though there were more than 22,000 annotated in the combination of the RefSeq, UniProtKB and Ensembl/GENCODE reference sets in 2018 [7].

Recently, several high-profile large-scale ribosome profiling analyses have been published that demonstrate evidence for tens of thousands of unannotated open reading frames (ORFs) in the human genome [18, 19]. A consortium has been formed that will investigate whether these regions are likely to code for proteins [10], with one of the early papers published by members of the consortium suggests that the more than 7,000 well-supported novel ORFs they find might expand coding gene numbers by 30% [20]. Such a large increase in protein coding genes would raise the question of what constitutes a coding gene.

The paper highlights the discovery of 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 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 all have obvious cross-species conservation. Seven can trace their ancestry back to lobe-finned fish at least, while the other two are conserved across all mammals. However, these 9 genes are not representative of the class of novel ORFs because most novel ORFs detected in ribosome profiling experiments have little evidence of cross-species conservation.

While it is certainly true that there will be other conserved novel coding ORFs beyond the nine cited examples, we have recently detected one that overlaps the *GRIN2A* gene that is conserved across eutherian mammals and that has peptide evidence [21] for example what links the large majority of the novel ORFs detected in ribosome profiling experiments is that they usually have little or no evidence of cross species conservation. Analyses of germline variants from other coding regions with little or no evolutionary track record shows that these are generally not under purifying selection [7, 21, 22-24] and that few to none of the variants found for these regions are pathogenic [25, 26]. Human variant data strongly suggests that very few non-conserved novel ORFs will be functionally relevant, although some recently evolved functionally important coding genes certainly do exist [27].

The proteomics evidence for novel ORFs is not strong. The same 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 [18], and although the other detected 541 peptides for hundreds of novel ORFs in two proteomics analyses, only five of the novel peptides (fewer than 1%) were found in both experiments [19]. The consortium themselves [10] found just 13 peptides for the novel ORFs in PeptideAtlas, with one peptide per coding gene and most of these peptides were not tryptic. The extent to which these novel ORFs produce stable proteins is not clear.

If these novel ORFs are being translated as the ribosome profiling experiments suggest, and these 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 are not amenable to detection in standard trypsin-based proteomics experiments [20], although this still leaves many thousands of undetected translated ORFs [21]. 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 [28]. If translated, it is also possible that many of these peptides are rapidly degraded [29]. There is certainly evidence for some degradation in the ribosome profiling-based analyses [18,19], since there are plenty of novel ORF peptides in proteasome-derived human leukocyte antigen class 1 (HLA-I) proteomics experiments [20, 30]. The final possibility is that the novel ORFs are translated rarely or in low quantities. If this was happening, a single proteomics analysis would be unlikely to find much evidence, but multiple large-scale proteomics experiments might be able to provide some support.

Here we carry out a manual analysis of the novel ORFs detected in the PeptideAtlas database [31]. PeptideAtlas is a database that maps annotated and predicted proteins to thousands of proteomics experiments. PeptideAtlas interrogates spectra from many varied proteomics experiments and its search database contains many predicted proteins that are not produced by genes in the reference databases. The combination of these three features (and manual curation possibilities) means that PeptideAtlas is a potential source of evidence for protein coding genes still missing from the reference annotations.

Instead of starting with a list of transcripts from ribosome profiling experiments, we downloaded all PeptideAtlas peptides with the aim of discovering protein coding regions that the Ensembl/GENCODE reference might have missed. We found evidence for hundreds of regions that are not annotated as coding in the GENCODE version of the human reference gene set. Although most of the non-reference peptides mapped to alternative variants of known proteins, there was convincing evidence for proteins from 34 novel coding ORFs, most of which have little cross species conservation. We found peptide evidence for seven of the short ORFs predicted by ribosome profiling experiments [18,19], and we believe that five of these represent coding ORFs that are novel to Ensembl/GENCODE. We predict that other two proteins are the result of aberrant translations.

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 by the PeptideAtlas Trans-Proteomic Pipeline (TPP, 32), a suite of locally installed tools. The TPP maps spectra from large and small-scale proteomics experiments to the PeptideAtlas human protein search database. 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, 33) is made up of sequences from the UniProtKB [6], NextProt [34], 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* [18] and van Heesch *et al* [19] were annotated in the January 2023 build of the THISP database.

The TPP validated 3,489,945 peptides from the January 2023 build and mapped these peptides to 62,245 protein entries. Peptides identified by the pipeline were mapped to a single entry in the PeptideAtlas file. For our analysis this had two drawbacks. Firstly, peptides might map to more than one protein entry, and many do because many of the entries in the THISP database have common protein sequences. Secondly, peptides that might have been produced from a single transcript are often mapped to two or more distinct entries (due again to the abundant regions of common protein sequences in the THISP database).

To detect which peptides mapped to proteins outside of the Ensembl/GENCODE reference set [3, 4], we remapped the 3,489,945 peptides validated by the Trans-Proteomic Pipeline to the GENCODE v43 reference set of proteins [Figure 1]. The GENCODE v43 reference set was downloaded from the GENCODE website. We excluded all peptides that were not fully tryptic. Peptides that mapped to GENCODE proteins were excluded from the remainder of the analysis. We also excluded decoy peptides, and peptides that mapped to common contaminants and to microbes.

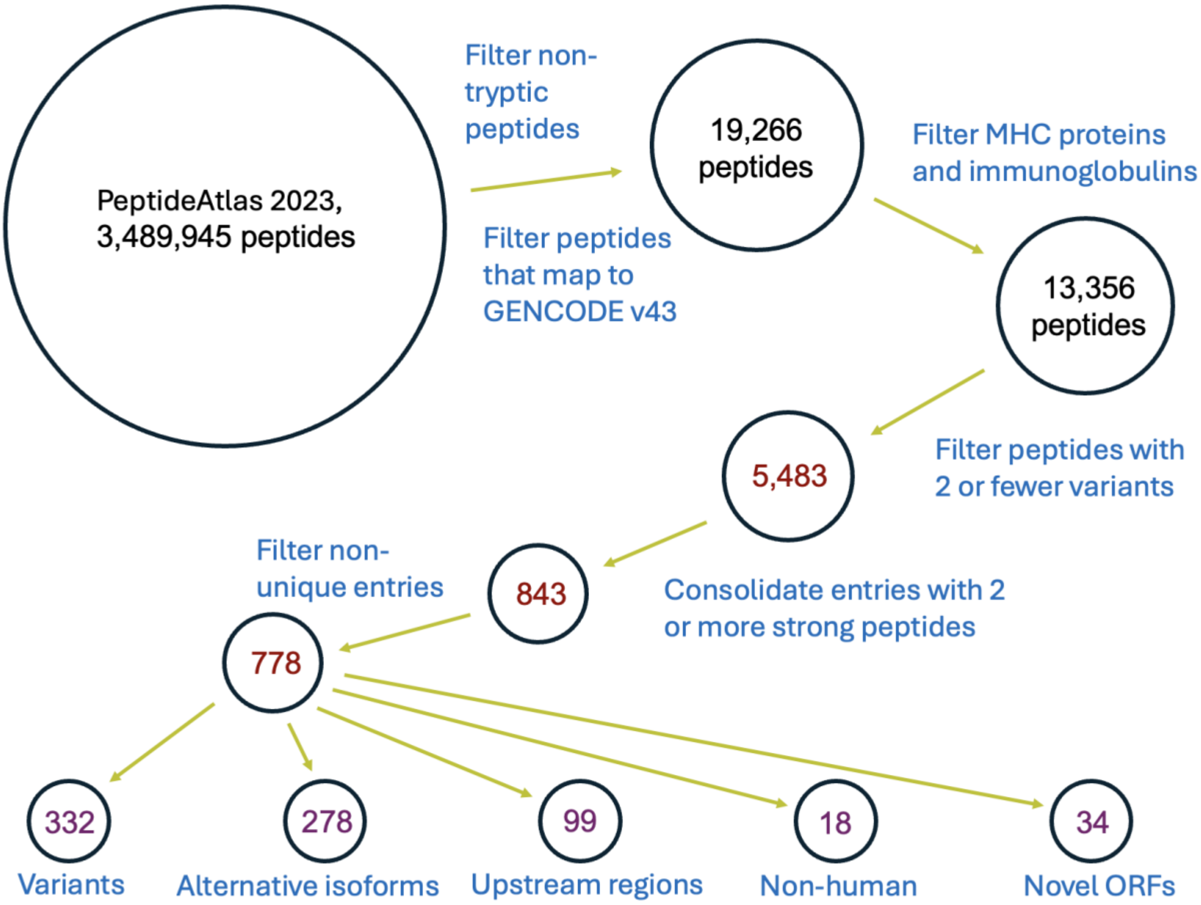

PeptideAtlas 2023,
3,489,945 peptides
Filter non-tryptic peptides
Filter peptides that map to GENCODE v43
19,266 peptides
Filter MHC proteins and immunoglobulins
13,356 peptides
Filter peptides with 2 or fewer variants
5,483
Consolidate entries with 2 or more strong peptides
843
Filter non-unique entries
778
332
Variants
278
Alternative isoforms
99
Upstream regions
18
Non-human
34
Novel ORFs

Figure 1. Workflow
The analysis of PeptideAtlas peptides that did not map to proteins in the Ensembl/GENCODE reference set and the five largest classes of "missing" proteins that were identified.

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

Defining strong peptides

We know that there are many probable pseudogenes in the THISP search database [**7**], and we know that many of the pseudogenes 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 theoretical peptides pseudogenes, largely because of single amino acid variations (SAAVs) and post-translational modifications [33]. Proteins often have many natural commonly occurring SAAVs. One way to get around these problems would be to just remove any peptide that was a SAAV of a GENCODE protein. However, this would make it impossible to uncover *bona fide* novel coding genes that are close homologues of other coding genes.

The downloaded list of PeptideAtlas peptides includes the number of times the peptide has been observed across the experiments interrogated in PeptideAtlas. The more observations, the more common the peptide. We made use of the number of observations for each peptide to score peptides that were just one or two amino acids different from peptides that mapped to GENCODE proteins. For each peptide we summed the observations for all peptides that differed by two or fewer amino acids. These we termed double amino acid variant (DAAV) peptides. The observation score for each peptide was then the number of observations for the peptide divided by the sum of observations from all DAAV peptides. We regarded peptides with observation scores above 0.9 as strong peptides. Peptides that had no DAAV peptides automatically had observation scores of 1. These strong peptides are referred to in the paper as strong discriminating peptides (SDPs).

