SUPPLEMENTAL FIGURES AND FIGURE LEGENDS

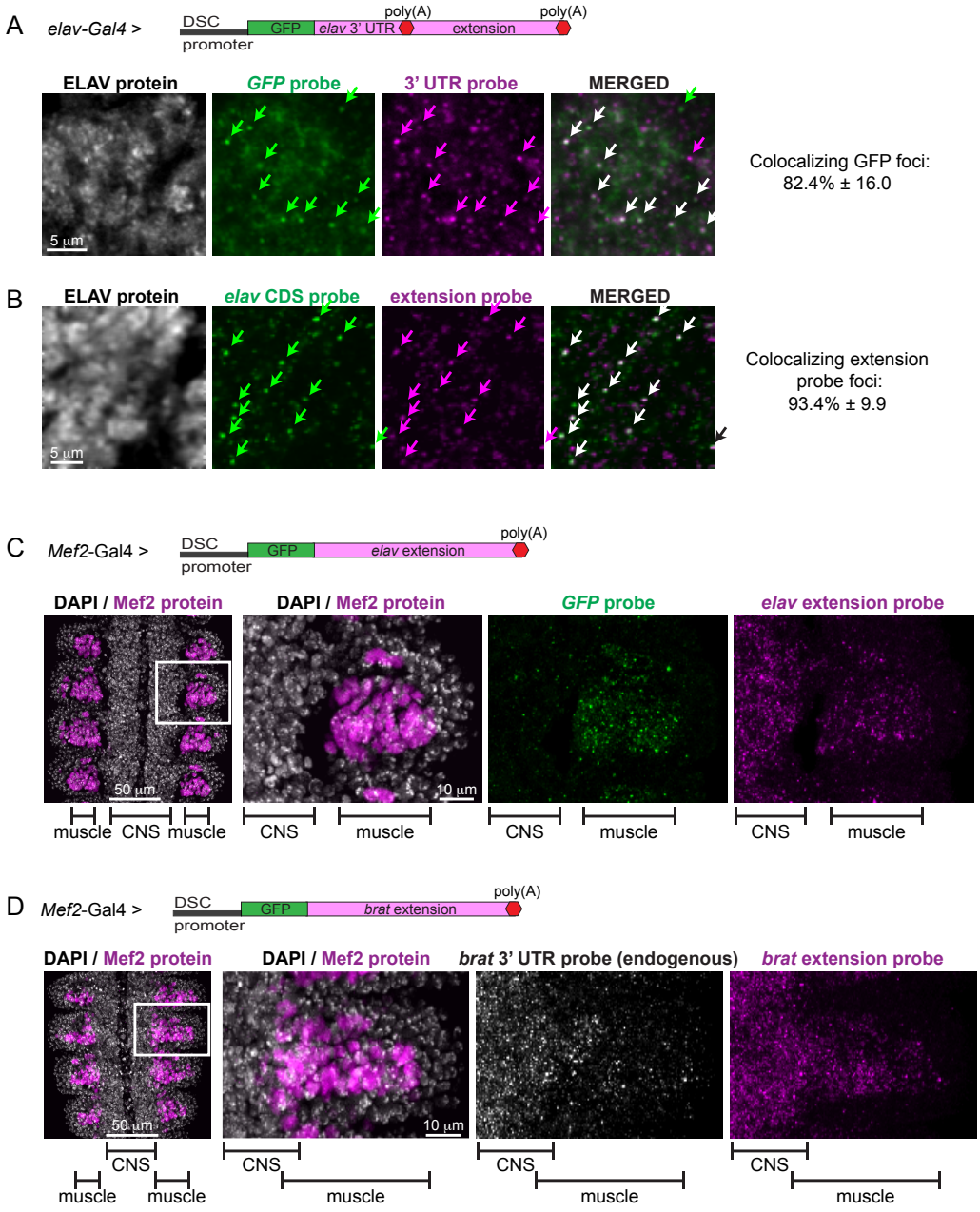

Figure S1. Transgene properties that promote expression of 3' extensions. Related to Figure 1.

