



Supporting Online Material for **The Landscape of *C. elegans* 3'UTRs**

Marco Mangone, Arun Prasad Manoharan, Danielle Thierry-Mieg, Jean Thierry-Mieg, Ting Han, Sebastian Mackowiak, Emily Mis, Charles Zegar, Michelle R. Gutwein, Vishal Khivansara, Oliver Attie, Kevin Chen, Kourosch Salehi-Ashtiani, Marc Vidal, Timothy T. Harkins, Pascal Bouffard, Yutaka Suzuki, Sumio Sugano, Yuji Kohara, Nikolaus Rajewsky, Fabio Piano,* Kristin C. Gunsalus,* John K. Kim*

*To whom correspondence should be addressed. E-mail: fp1@nyu.edu (F.P.); kcg1@nyu.edu (K.C.G.); jnkim@umich.edu (J.K.K.)

Published June 3 2010 on *Science Express*
DOI: 10.1126/science.1191244

This PDF file includes:

Materials and Methods
Figs. S1 to S14
Tables S1 to S10
References

Other Supporting Online Material for this manuscript includes the following: (available at www.sciencemag.org/cgi/content/full/science.1191244/DC1)

Datasets S1 to S7 as a zipped archive: [1191244_datasets_s1_to_s7.zip](#)

Table of Contents

Supplementary Materials and Methods	4
Figure S1. <i>Overview of the 3'UTRome.</i>	28
Figure S2. <i>Overview of 3'UTRome computational pipeline.</i>	30
Figure S3. <i>Workflow for polyA capture assay.</i>	31
Figure S4. <i>PolyA capture protocol.</i>	32
Figure S5. <i>Workflow for 3'RACE.</i>	33
Figure S6. <i>Distance between individual 3'ends and the representative polyA addition site for a cluster.</i>	34
Figure S7. <i>Distribution of the number of polyA sites per gene.</i>	35
Figure S8. <i>Introns in 3'UTR regions.</i>	36
Figure S9. <i>Distribution of the canonical AAUAAA and variant PAS elements relative to the cleavage and polyA addition site..</i>	37
Figure S10. <i>Distribution of variant PAS elements relative to the cleavage and polyA addition site.</i>	38
Figure S11. <i>Relationships between alternative polyA addition sites for the same transcript.</i>	39
Figure S12. <i>Polyadenylated 3'UTRs for histone genes.</i>	40
Figure S13. <i>PicTar miRNA target predictions and PAS conservation.</i>	41
Figure S14. <i>3'UTRs on opposite strands sometimes overlap.</i>	42
Table S1. <i>Sequence data in the 3'UTRome.</i>	43
Table S2. <i>Summary of the polyA capture 454 sequencing runs.</i>	43
Table S3. <i>Gene and 3'UTR isoform coverage for individual datasets and overlaps among datasets in the 3'UTRome using AceView gene models.</i>	44

Table S4. <i>Subset of 3'UTRome matching WS190 gene models.</i>	44
Table S5. <i>Identification of putative PAS elements.</i>	45
Table S6. <i>Cumulative list of polyadenylated 3'UTRs detected in histone genes.</i>	46
Table S7. <i>Summary statistics for PicTar miRNA target predictions and other conserved sequence blocks in genomic regions spanned by the 3'UTRome compendium.</i>	48
Table S8. <i>Number of genes present in multiple developmental stages but with stage-specific 3'UTR isoforms.</i>	49
Table S9. <i>Number of genes with two 3'UTR isoforms detected in the staged polyA capture dataset.</i>	49
Table S10. <i>3'UTR clones available in the 3'UTRome library.</i>	50
Dataset S1 Legend. <i>AceView gene models in the 3'UTRome compendium.</i>	51
Dataset S2 Legend. <i>The complete 3'UTRome dataset.</i>	51
Dataset S3 Legend. <i>3'UTR coordinates attached to AceView genes.</i>	51
Dataset S4 Legend. <i>Genes with overlapping 3'UTRs.</i>	51
Dataset S5 Legend. <i>Genes displaying changes in 3'UTR length between developmental stages.</i>	51
Dataset S6 Legend. <i>Genes displaying 3'UTR isoform switching during development.</i>	51
Dataset S7 Legend. <i>The 3'UTRome clone library.</i>	51
References	52

Supplementary Materials and Methods

PolyA capture

Strains: Worms were grown on NGM plates seeded with *E. coli* OP50 to adulthood. For collection of staged samples, the wild-type N2 strain was used. Embryos were isolated from gravid worms by standard alkaline/hypochloride treatment (1). A sample of embryos was frozen down in TriReagent (Ambion, Austin, TX), and the remainder hatched overnight in M9 buffer to yield synchronized L1 stage worms. Starved L1 larvae were plated and fed on NGM plates seeded with OP50 *E. coli* and raised at 20°C. Synchronized staged samples were collected at ~8 hr (L1), ~20 hr (L2), ~30 hr (L3), ~45 hr (L4), and ~70 hr (adult hermaphrodite). The developmental stage of each sample was verified by monitoring the seam cell lineage using Nomarski optics (Olympus, Center Valley, PA). For adult male isolation, the CB1489 *him-8(e1489)* strain was used, which increases the percentage of XO males to ~37% of the population versus ~0.2% males in the N2 wild-type strain (2). The *him-8(e1489)* embryos were synchronized by bleaching and incubated overnight at room temperature. Hatched L1s were aliquoted onto NGM plates seeded with *E. coli* OP50 and grown at 20°C for 4 days. Male adults were isolated by filtering through 35 μ m nylon mesh, resulting in >95% males in the final sample. For dauer larvae preparation, CB1370 *daf-2 (e1370)*, CB1372 *daf-7 (e1372)*, DR47 *daf-11 (m47)*, DR2281 *daf-9 (m540)* mutants from starved plates were collected, resuspended in M9 buffer (1) containing 1% SDS, and incubated for 20 min at room temperature. The suspension was then washed with M9 buffer and worms were placed on a fresh unseeded plate at 20°C for 12 h. Live worms that had

crawled away from the dead worms were collected as dauer larvae. Worms were washed off plates with M9, washed 5 times with M9 to remove residual bacteria, and frozen in TriReagent.

RNA preparation: Total RNA was extracted using TriReagent following the vendor's protocol with the following modification: three freeze-thaw cycles (freeze in liquid nitrogen / thaw at room temperature / vortex 1 min) were included to increase worm lysis efficiency; RNA was precipitated with isopropanol at -80°C for one hour. To subtract 72 most abundant ribosome subunit genes, 25µg total RNAs were mixed with antisense DNA oligos (IDT, Coralville, IA) targeting the last DpnII site of each of these genes and digested with RNaseH (Invitrogen, Carlsbad, CA), which only cleaves RNA in RNA:DNA duplex. After subtraction, PolyA⁺- selected mRNAs were isolated from total RNA using oligo(dT) magnetic beads (Invitrogen, Carlsbad, CA) using the manufacturer's protocol.

cDNA synthesis: First-strand synthesis was carried out using Superscript III reverse transcription kit (Invitrogen, Carlsbad, CA) with ~20 ng of PolyA⁺- selected mRNA and 10 pmol of biotinylated reverse primer at 50°C for 30 min followed by incubation at 42°C for 30 min. The following biotin-labeled primer was synthesized by Integrated DNA Technologies (Coralville, IA) and PAGE-purified: 5'Biotin-TAATAC-GGCGCGCCGCCTTGCCAGCCCGCTCAG-T₂₀-VN-3'. The poly(dT) and two nucleotide anchor (VN) target the proximal end of the mRNA polyA tail. The second strand was synthesized using DNA polymerase I in the presence of RNase H for 2.5 hr. The

double-stranded cDNA product was extracted twice with 200 μ L phenol/chloroform/isoamyl alcohol (25:24:1), ethanol precipitated, and dissolved in 20 μ L H₂O.

DpnII digestion: The resulting cDNA was digested with *DpnII* restriction enzyme (New England Biolabs, Ipswich, MA) at 37°C for 1 hr, extracted twice with 200 μ L phenol/chloroform/isoamyl alcohol (25:24:1), and then ethanol precipitated and dissolved in 20 μ L H₂O.

Binding biotinylated cDNA to magnetic beads: 100 μ L of Streptavidin-Dynabeads M-280 (Invitrogen, Carlsbad, CA) were prepared in a 1.5 mL Eppendorf tube and then washed twice with 1 mL TE (10mM Tris-HCl, PH7.5, 1mM EDTA) and twice with 200 μ L 1X B&W buffer (5mM Tris-HCl, PH7.5, 0.5mM EDTA, 1M NaCl). The beads were resuspended in 100 μ L 2X B&W buffer (10mM Tris-HCl, PH7.5, 1mM EDTA, 2M NaCl). 10 μ L of *DpnII*-digested cDNA fragments and 90 μ L H₂O were added to the beads. The tube was rotated for 30 min at room temperature and then the beads were washed twice with 200 μ L 1X B&W buffer and twice with 200 μ L TE.

Ligation of barcoded linkers to the bound cDNA: Immediately after binding to Dynabeads, cDNAs were ligated to 5 μ L Linker A (10 μ M) using T4 DNA ligase (Invitrogen, Carlsbad, CA) (5 U/ μ L) for 2 hr at 16°C with intermittent gentle mixing. The beads were washed twice with 200 μ L 1X B&W buffer, washed twice with 200 μ L TE, and resuspended in 200 μ L TE. Linker A was prepared by annealing the following two complementary oligonucleotides in TE plus 50 mM NaCl: 5'-GCCT-CCCTCGCGCCATCAG-XXXX-3' and 5'-phosphate-GATC-XXXX-CTGATGGCGCGAG GGAGGC-3', where *GATC* is the *DpnII* restriction sequence and XXXX represents a

four-base barcode tag specific to each developmental stage: CATG (embryo), TAGT (L1), GATC (L2), CACT (L3), TACG (L4), or GAGC (adult hermaphrodite).

3' cDNA recovery: 100 μ L beads were mixed with 100 μ L phenol/chloroform/isoamyl alcohol (25:24:1), incubated at 65°C for 30min, vortexed at full speed for 5min, and centrifuged at 15,000 rpm for 5 min. The supernatant was collected using Phase Lock Gel (5PRIME Inc., Gaithersberg, MD). DNA was ethanol precipitated and resuspended in 20 μ L H₂O.

PCR amplification: The ligation products from each developmental stage were used as template for two sequential rounds of PCR using 1 μ L of DNA, the forward primer set 5'-GCCT-CCCTCGCGCCATCAG-XXXX-3', and the reverse primer set 5'-GCCTTGCCAGCCCGCTCAG-X-TTTT-X-TTTT-X-TTTT-X-TTTT-3', where the four Xs represent the four nucleotides of the stage-specific barcode tag distributed in order along a polyA tail. The periodic insertion of the X nucleotides improves reliability of Roche/454 sequencing by decreasing homopolymerization of Ts. Samples were extracted with phenol/chloroform/isoamyl alcohol (25:24:1), ethanol precipitated, and resuspended in 50 μ L H₂O. DNA concentration was measured using a Nanodrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE).

454 GS FLX Sequencing: Deep sequencing was performed on the Genome Sequencer FLX system (Roche/454 Life Sciences, Branford, CT) following the manufacturer's protocol.

3'RACE

RNA extraction: Total RNA from *C. elegans* N2 mixed developmental stages was prepared using an adaptation of the RNeasy Mini kit (Qiagen, Valencia, CA). Worms were grown on NGM plates seeded with *E. coli* OP50, washed with M9 buffer, transferred to an RNase-free Eppendorf tube, and dipped into liquid nitrogen. Worms were ground using RNase-free pestles and incubated with RLT buffer (Qiagen) and beta-mercaptoethanol. The lysate was homogenized by aspiration through a 20-gauge needle fitted to a syringe and centrifuged at 13,000 rpm for 3 min. The supernatant was transferred to RNase-free tubes and treated as per the manufacturer's recommendations.

Primer Design: Forward primers were designed to target 7,077 CDS-specific regions from WormBase WS150 for CDSs also contained in the Promoterome (3) and the ORFeome (4-5) collections. For each CDS, in-frame sequence just upstream of and including the STOP codon (based on spliced transcript models) was selected to achieve a T_m of $60^\circ\text{C} \pm 5^\circ\text{C}$ during PCR amplification. Each CDS-specific sequence was preceded by the Gateway adaptor 5'-GGGGACAGCTTTCTTGTACAAAGTGGGA-3' to allow recombination into the pDONR P2R-P3 vector (Invitrogen, Carlsbad, CA). The primer list is available at <http://www.utrome.org>. A universal reverse primer was used, containing a Gateway adaptor (for recombination into pDONR P2R-P3) followed by poly(dT) and a two nucleotide anchor (VN) to target the proximal end of the mRNA polyA tail: 5'-GGGGACAACTTTGTATAATAAAGTTG-T₂₀-VN-3'. Primers were obtained from Invitrogen.

RT-PCR: Total RNA was incubated at 55°C for one hour with Superscript III reverse transcriptase (Invitrogen, Carlsbad, CA) and the universal reverse primer according to the manufacturer's specifications. PCR amplification of 3'UTRs from the single-stranded cDNA reaction was performed in 96-well plate format, using, in each well, the universal reverse primer and a different transcript-specific forward primer as follows: denaturation at 94°C for 30 sec, annealing at 60°C for 30 sec, extension at 72°C for 3 min.

Gateway BP recombination reaction and transformation: 3'UTRs were recombined into the pDONR P2R-P3 entry vector using the BP Clonase II Enzyme Mix kit (Invitrogen, Carlsbad, CA) following the manufacturer's specifications and transformed into MultiShot Stripwell TOP 10 plates (Invitrogen, Carlsbad, CA). The transformed bacteria were grown overnight at 37°C under kanamycin selection.

Sanger Sequencing: Aliquots from overnight cultures of 3'UTR minipools were used as templates for PCR with the M13 primer set as follows: denaturation at 94°C for 30 sec, annealing at 60°C for 30 sec, extension at 72°C for 3 min. 7,077 PCR amplicons were sequenced at Agencourt Bioscience Corporation (Beckman Coulter Genomics, Danvers, MA) using the ABI 3700 automated DNA sequencers.

Preparation of deconvolved 3'UTR libraries: 6,912 minipools containing 3'UTR isoforms were manually streaked onto LB kanamycin plates. From each minipool, eight single colonies were manually isolated and propagated as individual 3'UTR clonal isoforms in 96-well plates (for a total of 55, 296 colonies). Liquid aliquots of isolated clones were re-pooled into eight different super-pools using the Aquarius

automated multi-channel pipetting system (Tecan Trading AG, Switzerland), resulting in eight libraries that should each contain zero (if no insert was cloned) or one unique 3'UTR isoform per targeted CDS. These deconvolved libraries (labeled A-H) were sequenced using Solexa/Illumina and FLX Roche/454 platforms.

