Cloning and Expression Analysis of the Engrailed-family Genes from *Pedetontus unimaculatus* Machida (Insecta: Archaeognatha)

A Dissertation Submitted to
the Graduate School of Life and Environmental Sciences,
the University of Tsukuba
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Science
(Doctoral Program in Structural Biosciences)

Yasutaka NAKAGAKI

Contents

Abstract

It has been shown that segmentation in the short-germ insects proceeds by a two-step mechanism. The anterior region is simultaneously segmented in a manner similar to that in *Drosophila*, which is apparently unique to insects, and the rest of the posterior region is segmented sequentially by a mechanism involving a segmentation clock, which is derived from the common ancestor of arthropods. In order to propose the evolutionary scenario of insect segmentation, I examined segmentation in the jumping bristletail, the basalmost extant insect. Using probes for engrailed-family genes for *in situ* hybridization, I found no sign of simultaneous segmentation in the anterior region of the jumping bristletail embryos. All segments except the anteriormost segment are formed sequentially. This condition shown in the jumping bristletail embryos may represent the primitive pattern of insect segmentation. The intercalating formation of the intercalary segment is assumed to be a synapomorphic trait shared by all insects after the branching of the jumping bristletail, and an autoapomorphy of Dicondylia.

An evolutionary scenario of engrailed-family gene in arthropods is also discussed in present study. Paired engrailed-family genes have been isolated from most lineages of insects. I also found a pair of the genes from jumping bristletail. From the point of parsimony, it is reasonable to suppose single duplication in the ancestor, but phylogenetic analyses of engrailed-family genes in insects have always demonstrated multiple occurrence of duplication. Some previous studies proposed a possibility that gene

conversion occurred between the pair of these genes misleads the output of phylogenetic analyses. I examined engrailed-family genes in vast lineages of insects and in several non-insect arthropods by using a statistical method and found traces of gene conversion between the pairs of the genes in some insects. Phylogenetic analyses using different regions of the gene suggested that the phylogenetic trees were affected severely by gene conversion. In addition, comparison of conserved features in arthropod engrailed-family genes showed a possibility that gene duplication causing paired engrailed-family genes occurred prior to the branching of hexapods including insect lineage from a crustacean lineage.

General Introduction

Comparison between model animals has shown astonishing universality in genetic mechanism of morphogenesis of several structures. In case of segmentation, however, it does not go very well.

The most knowledge on genetic mechanism of segmentation is amassed in *Drosophila* among animals. In most insects including *Drosophila*, after the fertilized egg repeats nuclear division and becomes into syncytium, most nuclei migrate to the periphery to form blastoderm, and then the cellularization occurs. In *Drosophila*, embryo is subdivided by gap and pair-rule proteins, which are free to defuse in syncytial environment, and all segments are simultaneously determined in blastoderm stage. In contrast, segment formation of vertebrates and annelids progresses sequentially from anterior to posterior in cellularized embryo. Therefore, it is hard to assume that *Drosophila*-like molecular mechanism works in segmentation of these animals. It seems to be natural to consider that the fundamental difference in the regulatory mechanism means independent origins of segmentation in different phyla. In fact, segmentation in *Drosophila* is exceptional even in arthropods and cannot be a suitable model of arthropod segmentation for comparing segmentation among animal phyla.

Morphologists have categorized segmentation pattern of insects into short-, semi-long- (or intermediate-), and long-germ types (Krause 1939; Sander 1976) (Fig. 1). In long-germ type, all segments develop simultaneously as observed in *Drosophila*.

Long-germ type is apparently a derived trait, distributed in part of holometabolous insects. Other insects show short- or semi-long-germ types. In short- and semi-long-germ type, shorter germ rudiment is formed than in long-germ type and posterior segments are sequentially added as the germ-band elongates. Segments in the region of germ rudiment develop simultaneously prior to posterior sequential segmentation. Short-germ type bears particularly short-germ rudiment, which seems to contain only head segments and semi-long-germ type bears longer germ rudiment containing head to thoracic or to anterior abdominal segments. Therefore, there is no definite division between short- and semi-long-germ types. Though various segmentation patterns found in the other arthropods are less systematically interpreted than in insects, they are similar in that more or less segments arise sequentially. Segmentation in arthropods excluding long-germ insects, in which segments are sequentially produced in posterior growth zone, morphologically resembles vertebrate segmentation (or somitogenesis), in which somites develop sequentially in presomitic mesoderm.

Recently, genetic mechanism of segmentation has been examined in several non-model insects and arthropods, and it is revealed that similarity in segmentation mechanism between distinct phyla, which is lost in *Drosophila*, and difference between short-/semi-long- and long-germ types. Somitogenesis in vertebrate is controlled by a genetic mechanism named segmentation clock, in which cyclic expression of *Delta* and *hairy* homologues is observed under *Notch* signaling pathway. In *Drosophila* segmentation, *Notch* signaling pathway is not involved and, no cyclic gene expression observed. Stollewerk et al. (2003) examined genetic mechanism of segmentation in spider abdomen (opisthosoma) and reported its striking similarity with that of vertebrate somitogenesis. Expression of *Delta* and *hairy* homologues in spider appears and then clears repeatedly in the posterior growth zone. In addition, RNA interference targeting of *Notch* or *Delta* homologues leads to severe defects in segmentation. Peel and Akam (2003) predicted that a mechanism similar to segmentation clock plays a role in the sequential segmentation of short- and semi-long-germ insects, and proposed a hypothesis on evolutionary transition from short- to long-germ segmentation that segments patterned by a clock have been progressively reduced in number during evolution, being replaces with mechanisms analogous to those already operating in the anterior region. An examination of segmentation in cockroach (Pueyo et al. 2008) supported Peel and Akam's (2003) hypothesis. In this short-germ insect, the posterior sequential segmentation was inhibited by interference for *Notch* signaling but the anterior simultaneous segmentation was less affected. Additionally, recent studies reported cyclic expression of homologues of *Drosophila pair-rule* genes controlling segmentation in a short-germ insect, *Tribolium*, and suggested that clock mechanism underlies also in the insect segmentation.

From the point of the morphology of insect segmentation, Peel and Akam's (2003) hypothesis is tempting because it reasonably explains short-, semi-long-, and long-germ variation found in insect segmentation. According to the context of this hypothesis, plesiomorphic state in insect segmentation should exhibit the largest number of sequentially formed segments. Morphological studies have reported jumping bristletails (order Archaeognatha), which is the basalmost of extant insects, represent "extremely" short-germ segmentation (Larink 1969; Machida 1981). No segmental elements are observed in its germ rudiment, a tiny circular germ disc, and most segments appear sequentially. Inferring from the phylogenetic position and the segmentation pattern, jumping bristletails could be expected to represent the ancestral pattern of insect segmentation.

Because genetic background of animal segmentation is being revealed and the understanding of the origin of segmentation is changing, insect segmentation pattern, that had been interpreted morphologically, is required to be reevaluated. In addition, because a root of the phylogenetic tree of the insects has not been confirmed yet, information for the basal insects is important for discussing what the start point of evolutionary course of insect segmentation is and how the insect segmentation pattern derived in the evolutionary transition of arthropods.

In the present study, the expression pattern of engrailed-family genes, which are frequently used as segment marker for arthropods, was examined in the embryo of jumping bristletails for the purpose of making documentation comparable with the data accumulated in other insects and arthropods. Then I made discussion about the ancestral state and the evolutionary process of insect segmentation by comparison of the information from jumping bristletail and accumulated in other insects.

Besides, it was revealed that jumping bristletails possess two engrailed-family paralogs in the course of the gene cloning for investigating segmentation process in this insect. Paired engrailed-family paralogs are reported in many other insect species. Several studies have discussed how the paired engrailed-family paralogs in various lineages of insects had arisen. Sequence motifs shared in only one of the each pair of paralogs suggest duplication in common ancestor, while the topology of the trees drawn by the molecular phylogeny of the genes shows independent duplications in multiple lineages. Peel et al. (2006) showed a tendency that the similarity of gene sequence between the pair of the paralogs in a species is higher in the 3' region than in the 5' region. The authors pointed out that gene conversion can cause the unusual similarity and make the apparent evolutionary distance between the paired paralogs in a species shorter. They proposed that recurrent gene conversion in the evolution of insect engrailed-family genes leads to overestimate the frequency of the gene duplication on the phylogenic tree.

I analyzed engrailed-family genes in arthropods, adding the information of the genes becoming newly available in jumping bristletails, to detect gene conversion occurring in arthropod lineages and made a discussion on the duplication of the genes in arthropods.

Chapter 1: Expression of engrailed-family genes in the jumping bristletail and discussion on the primitive pattern of insect segmentation

Introduction

In insects (I use the word “insects” for Insecta *s. str.*, or Ectognatha), segmentation patterns are often categorized as either “long-germ” or “short-germ” segmentation (Krause 1939; Sander 1976). Long-germ segmentation only occurs in some holometabolous insects, which are the most derived group of insects. Other insects including apterygote (jumping bristletails [Archaeognatha], and silverfish, firebrat and bristletail [Zygentoma]), hemimetabolous and some holometabolous insects exhibit short-germ segmentation patterns.

In the long-germ insects, almost all segments simultaneously develop in the germ band directly formed from the blastoderm. In contrast, short-germ segmentation is characterized by sequential development of most segments from anterior to posterior as the germ band elongates. Strictly, not all segments are, however, formed sequentially in the short-germ insects, and a few or more anterior segments have already developed simultaneously before sequential segmentation and germ-band elongation begin (Rogers and Kaufman 1996; Peterson et al. 1998; Pueyo et al. 2008).

Among the short-germ insects, the number of simultaneously developing segments varies according to lineages. Germ types in which not all but more segments develop simultaneously are called semi-long-germ type or intermediate germ type by some authors (Krause 1939; Sander 1976). Because there is no fundamental difference between short-germ and semi-long-germ types, semi-long-germ type is not discriminated from short-germ type and only "short-germ" is used in the present study.

The genetic mechanism of insect segmentation has been studied in detail for *Drosophila*, one of the most important model organisms. Therefore, much knowledge about the genetic mechanism of long-germ segmentation has been amassed. On the other hand, studies on short-germ segmentation involving non-model insects have gradually revealed differences between short-germ and long-germ segmentations. Recent studies suggest that the segmentation clock, a genetic mechanism that generates reiterated patterns, is involved in sequential segmentation of short-germ insects (Pueyo et al. 2008; Sarrazin et al. 2012; El-Sherif et al. 2012). The segmentation clock was originally discovered as a regulator of somitogenesis in vertebrates (Pourquié 2003) and later described in a report on the segmentation clock in abdominal segmentation of spiders (Stollewerk et al. 2003). Somitogenesis in vertebrates, the segmentation in abdomen of spider, and the segmentation in the short-germ insects are similar in that metameric body patterns are sequentially developed. The segmentation clock is not involved in the segmentation of the long-germ insects *Drosophila* or honeybees (Wilson et al. 2010). The sharing of the segmentation clock in spider and short-germ insects suggests that the sequential segmentation found in the short-germ insects is a plesiomorphic trait and that lack of the segmentation clock is an apomorphic trait among insects (Pueyo et al. 2008; Sarrazin et al. 2012).

Notably, the short-germ insects possess anterior segments that form simultaneously and posterior segments that form sequentially, and the numbers of segments that form non-sequentially and sequentially differ depending on the lineage. For example, expression studies of the engrailed-family genes, a widely conserved group of genes involved in segmentation, have demonstrated that sequential segmentation begins with the labial segment in firebrat (an apterygote insect), the first abdominal segment in cricket (a hemimetabolous insect), and the seventh abdominal segment in flea (a holometabolous insect) (Rogers and Kaufman 1996; Peterson et al. 1998). The evolutionary transition from short-germ type to long-germ type is assumed to have involved the gradual decrease in number of sequentially formed segments and the resultant increase in number of non-sequentially formed segments (cf. Peel and Akam 2003).

Jumping bristletails or Archaeognatha are the basalmost taxon of insects (cf. Misof et al. 2014). Morphological studies have demonstrated that jumping bristletails possess an "extremely" short-germ pattern (Larink 1969; Machida 1981). No segmental elements are observed in the earliest germ band, a tiny circular germ disc, and most segments appear sequentially as the germ band elongates. Following the inference on the evolution of insect segmentation mentioned above, the number of segments that form sequentially should be larger for the basalmost insects than for other insects; notably, the segmentation pattern seen in the jumping bristletail is consistent with this hypothesis. Thus, the jumping bristletail has presumably retained the ancestral pattern of insect segmentation.

Morphological studies of jumping bristletail embryos indicate that sequential segmentation can be observed in the segments posterior to the labial segment (Larink 1969; Machida 1981). However, details are ambiguous for structures anterior to the labial segment because the segmental borders in the tiny embryos that represent the earliest stages of segmentation are too faint to determine the order of segmentation based only on morphological observation of external structures. Recent investigations of insect segmentation patterns often employ examination of the expression pattern of engrailed-family genes (or their protein products) because these genes are assumed to play a critical role in insect segmentation. *engrailed* is known as a segment polarity gene of *Drosophila*, a factor that determines anterior-posterior polarity in each segment, and shows a reiterated pattern of expression in the posterior portion of each segment (Kornberg 1981). This segmentally reiterated pattern of engrailed-family gene expression is broadly conserved in arthropods, including insects. Expression patterns of engrailed-family genes have been documented for many arthropod species, and abundant information is available, especially for various insect taxa. Documentation of the expression pattern of an engrailed-family gene in the jumping bristletail is crucial not only for elucidating the segmentation pattern of this group, but also for discussing the evolutionary transition of insect segmentation.

I examined the expression of engrailed-family genes during the segmentation process in the jumping bristletail, focusing on the anterior segments. I then discussed the evolutionary transition of segmentation pattern in insects.

Materials and methods

1. Animals

I collected adults of the jumping bristletail, *Pedetontus unimaculatus* Machida, 1980, in Shimoda, Shizuoka, Japan (34°40'08"N 138°56'34"E). Eggs of jumping bristletails were collected as described by Machida (1981). Developmental staging of embryos was also performed as described by Machida (1981). Embryos for SEM were processed according to Machida (2000).

2. Cloning of engrailed-family genes from the jumping bristletail

ISOGEN (Nippon Gene) was used to extract total RNA from eggs of jumping bristletails. First-strand cDNA Synthesis Kit (Amersham) was used to reverse-transcribe first-strand cDNA and synthesize double-stranded cDNA, which was then used as a template for degenerate PCR. Fragments of engrailed-family genes were amplified by PCR with the following degenerate primers: 5'-TGGCCNGCNTGGGTNTAYTGY-3' (outside 5' primer), 5'-RTTRTANARNCCYTGNGCCAT-3' (outside 3' primer), 5'-GYACRMGVTAYTCVGAYNGRC-3' (inside 5' primer) and 5'-TVGCNCKYTTRTTYTGRAACC-3' (inside 3' primer). TOPO TA Cloning® Kit

(Invitrogen) was used to clone the products of degenerate PCR. FirstChoice® RLM-RACE Kit (Ambion) was used for the 5' and 3' RACE experiments.

3. Whole-mount *in situ* hybridization

Before the dissection of eggs, pre-treatments were performed as described by Machida (1981). Eggs were soaked in PBT (phosphate-buffered saline [PBS] containing 0.1% Triton X-100) during dissection, and fine forceps were used to remove embryos from the eggs. The isolated embryos were fixed with 8% paraformaldehyde in PBS for 2~3 days at room temperature. Then, they were carefully washed in PBT, gradually dehydrated using methanol and stored in methanol at -20℃.

After gradual hydration with PTw (PBS contains 0.1% polyoxyethylene [20] sorbitan monolaurate), the embryos were partially digested in 4 μg/ml Proteinase K (Invitrogen) for 20~40 minutes at room temperature. Developed embryos that had already secreted embryonic cuticle were treated longer than younger ones with the embryonic cuticle not yet secreted. After the proteinase reaction, the embryos were washed in PTw before being fixed again for 20 minutes in 4% paraformaldehyde-0.2% glutaraldehyde that was dissolved in PTw. Embryos were washed again in PTw, and the specimens were then incubated in hybridization solution (50% formamide, 5 x SSC, 2% blocking reagent [Roche Diagnostics], 0.1% Triton X-100, 0.1% CHAPS, 1 μg/ml yeast tRNA, 2 mM EDTA·2Na, 50 μg/ml heparin sodium salt) at 60℃ for at least 1 hour.