A-B. Native promoters are required for expression of 3' extensions.

A. *elav-Gal4* drives expression of a *GFP* transgene in the nervous system. The promoter used for expression was the DSCP. The *GFP* coding sequence was

placed upstream of the entire extended 7.2 kb *elav* 3' UTR. CPA at the proximal poly(A) produces the short 3' UTR form of the mRNA, whereas CPA at the distal-most poly(A) produces the fully extended transcript. RNA probes directed against different regions of the transcripts were used to detect mRNAs. Shown are double fluorescent in situ hybridization assays. Single confocal sections of a portion of the developing CNS in stage 13 embryos. Colocalization of the *GFP* and 3' UTR probes indicates expression of the short transcript from the transgene.

B. Virtually all foci from the *elav* extension probe colocalize with the probe directed against the endogenous *elav* coding sequence, which indicates that the extension signal originates from the endogenous *elav* transcript. Numbers represent mean ± SD of six embryos for each sample.

C-D. Bypassing the requirement for ELAV recruitment allows for transcription of extension sequences from the DSCP.

Mef2-Gal4 drives expression of *GFP* transgenes in muscle cells. The promoter used for expression was the DSCP. The *GFP* coding sequence was placed upstream of the extended portion of the *elav* 3' UTR (C) or the extended portion of the *brat* 3' UTR (D), thereby excluding the respective short 3' UTRs and proximal poly(A) signals. CPA at the indicated poly(A) produces a transcript that was detected using RNA probes directed against the *GFP* coding sequence (C) as well as a distal region of the *elav* (C) or *brat* (D) 3' UTR extension. Shown are double fluorescent in situ hybridization assays combined with antibody staining against Mef2 protein as a muscle marker. Projections of consecutive confocal sections of stage 13 embryos. Ventral views; anterior is up.

C. The *GFP* mRNA signal in muscle cells shows muscle-specific expression of the transgene. Signal from the *elav* extension probe in the central nervous system (CNS) corresponds exclusively to the endogenous extended *elav* transcript. Detection of extension sequences in muscle cells indicates expression of extended transcripts from the transgene.

D. Signal from the *brat* 3' UTR (that is not present in the transgene mRNA) and *brat* extension probes in the CNS corresponds exclusively to endogenous *brat*

transcripts. Detection of extension sequences in muscle cells indicates expression of extended transcripts from the transgene.

