Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

Deep conservation does not amplify proteomic detection of reverse-strand shadow ORFs in vertebrates

Yongxian Chen[1], Qian Zhao[2], Yin Zhang[1], Yabin Guo[1,*]

[1] Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangdong-Hong Kong Joint Laboratory for RNA Medicine, Medical Research Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China 510120.

[2] Department of Gynecology, Guangdong Women and Children Hospital, Guangzhou, China.

[*] Correspondence: guoyb9@sysu.edu.cn

Keywords: shadow ORF; non-canonical open reading frame (ncORF); proteogenomics; mass spectrometry; false discovery rate; conserved sequences

Abstract

Reverse-strand overlapping open reading frames ("shadow ORFs") are pervasive antisense to protein-coding genes, but whether evolutionary conservation amplifies their translation has not been tested at the class level. We catalogued shadow ORFs in 318 deeply conserved human genes and ran a pre-specified mass-spectrometry enrichment test on three size- and length-matched peptide sets of 958 each: conserved candidates (T), low-conservation reverse controls (C), and shuffled decoys (S), searched under identical criteria across six public proteomes. T exceeded S (union OR = 2.63, $P = 1.1 \times 10^{-11}$), confirming assay resolution; however, T showed no class-level enrichment over C (OR = 0.81, P = 0.076; 80% power for OR ≥ 1.35). Cross-dataset recurrence did not separate conserved from control peptides (10 vs 12 confident in ≥ 3 datasets), and one shuffled decoy recurred in four of six datasets, quantifying an empirical false-positive floor. Strand-specific RNA-seq detected sparse antisense transcription (0.9–3.9% of

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

sense) that did not track recurring peptides. The chief output is a calibrated, three-arm false-positive floor for proteogenomic searches of non-canonical ORFs, plus ten recurring peptides (led by EGLN1 and BTBD3) as uncertainty-annotated leads for targeted validation.

Significance

Reverse-strand “shadow ORFs” overlapping protein-coding genes are widespread, deeply conserved, and repeatedly proposed as a hidden source of new proteins, yet proteomic evidence for their translation remains anecdotal. Using a length-matched, power-calibrated enrichment test across six public proteomes, we show that deep conservation does not increase proteomic detection: peptides from conserved shadow ORFs are detected no more often than low-conservation controls, although both exceed shuffled decoys. This result defines a quantitative detection ceiling for antisense-frame translation, calibrates false-discovery expectations for proteogenomic searches of non-canonical ORFs, and redirects the field toward targeted validation of the few reproducible candidates tabulated here.

1. Introduction

The reverse-complement strand of protein-coding genes frequently harbours sequences able to sustain a comparatively long open reading frame. The earlier literature refers to these variously as "reverse-strand overlapping ORFs" or "antisense ORFs"; here we use the single term "shadow ORF" because the reverse-strand frame shadows the host gene on the opposite strand. Shadow ORFs belong to the antisense branch of the broad class of non-canonical open reading frames (ncORFs) that has attracted intense recent attention [1,2]. Placing shadow ORFs within the ncORF framework has one direct consequence: judgements about whether such a frame is translated, and what its product is, must be held to the evidentiary standards of the mainstream ncORF field (Methods 2.4). A validated prokaryotic exemplar exists: in *E. coli* O157:H7, an open reading frame (*pop*) fully embedded within the antisense strand of *ompA* carries complete transcriptional and translational hallmarks — promoter, transcription start site, Shine–Dalgarno sequence and terminator — and was confirmed by ribosome profiling, Western blot and a pH-dependent phenotype, establishing that an antisense overlapping frame can be a real, translatable gene

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

rather than an annotation artefact [3]. In vertebrates, however, whether reverse shadow ORFs are genuinely translated, and whether their products are biologically functional, remains fragmentary and contested; most such frames lack independent translation evidence and are difficult to separate from the null hypothesis that the sequence merely happens to remain open.

Recent evolutionary theory motivates a closer look. Antisense overlap has been shown to raise the probability that a new ORF arises and to lower the probability that, once arisen, it is lost, with the effect concentrated in one particular reading frame rather than distributed uniformly across the three [4]. A reverse frame kept open across evolution therefore need not be pure noise: a fraction may sit on a trajectory from chance emergence toward retention, so that some conserved reverse ORFs could in principle be biologically meaningful. Consistent with this, human ribosome-profiling data reveal "dual-decoding regions" in which the sense and antisense frames of one sequence may be engaged by ribosomes simultaneously [5], developmentally regulated programmes such as fission-yeast meiosis show broad non-canonical translation including antisense frames [6], and resources such as OpenProt systematically predict "alternative proteins" (altProts) from these regions [7].

Every one of these leads, however, stalls at the same epistemological threshold: sequence openness is measurable, whereas translation is only inferable with orthogonal, protein-level evidence. That a frame stays open across 430 million years is a strong sequence-level observation, but it does not by itself constitute evidence of translation. What the field lacks is a systematic, class-level test of whether evolutionary conservation actually amplifies the probability that shadow ORFs are translated and detectable by mass spectrometry. We adopt from the outset a presumption of noise: absent orthogonal evidence, the most parsimonious identity of a shadow-ORF peptide is a "junk RNA yields junk peptide" by-product, and we pre-commit to

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

reporting a result in either direction rather than presupposing that conservation implies function. To resolve this threshold we designed a falsifiable, proteome-wide enrichment test on a set of genes conserved across ~430 million years of vertebrate evolution (the human–zebrafish divergence, per TimeTree). We assembled a conservative dual-null catalog of reverse shadow ORFs from 318 conserved genes, then built three size- and length-matched peptide sets — a high-conservation test set, a low-conservation reverse control set and a shuffled decoy set — and searched them under identical, stringent criteria across six public mass-spectrometry datasets. Genuine reverse sequences proved readily distinguishable from shuffled decoys, confirming that the test has resolving power; on that calibrated baseline, however, deeper conservation conferred no class-level increase in translational detectability. A small subset of peptides recurred across three or more datasets, but the low-conservation control produced an equal number, and one shuffled decoy peptide recurred in four of six datasets; we therefore present these recurring loci as the highest-priority targets for spectrum-level validation rather than as evidence that conservation predicts translation.

2. Methods

2.1 Conserved genes and orthologous sequences

We selected six species spanning the major vertebrate lineages: human, mouse, opossum (a marsupial mammal), chicken (birds), *Xenopus* (amphibians) and zebrafish (teleost fish). Human and zebrafish diverged roughly 430 million years ago (per TimeTree), so the "cross-species conservation" demanded by this species panel is a distinctly stringent threshold.

The 318 genes studied here were screened by the authors for deep conservation rather than taken from a published gene set. Starting from the HGNC catalogue of human protein-coding genes, we queried the NCBI ortholog API for each gene and retained those with an orthologue present

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

in at least 4 of the 5 non-human species of the six-species ladder, requiring in addition that each retained orthologous CDS be 40–250% of the length of the human CDS; 318 genes passed both filters.

For each gene we extracted the orthologous coding sequence (CDS) in each species. For the 318 genes studied here, the number of orthologues obtainable per species was: human 318, mouse 316, opossum 313, *Xenopus* 311, chicken 294, zebrafish 248 — that is, the evolutionarily most distant species, zebrafish, is missing the most, consistent with the expectation that "the more distant the species, the harder it is to retain an alignable orthologue". Multiple-sequence alignment was performed with MAFFT [8]; in the alignment, J served as a placeholder and * denoted a stop codon, both of which were specifically excluded in downstream peptide processing (see 2.4).

2.2 Six-frame shadow-ORF enumeration and quantifying "how open a frame stays"

For each gene we translated in all six frames — both strands, three frames each (fwd/rev × f0/f1/f2) — and located the longest open stretch in each (denoted longest_open_aa).

Whether a reading frame stays open more often than expected by chance depends entirely on the null. We therefore built two mutually independent null models, and this "dual-null" design is central to the openness quantification. In both nulls the forward (host) protein sequence is held fixed and only the synonymous codons encoding it are re-randomised, so that any constraint acting on the host protein is fully preserved and only the reverse-strand frame is free to vary:

Uniform null: each host codon is re-encoded by a uniformly chosen synonymous codon, without regard to species codon bias;

Codon-usage null: each host codon is re-encoded by a synonymous codon drawn in proportion to species-specific codon-usage frequencies, thereby controlling for compositional

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

biases such as "some species intrinsically generate fewer stop codons". The usage weights are learned by pooling codon counts across the input CDS set (i.e. they are species-level, not per-gene); the consequence of this choice, and its relation to the per-gene GC3 explanation of long antisense frames, is examined in §3.1 and the Limitations.

The three reverse frames are defined relative to the host CDS start: rev_f0, rev_f1 and rev_f2 denote the three reading frames of the reverse complement, offset by 0, +1 and +2 nt respectively; rev_f1 is the "reverse minus-one" interlock frame that is the primary hypothesis frame (see §3.2).

Under each null we estimated, respectively, the excess percentage points by which a frame stays open (open_excess_pp, i.e. how much more open a candidate is relative to random expectation) and an amino-acid-level conservation z-score (an_z). The two nulls give a concordant conclusion: the open-excess of reverse frames is robustly positive. For rev_f0, for example, the median excess under the uniform null is +18.4 percentage points, with a departure from zero significant at $P = 4.3 \times 10^{-69}$; the codon-usage null is likewise significant (Table S1). In other words, the degree to which reverse ORFs stay open is systematically higher than it would be were the sequence under no coding constraint at all.

On this basis we computed, for each candidate, a composite conservation score (composite) integrating the percentiles of open_excess, an_z, host-gene conservation (host_cons) and the ratio of open stretch to alignment length; the composite ranges over 7.57–94.88. In addition, we set a more stringent "hard gate" (hard_gate_hi) to flag the most prominent candidates, and a total of 23 candidates met this gate.

2.3 Composition of the candidate set

The pipeline above yields 318 genes × 6 frames = 1,908 candidates, 954 on each of the forward

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

and reverse strands. By provenance, candidates fall into two parts: panel (108, drawn from known developmentally relevant or highly conserved genes, serving as controls and a positioning reference) and background (1,800).

