

Supplementary Materials:

Materials and Methods:

C. riparius culture

Embryos of *Chironomus riparius* (Meigen) (Diptera: Chironomidae) were obtained from a laboratory culture at the University of Chicago (Illinois). Details on the life cycle and culture methods have been published elsewhere (24). The culture was obtained from Gerald K. Bergtrom (University of Wisconsin, Milwaukee, WI) in 2004 and has been maintained at 25°-30°C with 60% relative humidity and a constant 16/8-hour day/night cycle. Eggs, larvae, and pupae were reared in aerated reconstituted freshwater (1.14 mM NaHCO₃; 0.12 mM MgSO₄; 0.37 mM CaSO₄; 0.45 mM CaCl₂; 0.05 mM KCl). Larvae were fed with a suspension of sterilized food paste prepared from pulverized parsley (Aquatic Eco-Systems, Apopka, FL) supplemented with active dry baker's yeast (3.3% w/w, Red Star).

Cloning procedures and RNA synthesis

Cri-cad (KP769548), *Cri-nos* (KP769547), *Cri-hb* (KP769546), *Cri-tor* (KP769549), *Cri-tll* (KP769551), *Cri-oc* (KP769552), and *Cri-hkb* (KR076545) were initially identified using published degenerate PCR primers and RACE or PCR (6, 25, 26). The *panish* (KP769550) and *pangolin* (KP769545) transcripts were cloned with primers constructed from RNA-seq data. *Nco*I and *Xba*I sites were incorporated immediately upstream of the putative translation start site and downstream of the stop codon. These restriction sites were used for ligation into the expression vector pSP35T (27) to generate capped mRNAs using the Ambion® mMessage mMachin Kit with SP6 RNA polymerase (Life Technologies, NY, USA). Synthesized mRNA was dissolved in H₂O to yield a concentration of ~7 µg/µl. Aliquots were stored at -80 °C until use. *Eba-bcd* mRNA was generated as described in (6). Primer sequences are available upon request.

Double-stranded RNAs (dsRNAs) were generated from TOPO-TA-PCR II (Life Technologies, NY, USA), pGEM-T (Promega, WI, USA), or pSP35T plasmids containing gene-specific cDNA sequence using vector-specific primers with attached T7 promoter sequences. DsRNAs were prepared from partial gene sequences as follows (pos. -1 is the last nucleotide in ORF before the stop codon): *Cri-cad* pos. -213 to -1275, *Cri-hb* pos. -1026 to -26, *Cri-hkb* pos. -641 to -21, *Cri-nos* pos. -1411 to -584, *Cri-oc* pos. -1074 to -65, *panish* pos. 1 to 912, *Cri-tll* pos. -103 to -936, pos. -925 to -630, and pos. -492 to -93, and *Cri-tor* pos. -1210 to -269. *In vitro* transcription was done using the MEGAscript T7 transcription kit (Life Technologies, NY, USA). dsRNA products were dissolved in injection buffer (100 mM NaPO₄, pH=6.8; 5 mM KCl) and confirmed on a 1% agarose gel prior to use.

Injection procedures

Early embryos were treated with 3% commercial bleach for one minute and were immediately subjected to 6 washes with reconstituted fresh water, aligned on a glass slide along a thin glass capillary (outer diameter: 0.2 mm), dried briefly to remove water, and covered with a 1:1 mixture of 27-halocarbon oil (Sigma H8773) and 700-halocarbon oil (Sigma H8898). Embryos were injected prior to or during the two-pole-cells stage with an IM300 microinjector (Narishige, NY, USA) at 18 °C. Embryos were kept in a moist chamber at ~25 °C and allowed

to continue developing following injections. For initial reported *panish* mRNA rescue experiments, *panish* mRNA was injected with *panish* dsRNA and again at approximately one hour later to boost efficiency. Following modification of the expression vector with the kozak sequence most prevalent in the *Chironomus* transcriptome, "...AAAAATG...", only a single dose of *panish* mRNA was used, allowing for rescue of the double-abdomen phenotype with equal efficiency.

Fixation procedures

Embryos were fixed for 45 minutes in a mixture of 8% formaldehyde in PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na₂HPO₄, 2 mM KH₂PO₄, pH 7.4) and n-heptane (300 μ L PBS, 81 μ L 37% formaldehyde, 171 μ L n-heptane), treated with 0.9 mg/ml proteinase K in PBT for 30-60 minutes at 37 °C, and devitellinized in a 1:1 mixture of n-heptane and methanol.