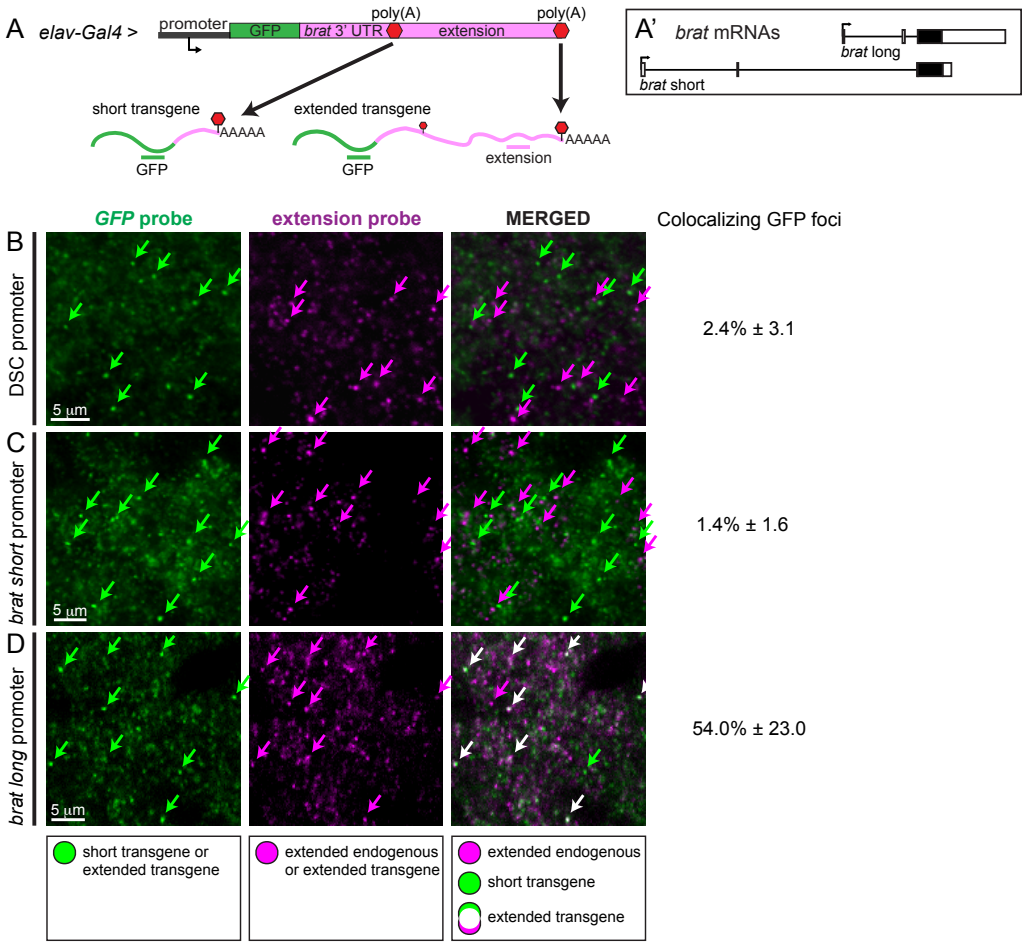

Figure S2. Native promoters are required for expression of 3' extensions. Related to Figure 1.

A. *elav-Gal4* drives expression of a *GFP* transgene in the nervous system. Three different promoter regions were used: DSCP (B), the native promoter producing the short form of *brat* (C), and the native promoter producing the extended form of *brat* (D). A' depicts the configuration of endogenous extended and short *brat* mRNAs with their respective promoters. The *GFP* coding sequence was placed upstream of the entire extended 8.5 kb *brat* 3' UTR. CPA at the proximal poly(A) produces the short 3' UTR form of the mRNA, whereas CPA at the distal-most

poly(A) produces the fully extended transcript. RNA probes directed against different regions of the transcripts were used to detect mRNAs.

B-D. Double fluorescent in situ hybridization assays using probes indicated in A. Single confocal sections of a portion of the developing CNS in stage 13 embryos. Note that the extension probe detects not only the transgene, but also the endogenous *brat* transcript, which is expressed in the nervous system. Colocalization of the *GFP* and extension probes indicates expression of extended transcripts from the transgene.

B,C. The reporter transgenes carrying the DSCP (B) or the native promoter of the short form of *brat* (C) do not exhibit colocalization of *GFP* and extension probes. Extension signals (magenta arrows in merged image) do not colocalize with the green *GFP* signals, indicating that they correspond to endogenous *brat* mRNAs.

D. Replacing the DSCP with the native promoter producing the extended form of *brat* induces 3' extension of the *GFP* transgene. There is extensive colocalization of the *GFP* (green arrows) and extension probes (magenta arrows), indicating expression of extension sequences from the transgene (white arrows in merged image). Non-colocalizing *GFP* signal (e.g., green arrow in merged image) corresponds to the short transgene, and non-colocalizing signal from the extension probe (e.g., magenta arrow in merged image) corresponds to the endogenously expressed extended *brat* mRNA. Numbers represent mean ± SD of six embryos for each promoter (except C: three embryos).

