Supplementary Figure 1: The percentage of sORFs with sequence similarity, divided by the amount of distinct species is represented.

Supplementary Figure 2: Overview of sORFs with additional PRIDE-ReSpin evidence, grouped by the different annotations used within the sORFs.org repository.
Supplementary Figure 3: The percentage of sORFs over the number of datasets wherein the sORF was identified.

Supplementary Figure 4: The noise filtering algorithm was validated on the HCT116 datasets (Crappé et al., 2014). The canonical protein coding transcript from Ensembl v86 were used as a positive control. The x-axis represents in-frame coverage percentage segments while the y-axis represents the average p-value and the gene percentage of that in-frame coverage segment. The majority of protein coding transcript do not show evidence of translation as represented by the lower in-frame coverage segments, the average p-value coherently suggests that these transcripts do not represent true translation events.
Supplementary Figure 5: The noise filtering algorithm was validated on the HCT116 datasets (Crappé et al., 2014). 3′-UTR regions of canonical protein coding transcript from Ensembl v86 were used as negative control. For the 3′-UTR regions in-frame is considered as a continuation of the CDS frame. The x-axis represents in-frame coverage percentage segments while the y-axis represents the average p-value and the gene percentage of that in-frame coverage segment. 98% of 3′-UTR regions have an in-frame coverage between 0-5%, indicating that no translational events happen as to be expected. Generally, the average p-value associated with these regions is high, except for several drops at higher in-frame coverage regions attributed to data scarcity, indicating that the noise filtering algorithm can correctly distinguish between true translational events and noise.
PRIDE ReSpin Methodology

PRIDE-ReSpin downloads publicly available datasets from the PRIDE database(1) and subsequently infers the initial experiment parameters by reverse engineering the identification and spectra files from PRIDE, using PRIDE-ASAP(2) and a module from Pladipus(3). This implies that only entries which contain peptide identification files can be processed by PRIDE, since the supplied experimental parameters are often absent or incomplete. Next, SearchGUI(4) is used as search engine management tool, launching the X!Tandem(5) and MS-GF+(6) search engines to match spectra against the UniProt Reference proteome(7) concatenated with the cRAP database(8) and our sORFs.org database. Output from these search engines is processed using PeptideShaker(9) to filter and validate spectra. To reduce false positive identifications, only spectra without higher confident PSM matches to non-sORF peptides are withheld at a 1% FDR and a spectrum coverage of at least 30% is required.


