The dsRBP and inactive editor, ADR-1, utilizes dsRNA binding to regulate A-to-I RNA editing across the *C. elegans* transcriptome

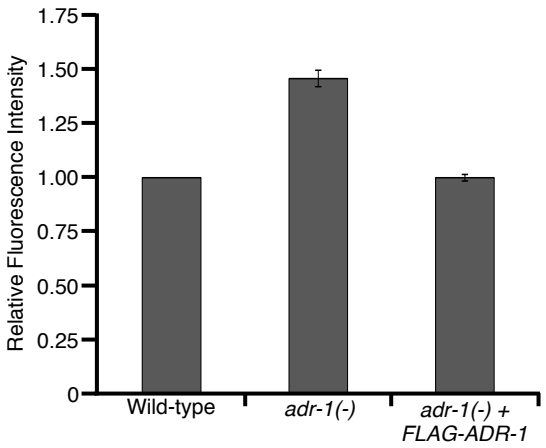
Michael C. Washburn[1,8], Boyko Kakaradov[2,3,8], Balaji Sundararaman[3], Emily Wheeler[4], Shawn Hoon[5,6], Gene W. Yeo[2,3,5,7,**], Heather A. Hundley[4*]

Figure S1

A

Editing Site	% Editing RT#1	% Editing RT#2	Std. Dev.
528	20.1	20.7	0.42
537	18.3	18.3	0.00
550	16.5	15.6	0.63
589	94.8	95.1	0.21
591	14.7	15.3	0.42
592	16.1	16.5	0.28
605	19.8	19.4	0.28
606	9.5	8.9	0.42
623	13.2	13.8	0.42
631	25.9	25.7	0.14
Average Std Dev.			0.32