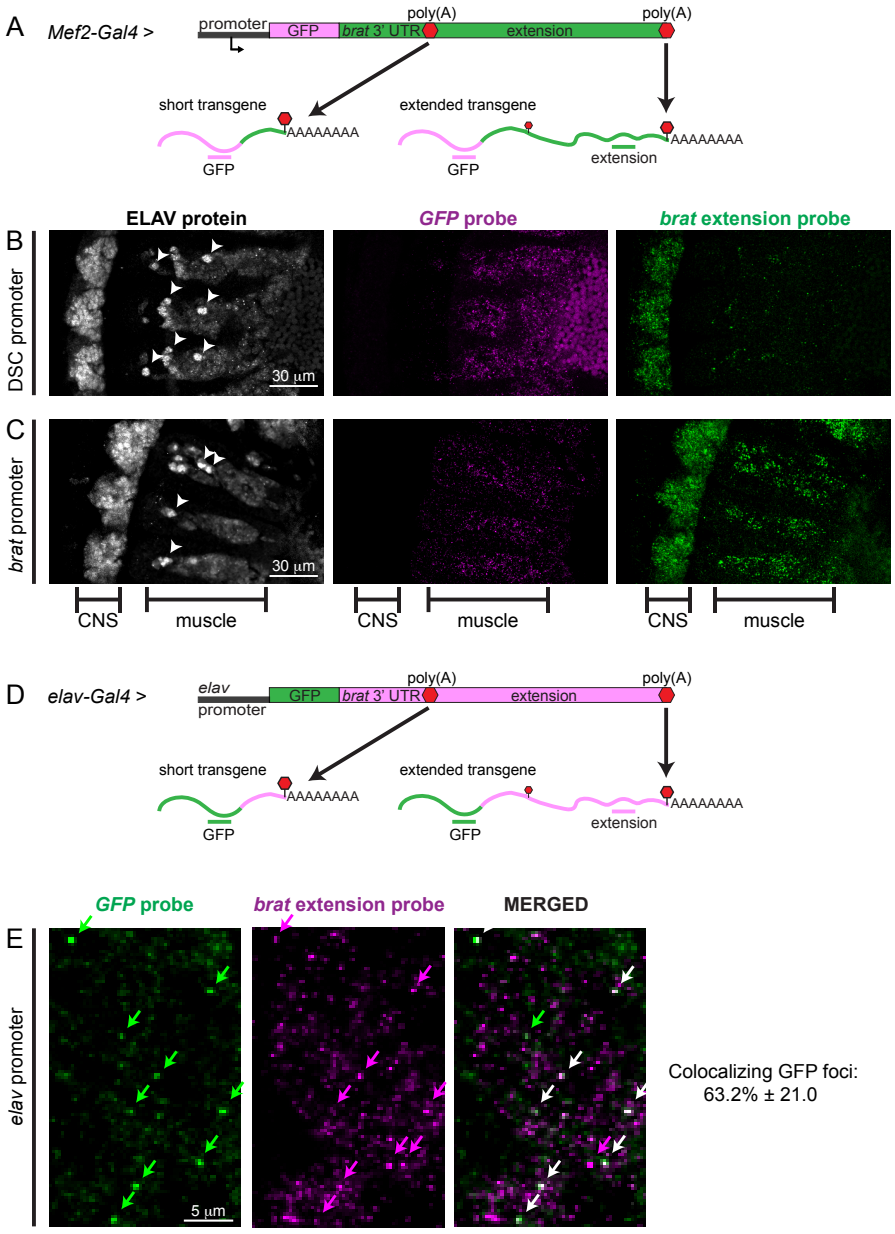

Figure S3. The native *brat* and *elav* promoters mediate *brat* 3' UTR extension. Related to Figure 2.

A-C: The native *brat* promoter mediates 3' UTR extension in ectopic tissues.

A: *Mef2-Gal4* drives expression of a *GFP* transgene in muscle cells. The promoter used for expression was either the DSCP, or the native *brat* promoter. The *GFP* coding sequence was placed upstream of the entire extended 8.5 kb *brat* 3' UTR. CPA at the proximal poly(A) produces the short 3' UTR form of the

mRNA, whereas CPA at the distal-most poly(A) produces the fully extended transcript. RNA probes directed against different regions of the transcripts were used to detect mRNAs.

