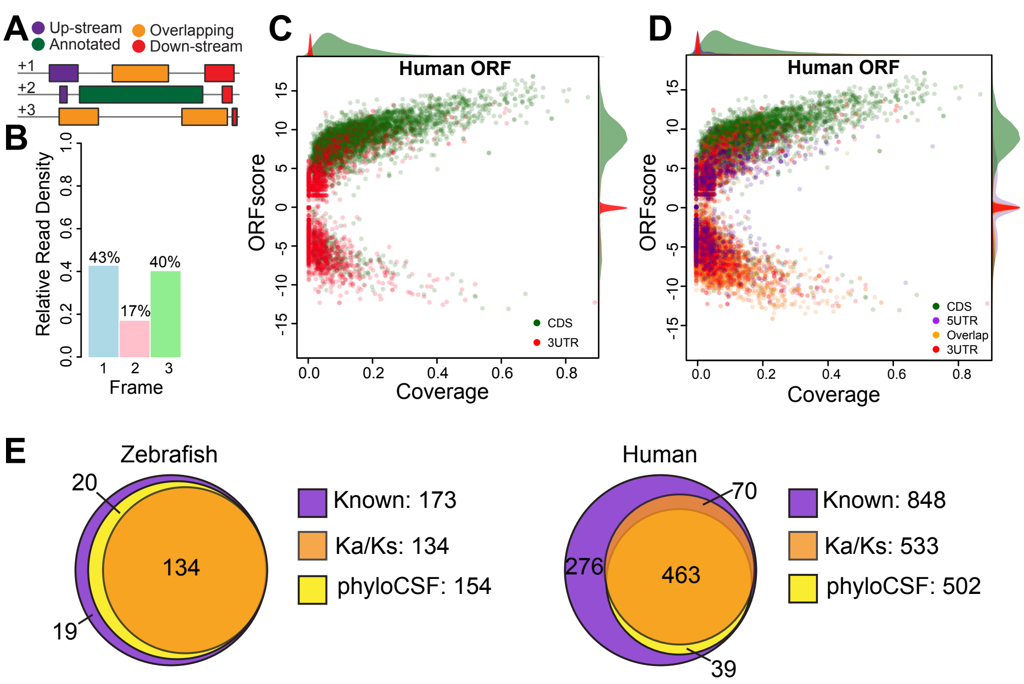

Figure S6. Defining translated ORFs in Human.

Using the Ka/Ks approach, we detected 134 of 173 annotated smORFs in zebrafish and 533 of 848 in human. Additionally, we recover 43 of 63 predicted smORFs found by phyloCSF in zebrafish and 88 of 173 predicted smORFs in human, indicating that their coding potential can be robustly identified.

A smORF among the top hits is a putative short ribosomal protein thus far not annotated in the zebrafish genome (Fig. 5B). Notably, we have observed that a region on chr9 rich in segmental duplications gives rise to several candidates, predicted computationally and/or detected via ORFscore, with nearly identical peptide sequences and homology to known proteins.

(A) Schematic of RefSeq ORFs: annotated CDS (green), 5‘UTR ORFs (purple), 3’UTR ORFs (red), and CDS-overlapping ORFs (orange).

(B) RPF reads (Guo et al., 2010) mapped to a composite human RefSeq transcript. RPFs map to the CDS showing a 3-nucleotide periodicity.

(C) Scatterplot of ORFscore versus coverage (position 1) for 25,000 randomly selected Ensembl annotated coding ORFs (green) and all ORFs in the 3’UTR for each time point.

(D) Scatterplot of ORFscore versus coverage (position 1) for all four types of ORF in 25,000 randomly selected protein-coding transcripts from Ensembl.

(E) Venn diagrams showing the number of known smORFs defined by Ka/Ks and phyloCSF methods.