Manual curation of the entries supported by strong discriminating peptides

We found 5,483 SDPs in PeptideAtlas that did not map to GENCODE reference set proteins. These peptides mapped to 3,774 distinct non-GENCODE entries. Most entries were supported by a single SDP, but 843 entries had two or more mapping SDPs [Figure 1].

We manually curated the novel peptides for the 843 entries that had two or more SDPs that did not map to the GENCODE 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 whole new ORF that was not annotated in GENCODE.

Since peptides from the same gene often map to more than one entry in the THISP database, we combined entries where the peptide data supported the same protein product. 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.

Results

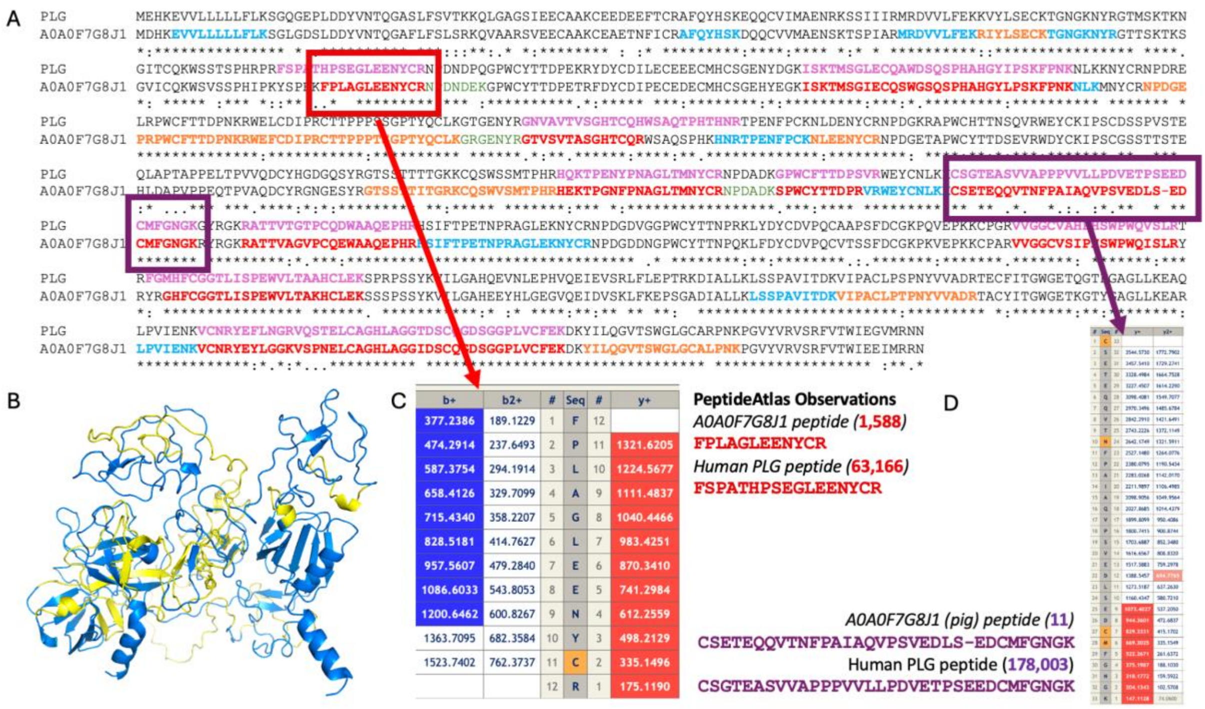
Most novel peptides are most likely to be variants, modifications or false positives

Most of the 13,356 novel peptides identified by PeptideAtlas were single or double amino acid variants of peptides from GENCODE proteins, so were likely to be either amino acid variants or post-translational modifications of known peptides. In this analysis we used the PeptideAtlas observation count to score all peptides and used only strong discriminating peptides (SDPs), those that were either substantially different from GENCODE peptides or that were similar GENCODE peptides but detected in much higher quantities (see methods section), to search for novel genes and coding regions.

Many apparently novel entries in PeptideAtlas are not novel

In total, we defined 5,483 of the 13,356 peptides as strong. We carried out an in-depth curation of those PeptideAtlas entries that had at least two SDPs, 778 entries in total (see methods section). We considered each of these novel protein coding ORFs on its merits. We found that many of the novel entries that had the support of SDPs were still most easily explained as variants or post-translational modifications of known proteins. This is largely because amino acid variants involving lysines or arginines can produce very different peptides due to trypsin cleavage. There were also peptides with insertions or deletions of one or more amino acids in low complexity regions that were most easily explained as variants.

Some entries were particularly complicated to decipher. For example, the novel protein from UniProtKB-Trembl, A0A0F7G8J1, is supported by 24 uniquely mapping PeptideAtlas peptides, which strongly suggests that it is produced from a novel ORF. This entry is a homologue of plasminogen (*PLG*) [Figure 2], to which it has 80.4% identity. There is considerable peptide evidence for the canonical plasminogen protein in PeptideAtlas, so it is conceivable (but unlikely) that all 24 peptides that map to A0A0F7G8J1 could be variants of plasminogen peptides.


A
PLG MEHKEVVLLLLLFLKSGQGEPLDDYVNTQGASLFSVTKKQLGAGSIEECAAKCEEDEEFTCRAFQYHSKEQQCVIMAENRKSSIIIRMRDVVLFEKKVYLSECKTGNGKNYRGTMSKTKN
A0A0F7G8J1 MDHKEVVLLLLLFLKSGLGDSLDDYVNTQGAFLFSLSRKQVAARSVEECAAKCEAETNFICRAFQYHSKDQQCVVMAENSKTSPIARMRDVVLFEKRIYLSECKTGNGKNYRGTTSKTKS
PLG GITCQKWSSTSPHRPRFSPATHPSEGLEENYCRNPDNDPQGPWCYTTDPEKRYDYCDILECEEECMHCSGENYDGKISKTMSGLECQAWDSQSPHAHGYIPSKFPNKNLKKNYCRNPDRE
A0A0F7G8J1 GVICQKWSVSSPHIPKYSPPKFPLAGLEENYCRNPDNDEKGPWCYTTDPETRFDYCDIPECEDECMHCSGEHYEGKISKTMSGIECQSWGSQSPHAHGYLPSKFPNKNLKMNYCRNPDGE
PLG LRPWCFTTDPNKRWELCDIPRCTTPPPSSGPTYQCLKGTGENYRGNVAVTVSGHTCQHWSAQTPHTHNRTPENFPCKNLDENYCRNPDGKRAPWCHTTNSQVRWEYCKIPSCDSSPVSTE
A0A0F7G8J1 PRPWCFTTDPNKRWEFCDIPRCTTPPPTSGPTYQCLKGRGENYRGTVSVTASGHTCQRWSAQSPHKHNRTPENFPCKNLEENYCRNPDGETAPWCYTTDSEVRWDYCKIPSCGSSTTSTE
PLG QLAPTAPPELTPVVQDCYHGDGQSYRGTSSTTTTGKKCQSWSSMTPHRHQKTPENYPNAGLTMNYCRNPDADKGPWCFTTDPSVRWEYCNLKKCSGTEASVVAPPPVVLLPDVETPSEED
A0A0F7G8J1 HLDAPVPPEQTPVAQDCYRGNGESYRGTSSTTITGRKCQSWVSMTPHRHEKTPGNFPNAGLTMNYCRNPDADKSPWCYTTDPRVRWEYCNLKKCSETEQQVTNFPAIAQVPSVEDLS-ED
PLG CMFGNGKGYRGKRATTVTGTPCQDWAAQEPHRHSIFTPETNPRAGLEKNYCRNPDGDVGGPWCYTTNPRKLYDYCDVPQCAAPSFDCGKPQVEPKKCPGRVVGGCVAHPHSWPWQVSLRT
A0A0F7G8J1 CMFGNGKRYRGKRATTVAGVPCQEWAAQEPHRHSIFTPETNPRAGLEKNYCRNPDGDDNGPWCYTTNPQKLFDYCDVPQCVTSSFDCGKPKVEPKKCPARVVGGCVSIPHSWPWQISLRY
PLG RFGMHFCGGTLISPEWVLTAAHCLEKSPRPSSYKVILGAHQEVNLEPHVQEIEVSRLFLEPTRKDIALLKLSSPAVITDKVIPACLPSPNYVVADRTECFITGWGETQGTFGAGLLKEAQ
A0A0F7G8J1 RYRGHFCGGTLISPEWVLTAKHCLEKSSSPSSYKVILGAHEEYHLGEGVQEIDVSKLFKEPSGADIALLKLSSPAVITDKVIPACLPTPNYVVADRTACYITGWGETKGTYGAGLLKEAR
PLG LPVIENKVCNRYEFLNGRVQSTELCAGHLAGGTDSCQGDSGGPLVCFEKDKYILQGVTSWGLGCARPNKPGVYVRVSRFVTWIEGVMRNN
A0A0F7G8J1 LPVIENKVCNRYEYLGGKVSPNELCAGHLAGGIDSCQGDSGGPLVCFEKDKYILQGVTSWGLGCALPNKPGVYVRVSRFVTWIEEIMRRN
B
C
PeptideAtlas Observations
A0A0F7G8J1 peptide (1,588)
FPLAGLEENYCR
Human PLG peptide (63,166)
FSPATHPSEGLEENYCR
D
A0A0F7G8J1 (pig) peptide (11)
CSETEQQVTNFPAIAQVPSVEDLS-EDCMFGNGK
Human PLG peptide (178,003)
CSGTEASVVAPPPVVLLPDVETPSEEDCMFGNGK

b+	b2+	#	Seq	#	y+
377.2386	189.1229	1	F	12	
474.2914	237.6493	2	P	11	1321.6205
587.3754	294.1914	3	L	10	1224.5677
658.4126	329.7099	4	A	9	1111.4837
715.4340	358.2207	5	G	8	1040.4466
828.5181	414.7627	6	L	7	983.4251
957.5607	479.2840	7	E	6	870.3410
1086.6033	543.8053	8	E	5	741.2984
1200.6462	600.8267	9	N	4	612.2559
1363.7095	682.3584	10	Y	3	498.2129
1523.7402	762.3737	11	C	2	335.1496
		12	R	1	175.1190

Figure 2. Analysing the SDPs mapping to a plasminogen homologue, A0A0F7G8J1

A. The alignment between A0A0F7G8J1 and human plasminogen with the PeptideAtlas peptides in bold and coloured by type. Red peptides are SDPs, orange peptides distinguishing peptides, but not SDPs and blue peptides are non-discriminating peptides. The tryptic peptides equivalent to the SDPs in plasminogen are marked in pink. B. The Alphafold [34] model of A0A0F7G8J1 with detected peptides marked in yellow. C. Part of the PeptideAtlas peptide spectrum match for the A0A0F7G8J1 peptide FPLAGLEENYCR showing that the y-series (in red) is a perfect match. D. Part of the peptide spectrum match for the A0A0F7G8J1 peptide CSETEQQVTNFPAIAQVPSVEDLSEDCMFGNGK showing that the y-series (in red) only matches the N-terminal residues that are identical in the equivalent *PLG* peptide.