DIG RNA Labeling Mix (Roche Diagnostics) was used according to the manufacturer's instructions to produce digoxigenin (DIG)-labeled antisense RNA probes. Embryos were treated with the probes overnight at 60ºC. The probes were removed by washing the embryos three times in 2 x SSCC (SSC containing 0.1% CHAPS) and then three times in 0.2 x SSCC; each wash was 20 minutes at 60ºC. Embryos were then washed three times in KTBT (50 mM Tris-HCl [pH 7.5], 150 mM NaCl, 10 mM KCl, 0.1% Triton X-100) for 10 minutes each at room temperature and then treated with KTBT containing 1.5% blocking reagent for 3 hours at room temperature. Embryos were then incubated overnight at 4ºC in blocking buffer containing a 1:2500 dilution of alkaline-phosphatase-conjugated anti-DIG Fab fragments (Roche Diagnostics). Next, embryos were washed for an hour with several changes of KTBT and then with three 10-minute washes in NTMT (0.1 M NaCl, 0.1 M Tris-HCl [pH 9.5], 50 mM $MgCl_2$, 0.1% Triton X-100); embryos were incubated in NTMT containing 2% BCIP/NBT stock solution (Roche Diagnostics). The development reaction was stopped with three washes in PBT. The specimens were soaked in 80% glycerol overnight and then mounted on microscope slides in 80% glycerol. Specimens were observed with a DM-6000B microscope (Leica Microsystems).

Results

1. Isolation of two engrailed-family genes from the jumping bristletail

Two distinct fragments (197 bp and 203 bp) were amplified from the jumping bristletail via degenerate PCR targeting engrailed-family genes. I named the corresponding genes *Pu-en1* and *Pu-en2*. I carried out 5' and 3' RACE experiments to determine the entire sequences of these engrailed-family genes.

The reconstituted *Pu-en1* (accession number LC031803) cDNA was 1297 bp long and contained an open reading frame predicted to encode a protein of 263 amino acids (Fig. 2). *Pu-en2* (accession number LC031804) cDNA with at least 2255 bp contained an open reading frame predicted to encode a protein of 314 amino acids (Fig. 2). Because neither an upstream stop codon nor a convincing Kozak consensus sequence (Kozak 1986) was found in the *Pu-en2* cDNA sequence, I could not dismiss the possibility that *Pu-en2* extended further in the 5' direction.

On the basis of the amino acid sequences deduced from the cDNA sequences, Pu-en1 and Pu-en2 each possessed all five domains that are conserved among vertebrate and arthropod engrailed homologues and called engrailed homology regions (Logan et al. 1992; Peel et al. 2006). Predicted Pu-en1 and Pu-en2 sequences could be aligned in the engrailed homology regions. In these regions, Pu-en1 and Pu-en2 shared 92.2% amino acid identity.

2. Embryonic expression of *Pu-en1* and *Pu-en2*

I synthesized probes for *in situ* hybridization in the cDNA regions at the nucleotide positions from 96 through 502 in *Pu-en1* and from 57 through 813 in *Pu-en2* (Fig. 2). Using these probes, I analyzed the expression patterns of *Pu-en1* and *Pu-en2* by whole-mount *in situ* hybridization.

The expression pattern of *Pu-en2* mRNA was analyzed for different developmental stages. In early stage 1, when the circular germ disc has just formed from the blastoderm, the expression of *Pu-en2* was not evident (Fig. 3a). The earliest expression of *Pu-en2* was observed in late stage 1 when the embryos are slightly elongated, assuming a triangular shape: *Pu-en2* mRNA formed a single stripe across the embryo (highlighted by an asterisk; Fig. 3b). In stage 2, the germ disc elongated to assume a pear-shape, and newly expressed *Pu-en2* mRNA formed a stripe located posterior to the original *Pu-en2* stripe and, separately, a chevron pattern with a midline gap located anterior to the original stripe (Fig. 3c). In stage 3, segmental boundaries became distinct externally (SEM photo in Fig. 3d), and *Pu-en2* was expressed in the posterior region of the antennal (AnS), intercalary (IS), mandibular (MdS) and maxillary segments (MxS) (Fig. 3d): the stripe in the maxillary segment was the last to appear; the original *Pu-en2* stripe that first appeared in late stage 1 (Fig. 3b) and the *Pu-en2* stripes expressed in stage 2 (Fig. 3c) respectively represent the intercalary segment and the antennal, intercalary and mandibular segments. In stage 3, a pair of regions with faint expression (arrow) in a chevron pattern also appeared anterior to the antennal expression (Fig. 3d). In stage 4 when the segmentation of thoracic segments starts, *Pu-en2* pattern elements were observed in the pre-antennal region (arrow) and in the posterior regions of the antennal, intercalary, mandibular, maxillary, labial and prothoracic segments (Fig. 3e): the *Pu-en2* expression in the pre-antennal region became more obvious, and this pattern element seems to be serially homologous to the antennal chevron pattern. In stage 6 when abdominal segmentation starts, *Pu-en2* was expressed in the pre-antennal region (arrow in Fig. 3f), head segments (from antennal to labial), three thoracic segments and the first two abdominal segments (Fig. 3f). The pre-antennal and antennal chevrons became faint. In gnathal and thoracic segments, expression was localized in the posterior compartments, not only for the sterna, but also for the appendages. In stage 11, all segmental boundaries in the abdominal region could be identified, based on external structures of embryos (Machida 1981). The *Pu-en2* expression pattern in the head was highly obscured and hardly recognized anymore due to a high-level background signal and increasing complexity of the cephalic structure, but expression elements located in the posterior portion of some head appendages were clearly observed (Fig. 3h-j). In the thorax and abdomen, *Pu-en2* was expressed in the posterior compartments of the sterna, terga and appendages of each segment (Fig. 3g, k, l). In summary, the first expression of *Pu-en2* appears in the intercalary segment; then, *Pu-en2* expression occurs sequentially in the following segments posterior to the intercalary one. *Pu-en2* expressions in the antennal segment and pre-antennal region, however, do not appear sequentially.

Despite testing various conditions for material processing and temperature variations during and after hybridization, I could not achieve a sufficient signal-to-noise ratio for reproducible *in situ* hybridization with any *Pu-en1* probe. However, one specimen, a late stage 3 embryo, did show obscure signals and expressions of *Pu-en1* in the posterior regions of the antennal, intercalary, mandibular, maxillary and labial segments (Fig. 3m). No significant difference could be found between the *Pu-en1* expression pattern (Fig. 3m) and the *Pu-en2* expression pattern in an embryo at a similar stage (Fig. 3d). Nevertheless, I could not determine whether the *Pu-en1* and *Pu-en2* expression patterns were equivalent only from this single specimen.

Discussion

1. Expression pattern of an engrailed-family gene in the jumping bristletail could be used as a segment marker

Previous studies have compared the expression patterns of two paralogous engrailed-family genes in individual insect species such as the firebrat (Peterson et al. 1998), *Drosophila* (Siegler and Jia 1999), cockroach (Marie and Bacon 2000) and grasshopper (Peel et al. 2006). In *Drosophila*, the *engrailed* paralogs (*engrailed* and *invected*) differ slightly with regard to the expression patterns in the central nervous system (Siegler and Jia 1999). In firebrat (Peterson et al. 1998) and grasshopper (Peel et al. 2006), the expression of one paralog precedes that of the other when expression begins in the pre-antennal region. Regarding segmental expression, Peterson et al. (1998) reported that intercalary expression of one engrailed-family paralog (*Td-en*-r1) precedes that of the other (*Td-en*-r2) in firebrat. However, in my close inspection of their figures, *Td-en*-r2 appeared to be faintly expressed in the intercalary segment at the stage when *Td-en*-r1 begins to be expressed in that segment (Fig. 4A and D in Peterson et al. 1998). If my interpretation of their images is correct, then the timing of the initial intercalary expression should be very similar for both paralogs in firebrat. Marie and Bacon (2000) compared expression patterns of two engrailed-family genes in cockroach and found no obvious difference between them. These reports indicate that segmental expression patterns are not remarkably different between the engrailed-family paralogs in insects. Simultaneously, all cases reported in these previous studies show that the segmental expression begins with segment formation and is localized to the posterior portion of each segment.

I isolated two engrailed-family genes, *Pu-en1* and *Pu-en2*, from the jumping bristletail and analyzed the expression patterns of each gene with whole-mount *in situ* hybridization. As for the *Pu-en1*, however, I could obtain only one stage 3 embyo as the specimen with its in situ hybridization signal detected. Nevertheless, as far as the stage 3 embryo was concerned, there was no obvious difference between the *Pu-en1* and *Pu-en2* expression patterns (Fig. 3d, m). Although the data that I obtained in the present study are insufficient for a comparison between the *Pu-en1* and *Pu-en2* expression patterns, it might be reasonable to use the *Pu-en2* expression pattern as a segment marker because: 1) segmental expression patterns do not differ substantially between any known engrailed-family paralog pair in insects, and 2) the

Pu-en2 expression pattern was typical of that for an engrailed-family gene, in that the *Pu-en2* expression accompanied segment formation and was restricted to the posterior portion of each segment.

2. Expression of engrailed-family gene, *Pu-en2*, represents modified pattern of ocular spots, or the vestige of pre-antennal segment?

Ocular spots, which are a paired expression element of engrailed-family gene in spot shape in pre-antennal region in insects, have been regarded as the expression in the clusters of cells developing into eyes and central nervous system (Rogers and Kaufman 1996), hence should not be homologous to segmental expression elements. In the present study, I found *Pu-en2* expression in a chevron pattern in the pre-antennal region of early embryos in jumping bristletail, which can be homologized with that of ocular spots. Otherwise, closely resembling the antennal chevron pattern, this pre-antennal expression seems to be serially homologous to the antennal chevron pattern. The existence of a pre-antennal segment has been argued extensively (cf. Rempel 1975), but it is likely that the *Pu-en2* expression of the chevron pattern revealed in the present study for jumping bristletail might represent one more vestigial segment anterior to the antennal segment or the pre-antennal segment.

3. Expression pattern of engrailed-family gene, *Pu-en2*, in the jumping bristletail may show

the primitive pattern of segmentation in Insecta

The intercalating formation of the intercalary segment is a widely conserved feature of anterior segmentation in insects. Rogers and Kaufman (1996) showed that engrailed-family genes are first expressed in the intercalary segment after stripes of expression have appeared in other head segments in two holometabolous insects (fly and fleas) and two hemimetabolous insects (milkweed bug and cricket). Peterson et al. (1998) have reported the intercalating expression of one of two engrailed-family paralogs (*Td-en*-r2) in the intercalary segment of firebrat, which belongs to the order Zygentoma, the second basalmost group of insects next to Archaeognatha. According to their figure of the youngest embryo that they analyzed, the intercalary stripe of the other engrailed-family paralog (*Td-en*-r1) is weaker than the antennal and gnathal stripes (Fig. 4A in Peterson et al. 1998). This weak expression suggests that the intercalary stripe appears after the antennal and gnathal stripes. Thus, it seems that both engrailed-family paralogs in firebrat are expressed in the intercalary segment after they are expressed in the other head segments. Rogers and Kaufman (1996) noted that the high conservation of the order in engrailed-family gene expression suggests that the mechanism of segment formation should also be highly conserved.

My observations indicate that the first expression of an engrailed-family gene in the jumping bristletail appeared in the intercalary segment. This finding reveals that the order of segment formation in the jumping bristletail differs from that shown in the other insects and suggests that the jumping bristletail does not share the segmentation mechanism for the intercalary and neighboring segments that seems to be conserved among all other insects hitherto examined.

Peel and Akam (2003) proposed a hypothesis that explains the evolutionary transition from short-germ to long-germ segmentation as follows. Ancestral short-germ insects possessed two types of mechanism controlling segmentation: one mechanism involved anterior determinants (the future gap genes) that controlled non-sequential segmentation in the anterior segments, and the other mechanism was the segmentation clock that controlled sequential segmentation in the posterior segments. During evolution, the number of segments controlled by the segmentation clock was reduced, beginning with the more anterior segments, and genes that were newly recruited as anterior determinants took over control of the formation of those anterior segments. In the most derived insects such as *Drosophila*, which represent the final state in this hypothesis, the segmentation clock is not involved in segmentation, and all segments are under the control of the gap genes. This hypothesis does not contradict the segmentation patterns of extant insects. However, I cannot entirely exclude the possibility that parallel transition from short-germ to long-germ occurred at some steps during the evolutionary history of insects.

The findings reported by Pueyo et al. (2008) reinforced Peel and Akam's (2003) hypothesis. They investigated the molecular mechanism of segmentation in cockroach, which is a hemimetabolous short-germ insect. They revealed that the Notch signaling pathway, which is involved in one of the components of the segmentation clock, plays a critical role in sequential segmentation and, in contrast, that the Notch is less important in the non-sequential segmentation in the anterior region.

In the present study, I revealed that segmentation in the jumping bristletail began with the intercalary segment and that segments posterior to the intercalary segment were sequentially formed. These findings are consistent with the hypothesis of Peel and Akam (2003) in that the more primitive forms possess more segments sequentially differentiated.

My results suggest that segmentation in the jumping bristletail represents the most ancestral state; specifically, most segments including the intercalary segment are under the control of the segmentation clock. In all insects examined to date, except jumping bristletail, the intercalary segment develops intercalatingly, not sequentially. This implies that the intercalary segment has escaped the control of the segmentation clock and been controlled by some newly acquired factor, which may be regarded as one of the anterior determinants hypothesized by Peel and Akam (2003). If this is the case, then the mechanism responsible for intercalating formation of the intercalary segment can be an autoapomorphy of the Dicondylia, all extant insects excluding the jumping bristletail or Monocondylia.

Chapter 2: Phylogenetic analysis of engrailed-family gene duplication in insects, with emphasis on the influence of gene conversion

Introduction

Insect species possessing at least two engrailed-family genes have been reported in many groups from ancestral orders to derived ones. Several recent studies have discussed the evolutionary relationships between engrailed-family genes in insects, including when and how many times gene duplications have occurred in the evolution of insects.

There are several scenarios for the evolution of engrailed-family genes that are roughly classified into two categories: gene duplication in a common ancestor and multiple independent duplications in some insect lineages. The existence of the regions that code motifs characteristic to only one of the two engrailed-family genes in each lineage of insects is regarded as evidence for duplication in a common ancestor. On the other hand, following points have been proposed as the evidence for the multiple duplications. Firstly, molecular phylogenetic analyses demonstrate tree topologies supporting multiple duplications (Peterson et al. 1998; Marie and Bacon 2000). Secondly, the pairs of engrailed-family genes in Diptera and Lepidoptera share a conservative intron that has not been found in any other taxa (Walldorf et al. 1989; Peel et al. 2006). It should mean that duplication have occurred in the common ancestor of Diptera and Lepidoptera independently from the other lineage.

Peel et al. (2006) hypothesized that gene conversion has repeatedly occurred between engrailed-family paralogs in insect. When gene conversion occurs between paralogs, parts of their sequences become homogenized and the apparent evolutionary distance between paralogs become smaller. If gene conversion of engrailed-family paralogs occurred in each lineage of insects, the phylogenetic trees of these genes would show evidence that the genes individually duplicated in each lineage at a timing later than the actual timing of duplication. In addition, the existence of introns conserved among paired engrailed-family genes in Diptera and Lepidoptera is evidence that gene conversion occurred and transferred an intron in one of the paired paralogs to the other.

Peel et al. (2006) strongly suggested that gene conversion had occurred in several insect species in which full-length sequences of engrailed-family genes were reported, although these were limited in the crown group of insects, or Neoptera. In the stem groups of insects, such as Apterygota and Paleoptera, only partial sequences of engrailed-family genes have been isolated. It is a possible that gene conversions occurred in the stem groups of insects, but previous studies did not show clear evidence to support this assumption. More information on the engrailed-family genes from basal insects is necessary to determine whether duplication of these genes had occurred in the common ancestor of all insects.

Here, I cloned engrailed-family genes in jumping bristletail, order

Archaeognatha, which is the basalmost of the insects. In addition, I searched for several engrailed-family genes in genomic sequences of arthropods in public databases. Based on the sequenced and obtained data, the conserved sequence regions and genomic structure of arthropod engrailed-family genes were examined, and statistical analyses for traces of gene conversion and phylogenetic analyses were conducted. Based on these findings, the origin and evolution of the paired engrailed-family genes in insects is discussed.

Materials and methods

1. Data

See Chapter1 for cloning of engrailed-family genes in jumping bristletail, *Pedetontus unimaculatus* Machida, 1980. cDNA sequence data of engrailed-family genes from other animals were obtained from Genbank or extracted from the genomic sequence data in Genbank. The cDNA and genomic DNA sequences of the engrailed-family genes listed in Table 1 were aligned by eye based on putative amino acid sequences and reconverted into nucleotide sequences having 396 nucleotides for analyses. Bioedit (Hall 1999) was used for editing sequence data.