In situ hybridization and immunohistochemistry

RNA *in situ* hybridizations were done essentially as described (28, 29), using digoxigenin (DIG)-labeled probes and Fab fragments from anti-DIG antibodies conjugated with alkaline phosphatase (AP) (Roche, IN, USA). Probes were prepared from partial gene sequences as follows (pos. -1 is the last nucleotide in ORF before stop codon): *Cri-cad* pos. -213 to -1275, *Cri-hb* pos. -1026 to -26, *Cri-hkb* pos. -641 to -34, *Cri-nos* pos. -1411 to -584, *Cri-oc* pos. -1074 to -65, *panish* pos. 1 to 912, *Cri-tll* pos. -103 to -936, pos. -925 to -630, and pos. -492 to -93, and *Cri-tor* pos. -1210 to -269

Cuticle preparations

Cuticles were prepared three to four days after injection. Following mechanical eggshell removal, the embryos were incubated at room temperature for at least 2 hours in 1:4 glycerol/acetic acid. They were then transferred onto a glass slide, oriented, mounted in 1:1 Hoyer's medium/lactic acid, and dried overnight at 65°C (30).

RNA sample preparation and sequencing

For each of three sample sets, approximately ten *Chironomus riparius* embryos at the "two-pole-cells" stage from the same egg package were bisected into anterior and posterior halves on dry ice using a razor blade or scalpel. Bisected halves were separated and immediately mixed with TRIzol reagent (Life Technologies, NY, USA), incubated for 5-10 minutes at room temperature, and frozen at -80° C. RNA was extracted from samples using the TRIzol protocol scaled to 50 μ L and total RNA was used for cDNA library construction using the TruSeq kit with PCR (Illumina, CA, USA) (conducted at the University of Chicago genomics core facility). cDNA libraries were barcoded and multiplexed for 100bp paired-end sequencing on one lane of a HiSeq Illumina 2000 sequencer. Sequencing data generated from this work are available at the NCBI Sequence Read Archives (www.ncbi.nlm.nih.gov/sra) with the project accession PRJNA229141.

RNA-seq data preprocessing

Raw read data was examined using FASTQC v0.10.1 (31) to determine sequence quality, overrepresented sequences, and other quality control metrics. CUTADAPT v1.1 (32) was used to remove barcodes and contaminating adapter sequences identified by at least 5 bp of similarity in flanking sequences. Paired end reads that overlapped were merged into single reads using

FLASH v1.2.2 (33). Low quality reads were removed and orphaned reads were separated from paired-end reads in the data sets using SICKLE (34).

Transcriptome assembly and annotation

The preprocessed RNA-seq read data were assembled using the de Bruijn di-graph networks (35) implemented in the ABYSS v1.3.5 assembler (36, 37). Tiled sub-read (kmer) lengths were varied from 20-55 and the transcriptome assembly with the greatest N50 was chosen, corresponding to a kmer length of 25 nucleotides.

Coding sequences and Ensembl gene IDs for *T. castaneum* (Tcas3) and *D. melanogaster* (BDGP5) were downloaded using BIOMART v2.14.0 (38). A BLAST database was created from these files and subsequent BLAST procedures were completed using BLASTSUITE v34 tools (39). Annotation was conducted in three steps. First, orthology searches were completed for every transcriptome contig in the Tribolium and Drosophila annotated databases using the alignment algorithm in tBLASTx, and a maximum e-value of 0.01. For a given *C. riparius* contig, the BLAST hits from the reference databases were merged if they existed within 20 basepairs of each other. For each *C. riparius* contig, the largest difference in consecutive hits based on \log_2 (e-value) was chosen as the cutoff and hits with e-values lower than the cutoff were retained. For contigs with greater than one BLAST hit after merging and filtering by e-value, a reciprocal tBLASTx was used for each hit from the respective database such that the hit with the lowest mean e-value for blast and reciprocal blast hits was chosen for the annotation. In the second step, the largest ORF from each remaining unannotated contig was searched in the Drosophila and Tribolium annotated databases using tBLASTn. In the third step, the remaining unannotated contigs were searched in the NCBI protein reference sequence database (RefSeq, 04-24-2013) using the above approach for the largest ORFs in the contigs. Transcript annotations are provided in the assembled transcriptome fasta file with the Ensembl gene IDs, gene descriptions, and either BLAST or average reciprocal BLAST (rb) e-values. Transcript annotations using the RefSeq database are provided in standard RefSeq format with either BLAST or reciprocal BLAST e-values.

RNA-seq alignment and differential expression

BOWTIE2 v2.0.0 (40) was used to align reads from each sample to the transcriptome assembly. The SAMTOOLS v0.1.18 suite (41) was used for sorting, converting, removing duplicate, and indexing reads. Read information was extracted using SAMTOOLS.

Differential expression analysis was conducted in R and sequence file manipulation was handled with the Bioconductor (42) package Biostrings v2.26.3 (43). Transcripts 200nt or shorter or with fewer than five reads in any of the three anterior-posterior sample pairs were excluded. To increase statistical sensitivity, transcripts that were unannotated or annotated with “ribosomal”, “hypothetical”, or “predicted” terms were also removed (N = 1,149) such that 6,604 annotated transcripts remained in the analysis. \log_2 RPKM (reads per kilobase of transcript per million mapped reads) values were calculated. Unsupervised hierarchical clustering with average linkage and Euclidean distance was conducted with variably expressed genes (IQR > 1.5) using GENEFILTER v1.40.0 (44) with bootstrapped node support (N = 10,000) by PVCLUST v1.2-2 (45). Samples were compared in a pairwise fashion for each biological replicate rather than grouping according to condition (A versus P) because much of the variability in the data appeared to be between biological replicates. A linear fit for each anterior-posterior sample pair of log-2 RPKM data was computed and the residuals were weighted by

mean RPKM values and Z-score normalized. To compare the weighted residual (differential expression score) for a given gene across samples, a bootstrapped t-statistic was used. P-values were thus generated by comparing the t-statistic for a given gene to a distribution of t-statistics from randomly chosen weighted and normalized residuals from the data (three at time, with replacement, N = 1,000).