B,C. Double fluorescent in situ hybridization assays using probes indicated in A, combined with antibody staining against ELAV protein. Projections of consecutive confocal sections of stage 13 embryos. Lateral views; anterior is up. The weak ELAV signal in muscle cells corresponds to ectopic expression driven by *Mef2-Gal4*, whereas the strong signal in the CNS corresponds to endogenous ELAV. Arrowheads indicate neurons of the PNS (strong ELAV signal). Ectopic expression of the short 3' UTR *GFP* transgene can be achieved from both DSC and *brat* promoters, as shown by detection of *GFP* probe signal in muscle cells in B and C (middle panels, magenta). Right panels exhibit hybridization signals with the *brat* extension probe (green). Signal in the CNS corresponds exclusively to the endogenous extended *brat* transcript, whereas expression in the muscle corresponds exclusively to reporter expression. Background staining in muscle is observed with the DSCP transgene (B), indicating little or no expression of the extended 3' UTR from the GFP transgene. In contrast, there is significant expression of extended transcripts from the transgene containing the *brat* promoter in muscle (C).

D,E. The native *elav* promoter mediates *brat* 3' UTR extension.

D. *elav-Gal4* drives expression of a *GFP* transgene in the nervous system. The promoter used for expression was the native *elav* promoter. The *GFP* coding sequence was placed upstream of the entire extended 8.5 kb *brat* 3' UTR. CPA at the proximal poly(A) produces the short 3' UTR form of the mRNA, whereas CPA at the distal-most poly(A) produces the fully extended transcript. RNA probes directed against different regions of the transcripts were used to detect mRNAs.

E. Double fluorescent in situ hybridization assays using probes indicated in D. Single confocal sections of a portion of the developing CNS in a stage 13 embryo. Note that the extension probe detects not only the transgene, but also the endogenous *brat* transcript, which is expressed in the nervous system. There is

extensive colocalization of the *GFP* (green arrows) and extension probes (magenta arrows), indicating expression of extended 3' UTR sequences from the transgene (white arrows in merged image). Non-colocalizing *GFP* signal (e.g., green arrow in merged image) corresponds to the short transgene, and non-colocalizing signal from the extension probe (e.g., magenta arrow in merged image) corresponds to the endogenously expressed extended *brat* mRNA. Numbers represent mean ± SD of six embryos.