2. Statistical analysis for gene conversion

GENECONV (Sawyer 1989) was used to seek possible sequence regions in which gene conversion had occurred within the nucleotide sequence dataset alignment.

GENECONV analysis was performed with the gscale set from 0 to 5. The gscale is a parameter indicating mismatch penalty between sequence pairs. Fragments in a certain pair of sequences, which represent the *p*-value < 0.05 in global comparison, not pairwise comparison, were listed as the candidate regions in which the gene conversion occurred.

3. Phylogenetic analysis

Based on the results of GENECONV analysis, the presumptive region in which gene conversion frequently occurred among all aligned genes was chosen, and the aligned sequences were divided into two datasets, one including the positions in the chosen region and the other excluding the positions. Then a pair of trees was obtained from two different phylogenetic analyses that were performed individually using datasets including and excluding these positions. The paired trees were compared for the purpose of evaluating the influence of gene conversion on the phylogenetic analyses using maximum likelihood (ML) phylogenetic analyses in MEGA version 6.06 (Tamura et al. 2013). Prior to the analysis, substitution models were tested using model selection in the MEGA 6.06 and models scoring the smallest BIC (Bayesian information criterion) were applied.

Results

1. Conserved features in the sequences and genomic structures of engrailed-family genes in insects and other arthropods

Regions of engrailed-family genes in genomic sequences were identified by eye using characteristic sequences and conservative introns as clues. Logan et al. (1992) compared the amino acid sequences of vertebrate engrailed-family proteins and identified five conserved regions (engrailed homology regions, EH1 to EH5). These five regions are also conserved in insect and arthropod engrailed-family proteins (Peel et al. 2006). An intron that is widely conserved among animals, referred to as the EH2 intron, was inserted between the first and second codon positions of a highly conserved glycine residue at the C-terminal region of EH2 (Logan et al. 1992; Peel et al. 2006). In insects, the genomic structures of engrailed-family genes have been reported in some holometabolous species (Coleman et al. 1987; Hui et al. 1992; Brown et al. 1994; Peel et al. 2006) (Fig. 4). In these reports, two engrailed-family paralogs are shown to be oriented in tail-to-tail orientation to each other. One of the paired engrailed-family genes has an EH2 intron, and the other has two introns and a hexanucleotide microexon, corresponding to the arginine-serin dipeptide motif (RS-motif), at the position of the EH2 intron. The region corresponding to the EH2 intron, or the two introns and a microexon flanked by them, is called the EH2 intronic region (Peel et al. 2006).

Insertion of RS-motif coding sequences was found in one of the two engrailed-family genes in all insect species that are known to possess two engrailed-family genes. Engrailed-family genes that do not have an RS-motif coding region are generally named *engrailed* (*en*) in holometabolous insect species or *engrailed1* (*en1*) in other insect taxa, and other genes including the RS-motif are generally named *invected* (*inv*) in holometabolous insects or *engrailed2* (*en2*) in others. Here, the convention for insect engrailed-family genes is used, although engrailed-family genes in some insects originally had other names. Peel et al. (2006) compared full-length sequences of engrailed-family proteins in insects and demonstrated that invected and engrailed2 proteins in these insects shared a 'leucine-serine-valine-glycin' tetrapeptide motif (LSVG-motif) on the N-terminal side of EH1. Here, the microexon within the EH2 intronic region, RS-motif and LSVG-motif is regarded as "invected-specific" characteristics, although, in a strict sense, they are "invected-and-engrailed2-specific".

The engrailed-family genes in 28 insect species and 12 non-insect arthropod species were examined (Table 1). Except for nine species that were examined in Peel et al. (2006), 31 species were newly examined in the present study. In the six of these 31 species, the cDNA sequences of engrailed-family genes were available, but the genomic sequences involving the engrailed-family genes have not been published. In the other 25 of the 31 species, the genomic sequence data involving predicted engrailed-family genes were registered in Genbank. The gene regions of engrailed-family genes in these 25 species were carefully reexamined by referring to the existence of sequences coding for five engrailed homology regions, RS-motif and LSVG-motif, and the conservative introns. Consequently, 12 putative engrailed-family genes were newly found from eight insects and one non-insect arthropod (Fig. 5).

Concerning insects, the newly determined findings for engrailed-family genes in the present study do not contradict the previously reported features of engrailed-family genes. More specifically, the full-length sequences of the genes were found to possess conserved regions coding for five engrailed homology regions (Fig. 6), and one of the two paralogs in each species possessed a LSVG-motif coding region in the 5' terminal region and a RS-motif coding region in the EH2 coding region (Figs 4 and 5a). However in *Aedes* (yellow fever mosquito, order Diptera) and *Plutella* (diamondback moth, order Lepidoptera), putative *engrailed* genes have already been deposited in Genbank, but the second engrailed-family gene for each of these species was not found in the published genomic sequences. In *Zootermopsis* (termite, order Isoptera), full-length *Zootermopsis_en1* was found in the genomic shotgun sequences but only a partial sequence of *Zootermopsis_en2* lacking exons on the 5' side of the EH2 intronic region was found (Fig. 5k).

Within newly examined insects in the present study, no information on the genomic structure of engrailed-family genes was available for *Pedetontus* as only cDNA and not genomic DNA was analyzed for this insect. All engrailed-family genes and their genomic structures were found to bear the EH2 intronic region. Only one of two paralogs in each insect species possessed the microexon corresponding to RS-motif in the EH2 intronic region. All engrailed-family genes in dipterans and lepidopterans possess the conserved intron in the EH4 coding region, which is called the homeobox intron (Peel et al. 2006).

Because each of the engrailed-family genes was included in distinct shotgun sequences, the relative positions of the paired paralogs in the genomic sequences are unknown in *Operophtera* (winter moth, order Lepidoptera) and *Ceratosolen* (fig wasp, order Hymenoptera). The genomic positions of *Pedetontus_en1* and *Pedetontus_en2* are also unknown due to the absence of genomic data. Second engrailed-family genes have not been found in *Aedes* and *Plutella*. In the insects newly examined in the present study, except for *Operophtera*, *Ceratosolen*, *Pedetontus*, *Aedes* and *Plutella*, the paired paralogs were revealed to be positioned in a tail-to-tail orientation in each genome.

Two engrailed-family genes in *Daphnia* (water flea, Crustacea, Branchiopoda) were also positioned in a tail-to-tail orientation in a single genomic scaffold sequence, and only one of the two paralogs included microexon in the EH2 intronic region. The LSVG-motif was found in the 5' terminal region in both of the paralogs (Fig. 7a).

2. Statistical analysis for gene conversion

Table 2 shows the list of fragments in specific pairs of genes as detected by GENECONV analysis that are the candidates in the region of gene conversion. Distribution of fragments with a putative converted region in aligned sequences is illustrated in Figure 8.

The results of GENECONV changed moderately depending on penalty level for mismatch, except the case of gscale = 1, or the smallest penalty. Fourteen fragments, the largest number of fragments associated with a gscale, were detected for gscale = 1, and the breakdown of number of detected fragments was as follows for the other gscale values: ten fragments for gscale = 2 and 3, nine for gscale = 4, seven for gscale = 5 and six for gscale = 0or the heaviest penalty in which no mismatch is allowed. Six fragments were detected solely for gscale = 1, while two fragments were detected in all conditions aside from gscale = 1. Looking at the condition of gscale = 2 and progressively heavier penalties, an increase in the penalty caused some pairs of genes to disappear from the list of detected pairs, while new pairs were never added. The fragment length became shorter or was unchanged as the penalty was raised. Most of the detected fragments begin within the EH4 coding region and end within the 3' terminal of EH5 coding region.

All five pairs of engrailed-family paralogs in lepidopterans were detected with gscale = 1 to 4, and three pairs in lepidopterans were even detected with gscale = 5 and 0. The pairs of paralogs in two species of Polyneoptera, one of the groups of hemimetabolous insects, *Schistocerca* (grasshopper, order Orthoptera) and *Periplaneta* (cockroach, order Blattodea) were detected with gscale = 2 to 5 and 0. The pair of *Schistocerca_en1* and *Periplaneta_en2* was also detected with gscale = 2 to 5 and 0.

3. Phylogenetic analysis

The ranges of the fragments detected by GENECONV analysis were overlapping but varied depending on gscale and sequence pair. Two sequence regions in which gene conversion were purported to occur frequently were chosen.

The first putative gene conversion region was from position 184 to 372 in the alignment. This region included most positions in fragments of lepidopterans except for the 5' 80-nucleotide long region of the fragment of *Danaus* (monarch butterfly), which is significantly longer than other fragments (Fig. 8). Substitution model selection was carried out for each of the regions, including and excluding positions 184 to 372, and the Tamura-Nei model (Tamura and Nei 1993) incorporating among-site rate variation approximated by a five-category discrete gamma distribution and a proportion of invariant sites (TN93+G+I model) was selected for both of the separate datasets.

Figure 9 shows the maximum likelihood trees inferred from the two datasets using sequences from 184 to 372 and other positions. In both trees, support values for the nodes were low, as a whole, except for some peripheral nodes. A small number of aligned positions seemed to be responsible for the insufficient resolutionin the trees.

In the tree inferred by sequences from positions 184 to 372, each pair of *engrailed* and *invected* genes in the same species in lepidopterans formed a highly supported clade, and all lepidopteran genes formed a clade. In contrast, in the other tree inferred from a dataset that excluded positions 184 to 372, all *engrailed* genes in lepidopteran species formed a clade, as did all *invected* in lepidopterans, and the clades of lepidopteran *engrailed* and *invected* were positioned on branches distant to each other. Similarly, while each pair of paralogs of *Drosophila* (fruit fly, order Diptera), *Nasonia* (jewel wasp, order Hymenoptera) and *Orussus* (parasitic wood wasp, order Hymenoptera) formed a clade in the tree based on sequences from positions 184 to 372, all *engrailed* genes of dipterans, all *invected* of dipterans and all *engrailed* of hymenopterans formed separate clades and all *invected* of hymenopterans were also positioned around closely related branches in the tree based on the positions other than at positions 184 to 372. Four genes of two coleopteran species also formed two clades containing two *engrailed* or two *invected* genes in the tree of the analysis using the sequences excluding positions 184 to 372, though these genes were scattered in the tree based on the sequences from position 184 to 372. Looking at the genes of non-holometabolans, however, pairs of *engrailed1* and *engrailed2* in *Pediculus* (louse, order Anoplura), *Schistocerca, Pedetontus* and *Daphnia* formed separate clades in both trees.

The other putative gene conversion region is between positions 154 to 374. This region is equivalent to a long fragment detected in *Schistocerca*, and it includes most positions in all fragments except for *Danaus*, which is the longest one (Fig. 8). Using positions 154 to 374, the TN93+G+I model scored the smallest BIC, followed by the Hasegawa-Kishino-Yano (HKY) (Hasegawa et al. 1985) +G+I model, while the HKY+G+I model scored the smallest BIC, followed by the TN93+G+I model excluding positions 154 to 374. When phylogenetic analyses were performed with each of the datasets applying these two models, the branching pattern did not change between the trees of the different models. Figure 10 shows the maximum likelihood trees based on these datasets with the TN93+G+I model. In both trees, support values for the nodes were generally low, except for at some peripheral nodes. In the genes of holometabolans, the differences in clustering patterns between trees based on sequences including and excluding positions 154 to 374 showed patterns similar as for trees based on sequences including and excluding positions 184 to 372. The five genes of polyneopteran species, including *engrailed1* and *engrailed2* in *Schistocerca* and *Periplaneta*, and *engrailed1* in *Zootermopsis*, formed a highly supported clade, and each pair of paralogs in *Schistocerca* and *Periplaneta* also formed highly supported clades in the tree based on sequences of positions 154 to 374. In contrast, in the tree excluding positions 154 to 374, two clades containing three *engrailed1* genes and two *engrailed2* genes in polyneopterans were respectively formed and positioned apart from each other. While each pair of paralogs in *Pedetontus* and *Daphnia* formed a highly supported clade in the tree using the positions 154 to 374, *engrailed1* and *engrailed2* in each species were positioned distantly and these two *engrailed1* genes clustered with *engrailed1* of polyneopterans and *engrailed* of hymenopterans in the tree excluding the positions 154 to 374. The pair of paralogs in *Pediculus* formed a clade in both trees.

Discussion

1. Gene conversion detected in 3' region of sequences, significantly in Lepidoptera

The results of GENECONV analysis changed moderately, depending on the selections of gscale = 2 or greater mismatch penalties, including gscale = 0. It is appropriate to regard fragments detected under these conditions as candidates for regions of gene conversion.

It is remarkable that the same fragments were detected in the pair of *Danaus_en* and *Danaus_inv* with all the different gscale settings examined. In addition, this fragment, which spans from the EH2 to EH5 coding region, was the longest. When genomic sequences of the paired paralogs in *Danaus* were aligned, only two nucleotide mismatches and one nucleotide gap were found in a continuous 807 bp region that included two exons and two introns (Fig. 11). The identity score between *Danaus_en* and *Danaus_inv* was 62.2% (79/127 bp) in the entire conserved region, excluding the fragment detected by GENECONV analysis; therefore, this pair of paralogs shows great similarity in this 807 bp genomic sequence region. This high similarity, which is restricted to the partial sequences of the genes, should be considered to be a consequence of the recent gene conversion in the pair of *Danaus* engrailed-family genes.

Paired engrailed-family genes in dipterans and lepidopterans share conserved introns called homeobox introns, which are inserted in the EH4 coding region or homeobox (Hui et al. 1992; Peel et al. 2006). Peel et al. (2006) pointed out the possibility that the homeobox intron was acquired after duplication of the engrailed-family gene and transferred from one of the duplicated paralogs to the other. The entire regions of homeobox introns in *Danaus_en* and *Danaus_inv* were included in the sequence region bearing high similarity, suggesting that gene conversion likely occurred (Fig. 11).

The 5' terminal ends of the fragments detected in three lepidopterans, *Papilio* (African swallowtail butterfly), *Bombyx* (silkworm) and *Operophtera*, extended to the fourth to seventh nucleotides from the 5' terminal position of the insertion point of homeobox intron (Fig. 8). The genomic sequences corresponding to these fragments were checked, and while a pair of engrailed-family genes in each species showed significant similarity in the exon region neighboring the 3' terminal of the homeobox intron, sequences in the 5' terminal region of homeobox introns were not at all similar to each other (Fig. 12). In addition, these homeobox introns showed different lengths between pairs of paralogs in each species. Due to high conservation in EH4 coding regions, GENECONV analysis chose longer fragments beyond the insertion point of the homeobox intron in the condition with lighter mismatch penalties. In the gene pairs of these three lepidopterans, significant traces of gene conversion were restricted in the 3' side to the homeobox introns.

Homeobox introns of the same length, which are found only in *Danaus* among the five lepidopterans, suggest that recent gene conversion in *Danaus* had homogenized a pair of homeobox introns that were originally of different lengths because *Danaus* and *Papilio* are classified into the same superfamily, Papilionoidea. Based on these observations in homeobox introns of lepidopteran engrailed-family genes, it is reasonable to hypothesize that homeobox introns inserted in one of the paired paralogs had been transferred into the other in the common ancestor of lepidopterans and dipterans.

The tendency for gene conversion to occur depends on sequence identity (Walsh 1987) and the distance between sequences (Leigh Brown and Ish-Horowicz 1981; Liao 1999). Because the majority of the conserved regions in engrailed-family genes are in exons located on the 3' side of the EH2 intronic region, the sequence identity between paired paralogs is higher on the 3' side of the EH2 intronic region than on the 5' side. Further, due to the tail-to-tail orientation and very large EH2 intronic region, a pair of the 3' exons of the two engrailed-family paralogs in insects is positioned closer to each other than is a pair of 5' exons. Based on these conditions, Peel et al. (2006) proposed that the regions in which gene conversion occurred were restricted to the homeobox and surrounding sequences—that is, the 3' region of insect engrailed-family genes. Most fragments detected in the analyses correspond to the exons located on the 3' side of the EH2 intronic region. Only one position in the 5' end of the fragment in *Danaus* belongs to an exon located on the 5' side of the EH2 intronic region, but this position is the first codon for the highly conserved glycine residue in EH2 and no polymorphism is found in this position among all of the genes analyzed. Hence, these identical nucleotides should not be taken to indicate that the converted region between *Danaus_en* and *Danaus_inv* extends over the entire EH2 intronic region. Inspecting the genomic sequences of *Danaus_en* and *Danaus_inv*, the 5' terminal regions in EH2 intronic regions were shown to not be similar between paired paralogs. Therefore, the traces of gene conversion detected in the present study were involved in the 3' flanking regions in the genomic sequences of the insect engrailed-family genes, which is consistent with the discussions of Peel et al. (2006).

