Supporting Information

for

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

Yongxian Chen, Qian Zhao, Yin Zhang, Yabin Guo*

*Correspondence: guoyb9@sysu.edu.cn

Contents

Tables S1–S10 — legends and contents (pages S-2–S-3; CSV files accompany this PDF)

Datasets S1–S2 — legends and contents (page S-3; CSV files accompany this PDF)

Figures S1–S5 — full figures with captions (pages S-4–S-8)

Supplementary Tables

Table S1. Dual-null summary of reverse-frame open excess. The open_excess median and significance for each frame under the uniform and usage nulls (e.g. rev_f0 uniform median excess +18.4 percentage points, $P = 4.3 \times 10^{-69}$). (Source file: Table_S1_background_frame_summary.csv; corresponding text 2.2, 3.1.)

Table S2. Dual-signal ranking table of functional candidates. Open-excess percentile × conservation percentile; top-ranking genes include HIVEP3, BTBD3, LSAMP and KALRN. (Source file: Table_S2_functional_candidates_ranked.csv; corresponding text 2.3.)

Table S3. OpenProt cross-check. The fraction of reverse candidates with mass-spectrometry evidence and with ribosome (TE) evidence in OpenProt for 28 hotspot genes, and the corresponding real antisense transcripts (e.g. NFIB-AS1, BTBD3-AS1). Only 7.2% of reverse candidates carry ribosome evidence, whereas 56.3% carry mass-spectrometry evidence (about 8:1). (Source file: Table_S3_openprot_crosscheck.csv; corresponding text 2.3.)

Table S4. Candidate tiering and composite-score ranking table. The open_excess percentile, conservation percentile, composite score (7.57–94.88) and tier label of all 1,908 candidates (318 genes × 6 frames). (Source file: Table_S4_candidates_tiered_ranked.csv, 1,908 rows; corresponding text 2.3, 3.1.)

Table S5. Positioning table of candidates relative to the background distribution, incl. NFIB rev_f1 open-excess at the 99th percentile. (Source file: Table_S5_panel_vs_background_positioning.csv; corresponding text 3.2.)

Table S6. Per-dataset mass-spectrometry search statistics and quality control: peptides searched, confident hits and rate with Wilson 95% CI, median PSM count and median peptide length per dataset. (Source file: Table_S6_ms_dataset_qc.csv, 6 rows; corresponding text 3.3.)

Table S7. Full mass-spectrometry enrichment statistics: hit counts, hit rates, odds ratios, Fisher P and BH-FDR for the T/C/S sets across datasets and pooled strata (all_datasets_union, healthy_union). Source data for main-text Tables 1 and 2. (Source file: Table_S7_shadow_ms_enrichment_summary.csv, 32 rows; corresponding text 3.3.)

Table S8. Threshold-sensitivity odds ratios: T-versus-control odds ratios across criterion-threshold sweeps. (Source file: Table_S8_threshold_sensitivity_or.csv, 38 rows; corresponding text 3.3.)

Table S9. Cross-dataset reproducibility of T/C/S peptides: number of peptides confident in ≥ 1, ≥ 2, ≥ 3, ≥ 4 and ≥ 5 independent datasets (T = 175 / 32 / 10 / 4 / 1; C = 207 / 50 / 12 / 0 / 0; S = 75 / 15 / 1 / 1 / 0). (Source file: Table_S9_reproducibility_by_set.csv, 15 rows; corresponding text 3.5.)

Table S10. Strand-specific antisense-RNA quantification over the 27 T-set host genes in four total-RNA-seq datasets (GM12878, lung, liver, colon): antisense RPKM, antisense fraction and sense RPKM per tissue. Antisense RNA is detectable at most loci but sparse (per-tissue median 0.0009–0.010 RPKM), and does not track the reproducible MS loci. (Source file: Table_S10_shadow_antisense_rnaseq.csv, 27 rows; corresponding text 3.7.)

Supplementary Datasets

Dataset S1. Full peptide-level mass-spectrometry search results (shadow_ms_pepquery_results.csv, 15,901 rows; corresponding text 3.3, 3.5.)