Twelve of the 24 peptides are only one or two amino acids different from the plasminogen peptides that PeptideAtlas detected and have many fewer observations. But the other 12 discriminating peptides were sufficiently dissimilar from the equivalent plasminogen peptides to be classed as SDPs. If the 12 peptide-spectrum matches (PSMs) were good quality, this would be more than enough to support the presence of protein A0A0F7G8J1 because it is hard to imagine how amino acids variants and PTMs could explain so many large differences.

However, the PSM evidence is mixed. For example, the A0A0F7G8J1 peptide “FPLAGLEENYCR” has 1,588 observations, is very different from the equivalent peptide in plasminogen, and has very good looking PSM evidence [Figure 2]. This peptide supports a novel ORF for this protein. On the other hand, other peptides do not have such good PSM. The peptide “CSETEQQVTNFPAIAQVPSVEDLSEDCMFGNGK” has just 11 observations (compared to 178,003 for the equivalent plasminogen peptide) and PSMs that only supports the final 9 amino acid residues. These final 9 amino acid residues are identical between the A0A0F7G8J1 peptide and the equivalent plasminogen peptide, so these PSM do not supports the theory that the A0A0F7G8J1 protein is translated.

In fact, A0A0F7G8J1 is 99.4% identical to pig plasminogen (UniProtKB entry P06867) and does not map anywhere in the human genome. It is possible that pig plasminogen or something similar may be a contaminant that has not yet been included in the PeptideAtlas database.

A total of 322 entries were classified as most easily explained as amino acid variants or PTMs of canonical human proteins. Many of these entries are actin-like proteins. There are many actin-like entries in the PeptideAtlas THISP search database [33], partly because there are multiple actin pseudogenes in the human genome. Canonical actin proteins are highly expressed and highly conserved; seven of the ten entries with most PSM in PeptideAtlas are actins. Peptides that map to actin-like entries from the THISP search database have many fewer supporting PSM, and it is highly likely that the spectra that map to these actin-like entries belong to canonical actin peptides and are mis-mapped to the actin-like pseudogenes instead. The mis-mapping may be because of modifications or variants, or simply because of false positive match from a low-quality spectrum. Like actin, other genes and gene families, including haemoglobins, ribosome proteins and *GAPDH*, also have multiple erroneously identified pseudogenes.

Multiple non-human proteins detected in human proteomics experiments

HEL-109, a novel epididymis luminal protein, is identified with 13 uniquely mapping tryptic peptides in our analysis and in the January 2024 version of PeptideAtlas there are spectra for 24 distinct fully tryptic peptides. This novel 711 amino acid protein [Figure 3] is annotated by UniProtKB and appears to be a homologue of tropomyosin; the C-terminal region has more than 60% identity to one of the human *TPM3* isoforms.

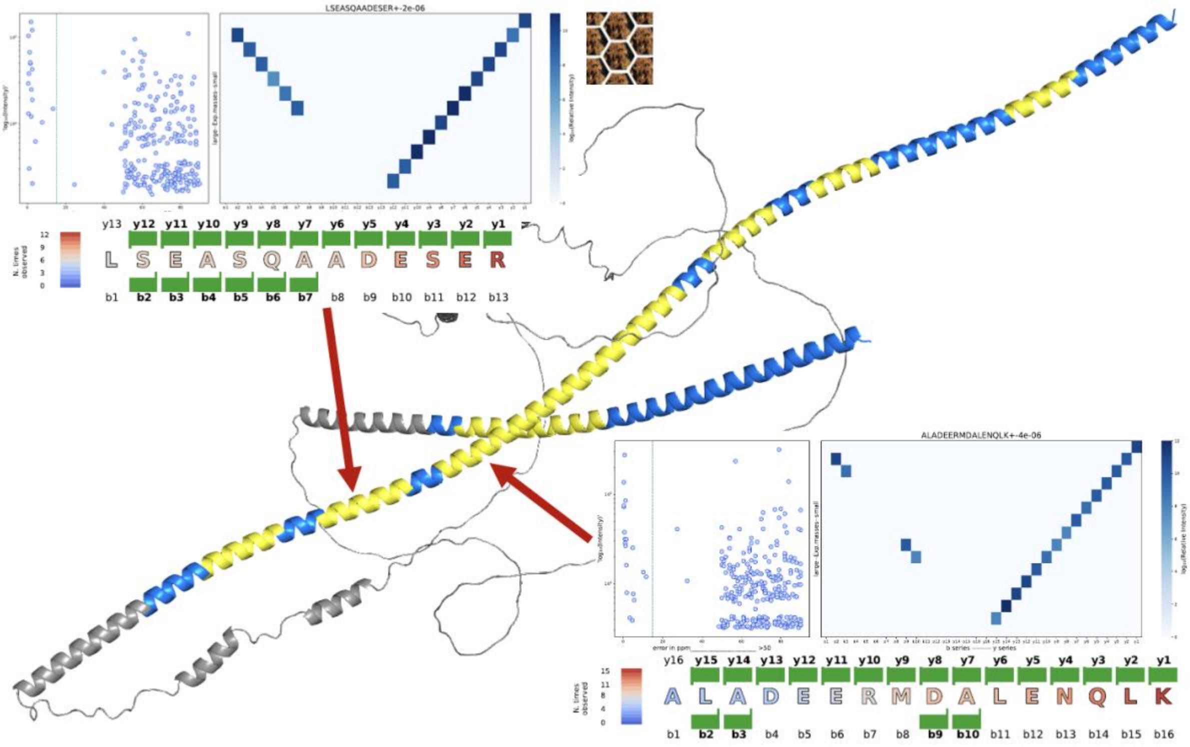

LSEASQAADESER+-2e-06
y13 y12 y11 y10 y9 y8 y7 y6 y5 y4 y3 y2 y1
L S E A S Q A A D E S E R
b1 b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12 b13
N. times observed
ALADEERMDALENQLK+-4e-06
error in ppm
y16 y15 y14 y13 y12 y11 y10 y9 y8 y7 y6 y5 y4 y3 y2 y1
A L A D E E R M D A L E N Q L K
b1 b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12 b13 b14 b15 b16

Figure 3. Validating peptides for *Drosophila melanogaster* "HEL-109"
The AlphaFold model of UniProtKB protein HEL-109 (V9HVX8) with the detected peptides mapped in yellow. HEL-109 is really the product of an extended *Drosophila melanogaster* alternative transcript (Tm1-RE). The regions of V9HVX8 that are identical to the *Tm1* principal isoform [36] are coloured blue (where no peptide matches) and yellow (where peptides from PeptideAtlas map). The structure is coloured grey for those regions of V9HVX8 that are substantially different from the *Tm1* principal isoform. There are no peptides for these 426 residues. Two of the peptides (positions shown by red arrows) were analysed using VSeq [37], and the VSeq output shows that these two peptides could not be mistaken for other peptides. Similar results were found for almost all the detected HEL-109 peptides.

However, HEL-109 is not a human protein, it is a minor isoform of *Drosophila melanogaster* tropomyosin, *Tm1*. The protein comes from a CDS deposited in GenBank [40] in 2008 (EU668323) that has been incorporated into UniProtKB TrEMBL. TrEMBL proteins form part of the THISP search database. The Flybase [38] Tm1-RE transcript almost certainly does not produce a functional 711 amino acid protein, it has TRIFID [39] score of 0.263 against 1.0 for the principal transcript Tm1-RG. All 24 SDPs map to the protein produced by Tm1-RG.

How did a fruit fly protein come to be identified by so many uniquely mapping peptides in human mass spectrometry experiments? One possibility is that the search engine simply mistook the *Drosophila melanogaster* peptides for human peptides. However, we have analysed the spectra for the 24 peptides and 21 of them are supported by at least one peptide-spectrum match that has a perfect y-series [Figure 3]. The identifications of the *Drosophila melanogaster* peptides are not erroneous.

Another possibility is that human tropomyosin genes have single amino acid variants or post-translational modifications that mimic their *Drosophila melanogaster* counterparts. However, almost all the 24 *Drosophila melanogaster* peptides have three or more amino differences from the human counterparts. Human tropomyosin genes have few non-synonymous variants and those missense variants that are known do not correspond to the amino acids that differ between the human and fruit fly peptides. These are not variants.

Post-translational modifications in human tropomyosin are less well annotated, but even here it is hard to imagine that they could be the reason why fruit fly proteins were identified. With 21 of the 24 peptides supported by PSMs with perfect y-series, there would need the perfect combination of multiple variants and post-translational modifications in the human peptide to confuse the PeptideAtlas search engine. And this needs to have happened not just for one peptide, but for all 21 peptides. This explanation is clearly impossible.

So, only one possibility remains. The PeptideAtlas search has identified a *Drosophila melanogaster* protein in human experiments. Ignoring explanations that belong in science fiction, it means that experiments were contaminated in some way with fruit fly proteins. Twenty three of the 24 peptides were detected in only four proteomics experiments. The final peptide detected had only a single PSM and may be a false positive identification. The four experiments covered a range of unrelated tissues, including ovarian tumours, placenta, and peritoneum. Each experiment identified between 7 and 10 distinct fruit fly peptides and all 21 perfect y-series came from these four experiments. These experiments were all somehow contaminated with insect proteins.