2. Phylogenetic trees affected by gene conversion between engrailed-family paralogs

It is not reasonable to consider that phylogenetic trees in the present study reliably represent the true evolutionary history of arthropod engrailed-family genes because the region of aligned sequences is so short that the support values for basal branchings are insufficient. It is notable that the phylogenetic inferences seem to be completely contradicted between different sequence regions in the homologous genes.

Positions from 184 to 372 in the aligned sequences include most of the regions that include traces of the gene conversion in lepidopterans (Fig. 8). The clustering pattern of lepidopteran engrailed-family genes is completely different between the phylogenetic trees made using datasets including and excluding these positions (Fig. 9). It is supposed that the evolutionary distances between the paralogs in each lepidopteran species are underestimated in the tree based on the dataset including the positions affected by gene conversions. As suggested from the tree made from the dataset excluding positions 184 to 372, it is probable that pairs of engrailed-family genes in lepidopterans have been duplicated prior to the divergence of each species.

Significant indication of gene conversion was also detected by GENECONV analysis in *Schistocerca*, which is a phylogenetically more basal insect than lepidopterans.

The paired engrailed-family genes in this insect formed a clade in both of the trees drawn by the datasets, including or excluding positions 184 to 372 (Fig. 9). When the other datasets including or excluding positions 154 to 374 were analyzed, however, this pair of genes clearly showed different clustering patterns between the two trees (Fig. 10). In the paired engrailed-family genes of *Schistocerca*, phylogenetic inference was affected by the signal of gene conversion involved in the 32 positions excluded from the 189 positions from position 184 to 372 but included in the 221 positions from 154 to 374.

While differences in clustering patterns in the paired paralogs in lepidopterans and polyneopterans depend on the sequence regions used in the analysis, similar tendencies are observed in pairs in other taxa, including species with no evidence of gene conversion based on GENECONV analysis. It is important to note that GENECONV analysis can find traces of gene conversion but it cannot be used to prove the absence of gene conversion. The differences of clustering pattern between the two trees (including and excluding the positions 154 to 374) suggest that gene conversion occurred not only in lepidopterans and polyneopterans but also in most of dipterans and hymenopterans, *Dendroctonus* (mountain pine beetle, order Coleoptera), *Pedetontus* and *Daphnia*. Though the pair of engrailed-family genes in *Pediculus* formed a clade in two pair of the trees (including and excluding the positions 184 to 372 or 154 to 374), even this robust result cannot be considered as evidence for the absence of gene conversion or independent gene duplication in this insect lineage. Focusing on the support value for the clade of *Pediclus* genes, it decreased from 100 in the analysis including positions 154 to 374 to under 50 in the analysis excluding these positions. This decrease seems to imply that the positions 154 to 374 in *Pediculus* genes are affected by gene conversion.

Comparing these two pairs of trees, it is still difficult to ascertain when and how many times gene conversion occurred during the course of insect evolution, but it is apparent that gene conversion affects phylogenetic analysis findings. Ignoring the low support values in the tree based on the dataset excluding the positions 154 to 374, its topology seems to indicate several duplication events; however, it is too hasty to consider that multiple duplication occurred in insect evolution because the influence of gene conversion must not be erased by this manipulation. At least 50 more positions should have to be removed from the dataset for phylogenetic analysis in order to exclude the region in which gene conversion probably occurred in *Danaus* (Fig. 8). Because the number of available nucleotide positions is limited, it is very difficult to determine when and how many times engrailed-family genes underwent duplication in insect evolution by excluding the influence of gene conversions.

3. Paralog-specific regions and conservation of genomic structures suggest gene duplication in the common ancestor of insects

Peel et al. (2006) compared engrailed-family genes in seven winged insect species belonging to four holometabolous orders (Diptera, Lepidoptera, Coleoptera and

Hymenoptera) and two hemimetabolus orders (Orthoptera and Blattodea). These were all of the available full-length sequences of insect engrailed-family genes at that time. Adding to these established sequences, here, I examined 40 engrailed-family genes in 21 insect species, including three newly examined orders: Anoplura, one of hemimetabolous group phylogenetically positioned relatively close to Holometabola; Isoptera, hemimetabolous insects closely related to Blattodea; and Archaeognatha, apterous insects positioned basalmost of the insects.

In the present study, pairs of engrailed-family paralogs coding five engrailed homology regions were found in all insects examined, except *Aedes* in Diptera, *Plutella* in Lepidoptera and *Zootermopsis* in Isoptera (Fig. 6). Although the second engrailed-family genes with invected-specific features in *Aedes* and *Plutella* and the 5' exon of *Zootermopsis_en2*, which are expected to code EH1 and the LSVG-motif, are not found in Genbank, there is no evidence for the absence of these sequences. Because these features are shared in multiple species that are closely related to each of these three insects, it seems highly probable that the conditions for these three insects are similar.

These findings showed that only one of the two engrailed-family genes possess LSVG-motif and RS-motif coding regions in various insect species (Figs 4 and 5a). These features in the pair of engrailed-family genes are shared in *Pedetontus*, one of the basalmost insects; many holometabolous insects, the most derived insects; and multiple hemimetabolous insects, members of paraphyletic stem group in insects. Consequently, it is tempting to conclude that paired engrailed-family genes found throughout insect lineages originated from a single duplication in a common ancestor. Here, it was demonstrated that gene conversions have affected the phylogenetic tree of the engrailed-family genes, although few findings contradict the hypothesis of duplication in the common ancestor. However, conclusive evidence for single duplication in a common ancestor has not yet been demonstrated.

Because parallel acquisition of the LSVG-motif or the RS-motif seems unlikely, the orthology of these motifs among insects is reliable. Nevertheless, the absence of these motifs is insufficient as proof of orthology among *engrailed* and *engrailed1* genes in insects. *Engrailed1*-like paralogs can occur independently by losing sequence regions that code invected-specific motifs. Furthermore, RS-motif can easily be lost by a single mutation of the splicing receptor site if the RS-motif is encoded by microexon in all insect species such as is the case in holometabolous insects (Peterson et al. 1998).

Although characteristics specific to *engrailed* in holometabolous insects and *engrailed1* in other insects (hereafter, “engrailed-specific”) are critical for the orthology of each paralog, engrailed-specific characteristics failed to be found in the vast insect lineages from basal to derived taxa. Hui et al. (1992) recognized two engrailed-specific domains by comparing engrailed-family proteins of *Bombyx* and *Drosophila*. Manzanares et al. (1993) insisted that the single engrailed-family protein of *Artemia* (brine shrimp, Crustacea, Branchiopoda) possessed one of these two engrailed-specific domains, though its conservance remains at a low level. Thereafter, comparisons of the engrailed-family proteins among insects, including hemimetabolous species and the other arthropods, have suggested that these engrailed-specific domains are not widely conserved (Marie and Bacon 2000; Peel et al. 2006). If engrailed-specific sequences do not exist, we cannot reject the possibility that *engrailed* and *engrailed1* occur independently in several lineages by losing invected-specific motif sequences.

In the present study, we found that one of the two engrailed-specific domains described by Hui et al. (1992), which was also found in *Artemia* by Manzanares et al. (1993), is conserved in *engrailed* in dipterans, lepidopterans and *Dendroctonus*, and *engrailed1* in *Pedetontus*, *Zootermopsis* and *Pediculus* (Fig. 7b). Although the sequence of this domain is not highly conserved between *Pedetontus* and *Drosophila*, *Bombyx* or *Dendroctonus*, high sequence conservation is observed between *Artemia* and *Pedetontus* in not only the engrailed-specific domain described by Hui et al. (1992) and Manzanares et al. (1993) but also in the C-terminal side sequence flanking region. The engrailed-specific domain of *Zootermopsis* and *Pediculus* is similar to that of *Drosophila, Bombyx* and *Dendroctonus* on the N-terminal side and to that of *Pedetontus* and *Artemia* on the C-terminal side. This similarity means that the domain found by Manzanares et al. (1993) shared among crustacean species and derived insects was not acquired independently but was evolutionarily inherited.

The present study has shown that both the basalmost and derived insects possess paired engrailed-family proteins, one which has an engrailed-specific domain and the other which has invected-specific motifs. This situation suggests that the common ancestor of all insects already had a pair of engrailed-family genes; one is orthologous to *engrailed* in holometabolous insects and the other to *invected*. This engrailed-specific domain seems to have been lost several times in the lineage of hymenopterans, in *Schistocerca* and *Periplaneta*. Compared to the other species in the same order, most of the engrailed-specific domains were lost and became gaps in *Tribolium* and *Danaus*.

Genomic information of engrailed-family genes in each insect lineage is necessary for confirming that the two engrailed-family genes in all insect species originate from a single duplication in a common ancestor. In a previous study, several insects from vast holometabolous lineages, Diptera, Lepidoptera, Coleoptera and Hymenoptera, were shown to share the characteristic genomic structure of engrailed-family genes, that is, two engrailed-family genes are positioned tail-to-tail on a chromosome and only one of the two genes has six nucleotide microexons corresponding to the RS-motif in EH2 intron (Peel et al. 2006). Adding to this, here, this genomic structure was identified in the published genome sequences of 14 holometabolous insects and 2 hemimetabolous insects, *Pediculus* and *Zootermopsis*. Thus, it seems probable that not only holometabolous insects but also some hemimetabolous insects possess the set of orthologs corresponding to *engrailed* and *invected* in *Drosophila*.

Although there is not yet genomic information on engrailed-family genes of the more basal hemimetabolous insects and apterygote insects, the genomic information of some non-insect arthropods has been published. Recently published genome sequence of water flea, *Daphnia* (Colbourne et al. 2011) includes two engrailed-family genes, and one of two has a microexon corresponding to a RS-motif like *invected* and *engrailed2* in insects. Furthermore, the two engrailed-family genes of *Daphnia* are positioned tail-to-tail in a genomic sequence. In the putative amino acid sequences, protein without a RS-motif (Daphnia_en1) was shown to have an engrailed-specific domain in the N-terminal end (Fig. 7b), and the other (Daphnia_en2) was shown to have a LSVG-motif. More precisely, Daphnia_en1 and Daphnia_en2 have an LSVG-motif on the N-terminal side of EH1, while the sequence around the LSVG-motif is highly conserved between Daphnia_en2 and Pedetontus_en2 but is not conserved between Daphnia_en1 and any engrailed2 or invected of insects (Fig. 7a). These similarities in engrailed-family genes of *Daphnia* and insects suggest that the gene duplication that established paired engrailed-family genes in insects probably occurred in the common ancestor of insects and *Daphnia*.

The single engrailed-family gene in *Artemia* is paradoxical under the supposition that duplication of the engrailed-family gene occurred in the common ancestor of insects and *Daphnia*. Because *Artemia* and *Daphnia* are branchiopods, we have to postulate that one of the duplicated genes was lost in the lineage of *Artemia*. However, strangely, the engrailed-family protein of brine shrimp has an engrailed-specific domain and an RS-motif (Fig. 6 and 7b), one of the invected-specific motifs (Manzanares et al. 1993). One possibilities is that the RS-motif in the engrailed-family protein of *Artemia* was acquired independently from the RS-motif of insects and *Daphnia*, based on differences in the genomic structure around the RS-motif coding region; there is no intron at the 5' terminal side of RS-motif coding region (Manzanares et al. 1993). Thus, the RS-motif is not coded by the microexon in *Artemia*, which is different from the situation in insects and *Daphnia*. Information on other branchiopods is required for further discussion on the evolutionary relationship between the two engrailed-family genes in *Daphnia* and single engrailed-family gene in *Artemia*.

While molecular phylogenetic studies have frequently shown insects branching from one of the crustacean lineages, we have no concrete evidence for which lineage of crustaceans is closest to insects. Brachiopods are one of the crustacean groups that are regarded as being the closest to insects (Glenner et al. 2006), but the most recent phylogenomic study on arthropods does not support this (Regier et al. 2010). If gene duplication which is the cause of paired engrailed-family genes in insects occurred in the common ancestor of insects and some crustaceans, not only the information on engrailed-family genes in various crustacean lineages but also inference related to the phylogenic tree of insects and crustaceans is important for determining the timing of gene duplication. Alternatively, the extremely unique genomic structure of engrailed-family genes might be one of the critical clues for determining the root of insects in crustacean lineages.

General Discussion

Segmentated body plan can be defined as repetition of the set of structural units along the anterior-posterior axis (Davis and Patel 1999). Each unit in a single segmented organism is homologous, that is called the serial homology. Although it is often judged by presence or absence of the common origin whether certain structures are homologous or not, entitative common origin cannot exist between serially homological structures: imagine the common origin of arm and leg, and then, the limb in your image had never existed in an ancestral animal. Therefore, a discussion on regarding certain structures serially homologous or not sometimes becomes that on how similar they are. In other words, a proof of serial homology is to discover critical similarity. The topic on preantennal segment in insects is the typical of issue for serial homology. From the point of comparative morphology and embryology, the existence of preantennal segment have been examined by relationship between ganglions and anterior neural structures, e.g. prosocerebrum, deutocerebrum and tritocerebrum; or between appendages and anterior appendicular structures, e.g. clypeolubrum, eye, and great appendage in Cambrian animals, however agreement cannot be reached. That is to say, some similarities are recognizable between preantennal region and following segments, but these similarities are not accepted as the critical evidence. I examined the expression pattern of engrailed-family genes in jumping bristletail, the basalmost insect, and observed similar expression patterns between preantennal region and antennal segment. I have not necessarily intended to insist that this similarity is critical. It has been known that part of engrailed-family expression in insects is not involved in segment formation, being expressed in developing neural systems, and also that expression pattern of engrailed-family genes is specialized in limited insect species, being expressed in labrum and hindgut as observed in *Drosophila*. Nevertheless, if the similarities between preantennal regions and following segments pointed by comparative morphology and embryology are not critical in extant insects, it is possible that the similarities originate from a common mechanism which used to exist in the ancestral insects or arthropods. In such a case, it is not very strange that some vestiges sharing common mechanism can be found in some primitive arthropods. The present observation in segmentation of jumping bristletail might provide an insight about relationship between mechanism of segmentation and segmental-looking structures in preantennal region. Now studies on the homologues of segmentation genes of *Drosophila* in other insects and arthropods are not a few, but information on basal insects or non-insect arthropods is still fragmentary in the point of phylogenetic linkage. Engrailed-family gene expression in preantennal region in crustaceans represents different pattern from insects (Scholtz, 1997). Information of segmentation genes in the jumping bristletail will be important to compare the statuses of preantennal region in insects and those in crustacean lineages that are closely related to insects, the information of which has remained ambiguous yet.

Now the crustaceans have come up on the present topic. It is important to define the status of monophyletic group Insecta for comparing insects with crustaceans, not only about preantennal region, but about any other themes on the morphological evolution. Traditionally, it was believed that the hexapod lineage, which is represented by insects, is closely related with myriapods, and morphologists often presupposed that plesiomorphic states of insects were found in myriapods. However, molecular phylogenic studies revealed that hexapod lineage had branched from crustaceans. Because the supposed sister group of hexapods have changed frequently, re-evaluation for evolutionary status of hexapod and insect morphologies is now required. However, there still remains difficulty with the resolution of the tree concerning on the sister group of hexapods; it is still unclear which group in crustaceans is the closest relative to hexapods. In addition, morphological diversity among crustaceans is quite large comparing to that within hexapods. In the present study, I have discussed the status of jumping bristletail as the start point of evolution of insect segmentation. The status of jumping bristletail may also be a milestone for discussion about evolutionary transition from crustaceans to insects.

Here I would like to further mention that jumping bristletail could link segmentation patterns between insects and crustaceans. Even though the range of the variation in segmentation pattern among crustaceans is still unclear, two types of segmentation obviously different each other can be recognized, that in the direct development and in indirect development. Indirectly developing crustaceans typically hatches out as nauplius larva, which possesses only three anteriormost segments, first antennal, second antennal and mandibular segments. Posterior segments are sequentially added as the larva develops toward adult. On the other hand, segmentation process was documented by antibody staining for an engrailed-family protein in one of the directly developing crustaceans, gammarids; the three anteriormost segments appear at first, and posterior segments are formed sequentially as the embryo grows (Scholtz et al. 1994). I observed that the only intercalary segment appears at first, and then in the following stage, the three anteriormost segments were observed in the jumping bristletail embryo. It is tempting to imagine that somewhat common mechanism underlies this similarity between some crustaceans and the basalmost insect, Archaeognatha.