Dataset S2. Per-tissue sense/antisense quantification for all 318 conserved genes (Dataset_S2_{gm12878,lung,liver,colon}_all318_antisense.csv, 318 rows each; corresponding text 3.7.)

Figure S1

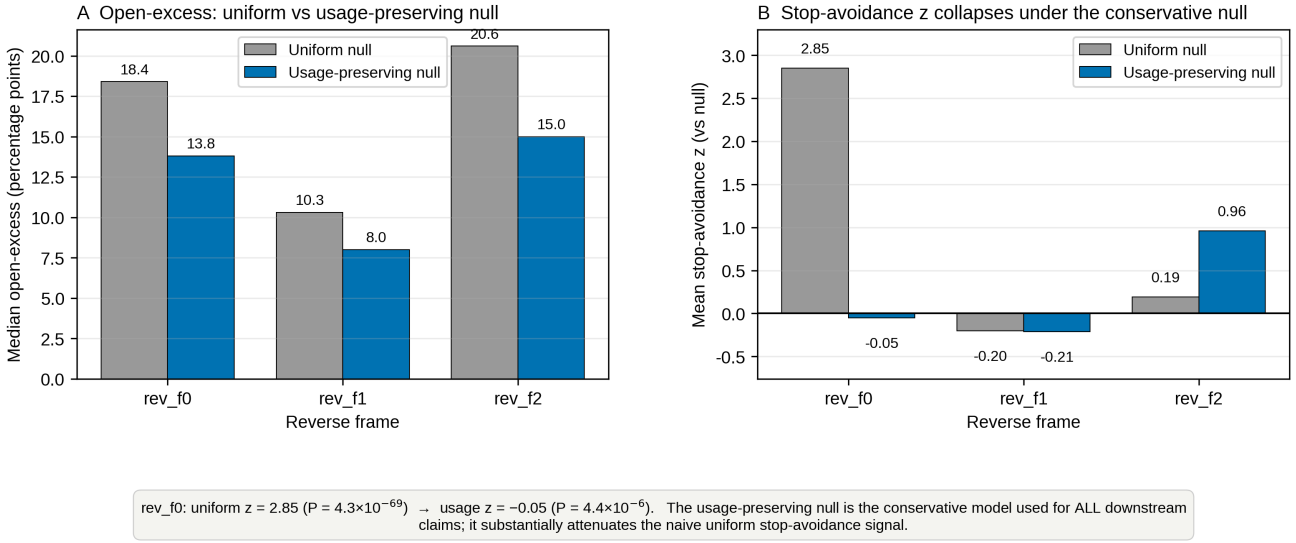


Figure S1. Robustness of the reverse-frame openness signal to null-model specification. (A) Median open-excess (percentage points) for the three reverse frames under the naive uniform-composition null (grey) and the conservative usage-preserving null (blue). (B) Mean stop-avoidance z-score for the same frames and nulls, relative to zero. For the interlocking rev_f0 frame the signal attenuates from uniform z = 2.85 (P = 4.3 × 10^{-69}) to usage z = −0.05 (P = 4.4 × 10^{-6}) once codon usage is preserved. The usage-preserving null is the conservative model used for all downstream claims.

