Supplementary methods

Datasets

The reads of Ribo-seq and RNA-seq used in this study were collected from previously published literatures. These datasets are publicly available at NCBI Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra) and the accession entries and sources are listed in Table S1.

Identification of SL-containing ribosome-protected footprints (RPFs) and footprints for translation initiation

RPFs were obtained by trimming off the adaptors on Ribo-seq reads. The reads without adaptor were discarded. SL-containing RPFs were identified by searching for RPFs start with full or partial SL sequences longer than 8 nt. RPFs were mapped to reference genome of nematodes using STAR with default parameters, and RPFs overlapped with the sites of annotated start codons were selected using bedtools.

Calculation of P-sites offsets

The offsets of P-sites were calculated in both genome-dependent and independent manners. We first counted the distance from the 5' and 3' end of RPFs to the most upstream AUG triplet in SL-RPFs as the offsets of P-sites. Genome-dependent approach had been applied to nematodes but not to dinoflagellates, for which the reference genome is not available yet. In this manner, we first mapped the RPFs to the genome and select the ones overlapped with the start codon of genes, followed by the counting of the distances from the 5' end of RPFs to the annotated start codons, the offsets of P-sites. Both analyses had resulted in the same offsets suggesting the reliability of the genome-independent strategy. Furthermore, the offsets are independent to the lengths of RPFs (Figure 2A).

Extraction of potential start codons and Kozak context

Because all the SL-RPFs with different sizes nearly have the same offsets from the codons at P-sites to the 5'end of the RPFs while the distances to the 3'end vary depending on the length of RPFs (Figure 2A), the offsets to the 5'end terminus (12 nt for *C. elegans*, *C. brennei* and *C. remanei*, and 12 and 13 nt for *L. polyedra*) were used to determine the codons being translated represented by the given RPFs. Given that SL-RPFs represent the first RPF of coding sequences being translated, the codons extracted from SL-PRFs can be the start codons. As incomplete or over digestion of RNA sometimes can result in the shift of offsets, we also extracted the triplet at one-nt down or upstream of the offsets and preferentially keep the triplets of NUG with AUG the highest priority. The potential start codons with usage higher than 2% were ploted on pie charts. The Kozak contexts were extracted separately for each potential start codon and the consensus sequences were plotted using WebLogo [1].

Reference:

1. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. Genome Res. 2004;**14:**1188-1190.