Sample preparation and sequencing with Illumina Genome Analyzer II:

Plasmid DNA was recovered using standard alkaline lysis from overnight cultures of the eight deconvolved libraries (A-H). Inserts from each library were amplified by PCR using common Forward (5'-GTTTCTCGTTCAACTTTCTTGTACAAAGTGGGA-3') and Reverse (5'-ATAATGCCAACTTTGTATAATAAAGTTGTTTTTTTTTTTT-3') primers. The eight amplicon libraries were purified using MinElute columns (Qiagen), treated to create blunt ends using T4 DNA polymerase (New England Biolabs, Ipswich, MA) and T4 polynucleotide kinase (New England Biolabs), incubated overnight with DNA ligase (New England Biolabs), and then sonicated using the Bioruptor UCD-200 (Diagenode Inc., Sparta, NJ) for 30 min in cycles of 30 sec ON, 30 sec OFF. The resulting 8 fragmented libraries were prepared for Illumina sequencing according to manufacturer's recommendations, and six of the libraries were sequenced using the Illumina Genome Analyzer II system (Illumina, Inc., San Diego, CA) in the Sachidanandam laboratory at the Mount Sinai School of Medicine (New York, NY).

Sample preparation and sequencing with 454 GS FLX: Plasmid DNA was recovered from overnight cultures of the eight deconvolved libraries (A-H) using the Wizard Plus miniprep kit (Promega, Madison, WI) and used as template for PCR amplification with eight barcode-matched primer pairs: AdaptorA::Barcode::Forward (5'-

GCCTCCCTCGCGCCATCAG-XXXX-Forward-3') and AdaptorB::Barcode::Reverse (5'-GCCTTGCCAGCCCGCTCAG-XXXX-Reverse-3'), where Forward and Reverse are the same sequences used for Illumina above and barcode tags, XXXX, for libraries A-H are A: CATG, B: TAGT, C: GATC, D: CACT, E: TACG, F: GAGC, G: CTGC, H: ATCG. Barcoded PCR amplicons from all eight libraries were combined and purified using the MinElute PCR purification kit (Qiagen). Because the FLX platform output for samples of variable length is biased toward shorter reads, the combined sample was split into two equal batches: (i) untreated, and (ii) treated with the Agencourt AMPure SPRI PCR purification kit (Beckman Coulter Genomics) to enrich for longer fragments by removing fragments shorter than 100 bp. AMPure library DNA was evaluated for quality and quantified using a BioAnalyzer DNA 1000 lab chip (Agilent, Santa Clara, CA). DNA concentration in ng/ μ l was converted to molecules/ μ l and adjusted to 2×10^5 molecules/ μ l in TE buffer. The resulting fragments were prepared for 454 sequencing according to the manufacturer's recommendations and sequenced using the Genome Sequencer FLX system.

cDNA libraries

Two sets of polyA⁺-selected cDNA libraries from the Kohara laboratory and prepared from various stages of *C. elegans* development were used (totaling 152,000 cDNA clones).

First, lambda-zap embryonic and *him-8* mixed stage libraries were prepared without any amplification or rationalization steps. These libraries are of very high quality,

with $\sim 10^{-4}$ mismatches per base relative to the genome (after removal of ~ 200 errors detected in the genome) and less than 3% structural defects or artifacts.

The second set consists of full-length L1, L2, L4 and mixed stage libraries prepared by S. Sugano Y. Suzuki and Y. Kohara using the oligo cap selection procedure (6). These libraries were designed to include the entire transcript, from 5' capped first base to poly A, and are validated by the fact that >99% of the clones with a *trans*-spliced leader in this collection contain the entire leader sequence (21 to 23 bases long). These collections allowed identification of 12 varieties of SL as well as 3,953 genes that are not *trans*-spliced.

Sequencing traces from a polyA⁺-selected library (n=14,811 cDNA clones), generously provided by Exelixis Inc. (San Francisco, CA), along other publicly available cDNAs and EST data obtained from the NCBI Trace and dbEST archives (in the form of either sequences or traces), were also manually curated at NCBI as part of the experimentally supported worm transcriptome project known as AceView (7).

The combined cDNA dataset provides experimental evidence for 16,659 distinct polyA sites in 11,180 genes. These data are all publicly available from <http://www.aceview.org> and <http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly>.

RNA-Seq datasets: Illumina data for staged samples (L2, L3, and L4 larvae and young adults) from the modENCODE transcriptome project, described in (8), were obtained from NCBI GEO (SRX001872-SRX001875). Additional published Roche/454 datasets for the L1 stage (9) were also analyzed. Together, these data provide support for 8,332 polyA sites for 7,461 genes.

Sequence analysis of primary datasets

Genome version: All data were aligned to *C. elegans* genome sequence version CE6 (on which WormBase WS190 gene annotations are also anchored).

PolyA capture libraries: 454 sequence data from three independent runs were pooled. Runs A and B (Run 1) comprised sequences from combined staged samples (Run A: embryo, L1-L4, adult hermaphrodite; Run B: embryo, L1-L4, adult hermaphrodite, adult male); Run C (Run 2) contained mixed sequences from four dauer mutants: *daf-2*, *daf-7*, *daf-9*, and *daf-11* (see Table S2 for read counts from each run). Forward reads were identified by the pattern 5'-XXXX-GATC-N_m-X'-AAAA-X'-AAAA-X'-AAAA-X'-AAAA-3', where GATC is the *DpnII* restriction site, N_m is a sequence of length *m* extending from the *DpnII* site to the end of the 3'UTR, and X'X'X'X' is the reverse complement of the matching 3'end barcode. Reads that did not contain a decipherable barcode tag were discarded. Barcodes were used to identify the library of origin for the remaining reads, and sequences were processed to remove the 5' and 3' adaptor sequences and barcode tags. Sequences retaining length ≥ 15 nt were aligned to the genome using BLAT (10), with a maximum intron size of 1000, minimum window size of 5, and maximum gap of 6. Best matches were selected, and multiple alignments reported if present in more than one genomic location. Alignments in PSL format were converted to SAM format using the psl2sam.pl script provided with SAMtools (11). Alignments for sequences that did not reach the polyA were set aside; the remaining alignments were further annotated.

3'RACE: RACE clones were sequenced by three different methods. Sequences from ABI or SCF files were trimmed of vector sequence and filtered for empty vectors and putative primer-dimer products. The remaining sequences were aligned to the genome using BLAT (10) and WU-BLAST 2.0 (12). Aligned regions were scanned for the presence of detectable CDS-specific primer and terminal polyA sequences (defined as 10 or more consecutive As with zero or one intervening nucleotide).

For Illumina data, 50 million sequence reads from six independently sequenced libraries were aligned to both the genome and to AceView transcripts using the AceView aligner (<http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/Software>). PolyA sites were identified by trimming reads beginning with at least 5 consecutive T's or ending in at least 5 consecutive A's, and then mapping either the full remaining tag sequence or a version lacking the last two nucleotides upstream of the polyA (since we had previously determined that the cloned RACE products contained a high proportion of T to C base changes at these positions, which pair with the two anchor nucleotides in the universal reverse primer). Overlapping mapped reads were assembled into contigs, and these were used for further annotation.

From the two 454 runs, a total of ~170,000 reads corresponding to ~85,000 unique sequences were produced. Initial processing, BLAT alignment, and conversion to SAM format were the same as described above for polyA capture data.

Alignments from all three platforms were then considered together and, where possible, alignments were assigned to the putative plate-well of origin based on the identity of the corresponding primer; for deconvoluted libraries, the combination of

primer and barcode, if detectable, was used to assign a putative location in the isolated clone library plates.

cDNAs and ESTs: cDNA clones from the yk collection were sequenced using the Sanger method. All cDNA and EST data from this collection and from other sources (as described above) were aligned to the genome and annotated using AceView tools; these were further hand-curated by visual inspection of multiply aligned ABI sequence traces, where available.

RNA-seq: Illumina and Roche/454 datasets (described above) were aligned to both the genome and AceView transcripts using the AceView software tools (<http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/Software>). PolyA sites were identified trimming reads beginning with at least 5 consecutive T's or ending in at least 5 consecutive A's, and mapping the remaining sequence tag as above. Overlapping mapped reads were assembled into contigs, and these were used for further annotation.

cDNA and transcriptome annotation

Annotation of independent datasets: Sequences from RACE and polyA capture with best-hit alignments or assembled contigs near the last exon of a (targeted, for RACE clones) CDS were defined as candidate 3'USTs (UTR Sequence Tags). USTs were initially assigned to the overlapping or immediately adjacent upstream CDSs from WormBase WS190 gene models (<http://www.wormbase.org>); these assignments were later revised using AceView genes (<http://www.aceview.org>), which in some cases revealed that the combined data were incompatible with existing WS190 (or WS150)

CDS models. In such cases, USTs from RACE experiments were retained as evidence of transcriptional activity but were removed from the final list of cloned 3'UTRs. USTs with a contiguous BLAT alignment extending through the STOP codon of a valid AceView CDS model and containing polyA sequences were considered to be bona fide complete 3'UTR isoforms with full-length coverage. Those with incomplete 3'UTR coverage and/or no detectable polyA sequence were annotated as partial 3'USTs and used to refine 3'UTR boundaries. Mapped tags from short read data were assembled into contigs and used together with cDNA, EST, UST data to define transcribed regions. The combined data were used to refine and extend existing AceView genes. Data mapping downstream of (but not overlapping) an existing gene were extended *in silico*, where possible, and assigned to the nearest gene upstream or else used to define new transcriptional units. All annotated 3'USTs and 3'UTRs were used for subsequent analyses.

Definition of representative polyA sites and 3'UTR isoforms: To define 3'UTR isoforms and assign a single representative polyA site per isoform, we combined evidence for polyA addition sites from all four independent data sources in the 3'UTR compendium into a single large dataset and performed an iterative local clustering procedure using their chromosomal coordinates. The clustering software is included in the AceView software, available from <http://www.ncbi.nlm.nih.gov/IEB/Research/AceView/Software>. When evidence sources were attached to a known gene model, clustering was performed on a per-gene basis. The local maximum for each cluster was computed and used as the position of the reported (“representative”) polyA addition site

for each 3'UTR isoform. The spread of the clusters extends from one up to around 20 nucleotides, with 86% of all individual data points falling within 4 nt of the representative polyA site (Fig. S6).

Using this clustering procedure, each 3'UTR isoform was then defined as a unique sequence span that extends from a specific CDS end and terminates downstream at a distinct “canonical” polyA addition site: 3'UTR sequences that share the same CDS end and terminate within the same polyA cluster were defined as examples of the same isoform, whereas 3'UTR sequences that terminate within different polyA clusters (even if linked to the same CDS) were defined as distinct isoforms. Isoforms of a gene that were represented by less than 5% of the total polyA counts for that gene, isoforms that were not supported by two or more independent pieces of evidence, and those that were shorter than 20 nt (which mostly contained dubious cloning artefacts) were removed from the final dataset. For reporting purposes and all downstream analyses involving isoforms, we considered only the “representative” polyA coordinate for each reported 3'UTR isoform.

Identification of PAS sites: The 50 nt regions immediately upstream of all polyA sites were scanned in an unbiased way for all possible 5 to 10-mer sequences to identify any statistically over-represented motifs. The only motifs returned from this exercise were the canonical PAS sequence (AAUAAA) and several closely related sequences. The distribution of all over-represented hexamers peaked at a start position of -19 nt from the polyA site, which was taken as the most likely position of the PAS site. All of the 3'UTR isoforms in the compendium were then scanned for the canonical PAS

sequence and any hexamer with an edit distance of 1 or 2 nt. Because it is not possible to definitively identify the "real" PAS site, we scanned for hexamers in a preferred order based on their observed frequency of occurrence in 3'UTRs between 10 and 30 nt upstream of the polyA site, and considered those occurring at a frequency of $\geq 1\%$ as putative PAS motifs. We used the first occurrence of a putative motif in the ordered list as the most likely functional PAS sequence. UTRs that did not contain one of the resulting 26 putative PAS motifs within this interval were termed "no PAS".

Analysis of genomic nucleotide frequencies in the 120 nt region spanning ± 60 nt of polyA sites showed that strongly supported PAS sites, which we consider the best candidates for recognition by CPSFs for 3'end-processing (13), also show an enrichment of T's that peaks at +5 nt downstream of the putative PAS site (Fig. 1D). These include nine principal motifs are: AATAAA (the canonical PAS hexamer), AATgAA, tATAAA, cATAAA, gATAAA, AtTAAA, tATgAA, AgTAAA and cATgAA (where upper-case letters are identical with the canonical hexamer, and lower-case letters indicate substitutions).

Comparison of 3'UTRome and WormBase annotations: Operon, Gene, CDS, and 3'UTR annotations for WS190 were obtained from WormBase. For comparative purposes, any 3'UTR in our compendium whose 5'end matched a WS190 CDS and whose 3'end was within 10 nt of an annotated WS190 3'UTR was considered identical; all others were labeled as "longer" or "shorter" than the WS190 3'UTR, as appropriate. 3'UTRs in our dataset that matched a WS190 CDS end but had no corresponding WS190 3'UTR were annotated as "new 3'UTRs". 3'UTRs that did not match a WS190

gene model, but matched an alternate transcript model that could be generated from experimental data, were annotated as 3'UTRs of "new AceView genes". These data are summarized in Fig. S1.

Intron analysis: Gapped sequence alignments were examined for the presence of putative splice signal consensus sequences, and introns were annotated as appropriate. Numerous gapped alignments of polyA capture data spanned bona fide splice junctions but were on the opposite strand and thus contained the reverse complement of known splice consensus signals. Such alignments were observed to occur most frequently within coding regions; these were determined most likely to represent mis-priming in A-rich regions and were discarded. A subset of gapped alignments for these data contained terminal segments <10 nt; these appeared to be alignment artifacts of degraded sequence data and were also discarded. A total of 363 3'UTRs for 192 genes were determined to contain bona fide introns, based on the presence of a strongly supported CDS upstream with no evidence for another CDS that could extend into the putative 3'UTR. The 3'UTRs with an intron that could also occur internally within the CDS of an alternative isoform were not counted in this set.

Operon and SL analysis: To compare the six categories of genes analyzed in Fig. 2, we selected a subset of trans-spliced and non-trans-spliced genes for which assignment to a unique category could be unambiguously determined. Among the SL1 trans-spliced genes, we identified 574 SL1 genes occupying the first position of an operon (genes fully supported from SL1 to polyA and separated by at most 300 bases from the next gene in cis, which is itself trans-spliced mostly to SL2) and 3,530 SL1-

genes undoubtedly not in an operon (selected as followed either by another SL1-gene (n=1,749) or by a confirmed non-trans-spliced gene (n=1781)); these two subsets were found to be indistinguishable and were merged in Fig. 2.

Directed RT-PCR assay for retained 3'UTR introns: Total RNA was extracted from mixed-stage worms and RT-PCR was performed essentially as described above. 1 μ g of total RNA from mixed-stage worms was used as template for a first strand reaction using the universal anchored poly(dT) reverse primer. PCR was performed using internal primer pairs flanking putative retained introns in the 3'UTRs of two genes: *par-5* (Forward: 5'-GAG GGA AAC CAG GAA GCT GGA AAC TAA-3'; Reverse: 5'-GAT GCT ATT GCG CAG TGT TGT ATG GAG TAT TGG) and *sams-1* (Forward: 5'-GCC ACA TCT GCT ATC GCT CAC TAA-3'; Reverse: 5'-CAA GAC AGC TCA GCG GGT AGC GGA AAC CG-3'). Products were separated on a 2% agarose gel and visualized with ethidium bromide.