Protein alignment and transcript mapping

Protein alignments were conducted in STRAP v1.0(46) using MUSCLE(47). *Pangolin/TCF4* homologous sequences from *T. castaneum* (AY800247.1), *D. melanogaster* (NP_726527.2) and *A. gambiae* (XP_320170.4) along with translated transcripts from *C. riparius* RNA-sequencing for *Pangolin* and *Panish* were used for the protein sequence alignment. Alignments were plotted with JALVIEW v2.8 (48) according to average distance using the BLOSUM62 substitution matrix. Transcripts were mapped to a *C. riparius* genomic contig sequence containing the *Cri-zap3* - *panish* locus using BLASTn. Open reading frames were predicted to be the longest uninterrupted coding sequences (following Methionine) within each transcript from the *C. riparius* transcriptome assembly.

Determination of phylogenetic occurrence of bicoid and panish in dipteran genomes

Bicoid homeoboxes were identified in dipteran genomes by calculating a position weight matrix using four iterations of PSI-BLAST (49) and up to 100 hits with the homeodomain: PRRTRTFTSSQIAELEQHFQGRYLTAPRLADLSAKLALGTAQVKIWFKNRRRRRHKIQS (4). The resulting position weight matrix was used in a pBLAST of each genome. A similar approach was used for the *panish* C-clamp. *D. melanogaster Antennapedia* complex genes, *pangolin*, and *Glut4EF* sequences were obtained from Ensembl Metazoa using BiomaRt. The top hit for each gene isoform was mapped to each dipteran genome using tBLASTn and isoform hits were merged if overlapping. Regions containing either a putative *bicoid*-like homeobox or *panish*-like C-clamp domain were investigated further if any of the *Antennapedia* complex genes or C-clamp genes, respectively, also mapped to them. These regions were extended to contain up to 50 additional residues on each flank, ignoring stop codons to allow for the possibility of spurious assembly artifacts. The potential homeobox and C-clamp containing regions were then submitted to a reciprocal-BLAST in the *D. melanogaster* transcriptome (BDGP5) and further interrogated in Dipteran protein sequences included in the the NCBI nr-protein database to identify the closest homologous gene match. Top hits were plotted on genomic contigs using the R package circlize (50).

Organisms known to possess the *bicoid* gene are represented in the following seven groups:

Multiple species (Drosophilidae)

Episyrphus balteatus, HM044914 (Syrphidae)

Lonchoptera lutea, EU589575 (Lonchopteridae)

Platypeza consobrina, EU589580 (Platypezidae)

Megaselia abdita, AJ133024 (Phoridae)

Musca domestica, AJ297850-AJ297854 (Muscidae)

Multiple species (Calliphoridae)

Note: where required, transcriptomes were assembled as detailed above.

Aedes aegypti (Culicidae): AAGE02

Anopheles species (Culicidae): APCK01, APCN01, APCM01, ADMH01, APCL01, APCJ01, APCI01, ABKP02, ABKQ02, AAAB01, APHL01, AT LZ01, APCH01, AT LV01, ALPR02
Bactrocera species (Tephritidae): JHQJ01.1, *B. oleae* transcripts (SRA Accession SRX265051)
Belgica antarctica (Chironomidae): JPYR01 and transcripts GAAK01
Ceratitis capitata (Tephritidae): AOHK01
Chironomus piger (Chironomidae): Thomas Hankeln, personal communication
Chironomus tentans (Chironomidae): CBTT01
Culex quinquefasciatus (Culicidae): AAWU01
Drosophila species (Drosophilidae): ACVV01, AAPP01, AFFD01, AFFE01, AFFF01, AAPQ01, AFPQ01, AFFG01, AAPT01, AFFH01, AABU01, AJMI01, AAPU01, AAIZ01, AADE01, AFPP01, AAKO01, GAHN01, AAGH01, CAKG01, AFFI01, AANI01, AAQB01, AAEU02
Glossina species (Glossinidae): *G. austeni*, *G. brevipalpis*, *G. fuscipes*, *G. morsitans*, *G. pallidipes* <https://www.vectorbase.org/>, and *G. morsitans* transcripts
Hermetia illucens (Stratiomyidae): see (51) and SRA Accession SRP021047
Lutzomyia longipalpis (Psychodidae): AJWK01
Mayetiola destructor (Cecidomyiidae): AEGA01
Musca domestica (Muscidae): AQPM01
Phlebotomus papatasi (Psychodidae): AJVK01
Sitodiplosis mosellana (Cecidomyiidae): GAKJ01