Figure S2

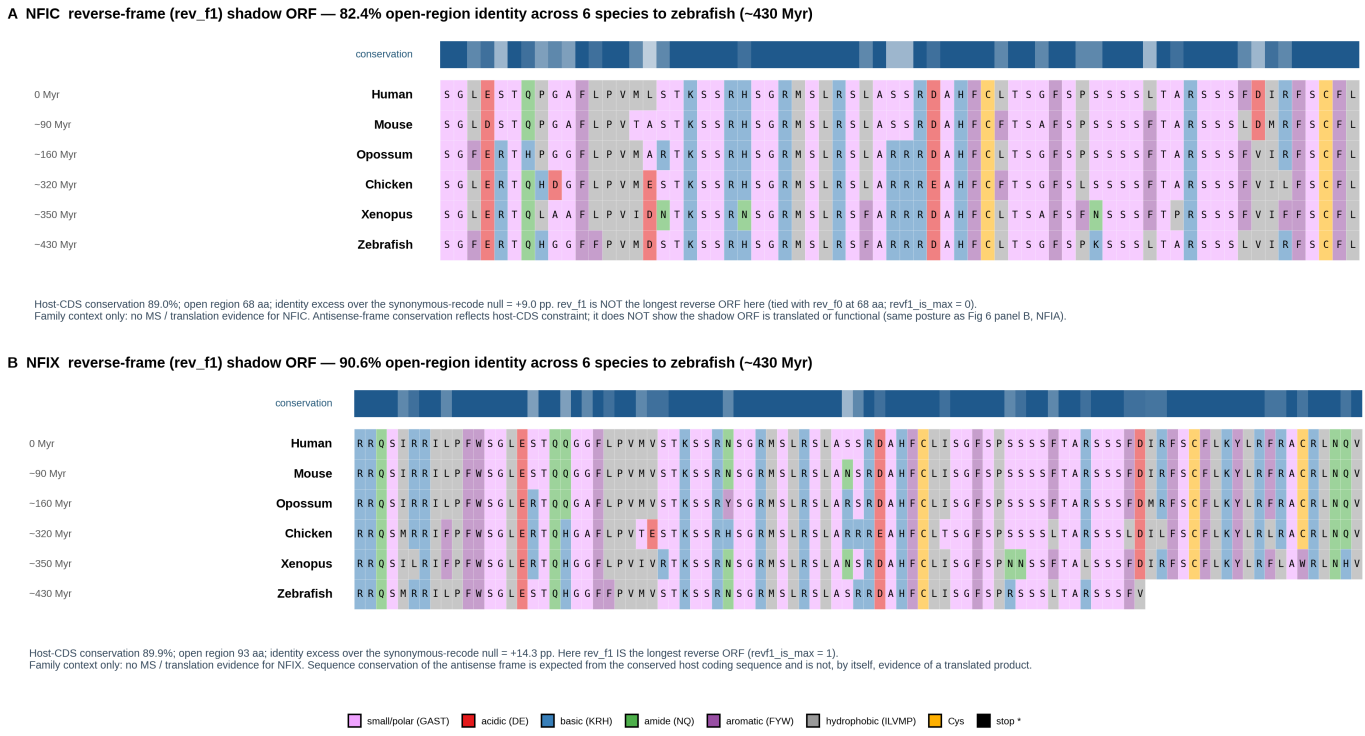


Figure S2. Antisense-frame conservation extends to the other zebrafish-reaching NFI members, NFIC and NFIX. Reverse-frame (rev_f1) shadow-ORF alignment across six vertebrate species (human to zebrafish, ~430 Myr). (A) NFIC: 82.4% open-region identity. (B) NFIX: 90.6% open-region identity. These panels are family context only: NFIC and NFIX have no mass-spectrometry or translation evidence, and antisense-frame sequence conservation does not by itself show that the shadow ORF is translated or functional.