PeptideAtlas also identifies multiple peptides for five mouse proteins that are misannotated as human in UniProtKB Trembl. We analysed these five proteins in detail and concluded that the identification of these mouse peptides was almost certainly due to cross-contamination of mouse proteins in human experiments. A small number of proteomics experiments identified multiple peptides for three or four of the different mouse proteins, but as many as 50 or 60 distinct human proteomics experiments had evidence for at least one of these mouse proteins.

There are also proteins from invasive bacteria and virus that are annotated as human by UniProtKB, though it is easier to imagine how these came to contaminate human proteomics experiments. For example, the novel protein with the seventh highest number of SDPs was derp12 (Dermal papilla derived protein 12), which a BLAST [41] search shows is a dihydrolipoyl dehydrogenase from *Mycoplasma hyorhinis*, a swine pathogen sometimes found on human skin [Figure 4]. Entries tagged as human by TrEMBL and validated by the PeptideAtlas human build include proteins from E. Coli, pseudomonas and papillomavirus, that clearly should not be annotated as human proteins. These proteins could be added as part of the “microbe” section of the PeptideAtlas THISP database.

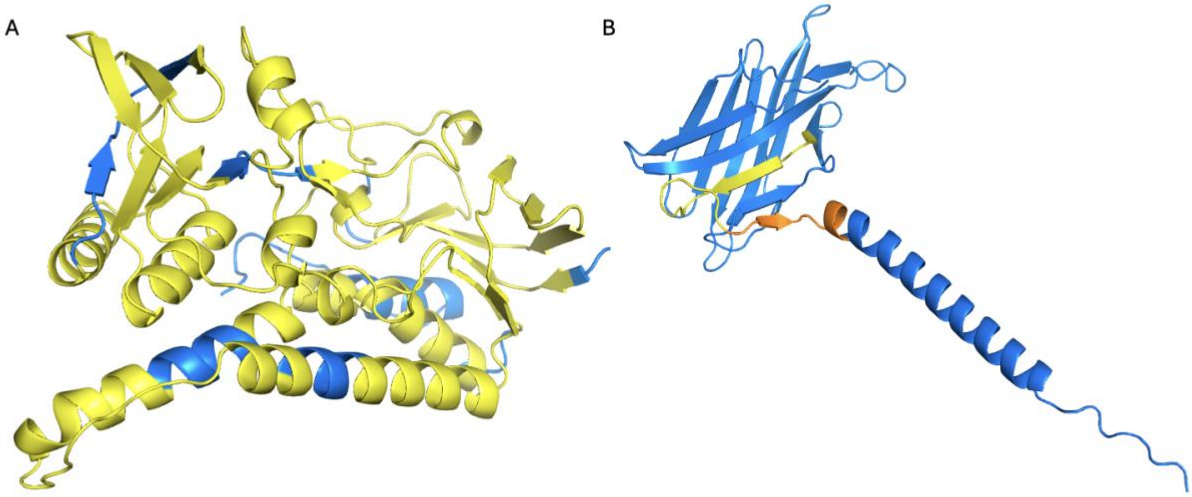


Figure 4, Identified peptides for Derp12 and the *LAMP3* alternative isoform.
A. The AlphaFold predicted structure for Derp12 painted yellow to indicate the peptides detected. B. The AlphaFold predicted structure for the *LAMP3* alternative isoform (XP_005247417.1) showing the alternative C-terminal helix. The region with PeptideAtlas peptides is painted yellow where it overlaps the *LAMP3* region common between the alternative and principal isoforms and orange where it overlaps the alternative C-terminal helix.

In total, we find at least two PeptideAtlas peptides for 18 non-human proteins that are annotated as human in UniProtKB. This result should be somewhat concerning for the UniProtKB TrEMBL database, since the misclassified proteins that PeptideAtlas detects are likely to only be the tip of the iceberg.

Non-reference proteins

Four of the ten novel entries with most supporting evidence were non-reference proteins. These are proteins that are not included in the Ensembl/GENCODE reference set. Three of the entries in the top ten of SDP support correspond to the loss of function genes, *GBA3*, *GPATCH4* and *PNLIPRP2*. These genes have high frequency alleles with frameshifts or stop codons. However, the proteins are clearly expressed in a large part of the population at least. *GPATCH4* has a high frequency frameshift allele that produces a different, extended C-terminal. PeptideAtlas has considerable peptide evidence for the C-terminal produced from the high frequency frameshift and none from the original C-terminal. GENCODE does not currently include full length loss of function proteins in its downloadable proteome.

LINE-1 ORF1

The entry with most SDP support (65 peptides) is the LINE-1 retrotransposable element ORF1 protein. The LINE-1 element that produces both this protein and an ORF2 protein that was practically undetected in PeptideAtlas [42] is present in multiple copies in the genome. The UniProtKB LINE-1 ORF1 protein maps to 37 different sites in the genome and there are more than a thousand full length LINE-1 ORF1 open reading frames that could produce ORF1 proteins with 95% or more identity to the UniProtKB protein. LINE-1 ORF1s make up more than a million bases and the sheer scale means that coordinate-based reference sets like Ensembl/GENCODE and RefSeq cannot annotate LINE-1 ORF1 as coding.

As many as 100 LINE-1 elements are still active [43]. These are responsible for retrotransposition of processed pseudogenes and SINE Alu and SINE Ventr Alu elements and are suppressed in normal tissues but become activated in tumours [44]. It is noticeable that the vast majority of observations for the 65 peptides in PeptideAtlas come from cancer cell lines. Given their clinical importance and abundance in cancer tissues [45, 46], it seems that diagnostic proteomics experiments ought to include the LINE-1 ORF1 protein in their search databases, something that could not happen with the RefSeq and/or Ensembl/GENCODE gene sets alone.

Unannotated alternative splice isoforms

PeptideAtlas has proteomics support for 278 alternative splicing events that are not annotated in GENCODE. One example is a novel C-terminal in *LAMP3* [Figure 4B] produced from an alternatively spliced tandem duplicated exon that can be traced back to vertebrates [47]. However, none of the 278 predicted splice isoforms have yet been validated by GENCODE manual curators, and we did not analyse any of the PSM for this entry class, so we cannot rule out the possibility that some of these entries may be variants or false positive identifications, or that some may be the result of aberrant or noisy translation.

Translation of upstream untranslated regions

There have been a number of recent publications highlighting translation from the untranslated regions (UTR) of protein coding genes, in particular translation from the 5’ UTR [21, 48, 49]. PeptideAtlas has multiple SDPs for 99 novel entries derived from the translation from 5’ UTR regions. Most of these translated upstream regions are in the same reading frame as the main transcript and would result in N-terminal extensions of the main protein isoform. We also found evidence for the translation of uORFs (ORFs that begin and end upstream of the canonical ATG), and uoORFs (ORFs that begin upstream of the canonical ATG and read through to coding exons, but in a different frame). Since these ORFs are found in the 5’ UTR regions of coding genes, we did not consider these regions as novel coding genes.

Most of these translated upstream regions have little cross species support and may be products of an aberrant or noisy translation initiation process [21**,** 49]. As with the alternatively spliced isoforms, these UTR decorations have not yet been validated by GENCODE manual curators, and we did not analyse the PSM for these regions, so some may be false positive identifications.

Genes that are not annotated as coding in the GENCODE reference set

As well as non-reference genes, alternative isoforms and translated upstream regions, PeptideAtlas also has SDP support for 34 possible coding genes that would be novel to the Ensembl/GENCODE gene set. We labelled these genes as coding, aberrant or unplaced after manual curation of the data associated with each novel coding region.

The 16 likely coding genes are listed in Table 1. Two of these we had detected in an earlier analysis [9]. *WASHC1* (12 SDPs) and *GPRIN2L* (3 SDPs) are both part of the new CHM13 Assembly. Neither the correct *WASHC1* gene nor *GPRIN2L* were annotated in UniProtKB, nor in the Ensembl/GENCODE or RefSeq reference assemblies that are based on the GRCh38 assembly.

Gene	**Strong Peps**	**UniProt**	**Annotated**	**Decision**
ANKRD26P1	9	Proteome	Pseudogene	Coding
C5orf60	16	Proteome	Pseudogene	Coding
EP400P1	6	Proteome	Pseudogene	Coding
ERVFRD-2	3	TrEMBL	Retrovirus	Coding
ETD-L (Q3ZM62)	3	TrEMBL	Pseudogene	Coding
FAM183BP	6	Proteome	Pseudogene	Coding
GPRIN2L	3	-	NA	Coding
LIPT2-AS1	3	TrEMBL	Pseudogene	Coding
LNCPRESS1	2	-	Retrovirus	Coding
LOC107986768	4	-	Retrovirus	Coding
MSL3P1	9	Proteome	Pseudogene	Coding
PLAC4	3	TrEMBL	Retrovirus	Coding
TSPY26P	2	Proteome	Pseudogene	Coding
TXLNGY	5	Proteome	Pseudogene	Coding
WASHC1	10	-	NA	Coding
ZNF840P	12	Proteome	Pseudogene	Coding

Table 1. Predicted likely coding genes

The 16 genes supported by PeptideAtlas that are not in the Ensembl/GENCODE reference set and are likely to be protein coding.

Novel ORFs from predicted pseudogenes

There are multiple peptides in PeptideAtlas for ten genes annotated in UniProtKB as part of the human reference proteome, but that were annotated as pseudogenes in both Ensembl/GENCODE and RefSeq. They are *ANKRD26P1* (9 SDPs, Figure 5A), *CFAP144P1* (6 SDPs), *EP400P1* (6 SDPs), *MSL3P1* (9 SDPs), *MYH16* (11 SDPs), *PMS2CL* (3 SDPs), *POM121L1P* (3 SDPs), *TSPY26P* (2 SDPs) *TXLNGY* (5 SDPs), and *ZNF840P* (12 SDPs).

ANKRD26P1 [Figure 5A], *EP400P1*, *MSL3P1* [Figure 5B] and *MYH16 are* highly truncated versions of their parent genes. We believe that 7 of these genes are likely to be coding genes.