B



C

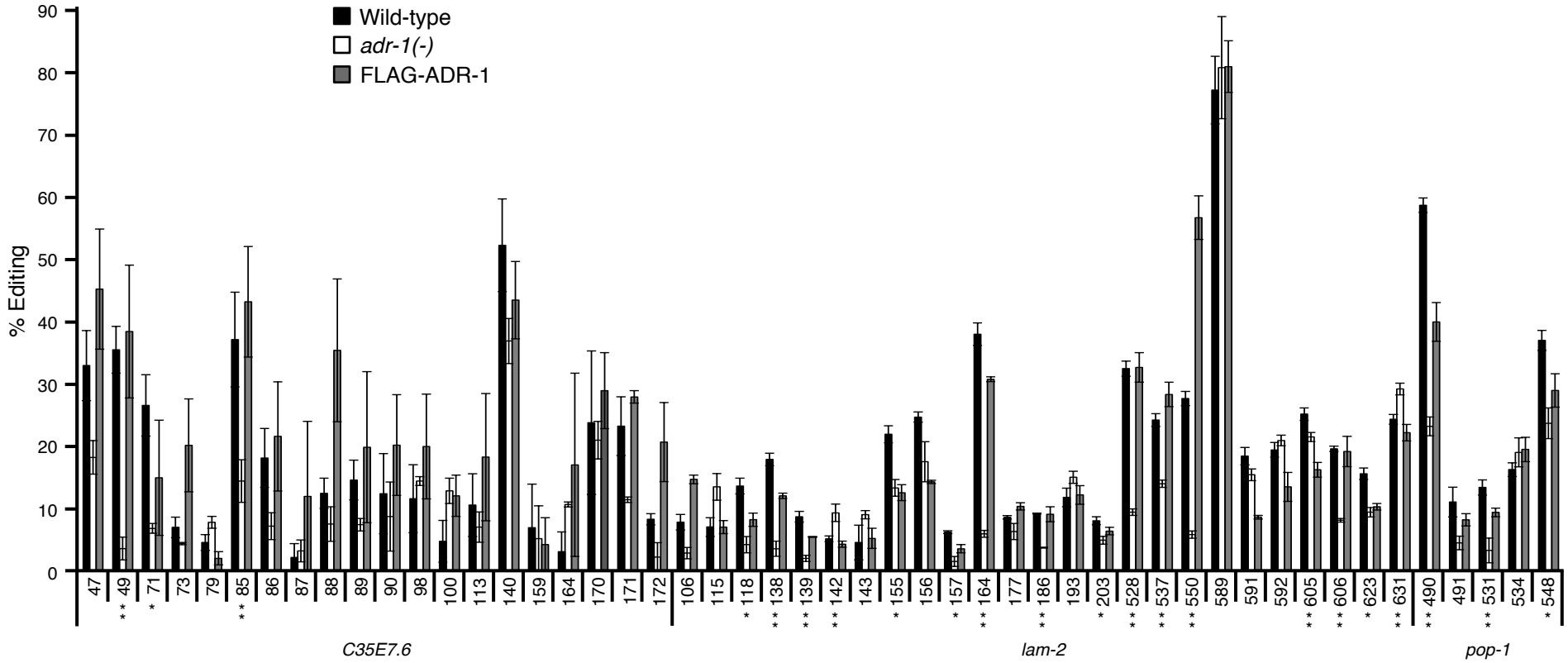
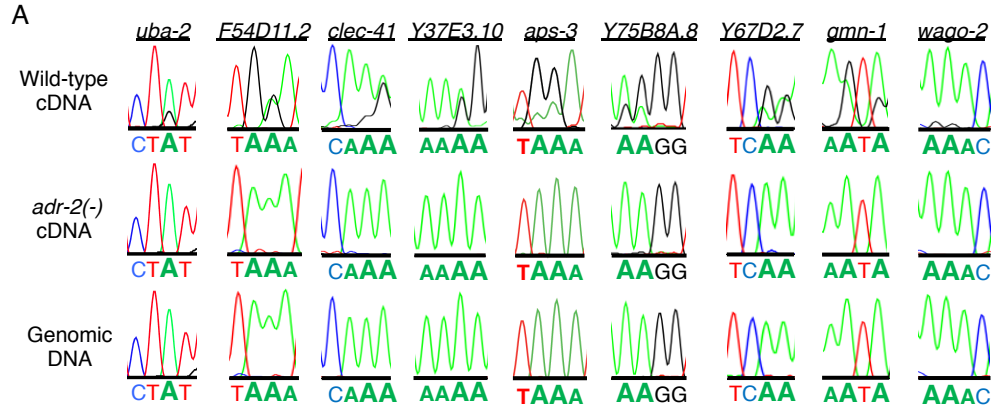
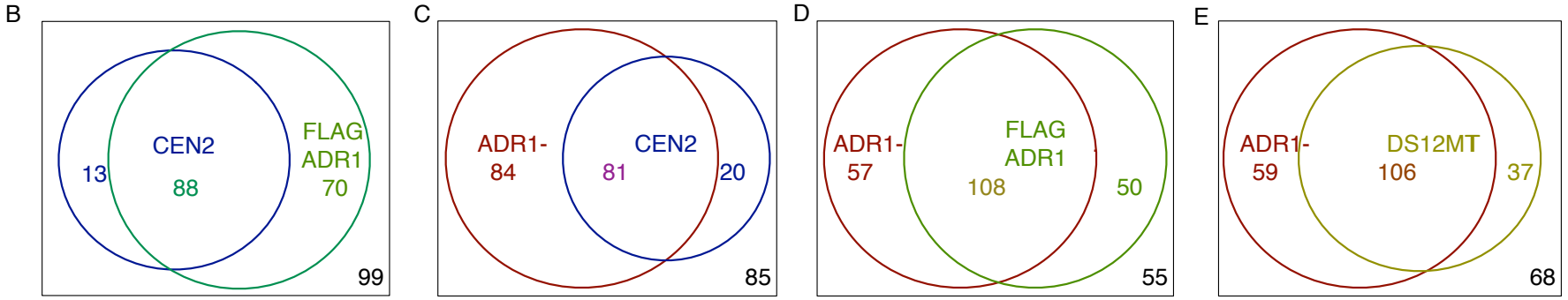