Fig. #:

Figure S1. Occurrence of *bicoid*-like homeoboxes in two Tephritidae genomes. Outer ring: Shaded boxes indicate genomic regions matching the *bicoid* homeobox position weight matrix. Red lines indicate gaps in the contig greater than 500 bp that are not shown. Inner ring: Annotation of each shaded region of outer ring. Also included are best BLAST hits for *Antennapedia* complex genes, and *bicoid* from *E. balteatus*, *M. domestica*, and *M. abdita*. For a given contig, strongest BLAST matches are on the outer edge. Full contig names are provided as Supplementary Data 1.

Figure S2. Occurrence of *bicoid*-like homeoboxes in five Glossinidae genomes. Outer ring: Shaded boxes indicate genomic regions matching the *bicoid* homeobox position weight matrix. Red lines indicate gaps in the contig greater than 500 bp that are not shown. Inner ring: Annotation of each shaded region of outer ring. Also included are best BLAST hits for *Antennapedia* complex genes, and *bicoid* from *E. balteatus*, *M. domestica*, and *M. abdita*. For a given contig, strongest BLAST matches are on the outer edge. Full contig names are provided as Supplementary Data 2.

Figure S3. RNA in situ hybridization for *Cri-nanos* in *C. riparius*. Anterior is to the left and dorsal is up; scale bar, 10 μ m.

Figure S4. Full sequence alignment for Pangolin homologs and Panish. Amino acid sequence alignment of Panish with Pangolin sequences from *T. castaneum* (Tca), *D. melanogaster* (Dme), *C. riparius* (Cri), and *A. gambiae* (Aga). Alignments were conducted with MUSCLE multiple sequence alignment algorithm. Conservation score is a measure of conserved physico-chemical properties and blue shading intensity indicates relative BLOSUM62 conservation score.

Figure S5. RNA in situ hybridization for *Cri-pangolin* in *C. riparius*. Anterior is to the left and dorsal is up; scale bar, 10 μ m.

Figure S6. Reduction of *panish* transcript and intermediate phenotypes from *panish* mRNA injection following *panish* RNAi. **A)** RNA *in situ* hybridization of *panish* 60 minutes after injection of *panish* dsRNA (compare to Fig. 2B, first panel). Controls exhibited normal *panish* expression. *Panish* transcript was evident infrequently (second panel). Anterior is to the left; scale bar, 10 μ m. **B, C)** Representative dark field (**B**) and bright field (**C**) images of incomplete larval cuticles obtained in 13 out of 219 *panish* mRNA injections into *panish* RNAi embryos. **D)** Close-up view of deformed larval head structures.

Figure S7. Expression of *Cri-hb* and *Cri-oc* and bright field images of representative cuticles following *Cri-hb*, *Cri-oc*, and *Cri-cad* RNAi. **A)** RNA *in situ* hybridization of *Cri-hb* at blastoderm stages before, during, and after cellularization, in lateral view. **B)** RNA *in situ* hybridization of *Cri-oc* during blastoderm cellularization, in lateral (left and middle panels) and dorsal views (right panel). Scale bars, 10 μ m. **C)** *Cri-hb* RNAi cuticle. Note that body appears similar to wild-type (WT) (compare to Figs. 3A, 4F) but possesses duplicated parapods in place

of maxillae. **D)** *Cri-oc* RNAi cuticle. Note that posterior body appears similar to wild-type (compare to Fig. 3A) but head structures are disrupted or absent (compare to Fig. 4F). **E)** *Cri-cad* RNAi cuticle. Anterior is to the left in all images.

Figure S8. Loss of *Cri-tll* transcript and double-head formation following *Cri-tll* RNAi. **A)** *Cri-tll* RNA *in situ* hybridization of a wild-type embryo at the blastoderm stage. **B)** Double RNA *in situ* hybridization of a *Cri-tll* RNAi embryo with *Cri-nos* and *Cri-tll* probes, showing reduced *Cri-tll* expression and normal *Cri-nos* expression (compare to Fig. S3 for *Cri-nos* wild-type expression). **C, D)** Dorsal view of live wild-type and *Cri-tll* RNAi embryos. Anterior is to the left; scale bars, 10 μ m.

Figure S9. *Cri-hkb* and *Cri-tor* expression and larval cuticles of *Cri-tor* RNAi and *panish*/*Cri-tll* double RNAi embryos. **A)** *Cri-hkb* RNA *in situ* hybridization of pre-blastoderm and blastoderm embryos compared to an extended germband stage. **B)** *Cri-tor* RNA *in situ* hybridization of pre-blastoderm and blastoderm embryos. Anterior is to the left; scale bars, 10 μ m. **C)** *Cri-tor* RNAi larval cuticles showing shortened body and deformed head. Note the lack of terminal structures, including missing labrum at the anterior side (compare to Fig 4F) and missing posterior parapods and anal setae (compare to Fig 3A). **D)** Representative dark field image of double-abdomen larval cuticle obtained from RNAi against both *panish* and *Cri-tll*.