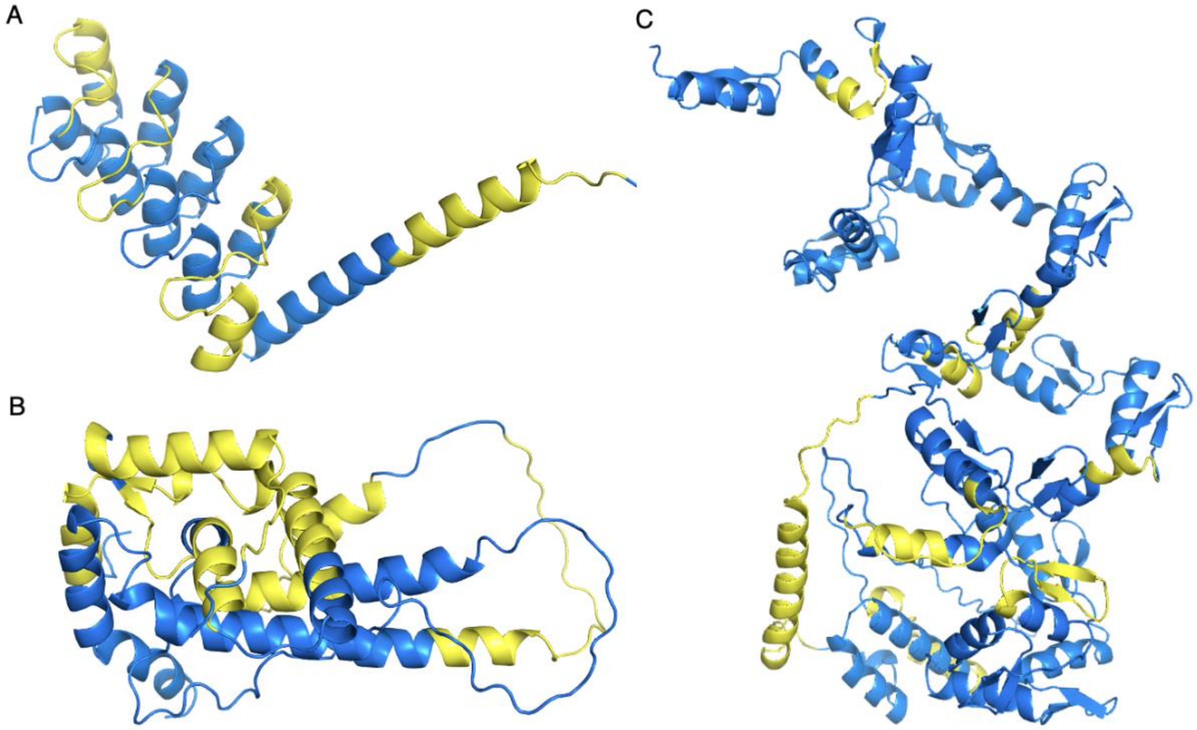


Figure 5. Discriminating peptides detected for *ANKRD26P1*, *MSL3P1* and *ZNF840P*
All structures predicted by AlphaFold for the UniProtKB database. Detected peptides mapped in yellow. A. *ANKRD26P1* protein with the disordered N- and C-termini removed. B. *MSL3P1* protein from the downstream ATG with the large, disordered loop removed for clarity. C. *ZNF840P* protein.

There were nine SDPs for *MSL3P1*, all of which are substantially different from the equivalent in the parent gene, *MSL3*. This is a primate duplication which has lost the N-terminal thanks to a frameshift [Figure 5B]. The peptide evidence suggests that *MSL3P1* uses a downstream ATG that would bypass the tudor knot domain in the equivalent region of *MSL3*. In addition, great apes have a different C-terminal because of a frameshift that affects the last two codons. Despite the recent changes, alignments suggest that *MSL3P1* is under protein-coding selection across all primates.

Although RefSeq has recently made *MSL3P1* coding, there is some evidence to suggest that this gene product may be the result of aberrant translation. Almost all the peptide evidence is observed in cancer cell experiments, with just a small number of unconvincing spectra found in lung. All four recent papers published on *MSL3P1* suggest that it may be an oncogene [50].

Peptides for *EP400P1* are mostly detected in cancer cell experiments, but there are also many observations in normal tissues, particularly testis. As with *MSL3P1* there are also four recent publications for *EP400P1*, and these suggest a biological role for this misannotated pseudogene [51]. *TSPY26P* has few supporting peptides but does have good conservation evidence. It is a conserved single exon gene that is under coding selection across primates, and that has orthologues in mammalian species including squirrels, whales and dolphins, and bats. Peptides were found for the N-terminal end of the protein that is almost entirely different from any paralogue.

ZNF840P has the most convincing evidence, though it may have a different N-terminal from the one annotated in the UniProtKB protein [Figure 5C]. *ZNF840P* would be a simian duplication that has evolved considerably. It has little more than 40% identity to other zinc finger proteins and is perhaps most similar to *ZNF780B*. The start and stop codons and the frame are conserved among old and new world monkeys. All PeptideAtlas peptides for *ZNF840P* were detected in oocytes, which fits with the RNAseq data.

The evidence *for MYH16* is conflicting. Superficially, 11 SDPs in both cancer cell experiments and in heart and muscle tissues is strong peptide support. However, the heart and muscle PSMs are poor and do not support the *MYH16* peptides. Only the PSMs from the cancer cell line experiments unequivocally support *MYH16*. In isolation, the peptide data suggests that this is an aberrant coding gene. *MYH16* is coding in primates and is specifically expressed in jaw muscles, but the human protein would have to start from a downstream start codon because of a stop codon in exon 18. In addition, human *MYH16* has lost the 5' splice signal of exon 26. Repurposed pseudogenes do exist, but *MYH16* has no support beyond peptides in cancer cell lines. It is upregulated in lung adenocarcinoma and other cancers [52].

Finally, it is hard to imagine how *PMS2CL* could be a *bona fide* protein coding gene. It has three SDPs and all are at least three amino acids different from the parent gene. However, *PMS2CL*, currently a pseudogene, would produce a protein that would be truncated at both ends. It is also only conserved in human. The PSM found for *PMS2CL* are only found in cancer cell lines, so this may be a protein that is expressed under aberrant conditions [Table 2].

Gene	Strong Peps	UniProt	Annotated	Decision
ARRDC1-AS1	3	Proteome	LncRNA	Aberrant
DOCK8-AS1	11	Proteome	LncRNA	Aberrant
ERVS71-1	2	Proteome	Retrovirus	Aberrant
FAM230A	6	TrEMBL	LncRNA	Unplaced
FAM230C	19	Proteome	LncRNA	Aberrant
HCA25a	2	-	LncRNA	Unplaced
LINC00839	12	TrEMBL	LncRNA	Aberrant
LINC02693	2	Proteome	LncRNA	Aberrant
LINC03040	2	Proteome	LncRNA	Aberrant
LOC124905049	2	-	LncRNA	Aberrant
MHENCR	2	-	LncRNA	Aberrant
MNX-AS1	2	-	LncRNA	Aberrant
MYH16	11	Proteome	Pseudogene	Aberrant
NECTIN3-AS1	2	Proteome	LncRNA	Aberrant
PMS2CL	3	Proteome	Pseudogene	Aberrant
POM121L1P	3	Proteome	Pseudogene	Unplaced
PTPN18	2	-	LncRNA	Aberrant
ZNF252P	9	TrEMBL	Pseudogene	Aberrant

Table 2. Predicted aberrant and unplaced coding genes

The 18 genes with peptide support in PeptideAtlas that are not in the Ensembl/GENCODE reference set and that are either likely to be aberrant coding genes or do not map to the genome

Proteins from long non-coding genes

We found peptides for seven genes that were part of the UniProtKB reference proteome, but that were classified as non-coding in both Ensembl/GENCODE and RefSeq reference sets. *C5orf60* was supported by 16 SDPs, *DOCK8-AS1* by 8, *ARRDC1-AS1* by 3, and *ERVS71-1*, *LINC02693*, *NECTIN3-AS1*, and *LINC03040* by 2 SDPs each.

An eighth entry was included in the UniProtKB proteome but is not part of the GENCODE reference set. It was supported by 19 SDPs in our analysis. The protein sequence of Q9UF83 is made up almost entirely of nine amino acid repeats and there are many highly similar repeat regions of the genome, especially on chromosomes 13 and 22. Despite the overwhelming peptide evidence in PeptideAtlas, it is noticeable that all 19 peptides were detected in cancer cell lines.

Of the eight ORFs, we believe that only *C5orf60* is coding. Even though *C5orf60* is annotated as “TEC” in Ensembl/GENCODE and non-coding RNA in RefSeq, it has coding ancestry because it is a truncated copy of parent gene *SPATA31E1*. All 16 SDPs were detected in testis.

If *DOCK8-AS1* were a coding gene, it would be a *de novo* emergence from a non-coding sequence. Multiple stop codons and frame shifts means this ORF would not be translated in any other species. The peptides are all supported by their spectra, and while most are found in kidney cancer cell lines, one is found in kidney tissue. The kidney-specific expression fits with the RNAseq data from GTex [53], which shows that kidney is the most common tissue for *DOCK8-AS1*. There is no doubt that this protein is translated, but the translation is likely to be aberrant. Not only would *DOCK8-AS1* have to be a novel human-specific coding gene, but it also overlaps the first coding exon of *DOCK8* on the other strand. The expression of *DOCK8-AS1* has been shown to be related to survivability in renal cancers [54-55].

LINC03040 was protein-coding for many years, but the CDS was removed because it had little or no evolutionary support. However, there are multiple distinguishing peptides for this gene, including ten non-tryptic peptides that we filtered out at the start of the automatic analysis. These ten peptides and the only strong tryptic peptide with a good-looking PSM were all detected in HLA-I experiments. So, *LINC03040* does have considerable evidence for translation, but the most reliable peptides are detected exclusively in HLA proteomics experiments, most of which were carried out on cell lines. We still do not have a clear idea why some peptides are detected in HLA experiments, but not in normal tissue, but *LINC03040* seems most likely to be an aberrant rather than a functional coding gene. It was recently suggested that *LINC03040* is an oncogene in colon carcinoma [56].

Possible novel genes annotated only in the TrEMBL database

Eight ORFs with peptide evidence were novel to GENCODE but found in the catch-all UniProtKB TrEMBL database. *LINC00839* was supported by 12 SDPs, *ZNF252P* by 9, *FAM230A* by 6 and Q3ZM62 by 5. Three other cases (*ERVFRD-2*, *LIPT2-AS1*, and *PLAC4*) were supported by three SDPs each, and HCA25a by just two peptides.