Developmental stage analysis: The staged polyA capture dataset was used for this analysis, since this dataset can provide specific information on the abundance of alternative 3'ends expressed in different stages. Since the total polyA tag count differed between libraries, the total number of read counts from each stage was normalized to match the total counts in embryo, and counts for individual isoforms scaled proportionally to reflect the relative expression level in different lifestages. The number of isoforms detected per gene was evaluated for each developmental stage and across all stages. To study the expression of long vs. short isoforms we identified genes showing exactly two distinct 3'UTR isoforms (2,295 in total) and restricted our analysis

to a stringent subset of 1,960 genes showing at least 5 read counts for the most abundant isoform (Supplementary Dataset S5). To identify genes showing preferential isoform usage, we further selected a subset of genes that showed, in the cumulative dataset, at least twice as many total counts for one isoform as the other (915 genes for long>short; 615 genes for short>long). The per-stage relative expression of a particular isoform of a gene was calculated by dividing the counts for that isoform by the total counts for both isoforms expressed during that stage. The relative expression of an isoform across all stages was calculated as the ratio of the normalized counts of the isoform in a single stage to the total normalized counts of both isoforms of the gene across all developmental stages.

To identify genes that exhibit a differential preference for 3'UTR isoforms during development (i.e. 3'UTR isoform "switching"), we filtered the 1,960 genes described above using the following criteria: 1) isoform 'a' was more abundant than the isoform 'b' in one developmental stage, and isoform 'b' was more abundant than isoform 'a' in any other developmental stage; 2) the total abundance of all isoforms for the same was ≥ 20 counts (abundance was based on normalized polyA capture counts). We identified 612 genes exhibiting such 3'UTR isoform switching (see Supplementary Datasets S5, S6). To obtain a "high-confidence" subset of these genes, we imposed two additional criteria: 1) the ratio of counts for isoform 'a' to counts for isoform 'b' (a/b) was ≥ 2 fold in one stage, and the ratio of isoform 'b' to isoform 'a' counts (b/a) was ≥ 2 fold in another stage; 2) the difference in support between isoform 'a' and 'b' was ≥ 5 counts within each developmental stage in which switching occurred. Of the 612 genes, 263 genes passed

these filters (see Supplementary Datasets S5, S6).

miRNA target prediction and 3'UTR conservation analysis

3'UTR alignments: We used the Galaxy server processing pipeline (14) and the UCSC Table Browser (15) to prepare a multiple alignment file (MAF) for *C. elegans* (WS190/CE6), *C. remanei*, *C. briggsae*, *C. brenneri*, and *C. japonica*. The MAF file did not contain overlapping blocks or gaps in the *C. elegans* sequence. We then extracted a MAF file for each of the initial 33,909 3'UTRs from the 3'UTRome. Overlapping 3'UTRs were fused to yield 15,685 unique 3'UTR regions that were used for subsequent analyses.

miRNA sequences: We used for our analyses 174 *C. elegans* mature miRNA sequences downloaded from miRBase version 14 (16) and 9 novel miRNAs determined by miRDeep2 (17). These miRNAs were grouped into 124 miRNA families sharing the same seed sequence at nucleotides 2-7 in each miRNA.

Identification of miRNA seeds in 3'UTRs: The PicTar algorithm (18-19) was used to identify non-conserved and conserved miRNA seeds in mRNA sequences, which were defined as regions in mRNA sequences with perfect base complementarity to miRNA 6-mer seeds (nucleotides 1-6 or 2-7 at the miRNA 5' end). Seeds conserved in 3 species (*C. elegans*, *C. remanei*, *C. briggsae*) and those conserved in 5 species (*C. elegans*, *C. remanei*, *C. briggsae*, *C. brenneri*, *C. japonica*) were identified. PicTar was further used to predict and assign scores for full miRNA binding sites, as described (19). The probability of a conserved predicted miRNA target seed site being functional in 3-

way or 5-way species comparisons is 2.7 and 3.1, respectively. The comprehensive list of PicTar predictions is available from the UTRome (<http://www.utrome.org>) and modENCODE (<http://www.modencode.org>) websites.

Comparison with Lall et al., 2006: We compared our updated miRNA target predictions within our previous predictions for *C. elegans* (19). For this comparison, we considered only those miRNAs that were analyzed in Lall et al. and the set of unique (non-overlapping) 3'UTRs contained in the UTRome to which the Lall et al. target site predictions map; thus, any predicted sites from either study that were not contained in 3'UTRs considered in the other study were not included in this comparison. In addition, we excluded from the comparison the two miRNAs cel-miR-68 and cel-miR-69 used in the Lall et al. analysis (because they are currently annotated as siRNAs in WormBase), and the seven miRNAs cel-miR-42, cel-miR-239b, cel-miR-248, cel-miR-250, cel-miR-252, cel-miR-253 and cel-miR-358 (because the reported sequences of their seed regions, i.e. positions 1-7 or 2-8 in the mature miRNA, were different according to Rfam version 6 and miRBase version 14).

We then compared the number of predicted sites from this study with the previous set of predictions within the sequence space analyzed in both studies (summarized in Table S7). From our new prediction set, 5,943 predicted miRNA target sites fall in this intersecting sequence space, of which 580 sites (9.8%) were not identified in the Lall et al. study. We attribute the identification of these new sites to improved multi-species alignments and the inclusion of newly sequenced species in the alignments.

Of the 11,131 miRNA target sites predicted in the Lall et al. study, 6,474 sites were located in the intersecting sequence space. In the current study, we recovered 5,363 of those sites, or 82.8%; the remaining 1,111 sites from Lall et al. (17.2%) could not be recovered. The loss of these sites is explained by the fact that the Lall et al. study used some sequence regions outside the 3'UTRome for the initial predictions; if conserved sites were identified in these regions, then non-conserved sites falling within shorter 3'UTRs would also be designated as candidate target sites due to the presence of the initial conserved site. However, if this sequence region is not used for the initial identification, and no other conserved sites are identified within the sequence space analyzed, then non-conserved sites will not be considered by the algorithm as potential target sites, and previously predicted sites would then be lost.

We note that many previously predicted target sites from Lall et al. that fall outside the spans of our 3'UTR annotations (either because they targeted genes for which we have no 3'UTR annotation, or because we previously used up to 500nt spans downstream of any CDS if no 3'UTR was available) are not currently supported by empirically defined 3'UTR regions.

Conserved blocks not explained by miRNA seeds: To identify conserved sequence blocks that do not correspond to conserved miRNA seed sequences, all (reverse complemented) miRNA seeds were masked with Ns in the 3'UTR multiple alignment files (MAFs), and all remaining k-mers ($k \geq 6$) conserved in 3 species (*C. elegans*, *C. remanei*, *C. briggsae*) or in 5 species (*C. elegans*, *C. remanei*, *C. briggsae*,

C. brenneri, *C. japonica*) were identified. The alignment of any conserved 6-mer was extended as far as possible in both directions.

Distribution of conserved PAS motifs and sequence blocks: We excluded from this analysis all 3'UTRs shorter than 10 nt and those contained within coding sequences of alternative CDSs, resulting in a final set of 24,858 3'UTRs, of which 8,319 genes have a single isoform, 3,320 genes have exactly two isoforms, and 2,616 have more than two isoforms. All conserved miRNA seeds in 3'UTRs, all 29 putative PAS motifs, and all conserved sequence blocks as defined above were investigated with respect to their positions relative to UTR ends. A PAS site was considered as “conserved” in this analysis if it was found in *C. elegans* and the same or another PAS motif was found within a window of ± 5 nucleotides in aligned *C. briggsae* and *C. remanei* sequences. Only PAS sites in genes with one isoform or exactly two isoforms, where the longest isoform was at least 100 nt, were considered. The set of genes with 2 isoforms was further filtered to require a length difference of at least 50 nt between the short and long isoform; if this requirement was not met, the short isoform was discarded and the gene was treated as having a single long isoform for this analysis.

Analysis of overlaps between experimentally determined ALG-1 binding sites and conserved sequence motifs: We compared recently published *in vivo* Argonaute (ALG-1) binding sites (20) with our conserved sequence motifs (predicted miRNA target sites and conserved sequence blocks). For this analysis we considered only those 3'UTRs containing or overlapping at least one ALG-1 binding site. The probability of predicted miRNA target seed sites from 3-way species alignments (*C.*







elegans, *C. briggsae*, *C. remanei*) occurring within an ALG-1 binding site was 0.75. As a control, we calculated the overlap between ALG-1 sites and 6-mers (the length of predicted miRNA seed sites) placed at random positions along the length of annotated 3'UTRs ($p=0.43$), which represents a lower bound to the resolution at which we could discern meaningful correlations with ALG-1 sites. The overlap was not significant for the thousands of other conserved blocks that are not explained by predicted miRNA target sites or by conserved PAS sites (0.54 vs. 0.48 for random controls). These results indicate that the overlap between ALG-1 sites and predicted miRNA target sites is highly significant, and that while other conserved sequence blocks are likely functional, they are not, overall, directly related to microRNA function.

Data Availability

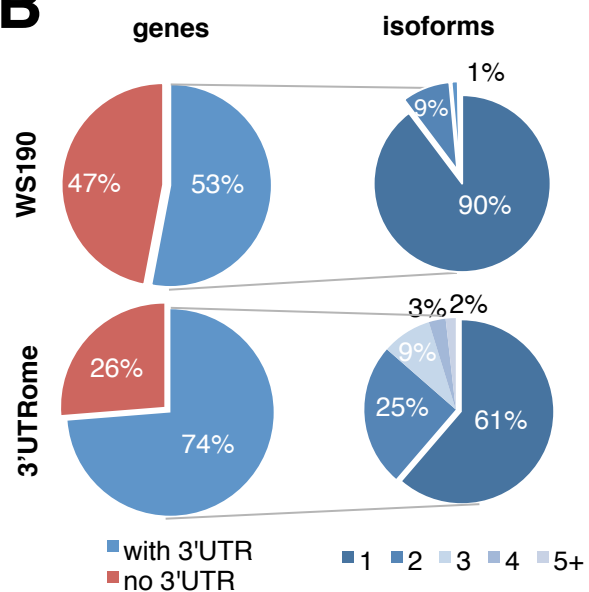
Raw data from Roche/454 and Illumina sequencing were deposited at NCBI Short Read Archive (accession numbers: GSM443959-GSM443964, GSM446651-GSM446661, GSM469439, GSM469976) and GEO (accession number: GSE17781). ABI traces and UST sequences were deposited in the NCBI Trace (trace IDs: 2216286010-2216288816) and dbEST (dbEST IDs: 63366486-63366494) archives. Genome alignments and annotations for 3'UTRs, polyA sites, and PAS sites were deposited with the modENCODE DCC (accession numbers: 515, 896, 992, 2327-2337, 2455-2465, 2482, 2484, 2501 and 2745), along with metadata describing experimental and bioinformatic protocols and links to raw datasets in NCBI public repositories. See also

Datasets S1-S7. Multiple web portals will provide access to 3'UTRome data, including UTRome.org, AceView.org, modENCODE.org, and WormBase.org.

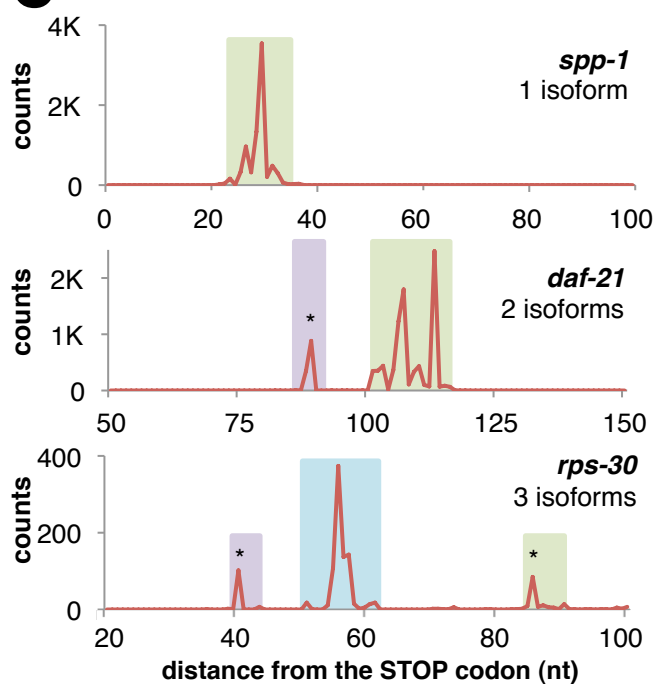
A

genes (polyA counts) with annotated 3'UTRs		new 3'UTRs	new AceView genes
WS190	 10,802 (12,877)	—	—
3'UTRome	same (+/-10nt) 	 4,466 (6,177)	 1,031(1,490)
	longer 		
	shorter 		

B



C



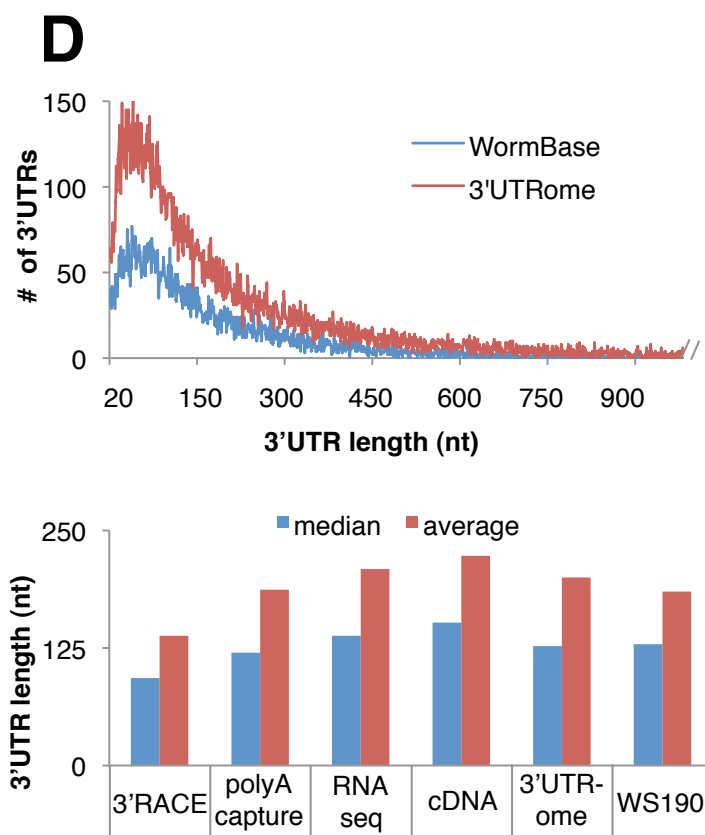


Figure S1. Overview of the 3'UTRome.