Figure S10. *Panish* operates differently than *Drosophila bicoid* and is conserved in Chironomus. **A)** Comparison of proposed Chironomus and established *Drosophila* axis specification models. **B)** Protein alignment based on *panish* cDNA sequences from *C. riparius* and *C. tentans* and predicted *panish* transcript sequence from *C. piger* genomic sequence. Conservation score is a measure of conserved physico-chemical properties, blue shading intensity indicates the relative BLOSUM62 conservation score, and red lines indicate exon boundaries.

Figure S1

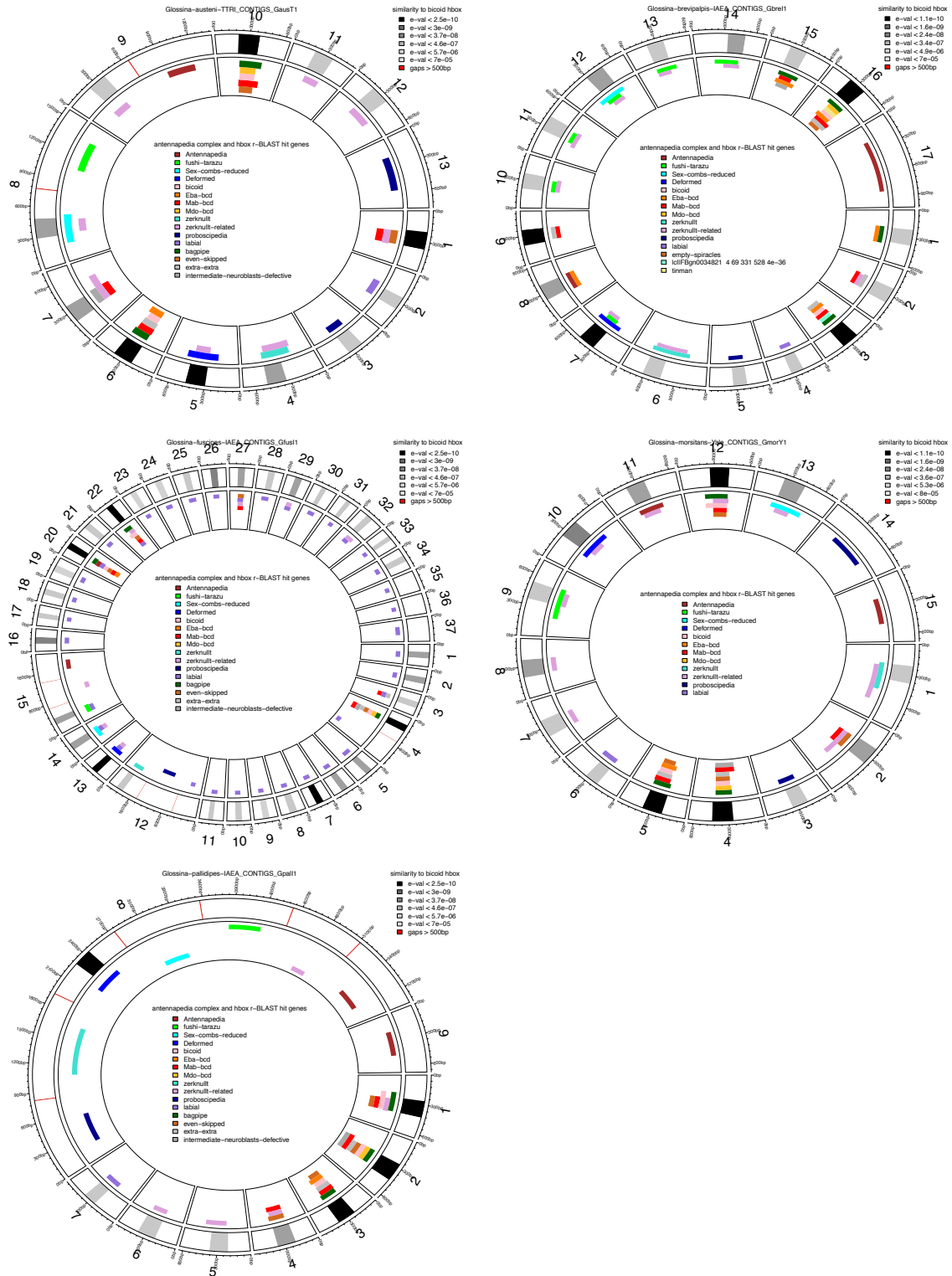


Figure S2

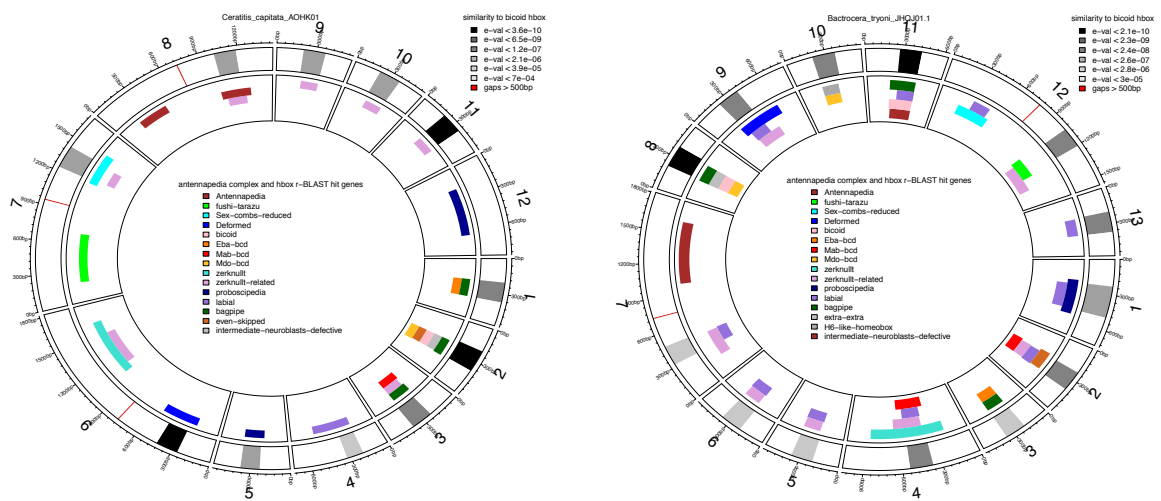


Figure S3

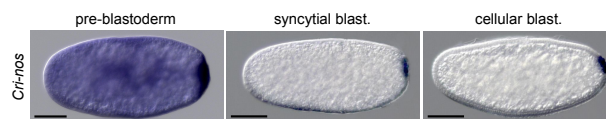


Figure S4

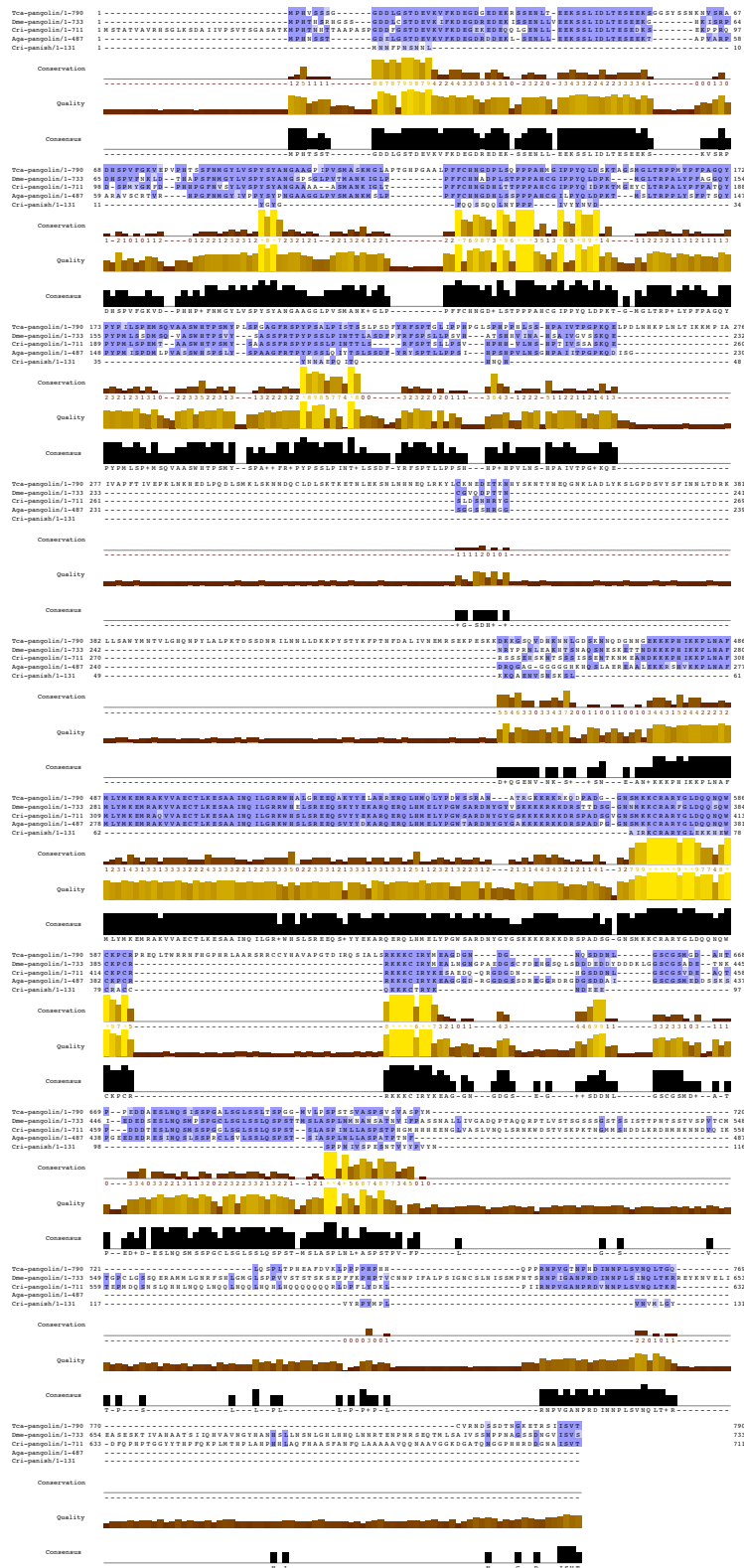


Figure S5