To provisionally pinpoint the "most noteworthy" candidates at the sequence level, we ranked genes by a dual-signal ordering (open-excess percentile × conservation percentile); those ranking highly include HIVEP3, BTBD3, LSAMP and KALRN (Table S2). We further cross-checked candidates against OpenProt and found that most hotspot genes do indeed have real antisense lncRNAs (e.g. NFIB-AS1, BTBD3-AS1, LSAMP-AS1; Table S3) — consistent with "the reverse strand of these loci does carry transcriptional activity", though, equally, transcriptional activity per se does not amount to translation.

2.4 A falsifiable mass-spectrometry test of translation evidence

This is the core test of the study. It is designed to turn "are conserved candidates translated more often?" into a falsifiable proposition: if conservation genuinely improves translational detectability, we should see a higher confident-hit rate for conserved candidates than for matched controls; if we do not, the hypothesis is falsified at the class level. To this end we built three sets of size- and length-matched peptides, 958 each:

T — the conserved-candidate set (test set): reverse shadow ORFs from conserved candidates (hard-gate hits plus the highest reverse-composite tier), spanning 27 genes with composite scores 73.3–94.9;

C — the low-score reverse control set (low-conservation control): reverse shadow ORFs from low-composite candidates (composite < 50), spanning 190 genes and sampled to match T's peptide-length distribution;

S — the sequence-shuffled decoy set (negative control): T peptides with their amino acids

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

randomly shuffled (preserving length and composition), excluding any sequence identical to the reference proteome, to T, or to C.

Length-matching is a key step for excluding confounds: the peptide-length distributions of T and C are almost perfectly identical (KS D = 0.000, P = 1.000, median 14 aa), thereby ruling out the potential bias that "longer peptides are more readily detected". All three peptide sets were processed by identical rules: theoretical trypsin digestion allowing ≤ 2 missed cleavages, peptides of 7–30 aa retained, peptides containing a stop codon removed, and sequences identical to the tryptic peptides of the GENCODE v50 reference proteome [9] removed, to ensure that any detection is necessarily candidate-specific and not already known from the reference.

Searches were run with PepQuery2 [10] against six public mass-spectrometry datasets from PepQueryDB: Deep-29 and GTEx were grouped as the healthy-tissue subset; the other four datasets were Colon, LUAD, HCC and CCLE. All three sets (T/C/S) used entirely identical criteria: the primary criterion was confident (at least one PSM passing FDR and the non-synonymous-modification competition, UMS); the secondary criterion was robust (reference-database hit count n_db == 0 and $P < 1 \times 10^{-3}$, taken before UMS). We used confident as the primary criterion.

As an internal reproducibility check, before the formal test we asked whether the pipeline could re-recover the 13 reverse-shadow peptides of the NFI family that had themselves been called confident by the same PepQuery workflow on three of the same databases (Deep-29, GTEx and HCC; see §3.2). In the rev_f1 frame the pipeline re-recovered 11 of these 13 peptides. Because these 13 peptides were derived from the same search procedure and overlapping data, this gate demonstrates the reproducibility of the digestion–search–modification-competition pipeline, not its sensitivity to independently validated translation products; a genuinely external positive

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

control (e.g. synthetic-peptide spike-in or an orthogonally confirmed ncORF peptide) remains a necessary future addition. Statistically, the T vs C and T vs S comparisons used Fisher's exact test, with the effect size expressed as an odds ratio (OR) and multiple comparisons corrected by BH-FDR; for datasets involving TMT labelling, the relevant parameter bindings were separately validated.

Throughout, "pooled" denotes the union of peptides confident in at least one of the six datasets, analysed as a single 2 × 2 table; it is not a Mantel–Haenszel or random-effects combination of the six per-dataset tables. Because the six datasets share peptides, union counts fall below the independence expectation (observed 175, 207 and 75 for T, C and S against Poisson-binomial expectations of approximately 202, 240 and 89), so the union is not a sum of independent observations. With 958 peptides per set and a two-sided Fisher exact test at $\alpha = 0.05$, the design has 80% power to detect an increase from the observed C union rate of 21.6% to 27.1% (OR = 1.35); within a single dataset (baseline ≈ 4.7%) the corresponding minimum detectable effect is OR ≈ 1.71. Enrichments smaller than these are not excluded by this test.

Our translation-evidence criterion is deliberately more sensitive than the annotation-grade ncORF standard, because the two serve different aims. The authoritative annotation-grade benchmark is the TransCODE Consortium consensus (GENCODE/PeptideAtlas/HUPO-HPP) over 7,264 human Ribo-seq ncORFs [2,11]: recognising an ncORF as a protein requires the HUPO-HPP criteria — ≥ 2 uniquely mapping tryptic peptides, each ≥ 9 amino acids, protein coverage ≥ 18 residues, and protein-level FDR < 0.1% (that work achieved a peptide-level FDR of 0.0009%) — supplemented by Ribo-seq and manual spectral inspection, with a tiered system arraying weaker evidence below [2]. This study asks a class-level relative question — are conserved candidates (T) detected more often than matched controls (C) and than shuffled sequences (S) —

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

for which the requirement is that all three sets be scored on one identical benchmark, in equal numbers and matched lengths, not that any single ORF earn a protein-coding-gene annotation. We therefore set the primary criterion at ≥ 1 PSM passing FDR and the modification competition. That the annotation-grade threshold is unfriendly to short peptides is well documented: the consensus work notes that many ncORFs confirmed as translated by HLA immunopeptidomics carry no tryptic-peptide evidence at all, so that "no tryptic-peptide evidence ≠ not coding" [2]. Accordingly, the reproducible loci reported in §3.5 correspond in mainstream vocabulary to the "peptide detected, function not established" tier (the *putative* level of the Prensner system): they carry MS peptide evidence but lack strand-resolved Ribo-seq and functional data, and are leads for targeted validation rather than proteins.

Two orthogonal axes were considered but not applied: strand-resolved Ribo-seq (precomputed ribosome evidence is intrinsically sparse — in OpenProt only 7.2% of the reverse candidates carry ribosome evidence versus 56.3% with mass-spectrometry evidence, and of the 27-gene T set only 1 gene carries reverse ribosome evidence), and standard dN/dS on the dual-coding frames. The supporting rationales are given in the Discussion (Limitations), and dedicated selection testing with an overlapping-gene model is listed under Future directions.

2.5 Strand-specific total-RNA-seq test of antisense transcription

To ask whether the conserved shadow-ORF loci carry a corresponding RNA at the transcription level, we quantified antisense transcription over the 27 T-set host genes using public strand-specific total-RNA-seq. Four ENCODE [12] (Gingeras lab, CSHL) rRNA-depleted total-RNA-seq datasets on the > 200 nt fraction were used — GM12878 (ENCSR000AEC), colon (ENCSR403SZN), lung (ENCSR425RGZ) and liver (ENCSR504QMK) — all stranded paired-end 101 nt libraries. Reads were aligned to GRCh38 with STAR 2.7.11b [13]; uniquely mapping

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

rates were 85.5–93.7% (105.1 M, 74.8 M, 47.8 M and 52.8 M input reads for GM12878, colon, lung and liver respectively). For each gene (GENCODE v50 coordinates [9], exon union) we counted uniquely mapping reads (MAPQ ≥ 1) whose inferred transcript strand — taken from READ1 under the reverse (dUTP) convention, one count per read pair — overlapped the sense or the antisense strand, and computed RPKM and the antisense fraction. The dUTP (reverse) library orientation was verified empirically for every sample: the highly expressed housekeeping gene GAPDH (present in the 318-gene panel) showed sense ≫ antisense only under the reverse assignment (antisense fraction 0.0001–0.0031 across the four datasets), and the mirror image under the forward assignment; the same reverse assignment also reproduces STAR's own stranded GeneCounts (sense-count Pearson r = 0.97–0.999 per tissue). This assay reads the transcription layer only: an antisense RNA is necessary but not sufficient for a shadow peptide, and its detection makes no translation claim, let alone a function claim.

3. Results

3.1 Reverse shadow ORFs remain open far beyond dual-null expectations across 430 million years of vertebrate evolution

The 318 genes analysed here are not drawn from any published gene set: they were screened by the authors for deep cross-species conservation from the HGNC human protein-coding catalogue via the NCBI ortholog API (selection criteria in 2.1). Across the 318 genes conserved over 430 million years, reverse-strand shadow ORFs remain structurally open to a degree significantly beyond random genomic expectation. That is, at the sequence level, the "staying open" of reverse-strand frames (i.e. a longest open reading frame significantly longer than random expectation) is not accidental. Having translated each gene, strand and frame (fwd/rev × f0/f1/f2) and located the longest open stretch, we quantified "open excess" with two mutually independent

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

null models: one shuffling under uniform amino-acid composition (uniform), the other under species-specific codon-usage frequencies (usage). The two nulls give a concordant conclusion — the open-excess (open_excess) of reverse frames is robustly positive. For rev_f0, for example, the median open-excess relative to random expectation under the uniform null is +18.4 percentage points, with a departure from zero significant at $P = 4.3 \times 10^{-69}$ (Table S1; the codon-usage null is likewise significant). In other words, the degree to which reverse ORFs stay open is systematically higher than it would be were the sequence under no coding constraint. The full per-candidate tiering and composite-score ranking is given in Table S4, and the robustness of the open-excess conclusion to the choice of null model in Figure S1.

It is important to distinguish this result from translation: the observation is at the sequence / transcription level, answering whether the frame is kept open, not whether it is translated into protein. A reverse frame kept open across evolution is therefore worth asking further whether it has left any trace at the protein level. The length ratios and tiered distributions are shown in Figure 1, and the positioning of candidates within the background distribution in Figure 2.