The lncRNA *LIPT2-AS1* has 42% identity to the *JRK* gene, which is a Tigger transposon domestication event, but the *LIPT2-AS1* protein would not have the DDE endonuclease domain central to the *JRK* protein. The Cactus 470 mammal alignments [57] show clear protein coding conservation right across all primates for *LIPT2-AS1* and UniProtKB annotates highly similar proteins for rabbit, naked mole rat, beaver, green anole and bamboo shark, so this gene may be a lot more ancient than it appears.

We found 3 SDPs for this gene in the 2023 build of PeptideAtlas 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 ties in with transcriptomics support for *LIPT2-AS1* in limited brain tissues. In addition, the gene model predicted by the UniProtKB protein is probably not quite right; we find another ATG 51 codons upstream that would complete a CENP-B N-terminal DNA-binding domain and that are even conserved in bamboo shark [Figure 6A].

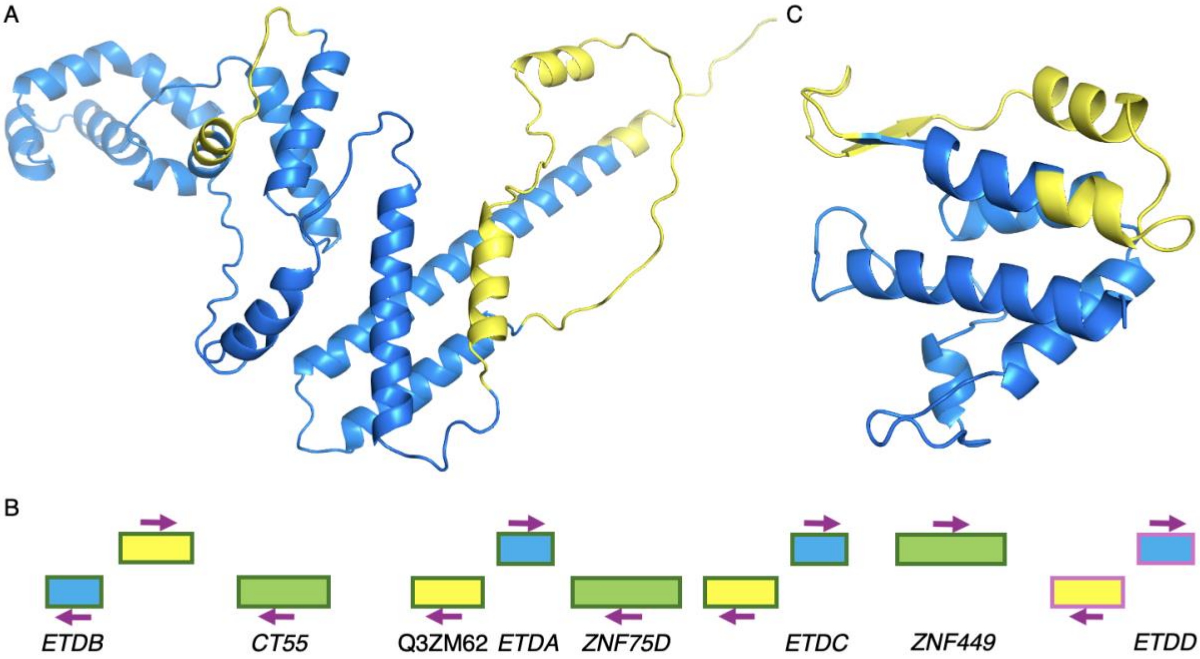


Figure 6. Peptides detected for *LIPT2-AS1*, *LOC107986768* and *the ETD family*
A. The AlphaFold structure of macaque *LIPT2-AS1* with the peptides detected in PeptideAtlas marked in yellow on the structure. B. A schematic representation of the human ETD family on chromosome X. ETD genes are coloured in blue, antisense ETD-like genes are shown as yellow boxes. Other intervening genes are marked in green. Gene sense is marked with arrows. Probable pseudogenes have a pink border. C. The structure of the C-terminal 134 residues of *LOC107986768* from the HHPRED server [58] with the peptides detected in PeptideAtlas marked in yellow on the structure.

UniProtKB TrEMBL protein Q3ZM62 is not annotated as an ORF in either the Ensembl/GENCODE or the RefSeq reference set. The ORF would overlap a section of 3' UTR in *ZNF75D* just downstream of *ETDA* on chromosome X [Figure 6B]. The peptides in the most recent build of PeptideAtlas cover the protein sequence from residues 2 to 73, but the sequence of Q3ZM62 is incorrect because it would require a change in the reading frame beyond codon 80. Q3ZM62 has just over 40% identity to *ETDA.*

In fact, both *ETDA* and the Q3ZM62 ORF have three paralogues each and all on are chromosome X [Figure 6B]. As with the Q3ZM62 ORF, each of the Q3ZM62 paralogues are antisense to an EDT paralogue (*ETDB*, *ETDC*, and an unannotated ETD pseudogene). This suggests that one of the Q3ZM62 ORFs duplicated from an ETD family member, and the remaining paralogues are the result of segmental duplications of the original pair of genes. The original ETD gene duplication occurred prior to the primate clade because there is an orthologue of the Q3ZM62 ORF in mouse, though it is not antisense to mouse *Etda*. Two segmental duplications appear to have involved *ETDA* and the Q3ZM62 ORF, one pair of genes (possibly the pair involving the *ETDC* gene) duplicated right at the base of primates, the second duplication (the *ETDB* pair) is a human specific duplication. The *ETDD* pair duplicated from the *ETDC* pair after the split between old world and new world monkeys. *ETDD* is a pseudogene in great apes and its Q3ZM62-like ORF appears to be a pseudogene outside of great apes. *ETDA*, *ETDC* and Q3ZM62 are clearly protein coding across primates, but evidence for the other three ORFs is not quite so clear. In any case, all four ETD-like ORFs have been annotated as coding genes by the GENCODE annotators.

The aberrant expression patterns of *LINC00839* in cancer tissues and cell lines have been well documented [59]. The reading frame of this ORF is not conserved beyond human and all 12 peptides in PeptideAtlas are detected solely in cancer cell lines. This is a clear case of aberrant translation.

The entry HCA25a (hepatocellular carcinoma-associated antigen 25a) does not map to the human genome, but the central region that is identified by SDPs does map to multiple regions (14 in total) across the genome. This region of the HCA25a sequence is highly similar to the theoretical protein sequences of oesophageal carcinoma antigens [60]. Peptides are observed in cancer cell lines and in all likelihood whichever ORF is produces these peptides is likely to be undergoing aberrant translation. HCA25a has been tagged as unplaced because we do not know which ORF the peptides belong to.

Possible novel ORFs not in UniProtKB

Six ORFs with peptide evidence were novel to GENCODE and not in any UniProtKB database. *LOC107986768* was supported by four SDPs, the others (*LNCPRESS1, LOC124905049, MHENCR, MNX-AS1*, and an ORF in the 3' UTR of *PTPN18*) had two SDPs each. We believe that two of these ORFs may code for proteins in normal tissues.

LOC107986768 is a retroviral GAG-30 derived ORF on the opposite strand from the *SCIN* coding gene and is annotated by RefSeq as coding. RefSeq notes that expression is restricted to placenta, and all the peptides in PeptideAtlas were detected in placental tissue. In the 2024 build of PeptideAtlas there are 11 peptides (5 strong and 6 non-tryptic) for this gene and all peptides map between residues 140 and 270 [Figure 6C]. This strongly suggests that there may be translation from a downstream start site, but if this is the case, it would have to be non-canonical. The reading frame is undisturbed among great apes, but if the downstream start site were used it would be undisturbed across old world monkeys. Another two novel ORFs, *PLAC4* and *ERVFRD-2*, are also retroviral and appear to be expressed mostly or solely in placenta.

LNCPRESS1 is another retroviral ORF, this time derived from a L1 transposon. The ORF covers less than half of the original LINE-1 ORF, but what is left is largely maintained across old world monkeys. All the peptides are detected in embryonic stem cells, which makes sense given that extant LINE-1 ORFs are expressed in the early stages of development [44]. At the same time, *LNCPRESS1* RNA has been shown to be mainly nuclear and it has been postulated that the *LNCPRESS1* transcript (not protein) is a crucial part of maintaining pluripotency of cells [61]. *MHENCR* is a melanoma oncogene [62]. Peptides are detected in cancer cells and cell lines, with the majority of observations in HLA proteomics experiments. The combination of these two facts suggests that the translation of *MHENCR* is aberrant.

Discussion

We have carried out an in-depth analysis of PeptideAtlas peptides that do not map to Ensembl/GENCODE coding genes. With the peptides that did not map to the GENCODE reference set we identified more than 250 alternative isoforms, almost 100 translated upstream regions and 34 possible coding genes that are not currently annotated by Ensembl/GENCODE. These novel coding regions are currently being analysed by GENCODE curators.

A large majority of the almost 20,000 peptides that do not map to GENCODE coding genes are most likely to be amino acid variants or post-translational modifications of peptides from known coding genes rather than novel coding regions. These peptide variants map to immunoglobulins, major histocompatibility complex proteins or other proteins known to have many variants. Many map erroneously to pseudogenes of known coding genes [33].

One surprising finding was that there were many valid PSMs for non-human proteins in the PeptideAtlas human build. The TPP identifies peptides for vertebrate, invertebrate and microbial proteins that were erroneously annotated as human by UniProtKB. The microbial proteins, particularly those of *Mycoplasma hyorhinis*, probably need to be added to the list of contaminants. We also found valid peptides for fruit fly, mouse and pig proteins in the PeptideAtlas build. While it is possible to imagine insect proteins as floating contamination similar to other contaminants such as keratin and wool, mouse and pig proteins cannot possibly be airborne. In addition, although tropomyosin is a common, ubiquitous cytoskeletal protein, there is nothing special about the mouse proteins that are misannotated as human and identified in experiments. They include a geranylgeranyl transferase, a PIDDosome adapter protein, and a substrate adapter for ufmylation. This means that the most likely explanation for the detection of these mouse proteins is large-scale cross-contamination or some kind of sample mix up.