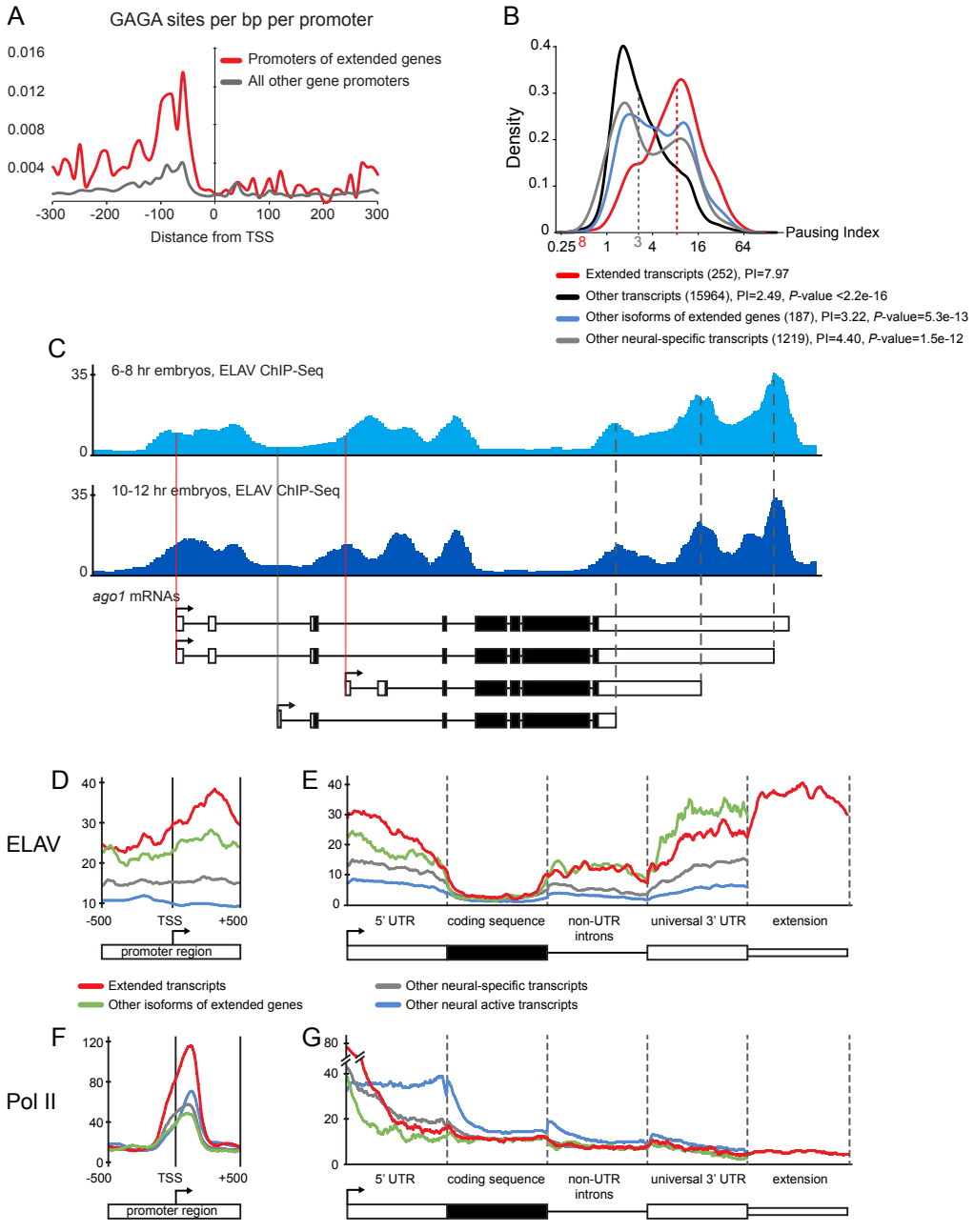

Figure S4. Promoters of extended genes contain the GAGA motif and paused Pol II and are bound by ELAV. Related to Figure 3 and Figure 4.

A. Frequency of occurrence and distribution of identified GAGA motifs in promoters of extended or control genes relative to the TSS. GAGA motifs are most often located between -100 bp and the TSS in both groups of promoters and occur significantly more frequently in promoters of extended genes.

B. Pausing index (PI) distribution and median pausing index values of the promoters of the indicated groups of transcripts in muscle tissues (see Supplemental Experimental Procedures), where ELAV is absent. The numbers in parentheses denote the number of transcripts in each group. Promoters of extended transcripts are significantly more paused than promoters of any control group. Wilcoxon rank sum test *P*-values were calculated by comparing the pausing index of extended transcripts with each group of controls.

C. Normalized ELAV ChIP-Seq reads at the *ago1* locus in 6-8 hr and 10-12 hr embryos. Shown are merged tracks of duplicate experiments. ELAV peaks at each proximal poly(A) site (dotted lines) are found in both 6-8 hr and 10-12 hr embryos.

D,E. Meta-gene plots of ELAV ChIP-Seq datasets at the promoter region (D) (±500 bp relative to the TSS) or across the entire transcription unit (E) in 10-12 hr embryos. Each line (meta-gene) averages the ChIP-Seq data of all indicated transcripts. ELAV binding is higher in extended transcripts compared to other transcripts (see exception below) at the promoter region, 5' UTR, introns and the 3' UTR. In all genes, ELAV binding is excluded from the coding sequence. Differences in ELAV binding between extended transcripts and 'other isoforms of extended genes' are not significant. We think the reason is that transcripts from these two groups share many gene regions including sequences as close as ±100 bp relative to the TSS, introns and the universal 3' UTR. Moreover, both groups of transcripts are relatively small (252 and 187 transcripts, respectively).