Figure S2


A
uba-2
F54D11.2
clec-41
Y37E3.10
aps-3
Y75B8A.8
Y67D2.7
gmn-1
wago-2
Wild-type cDNA
adr-2(-) cDNA
Genomic DNA
CTAT
TAAA
CAAA
AAAA
TAAA
AAGG
TCAA
AATA
AAAC


B
CEN2
FLAG ADR1
13
88
70
99
C
ADR1-
CEN2
84
81
20
85
D
ADR1-
FLAG ADR1
57
108
50
55
E
ADR1-
DS12MT
59
106
37
68

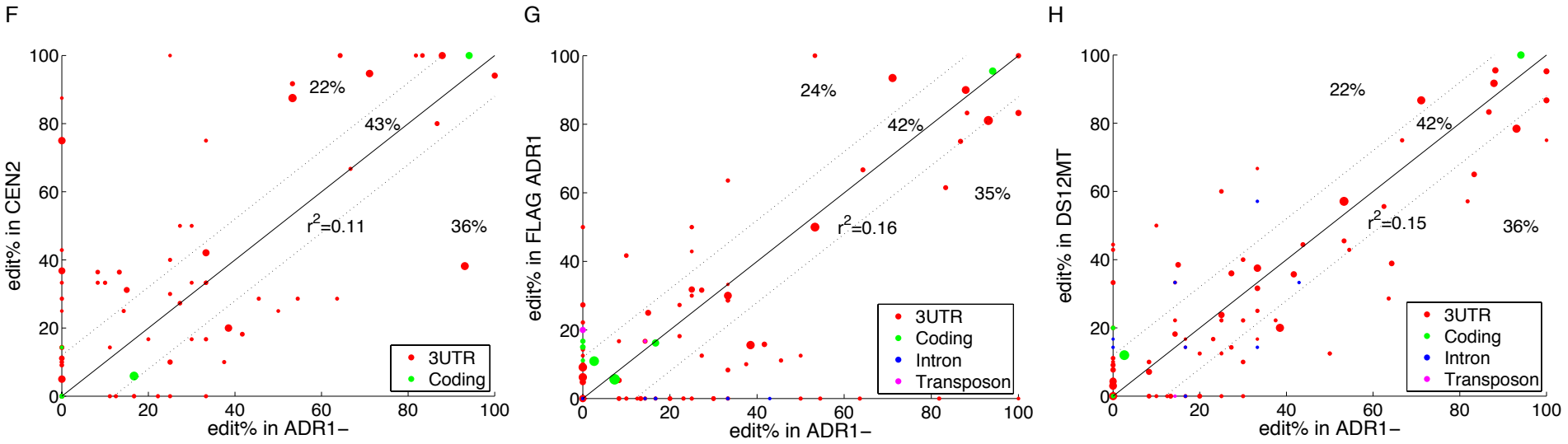
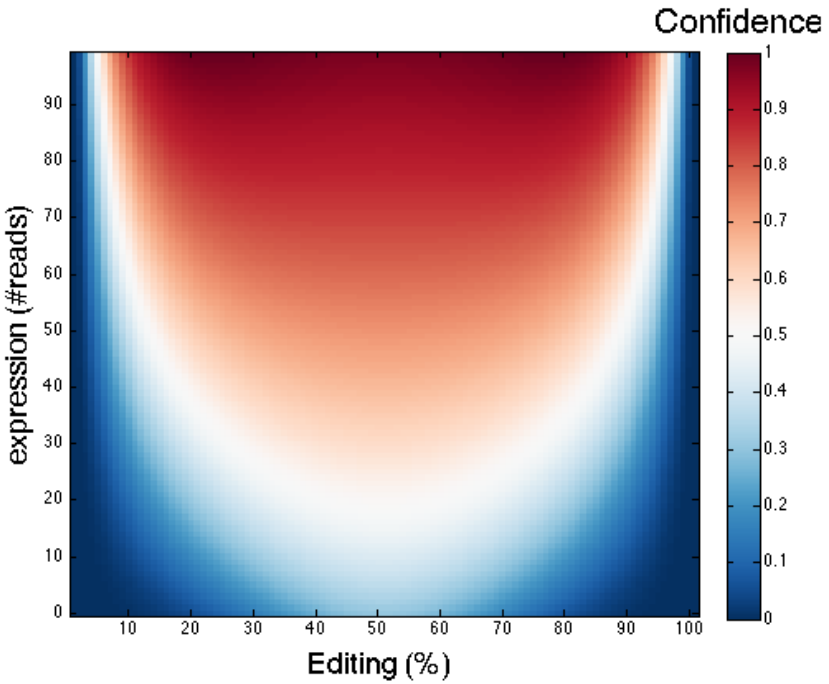
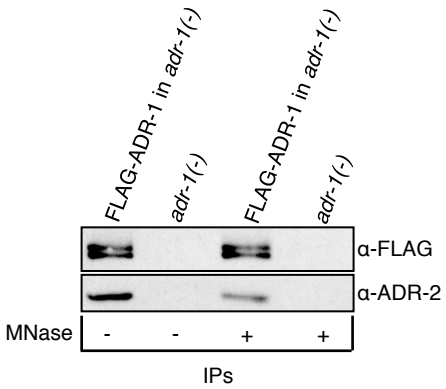