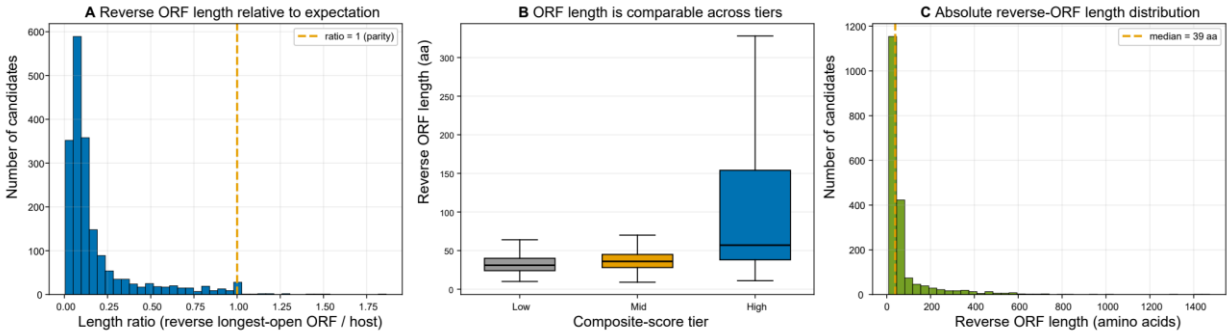


Figure 1. Length ratios and tiered length distributions of reverse shadow ORFs. (A) Length ratio of each candidate's reverse longest open reading frame relative to random expectation. (B) ORF amino-acid-length distribution within each tier after tiering by composite score. (C) Absolute reverse-ORF amino-acid-length distribution across all 1,908 candidates. (Source data: Table_S4_candidates_tiered_ranked.csv; corresponding text 2.3, 3.1.)

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

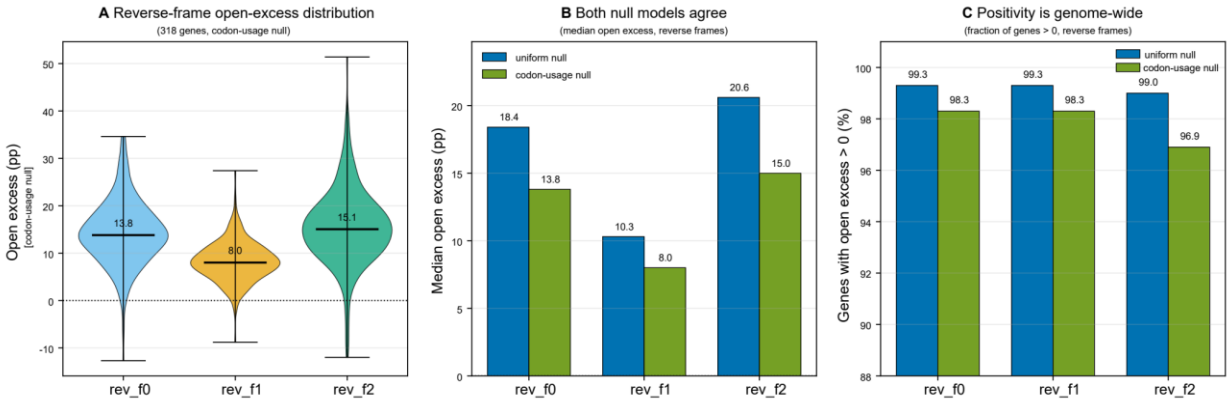


Figure 2. Background survey of how "open" reverse frames stay. For 318 genes conserved across 430 million years, the longest open reading frame was located in each of the six frames (fwd/rev × f0/f1/f2), and "open excess" (open_excess) was quantified under two independent null models (uniform: shuffling under uniform amino-acid composition; usage: shuffling under species-specific codon-usage frequencies). This figure shows the open-excess of the three reverse frames; the open-excess of reverse frames is robustly positive, systematically higher than the random expectation under no coding constraint. (A) Per-candidate open-excess distribution of the three reverse frames under the codon-usage null (medians 13.8 / 8.0 / 15.0 pp for rev_f0 / rev_f1 / rev_f2). (B) Median open excess of the three reverse frames under both null models (uniform 18.4 / 10.3 / 20.6 pp; usage 13.8 / 8.0 / 15.0 pp): the two nulls give a concordant, robustly positive result. (C) Fraction of genes with positive open excess per reverse frame (96.9–99.3%): the positivity is genome-wide, not driven by a few outliers. (Source data: Table_S1_background_frame_summary.csv, Table_S4_candidates_tiered_ranked.csv; corresponding text 3.1.)

3.2 The NFI family as the motivating case and an internal reproducibility check

The NFI family (nuclear factor I [14]; four deeply conserved vertebrate members NFI-A/B/C/X) was the motivating source of this study: in our tiering, the open-excess of NFIB's rev_f1 frame lies at the 99th percentile of all candidates (Table S5), making NFIB one of the most prominent candidates if one takes "an anomalously well-kept-open reverse frame" as the lead. The antisense-frame conservation seen in NFIA/NFIB extends to the other zebrafish-reaching family members NFIC and NFIX (Figure S2), which serve as family context only.

The NFI family also provided an internal reproducibility check for the mass-spectrometry pipeline. The 13 NFI reverse-shadow peptides used for this check were themselves called

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

confident by the same PepQuery workflow on three of the same databases (Deep-29, GTEx and HCC); in the rev_f1 frame the pipeline re-recovered 11 of these 13. Because these peptides derive from the same search procedure and overlapping data, this check demonstrates the pipeline's reproducibility — that it does not return a blanket negative for every input — but it is not an independent positive control and does not by itself establish sensitivity to genuine reverse translation products (an external spike-in or orthogonally validated control remains future work; see Limitations). One of these peptides, HTLSAWR, is only 7 amino acids long and is correspondingly down-weighted in our interpretation (see §3.5 and Limitations).

The NFI family also provides an intuitive illustration of a reverse frame that is both kept open and highly conserved across species (Figure 3). Taking the reverse rev_f1 translation of each species' orthologous CDS and aligning with MAFFT, one can see that NFIB's reverse open reading frame reaches a mean pairwise identity of 95.6% across the open-reading-frame region among the 5 species from human to *Xenopus* (about 350 million years); and 5 of the 8 NFIB peptides confident in at least two datasets — including the positive-control peptide SSSLTALSSSFDIR — fall exactly within the displayed human open reading frame (red bars in Figure 3A). We note that NFIB, despite being the sequence-level extreme of the catalogue, contributes no peptide confident in three or more datasets (Table 4); its prominence here is at the level of frame openness and reverse-frame conservation, not of proteomic recurrence. Two boundaries must be stressed. First, NFIB has no zebrafish orthologue in Ensembl Compara, so its alignment stops at *Xenopus*; "human to zebrafish, about 430 million years" is a span at the whole-catalogue level, not for NFIB individually. To display the conservation depth that "reaches the fishes" intuitively, Figure 3B instead uses NFIA as an example — its rev_f1 open reading frame extends to zebrafish (about 430 million years), with an open-reading-frame-region identity

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

of 89.9% across 6 species, but NFIA's confident mass-spectrometry peptides map outside the rev_f1 frame, so Figure 3B illustrates sequence conservation only and makes no translation claim. Second, sequence conservation is only a visualisation of the measurable fact that "the frame is kept open"; it does not amount to that reverse frame being under functional selection — the distinction between the two is given in the Discussion (Limitations).

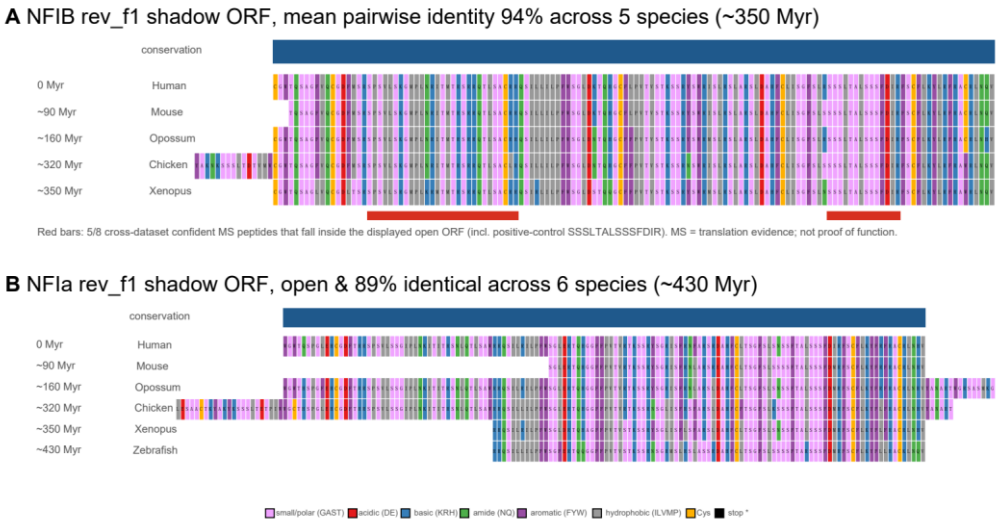


Figure 3. A cross-species alignment example of NFI-family reverse shadow ORFs. The reverse second frame (rev_f1) translation of each species' NFI orthologous CDS was taken and aligned with MAFFT; amino acids are coloured by physicochemical class, and species are ordered top to bottom by their approximate divergence time from human (Myr). (A) NFIB, 5 species from human → mouse → opossum → chicken → Xenopus (about 350 million years); the mean pairwise identity of its reverse open-reading-frame region is 95.6%. Red bars mark those of the 5 (of 8) NFIB peptides confident in at least two datasets that fall within the displayed human open reading frame (including the positive-control peptide SSSLTALSSSFDIR); because some peptides are head-to-tail adjacent, they visually merge into a few segments. In this panel, MS = translation evidence, not proof of function. (B) NFIA, extending to zebrafish across 6 species (about 430 million years); the reverse open-reading-frame-region identity is 89.9%. NFIA's confident mass-spectrometry peptides map outside the rev_f1 frame, so this panel illustrates sequence conservation only and makes no translation claim. (Source data: rev_f1 translations of each species' orthologous CDS aligned with MAFFT, peptide mapping taken from shadow_ms_pepquery_results.csv; image file fig_nfib_shadow_alignment.svg; corresponding text 3.2.)

3.3 Deep evolutionary conservation does not confer class-level translational amplification

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

on shadow ORFs

On a power-calibrated mass-spectrometry test across six public proteomes, deep sequence conservation does not raise the class-level translational detectability of reverse shadow ORFs (T vs C pooled OR = 0.81, P = 0.076, n.s.). With a sensitive detection pipeline (§3.2) and confirmed sequence-level openness (§3.1) in hand, this was the study's central falsifiable question — does deeper conservation translate into higher translational detectability? — and where Figure 4 sits. We built three sets of size- and length-matched peptides, 958 each: T (test set, reverse shadow ORFs of super-conserved candidates, 27 genes, composite 73.3–94.9), C (control set, reverse shadow ORFs of low composite, 190 genes, composite < 50, sampled to match T's peptide-length distribution) and S (negative control, T peptides randomly shuffled by amino acid, preserving length and composition). The peptide-length distributions of T and C are almost perfectly identical (KS D = 0.000, P = 1.000, median 14 aa), thereby ruling out the confound that "longer peptides are more readily detected". The three sets were searched with entirely identical criteria across 6 public mass-spectrometry datasets, the primary criterion being confident (≥ 1 PSM passing FDR and the non-synonymous-modification competition).