F,G. Meta-gene analysis of Pol II binding at promoter region (F) or across the entire transcription unit (G) in 12-16 hr embryos. Promoter regions of extended transcripts show significantly higher Pol II binding than other control groups of transcripts. Other regions downstream of the TSS do not differ in their Pol II binding profile between the four groups, except the "other neuronal active transcripts" that show higher Pol II binding at the 5' UTR and the coding sequence due to their high level of expression. See also Table S1 and Table S2 for ELAV peak coordinates.

Table S1. Listing of chromosomal coordinates (UCSC dm3 release) of 6879 ELAV binding peaks in 6-8 hr embryos identified by ChIP-Seq.

Table S2. Listing of chromosomal coordinates (UCSC dm3 release) of 8076 ELAV binding peaks in 10-12 hr embryos identified by ChIP-Seq.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Primers

Native promoter sequences (300 bp surrounding the transcription start site) were amplified from genomic DNA and cloned into pBID-UASc-eGFP SacI/BglII, thus removing the DSC promoter and maintaining the UAS repeats.
Primers for *brat short*:
forward: 5'- TCCCATTTTGAATTTAAGTAAAACTTAGCC
reverse: 5'- AATTGGCCACAGAACAAAGCG
Primers for *brat long*:
forward: 5'- GTTGAGTGAGTTTTTTCGGCTG
reverse: 5'- ATAGGCTAGGTATGTGTATGTTTCTGTTG
Primers for *elav*:
forward: 5'-CTCGAGAGGCAACTATGAGATATGAG,
reverse: 5'- ATTTCGCTCGGTGTGAGATGA.

Extended 3' UTR sequences were amplified from genomic DNA and cloned into the modified pBID-UASc-eGFP NotI/XbaI.
Primers for *brat*:
forward: 5'-CACACAGACACACACACTCCATG
reverse: 5'- TGCCAAAACTCACTGATCGAATAA
Primers for *elav*:
forward: 5'-AGCGGCCCAAATGGAAG

reverse: 5'-TCGGTCATAGTGTCATTTATTCCAT

Extension sequences lacking the short 3' UTR and the proximal poly(A) were cloned in the same way.
Primers for *elav*:
forward: 5'-CCAATTTACCTATTAAGTAAGCAAAGAGC
reverse: 5'- CTTCGTAATTAAAACAACCCTTTCAGT
Primers for *brat*:
forward: 5'- AAAACAAGGCGATATTTATGTGCA
reverse: 5'- AGTCATTTAAGTCATTTATGCTTGCC

Computational filtering of the extended transcripts and the control groups

We focused our analysis on known neural-specific extended genes. We used the RefSeq (release 65) annotation and published RNA-seq data (Gaertner et al., 2012; Graveley et al., 2011) to filter the 401 described genes that undergo 3' UTR extension (Hilgers et al., 2011; Smibert et al., 2012). The transcripts of these 401 genes which satisfied all the following 3 criteria were included in this study: (1) had a 3' UTR extension of at least 200 bp, (2) had at least 1 read per kilobase of exon model per million mapped reads (RPKM) in either 10-12 hr or 14-17 hr embryo neurons, and (3) had significantly more expression in neurons than in muscle cells in 10-12 hr and 14-17 hr embryos (*P*-value ≤ 0.05 by RankProd (Hong et al., 2006)). After these filters, if several extended isoforms shared the same transcription start site (TSS), the longest extended isoform was used. If several isoforms shared the same transcription termination site (TTS), the isoform with the highest expression in neuron cells was used. 252 transcripts of 219 genes with 3' UTR extensions were included in the analysis.