F
edit% in CEN2
edit% in ADR1−
22%
43%
36%
r2=0.11
3UTR
Coding
G
edit% in FLAG ADR1
edit% in ADR1−
24%
42%
35%
r2=0.16
3UTR
Coding
Intron
Transposon
H
edit% in DS12MT
edit% in ADR1−
22%
42%
36%
r2=0.15
3UTR
Coding
Intron
Transposon

Figure S2 (continued)

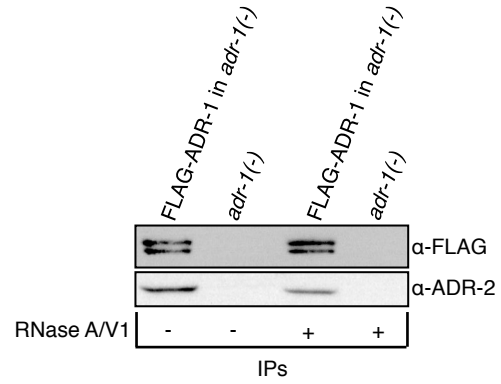
I



J



K



EXTENDED EXPERIMENTAL PROCEDURES

Transgenic Generation and Strains

Transgenic worm lines were generated by microinjection into the gonads of young adult worms of the appropriate genetic background. The injection mix used for generating transgenics contained the following: 1ng/μl of the transgene of interest, 20ng/μl of the dominant marker, and 79ng/μl of 1kb DNA ladder (NEB). Transgenic strains were maintained by passaging only worms with the dominant marker. The dominant markers used in this study were *rab3::gfp::unc-54* and the *rol-6(su1006)* plasmid, both of which were previously described (Hundley et al., 2008; Kramer et al., 1990; Mello et al., 1991). The transgenes expressing *adr-1* were injected with a modified pBluescript SK plasmid that contained the genomic *adr-1* locus, including the *adr-1* promoter (1245nt upstream of the start codon) and 3' UTR (1560 nt downstream of the stop codon) and three copies of the FLAG epitope (DYKDHD) immediately after the start codon. Mutations to the dsRBDs of adr-1 were generated by PCR and confirmed by Sanger sequencing.