Acknowledgments

I gratefully acknowledge Prof. Tetsuo Hashimoto of the University of Tsukuba and Prof. Ryuichiro Machida of the Sugadaira Montane Research Center, University of Tsukuba for their constant guidance as well as invaluable suggestions and advices given to me throughout the present study and critical reading of the manuscript.

For the valuable advice and critical reading of the manuscript, I wish to express my sincerest thanks to Prof. Katsuo Furukubo-Tokunaga and Assoc. Prof. Yuji Inagaki of the University of Tsukuba. Thanks are also due to Dr. Shigeo Hayashi and Dr. Nao Niwa of RIKEN CDB, Dr. Sohta Ishikawa of the Center for Computational Sciences, University of Tsukuba, and Dr. Masashi Sakuma for their valuable support and advice for the experimentation, and Prof. Osamu Numata of the University of Tsukuba for his constant encouragements.

I also thank Dr. Toshiki Uchifune, Dr. Mika Masumoto, Ms. Mariko Kusakari, Dr. Yoshie Uchifune, Dr. Makiko Fukui, Dr. Shota Shimizu, Dr. Shigekazu Tomizuka, Dr. Yuta Mashimo, Ms. Mari Fujita, the staff of the Sugadaira Montane Research Center for their kind hospitality and help, and Dr. Kaoru Sekiya for her valuable suggestions and continuous encouragements.

This work was supported by a Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Fellows Grant Number 20·6763.

References

Brown SJ, Patel NH, Denell RH (1994) Embryonic expression of the single *Tribolium* engrailed homolog. Dev Genet 15:7–18

Colbourne JK et al. (2011) The ecoresponsive genome of *Daphnia pulex*. Science 331:555–561

Coleman KG, Poole SJ, Weir MP, Soeller WC, Kornberg TB (1987) The *invected* gene of *Drosophila*: sequence analysis and expression studies reveal a close kinship to the engrailed gene. Genes Dev 1:19–28

Davis GK, Patel NH (1999) The origin and evolution of segmentation. Trends Genet 15:M68–M72

El-Sherif E, Averof M, Brown SJ (2012) A segmentation clock operating in blastoderm and germband stages of *Tribolium* development. Development 139:4341–4346

Glenner H, Thomsen PF, Hebsgaard MB, Sørensen MV, Willerslev E (2006) Evolution. The origin of insects. Science 314:1883–1884

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98

Hasegawa M, Kishino H, Yano T (1985). Dating the human-ape split by a molecular clock of mitochondrial DNA. J Mol Evo 22:160–174

Hui C, Matsuno K, Ueno K, Suzuki Y (1992) Molecular characterization and silk gland expression of *Bombyx engrailed* and *invected* genes. PNAS USA 89:167–171

Kornberg T (1981) *engrailed*: a gene controlling compartment and segment formation in *Drosophila*. PNAS USA 78:1095–1099

Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283–92

Krause G (1939) Die Eitypen der Insekten. Biol Zbl 59:495–536

Larink O (1969) Zur Entwicklungsgeschichte von *Petrobius brevistylis* (Thysanura, Insecta). Helgoländer Wiss Meeresunters 19:111–155

Leigh Brown AJ, Ish-Horowicz D (1981) Evolution of the 87A and 87C heat-shock loci in *Drosophila*. Nature 290:677–682

Liao D (1999) Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet 64:24–30

Logan C, Hanks MC, Noble-Topham S, Nallainathan D, Provart NJ, Joyner AL (1992) Cloning and sequence comparison of the mouse, human, and chicken engrailed genes reveal potential functional domains and regulatory regions. Dev Genet 13:345–358

Machida R (1981) External features of embryonic development of a jumping bristletail, *Pedetontus unimaculatus* Machida (Insecta, Thysanura, Machilidae). J Morphol 168:339–355

Machida R (2000) Serial homology of the mandible and maxilla in the jumping bristletail *Pedetontus unimaculatus* Machida, based on external embryology (Hexapoda:

Archaeognatha, Machilidae). J Morphol 245:19–28

Manzanares M, Marco R, Garesse R (1993) Genomic organization and developmental pattern of expression of the *engrailed* gene from the brine shrimp *Artemia*. Development 118:1209–1219

Marie B, Bacon JP (2000) Two *engrailed*-related genes in the cockroach: cloning, phylogenetic analysis, expression and isolation of splice variants. Dev Genes Evol 210:436–448

Misof B et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767

Peel A, Akam M (2003) Evolution of segmentation: rolling back the clock. Curr Biol 13:708–710

Peel AD, Telford MJ, Akam M (2006) The evolution of hexapod engrailed-family genes: evidence for conservation and concerted evolution. Proc R Soc B 273:1733–1742

Peterson MD, Popadić A, Kaufman TC (1998) The expression of two *engrailed*-related genes in an apterygote insect and a phylogenetic analysis of insect *engrailed*-related genes. Dev Genes Evol 208:547–557

Pourquié O (2003) The segmentation clock: converting embryonic time into spatial pattern. Science 301:328–330

Pueyo JI, Lanfear R, Couso JP (2008) Ancestral Notch-mediated segmentation revealed in the cockroach *Periplaneta americana*. PNAS USA 105:16614–16619

Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463:1079–1083

Rempel JG (1975) The evolution of the insect head: the endless dispute. Quaest Entomol 11:7–25

Rogers BT, Kaufman TC (1996) Structure of the insect head as revealed by the EN protein pattern in developing embryos. Development 122:3419–3432

Sander K (1976) Specification of the basic body pattern in insect embryogenesis. Adv Insect Physiol 12:125–238

Sarrazin AF, Peel AD, Averof M (2012) A segmentation clock with two-segment periodicity in insects. Science 336:338–341

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Scholtz G (1997) Cleavage, germ band formation and head segmentation: the ground pattern of the Euarthropoda. In: Fortey RA, Thomas RH (eds) Arthropod Relationships. Chapman and Hall, London, pp 317–332

Scholtz G, Patel NH, Dohle W (1994) Serially homologous engrailed stripes are generated via different cell lineages in the germ band of amphipod crustaceans (Malacostraca, Peracarida). Int J Dev Biol 38:471-478

Siegler MV, Jia XX (1999) Engrailed negatively regulates the expression of cell adhesion molecules connectin and neuroglian in embryonic *Drosophila* nervous system.

Neuron 22:265–276

Stollewerk A, Schoppmeier M, Damen WGM (2003) Involvement of Notch and Delta genes in spider segmentation. Nature 427:863–865

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Walldorf U, Fleig R, Gehring WJ (1989) Comparison of homeobox-containing genes of the honeybee and *Drosophila*. PNAS USA 86:9971–9975

Walsh JB (1987) Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion? Genetics 117:543–557

Wilson MJ, McKelvey BH, van der Heide S, Dearden K (2010) Notch signaling does not regulate segmentation in the honeybee, *Apis mellifera*. Dev Genes Evol 220:179–190

Tables

Table 1 List of sequences analyzed in the present study. Some genes predicted in genomic sequences were not registered as individual cDNA sequences but were registered as putative amino acid sequences of their products. Accession numbers for proteins are indicated with “protein_id”. *: Predicted exon regions for *Papilio_inv* and *Danaus_inv* seem complete except for the microexon corresponding to the RS-motif. On careful examination of the genomic sequence, a microexon for each of these two genes was found.

Table 1

Taxonomy	Species	Abbreviation for gene name	Accession number of gene (if not registered, of protein)	Accession number of genomic sequence
Insecta				
Holometabola				
Diptera	*Drosophila melanogaster*	*Drosophila_en*	M10017.1	–
		Drosophila_inv	X05273.1	
	Ceratitis capitata	*Ceratitis_en*	XM_012299283.1	NW_004523309.1
		Ceratitis_inv	XM_012299464.1	
	Lucilia cuprina	*Lucilia_en*	protein_id="KNC30840.1"	JRES01000487.1
		Lucilia_inv	–	
	Musca domestica	*Musca_en*	XM_011295369.1	NW_004765572.1,
		Musca_inv	–	NW_004764500.1
	Anopheles gambiae	*Anopheles_en*	U42429.1	NT_078268.4
		Anopheles_inv	–	
	Aedes aegypti	*Aedes_en*	XM_001649112.1	CH478380.1
Lepidoptera	*Papilio dardanus*	*Papilio_en*	protein_id="CAX36786.1"	FM995623.2,
		Papilio_inv	protein_id="CAW30923.1"*	FM955425.1
	Danaus plexippus	*Danaus_en*	–	JH384663.1,
		Danaus_inv	protein_id="EHJ77829.1"*	JH381360.1
	Bombyx mori	*Bombyx_en*	M64335.1	NW_004582034
		Bombyx_inv	M64336.1	
	Operophtera brumata	*Operophtera_en*	–	JTDY01002589.1,
		Operophtera_inv	–	JTDY01003936.1,
				JTDY01000275.1,
				JTDY01004450.1
	Amylois transitella	*Amylois_en*	XM_013332814.1	NW_013535427.1
		Amylois_inv	XM_013332817.1	
	Plutella xylostella	*Plutella_en*	XM_011565742	NW_011952501.1
Coleoptera	*Tribolium castaneum*	*Tribolium_en*	XM_967795.3	NW_001092855.1
		Tribolium_inv	S73225.1	
	Dendroctonus ponderosae	*Dendroctonus_en*	protein_id="ENN81669.1"	KB740073.1
		Dendroctonus_inv	–	
Hymenoptera	*Orussus abietinus*	*Orussus_en*	XM_012419053.1	NW_012191647.1
		Orussus_inv	XM_012419165.1	
	Athalia rosae	*Athalia_en*	XM_012406192.1	NW_012162344.1
		Athalia_inv	XM_012406191.1	
	Apis mellifera	*Apis_en*	XM_001121029.3	NW_003378174.1
		Apis_inv	XM_003251642.2	
	Megachile rotundata	*Megachile_en*	XM_003704983.2	NW_003797249.1
		Megachile_inv	XM_012289337.1	
	Ceratosolen solmsi	*Ceratosolen_en*	XM_011503306.1	NW_011948412.1,
		Ceratosolen_inv	XM_011505324.1	NW_011948456.1
	Nasonia vitripennis	*Nasonia_en*	XM_001607624.3	NW_001816793.1
		Nasonia_inv	XM_001607622.3	
	Microplitis demolitor	*Microplitis_en*	XM_008556512.1	NW_007541599.1
		Microplitis_inv	XM_008556513.1	
	Harpegnathos saltator	*Harpegnathos_en*	XM_011140703.1	NW_011649479.1
		Harpegnathos_inv	–	
	Camponotus floridanus	*Camponotus_en*	XM_011252308.1	NW_011878336.1
		Camponotus_inv	XM_011252306.1	
Paraneoptera				
Phthiraptera	*Pediculus humanus*	*Pediculus_en1*	–	DS235005.1
		Pediculus_en2	–	
Polyneoptera				
Orthoptera	*Schistocerca gregaria*	*Schistocerca_en1*	DQ323891.1	–
		Schistocerca_en2	DQ323892.1	
Blattodea	*Periplaneta americana*	*Periplaneta_en1*	AJ243883.1	–
		Periplaneta_en2	AJ243884.1	
Isoptera	*Zootermopsis nevadensis*	*Zootermopsis_en1*	–	KK852970.1,
		Zootermopsis_en2	–	KK852861.1
Apterygota				
Archaeognatha	*Pedetontus unimaculatus*	*Pedetontus_en1*	LC031803	–
		Pedetontus_en2	LC031804	
Crustacea				
Branchiopoda	*Daphnia pulex*	*Daphnia_en1*	protein_id="EFX69935.1"	GL732628.1
		Daphnia_en2	protein_id="EFX69893.1"	
	Artemia franciscana	*Artemia_en*	X70939.1	–
Maxillopoda	*Caligus rogercresseyi*	*Caligus_en*	GAZX01033554.1	–
	Argulus siamensis	*Argulus_en*	JW968661.1	–
	Sacculina carcini	*Sacculina_en-a.E9*	AF057692.1	–
		Sacculina_en-a.E20	AF057693.1	
		Sacculina_en-b	AF171074.1	
Chelicerata				
Araneae	*Cupiennius salei*	*Cupiennius_en1*	AJ007437.1	–
	Stegodyphus mimosarum	*Stegodyphus_en*	protein_id="KFM60502.1"	KK113577.1
	Parasteatoda tepidariorum	*Parasteatoda_en*	AB125741.1	–
Acari	*Ixodes scapularis*	*Ixodes_en*	–	DS924853.1
	Metaseiulus occidentalis	*Metaseiulus_en*	XM_003737361.1	NW_003803498.1
	Archegozetes longisetosus	*Archegozetes_en*	JQ700300.1	–
Xiphosurida	*Limulus polyphemus*	*Limulus_en-1Ba*	XM_013933153.1	NW_013665731.1
		Limulus_en-1Bb	XM_013921195.1	NW_013666557.1
		Limulus_en-1Bc	XM_013917449.1	NW_013666013.1
		Limulus_en-1Bd	XM_013926423.1	NW_013667790.1

Table 2 List of fragments detected in GENECONV analysis.

Table 2

Sequences names	Pvalue	Aligned Offsets			Number of	Number of	Total	Mismatch
		Begin	End	Length	polymorphic site	differences	differences	penalty
gscale=1								
Amyelois_en; Amyelois_inv	0.0000	200	368	169	116	2	84	4
Operophtera_en; Operophtera_inv	0.0000	172	379	208	143	9	84	4
Danaus_en; Danaus_inv	0.0000	103	372	270	188	0	48	7
Bombyx_en; Bombyx_inv	0.0001	187	380	194	133	7	78	4
Papilio_en; Papilio_inv	0.0010	184	377	194	132	6	77	4
Papilio_inv; Danaus_en	0.0027	185	353	169	117	13	93	4
Papilio_inv; Plutella_en	0.0076	187	315	129	82	7	95	4
Schistocerca_en1; Schistocerca_en2	0.0093	154	386	233	162	2	47	7
Anopheles_en; Anopheles_inv	0.0184	323	384	62	48	2	110	3
Harpegnathos_en; Harpegnathos_inv	0.0232	227	368	142	97	6	80	4
Aedes_en; Nasonia_inv	0.0240	323	396	74	60	6	114	3
Danaus_en; Amyelois_inv	0.0299	230	365	136	93	15	118	3
Plutella_en; Schistocerca_en2	0.0365	185	332	148	100	16	113	3
Bombyx_inv; Schistocerca_en1	0.0365	246	383	138	92	14	113	3
gscale=2								
Amyelois_en; Amyelois_inv	0.0000	200	368	169	116	2	84	7
Danaus_en; Danaus_inv	0.0000	103	372	270	188	0	48	13
Papilio_en; Papilio_inv	0.0000	184	360	177	122	4	77	8
Operophtera_en; Operophtera_inv	0.0000	185	379	195	133	7	84	7
Schistocerca_en1; Schistocerca_en2	0.0000	154	374	221	152	1	47	13
Bombyx_en; Bombyx_inv	0.0000	187	332	146	99	3	78	8
Anopheles_en; Anopheles_inv	0.0007	323	380	58	44	1	110	6
Periplaneta_en1; Periplaneta_en	0.0035	192	321	130	83	1	61	10
Harpegnathos_en; Harpegnathos_inv	0.0038	253	359	107	70	2	80	8
Schistocerca_en1; Periplaneta_en2	0.0415	259	347	89	56	1	78	8
gscale=3								
Danaus_en; Danaus_inv	0.0000	103	372	270	188	0	48	19
Amyelois_en; Amyelois_inv	0.0000	226	368	143	98	0	84	11
Schistocerca_en1; Schistocerca_en2	0.0000	154	374	221	152	1	47	19
Papilio_en; Papilio_inv	0.0000	184	360	177	122	4	77	12
Operophtera_en; Operophtera_inv	0.0000	185	315	131	83	2	84	11
Bombyx_en; Bombyx_inv	0.0000	220	332	113	76	1	78	12
Anopheles_en; Anopheles_inv	0.0034	323	380	58	44	1	110	8
Periplaneta_en1; Periplaneta_en2	0.0055	192	321	130	83	1	61	15
Harpegnathos_en; Harpegnathos_inv	0.0205	253	359	107	70	2	80	11
Schistocerca_en1; Periplaneta_en2	0.0312	259	333	75	46	0	78	12
gscale=4								
Danaus_en; Danaus_inv	0.0000	103	372	270	188	0	48	25
Amyelois_en; Amyelois_inv	0.0000	226	368	143	98	0	84	14
Schistocerca_en1; Schistocerca_en2	0.0000	154	374	221	152	1	47	25
Bombyx_en; Bombyx_inv	0.0001	220	332	113	76	1	78	15
Operophtera_en; Operophtera_inv	0.0001	185	315	131	83	2	84	14
Papilio_en; Papilio_inv	0.0019	184	360	177	122	4	77	16
Periplaneta_en1; Periplaneta_en2	0.0130	192	321	130	83	1	61	19
Anopheles_en; Anopheles_inv	0.0150	323	380	58	44	1	110	11
Schistocerca_en1; Periplaneta_en2	0.0279	259	333	75	46	0	78	15
gscale=5								
Danaus_en; Danaus_inv	0.0000	103	372	270	188	0	48	31
Amyelois_en; Amyelois_inv	0.0000	226	368	143	98	0	84	18
Schistocerca_en1; Schistocerca_en2	0.0000	154	335	182	124	0	47	31
Bombyx_en; Bombyx_inv	0.0003	220	332	113	76	1	78	19
Operophtera_en; Operophtera_inv	0.0019	185	290	106	68	1	84	18
Schistocerca_en1; Periplaneta_en2	0.0265	259	333	75	46	0	78	19
Periplaneta_en1; Periplaneta_en2	0.0338	229	321	93	59	0	61	24
gscale=0								
Danaus_en; Danaus_inv	0.0000	103	372	270	188	0	48	None
Amyelois_en; Amyelois_inv	0.0000	226	368	143	98	0	84	None
Schistocerca_en1; Schistocerca_en2	0.0000	154	335	182	124	0	47	None
Operophtera_en; Operophtera_inv	0.0018	214	290	77	49	0	84	None
Schistocerca_en1; Periplaneta_en2	0.0264	259	333	75	46	0	78	None
Periplaneta_en1; Periplaneta_en2	0.0338	229	321	93	59	0	61	None