Four control groups of transcripts were used in this study: (A) all transcripts except the known 3' UTR extended transcripts, (B) all non-extended isoforms of the 219 extended genes, (C) all non-extended transcripts that are active in

neurons (RPKM > 10 in neurons) and (D) all non-extended transcripts that are specifically expressed in neurons (RPKM > 1 in neurons and RankProd *P*-value ≤ 0.05 when comparing their expression in neurons vs. muscle cells). All isoforms containing an annotated 3' UTR extension of at least 200 bp were excluded from all control sets. If several isoforms shared the same TSS, only the one with the highest expression in whole embryos (Negre et al., 2011) was included. The four control sets consist of (A) 15964, (B) 187, (C) 5841 and (D) 1219 transcripts.

Determination of pausing indexes

We defined promoter regions as 200 bp surrounding +30 bp from the annotated TSS in RefSeq (release 65), and defined gene body regions as TSS +400 bp to the 3' end of the genes. Genes whose size did not exceed 400 bp were excluded. Pol II enrichment at promoters and gene bodies was calculated as the fold enrichment of the normalized Pol II reads over input (Negre et al., 2011). Pausing index was defined as the Pol II enrichment at promoter regions divided by the Pol II enrichment within gene bodies. For each promoter, the maximum pausing index in 4-24 hr embryos was used. For analysis of pausing indexes in muscle tissue, the maximum pausing index in the following samples was used: 2-4h Toll10b mutant embryos, 6-8h, 8-10h, 10-12h, and 14-17h Mef2-sorted muscle cells (Gaertner et al., 2012).

Chromatin Immunoprecipitation and sequencing

Chromatin was prepared from 0.5 g of dechorionated 6-8 hr and 10-12 hr wild-type (*yw*) embryos. Two independent chromatin preparations (biological replicates) were done for each time point. Sonication of chromatin was performed with a Bioruptor (Diagenode) yielding genomic DNA fragments with an average size of ~200 bp. ChIP assays were done using a mix of two antibodies (5 ug each) raised against the entire ELAV protein (483 aa): mouse anti-ELAV-9F8A9

and rat anti-ELAV-7E8A10 (DSHB). ChIP DNA was resuspended in 40 ul, 8 ul of which were used to evaluate specific enrichment by qPCR. The remaining 32 ul were used to construct ChIP-Seq libraries.

ELAV and Pol II ChIP-Seq data processing

All sequencing reads were aligned to the *Drosophila melanogaster* reference genome (UCSC dm3 release) using Bowtie version 0.12.7 (Langmead et al., 2009). The following Bowtie parameters were used to select only uniquely aligning reads with a maximum of two mismatches:

-k 1 -m 1 -n 2 --best --strata

We used the model-based analysis of ChIP-Seq (MACS) peak-finding algorithm version 1.4.1 to identify the regions of ELAV and Pol II ChIP-Seq enrichment over input DNA. The default MACS parameters were applied. For each time point, only the overlapped ELAV-binding regions between two biological replicates were used for further analysis. In total, 6879 and 8076 ELAV-binding regions were identified in 6-8 hr and 10-12 hr embryos, respectively.

SUPPLEMENTAL REFERENCES

Gaertner, B., Johnston, J., Chen, K., Wallaschek, N., Paulson, A., Garruss, A.S., Gaudenz, K., De Kumar, B., Krumlauf, R., and Zeitlinger, J. (2012). Poised RNA polymerase II changes over developmental time and prepares genes for future expression. Cell reports *2*, 1670-1683.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W.*, et al.* (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473-479.

Hilgers, V., Perry, M.W., Hendrix, D., Stark, A., Levine, M., and Haley, B. (2011). Neural-specific elongation of 3′ UTRs during Drosophila development. Proceedings of the National Academy of Sciences *108*, 15864-15869.

Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L., and Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics (Oxford, England) *22*, 2825-2827.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Negre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., Sealfon, R.*, et al.* (2011). A cis-regulatory map of the Drosophila genome. Nature *471*, 527-531.

Smibert, P., Miura, P., Westholm, J.O., Shenker, S., May, G., Duff, M.O., Zhang, D., Eads, B.D., Carlson, J., Brown, J.B.*, et al.* (2012). Global patterns of tissue-specific alternative polyadenylation in Drosophila. Cell reports *1*, 277-289.