The following strains were utilized in this study: Bristol strain N2, BB19 *adr-1(tm668)*, BB20 *adr-2(ok735)*, BB21 *adr-1(tm668); adr-2(ok735)*, BB21 *adr-1(tm668)* + blmEx1[3XFLAG-*adr-1* genomic, *rab3::gfp::unc-54*], BB21 *adr-1(tm668)* + blmEx2[3X FLAG-*adr-1* genomic with mutations in dsRBD1 (K223E, K224A, and K227A) and dsRBD2 (K583E, K584A and K587A), *rab3::gfp::unc-54*], BB76 xtIs4[*rab3::rfp::C35E7.6* (3' UTR), *rab3::gfp::unc-54* (3' UTR), *unc-119 genomic rescue],* BB76 *xtIs4 +* blmEx4[*rol-6(su1006)*], BB77 *adr-1(tm668); xtIs4 +* blmEx4, BB77 *adr-1(tm668); xtIs4* + blmEx5[3X FLAG::*adr-1 genomic*, *rol-6(su1006)*].

Detailed description of each step in the Bioinformatics Pipeline

Mirrors Figure 4A and Table S1

1) ssRNAseq: The *adr-1(-);adr-2(-)* sample was sequenced on one lane of Illumina's HiSeq 2000 yielding 216 million single-end 76nt reads. Each other sample was sequenced on a lane of Illumina GAII yielding between 37 and 42 million reads of the same type.

2) Mapping: Sequenced reads were mapped to the *C. elegans* reference genome (ce10, WS220) with the spliced aligner TopHat (version 2.0.6) allowing only uniquely-mapped reads with up to two mismatches each with command line options -Mx 1 and -N 2.

3) Variant calling: sites with RNA-DNA differences were identified by SAMtools mpileup (version 0.1.18) tallying up to 1000 alignments per site. Additional command line options used were -D -I and -g.

4) Site filters: Annotated SNPs were obtained from Illumina's iGenomes collection for C. elegans (ce10) and unannotated variants were extracted from the *adr-1(-);adr-2*(-)

RNA-Seq dataset. These genomic variants were filtered from the putative sites in all other strains reducing the number of false-positive predictions.

5) Read filters: Each read aligned to one the remaining putative sites was filtered out if:
   a) it was a suspected PCR duplicate, according to SAMtools rmdup (version 0.1.18)
   b) it had a junction overhang < 10nt according to its SAMtools CIGAR string
   c) it had > 1 non-A2G or non-C2T mismatch or any short indel, per its MD tag
   d) it had a mismatch less than 25nt away from either end of the read
   (this was changed to 5nt in the relaxed version used for quantification)

6) Identify sites: Putative RNA editing sites were identified from A2G variants on the sense strand and T2C variants on the antisense strand that were covered by more than 5 reads which passed the filters in step 5, including the stringent 25nt threshold for filter 5d).

7) Quantify sites: The extent of editing at each site and our confidence in that prediction were quantified by a novel extension of the classical Bayesian model used for genomic variants, which is described in more detail in the next section.

7.5) To increase the accuracy and confidence of our predictions, we used additional reads from the relaxed version of filter 5d) that overlap the sites identified in step 6. Moreover, we dropped sites that exhibited editing in 100% of the reads (suggesting a genomic variant not filtered out by step 4) and those with very low editing (less than 10%), which would have been hard to distinguish from sequencing errors.

8) The predicted RNA editing sites from each strain were characterized according to their position in annotated genic regions (introns, exons, 3'/5' UTRs, etc.) and according to their overlap with other strains.

Bayesian quantification model

Also known as the ``inverse probability model'' in the SNP calling community (Li et al., 2008) a Bayesian model for identifying DNA polymorphisms from error-prone sequencing data has been shown to perform favorably to other discrete and discriminative models (Bahn et al., 2012). In general, the power of a Bayesian approach is its combination of prior knowledge and observed data into a posterior estimate. The prior knowledge encodes general domain-specific information like biases in the sequencing technology, while the observed data contain signals specific to editing sites in a particular experiment. In this exposition, we will use a simple context-independent prior for all editing sites, which consist of pseudo-counts of edited and non-edited reads: β and α, respectively. For sequence alignments in particular, the benefit of a Bayesian approach is that even low-coverage regions can give reasonable posterior estimates of the editing efficiency with low confidence, while high-coverage regions will give very accurate posterior estimates with high confidence.