Figures

Fig. 1 Segmentation patterns in insects. In short- and semi-long-germ insects, several anterior segments develop simultaneously on the germ rudiment just differentiated from blastoderm (the left line), and remaining posterior segments appears sequentially as the germ band elongating. Semi-long-germ insects bear longer germ rudiments and larger number of simultaneously formed segments than short-germ insects bear, but there is no definite division between them. In long-germ insects, all segments begin to develop almost simultaneously on the embryo differentiated from blastoderm and sequential segmentation does not occur.

Fig. 1

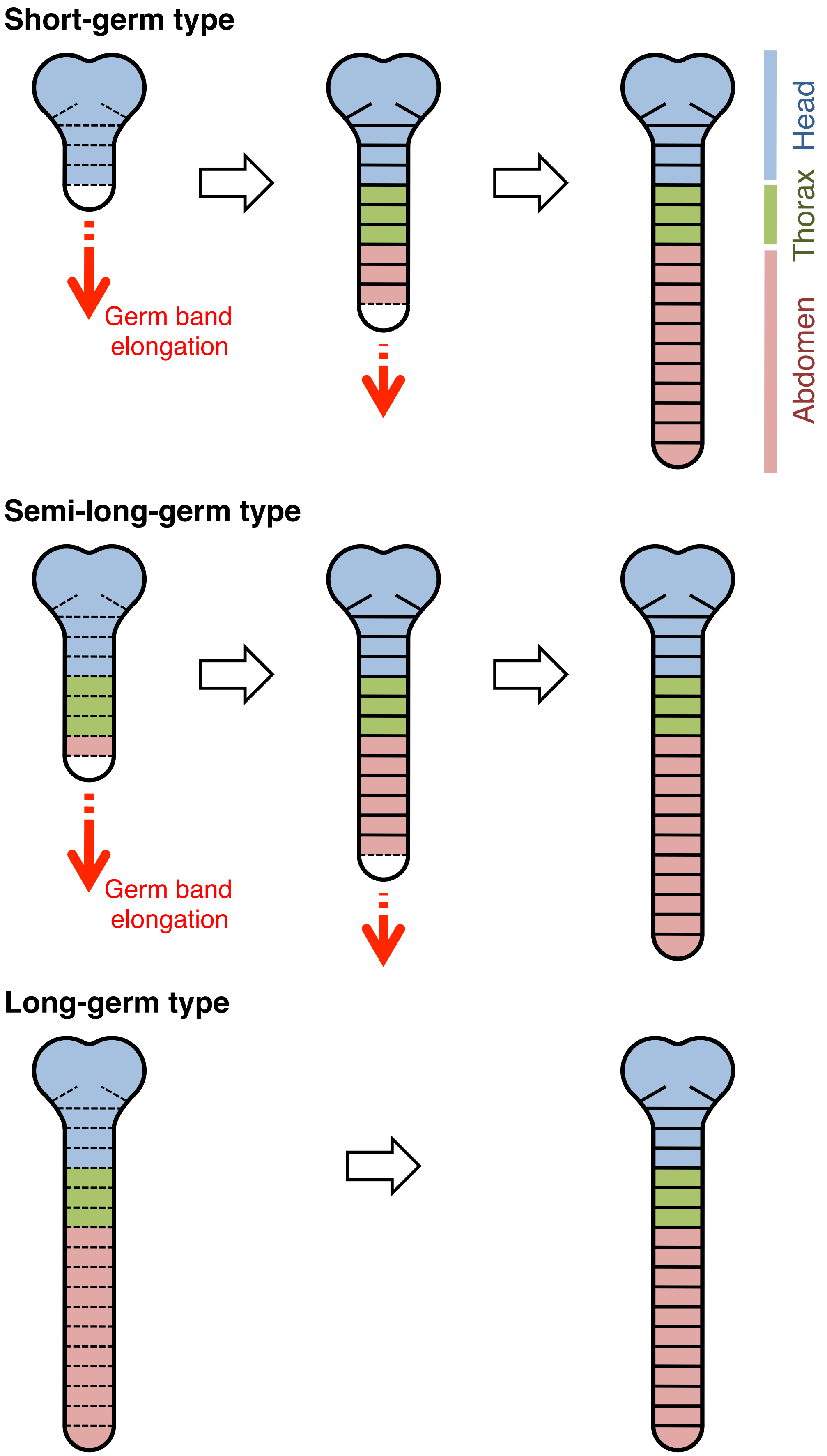

Short-germ type
Germ band elongation
Head
Thorax
Abdomen
Semi-long-germ type
Germ band elongation
Long-germ type

Fig. 2 cDNA sequence of *Pedetontus unimaculatus engrailed1* (*Pu-en1*, or *pedetontus_en1*) and *engrailed2* (*Pu-en2*, or *pedetontus_en2*). The coding region is in larger font and accompanied by the translated amino acid sequence. 5' and 3' untranslated regions are in small font. In-frame upstream codons are underlined. I am unsure whether the first methionine in Pu-en2 (indicated by an asterisk) is the start codon due to the lack of an upstream stop codon and Kozak consensus sequence. EH1 to EH5 stand for five engrailed homology regions. EH1, EH2, EH3, EH5 and corresponding cDNA regions are boxed. EH4, which is equal to Homeodomain, and corresponding cDNA region are shaded. RNA probes corresponding to the DNA sequence regions in bold font are synthesized for *in situ* hybridization analysis. Arrows above DNA sequence shows primer binding sites. Light gray arrows show binding sites of outer primers for degenerate PCR. Dark gray arrows show binding sites of inner primers for degenerate PCR. Black arrows show binding sites of gene specific primers for RACE PCR

Fig. 2

Pu-en1 (*Pedetontus_en1*)

```
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACGTGTTTATTAAATATGAAAAAAATTGTTGTTTAAAATAACAAACAAAATAAAAAAACTTAAGTGATAATTCATGTCAATGAA

ATG GCT TTA GAA GTC GAA AGA GAG ACT GCA GGT AGC CCT TCA GGG GCA AGT AGC CCT GGA CCA AGT CCA GGC CGG
 M   A   L   E   V   E   R   E   T   A   G   S   P   S   G   A   S   S   P   G   P   S   P   G   R

CCA GCA TCT GCA AAT CCA AAC GCT GGA ACA CCG TCA ACA GCA TCT CCG ACT AGC CCC GAG CCT TCC CCG CGG CCA
 P   A   S   A   N   P   N   A   G   T   P   S   T   A   S   P   T   S   P   E   P   S   P   R   P
                                                                    EH1
GTT CTC GCC CAG CCT GTT GTT GTT GTT GTA CCC CAG CCG CGG TTG TCC CTA CCC TTT TCG GTG GAG AAC ATC TTG
 V   L   A   Q   P   V   V   V   V   V   P   Q   P   R   L   S   L   P   F   S   V   E   N   I   L

AAA CCA GAA TTT GGG CGC AGG GCC ATC CAA CAA CGC CCC GTC GTC CTA CAA GAA TCC CCT ACC AAC CTT CCG AGA
 K   P   E   F   G   R   R   A   I   Q   Q   R   P   V   V   L   Q   E   S   P   T   N   L   P   R

TCT TTG ACA CCT CCA CGA AAG TCG ACA GAT CCG CCA GGG AAA AAG GAA AAA CTT GAC CCT CAG CTA CCC AAA GGA
 S   L   T   P   P   R   K   S   T   D   P   P   G   K   K   E   K   L   D   P   Q   L   P   K   G
                EH2                                                                         EH3
CCT CTC GTT TGG CCA GCG TGG GTC TAT TGC ACG AGA TAT TCG GAC AGA CCC TCT TCA GGT CCT CGT TCT CGA CGT
 P   L   V   W   P   A   W   V   Y   C   T   R   Y   S   D   R   P   S   S   G   P   R   S   R   R
                                        EH4
ATG AAG AAA AGA GAA CGC CGA CCT GAA GAA AAG AGG CCT CGA ACG GCG TTC ACT CAA GAG CAG TTA GCG CGT CTA
 M   K   K   R   E   R   R   P   E   E   K   R   P   R   T   A   F   T   Q   E   Q   L   A   R   L

CGG CGT GAG TTC GAA GAA AAT CGC TAC CTG ACA GAA CGT CGG CGC CAG GAT TTG GCG AGA GAA CTT CAT CTC CAC
 R   R   E   F   E   E   N   R   Y   L   T   E   R   R   R   Q   D   L   A   R   E   L   H   L   H
                                                                                        EH5
GAA AAT CAG ATT AAA ATC TGG TTT CAA AAC AAG CGA GCA AAA ATT AAG AAA TCT ACT GGT CAA AAG GGT GGG TTG
 E   N   Q   I   K   I   W   F   Q   N   K   R   A   K   I   K   K   S   T   G   Q   K   G   G   L

GCG CTG CAG CTC ATG GCT CAG GGA CTC TAT AAT CAC AGC ACG GTC TCA GTG GAT GAG GAC GAG TCA AAT CCT ATG
 A   L   Q   L   M   A   Q   G   L   Y   N   H   S   T   V   S   V   D   E   D   E   S   N   P   M

CCC CTA TCA CCA ACA TCC GTT GGA CAC CAG TCG GAC ATT
 P   L   S   P   T   S   V   G   H   Q   S   D   I

TAGAGTCACACAAATTTTCAATGATCTCGTTGGCAGTAGCCGTAAGGACCAGACAAATGTTTGAAGGTGTGAGATGTGCATCAGTGTTGTTTTTAATGACTGAAGAAGTGTGATTTGTCCTGCCACATACAT
ACATTACATACAAATGTCAAATAATTTTTTTATGTAAATAAACTTTATTTGCCAGTGTAAATAGACGATAGATAAGTTATTTCCAAAGCACAAGAGCTAGTGTTCAGTTTCCTCGTGCAAAGACGTTAGTTA
CTAAGTGTTGTGTCTCTTTGTACATATAAAATATAACATATATATTACTCCTGTGGCGCCTCTATTGCCATGGGTGTGAACTATAATTCTCACATATGTGTCAAATATATTATCATACAATAATCAAAAAAA
AAAAAAAAAA
```

Pu-en2 (*Pedetontus_en2*)

```
GTG GTT TCT GGG AAA TCG ACT GCC GCG GTT GTG TTT GGA ACG TCG TGT GAG GCC ATT CCG TGC AAC TCG GTG AAT
 V   V   S   G   K   S   T   A   A   V   V   F   G   T   S   C   E   A   I   P   C   N   S   V   N

GTA GGG GTC GAG AAT TTG TCG ACT GTA ACG TCA AAA TTG GGC GTC GAG TCA CCA TGT ATT TTG TGT GAG GGC ACC
 V   G   V   E   N   L   S   T   V   T   S   K   L   G   V   E   S   P   C   I   L   C   E   G   T

GGC GGT GTT ATG GAG GAC GCT GGT GTT TTG GTG TAC CGT CCT CCG GGT CTA CGA GAA GCA CCT CTC ATG TCA CTT
 G   G   V   M*  E   D   A   G   V   L   V   Y   R   P   P   G   L   R   E   A   P   L   M   S   L

CAA GAA TCT GAT GAA GAC GAA GAA CTT AGT GTT GGC AGT GAA TCT CCA CCA CCC TTA CCC CCA TTG TCG AAT GAG
 Q   E   S   D   E   D   E   E   L   S   V   G   S   E   S   P   P   P   L   P   P   L   S   N   E

GGG GTG GAT GAT ACT ACA CAG GAC TAT GAC ACG TTA TCG ACA CCA TCC ACC CCA CCG ACA TCA ACA ACG ACT TCA
 G   V   D   D   T   T   Q   D   Y   D   T   L   S   T   P   S   T   P   P   T   S   T   T   T   S

ACA ACA ACA ACA ACA GAT CAA CAT GCG ATC AAC AAC GCG TGT GCG CCG TTA AAT AAT AAC AAT AAC GAA CCA GTG
 T   T   T   T   T   D   Q   H   A   I   N   N   A   C   A   P   L   N   N   N   N   N   E   P   V
                  EH1
ACG AGG ACA CTC AAG TTT TCA ATT GAT AAT ATA CTC AAA CCG GAC TTT GGT TTT ACA CGG AGC ATT GTA ACG CAT
 T   R   T   L   K   F   S   I   D   N   I   L   K   P   D   F   G   F   T   R   S   I   V   T   H

AGT ACT AAC AGT AAT AAT GCT CTA CCT CAA CCT CTG GTC AGT AAC AAT AGT AGA AAC TGT AAC AGC AGC AGT AGC
 S   T   N   S   N   N   A   L   P   Q   P   L   V   S   N   N   S   R   N   C   N   S   S   S   S

AGT AGC TGC AGC AGC AGT AGT AGC AGT GCG AGT AGC GTT GAA CCT GTG GAC TTA TCG CGA CAA GAG ACT ACC GGT
 S   S   C   S   S   S   S   S   S   A   S   S   V   E   P   V   D   L   S   R   Q   E   T   T   G
                                              EH2
AAT AAA AAG TCC GGC TCT TCG CAG CCG TTA GTG TGG CCA GCG TGG GTG TAT TGC ACG AGG TAT TCG GAC CGG CCT
 N   K   K   S   G   S   S   Q   P   L   V   W   P   A   W   V   Y   C   T   R   Y   S   D   R   P
                              EH3                                             EH4
TCT TCA GGA AGA AGT CCT CGA TCT CGT CGT ACC AAG AAG AAA GAA CGC CGA CCG GAG GAT AAA AGG CCG CGA ACA
 S   S   G   R   S   P   R   S   R   R   T   K   K   K   E   R   R   P   E   D   K   R   P   R   T

GCG TTT ACG CAG GAA CAA TTG GCG CGT CTT CGG CGT GAA TTC GAA GAG AAC CGC TAC CTG ACG GAA CGC CGA CGT
 A   F   T   Q   E   Q   L   A   R   L   R   R   E   F   E   E   N   R   Y   L   T   E   R   R   R

CAA GAT TTG GCA CGA GAC TTG AAC CTT CAC GAA AAC CAG ATC AAA ATA TGG TTC CAA AAC AAA CGG GCC AAA ATT
 Q   D   L   A   R   D   L   N   L   H   E   N   Q   I   K   I   W   F   Q   N   K   R   A   K   I
                              EH5
AAG AAG GCT ACT GGT CAA AAG GGA GGT TTA GCG CTG CAA CTT ATG GCG CAA GGA CTG TAC AAT CAC AGC ACA ATA
 K   K   A   T   G   Q   K   G   G   L   A   L   Q   L   M   A   Q   G   L   Y   N   H   S   T   I

CCT TTA CGC GAT GGT GAA GAA GAT GAT TCC ACG TGT TCG CCC CCT CCA ACG
 P   L   R   D   G   E   E   D   D   S   T   C   S   P   P   P   T

TAGCAGTGGTTGACCTCCCAAAAGTGACTGTTGTGCAGTGACAAACTGAAGTGAATGTCATATCCCACATCATTGTTCATATGTATACCCATATTTCTTTCTTATATTCAGATGTTGTCAAGTGATGTACCA
TTTCATGGGAATATCAACGTTTTCCCGTATTCCACGTTTGCCCGCATTATAGGCCTGCGTTTTTCTAAGGTATCGACCTTATTTTCGAATAAACCACTATGATTTCAGGGGTTCGTTTGTTTCTTGAATTTT
GCACATTATATTAGGCTCCTCTCTAAGAGTCCGTGACAACTATAAAGCTCTAAATACGTTTATGCCAAAAATGTGACTCTATTACTAGCTTTTAATAAAATAAATTGAAAAAATGCTAACATATTTAAATAG
AATTTATTGTTATGCTTACAATAATAGACATTTAATGCCTTTTAATATTCGTGTATTTGTAATAGGGGTAAAATAGAAACAAAAACTGAAACTCATTTTCTTTTGCACTTATGAGTTTAAACGATACAAACA
GGGTGCGTCACCAGAATCCGCACCCTGTATATAACATTTCAAACATAACACTATATTTTAATGAAGTATTTATAACTTTTATTACCTTGTTTTGTTTTCCGTAAACGGGTTTCGCGATTCGCAGATATCACG
GACTCTAAAATCTTTTACGAAATATTTTAAATCATTGGGAAATTTCCTAACACGAAAATAGCCTACTATATTATGTTTATTATCTACACACAGAATAAAAAAAGCAATCGAACTGGCCTCTTTTCTATAGAA
TGTTATTCAATCTTCACATTTGTAAGTATTTTGTAAAATTTGGTAATATTAAAATAATGTGAAAACTGATATTAAAATTTTCATGTGTTTTAATAATAATTATAGGAAGTAAATATATTAGGCTAATGCTAT
TTACTCTAGTACAAGTATAAATATATTTGTCTTATAAAATTATAGGTACCATACGAATTATTTACAATTTTCATGATTGTATTAAATCAATAAGATTTGTTTTCATAAATTATGTTAAACCTTTTTAATATT
TAATAATTAAAAATAATGTGAATATAGTTTAAATATATTTTTAACAAATTTTGTGAAAGTTTAAACAGTGGATATAATTATACATATCTTTGGTAATGAAAAAAAAAAAAA
```