The per-dataset confident-hit rates (denominator 958 throughout) are as follows (per-dataset search statistics and quality-control metrics in Figure S3 and Table S6):

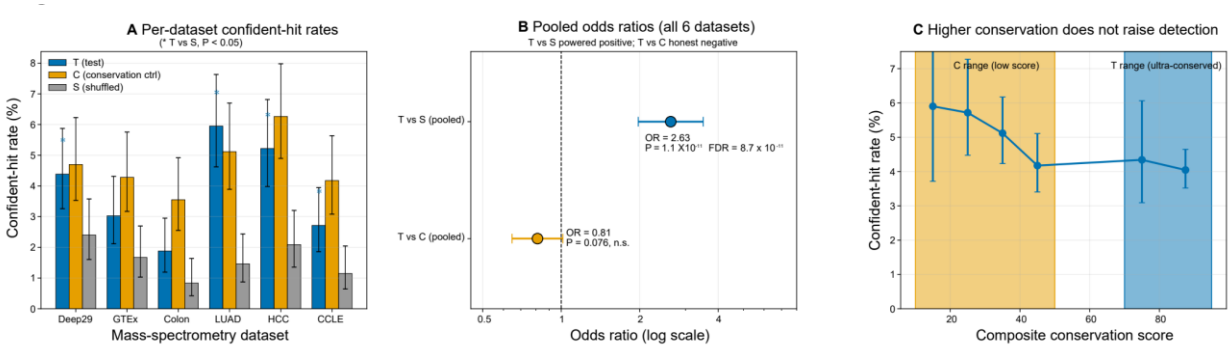


Figure 4. The mass-spectrometry enrichment test: genuine reverse sequences carry a detectable

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

signal, but conservation does not amplify translational detectability at the class level. T (test set), C (low-conservation control) and S (shuffled negative control), 958 size- and length-matched peptides each, were searched under entirely identical criteria (confident: ⩾ 1 PSM passing FDR and the non-synonymous-modification competition) across 6 public mass-spectrometry datasets. (A) Per-dataset confident-hit rates for the three sets; error bars are Wilson 95% confidence intervals, and asterisks mark a T-versus-S difference at $P < 0.05$. (B) Pooled odds ratios for T versus S and T versus C across the six datasets: T vs S gives OR = 2.63 ($P = 1.1 \times 10^{-11}$, FDR = 8.7×10^{-11}) and T vs C gives OR = 0.81 (P = 0.076, n.s.). (C) Confident-hit rate vs composite conservation score: the detection rate does not rise, but slightly falls, with conservation (Spearman $\rho = -0.030$, P = 0.0016); error bars are Wilson 95% CIs. (Source data: Table_S7_shadow_ms_enrichment_summary.csv, Table_S4_candidates_tiered_ranked.csv; corresponding text 3.3, 3.6.)

Table 1. Per-dataset confident-hit rates. Confident-hit rates for the T, C and S sets in each of the six public mass-spectrometry datasets (denominator 958 peptides per set), with T-versus-C and T-versus-S odds ratios and Fisher P per dataset. (Source file: Table_S7_shadow_ms_enrichment_summary.csv)

Dataset	T%	C%	S%	T vs C	T vs S
Deep-29 (healthy)	4.38	4.70	2.40	OR = 0.93, P = 0.83	OR = 1.86, P = 0.022
GTEx (healthy)	3.03	4.28	1.67	OR = 0.70, P = 0.18	OR = 1.84, P = 0.069
Colon	1.88	3.55	0.84	OR = 0.52, P = 0.034	OR = 2.27, P = 0.074
LUAD	5.95	5.11	1.46	OR = 1.17, P = 0.48	OR = 4.27, $P = 1.6 \times 10^{-7}$
HCC	5.22	6.26	2.09	OR = 0.82, P = 0.38	OR = 2.58, $P = 3.4 \times 10^{-4}$
CCLE	2.71	4.18	1.15	OR = 0.64, P = 0.10	OR = 2.40, P = 0.019

Pooling all 6 datasets, and looking only at the two healthy-tissue datasets (Deep-29 + GTEx), gives (full statistics in Table 2; full per-dataset and pooled statistics in Table S7):

Table 2. Pooled T/C/S confident-hit rates and enrichment statistics. Hit counts, hit rates, odds ratios, Fisher P and BH-FDR for the T/C/S sets in the pooled strata (all_datasets_union, healthy_union). (Source file: Table_S7_shadow_ms_enrichment_summary.csv)

Stratum	Comparison	T	Control	OR	Fisher P	BH-FDR
All 6 pooled	T vs C	18.27%	21.61%	0.81	0.076	0.27

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

All 6 pooled	T vs S	18.27%	7.83%	2.63	1.1×10^{-11}	8.7×10^{-11}
Healthy (Deep-29 + GTEx)	T vs C	7.31%	8.56%	0.84	0.35	0.50
Healthy (Deep-29 + GTEx)	T vs S	7.31%	3.97%	1.91	0.002	0.004

T significantly exceeds S, and does so with a consistent direction across all 6 datasets (OR 1.86–4.27), with a pooled OR = 2.63 ($P = 1.1 \times 10^{-11}$, FDR = 8.7×10^{-11}). The shuffled set S preserves T's length and amino-acid composition, yet is systematically harder to detect. Two things must be separated here. First, this comparison establishes that the search separates real reverse-frame sequence from composition-matched shuffled sequence, which bounds the assay's statistical resolution; because C (real reverse sequence) is detected at least as often as T, the parsimonious interpretation of the T-over-S excess is a sequence-authenticity effect — real sequences match paralogues, isoforms, SNV/PTM variants and other near-cognate spectra more readily than random shuffles — rather than evidence of translation per se. Second, the shuffled set S is by construction absent from any proteome, yet 7.83% of S peptides (75/958) still pass the confident criterion: this is a substantial assay-level false-positive floor, and we treat it as the empirical estimate of the class-level FDR for this search (see §3.5 and Limitations).

T is not higher than C. Of the 6 datasets, 5 show T < C and only 1 (LUAD) shows T > C, and none attains a significant T > C; the pooled OR = 0.81 (P = 0.076, not significant), and the secondary criterion robust likewise shows no enrichment of T (threshold-sensitivity sweeps in Figure S4 and Table S8; full peptide-level results in Dataset S1). That is, "higher conservation / higher composite" did not, at the class level, push the translational detectability of reverse shadow ORFs above that of ordinary reverse ORFs.

Three robustness checks bear on this null. First, power: with 958 peptides per set the design has 80% power to detect a union-level OR of 1.35 or greater (§2.4), so the test excludes moderate-to-large enrichment but not OR in the 1.1–1.3 range; the null should be read as "no class-level

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

amplification at or above OR ≈ 1.35", not as "no difference of any size". Second, length: restricting to peptides ≥ 9 aa (the HUPO annotation-grade length floor) leaves the null unchanged (T 133/810 = 16.4% vs C 158/810 = 19.5%, OR = 0.81, P = 0.12), while T still exceeds S (OR = 2.41, P = 4.1×10^{-8}). Third, clustering: because the 958 T peptides derive from only 27 genes (versus 190 for C), peptide-level Fisher tests could overstate significance; a gene-cluster-preserving permutation test (20,000 permutations of gene labels) confirms that the T-versus-C null is robust to this clustering (two-sided P = 0.26). The human default prior for the reverse strand is "mostly empty", so "conserved candidates are translated no more often than ordinary reverse ORFs" contradicts no established prior; the value of this calibrated baseline is that it sets a quantitative floor against which the handful of reproducible single loci below must be weighed (§3.5).

3.4 Higher host-gene expression in conserved candidates refutes abundance-driven signal masking

A potential confound for this class-level null is the baseline expression of the host genes: if the host genes of the T set were, overall, expressed at lower levels, then T peptides would be intrinsically harder to detect by mass spectrometry, potentially masking an enrichment that ought to be present. We ruled this out using the "RNA tissue distribution" field of the Human Protein Atlas [15] (coverage 100%), and the direction is exactly the opposite:

Table 3. Expression breadth of T and C host genes (Human Protein Atlas). Fraction of T and C host genes detected in all / most / some tissues in the Human Protein Atlas RNA tissue-distribution field (T 74.1% vs C 51.6% detected in all tissues, Mann-Whitney P = 0.038).

Expression breadth (HPA)	T host genes	C host genes
Detected in all tissues	74.1%	51.6%

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

Detected in most tissues	14.8%	30.0%
Remainder	11.1%	18.4%

The host genes of T are more broadly expressed (74.1% detected in all tissues, versus 51.6% for C; Mann–Whitney P = 0.038; the HPA classifies each gene as detected in all, most, or a subset of the tissues assayed — "most" means detected in more than half but not all, and the remainder are detected in fewer than half). Consistent with this, among the three sets' raw spectrum-match rates, T (88–99%) and C (90–99%) very nearly coincide dataset by dataset, whereas S is clearly lower (79–97%) — indicating that the input peptides of T and C are equivalent at the basal level of "being matchable at all", and that S is the true noise floor.

Putting the two together: T sits on genes that are more highly expressed and ought therefore to be more readily detected, yet is still confirmed no more often than C. This means the direction of the expression-abundance confound is opposite to "having masked T's enrichment" — far from explaining away the "no class-level amplification", it makes that judgement more robust. See Figure 5.