For example, consider two candidate-editing sites: site L has low coverage and site H has high coverage. Let the number of reads from edited (*g*) and unedited (*a*) transcripts containing those sites be: g_L = 1, a_L= 9 for site *L* and g_H=10, a_H=90 for site *H*. The observed counts suggest that both sites are edited with 10% efficiency, but we are inclined to believe that site *H* really is edited while site *L* is not and its single edited read could have easily been produced by a sequencing error. While filter-based approaches require manual fine-tuning to be able to filter out site *L* while keeping site *H*, the Bayesian approaches will simply have a lot more confidence that site *H* is edited. To formalize the notion of confidence, we introduce a latent binary variable *γ* which indicates whether a nucleotide is edited *γ=1* or not *γ=0*. Given a prior belief in the occurrence of RNA editing $P(\gamma_S) = \frac{\beta\gamma_S + \alpha(1-\gamma_S)}{\alpha+\beta}$ at a particular site *S* (which is currently site-independent but can be extended to differ depending on the genomic context or read position of *S*), and the likelihood of observing the RNA-seq reads at site *S* conditioned on the hypothesis of editing $P(a,g \mid \gamma_S=1)$ versus no editing $P(a,g \mid \gamma_S=0)$ which captures the probability of a sequencing error *ε*, Bayesian models for DNA-RNA differences use the ``inverse probability'' rule to produce a posterior belief on whether site *S* is edited or not:

$$P(\gamma_S|a,g) = \frac{P(\gamma_S)P(a,g|\gamma_S)}{P(a,g|\gamma_S=0)+P(a,g|\gamma_S=1)} = \frac{1}{\epsilon^a+(1-\epsilon)^g}\begin{cases}\alpha(1-\epsilon)^g & \text{if } \gamma_S=0\\ \beta\epsilon^a & \text{if } \gamma_S=1\end{cases}$$

Thus, instead of relying on a stringent threshold on the coverage to identify editing sites or completely excluding particular genomic loci such as splice junctions, we will compare our confidence in the editing hypothesis $P(\gamma_S=1 \mid a,g)$ to that of the no-editing hypothesis $P(\gamma_S=0 \mid a,g)$. A convenient way to measure the difference in these two hypotheses as a particular genomic site S is to take their log-ratio, which causes the partition function $P(a,b) = \varepsilon^a + (1-\varepsilon)^g$ to cancel out from top and bottom:

$$LLR(a,g) = \log\frac{P(\gamma_S=1|a,g)}{P(\gamma_S=0|a,g)} = \log\frac{\alpha(1-\epsilon)^g}{\beta\epsilon^a}$$

This measure depicted by the heatmap in Figure S4 has the desirable property of extracting the maximum confidence from the coverage at a given editing site. However, *LLR* alone is not sufficient to accept or reject either hypothesis in the way p-values are often used and misused (Simmons et al., 2011). However, it is very useful in ranking different sites in order of relative confidence that editing occurs at each. Given a ranked list of potentially edited sites, this approach still requires a cutoff in order to make actual predictions subject to validation. However, compared to the multiple thresholds for each filter in pipeline-based approaches, it is easier to manually pick or learn this parameter from training data. We tried three confidence cutoffs (0.95, 0.995, and 0.999) and chose the 0.995 based on two factors: the number of sites predicted in the *adr-1(-)* and N2 strains (141 and 59, respectively) was sufficiently large, but the number of sites in the *adr-2(-)* strain remained relatively low (only 6).