Fig. 3 Expression patterns of *Pu-en2* (**a-l**) and *Pu-en1* (**m**) in embryos of the jumping bristletail *Pedetontus unimaculatus* shown by whole mount *in situ* hybridization. **a** shows ventral view of early stage 1 embryo and serosa around the embryo. The embryonic area is shown by arrowheads. **b-f** and **m** show ventral views of embryos, oriented their heads to top. **b** is late stage 1 embryo. **c** is stage 2 embryo. **d** is early stage 3 embryo (left) and SEM image of the same stage embryo (right). The segment boundaries were first observed as external structures on embryo surface in this stage. **e** is stage 4 embryo. **f** is stage 6 embryo. **g** shows lateral view of an embryo of stage 11, with yolk mass on its back, oriented their head to left. **h-l** close up to the appendicular parts of the same specimen as **g,** showing the antenna, maxilla, maxillary palp, metathoracic appendage and styli, respectively. **m** is late stage 3 embryo, which was torn into anterior and posterior parts by bad manipulation. Asterisks in **b-f** and **m** show stripes of *Pu-en1* or *Pu-en2* expressing region, representing the posterior part of intercalary segments. Arrows in **d-f** show expression of *Pu-en2* in preanttenal region. An, antenna; AnS, antennal segment; IS, intercalary segment; Lb, labium; LbS, labial segment; Md, mandible; MdS, mandibular segment; Mx, maxilla; MxP, maxillary palp; MxS, maxillary segment; S, serosa; Sty, stylus; T, tergum; Th1–3, pro-, meso-, and metathoracic segments; ThL1–3, pro-, meso- and metathoracic appendages; Y, yolk; I–X, first to 10th abdominal segments. Scale bars, 100 μm

Fig. 3

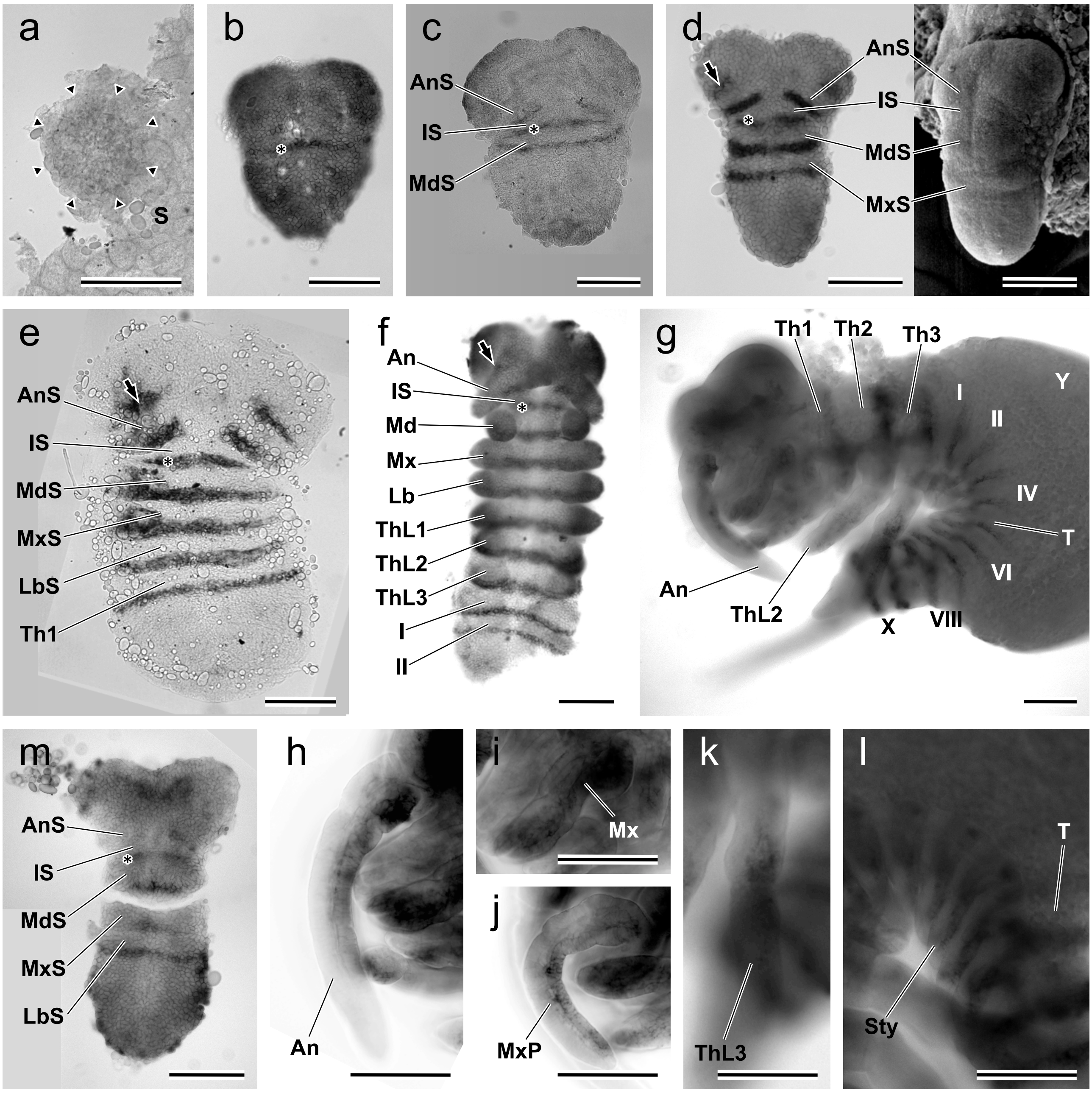

a
S
b
c
AnS
IS
MdS
d
AnS
IS
MdS
MxS
e
AnS
IS
MdS
MxS
LbS
Th1
f
An
IS
Md
Mx
Lb
ThL1
ThL2
ThL3
I
II
g
Th1
Th2
Th3
I
II
IV
VI
VIII
X
Y
T
An
ThL2
m
AnS
IS
MdS
MxS
LbS
h
An
i
Mx
j
MxP
k
ThL3
l
T
Sty

Fig. 4 Genomic structures of engrailed-family genes previously reported in holometabolous insects: Drosophila, Anopheles, Bombyx, Tribolium and Apis (Coleman et al. 1987; Hui et al. 1992; Brown et al. 1994; Peel et al. 2006). Boxes indicate the exon regions and horizontal lines indicate the UTR. Arrows above the gene regions show the direction of transcription. The *engrailed* and *invected* genes in each of these insects are positioned on the same chromosome and oriented their 3' terminal facing each other. EH2 intron in engrailed genes and two introns involved in the EH2 intronic region of invected genes are conserved among all of these insects. Other introns, which exist in not all of these insects, are not indicated in this figure.

Fig. 4

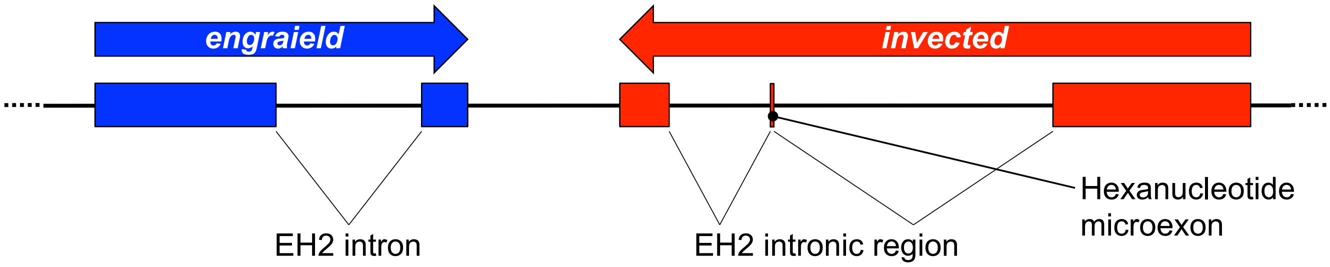

engraield
invected
EH2 intron
EH2 intronic region
Hexanucleotide
microexon

Fig. 5 cDNA sequence and deduced amino acid sequence of *Lucilia_inv* (**a**), *Musca_inv* (**b**), *Danaus_en* (**c**), *Operophtera_en* (**d**), *Operophtera_inv* (**e**), *Dendroctonus_inv* (**f**), *Harpegnathos_inv* (**g**), *Pediculus_en1* (**h**), *Pediculus_en2* (**i**), *Zootermopsis_en1* (**j**), *Zootermopsis_en2* in partial (**k**), and *Ixodes_en* (**l**), predicted from the genomic sequence. LSVG-motif or engrailed-specific domain is underlined by dashed lines. EH1 and EH1 coding region is in red. EH2 in Orange. The RS-motif and six-nucleotide sequence corresponding to the microexon are in violet. EH3 in dark red. EH4 or Homeobox/Homeodomain in blue. EH5 in green.

Fig. 5a

Lucilia_inv

```
atg cac ccg gac ttt agt ctg gtg ttg caa aag tgt gct tca aat tta agt tta aag cat aaa caa ccc ata cta
 M   H   P   D   F   S   L   V   L   Q   K   C   A   S   N   L   S   L   K   H   K   Q   P   I   L
tat agc gcc tat att aac ggt aaa act atg gcc aat acg gcg tcg gaa ata acc aat aat tct tca gag gat aca
 Y   S   A   Y   I   N   G   K   T   M   A   N   T   A   S   E   I   T   N   N   S   S   E   D   T
gag gat ctg gtg cgc ata gta agc gat gag gag gat gaa act gaa ata gag gaa aac tac cca cca cat cat gtg
 E   D   L   V   R   I   V   S   D   E   E   D   E   T   E   I   E   E   N   Y   P   P   H   H   V
cag caa gaa gat gat gat aat att agt tta tgc agt gaa cta tcg gtg ggt aag gaa aat cct gga gag gag ccg
 Q   Q   E   D   D   D   N   I   S   L   C   S   E   L   S   V   G   K   E   N   P   G   E   E   P
tca gca gcc agt gta gga gag gat gag gaa att ttg gat att agt gat act cat tcg aat tct tcc aat gag gaa
 S   A   A   S   V   G   E   D   E   E   I   L   D   I   S   D   T   H   S   N   S   S   N   E   E
atg caa cat tct ccc tcg gtt tct tcg caa gaa acg ggc ata aat ccg ttt gct ttg gcc cat aac tta aga ttt
 M   Q   H   S   P   S   V   S   S   Q   E   T   G   I   N   P   F   A   L   A   H   N   L   R   F
ccg gtg cct ttt ggt tta cag cct aat gct gca act aca gca gta aat ata tct cct ttt caa gag gaa ttt cta
 P   V   P   F   G   L   Q   P   N   A   A   T   T   A   V   N   I   S   P   F   Q   E   E   F   L
agg aaa tct cat ctt tat gct gag gaa cta atg aaa cat caa atg caa tta atg gct gcc gca cga gcc agt gcc
 R   K   S   H   L   Y   A   E   E   L   M   K   H   Q   M   Q   L   M   A   A   A   R   A   S   A
ttt tcg ttg cgt tcc aat agc cta aca caa ctg cat cat aca cat caa att tcg cca act gcc aat cta gtg cat
 F   S   L   R   S   N   S   L   T   Q   L   H   H   T   H   Q   I   S   P   T   A   N   L   V   H
cat ctt aat cct cta gct aaa ata gga caa tta agt gca gca gct gct gct gct ctc tcc gtg gcg gcg caa caa
 H   L   N   P   L   A   K   I   G   Q   L   S   A   A   A   A   A   A   L   S   V   A   A   Q   Q
aat cat aat aaa cct tta aat cac aca caa cat acc cag cat atg cta caa cag caa caa cag cag cag caa cat
 N   H   N   K   P   L   N   H   T   Q   H   T   Q   H   M   L   Q   Q   Q   Q   Q   Q   Q   Q   H
caa cac ctg cat cag caa tat gat acc ttg gcc aaa tta act gat cta agt aag caa cat tcg cca gcc tcc aca
 Q   H   L   H   Q   Q   Y   D   T   L   A   K   L   T   D   L   S   K   Q   H   S   P   A   S   T
acc agt acc caa cac atg atg tcc aca ctc aat caa ttg cag acc caa atg cag gct cat tta ccc ggt ctg ctg
 T   S   T   Q   H   M   M   S   T   L   N   Q   L   Q   T   Q   M   Q   A   H   L   P   G   L   L
cat gac aat aat gat aat ctt cat gaa agg gct tta aaa ttt agc ata gac aat ata ctt aaa gct gac ttt gga
 H   D   N   N   D   N   L   H   E   R   A   L   K   F   S   I   D   N   I   L   K   A   D   F   G
cgt tcc aat agt ctg gac tcc ccg cct cat gtt cga aaa agt ggc caa cat aaa tca cac act caa agg caa aat
 R   S   N   S   L   D   S   P   P   H   V   R   K   S   G   Q   H   K   S   H   T   Q   R   Q   N
cgt ctt aac act aca tcg tcc tcc tcc tca tcc gta gcc tcc aac agc gcg cat acc atg tta acc gat tca ctg
 R   L   N   T   T   S   S   S   S   S   S   V   A   S   N   S   A   H   T   M   L   T   D   S   L
gaa cat tcg gcc aaa tcg tta gtt cat aat att tca cct tcg aca caa aca ttt tcc act tca ttg gcc acc ata
 E   H   S   A   K   S   L   V   H   N   I   S   P   S   T   Q   T   F   S   T   S   L   A   T   I
tgt acg aat agc aat gat tcg agt agt aca gct gtt agc agt agt tat agt agt gct aat ggt gct gct gat tta
 C   T   N   S   N   D   S   S   S   T   A   V   S   S   S   Y   S   S   A   N   G   A   A   D   L
gta aaa tca tct ccc tct caa aca cct aca cta tca cca gga ttt tcg acg tct ttg gag aat ggc gct gct ggc
 V   K   S   S   P   S   Q   T   P   T   L   S   P   G   F   S   T   S   L   E   N   G   A   A   G
agt ggg aca aat aaa aat gtt aat act tcg agc aaa tca gag gag agt aca aca acg gct acc tcg act tcg gct
 S   G   T   N   K   N   V   N   T   S   S   K   S   E   E   S   T   T   T   A   T   S   T   S   A
acg ggc acg ggt aat ggt cct ata gta tgg cct gct tgg gtt tat tgt acc cgt tat agt gat cgt cct agt tcg
 T   G   T   G   N   G   P   I   V   W   P   A   W   V   Y   C   T   R   Y   S   D   R   P   S   S
ggc aga agt cct aga gta aga aaa cct aaa gcg cca aaa tca acc aac tcc tca tcg gcc aca gca gca gcg tca
 G   R   S   P   R   V   R   K   P   K   A   P   K   S   T   N   S   S   S   A   T   A   A   A   S
gca tct agt gca gca ggg gtc gaa aaa gcc aac tca ccc agt tca aca tcg tct gca aat aac aac gaa gat aaa
 A   S   S   A   A   G   V   E   K   A   N   S   P   S   S   T   S   S   A   N   N   N   E   D   K
cga ccg cgt aca gca ttt agt ggt tca caa tta gca aga ctg aag cat gaa ttt aat gaa aat cgc tat tta acc
 R   P   R   T   A   F   S   G   S   Q   L   A   R   L   K   H   E   F   N   E   N   R   Y   L   T
gaa aaa cgt cgt caa caa cta agc tcc gaa ttg ggt tta aat gaa gca caa att aag att tgg ttt caa aat aaa
 E   K   R   R   Q   Q   L   S   S   E   L   G   L   N   E   A   Q   I   K   I   W   F   Q   N   K
cgt gcc aaa ttg aaa aag tct agt ggc gtt aag aat ccc ttg gcc ctg caa cta atg gct caa ggc ctc tac aat
 R   A   K   L   K   K   S   S   G   V   K   N   P   L   A   L   Q   L   M   A   Q   G   L   Y   N
cac tcc acc ata cct ttg act aga gag gaa gag gaa ctg caa gag ttg caa gag cgc gaa aag aac aac agc aat
 H   S   T   I   P   L   T   R   E   E   E   E   L   Q   E   L   Q   E   R   E   K   N   N   S   N
aat aat aca aat agc ttg caa caa cag caa cag gca gca agt gca gtt act tct tag
 N   N   T   N   S   L   Q   Q   Q   Q   Q   A   A   S   A   V   T   S   *
```