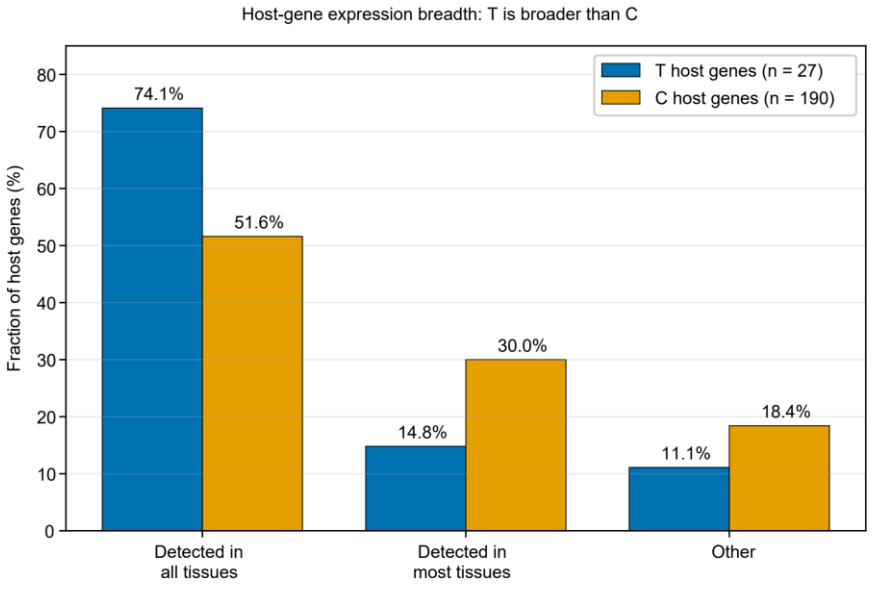


Figure 5. Screening for the expression-abundance confound: the direction is opposite to

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

"masking". Based on the RNA tissue distribution of the Human Protein Atlas, the expression breadth of T and C host genes is compared. T host genes have a higher fraction "detectable in all tissues" (T 74.1% vs C 51.6%, Mann–Whitney P = 0.038), i.e. T sits on more highly expressed genes that ought to be more readily detected, yet is still confirmed no more often than C. (Source data: Table_S7_shadow_ms_enrichment_summary.csv and the Human Protein Atlas RNA tissue-distribution field 15; corresponding text 3.4.)

3.5 Cross-dataset reproducibility does not separate conserved from non-conserved candidates

We next asked whether the subset of peptides confident in more than one dataset behaves differently from the class as a whole (full counts in Table S9; per-peptide results in Dataset S1). Of the 958 T peptides, 175 were confident in at least one dataset, 32 in at least two, 10 in at least three, 4 in at least four and 1 in five; the corresponding C counts were 207, 50, 12, 0 and 0, and the S counts 75, 15, 1, 1 and 0 (Table S9, Figure S5).

At every level up to three datasets, C equals or exceeds T: the odds ratio for reaching at least three datasets is 0.83 (T 10/958 vs C 12/958, P = 0.83), and conditional on being confident in at least one dataset the proportions reaching three or more are indistinguishable (T 5.7%, C 5.8%). Recurrence across datasets therefore does not discriminate high- from low-conservation candidates.

Recurrence is also not, by itself, evidence against chance. Under a Poisson-binomial model assuming independent datasets, the expected number reaching three or more datasets is 0.9 (T), 1.7 (C) and 0.07 (S); the observed counts exceed these expectations by 10.7-, 6.9- and 14.1-fold respectively. The excess is largest for the shuffled set, and one S peptide — a sequence that by construction is present in no proteome — was confident in four of the six datasets. Recurrence across datasets thus reflects peptide-level heterogeneity in detectability (ionisation efficiency, low sequence complexity, shared spectra) that applies equally to true and decoy-like sequences, and the independence expectation is not a valid null for it.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

The one asymmetry favouring T is confined to the upper tail: 4 T peptides reached four datasets and 1 reached five (ALAAGGR, EGLN1 rev_f0), whereas no C peptide exceeded three. This difference in the shape of the distribution is in the expected direction but rests on four peptides, does not reach significance (P = 0.12 vs C), and is accompanied by one S peptide at the same level. A further caveat is length: 6 of the 10 T peptides confident in ≥ 3 datasets are shorter than 9 aa (including ALAAGGR, 7 aa, and EAASGGR, 7 aa), below the HUPO annotation-grade length floor and rich in low-complexity Gly/Ala/Ser. Short, low-complexity peptides are exactly the class most prone to spurious matching, and proteogenomic searches against nucleotide-derived databases are known to require class-specific FDR control because database inflation distorts the global FDR [16,17,18]; the 7.83% shuffled-set floor (§3.3) is our empirical estimate of that class-level rate. We therefore treat these peptides as the highest-priority candidates for targeted validation rather than as class-level evidence of translation. Summarised by gene, the confident-peptide counts rank as SEC61A1 (18), BTBD3 (17), BMAL1 (17), NFIA (15) and EGLN1 (12); the ten ≥ 3-dataset peptides are listed individually in Table 4. These two orderings do not agree: SEC61A1 and NFIA lead on total confident peptides but contribute none of the ten recurrent peptides, whereas CTCF, HNRNPA1 and LHFPL4 contribute recurrent peptides without entering the top five. Total confident-peptide count scales with the length of the reverse frame and hence with the number of searchable peptides; it is not a measure of reproducibility, and the two should not be conflated.

It should be noted that the most reproducible peptides — ALAAGGR, AAGSTSGIR, EAASGGR — are all low-complexity sequences rich in Gly/Ala/Ser, which are themselves more prone to chance matching, and the low-conservation C set produces a comparable number of recurring peptides (12 at ≥ 3 datasets). Cross-dataset confidence raises their credibility above the

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

shuffled baseline but does not amount to confirmed translation, let alone function, and does not by itself mark the conserved candidates as special. We therefore present these recurring peptides (and their loci, represented by EGLN1 and BTBD3) as calibrated, uncertainty-annotated candidate leads for targeted validation, not as confirmed functional translations.

Table 4. The ten T peptides confident in ⩾ 3 of the six datasets. (Source file: shadow_ms_pepquery_results.csv)

Peptide	Gene	Frame	Length (aa)	Datasets confident (n)	Datasets
ALAAGGR	EGLN1	rev_f0	7	5	CCLE, Colon, GTEx, HCC, LUAD
AAGSTSGIR	BTBD3	rev_f1	9	4	CCLE, GTEx, HCC, LUAD
EAASGGR	CTCF	rev_f2	7	4	CCLE, GTEx, HCC, LUAD
DGAEVGEGAPAGELAAR	LHFPL4	rev_f0	17	4	CCLE, Colon, Deep29, LUAD
SVGPGDMK	BMAL1	rev_f1	8	3	CCLE, HCC, LUAD
VGSPPSR	BMAL1	rev_f1	7	3	CCLE, GTEx, HCC
KMSLVSLR	BTBD3	rev_f1	8	3	Colon, Deep29, LUAD
RLLLAGLGFGSHR	EGLN1	rev_f0	13	3	CCLE, Colon, HCC
LIVVIPK	HNRNPA1	rev_f0	7	3	CCLE, HCC, LUAD
VLASTAIGAR	HNRNPA1	rev_f0	10	3	CCLE, Colon, HCC

3.6 Detection rate is flat to slightly declining across conservation bins

We also asked whether the strict conservation filter itself suppressed the yield of detectable signal — that is, whether the largely class-level-null result is an artefact of an over-stringent threshold. It is not, and relaxing the threshold is expected to recover rather than lose candidate signal (Figure 4C, Figure 6). Two observations, both read off the existing data, point the same way.

Detection rate does not rise with conservation — if anything it slightly falls. Pooling the T and C peptides and binning by composite conservation score from low to high, the confident-hit rate

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

follows a trend of 5.9% → 5.7% → 5.1% → 4.2% (low bins) … down to 4.0% at the highest bin; the row-level Spearman $\rho = -0.030$ ($P = 0.0016$), and this weak negative also holds within the C set ($\rho = -0.033$, $P = 0.016$). The detection rate in the low-conservation C region is thus even slightly higher than in the super-conserved T region.

The candidate pool only expands as the required conservation depth decreases. Even within this 318-gene set already prescreened as deeply conserved, relaxing the requirement from "all vertebrates" to "mammals only" increases the number of genes that stay open from 215 to 312 (+97, +45%); because this set is already highly conserved, this increase is only a lower bound, and the true genome-level expansion would be larger still.

Putting the two together, on relaxing to the amniote/mammal range "the per-candidate detection rate does not fall" and "the candidate pool expands" point in the same direction, which motivates re-mining at the amniote/mammal range (Future directions). The effect size of the trend is weak ($\rho \approx -0.03$, and its P value is inflated by peptide-level pseudoreplication across the 27/190 genes; see §3.3) and the candidate-pool expansion is a lower-bound estimate within an already deeply conserved set, so this is a candidate direction, not an established conclusion.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

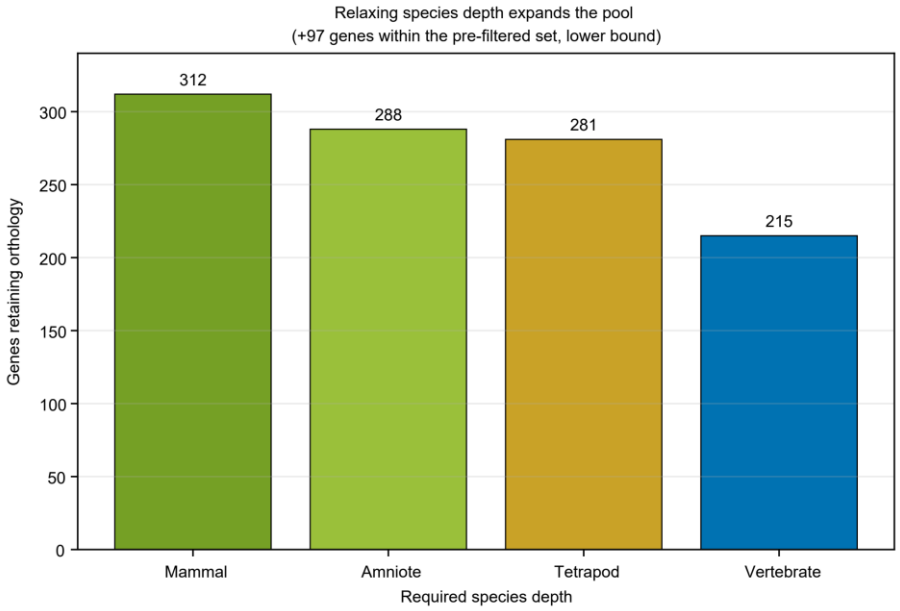


Figure 6. Directional check on the conservation threshold (supporting the "relax conservation" future direction). As the required species depth is progressively relaxed across four conservation tiers — mammals only / amniotes / tetrapods / all vertebrates — the number of candidate genes passing the conservation gate expands (312 / 288 / 281 / 215 respectively). (Source data: Table_S4_candidates_tiered_ranked.csv; corresponding text 3.6.)