A,B,D) Comparison with WormBase (WS190) gene models. A) The 3'UTRome contains 3'UTRs of similar, longer, and shorter length for WS190 genes with annotated 3'UTRs (left column); 3'UTRs for WS190 genes with no annotated 3'UTRs (middle column); and 3'UTRs for transcriptional units not annotated in WS190 (AceView genes) (right column). B) WormBase WS190 contains 3'UTR annotations for 10,802 protein coding genes (53% of total); of these, only 10% are annotated with two or more 3'UTR isoforms. Our 3'UTRome covers 14,918 WS190 coding genes (74%), 39% of which possess two or more isoforms. C) Observed counts of polyA sites from independent sequence reads cluster together, defining one or more 3'UTR isoforms. Variability within polyA clusters (colored boxes) spans up to ~20 nt. Asterisks denote newly identified 3'UTR isoforms. D) Top panel: The length distribution of 3'UTRs in WS190 and 3'UTRome datasets are homothetic. Bottom panel: median (blue bar) and average (red bar) length of 3'UTRs detected in each dataset.

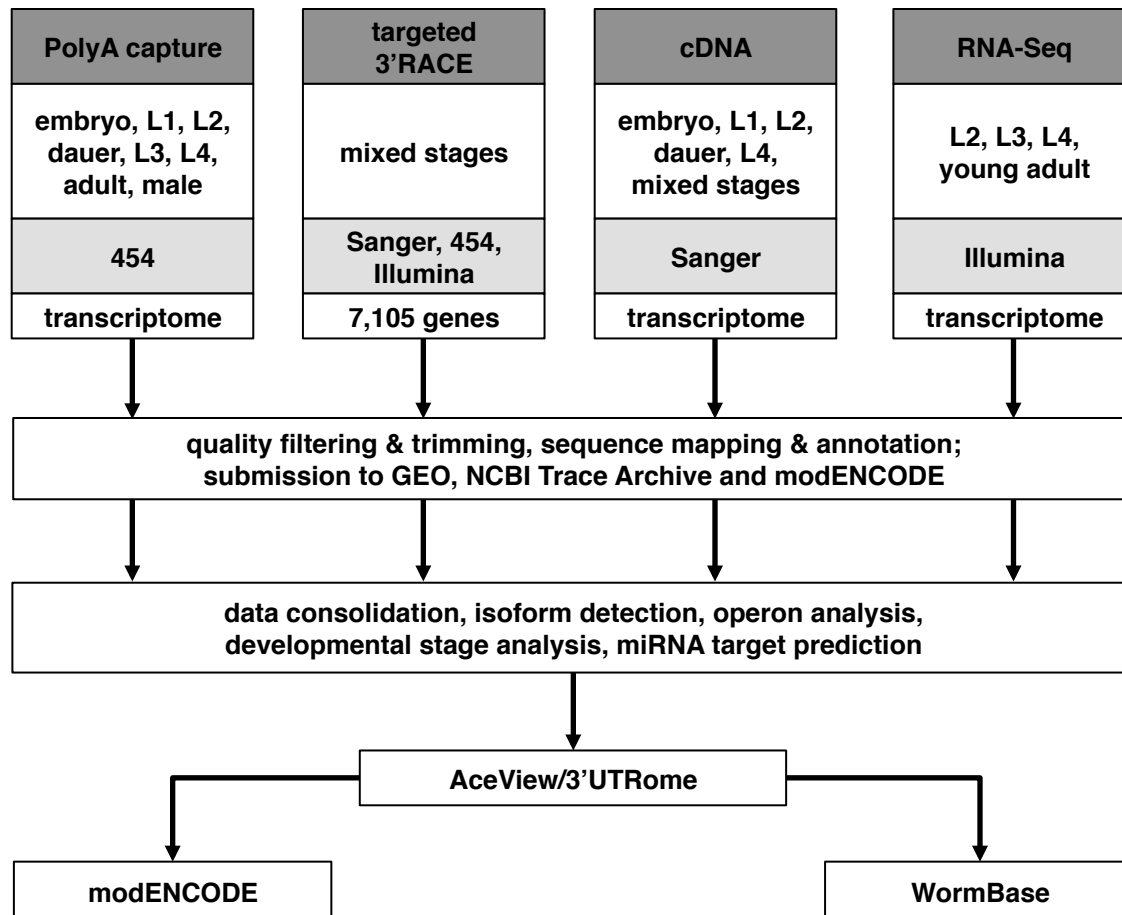


Figure S2. Overview of 3'UTRome pipeline.

The 3'UTRome project is composed of four datasets. PolyA capture and targeted 3'RACE were generated in this study, while publicly available cDNA and RNA-Seq data were reanalyzed and curated as part of this effort. Barcoded polyA capture tags contain the 3' end portions of 3'UTRs from staged samples; 3'RACE products directed at 7,105 coding genes were cloned from mixed-stage samples. The cDNA dataset represents AceView-curated cDNA and EST sequences using, where possible, the original traces from cDNA libraries produced by the Kohara laboratory, Exelixis, and others obtained from the NCBI trace repository, as well as cDNA sequences from NCBI sequence repositories (GenBank, dbEST). The RNA-Seq dataset consists of published data for staged mRNA samples from the modENCODE *C. elegans* transcriptome project (8) and previously reported L1-stage data (9). Datasets were sequenced as indicated (gray shading). Sequences were processed (to remove vector, linker, barcode, and polyA sequences), filtered for read quality, and aligned to the *C. elegans* WS190/CE6 genome. The consolidated datasets were used to define a compendium of 3'UTR isoforms, which was used for downstream analyses of 3'UTR structure and function. Raw data and annotations for the compendium are available in public repositories, including NCBI GEO and Trace Archive, the 3'UTR-centric 3'UTRome database (<http://www.utrome.org>), AceView (<http://www.aceview.org>), modENCODE (<http://www.modencode.org>), and WormBase (<http://www.wormbase.org>). Supplementary Materials and Methods provide additional details on data production and analysis.

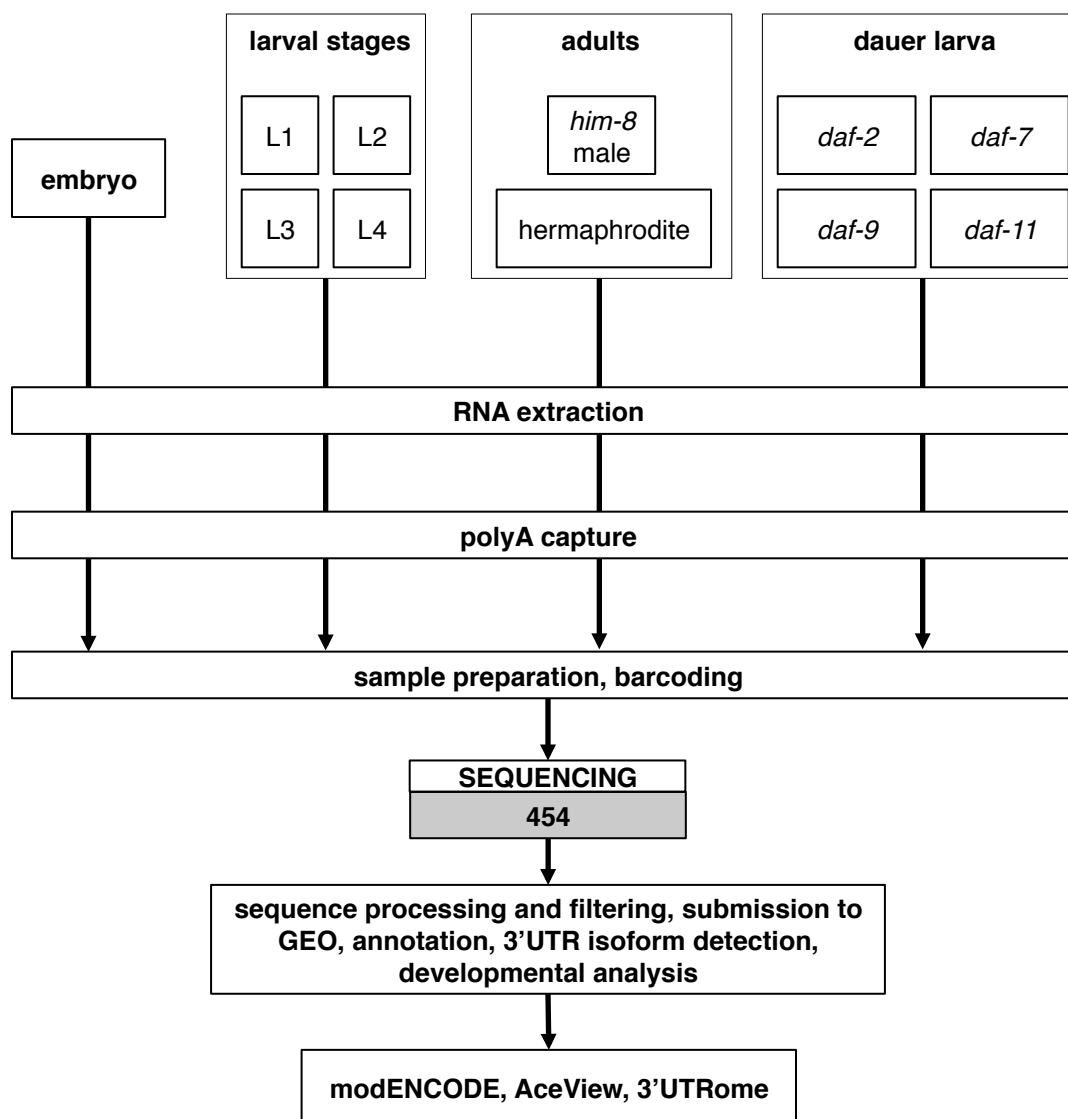


Figure S3. Workflow for polyA capture assay.

Barcoded polyA capture libraries were prepared using total RNA from staged animals and sequenced by Roche/454. Reads were filtered for quality, processed to remove adaptor and barcode sequences, and aligned to the WS190/CE6 genome build. Raw and processed sequence files were submitted to GEO. Alignments were consolidated with the other 3'UTR datasets and annotated with respect to WS190 and AceView gene models. Data and annotations are available in AceView, 3'UTRome, and modENCODE databases (see Supplementary Materials and Methods for details).

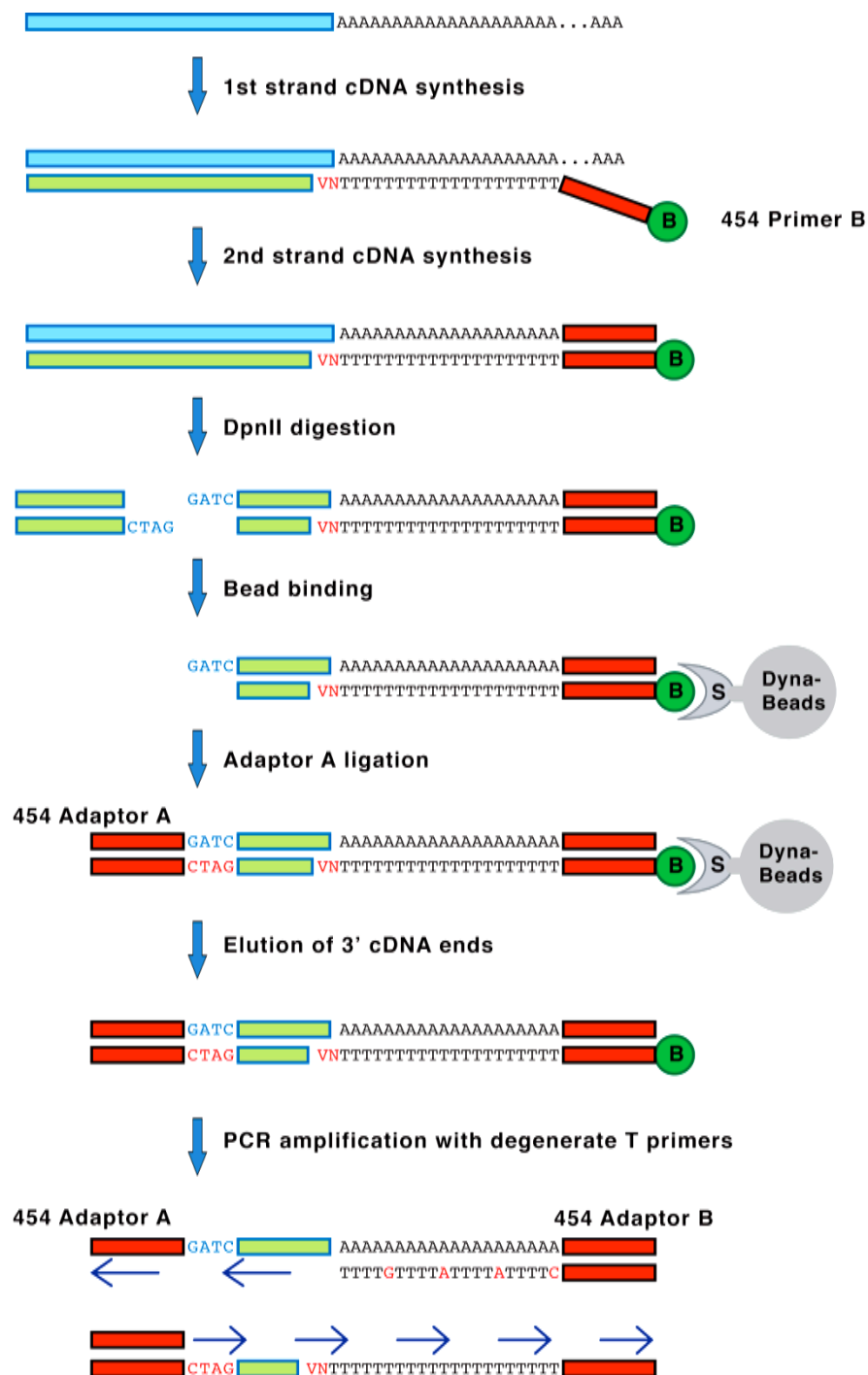


Figure S4. PolyA capture protocol.

Total RNA from staged samples (Figure S3) served as template for a first-strand reverse transcriptase (RT) reaction with an anchored, biotinylated poly-dT primer. Second-strand synthesis with T4 DNA polymerase produced dsDNA products that were digested with *DpnII*. Three-prime terminal fragments were recovered using streptavidin beads, ligated with barcoded 454 sequencing primers, PCR amplified, and subjected to pyrosequencing (see Supplementary Materials and Methods for details).