RNase Treatment of FLAG-ADR-1 Immunoprecipitations (IPs)
IPs were performed as previously stated except worms were not subjected to UV-crosslinking and only light salt washes were employed. For Micrococcal Nuclease (MNase) treatment, IPs were washed twice with MNase reaction buffer (50mM Tris-Cl, 5mM $CaCl_2$, pH 7.9), resuspended in MNase reaction buffer and treated with MNase (NEB) at a concentration of 20U/µl for 30 minutes at 37°C. For RNase A/VI treatment, IPs were washed twice with RNA Structure Buffer (10mM Tris-Cl 100mM KCl, 10mM $MgCl_2$, pH 7.0), resuspended in RNA Structure Buffer and treated with both RNase A (5 Prime) and RNase V1 (Ambion) at .07U/µl and .001U/µl respectively. Following MNase or RNase A/V1 treatment, IPs were washed twice with light salt wash buffer and analyzed by SDS-PAGE and western blotting.

SUPPLEMENTAL REFERENCES
Bahn, J.H., Lee, J.H., Li, G., Greer, C., Peng, G., and Xiao, X. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Genome Res *22*, 142-150.

Hundley, H.A., Krauchuk, A.A., and Bass, B.L. (2008). C. elegans and H. sapiens mRNAs with edited 3' UTRs are present on polysomes. RNA *14*, 2050-2060.

Kramer, J.M., French, R.P., Park, E.C., and Johnson, J.J. (1990). The Caenorhabditis elegans rol-6 gene, which interacts with the sqt-1 collagen gene to determine organismal morphology, encodes a collagen. Mol Cell Biol *10*, 2081-2089.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res *18*, 1851-1858.

Mello, C.C., Kramer, J.M., Stinchcomb, D., and Ambros, V. (1991). Efficient gene transfer in C.elegans: extrachromosomal maintenance and integration of transforming sequences. EMBO J *10*, 3959-3970.

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci *22*, 1359-1366.

SUPPLEMENTAL LEGENDS

Figure S1, Confirmation of Editing Assay Reproducibility and ADR-1 transgene viability, Related to Figure 1.

(A) Accuracy of Sanger Sequencing Editing Assay. Wild-type cDNAs from the *lam-2* 3' UTR were generated in two independent reverse transcription (RT) reactions. After PCR amplification, the cDNAs were subjected to Sanger sequencing and editing was quantified using the Bio-Edit program.

(B) FLAG-ADR-1 rescues neuronal reporter phenotype. Bar heights represent relative fluorescent intensity of a red fluorescent protein reporter with the *C35E7.6* 3' UTR divided by the fluorescent intensity of a green fluorescent protein reporter with an unstructured 3' UTR. For each worm strain, fluorescent intensities were measured for 25 young adult worms in three biological replicates. The average values for each strain were normalized to the wild-type, error bars represent SEM.

(C) Editing levels at individual nucleotides within the 3' UTRs were measured for 3 biological replicates. Error bars represent standard error of the mean (SEM). Significant changes ($p \leq 0.05$) in editing levels between wild-type and *adr-1(-)* are marked with a single asterisk. Those that are also significant changes ($p \leq 0.05$) in editing levels between *adr-1(-)* and FLAG-ADR-1 are marked with an additional asterisk.

Figure S2: Transcriptome-wide identification and quantitation of A-to-I editing in *C. elegans,* Related to Figure 4.

(A) cDNAs from the 3' UTRs of the indicated endogenous genes were amplified from wild-type and *adr-2(-)* worms and subjected to Sanger sequencing. In addition, for the noncoding regions, genomic DNA was amplified and sequenced. Chromatograms of the editing event(s) (bold A) identified from the RNA-Seq data are shown, but editing was examined for the entire RT-PCR product (Table S2).

(B-E) Venn diagrams show the overlaps in quantified sites between key pairs of strains. The number of sites in each strain was covered by more than 5 reads are shown and the remainder of all 270 editing sites in other strains is depicted in the black rectangles.

(B) The wild type (CEN2) and FLAG-ADR-1 strains overlap in 88 out of 170 sites.

(C) The *adr-1(-)* and wild type (CEN2) strains overlap in 81 out of 185 sites.

(D) The *adr-1(-)* and FLAG-ADR-1 strains overlap in 108 out of 225 sites.