3.7 Strand-specific total-RNA-seq detects antisense RNA at conserved loci, but at low abundance that does not track the reproducible MS loci

The mass-spectrometry test reads the protein layer; we therefore asked whether the same loci carry a corresponding RNA at the transcription layer, quantifying antisense transcription over the 27 T-set host genes in four strand-specific total-RNA-seq datasets (GM12878, colon, lung, liver; §2.5). Antisense RNA is detectable but sparse: 26 of the 27 loci show at least one antisense read in at least one tissue, and 13 show antisense reads in all four — yet the abundance is low, with a per-tissue median antisense level of 0.0009–0.010 RPKM (pooled median 0.0061), i.e. only 0.9–3.9% of the corresponding sense level. Only 8 of the 27 loci exceed 0.1 RPKM in any tissue (maximum 0.456, RPS21 in lung); the highest mean antisense levels are at CHMP5 (0.215) and RPS21 (0.186) (Table S10; per-tissue values for all 318 genes in Dataset S2).

Critically, the antisense-RNA level does not track the reproducible mass-spectrometry loci. The

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

six loci with peptides confident in ≥ 3 datasets (§3.5) show, if anything, *lower* antisense RNA than the rest of the T set (pooled median 0.0037 vs 0.0090 RPKM; lower in each of the four tissues). EGLN1 — the single most reproducible MS locus — is detectable in all four tissues but never exceeds 0.027 RPKM; NFIB never exceeds 0.0014. The one MS locus with a higher antisense-RPKM value, HNRNPA1 (0.124 in GM12878), is a very highly expressed gene whose antisense reads amount to only 0.67% of its own sense transcription. Thus the loci that recur at the protein layer are not the loci with abundant antisense RNA.

This is a transcription-level observation and is read accordingly. That antisense RNA is detectable at most conserved loci is consistent with the OpenProt cross-check (§2.3) that these loci carry real antisense transcripts; that it is sparse, and that it does not track the reproducible MS loci, is consistent with the paper's presumption of noise — most antisense transcription is expected to be non-productive. An antisense RNA is necessary but not sufficient for a shadow peptide: its presence makes no translation claim, and its scarcity at the MS-reproducible loci neither confirms nor refutes their peptides, which remain calibrated candidates for targeted validation (§3.5).

4. Discussion

The central result of this study is that "staying open" and "being translated" are two questions at different levels, each with its own answer. At the sequence level, reverse frames stay open to a systematically positive excess (§3.1) — a robust, measurable premise. At the protein level we find no class-level amplification of "conserved → translated more often" (§3.3), and cross-dataset recurrence — the property one would intuitively use to separate signal from noise — does not distinguish conserved from low-conservation candidates either (§3.5). The study therefore rests on a negative result that is calibrated rather than merely null, and it does not

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

equate the openness of the sequence layer with translation at the protein layer. This distinction also fixes the relationship to the theory of de novo gene birth [4]. Antisense overlap has been argued to raise the probability that a new ORF arises and lower the probability of its loss, concentrated in one frame — a sequence-level mechanism. Our observation is not in tension with this: reverse frames do stay open more often. But sequence-level openness does not automatically become class-level translational enrichment, and the two levels must be argued separately. Whether any individual locus has completed the path from emergence through retention to occasional translation cannot be decided from recurrence statistics alone, and is left to targeted validation.

Our class-level null must also be read against the evolutionary analysis of the TransCODE consensus itself [2], which we cite for the annotation standard. That work introduced an ORF-level relative-branch-length (ORBL) measure and reported that evolutionary constraint is common among ncORFs and correlates positively with peptide observability — superficially the opposite of our class-level null. Three considerations reconcile the two. First, the conservation metrics differ: ORBL is a branch-length measure at the ORF level, whereas our composite mixes open-excess, reverse-frame conservation, host-gene conservation and a length ratio, so the two studies rank candidates on different axes. Second, the ORF classes differ: TransCODE is dominated by 5′-UTR and lncRNA ncORFs, whereas we test only antisense-strand overlapping frames, which are a minority with distinct constraints. Third, mass spectrometry is intrinsically insensitive to most noncanonical proteins [19], so a positive ORBL–observability correlation across all ncORFs can coexist with a null enrichment within the harder antisense subclass. We therefore do not read our null as contradicting the consensus, but as delimiting where within it antisense shadow ORFs fall.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

The class-level null is power-calibrated, and the reproducible subset must be weighed against the false-positive floor rather than assumed genuine. First, self-demonstrated resolution: in the same data the T-versus-S difference is strongly significant (pooled $P = 1.1 \times 10^{-11}$), so "no class-level difference between T and C" is a real null effect rather than a false negative — though this resolution reflects sequence authenticity, not sensitivity to translation (§3.3). Second, a reversed confound: T's host genes are more broadly expressed and should be more readily detected, yet still show no class-level enrichment, so the null was not manufactured by low expression (§3.4). Third, criterion consistency: T/C/S used identical criteria, and the pipeline re-recovered 11/13 NFI peptides in an internal reproducibility check (§3.2). Against this, however, stands the shuffled-set floor: one S peptide recurred in four of six datasets, and 6 of the 10 T peptides reproducible in ≥ 3 datasets are shorter than 9 aa (§3.5). The recurring loci are therefore best regarded as the highest-priority candidates for targeted validation — the subset hardest to attribute to single-dataset chance — rather than as a class-level demonstration of translation.

The class-level null does not preclude translation at individual loci. The reproducible peptides in EGLN1 and BTBD3 remain candidates for targeted validation, but they do not support a general claim that conserved reverse ORFs are broadly translated, nor that their function is proven. The strand-specific RNA assay (§3.7) adds the transcription layer to this picture without disturbing it: antisense RNA is detectable at most conserved loci but is sparse, and it does not track the reproducible MS loci — so the transcription and protein layers each read the loci differently, and neither amounts to evidence of function. No amplification at the class level coexists with reproducible signal at particular loci; whether those signals correspond to real functional proteins is left to targeted validation, with the chance matching of low-complexity peptides a residual risk. Class-level detectability and the prior carried by conservation are likewise distinct and not

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

in conflict. That T peptides are confirmed no more often than C peptides — despite T host genes being the more broadly expressed of the two — shows only that cross-species conservation did not amplify reverse-frame detectability at the class scale; it does not negate the separate proposition that conservation — especially a longer open reading frame — raises the prior probability that a sequence has function. Long non-coding RNAs offer an instructive analogy: most lncRNAs are gained and lost rapidly and are seldom sequence-conserved, yet comparative-genomics evidence consistently shows that conserved lncRNAs are more likely than non-conserved ones to be functional, even though they are not necessarily more highly expressed [20]; conserved lncRNAs tend to have longer gene bodies, more exons and stronger disease associations [21], and conserved segments within lncRNAs carry about threefold more translational evidence than non-conserved segments [22]. Transposed to shadow ORFs, the reading is cautious and concrete: candidates that are both anomalously well-kept-open and deeply conserved, and that are longer (such as the NFIB-related locus), warrant prior suspicion of carrying function, while non-conserved short frames are more likely noise — but this only raises the prior and sets priority, and is not proof of function. Conservation and greater length move certain candidates to the front of the validation queue; whether they are genuinely translated and functional is adjudicated only by targeted experiments (below).

These results sit squarely within the paper's presumption of noise. Under the rigor gradient replication > transcription > translation, translation is the least precise of the three processes — its error rate exceeds that of transcription by roughly two orders of magnitude [23], and transcription itself is already error-prone at the level of RNA polymerase fidelity and proofreading [24,25,26,27] — so "junk RNA yields junk peptide" is the reasonable null, and absent orthogonal evidence a shadow peptide is treated as noise. The mass-spectrometry test advances

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

the evidence locus by locus: for the vast majority of candidates the null is not overturned, fully consistent with the prior that the reverse strand is mostly empty and not to be read as a failure; for a handful of loci the peptides recur across datasets, but because the low-conservation control recurs at the same rate, recurrence alone does not advance them beyond "predictable" — only orthogonal, spectrum-level evidence can. This framework is structurally analogous to the evidence standard the mainstream ncORF field has adopted, which insists on a distinction word-for-word consistent with our pre-specified threshold: *protein identification (detecting a polypeptide molecule) ≠ protein-coding gene annotation (that product having biological function)* [2,11]. The consensus does not deny translational noise — it introduces the term peptidein for the intermediate state of "confirmed translated, function not yet determinable", noting that this class may include transient products of cellular stress or defective-ribosome translation [2]; noise is not counted as expression or function but is listed separately and left for further evidence to upgrade or exclude. In that vocabulary the mapping is exact: for most conserved candidates the evidence rests at the "translational trace only / undetermined" tier — a field expectation, not a failure — while for the reproducible loci it reaches the *putative / peptidein* magnitude: peptide detected, function undetermined, awaiting validation.

Limitations

Two limitations bear on the openness result specifically. First, long non-stop frames on the antisense strand have been reported since the 1990s and were attributed largely to codon-usage and GC3 bias of the sense strand rather than to selection on the antisense frame [28,29,30,31,32]. Our codon-usage null holds the host protein fixed and randomises only synonymous codons, but it learns usage weights by pooling across the CDS set (species-level) rather than per gene (§2.2); a per-gene GC3-matched null would be required to exclude the compositional explanation fully.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

We therefore tested the compositional account directly: across the 300-gene background, open excess is essentially uncorrelated with host GC3 under the uniform null (Spearman $\rho \approx 0.0$, n.s.) and is, if anything, *negatively* correlated with GC3 under the codon-usage null ($\rho = -0.25$ to -0.38, $P < 10^{-5}$) — the opposite direction from the compositional prediction that high-GC3 genes should yield longer antisense frames. Moving from the uniform to the codon-usage null does reduce the median open excess by 22–27% (18.4 → 13.8, 10.3 → 8.0 and 20.6 → 15.0 pp for rev_f0, rev_f1 and rev_f2), so a partial compositional contribution cannot be excluded, but the sign of the GC3 correlation argues against GC3 bias being the dominant driver. Second, mass spectrometry is known to miss the majority of noncanonical proteins for reasons that are partly biological and partly statistical [19]; our class-level null should therefore be read as an upper bound on detectable translation, not as evidence of absent translation.