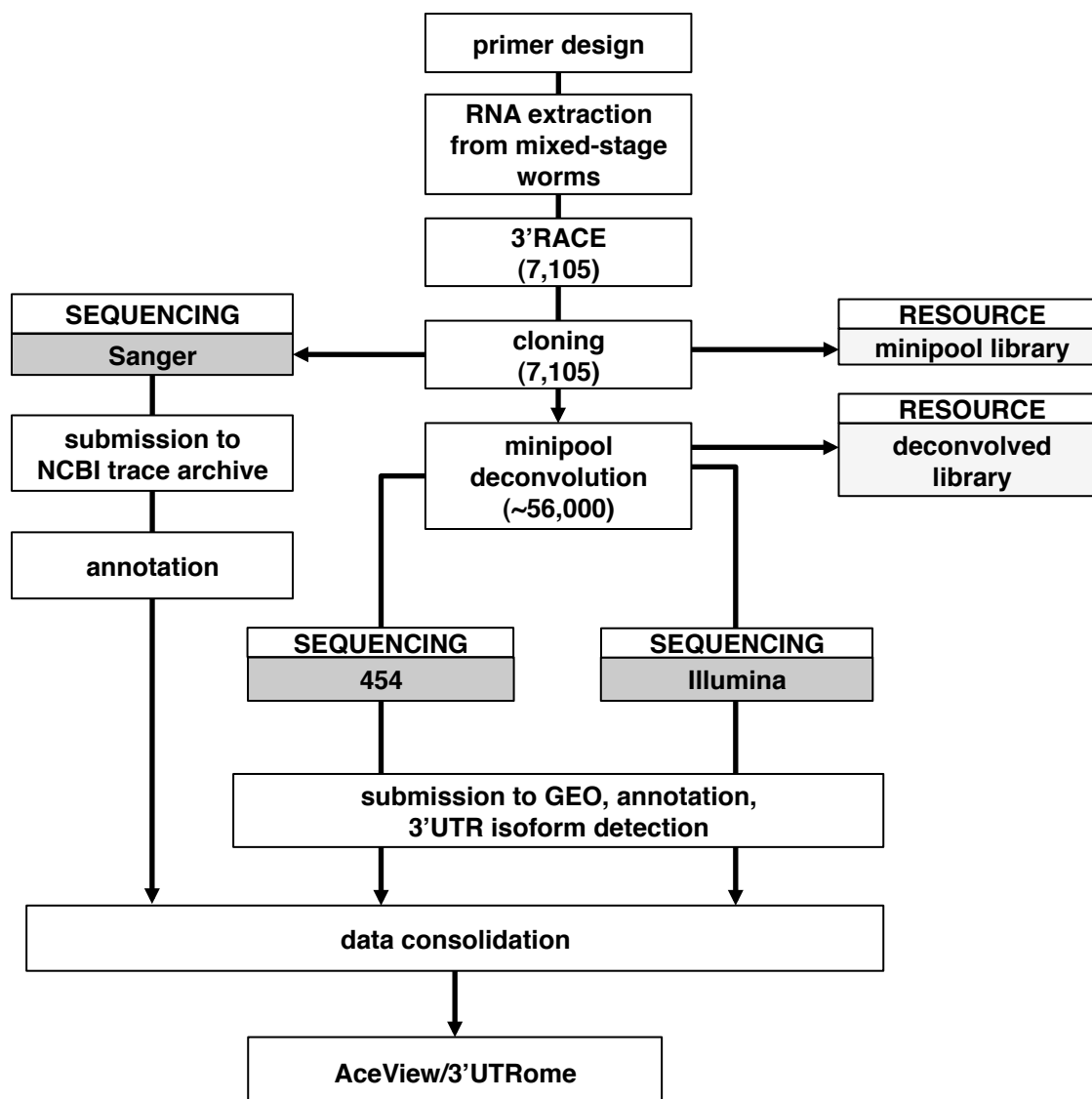


Figure S5. Workflow for 3'RACE.

A 3'RACE cloning pipeline was designed to target 3'UTRs of 7,105 CDSs for 6,741 genes previously included in the Promoterome (3) and ORFeome (4-5) collections. 3'RACE products were generated from total RNA isolated from mixed developmental stages, cloned into Gateway™ vectors, and collected as minipools of products for each target. Minipools were sequenced using the Sanger method. Eight individual colonies per minipool were isolated and re-pooled into eight bar-coded libraries containing one individual clone per targeted gene. Barcoded libraries were sequenced using Illumina and Roche/454 platforms. Minipool and deconvolved single-clone sequences were trimmed for vector and barcode sequences, filtered for quality, and aligned to the WS190/CE6 genome sequence. Alignments that extended beyond the CDS-specific primer were annotated and consolidated with other 3'UTRome datasets in AceView and 3'UTRome databases (see Supplementary Materials and Methods for details).

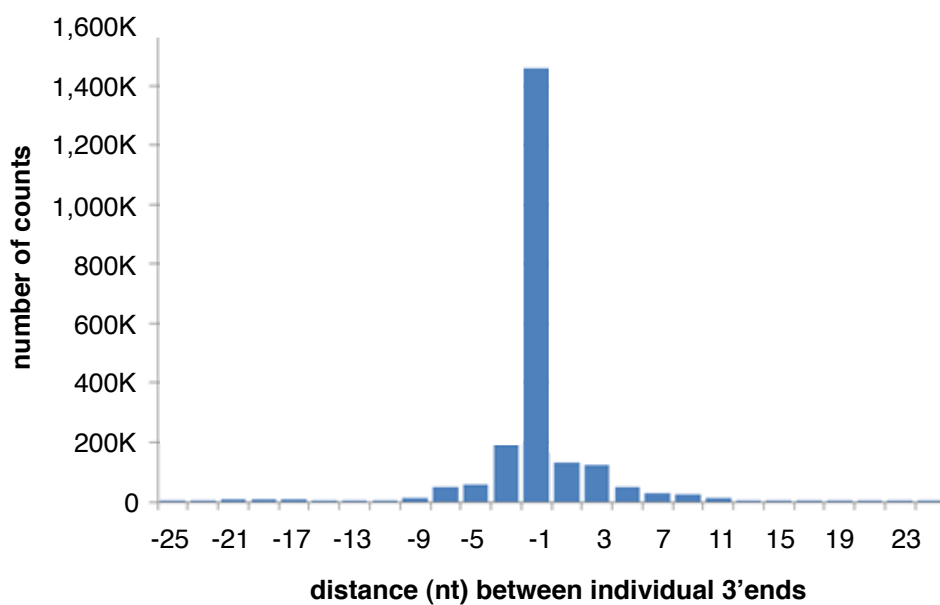


Figure S6. Distance between individual 3' ends and the representative polyA addition site for a cluster.

Frequency distribution of distance (in nucleotides) between the representative polyA site in a cluster and all other polyA sequence tags in the same cluster. Data are cumulative for all polyA clusters in the 3'UTRome. 86% of individual polyA tags fall within 4nt of the representative polyA site.

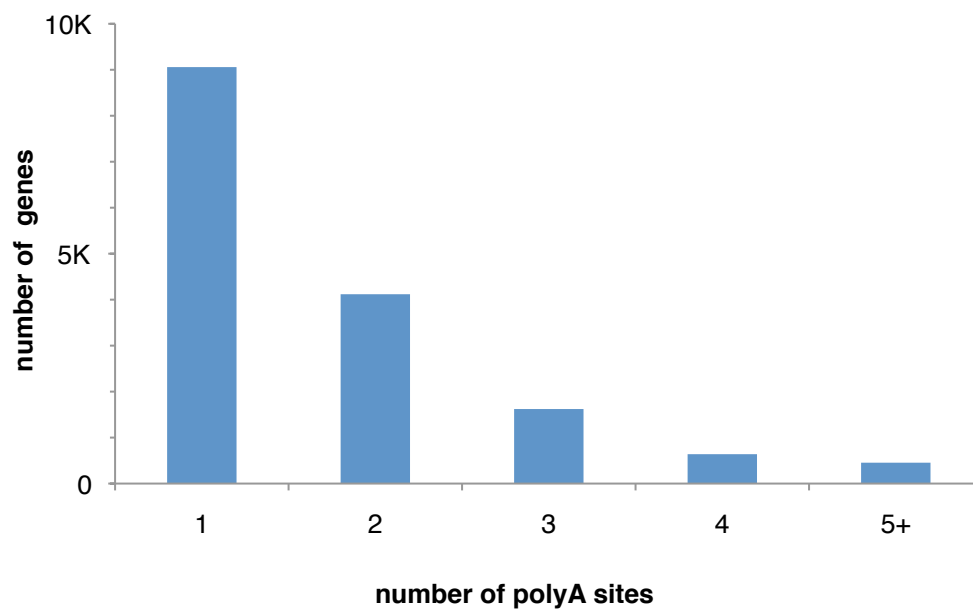


Figure S7. Number of polyA sites per gene.

The frequency distribution of distinct representative polyA sites per gene in the 3'UTRome. Around 40% of all genes with an annotated 3'UTR contain more than one alternative polyA site. Among genes with a large number of alternative 3'UTR isoforms are those encoding the small GTPase RAB-11.1 (6 isoforms), the LIN-61 paralog MBTR-1 (7 isoforms), and the RNA helicase VBH-1 (8 isoforms).

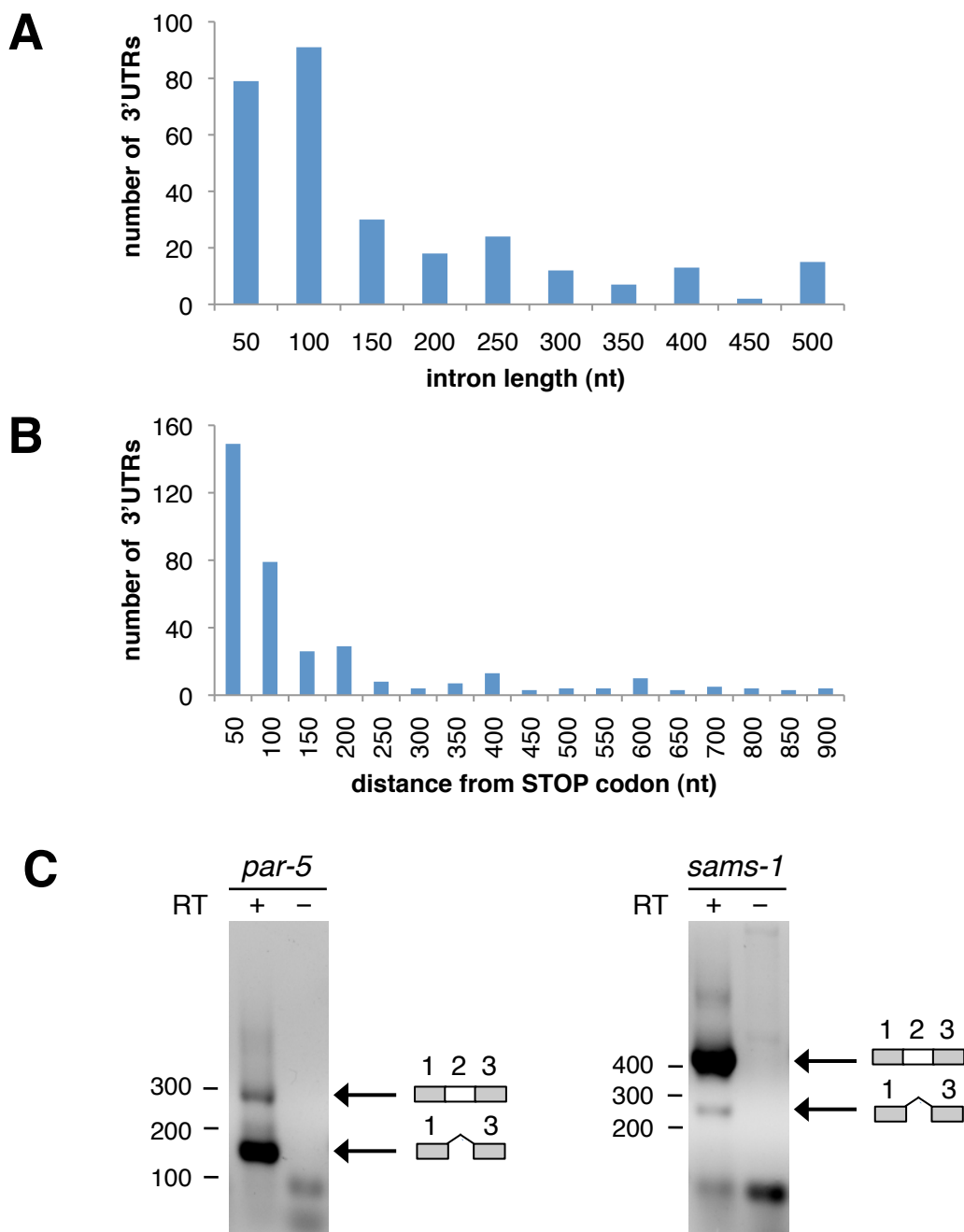


Figure S8. Introns in 3'UTR regions.

363 intron-containing 3'UTRs for 192 unique genes were used in this analysis. A) Length distribution (in nucleotides) of introns in 3'UTRs. B) Length distribution of the distance from the STOP codon to the intron start position. In both A and B, intron length is shown in 50 nt bins for simplification. C) Examples of facultative introns. Shown are 3'RACE products from *par-5* and *sams-1* 3'UTRs using mixed-stage total RNA and gene-specific primer pairs flanking the intron (regions 1 and 3), with (+) or without (-) inclusion of reverse transcriptase (RT) in the reaction. Agarose gel electrophoresis lanes with RT each produce two products consistent in size with the retention (top band) or excision (lower band) of region 2. Small bands below 100 nt represent unamplified primers and primer dimers (see Supplementary Online Materials and Methods for details). We observe that in some of these 3'UTRs, putative binding sites for miRNAs or ALG-1 (20) are contained within an intronic sequence.

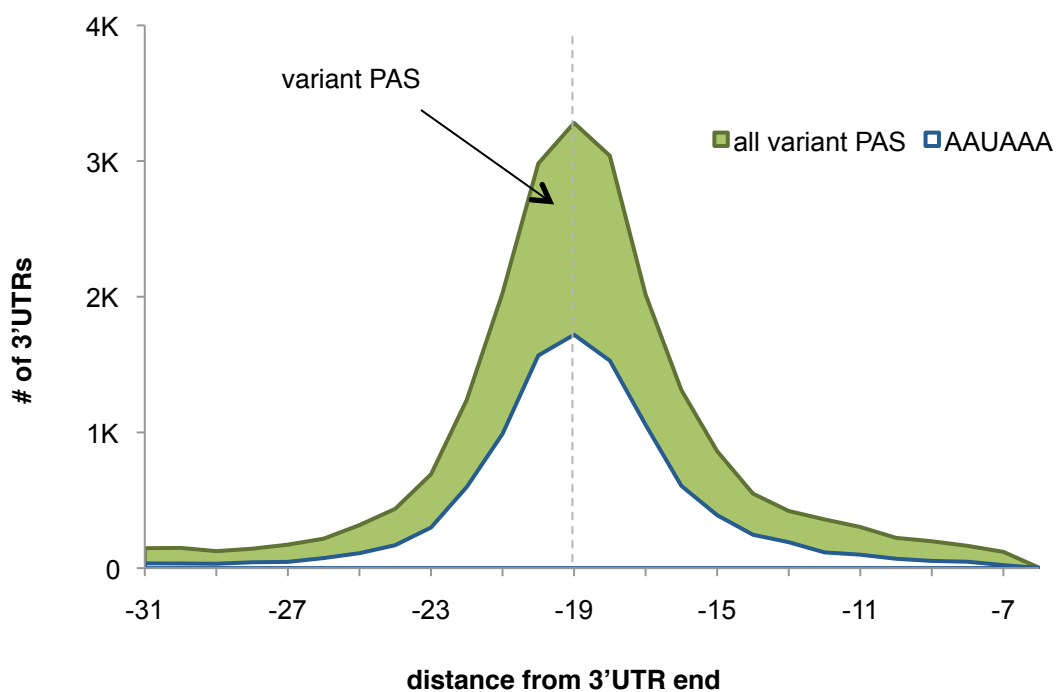


Figure S9. Distribution of the canonical AAUAAA and variant PAS elements relative to the cleavage and polyA addition site.