Although we validated PeptideAtlas peptides for 34 genes that are not in the GENCODE reference set, none of them can be regarded as novel discoveries since entries in the PeptideAtlas THISP database have all been discovered previously. Just over half the entries that we found peptides for are annotated by UniProtKB as part of their human reference proteome, while another two are produced by genes annotated as coding by RefSeq. Seven of the 34 entries that we detected peptides for were identified in the large-scale ribosome profiling experiments. *MSL3P1*, *TSPY26P*, *TXLNGY*, *ZNF252P* and *LNCPRESS1* were detected by Chen *et al* [18], while *LIPT2-AS1, and AARSD-AS1* were reported by Van Heesch *et al* [19].

We analysed the PSMs, tissue expression patterns and cross-species conservation for all 34 genes in this analysis. We believe that almost half are likely to be coding genes, including 8 of the 18 genes from the UniProtKB human proteome and 5 of the 8 proteins from the UniProtKB TrEMBL database. Most of these 16 likely coding genes appear to be recent evolutionary innovations, but still have clear evidence of coding conservation. This coding conservation is clear even in those ORFs that have novel stop codons in great apes (*ERVFRD-2*, *LIPT2-AS1*).

We believe that 14 of the ORFs are probably producing peptides as a result of aberrant translation. These ORFs have peptides detected only in HLA proteomics experiments

(*LINC03040*), only in cancer cell experiments (*ZNF252P*), or are known to be cancer antigens (*MHENCR*). The human reading frames of 12 of these 14 ORFs are not conserved in any species and the other two are not conserved beyond great apes. Six are already described as cancer relevant genes.

The remaining 4 proteins with multiple SDPs were unplaced in the genome. *FAM230A*, Q9UF83, HCA25a and *POM121L1P* mapped to multiple regions in the human genome, and it would be almost impossible to decide which of the regions was being translated. Two of these entries only had peptides in cancer cell line experiments, but almost all peptides for *FAM230A* are found in testis or sperm, and there are experimental peptides from sperm for *POM121L1P*, though the PSMs are poor for these peptides. Although something similar to the *POM121L1P* peptides may be sperm expressed, there are more than 150 possible different POM121 pseudogene locations in the human genome, scattered mostly across chromosomes 5, 6 and 22.

LINE-1 ORF1 could conceivably be added to the list of unplaced ORFs. As with the POM121 pseudogenes, LINE-1 ORF1 is present in multiple copies in the genome. It is also the unannotated ORF with most peptide evidence in PeptideAtlas, and almost all peptide-spectrum matches are from cancer cells or cell lines [21]. Between LINE-1 ORF1 and the unplaced and likely aberrant coding genes there were 17 unannotated non-conserved ORFs that are supported only by peptide evidence from cancer cells or cell lines, showing that aberrant translation products may be common in cancer cells. This adds to the evidence of the substantial dysregulation of translation in cancer cells [63-65].

The peptides for novel ORFs that have no cross-species support and expression limited to cancer cells illustrates what may become a whole new level of complexity in the annotation of coding genes if many of these ORFs are present in ribosome profiling experiments. The first question is whether the proteins produced by these ORFs have functional roles. Despite the peptides, the conservation evidence suggests that this is unlikely, but it cannot be ruled out completely. The second question is, if these proteins 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 like *HMHB1* [66] and *MYEOV* [67], or should they be labelled separately so that they cannot be mistaken for genes that produce biologically relevant proteins?

Funding

This work was funded by the National Human Genome Research Institute of the National Institutes of Health (grant number U41 HG007234), by grant PGC2018-097019-B-I00 from the Spanish Ministry of Science, Innovation and Universities, by grant IPT17/0019 from the Carlos III Institute of Health-Fondo de Investigación Sanitaria and by HR17-00247 from the ’la Caixa’ Foundation.

References

1. Nurk,S., Koren,S., Rhie,A., Rautiainen,M., Bzikadze,A.V., Mikheenko,A., Vollger,M.R., Altemose,N., Uralsky,L., Gershman,A., *et al.* (2022) The complete sequence of a human genome. *Science*, 376, 44–53.
2. Rhie,A., Nurk,S., Cechova,M., Hoyt,S.J., Taylor,D.J., Altemose, N., Hook, P. W., Koren,S., Rautiainen,M., Alexandrov,I.A., *et al.* (2023) The complete sequence of a human Y chromosome. *Nature*, 621, 344–354.
3. Frankish,A., Carbonell-Sala,S., Diekhans,M., Jungreis,I., Loveland,J.E., Mudge,J.M., Sisu,C., Wright,J.C., Arnan,C., Barnes,I., *et al*. (2023) GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res*., 51, D942–D949.
4. Martin,F.J., Amode,M.R., Aneja,A., Austine-Orimoloye,O., Azov,A.G., Barnes,I., Becker,A., Bennett,R., Berry,A., Bhai,J., *et al.* (2023) Ensembl 2023. *Nucleic Acids Res*., 51, D933–D941.
5. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Farrell,C.M., Feldgarden,M., Fine,A M., Funk,K., *et al.* (2023) Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res*., 51, D29–D38.
6. UniProt Consortium (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*., 51, D523–D531.
7. Abascal,F., Juan,D., Jungreis,I., Kellis,M., Martinez,L., Rigau,M., Rodriguez,J.M., Vazquez,J. & Tress,M.L. (2018) Loose ends: almost one in five human genes still have unresolved coding status. *Nucleic Acids Res*., 46, 7070–7084.
8. Pertea,M., Shumate,A., Pertea,G., Varabyou,A., Breitwieser,F.P., Chang,Y.C., Madugundu,A.K., Pandey,A., & Salzberg, S.L. (2018) CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol*., 19, 208.
9. Cerdán-Vélez,D. & Tress,M.L. (2024) The T2T-CHM13 reference assembly uncovers essential WASH1 and GPRIN2 paralogues. *Bioinfo. Adv*., 4, vbae029.
10. Mudge,J.M., Ruiz-Orera,J., Prensner,J.R., Brunet,M.A., Calvet,F., Jungreis,I., Gonzalez,J.M., Magrane,M., Martinez,T.F., Schulz,J.F., *et al.* (2022) Standardized annotation of translated open reading frames. *Nat Biotech*., 40, 994–999.
11. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.
12. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, 291, 1304–1351.
13. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res*., 30, 38–41.
14. Southan,C. (2004) Has the yo-yo stopped? An assessment of human protein-coding gene number. *Proteomics*, 4, 1712–1726.
15. Clamp,M., Fry,B., Kamal,M., Xie,X., Cuff,J., Lin,M.F., Kellis,M., Lindblad-Toh,K. & Lander,E.S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci. USA*,104, 19428–19433.
16. Church,D.M., Goodstadt,L., Hillier,L.W., Zody,M.C., Goldstein,S., She,X., Bult,C.J., Agarwala,R., Cherry,J.L., DiCuccio,M., *et al.* (2009) Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS Biology*, 7, e1000112.
17. Ezkurdia,I., Juan,D., Rodriguez,J.M., Frankish,A., Diekhans,M., Harrow,J., Vazquez,J., Valencia,A. & Tress,M.L. (2014) Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet*., 23, 5866–5878.
18. Chen,J., Brunner,A.D., Cogan,J.Z., Nuñez,J.K., Fields,A.P., Adamson,B., Itzhak,D.N., Li,J.Y., Mann,M., Leonetti,M.D., & Weissman,J S. (2020) Pervasive functional translation of noncanonical human open reading frames. *Science*, 367, 1140–1146.

19. van Heesch,S., Witte,F., Schneider-Lunitz,V., Schulz,J.F., Adami,E., Faber,A.B., Kirchner,M., Maatz,H., Blachut,S., Sandmann,C.L., *et al.* (2019) The translational landscape of the human heart. *Cell*, 178, 242–260.e29.
20. Prensner,J.R., Abelin,J.G., Kok,L.W., Clauser,K.R., Mudge,J.M., Ruiz-Orera,J., Bassani-Sternberg,M., Moritz,R.L., Deutsch,E.W., & van Heesch,S. (2023) What can ribo-seq, immunopeptidomics, and proteomics tell us about the noncanonical proteome?. *Mol. Cell. Proteo.*, 22, 100631.
21. Rodriguez,J.M., Abascal,F., Cerdán-Vélez,D., Gómez, L.M., Vázquez,J., & Tress,M.L. (2024) Evidence for widespread translation of 5' untranslated regions. *Nucleic Acids Res.*, 52, 8112-8126.
22. Liu,T., & Lin,K. (2015) The distribution pattern of genetic variation in the transcript isoforms of the alternatively spliced protein-coding genes in the human genome. *Mol Biosyst.*, 11, 1378–1388.
23. Martinez-Gomez,L., Abascal,F., Jungreis,I., Pozo,F., Kellis,M., Mudge,J.M., & Tress, M.L. (2020) Few SINEs of life: Alu elements have little evidence for biological relevance despite elevated translation. *NAR Genom Bioinfo.*, 2, lqz023.
24. Pozo,F., Rodriguez,J.M., Martínez Gómez,L., Vázquez,J., & Tress,M.L. (2022) APPRIS principal isoforms and MANE Select transcripts define reference splice variants. *Bioinformatics*, 38, ii89–ii94.
25. Sullivan,P.F., Meadows,J.R.S., Gazal,S., Phan,B.N., Li,X., Genereux,D.P., Dong,M.X., Bianchi,M., Andrews,G., Sakthikumar,S., *et al.* (2023) Leveraging base-pair mammalian constraint to understand genetic variation and human disease. *Science*, 380, eabn2937.
26. Pozo,F., Rodriguez,J.M., Vázquez,J., & Tress,M.L. (2022) Clinical variant interpretation and biologically relevant reference transcripts. *NPJ genomic medicine*, 7, 59.
27. Schittek,B., Hipfel,R., Sauer,B., Bauer,J., Kalbacher,H., Stevanovic,S., Schirle,M., Schroeder,K., Blin,N., Meier,F., *et al.* (2001) Dermcidin: a novel human antibiotic peptide secreted by sweat glands. *Nat Immunol.*, 2, 1133–1137.
28. Inada,T., & Beckmann,R. (2024) Mechanisms of translation-coupled quality control. *J Mol Biol*., 436, 168496.
29. Yewdell,J., Antón,L.C., Bacik,I., Schubert,U., Snyder,H.L., & Bennink,J.R. (1999) Generating MHC class I ligands from viral gene products. *Immun. Rev.*, 172, 97–108.
30. Kesner,J.S., Chen,Z., Shi,P., Aparicio,A.O., Murphy,M.R., Guo,Y., Trehan,A., Lipponen,J.E., Recinos,Y., Myeku,N., & Wu,X. (2023) Noncoding translation mitigation. *Nature*, 617, 395–402.
31. Desiere,F., Deutsch,E.W., King,N.L., Nesvizhskii,A.I., Mallick,P., Eng,J., Chen,S., Eddes,J., Loevenich,S.N., & Aebersold,R. (2006) The PeptideAtlas project. *Nucleic Acids Res.,* 34, D655-D658.
32. Deutsch,E.W., Mendoza,L., Shteynberg,D.D., Hoopmann,M.R., Sun,Z., Eng,J., & Moritz, R.L. (2023) Trans-Proteomic Pipeline: Robust mass spectrometry-based proteomics data analysis suite. *J Proteome Res*., 22, 615-624.
33. Deutsch,E.W., Sun,Z., Campbell,D.S., Binz,P.A., Farrah,T., Shteynberg,D.D., Mendoza,L., Omenn,G.S., & Moritz,R.L. (2016) Tiered human integrated sequence search databases for shotgun proteomics. *J Proteome Res*., 15, 4091-4100.
34. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*., 50, D439-D444.
35. Zahn-Zabal,M., Michel,P.A., Gateau,A., Nikitin,F., Schaeffer,M., Audot,E., Gaudet,P., Duek,P.D., Teixeira,D., Rech de Laval,V., *et al.* (2020) The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.*, 48, D328–D334.
36. Rodriguez,J.M., Pozo,F., Cerdán-Vélez,D., Di Domenico,T., Vázquez,J., & Tress,M.L. (2022) APPRIS: selecting functionally important isoforms. *Nucleic Acids Res*., 50, D54–D59.