This study has several further limitations that must be made explicit. First, the range of frames: this paper covers only reverse frames (the present study); forward shadow ORFs are a planned follow-up and are not included here. Second, the intrinsic limitations of shared mass-spectrometry data: whether or not something is detected is affected by sample type, modification-parameter settings, spectral depth and so on, and different datasets cannot be directly equated; the chance matching of low-complexity peptides is a residual risk throughout. Third, the Ribo-seq orthogonal axis: the ribosome footprints of a reverse overlapping ORF must be separated from those of the sense strand at the same locus, whereas the off-the-shelf human aggregate ribosome-profiling tracks [33,34] are strand-agnostic (a methodological ceiling), and precomputed ribosome evidence is intrinsically sparse (see below) — we therefore report it as an already-verified methodological boundary rather than a full-scale reprocessing. Fourth, the conservation threshold: the paper adopts the extremely stringent threshold of "conserved across

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

all vertebrates over 430 million years", which compresses the candidate pool; relaxing the threshold changes the number of detectable loci — both a limitation and a direct lead into the future directions below (reported as a result in §3.6). Fifth, the reproducible-locus evidence is not yet spectrum-validated: the 10 peptides confident in ≥ 3 datasets (§3.5) rest on the same automated database-search FDR as the rest of the catalogue; we have neither manually inspected their individual spectra nor computed a dedicated peptide-level (rather than run-level) FDR for this subset, and that spectrum-level confirmation is a necessary next step before any of these leads is treated as more than a calibrated candidate. Sixth, the transcription-layer assay is gene-body-level and bulk: the strand-specific RNA-seq quantification (§3.7) averages over the whole gene body and over bulk tissue, so it cannot resolve whether antisense reads initiate specifically over the shadow ORF (rather than elsewhere in the locus) nor capture cell-type-specific antisense transcription; it is a transcription-level probe only and is not used to make any translation claim. Selection pressure is deliberately left unquantified, and the reason is methodological rather than expedient. Standard dN/dS (Ka/Ks) is not applied to the dual-coding frames because it fails on overlapping genes in a specific, directional way. Its foundation is to sort each mutation independently as synonymous or non-synonymous and to calibrate dN against a neutral synonymous rate dS; but every nucleotide of a shadow ORF is simultaneously read by the host protein on the antisense strand, so whether a substitution is synonymous in the shadow frame and whether it is synonymous in the host frame are two entangled constraints that cannot take their values freely at once. Treating one frame as independent lets the dual-coding constraint systematically depress the usable count of synonymous sites and distort dS, misreading background as selection [35,36]. The consequence is not random noise but a directional artefact: Sabath, Landan and Graur (2008), modelling the two frames jointly, showed that analysing

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

overlapping influenza gene pairs as independent frames conjures a positive-selection signal out of nothing, which vanishes once the joint model removes the dual-coding constraint, leaving only purifying selection [35]; Monit et al. (2015) call using dN/dS to infer positive selection on overlapping frames outright invalid [36]. A reported "low dN/dS" or "positively selected site" for a shadow ORF could not be distinguished from the host protein's strong purifying selection projected onto the antisense strand — the default expectation. Dedicated methods do exist (the estimator of Wei & Zhang 2014, and OLGenie of Nelson, Ardern & Wei 2019), which model the other frame's coding constraint into the background rate [37,38]; their very existence corroborates that standard dN/dS cannot be applied directly. But they require sufficient within-species polymorphism or dense between-species substitutions to estimate the dual-coding rates separately, and the loci here — only 6 species, a span so deep that substitutions approach saturation, and each shadow ORF of limited length — do not supply enough information for reliable per-site estimation. This is consistent with the pre-specified threshold throughout: a sequence staying open (measurable) and a sequence being under functional selection (inferable only with the correct method) are two different things, and until the data can correctly remove the dual-coding constraint we will not substitute an easily misread statistic for evidence that does not yet exist.

Future directions

First, targeted validation of the handful of reproducible candidates. Perform targeted PRM (parallel reaction monitoring) on the 10 peptides in §3.5 reproducible in ≥ 3 datasets, comparing synthetic-peptide spectra against endogenous spectra one by one, with EGLN1 and BTBD3 first — the most direct step toward pushing a handful of loci from "credible" to "certain", and the route to the spectrum-level confirmation named in the limitations above.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

Second, an external, bioinformatics-only positive control. To establish the sensitivity of the search to genuinely translated non-canonical peptides without new wet-lab experiments, re-run the identical PepQuery2 workflow on a set of peptides that carry independent orthogonal evidence (e.g. high-confidence ncORF/microprotein peptides from the TransCODE consensus or OpenProt with Ribo-seq and immunopeptidomics support) and report their recovery rate against the same six datasets. Unlike the NFI internal check (§3.2), such peptides derive from independent evidence chains and would constitute a true sensitivity benchmark.

Third, strand-specific ribosome evidence. When custom Ribo-seq data from matched tissues with strand-specific library construction can be obtained, test whether the reverse frames exhibit triplet periodicity, circumventing the strand-agnostic ceiling of the off-the-shelf aggregate tracks.

Fourth, formally test selection pressure with a dedicated overlapping-gene model. Because standard dN/dS cannot be applied to dual-coding frames (above), test purifying/positive selection with a joint/dedicated model (OLGenie [38] or the joint estimation of Sabath et al. [35]) on data able to estimate the dual-coding rates separately — for example the shallower-divergence sequences obtained by widening the range to amniotes/mammals, or dedicated sampling of close relatives / population polymorphism for key loci (EGLN1, NFIB) — which is the correct way to fill in the selection dimension currently left blank rather than reporting a dN/dS number whose premises are violated.

Fifth, re-mine at a relaxed conservation range (amniote/mammal). As reported in §3.6, the per-candidate detection rate does not rise with conservation depth (Spearman $\rho = -0.030$, $P = 0.0016$) while the candidate pool expands as the depth requirement is relaxed (215 → 312 genes, +45%, a lower bound), so the two point the same way and the absolute yield of positive loci would very likely increase on relaxing to the amniote/mammal range. The concrete next stage is

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

to repeat the enumeration and the mass-spectrometry test on this larger, shallower-divergence pool — a data-backed candidate direction rather than an established conclusion.

Conclusions

Deep conservation keeps reverse-strand shadow ORFs open across vertebrates, but it does not amplify their detection by mass spectrometry at the class level: conserved candidates (T) are detected no more often than low-conservation reverse controls (C), and cross-dataset recurrence fails to separate the two, while a shuffled decoy recurs in four of six datasets. The chief reusable output is therefore methodological — a calibrated, three-arm false-positive floor for proteogenomic searches of non-canonical open reading frames — rather than a catalogue of confirmed shadow-ORF proteins. The ten most recurrent peptides (led by EGLN1 and BTBD3) are best treated as the highest-priority targets for spectrum-level validation.

Data Availability

All data supporting the conclusions of this study are provided as supplementary material or derive from public databases, as detailed below. The 12 processed data files listed in the table below have been deposited in a public repository under DOI [10.5281/zenodo.21871870](https://doi.org/10.5281/zenodo.21871870).

Data generated in this study:

File	Content	Size	Corresponding text
Table_S1_background_frame_summary.csv	Median open-excess and significance for each frame under the uniform / usage nulls	6 rows	Table S1, Figure 2; §2.2, 3.1
Table_S2_functional_candidates_ranked.csv	Dual-signal (open-excess percentile × conservation	318 rows	Table S2; §2.3

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

	percentile) ranking of functional candidates		
Table_S3_openprot_crosscheck.csv	Fraction of reverse candidates with mass-spectrometry / ribosome (TE) evidence in OpenProt for 28 hotspot genes, and the corresponding antisense transcripts	28 rows	Table S3; §2.3
Table_S4_candidates_tiered_ranked.csv	Six-frame enumeration, dual-null open_excess, amino-acid conservation z, composite score and tier labels for all candidates	1,908 rows (318 genes × 6 frames)	Table S4; §2.2–2.3, 3.1
Table_S5_panel_vs_background_positioning.csv	Positioning of candidates relative to the background distribution (incl. NFIB rev_f1 at the 99th percentile)	18 rows	Table S5; §3.2
Table_S6_ms_dataset_qc.csv	Per-dataset mass-spectrometry search statistics and quality control (peptides searched, confident hits and rate with Wilson 95% CI, median PSM count, median peptide length)	6 rows	Table S6; §3.3
Table_S7_shadow_ms_enrichment_summary.csv	Hit counts, hit rates, odds ratios, Fisher P and BH-FDR for the T/C/S sets across datasets and pooled strata	32 rows	Table S7, Table 1, Table 2, Figure 4; §3.3
Table_S8_threshold_sensitivity_or.csv	Threshold-sensitivity odds ratios (T-versus-control odds ratios across criterion-threshold sweeps)	38 rows	Table S8; §3.3
Table_S9_reproducibility_by_set.csv	Cross-dataset reproducibility of T/C/S peptides (number confident in ≥ 1, ≥ 2, ≥ 3, ≥ 4 and ≥ 5 independent datasets)	15 rows	Table S9; §3.5
Table_S10_shadow_antisense_rnaseq.csv	Strand-specific antisense-RNA quantification over the 27 T-set host genes in four total-RNA-seq datasets (antisense/sense	27 rows	Table S10; §3.7

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

	RPKM and antisense fraction per tissue)		
shadow_ms_pepquery_results.csv	Full peptide-level mass-spectrometry search results (per-peptide × per-dataset decision fields)	15,901 rows	Dataset S1; §3.3, 3.5
Dataset_S2_{gm12878,lung,liver,colon}_all318_antisense.csv	Per-tissue sense/antisense quantification for all 318 conserved genes (4 files)	318 rows each	Dataset S2; §3.7

Public data reused in this study. The mass-spectrometry evidence search was run against 6 public proteome datasets from PepQueryDB: Deep-29, GTEx (healthy-tissue subset), Colon, LUAD, HCC and CCLE (searched with PepQuery2 [10]). Orthologous alignment and the reference proteome are based on the GENCODE human reference annotation v50 [9]; the tissue-expression-breadth confound test uses the Human Protein Atlas [15]; the antisense-transcript / alternative-protein cross-check uses OpenProt [7]; the ribosome-profiling feasibility check refers to the GWIPS-viz human aggregate tracks [33,34]; the transcription-layer assay uses four ENCODE strand-specific total-RNA-seq datasets (GM12878 ENCSR000AEC, colon ENCSR403SZN, lung ENCSR425RGZ, liver ENCSR504QMK), aligned with STAR [13]. All of the above databases are publicly accessible, with access routes given in the corresponding references.