Figure S3

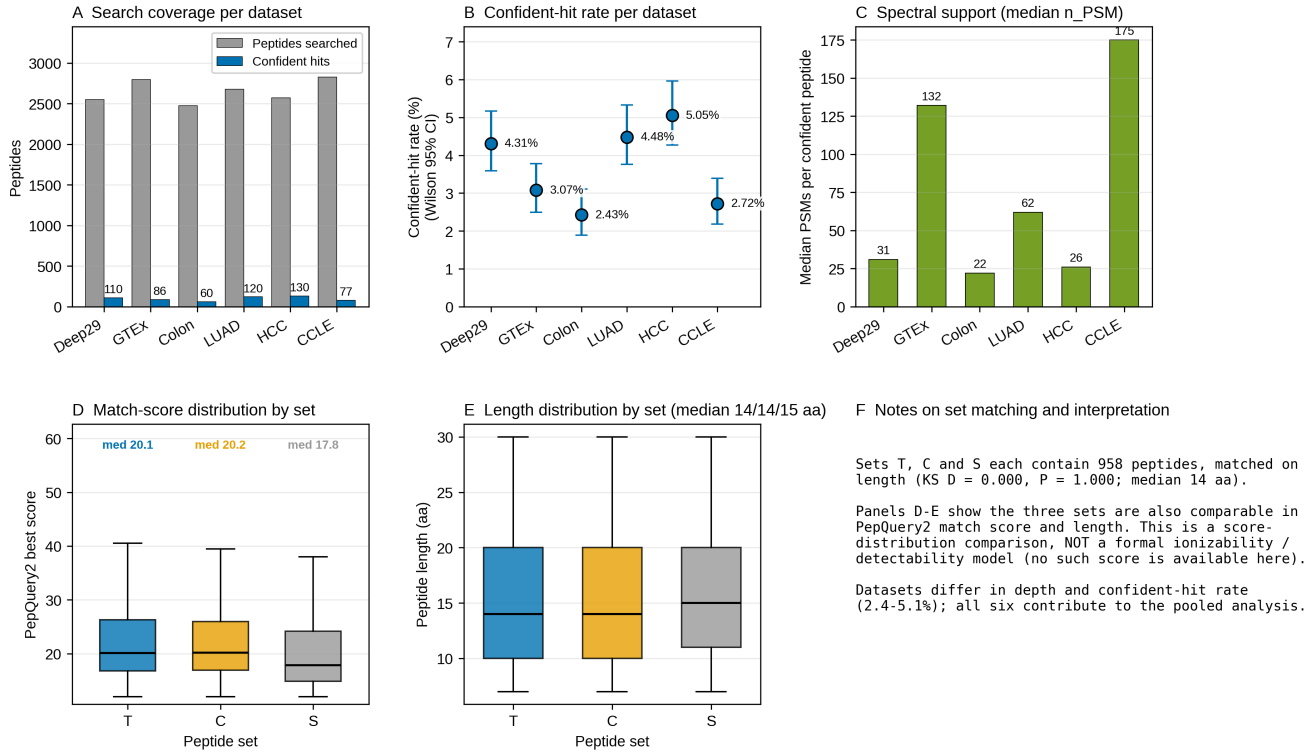


Figure S3. Quality control of the six mass-spectrometry datasets and comparability of the T, C and S peptide sets. (A) Peptides searched and confidently detected per dataset. (B) Confident-hit rate per dataset with Wilson 95% CIs (2.43%–5.05%). (C) Median PSM count per confident peptide. (D) PepQuery2 best-match score distributions (medians T 20.1, C 20.2, S 17.8). (E) Peptide-length distributions (median 14/14/15 aa). (F) Each set contains 958 length-matched peptides (KS D = 0.000, P = 1.000).