37. Cogliati,S., Calvo,E., Loureiro,M., Guaras,A.M., Nieto-Arellano,R., Garcia-Poyatos,C., Ezkurdia,I., Mercader,N., Vázquez,J., & Enriquez,J.A. (2016) Mechanism of super-assembly of respiratory complexes III and IV. *Nature*, 539, 579–582.
38. Öztürk-Çolak,A., Marygold,S.J., Antonazzo,G., Attrill,H., Goutte-Gattat,D., Jenkins,V.K., Matthews,B.B., Millburn,G., Dos Santos,G., Tabone,C.J., & FlyBase Consortium (2024) FlyBase: updates to the Drosophila genes and genomes database. *Genetics*, 227, iyad211.
39. Pozo,F., Martinez-Gomez,L., Walsh,T.A., Rodriguez,J.M., Di Domenico,T., Abascal,F., Vazquez,J., & Tress,M.L. (2021) Assessing the functional relevance of splice isoforms. *NAR Genom. Bioinfo.*, 3, lqab044.
40. Sayers,E.W., Cavanaugh,M., Clark,K., Pruitt,K.D., Sherry,S.T., Yankie,L., & Karsch-Mizrachi,I. (2024) GenBank 2024 Update. *Nucleic Acids Res*., 52, D134–D137.
41. Altschul,S.F., Gish,W., Miller,W., Myers,E.W., & Lipman,D.J. (1990) Basic local alignment search tool. *J Mol Biol*., 215, 403-410.
42. Goodier,J.L., Cheung,L.E., & Kazazian,H.H.Jr. (2013) Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. *Nucleic Acids Res*., 41, 7401-7419.
43. Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V., & Kazazian,H.H.Jr. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad. Sci. USA,* 100, 5280–5285.
44. Mendez-Dorantes,C., & Burns,K.H. (2023) LINE-1 retrotransposition and its deregulation in cancers: implications for therapeutic opportunities. *Genes and development*, 37, 948–967.
45. Xiao-Jie,L., Hui-Ying,X., Qi,X., Jiang,X., & Shi-Jie,M. (2016) LINE-1 in cancer: multifaceted functions and potential clinical implications. *Genetics in medicine*, 18, 431–439.
46. Bowers,E.C., Cavalcante,A.M., Nguyen,K., Li,C., Wang,Y., El-Zein,R., Chen,S.H., Kim,M.P., McKay,B.S., & Ramos,K.S. (2024) Long interspersed nuclear element-1 analytes in extracellular vesicles as tools for molecular diagnostics of non-small cell lung cancer. *Int J Mol Sci.*, 25, 1169.
47. Martinez-Gomez,L., Cerdán-Vélez,D., Abascal,F., & Tress, M.L. (2022) Origins and evolution of human tandem duplicated exon substitution events. *Genome Biol Evol.*, 14, evac162.
48. Zhu,Y., Orre,L.M., Johansson,H.J., Huss,M., Boekel,J., Vesterlund,M., Fernandez-Woodbridge,A., Branca,R.M.M., & Lehtiö,J. (2018) Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat Commun.*, 9, 903.
49. Fedorova,A.D., Kiniry,S.J., Andreev,D.E., Mudge,J.M., & Baranov,P.V. (2022) Thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. *Nat Commun*., 13, 7910.
50. Shao,M.M., Li,X., Wei,R.Q., Chen,Q.Y., Zhang,X., Qiao,X., & Li,H. (2024) Long non-coding RNA MSL3P1 regulates CUL3 mRNA cytoplasmic transport and stability and promotes lung adenocarcinoma metastasis. *Mol Cancer Res.*, 22, 746–758.
51. Devoucoux,M., Roques,C., Lachance,C., Lashgari,A., Joly-Beauparlant,C., Jacquet,K., Alerasool,N., Prudente,A., Taipale,M., Droit,A., *et al.* (2022) MRG proteins are shared by multiple protein complexes with distinct functions. *Mol Cell. Proteomics*, 21, 100253.
52. Zhang,L., Liu,J., Wang,H., Xu,Z., Wang,Y., Chen,Y., & Peng,H. (2023) MYH16 upregulation is associated with lung adenocarcinoma aggressiveness and immune infiltration. *J Biochem Mol Tox.*, 37, e23490.
53. GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369, 1318–1330.
54. Shu,X., Zhang,Z., Yao,Z.Y., & Xing,X.L. (2022) Identification of five ferroptosis-related lncRNAs as novel prognosis and diagnosis signatures for renal cancer. *Front Mol Biosci.*, 8, 763697.
55. Chen,L., Xiang,Z., Chen,X., Zhu,X., & Peng,X. (2020) A seven-gene signature model predicts overall survival in kidney renal clear cell carcinoma. *Hereditas*, 157, 38.

56. Bai,H., Yan,D.S., Chen,Y.L., Li,Q.Z., & Qi,Y.C. (2024) Potential biomarkers: The hypomethylation of cg18949415 and cg22193385 sites in colon adenocarcinoma. *Comp Biol Med.*, 169, 107884.
57. Armstrong,J., Hickey,G., Diekhans,M., Fiddes,I.T., Novak,A.M., Deran,A., Fang,Q., Xie,D., Feng,S., Stiller,J., *et al.* (2020) Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature, 587, 246–251.
58. Zimmermann,L., Stephens,A., Nam,S.Z., Rau,D., Kübler,J., Lozajic,M., Gabler,F., Söding,J., Lupas,A.N., & Alva,V. (2018) A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. *J Mol Biol*., 430, 2237–2243.
59. Hu,Y., Hu,Y., Lu,X., Luo,H., & Chen,Z. (2024) LINC00839 in human disorders: Insights into its regulatory roles and clinical impact, with a special focus on cancer. *J. Cancer*, 15, 2179–2192.
60. Kagaya,A., Shimada,H., Shiratori,T., Kuboshima,M., Nakashima-Fujita,K., Yasuraoka,M., Nishimori,T., Kurei,S., Hachiya,T., Murakami,A., *et al.* (2011) Identification of a novel SEREX antigen family, ECSA, in esophageal squamous cell carcinoma. *Proteome Sci.*, 9, 31.
61. Jain,A.K., Xi,Y., McCarthy,R., Allton,K., Akdemir,K.C., Patel,L.R., Aronow,B., Lin,C., Li,W., Yang,L., & Barton,M.C. (2016) LncPRESS1 is a p53-regulated lncRNA that safeguards pluripotency by disrupting SIRT6-mediated de-acetylation of histone H3K56. *Mol. Cell*, 64, 967–981.
62. Chen,X., Dong,H., Liu,S., Yu,L., Yan,D., Yao,X., Sun,W., Han,D., & Gao,G. (2017) Long noncoding RNA MHENCR promotes melanoma progression via regulating miR-425/489-mediated PI3K-Akt pathway. *Am J Transl Res.*, 9, 90–102.
63. Kahles,A., Lehmann,K.V., Toussaint,N.C., Hüser,M., Stark,S.G., Sachsenberg,T., Stegle,O., Kohlbacher,O., Sander,C., Cancer Genome Atlas Research Network, & Rätsch,G. (2018) Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell*, 34, 211–224.e6.
64. Ouspenskaia,T., Law,T., Clauser, K.R., Klaeger,S., Sarkizova,S., Aguet,F., Li,B., Christian,E., Knisbacher,B.A., Le,P.M., *et al.* (2022) Unannotated proteins expand the MHC-I-restricted immunopeptidome in cancer. *Nature Biotech.*, 40, 209–217.
65. Wernaart,D., Fumagalli,A., & Agami,R. (2024) Molecular mechanisms of non-genetic aberrant peptide production in cancer. *Oncogene*, 43, 2053–2062.
66. Tang,R., Ji,J., Ding,J., Huang,J., Gong,B., Zhang,X., & Li,F. (2020) Overexpression of MYEOV predicting poor prognosis in patients with pancreatic ductal adenocarcinoma. *Cell Cycle*, 19, 1602–1610.
67. Dolstra,H., Fredrix,H., Maas,F., Coulie,P.G., Brasseur,F., Mensink,E., Adema,G.J., de Witte,T.M., Figdor,C.G., & van de Wiel-van Kemenade,E. (1999) A human minor histocompatibility antigen specific for B cell acute lymphoblastic leukemia. *J Exp Med.*, 189, 301–308.