Start position for all PAS motifs (green line), AAUAAA (blue line), and variant PAS (green shading) peak at 19 nt upstream of the polyA addition site. See Supplementary Materials and Methods for details on the identification of PAS motifs and assignment of the most likely PAS motif for each 3'UTR.

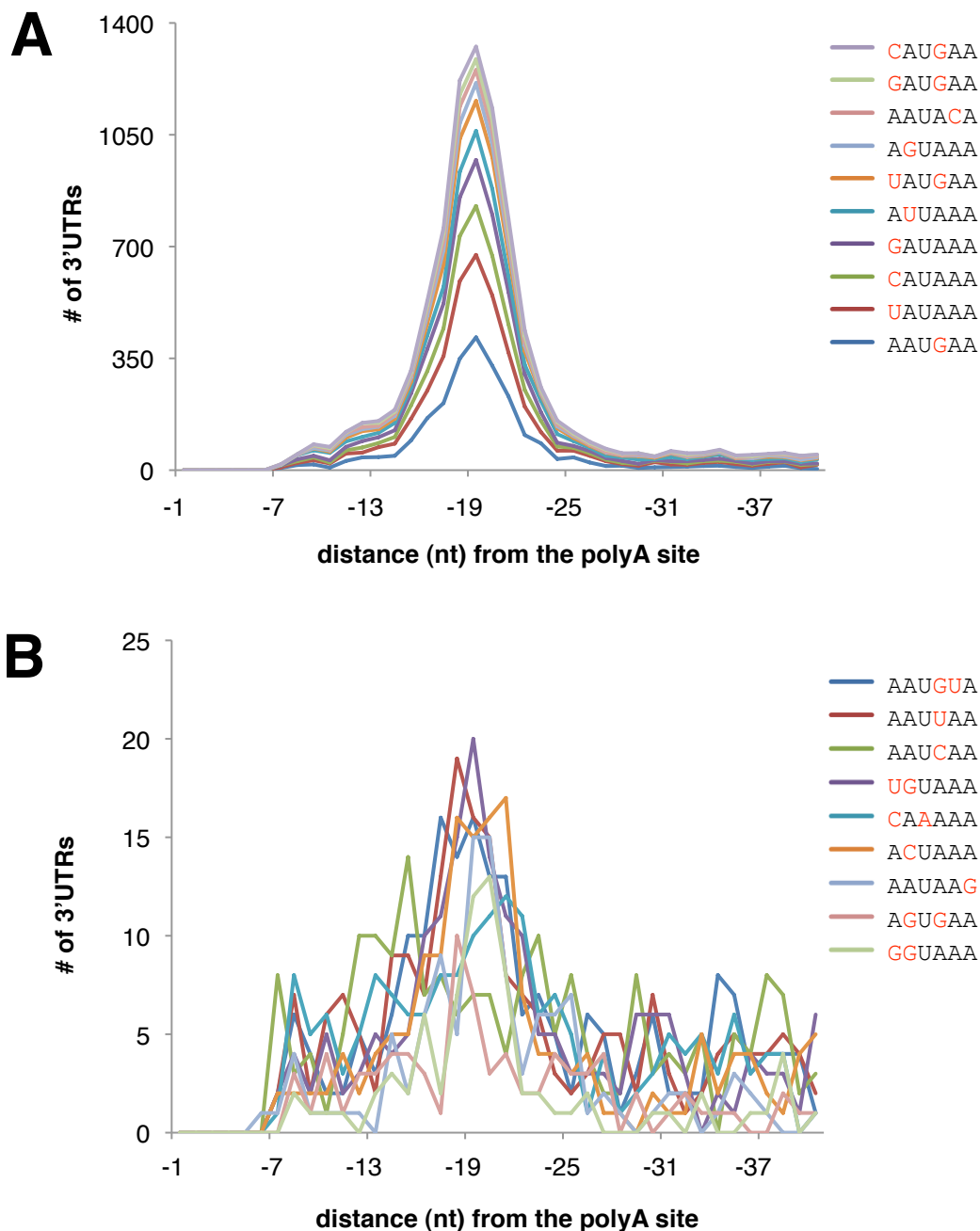


Figure S10. Distribution of variant PAS elements relative to the cleavage and polyA addition site.

In an unbiased search of all possible hexamers in the regions upstream of polyA sites in the 3'UTRome, the most common variant PAS hexamers show an enrichment that peaks at 19-20 nucleotides upstream of the polyA site. Using this as a guide, the most likely PAS motif for each polyA site was assigned using an ordered list of motifs according to the frequency of each motif in this region (see Supplementary Materials and Methods for details). The distribution of the most common motif, the canonical AAUAAA, which peaks at position -19, is not shown in this figure.

A) Ten of the most common variant PAS motifs (each assigned to $\geq 1\%$ of all polyA sites). The most common PAS variants contain a U in the third position and an A in the sixth position. B) Nine of the least common variant PAS motifs (each assigned to $\leq 1\%$ of all polyA sites). Total counts for each motif are given in Table S5.

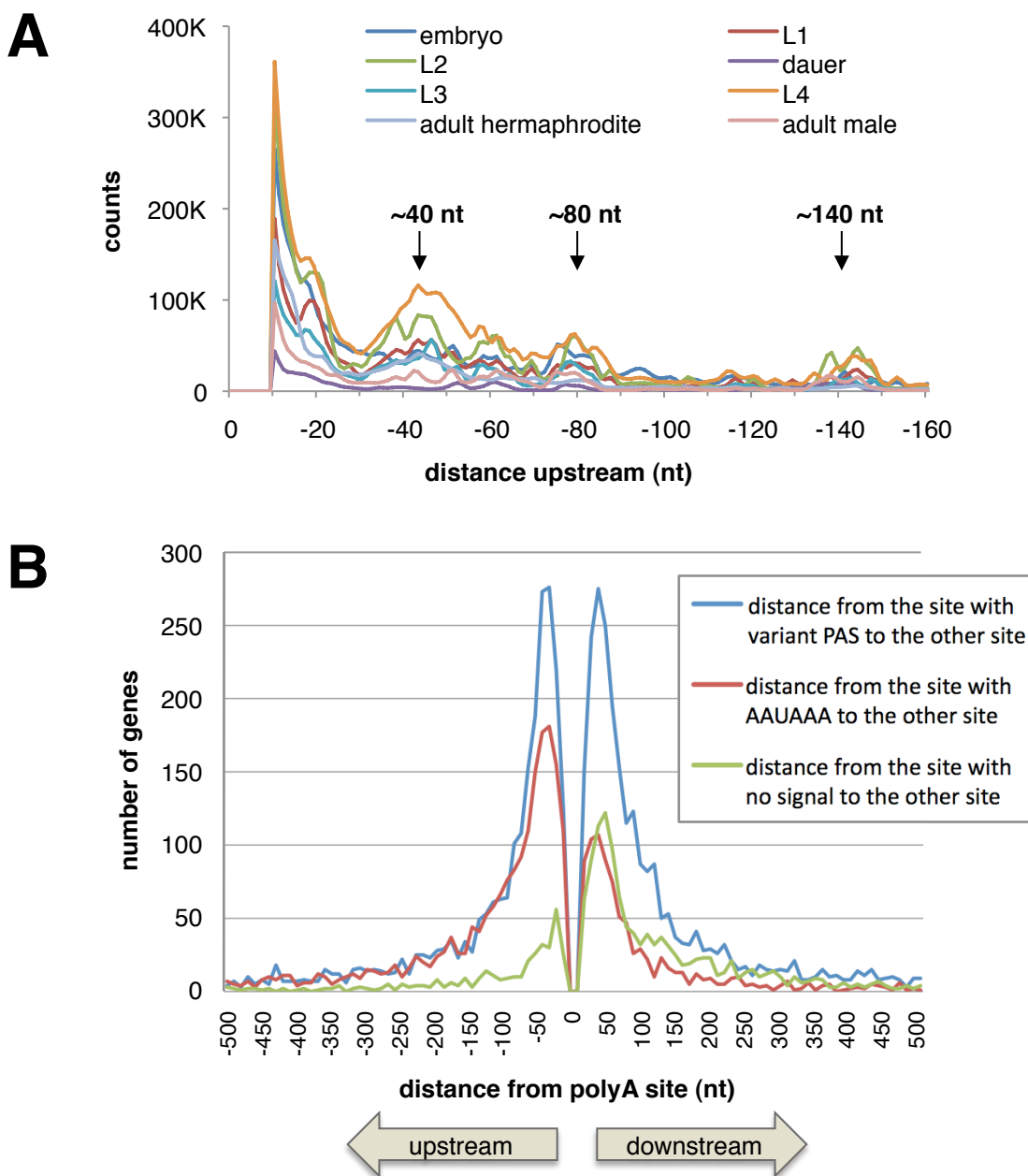


Figure S11. Relationships between alternative polyA addition sites for the same transcript.

A) The autocorrelation of polyA addition sites, pooled by stage, showing the average support count at each position relative to the most highly supported polyA site (aligned at 0 nt). The data show a main peak (arrow) ~40-45 bases upstream of the dominant polyA site. B) The distance between adjacent polyA sites peaks at ± 45 nt. PolyA addition sites with the canonical AAUAAA PAS motif (red) show a propensity to have a neighboring polyA site upstream; conversely, sites with no detectable PAS (green) tend to have a neighboring site downstream. Sites with a variant PAS (blue) are equally likely to have a neighboring site upstream or downstream.

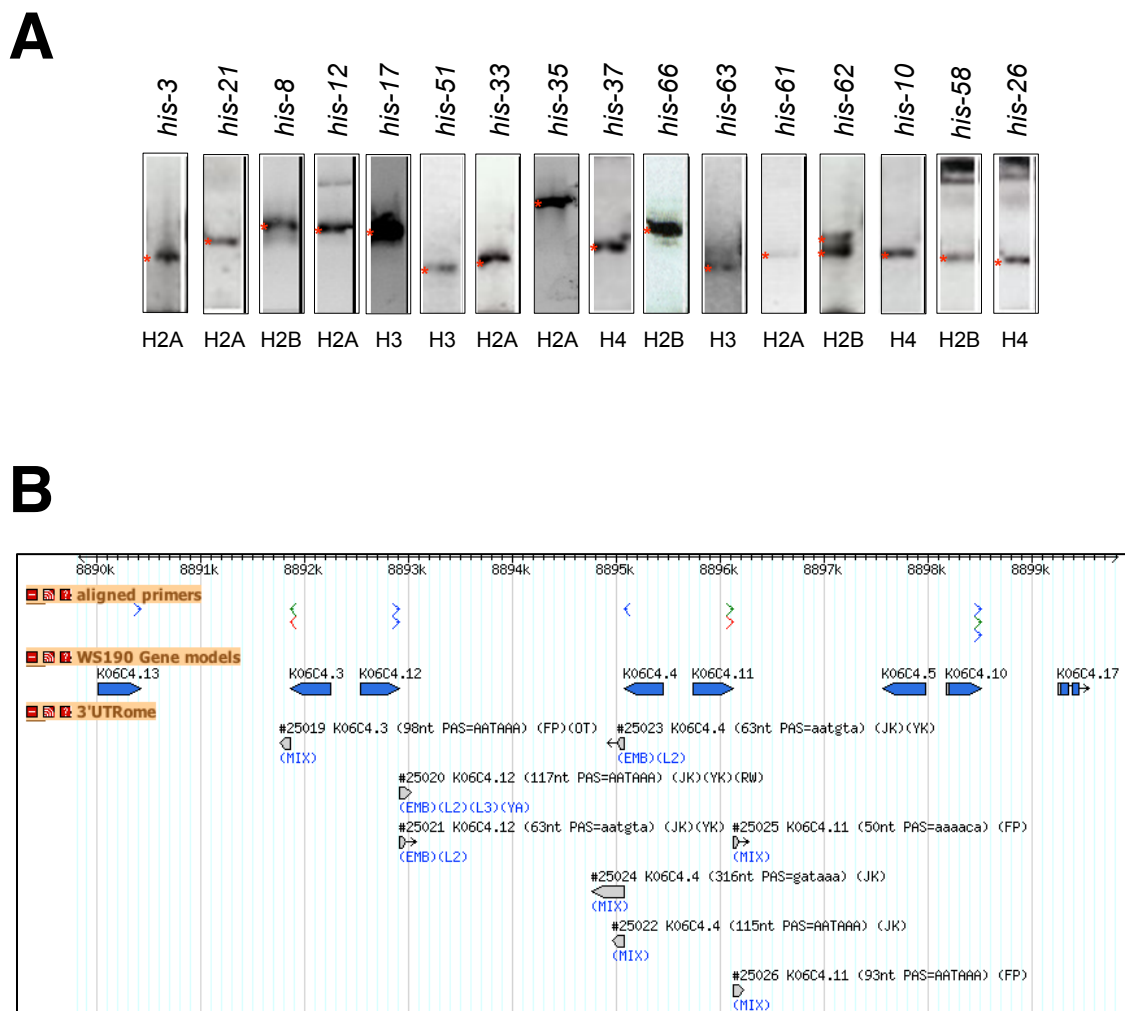


Figure S12. Polyadenylated 3'UTRs for histone genes.

A) The electrophoretic analysis on 2% agarose E-Gels of selected 3'RACE clones corresponding to 3'UTRs of histone genes obtained with the 3'RACE pipeline. PCR amplicons (red asterisks) correspond to unique or multiple 3'UTR isoforms. B) Histone gene cluster on chromosome V. Several histone genes with corresponding 3'UTRs detected in multiple developmental stages are shown. See Table S6 for the comprehensive list of histone 3'UTRs and PAS usage.

Combined with the observation that depletion of the SLBP homolog CDL-1 by RNAi severely depletes histone protein but not mRNA levels (21), our data lend support to the hypothesis that replication-dependent histone transcripts in *C. elegans* are first cleaved and polyadenylated using a PAS-directed mechanism, and are later post-processed to their final stem-loop form and regulated at the translational level by factors including CDL-1.