Figure S4

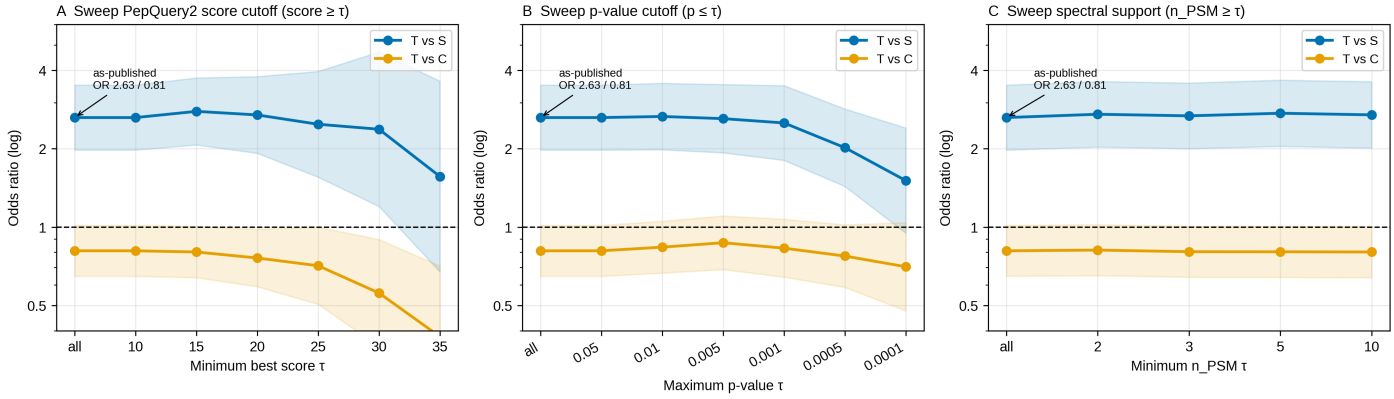
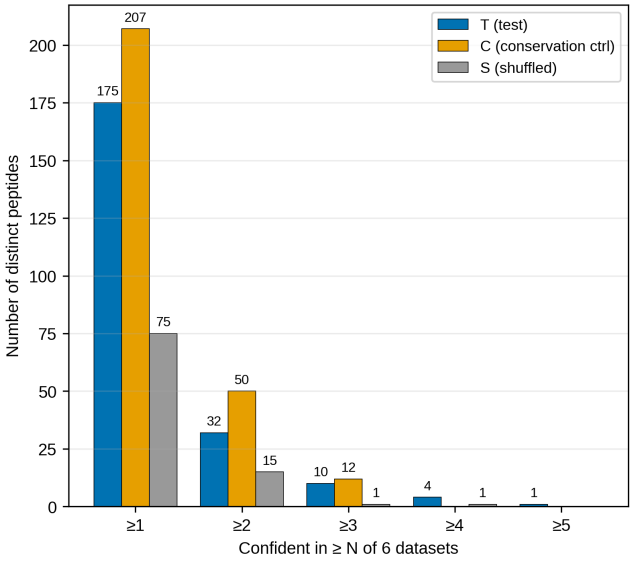


Figure S4. Enrichment odds ratios are robust to the peptide-filtering thresholds. Pooled (all-datasets-union) odds ratios for T vs S and T vs C recomputed across a sweep of secondary filtering thresholds.

Figure S5

A Reproducibility funnel
The ≥3-dataset tier is rare in every set; T (10) does not exceed C (12)



B The 10 T peptides confident in ≥3 datasets

Peptide	Host gene	Frame	#datasets
ALAAGGR *	EGLN1	rev_f0	5
AAGSTSGIR *	BTBD3	rev_f1	4
DGAEVGEGAPAGELAAR	LHFPL4	rev_f0	4
EAASGGR *	CTCF	rev_f2	4
KMSLVSLR	BTBD3	rev_f1	3
LIVVIPK	HNRNPA1	rev_f0	3
RLLLAGLGFGSHR	EGLN1	rev_f0	3
SVGPGDMK	BMAL1	rev_f1	3
VGSPPSR	BMAL1	rev_f1	3
VLASTAIGAR	HNRNPA1	rev_f0	3

* low-complexity (Gly/Ala/Ser ≥ 60%): ALAAGGR, AAGSTSGIR, EAASGGR. These dominate the most-reproducible T peptides and are flagged as a CAVEAT (compositional bias / possible mismapping), not as strong evidence of coding. Most reproducible: ALAAGGR (EGLN1 rev_f0), 5/6.

Figure S5. Cross-dataset reproducibility of confident peptide detections. (A) Reproducibility funnel: distinct peptides confidently detected in ≥ N of the six datasets for T, the low-conservation control C and the shuffled control S. T = 175 / 32 / 10 at ≥ 1 / ≥ 2 / ≥ 3 datasets; C = 207 / 50 / 12; S = 75 / 15 / 1. The ≥ 3-dataset tier is rare in every set, and T (10) does not exceed C (12) — consistent with the main-text finding that T is not greater than C, while both exceed S. (B) The 10 T peptides confident in ≥ 3 datasets, with host gene and reverse frame; the most reproducible is ALAAGGR (EGLN1 rev_f0, 5 of 6 datasets). Low-complexity peptides (Gly/Ala/Ser ≥ 60%: ALAAGGR, AAGSTSGIR, EAASGGR) are flagged as a caveat, not as strong evidence of coding.