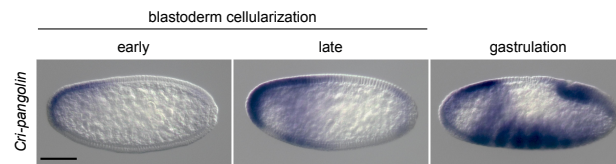


Figure S6

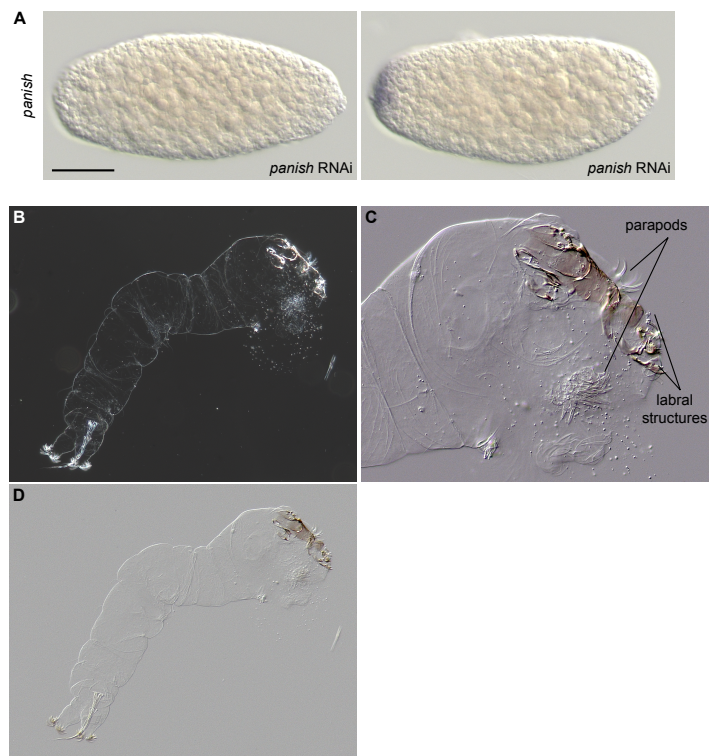


Figure S7

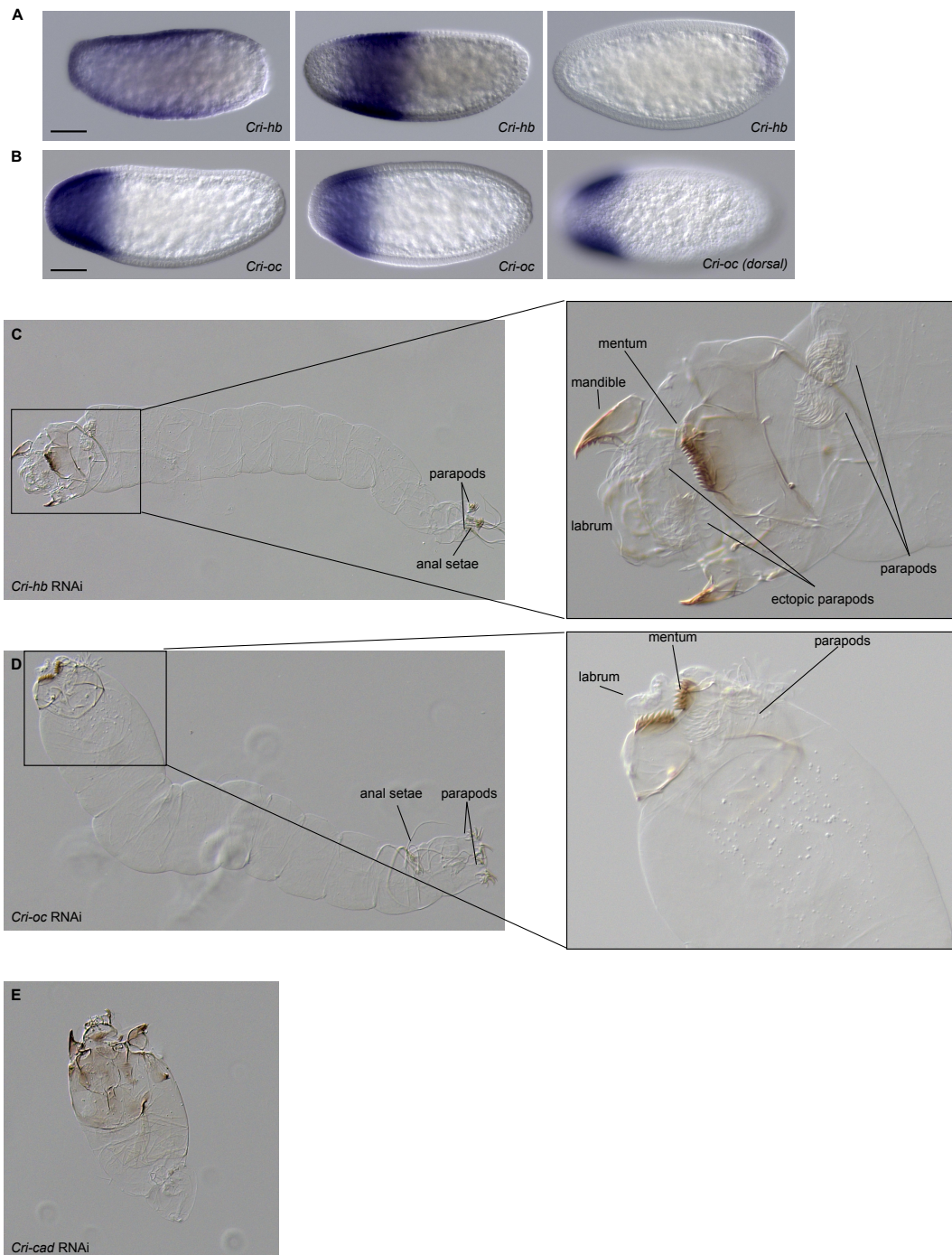


Figure S8

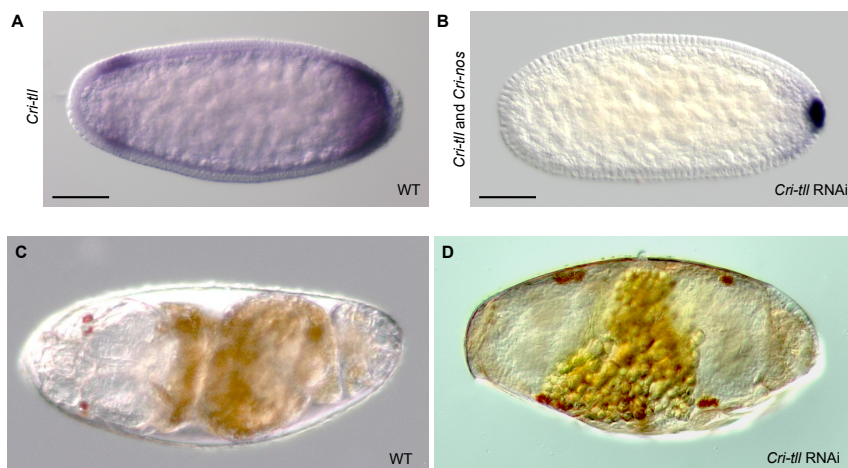


Figure S9

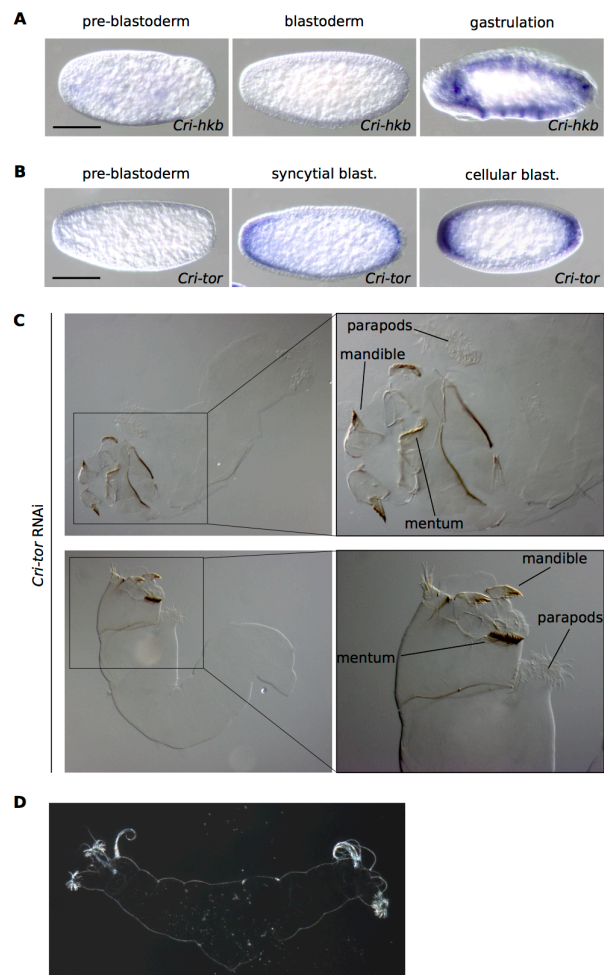


Figure S10