Code availability. The analysis scripts used for six-frame enumeration, dual-null openness quantification, construction and length-matching of the T/C/S sets, and the enrichment statistics on the PepQuery search results are provided as supplementary material with the paper; the third-party software used is MAFFT v7 [8] and PepQuery2 [10]. The mass-spectrometry datasets re-analysed here are publicly available under the accessions listed in Methods 2.4 (Deep-29 PXD010154, GTEx PXD016999, Colon PDC000109, LUAD PDC000219, HCC PDC000198, CCLE MSV000085836). The RNA-seq datasets are available from the ENCODE portal under ENCSR000AEC, ENCSR403SZN, ENCSR425RGZ and ENCSR504QMK.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

Supplementary data

Supplementary data. Supplementary data (Tables S1–S10, Datasets S1–S2 and Figures S1–S5) associated with this article can be found in the online version.

CRediT authorship contribution statement

Yabin Guo: Conceptualization; Writing – original draft. Yongxian Chen, Qian Zhao, and Yin Zhang: Methodology; Investigation; Formal analysis. All authors discussed the results, reviewed the manuscript, and approved the final version.

Funding

This work was supported by Guangdong Basic and Applied Basic Research Foundation (2026A1515012932 to YG) and the Science and Technology Planning Project of Guangdong Province (2023B1212060013 and 2020B1212030004).

Declaration of competing interest

The authors declare no competing financial interest.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Biomni and Kimi for data analysis, figure generation, and language editing. The authors reviewed and verified all analyses and results, take full responsibility for the accuracy and integrity of the work, and are accountable for its content. No AI tool is listed as an author.

References

(1) Mudge, J. M.; Ruiz-Orera, J.; Prensner, J. R.; et al. Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **2022**, *40* (7), 994–1003. DOI: 10.1038/s41587-

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

022-01369-0

(2) Deutsch, E. W.; Kok, L. W.; Mudge, J. M.; et al. Expanding the human proteome with microproteins and peptideins. *Nature* **2026**, *654*, 813–825. DOI: 10.1038/s41586-026-10459-x

(3) Zehentner, B.; Ardern, Z.; Kreitmeier, M.; Scherer, S.; Neuhaus, K. A novel pH-regulated, unusual 603 bp overlapping protein coding gene pop is encoded antisense to ompA in Escherichia coli O157:H7 (EHEC). *Front. Microbiol.* **2020**, *11*, 377. DOI: 10.3389/fmicb.2020.00377

(4) Iyengar, B. R.; Grandchamp, A.; Bornberg-Bauer, E. How antisense transcripts can evolve to encode novel proteins. *Nat. Commun.* **2024**, *15* (1), 6187. DOI: 10.1038/s41467-024-50550-3

(5) Michel, A. M.; Choudhury, K. R.; Firth, A. E.; Ingolia, N. T.; Atkins, J. F.; Baranov, P. V. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* **2012**, *22* (11), 2219–2229. DOI: 10.1101/gr.133249.111

(6) Duncan, C. D. S.; Mata, J. The translational landscape of fission yeast meiosis and sporulation. *Nat. Struct. Mol. Biol.* **2014**, *21* (7), 641–647. DOI: 10.1038/nsmb.2843

(7) Brunet, M. A.; Brunelle, M.; Lucier, J. F.; et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* **2019**, *47* (D1), D403–D410. DOI: 10.1093/nar/gky936

(8) Katoh, K.; Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30* (4), 772–780. DOI: 10.1093/molbev/mst010

(9) Frankish, A.; Carbonell-Sala, S.; Diekhans, M.; et al. GENCODE: reference annotation for

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

the human and mouse genomes in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D942–D949. DOI: 10.1093/nar/gkac1071

(10) Wen, B.; Zhang, B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. *Nat. Commun.* **2023**, *14* (1), 2213. DOI: 10.1038/s41467-023-37747-8

(11) Prensner, J. R.; Abelin, J. G.; Kok, L. W.; et al. What can Ribo-seq, immunopeptidomics, and proteomics tell us about the noncanonical proteome? *Mol. Cell. Proteomics* **2023**, *22* (9), 100631. DOI: 10.1016/j.mcpro.2023.100631

(12) ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489* (7414), 57–74. DOI: 10.1038/nature11247

(13) Dobin, A.; Davis, C. A.; Schlesinger, F.; et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29* (1), 15–21. DOI: 10.1093/bioinformatics/bts635

(14) Gronostajski, R. M. Roles of the NFI/CTF gene family in transcription and development. *Gene* **2000**, *249* (1–2), 31–45. DOI: 10.1016/S0378-1119(00)00140-2

(15) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; et al. Tissue-based map of the human proteome. *Science* **2015**, *347* (6220), 1260419. DOI: 10.1126/science.1260419

(16) Aggarwal, S.; Yadav, A. K. The Achilles' heel of proteogenomics: false discovery rate. *Brief. Bioinform.* **2022**, *23* (4), bbac163. DOI: 10.1093/bib/bbac163

(17) Li, Q.; Shortreed, M. R.; Wenger, C. D.; et al. A global approach for proteogenomic discovery of noncanonical reading frames and frame-shifted peptides. *BMC Genomics* **2016**, *17*, 1032. DOI: 10.1186/s12864-016-3327-5

(18) Blakeley, P.; Overton, I. M.; Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* **2012**, *11* (11), 5221–5234. DOI: 10.1021/pr300411q

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

(19) Wacholder, A.; Carvunis, A. R. Biological factors and statistical limitations prevent detection of most noncanonical proteins by mass spectrometry. *PLoS Biol.* **2023**, *21* (8), e3002409. DOI: 10.1371/journal.pbio.3002409

(20) Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* **2016**, *17* (10), 601–614. DOI: 10.1038/nrg.2016.85

(21) Zhou, Q.; Jiang, Y.; Cai, C.; et al. Multidimensional conservation analysis decodes the expression of conserved long noncoding RNAs. *Life Sci. Alliance* **2023**, *6* (6), e202302002. DOI: 10.26508/lsa.202302002

(22) Ruiz-Orera, J.; Albà, M. M. Conserved regions in long non-coding RNAs contain abundant translation and protein–RNA interaction signatures. *NAR Genomics Bioinform.* **2019**, *1* (2), lqz002. DOI: 10.1093/nargab/lqz002

(23) Willensdorfer, M.; Bürger, R.; Nowak, M. A. Phenotypic mutation rates and the abundance of abnormal proteins in yeast. *PLoS Comput. Biol.* **2007**, *3* (11), e203. DOI: 10.1371/journal.pcbi.0030203

(24) Sydow, J. F.; Cramer, P. RNA polymerase fidelity and transcriptional proofreading. *Curr. Opin. Struct. Biol.* **2009**, *19* (6), 732–739. DOI: 10.1016/j.sbi.2009.10.009

(25) Thomas, M. J.; Platas, A. A.; Hawley, D. K. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell* **1998**, *93* (4), 627–637. DOI: 10.1016/S0092-8674(00)81191-5

(26) Jeon, C.; Agarwal, K. Fidelity of RNA polymerase II transcription controlled by elongation factor TFIIS. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93* (24), 13677–13682. DOI: 10.1073/pnas.93.24.13677

(27) Traverse, C. C.; Ochman, H. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (12), 3311–

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

3316. DOI: 10.1073/pnas.1525329113

(28) Merino, E.; Balbas, P.; Puente, J. L.; Bolivar, F. Antisense overlapping open reading frames in genes from bacteria to human. *Trends Genet.* **1994**, *10* (5), 160–164. DOI: 10.1016/0168-9525(94)90125-2

(29) Silke, J. The majority of long non-stop reading frames on the antisense strand can be explained by biased codon usage. *Gene* **1997**, *194* (1), 143–155. DOI: 10.1016/S0378-1119(97)00199-6

(30) Rother, K. I.; Silke, J.; Georgiev, O.; Schaffner, W.; Matsuo, K. Influence of DNA sequence and methylation status on the open reading frames in the antisense strand of the Hsp70 and PrP genes. *Biol. Chem.* **1997**, *378* (12), 1521–1530. DOI: 10.1515/bchm.1997.378.12.1521

(31) Forsdyke, D. R. Sense in antisense? *J. Mol. Evol.* **1995**, *41* (5), 582–586. DOI: 10.1007/bf00175816

(32) Merino, E.; Balbás, P.; Puente, J. L.; Bolívar, F. Antisense overlapping open reading frames in genes from bacteria to humans. *Nucleic Acids Res.* **1994**, *22* (10), 1903–1908. DOI: 10.1093/nar/22.10.1903

(33) Michel, A. M.; Fox, G.; Kiran, A. M.; et al. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res.* **2014**, *42* (D1), D859–D864. DOI: 10.1093/nar/gkt1035

(34) Michel, A. M.; Kiniry, S. J.; O'Connor, P. B. F.; Mullan, J. P. A.; Baranov, P. V. GWIPS-viz: 2018 update. *Nucleic Acids Res.* **2018**, *46* (D1), D823–D830. DOI: 10.1093/nar/gkx790

(35) Sabath, N.; Landan, G.; Graur, D. A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS One* **2008**, *3* (12), e3996. DOI: 10.1371/journal.pone.0003996

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.08.13.000310. This version posted November 10, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution 4.0 https://creativecommons.org/licenses/by/4.0

(36) Monit, C.; Goldstein, R. A.; Towers, G. J.; et al. Positive selection analysis of overlapping reading frames is invalid. *AIDS Res. Hum. Retroviruses* **2015**, *31* (11), 1147–1148. DOI: 10.1089/aid.2015.0150

(37) Wei, X.; Zhang, J. A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biol. Evol.* **2014**, *6* (9), 2453–2461. DOI: 10.1093/gbe/evu294

(38) Nelson, C. W.; Ardern, Z.; Wei, X. OLGenie: estimating natural selection to predict functional overlapping genes. *Mol. Biol. Evol.* **2020**, *37* (1), 265–279. DOI: 10.1093/molbev/msaa087