(E) The *adr-1(-)* and DS12MT strains overlap in 106 out of 202 site total.

(F-H) Scatter plot of percent editing in one strain vs another for the quantified sites that overlap in the two indicated stains. The r^2 value describes the least-squares fit of the scatter cloud to the y=x line (black diagonal).

(I) A priori distribution over the editing model's confidence**.** For a fixed expression level, our confidence in the validity of an editing site is higher (red) for editing near 50% where the evidence from variant (G) and non-variant (A) reads is balanced, as compered to less confident sites (blue) with editing near 0% or 100%, where read evidence can be explained away as a mutation or a sequencing error. For each putative editing site, the numbers of variant (G) and non-variant (A) reads that overlap it are tallied and the likelihood of the observed totals is multiplied by the prior shown above. The result of this multiplication is a posterior distribution over the editing for this site, with maximum

confidence reached at the most likely editing % for that site.
(J and K) Immunoblotting analysis of FLAG immunoprecipitations (IPs) from the indicated strains. IPs were performed as previously stated except worms were not subjected to UV-crosslinking and only light salt washes were employed. After washing, a subset of IPs were treated with (J) Micrococcal Nuclease (MNase +) or (K) RNase A and RNase VI (RNase A/VI +). All IPs were then subjected to immunoblotting for the FLAG epitope and ADR-2.

Table S1: 270 high confidence editing sites identified by the bioinformatics pipeline, Related to Figure 4.
(Sheet 1) All unique editing sites from the pipeline are numbered (editing site #) and given an assigned gene based on wormbase annotations (Gene). In addition to site number, the editing sites are organized by transcript number (Transcripts #) to demonstrate the number of edited RNAs identified. Chromosomal number (CHROM) and location (POS) for all editing sites identified from the pipeline. In addition, the strand that the assigned genes are located is listed (plus and minus). Last, genes that were identified as *C. elegans* Staufen targets by LeGendre, et.al., JBC, 2013 are identified.
(Sheets 2-6) Tables candidate-editing sites identified for each of the indicated strains. In addition to chromosomal number (CHROM), location (POS) and the strand that the assigned genes are located is listed (plus and minus), the single nucleotide change from the wormbase genome (REF) and the RNA-Seq data (ALT) is given. Please note that both A-G and T-C changes on the plus and minus strands, respectively are indicative of A-to-I editing. Last, the number of reads (NUM_READS) for each site and the confidence calculated by the pipeline (CONFIDENCE) is listed.

Table S2: List of editing sites identified with Sanger editing assays from edited mRNAs identified by the bioinformatics pipeline, Related to Figure 4. Gene-specific reverse transcription, PCR amplification and Sanger sequencing of the indicated genes (Gene) were performed. The chromosomal location of each gene (Chromosome) and all A-to-I editing sites (Editing site position) are listed. Editing sites predicted by the pipeline are noted.

Table S3: 81 common predicted editing sites between WT and *adr-1(-),* Related to Figure 4. The chromosomal location of the high confidence editing sites that are covered by more than 5 reads in both wild-type (WT) and *adr-1(-)* RNA-Seq datasets. For both strains, the percent editing (Editing %) for each site was calculated by the number of reads containing an A-I editing event divided by the total number of reads at a given site. Two methods were used to determine if editing sites were regulated by ADR-1. Sites were determined to be regulated by ADR-1 if editing levels were between the two strains were greater than 12% (method 1). In method 2, the cutoff between regulated and non-regulated was determined by the read density at that editing site. To be regulated by ADR-1 in method 2, the difference in editing levels between the two strains needed to be greater than the editing percentage calculated if 1 read were edited out of the total reads covering that site for the strain that had the lowest read coverage.

The last two columns (G and H) list the co-occurrence of each editing site with additional sites in both the WT (N2) and FLAG-ADR-1 RNA-seq datasets.

Table S4: Sequences of all primers, Related to Experimental Procedures. Primers used in this study for Sanger editing assays and qRT-PCR analysis.