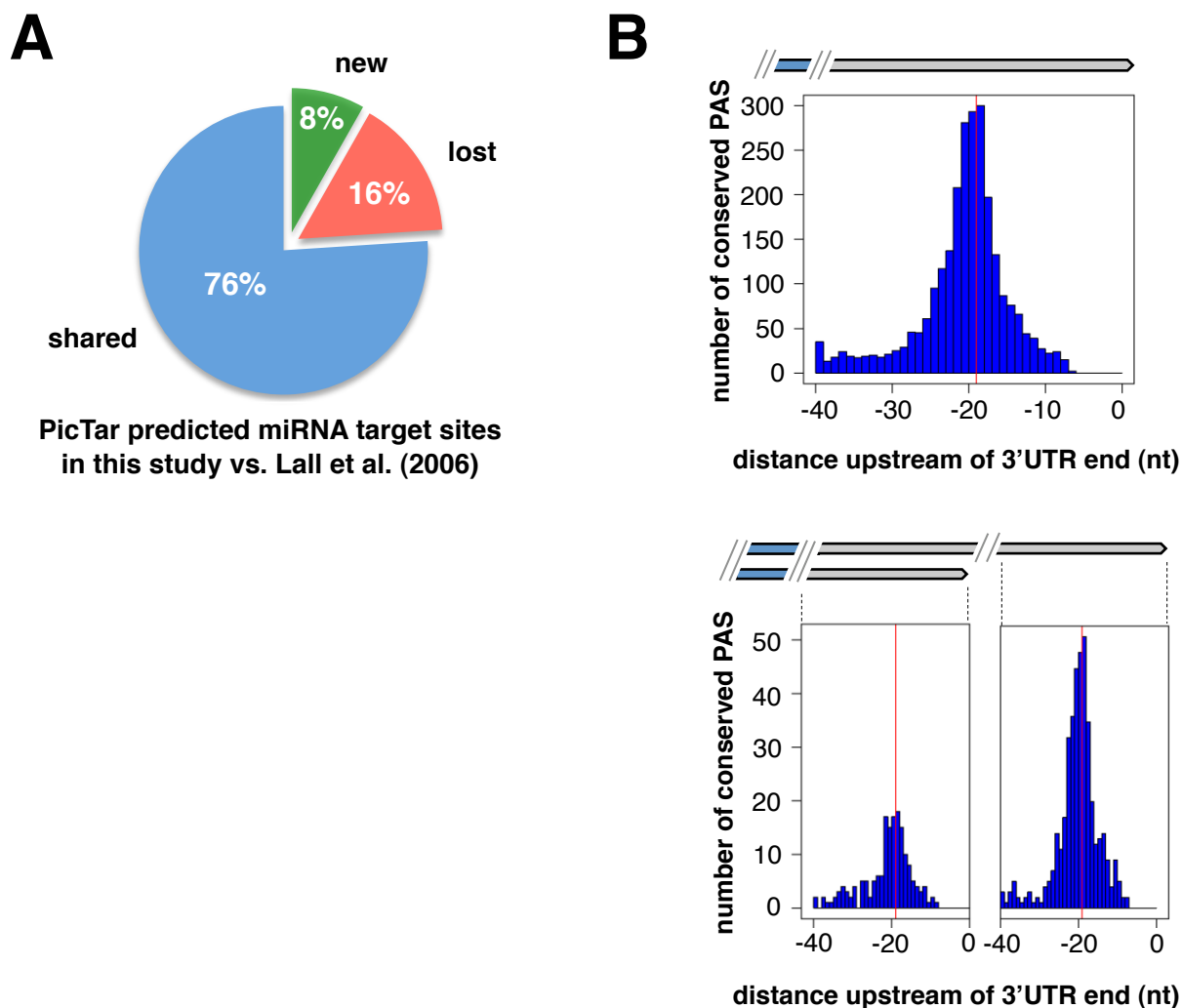


Figure S13. PicTar target predictions and PAS conservation in UTRome 3'UTRs.

A) Differences in PicTar predicted miRNA target sites within sequences spanned by the 3'UTRome, from this study in comparison with our previous predictions for *C. elegans* (19), as a percentage of the total number of predictions from both studies. See also Table S7. B) Distribution of conserved PAS motifs within 40 nt upstream of 3'UTR ends in three-way alignments between *C. elegans*, *C. briggsae*, and *C. remanei*, for (top) genes with one isoform (n=2,573 3'UTRs) or (bottom) exactly two isoforms (short, n=173; long, n=419). Red lines indicate the peak at -19 nt from the 3'UTR polyA addition site. See Supplementary Materials and Methods for additional details.

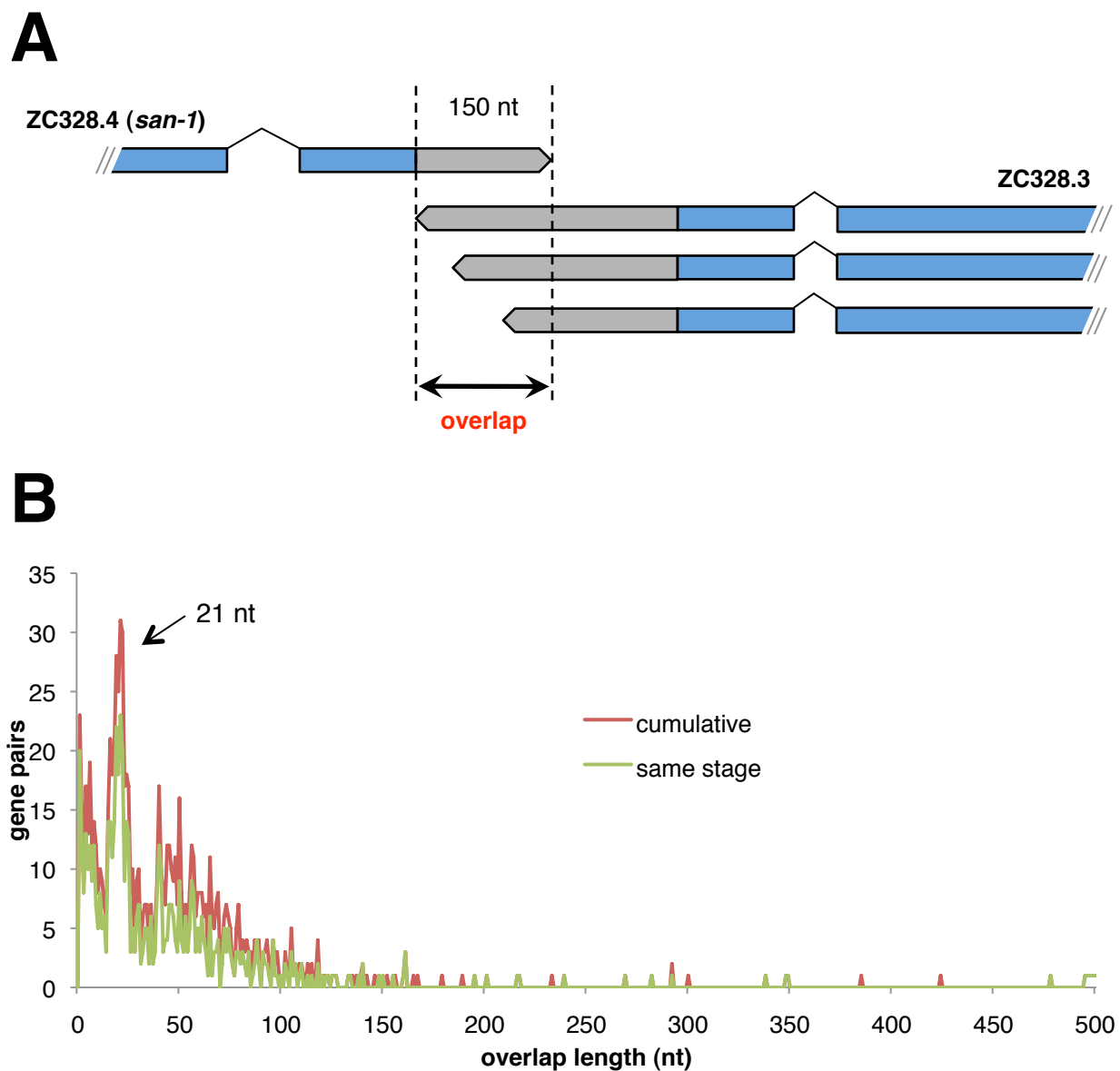


Figure S14. 3'UTRs on opposite strands sometimes overlap.

The 3'UTRome contains 1,876 convergently transcribed neighboring genes with overlapping regions that extend from the distal end of each putative transcript into the 3'UTR or CDS of the neighboring gene (see also Supplementary Dataset S4). For 1,240 of these genes, overlapping 3'UTR isoforms are co-expressed during at least one developmental stage. If both genes are transcribed simultaneously in the same cell, their 3'UTRs could potentially pair as dsRNA and trigger the production of endogenous siRNAs (endo-siRNAs) (22), which could down-regulate their mRNA levels.

A) Example of a 3'UTR overlap between the gene encoding mitotic spindle checkpoint protein ZC328.4 (*san-1*) and the uncharacterized gene ZC328.3. B) Length distribution (nt) of overlapping 3' end annotations for gene pairs on opposite strands, for cumulative overlapping pairs (red, n=938 pairs) or pairs detected in the same developmental stage (green, n=620 pairs). Overlapping pairs involve ~10% of genes in the 3'UTRome. Overlaps range from 1 to 495 nt, with an average overlap length of ~44 nt and median overlap length of ~28 nt. The peak in the overlap distribution at ~21 nt suggests that longer overlaps generally may be disfavored to limit recruitment of cellular machinery that could lead to endo-siRNA production (22).

Supplementary Tables

datasets	platform	total sequences	mapped sequences	developmental stage data	distinct polyA supported
PolyA capture	454	2,532,433	2,138,657	YES	165,538
RACE clones	Sanger	7,105	5,139	No	44,807
	454	166,112	86,577	No	
	Illumina	49,958,257	9,693,792	No	
cDNA	Sanger	—	119,434	YES	57,048
RNA-Seq	Illumina	291,573,831	84,771	YES	37,220

Table S1. Sequence data in the 3'UTRome.

Total number of raw and mapped sequences and the number of distinct polyA clusters supported for each data stream. Three of the datasets, polyA capture, cDNA and RNA-seq, provide developmental stage information allowing us to link distinct 3'UTR isoforms to specific developmental stages. See Figures S2-S5 for details on the different pipelines.

RUN 1	embryo	L1	L2	L3	L4	adult	male
total sequences	631,599	277,370	424,818	289,673	341,573	151,172	206,624
barcode detected	565,640	265,441	422,739	272,074	336,304	150,835	202,348
usable	yes	560,522	262,071	417,695	269,815	332,494	201,236
	no	5,118	3,370	5,044	2,259	3,810	1,112

RUN 2	daf-2	daf-7	daf-9	daf-11
total sequences	87,880	60,335	51,781	64,931
barcode detected	76,729	53,798	50,551	53,779
usable	yes	76,200	53,429	50,234
	no	529	369	317

Table S2. Summary of the polyA capture 454 sequencing runs.

Roche/454 reads produced by the polyA capture in individual developmental stages, males, and dauer mutants. The sequences obtained (total sequences) were scanned for the detection of a barcode (barcode detected). Reads containing a sequence contiguous with a polyA site were classified as 'usable'.

	PolyA capture	3'RACE	cDNA	RNA-seq
PolyA capture	11,606 (17,131)	–	–	–
3'RACE	3,879 (4,419)	5,929 (7,707)	–	–
cDNA	7,845 (9,808)	3,981 (4,475)	11,447 (16,986)	–
RNA-seq	5,445 (5,945)	2,732 (2,878)	6,040 (6,686)	7,442 (8,332)
total	15,683 (26,942)			
specific genes	1,858 (5,453)	632 (1,955)	1,358 (4,714)	314 (549)

Table S3. Gene and 3'UTR isoform coverage for individual datasets and overlaps between datasets in the 3'UTRome using AceView gene models.

Diagonal cells show the total number of coding genes and distinct polyA ends (in parentheses) for each of the four independent datasets; off-diagonal cells show intersections between each pair of datasets. The last row shows the total number of coding genes and distinct polyA ends that are specific to each individual dataset.

	PolyA capture	3'RACE	cDNA	RNA-seq
PolyA capture	11,007 (16,151)	–	–	–
3'RACE	3,878 (4,399)	5,919 (7,641)	–	–
cDNA	7,853 (9,724)	3,994 (4,469)	11,387 (16,710)	–
RNA-seq	5,394 (5,841)	2,743 (2,875)	6,070 (6,652)	7,322 (8,130)
total	14,986 (25,650)			
specific genes	1,382 (4,658)	635 (1,915)	1,300 (4,547)	253 (473)

Table S4. Subset of 3'UTRome matching WS190 gene models.

The subset of data from Table S3 that are compatible with WormBase WS190 gene models. See Table S3 legend for additional details.































name	3'UTRs	frequency (%)
AAUAAA	10,797	38.9 
no PAS	3,658	13.2 
AAUGAA	2,576	9.3 
UAUAAA	1,731	6.2 
CAUAAA	1,021	3.7 
GAUAAA	974	3.5 
UAUGAA	759	2.7 
AUUAAA	746	2.7 
AAAAAA	660	2.4 
UUUAAA	487	1.8 
AGUAAA	416	1.5 
AAUACA	387	1.4 
AAUAUA	353	1.3 
GAUGAA	313	1.1 
AAUAAU	311	1.1 
CAUGAA	310	1.1 
AAAUAA	307	1.1 
UGUAAA	302	1.1 
UCUAAA	231	0.8 
AAUGUA	229	0.8 
AAUUAA	176	0.6 
ACUAAA	174	0.6 
AAGAAA	168	0.6 
CAAAAA	167	0.6 
GAAAAA	146	0.5 
AACAAA	115	0.4 
AAUAAG	93	0.3 
GGUAAA	92	0.3 
AGUGAA	55	0.2 
AAACAA	35	0.1 

Table S5. Identification of putative PAS elements.

An unbiased search for over-represented hexamers in the last 50 nt of 3'UTRs in the 3'UTRome identified a handful of sequences whose start positions all peaked at around 19 nt upstream of the polyA cleavage site. Using these results as a guide, we searched all 3'UTRs recursively for the most likely PAS site utilized by each 3'UTR (see Supplementary Materials and Methods for details). The most common motif, the “canonical” PAS element AAUAAA, is observed in 39% of 3'UTRs; the other elements consist of variations of this motif differing by one or two nucleotides. This apparent diversity of PAS motifs suggests that the recognition of PAS sites in worms is more flexible than higher eukaryotes, where mutation in any position of the canonical AAUAAA element disrupts the 3' end processing of mRNAs (23), and may perhaps be more akin to the 3' end processing mechanism of yeast, where presence of an AU rich region is sufficient to allow docking of the processing machinery (24).

name	CDS	isoforms	3'UTR length (PAS)
<i>his-2</i>	T10C6.13	3	48 (no signal), 127 (AAUAAA), 365 (no signal)
<i>his-3</i>	T10C6.12	1	97 (AAUAAA)
<i>his-4</i>	T10C6.11	2	14 (no signal), 112 (AAUAAA)
<i>his-6</i>	F45F2.13	1	128 (AAUAAA)
<i>his-8</i>	F45F2.12	1	66 (no signal)
<i>his-9</i>	ZK131.3	2	120 (AAUAAA), 180 (AAUAAA)
<i>his-10</i>	ZK131.4	1	114 (AAUAAA)
<i>his-11</i>	ZK131.5	1	108 (GAUAAA)
<i>his-12</i>	ZK131.6	1	97 (AAUAAA)
<i>his-13</i>	ZK131.7	1	120 (AAUAAA)
<i>his-14</i>	ZK131.8	1	114 (AAUAAA)
<i>his-15</i>	ZK131.9	1	111 (GAUAAA)
<i>his-16</i>	ZK131.10	2	58 (no signal), 119 (AAUAAA)
<i>his-19</i>	K06C4.11	2	50 (AAACA), 93 (AAUAAA)
<i>his-20</i>	K06C4.4	3	63 (AAUGUA), 115(AAUAAA), 316 (GAUAAA)
<i>his-21</i>	K06C4.3	1	98 (AAUAAA)
<i>his-22</i>	K06C4.12	2	63 (AAUGUA), 117 (AAUAAA)
<i>his-24</i>	M163.3	1	236 (AAUAAA)
<i>his-25</i>	ZK131.2	3	98 (UGUAAA), 124 (AAUAAA), 151 (GAUAAA)
<i>his-26</i>	ZK131.1	1	114 (AAUAAA)
<i>his-27</i>	K06C4.13	1	222(AAUAAA)
<i>his-28</i>	K06C4.2	2	108 (AAUAAA), 219 (GAUGAA)
<i>his-32</i>	F17E9.10	2	115 (AAUAAA), 146 (AAUAAA)
<i>his-34</i>	F17E9.9	1	59 (AAUAAA)
<i>his-35</i>	C50F4.13	1	116 (AAUAAA)
<i>his-36</i>	C50F4.6	3	92 (AAUAAA), 100 (AAUAAA), 622 (AAUAAA)
<i>his-37</i>	C50F4.7	1	88 (AAUAAA)
<i>his-40</i>	NULL	1	128 (AAUAAA)
<i>his-41</i>	C50F4.5	3	92 (AAUAAA), 100 (AAUAAA), 622 (AAUAAA)
<i>his-42</i>	F08G2.3	1	276(AAUAAA)
<i>his-43</i>	F08G2.2	1	97 (AAUAAA)
<i>his-44</i>	F08G2.1	1	111 (GAUAAA)
<i>his-45</i>	B0035.10	1	116 (AAUGAA)
<i>his-46</i>	B0035.9	4	29 (no signal), 67 (no signal), 114 (AAUAAA), 155 (AAUCA)
<i>his-47</i>	B0035.7	2	115 (AAUAAA),171 (UAUAAA)
<i>his-48</i>	B0035.8	2	103 (AAUAAA), 115 (AAUAAA)
<i>his-49</i>	F07B7.5	1	120 (AAUAAA)
<i>his-50</i>	F07B7.9	2	108 (AAUAAA), 219 (GAUGAA)
<i>his-51</i>	F07B7.10	1	93 (AAUAAA)
<i>his-52</i>	F07B7.4	2	63 (AAUGUA), 117 (AAUAAA)
<i>his-53</i>	F07B7.3	2	50 (AAACA), 93 (AAUAAA)
<i>his-54</i>	F07B7.11	3	63 (AAUGUA), 115 (AAUAAA), 316 (GAUAAA)
<i>his-56</i>	F54E12.3	3	29 (no signal), 67 (no signal), 114 (AAUAAA)
<i>his-57</i>	F54E12.5	1	104 (AAUAAA)
<i>his-58</i>	F54E12.4	1	103 (AAUGAA)
<i>his-59</i>	F55G1.2	1	295 (AAUGAA)

name	CDS	isoforms	3'UTR length (PAS)
<i>his-60</i>	F55G1.11	2	66 (no signal), 120 (AAUAAA)
<i>his-61</i>	F55G1.10	1	98 (AAUAAA)
<i>his-62</i>	F55G1.3	2	31 (no signal), 107 (AAUAAA)
<i>his-63</i>	F22B3.2	1	116 (AAUAAA)
<i>his-66</i>	H02I12.6	1	107 (AAUAAA)
<i>his-68</i>	T23D8.6	2	15 (AAUAAA), 100 (AAUAAA)
<i>his-69</i>	E03A3.3	1	90 (GAUAAA)
<i>his-70</i>	E03A3.4	1	106 (AAUAAA)
<i>his-71</i>	F45E1.6	1	163 (AAUAAA)
<i>his-72</i>	Y49E10.6	2	104 (AAUAAA), 213 (AAUAAA)
<i>his-74</i>	W05B10.1	1	162 (AAUAAA)

Table S6: Cumulative list of polyadenylated 3'UTRs detected in histone genes.

Summary of 3'UTR isoforms detected in histone genes, showing the putative PAS element for each representative 3'UTR. Nucleotides that deviate from the canonical PAS motif are highlighted in red.

A	# of 3'UTR isoforms	26,942	
B	# of unique 3'UTR regions	15,685	
C	average 3'UTR length	250 nt	
D	total 3'UTRome length	3,898,952 nt	
E	per nucleotide conservation rate of 3'UTR (3 species)	0.3	
F	probability of a conserved seed being functional	3 species	5 species
		0.56 ±0.01	0.64 ±0.03
G	# of unique conserved seeds identified	3 species	5 species
		5,673	1,744
H	# of unique miRNAs used for analysis (# of families)	183 (124)	
I	probability of a conserved miRNA seed site occurring inside an ALG-1 site	3 species	5 species
		0.75	0.76
J	probability of a randomly positioned 6-mer in a 3'UTR occurring inside an ALG-1 site	3 species	5 species
		0.43	0.45
K	# of conserved blocks not explained by predicted miRNA seeds or conserved PAS (5 species)	4,758	
L	# of 3'UTRs with at least one conserved block (5 species)	2,887	
M	probability of a conserved (randomly shuffled) sequence block of the same length inside an ALG-1 site	0.54 (0.48)	
M	fraction of Lall et al. 3'UTR:miRNA interactions recovered	0.83	
O	# of Lall et al. 3'UTR:miRNA interactions lost	1,111	
P	# of unique new interactions vs. Lall et al. miRNAs/3'UTRs	580	

Table S7. Summary statistics for PicTar miRNA target predictions and other conserved sequence blocks in genomic regions spanned by the 3'UTRome compendium.

A) Total number of 3'UTRs used for miRNA target predictions. **B)** Number of unique 3'UTR regions, obtained by merging 3'UTRs with overlapping genomic coordinates. **C)** Average length of all unique 3'UTRs. **D)** The unique 3'UTRome comprises ~4M nucleotides. **E)** 30% percent of nucleotides in *C. elegans* 3'UTR are conserved in *C. remanei* and *C. briggsae*. Nucleotides in CDS, 5'UTR or intergenic regions were not considered in this analysis. **F)** Probability of a conserved miRNA seed being functional based on alignments of three or five species, obtained by creating artificial miRNAs resembling the original miRNAs (18) and comparing the number of target sites for the artificial miRNAs with the “real” target sites. **G)** Number of unique conserved miRNA seeds in the genome of three or five species. **H)** In total, 183 miRNAs were used. They comprise 174 miRBase (database release 14) miRNAs and 9 novel miRNAs determined by miRDeep2 (17), grouped in 124 miRNA families. **I)** The probability of a conserved miRNA seed within an ALG-1 binding site (20) in three or five species, calculated as the ratio of all miRNA target sites located in an ALG-1 binding site when considering only 3'UTRs that have an ALG-1 site and at least one miRNA target site. **J)** Probability of a shuffled seed sites (randomly positioned with the same 3'UTR) occurring within an ALG-1 binding site for three or five species. The probability is 30% less for shuffled sites than for the original miRNA seed position, signifying that miRNA seeds located in ALG-1 sites are indeed accurate signals. **K)** Number of conserved blocks, defined as at least 6 nt long and present in five species, that cannot be explained by a conserved predicted miRNA target seed site or a conserved PAS. **L)** Number of 3'UTR regions that contain at least one of such conserved blocks. **M)** Probability of a conserved block occurring within an ALG-1 binding site vs. randomly positioned blocks of the same length distribution within 3'UTRs is not significantly different. For analyses in K-M, regions overlapping a CDS in an alternative transcript were excluded. **N,O,P)** For the same miRNAs and 3'UTR regions, 83% of previously predicted miRNA target sites from Lall et al. (19) are identical with predictions using the empirically defined 3'UTRs in the 3'UTRome; 1,111 miRNA target sites are exclusively found in Lall et al., and 580 sites are newly predicted. Three species alignments always included *C. elegans*, *C. remanei*, and *C. briggsae*. Five species alignments also included *C. brenneri* and *C. japonica*. See Supplementary Materials and Methods for additional details.

stage	genes	isoforms
embryo	966	1,320
L1	325	353
L2	252	268
dauer	264	304
L3	131	134
L4	150	157
adult	84	88
male	374	447
total	2,049	3,071

Table S8. Number of genes present in multiple developmental stages but with stage-specific 3'UTR isoforms.

We have scanned the 3'UTRome for genes expressed in 1) at least two developmental stages, 2) with at least two 3'UTR isoforms, and 3) where one of these isoforms was stage-specific. The results shown here were used for the analysis described in Figure 4B.

stage	long 3'UTR more abundant	short 3'UTR more abundant
embryo	315	169
L1	80	58
L2	33	37
dauer	184	104
L3	59	34
L4	27	39
adult	80	45
male	94	97
total	915	610
total genes	1,960	

Table S9. Number of genes with two 3'UTR isoforms detected in the staged polyA capture dataset.

A subset of annotated genes from the polyA capture dataset with two 3'UTR isoforms used for the analyses in Figure 4. A 3'UTR isoform is defined as abundant if: 1) the total number of counts across all stages is larger than 5, and 2) if it is supported by at least twice the number of counts than the other 3'UTR isoform (see Supplementary Materials and Methods for details).

	minipools	deconvolved library
96-well plates	75	39
unique genes	7,105	3,750
unique isoforms	—	5,774

Table S10. 3'UTR clones available in the 3'UTRome library.

The 3'RACE approach produced sequence-validated 3'UTR clones that are available to the community to study 3'UTR biology. The UTR library collection will be updated on an ongoing basis and will expand to contain minipools and unique 3'UTR isoforms for all *C. elegans* 3'UTRs for protein-coding transcripts. See Supplementary Dataset S7 and the 3'UTR data repository <http://www.utrome.org> for clone availability.

Supplementary Datasets

Supplementary Dataset S1. AceView genes in the 3'UTRome.

Comprehensive list of AceView genes with annotated 3'UTRs in the 3'UTRome. All gene names are linked to the current AceView annotation at NCBI (<http://www.aceview.org>). The file can be downloaded in HTML format.

Supplementary Dataset S2. The complete 3'UTRome dataset.

A key is enclosed with Dataset S2 that describes all of the individual components.

Supplementary Dataset S3. 3'UTR coordinates attached to AceView genes.

We used AceView gene annotations (<http://www.aceview.org>) (7) to map 1,490 unique, fully supported 3'UTR isoforms in genomic regions with either no annotated gene models or no compatible CDS ends in WormBase WS190. This table contains genome coordinates of 3'UTRs for these new genes. The file can be downloaded in Microsoft® Excel format.

Supplementary Dataset S4. Convergently transcribed genes with overlapping 3'UTRs.

A list of genes in the 3'UTRome whose transcripts overlap (1 nt to 495 nt), indicating gene names, overlap length (nt), genome coordinates, and whether the two overlapping 3'UTRs are co-expressed in the same developmental stage. These data were used for the analysis described in Figure S14. The file can be downloaded in Microsoft® Excel format.

Supplementary Dataset S5. List of genes displaying changes in 3'UTR length between developmental stages.

A comprehensive list of genes with two 3'UTR isoforms showing a change in the expression of long vs. short 3'UTR isoforms between developmental stages. All data are derived from the polyA capture dataset and are based on the number of Roche/454 read counts identified per 3'UTR end. The file contains two worksheets: The worksheet labeled "All genes—counts" lists the raw tag counts and counts normalized to the total counts in the embryo dataset. The worksheet labeled "All genes—relative abundance" shows the number of reads normalized within and across all developmental stages. Genes that exhibit 3'UTR isoform switching across developmental stages (shown individually in Supplementary Dataset S6) are indicated in the last two columns, labeled "potential isoform switch" and " 'high-confidence' isoform switch" (defined as a difference of ≥ 2 -fold). See Supplementary Materials and Methods for details. The file can be downloaded in Microsoft® Excel format.

Supplementary Dataset S6. Individual graphs of genes displaying 3'UTR isoform switching during development.

Individual graphs for 612 genes with two 3'UTR isoforms that exhibit a detectable switch in the expression of the long vs. short isoform across developmental stages, and with at least 20 total Roche/454 polyA tag counts per gene. All data are derived from the polyA capture dataset and are based on the number of Roche/454 read counts identified per 3'UTR end. For each graph the gene name, chromosome location, strand (in parentheses), genomic coordinate of the 3'UTR start, and lengths of the two 3'UTR isoforms are indicated. Green boxes highlight genes for which the relative abundance of 3'UTR isoform 'a' vs. 'b' is ≥ 2 -fold in at least one particular stage and then "switches" so that the ratio of 'b' vs. 'a' is ≥ 2 -fold in another stage; in addition, the difference in expression between isoform 'a' and 'b' was required to be ≥ 5 counts. The cumulative list is given in Supplementary Dataset S5. See Supplementary Materials and Methods for details of the analysis. This file can be downloaded in Adobe® PDF format.

Supplementary Dataset S7. The 3'UTRome clone library.

List of 3'UTR clones released. The clones are available to the community in the form of bacterial minipools and isolated 3'UTR isoforms. The library is cloned into the Gateway™ entry vector P2R-P3 and is compatible with the Promoterome (3) and ORFeome (4,5) libraries. The file can be downloaded in Microsoft® Excel format.

References for Supplementary Online Materials

1. T. Stiernagle, *WormBook*, 1 (2006).
2. J. Hodgkin, H. R. Horvitz, S. Brenner, *Genetics* **91**, 67 (1979).
3. D. Dupuy *et al.*, *Genome Res* **14**, 2169 (2004).
4. P. Vaglio *et al.*, *Nucleic Acids Res* **31**, 237 (2003).
5. J. Reboul *et al.*, *Nat Genet* **34**, 35 (2003).
6. Y. Suzuki, K. Yoshitomo-Nakagawa, K. Maruyama, A. Suyama, S. Sugano, *Gene* **200**, 149 (1997).
7. D. Thierry-Mieg, J. Thierry-Mieg, *Genome Biol* **7 Suppl 1**, S12 1 (2006).
8. L. W. Hillier *et al.*, *Genome Res* **19**, 657 (2009).
9. H. Shin *et al.*, *BMC Biol* **6**, 30 (2008).
10. W. J. Kent, *Genome Res* **12**, 656 (2002).
11. H. Li *et al.*, *Bioinformatics* **25**, 2078 (2009).
12. R. Lopez, V. Silventoinen, S. Robinson, A. Kibria, W. Gish, *Nucleic Acids Res* **31**, 3795 (2003).
13. K. G. Murthy, J. L. Manley, *J Biol Chem* **267**, 14804 (1992).
14. D. Blankenberg *et al.*, *Curr Protoc Mol Biol* **Chapter 19**, Unit 19 10 1 (2010).
15. D. Karolchik *et al.*, *Nucleic Acids Res* **32**, D493 (2004).
16. S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, A. J. Enright, *Nucleic Acids Res* **34**, D140 (2006).
17. M. R. Friedlander *et al.*, *Nat Biotechnol* **26**, 407 (2008).
18. A. Krek *et al.*, *Nat Genet* **37**, 495 (2005).

19. S. Lall *et al.*, *Curr Biol* **16**, 460 (2006).
20. D. G. Zisoulis *et al.*, *Nat Struct Mol Biol* **17**, 173 (2010).
21. R. Keall, S. Whitelaw, J. Pettitt, B. Muller, *BMC Mol Biol* **8**, 51 (2007).
22. K. Okamura, S. Balla, R. Martin, N. Liu, E. C. Lai, *Nat Struct Mol Biol* **15**, 581 (2008).
23. M. D. Sheets, S. C. Ogg, M. P. Wickens, *Nucleic Acids Res* **18**, 5799 (1990).
24. S. Henikoff, J. D. Kelly, E. H. Cohen, *Cell* **33**, 607 (1983).