Fig. 5b

Musca_inv

```
atg tca tgt tcc atg gct gag ctg gaa ttc cag agc ccc aag aaa tcc aac gaa aat ggc aat gat gat gat gac
 M   S   C   S   M   A   E   L   E   F   Q   S   P   K   K   S   N   E   N   G   N   D   D   D   D
gat att gcc gag gat ttg gtg cga att gtc agc gat gat gaa gag gat gag ggt gga gat ggt gga gag gcc tcg
 D   I   A   E   D   L   V   R   I   V   S   D   D   E   E   D   E   G   G   D   G   G   E   A   S
gaa gca cct caa aaa cat cat caa ttg gat cta gat gaa aat gcc agc atg tgc agt gag ctg tcg gtg ggc cag
 E   A   P   Q   K   H   H   Q   L   D   L   D   E   N   A   S   M   C   S   E   L   S   V   G   Q
gaa cat gct ttg cat cat gaa ata ccg cca gcc gat gcc cag gag gat gaa gaa att ttg gat gtc agt gat acc
 E   H   A   L   H   H   E   I   P   P   A   D   A   Q   E   D   E   E   I   L   D   V   S   D   T
cat tca aat tct tcg gtg gat gaa aat aat cgc act ccc tcc aca cca gga gaa act atg tca tcc cta agc cct
 H   S   N   S   S   V   D   E   N   N   R   T   P   S   T   P   G   E   T   M   S   S   L   S   P
ttt act ctg gcc agt acc tta agg ttt ccc gtg cca ttt tct ctc caa gca ccc acg gcc aat gtg atg acc cct
 F   T   L   A   S   T   L   R   F   P   V   P   F   S   L   Q   A   P   T   A   N   V   M   T   P
ggg gct ccc aat gcc aat att tca cca ttt cag gag gaa ttc tta cgg aaa tcc cat tta tat gcc gag gaa cta
 G   A   P   N   A   N   I   S   P   F   Q   E   E   F   L   R   K   S   H   L   Y   A   E   E   L
atg aaa cat caa atg cat tta atg gcc gcc gct agg gca agt gca ttt agt ttg cgt aat caa aat cta cac cct
 M   K   H   Q   M   H   L   M   A   A   A   R   A   S   A   F   S   L   R   N   Q   N   L   H   P
ggc att ggg ggt cat cct tca cac ctg gtg cac att aat ccg ctg acg aaa ctg ggg caa att agt gca gcg gcg
 G   I   G   G   H   P   S   H   L   V   H   I   N   P   L   T   K   L   G   Q   I   S   A   A   A
gcg gct gca gcg gcc tta tca gcg gcg gcg gtg caa aat tcc aaa acc caa ggt cat ggt cag caa tcc caa att
 A   A   A   A   A   L   S   A   A   A   V   Q   N   S   K   T   Q   G   H   G   Q   Q   S   Q   I
ata tcc cca cac aat gcc gcc agt cat gtt cta cct cca caa cag cac cat cct cac tct cac ccc cat gcc cat
 I   S   P   H   N   A   A   S   H   V   L   P   P   Q   Q   H   H   P   H   S   H   P   H   A   H
ctt caa cat cac ctc cca caa tcg gcc agc acc aca tca gat acc tta gcc aaa ttg aca gcg cta agt aaa caa
 L   Q   H   H   L   P   Q   S   A   S   T   T   S   D   T   L   A   K   L   T   A   L   S   K   Q
acc tca ccc gcc acc agc cac atg atg tcc acc ctg aat caa ctg cag acc caa atg cag gcc cac ctg ccg cgg
 T   S   P   A   T   S   H   M   M   S   T   L   N   Q   L   Q   T   Q   M   Q   A   H   L   P   R
ctg ttc aat gac aat aat gat aat cta cat gaa aga gct tta aaa ttt agc ata gat aat ata ctc aaa ggc gac
 L   F   N   D   N   N   D   N   L   H   E   R   A   L   K   F   S   I   D   N   I   L   K   G   D
ttt ggc cgg ggt caa tcc cct cca gcc acg gcc acc tcg cat agt aaa aaa tcc ctg cac ttt tta tcc cat gcc
 F   G   R   G   Q   S   P   P   A   T   A   T   S   H   S   K   K   S   L   H   F   L   S   H   A
aat aaa caa ttt ttg aat att gcg aag caa aaa tca cat tgt tcc tca tca tcg tca tcg gca gta tcg gat act
 N   K   Q   F   L   N   I   A   K   Q   K   S   H   C   S   S   S   S   S   S   A   V   S   D   T
ttg gat cct cag caa cat caa cac caa caa caa cag cat cac caa cat caa tct tca gcg aat tca aca cca gcc
 L   D   P   Q   Q   H   Q   H   Q   Q   Q   Q   H   H   Q   H   Q   S   S   A   N   S   T   P   A
ttt tcc gct tcc ttg gcc tcg att tgt acc aat agc aat gac tcg aat agc acg gcg gtc agt agt agt tac agc
 F   S   A   S   L   A   S   I   C   T   N   S   N   D   S   N   S   T   A   V   S   S   S   Y   S
agt gca aat ggc acg gga gat tta att aaa tca tca ccc tct cag aca cca act ctt tcg ccg ggt tta aat gaa
 S   A   N   G   T   G   D   L   I   K   S   S   P   S   Q   T   P   T   L   S   P   G   L   N   E
acg gct ggt agt ggg aga gcg ggt ggt ggt ggt gcg gtg ggg aaa gat gat acg ggt tcg gca aca aca gcg gga
 T   A   G   S   G   R   A   G   G   G   G   A   V   G   K   D   D   T   G   S   A   T   T   A   G
gcg ggg aca agt ggc agc agt agc ggt ggt acc agc ggc ggc aat ggt gga cct atc gtc tgg cct gcc tgg gtc
 A   G   T   S   G   S   S   S   G   G   T   S   G   G   N   G   G   P   I   V   W   P   A   W   V
tat tgt acc cgt tac agt gat cgt cca agt tca gga aga agt cca cgg gtg cga aaa cct aag aaa gcc act ggg
 Y   C   T   R   Y   S   D   R   P   S   S   G   R   S   P   R   V   R   K   P   K   K   A   T   G
cca aat gat aaa gcc aac tca cca act gga aca tcc tca tcc aca tca gcc ggc agt ggt ggt tca gcg tcc gca
 P   N   D   K   A   N   S   P   T   G   T   S   S   S   T   S   A   G   S   G   G   S   A   S   A
gcg gcg gcg tca gcg gca tca tcc tca tca tcg tcg gaa gat aaa cga ccc cgg acg gcg ttt agt ggt tcg caa
 A   A   A   S   A   A   S   S   S   S   S   S   E   D   K   R   P   R   T   A   F   S   G   S   Q
tta gca aga ctg aag cat gaa ttt aat gaa aat cgt tat ttg aca gaa aaa cga aga caa cag cta agc tct gag
 L   A   R   L   K   H   E   F   N   E   N   R   Y   L   T   E   K   R   R   Q   Q   L   S   S   E
ttg ggc ttg aat gag gcc caa atc aaa ata tgg ttt caa aac aaa cga gcc aaa ttg aaa aag tca agt ggt gtt
 L   G   L   N   E   A   Q   I   K   I   W   F   Q   N   K   R   A   K   L   K   K   S   S   G   V
aaa aat ccc ctg gcc ctg cag cta atg gcc cag ggt ctg tat aat cat tcg acc ata cca ttg acc cgc gaa gaa
 K   N   P   L   A   L   Q   L   M   A   Q   G   L   Y   N   H   S   T   I   P   L   T   R   E   E
gaa gag ctc caa gag cta cag gaa cgt gag aaa tcg gcc aac aac aat aac ctg aca caa cca acc gcc agt gcg
 E   E   L   Q   E   L   Q   E   R   E   K   S   A   N   N   N   N   L   T   Q   P   T   A   S   A
gtg tcc tct taa
 V   S   S   *
```

Fig. 5c

Danaus_en

```
atg gcg tac gag gac agg tgc agc ggc cac gcc gac atc aca cag gtc aac cag acc cag tac acc tgc act atc
 M   A   Y   E   D   R   C   S   G   H   A   D   I   T   Q   V   N   Q   T   Q   Y   T   C   T   I
aac cct agg aac atc aaa gta cag ccc gcg tcg ccg ccg ccc agc ccc gag tac tac cgg ccg gag act ccg gac
 N   P   R   N   I   K   V   Q   P   A   S   P   P   P   S   P   E   Y   Y   R   P   E   T   P   D
gtg aag ccc gtc atc gag gac gag cgc cgg aac ccg ata gct ttc tcc atc agt aac ata ctg cgt cca gag ttc
 V   K   P   V   I   E   D   E   R   R   N   P   I   A   F   S   I   S   N   I   L   R   P   E   F
ggt gtg acc gcc ctg agg aac tcc aag aag ata gag ggt cct aaa ccg ctc ggg ccc aac cac agc atc ctc tac
 G   V   T   A   L   R   N   S   K   K   I   E   G   P   K   P   L   G   P   N   H   S   I   L   Y
aag ccg tac gag ata acc aag gag ttg agt caa tat ggt tac gag tat gtg aag acg aaa gag gat ttc aac ctg
 K   P   Y   E   I   T   K   E   L   S   Q   Y   G   Y   E   Y   V   K   T   K   E   D   F   N   L
ccg ccg ctg gga ggg ttg agg cag acg gtg tcc agc atc ggg gag aaa gag tcc ccg aag gtc gtg gaa cag aag
 P   P   L   G   G   L   R   Q   T   V   S   S   I   G   E   K   E   S   P   K   V   V   E   Q   K
aga ccg gac tcg gcc agc tcg ata gta tcc tcc acc tcg agc ggc gcc gtc tcc tgc ggc agc acc gac aac agc
 R   P   D   S   A   S   S   I   V   S   S   T   S   S   G   A   V   S   C   G   S   T   D   N   S
tcg cag agc tcc cag ctg tgg ccg gcc tgg gtg tac tgc acc cgg tac agc gac aga ccg agc tca ggt ccc agg
 S   Q   S   S   Q   L   W   P   A   W   V   Y   C   T   R   Y   S   D   R   P   S   S   G   P   R
agt aga cgg gtg aag aag aag gcg agc cct gag gag aag aga ccg agg act gcc ttc agc gcc tcg cag cta aca
 S   R   R   V   K   K   K   A   S   P   E   E   K   R   P   R   T   A   F   S   A   S   Q   L   T
aga tta aag cac gag ttc gcg gag aac cgc tac ctg acg gag agg agg agg cag gcg ctg gcc gcg gag ctg ggg
 R   L   K   H   E   F   A   E   N   R   Y   L   T   E   R   R   R   Q   A   L   A   A   E   L   G
ctg gcg gag gct cag atc aag atc tgg ttc cag aac aag agg gcc aag atc aag aag gcc tcg ggc cag agg aac
 L   A   E   A   Q   I   K   I   W   F   Q   N   K   R   A   K   I   K   K   A   S   G   Q   R   N
ccg ctg gcg ctg cag ctc atg gcg cag ggg ctg tac aac cac gcc aca gtc acc gag agc gag gac gag gag atc
 P   L   A   L   Q   L   M   A   Q   G   L   Y   N   H   A   T   V   T   E   S   E   D   E   E   I
agc gtc acg tag
 S   V   T   *
```

Fig. 5d

Operophtera_en

```
atg gcg ttc gag gac cgt tgc agc cca aac cag ggc acc agc cca ggt cct gtg tta ggc aga gtg ccc gca cca
 M   A   F   E   D   R   C   S   P   N   Q   G   T   S   P   G   P   V   L   G   R   V   P   A   P
cac ggc atg aac cag caa tac tat ccg cca agc caa tac aca tgt acc act att gat tca agg tac gaa aga acc
 H   G   M   N   Q   Q   Y   Y   P   P   S   Q   Y   T   C   T   T   I   D   S   R   Y   E   R   T
ccc agc atg acg ctt gtg aaa gtc caa ccg aac tca cct cca ccc agc ccc aac agc aac gaa gga tac caa aac
 P   S   M   T   L   V   K   V   Q   P   N   S   P   P   P   S   P   N   S   N   E   G   Y   Q   N
tac tac aga cct gaa aca cca gac gtc aaa ccg caa atc agt gaa cag cgt ttt gag aat aaa cag ccg tcg gcg
 Y   Y   R   P   E   T   P   D   V   K   P   Q   I   S   E   Q   R   F   E   N   K   Q   P   S   A
ccg gtt gct ttc tca atc agc aat atc ctg cac cca gaa ttt ggc ttg aat gct tta cga aaa act aac aaa atc
 P   V   A   F   S   I   S   N   I   L   H   P   E   F   G   L   N   A   L   R   K   T   N   K   I
gag gga cct aag tct gtc gga cct aac cac agc att ctg tac aaa cct tat gat cta tca aag cag caa aat ggg
 E   G   P   K   S   V   G   P   N   H   S   I   L   Y   K   P   Y   D   L   S   K   Q   Q   N   G
agc tct gta cag ttt cag aag tat aac ttt gac tat ttg aaa tca aaa gaa tcg aat gac ttc aac cct ttg ccg
 S   S   V   Q   F   Q   K   Y   N   F   D   Y   L   K   S   K   E   S   N   D   F   N   P   L   P
ccg ctt ggc gga ctg aga caa aca gtg tca caa ata gga gag agt aga gaa aga gag cag ccg aaa gtg gtc gag
 P   L   G   G   L   R   Q   T   V   S   Q   I   G   E   S   R   E   R   E   Q   P   K   V   V   E
gca cag aag aga cca gat tca gca agt tct atg gtg tct tca acc tcg agt ggg gcg ctg tcg aat tgt ggc agc
 A   Q   K   R   P   D   S   A   S   S   M   V   S   S   T   S   S   G   A   L   S   N   C   G   S
aca gat acg aac agt cag agc ggg aat cca tca ctg tgg cca gct tgg gtg tat tgt acg aga tat agc gat cga
 T   D   T   N   S   Q   S   G   N   P   S   L   W   P   A   W   V   Y   C   T   R   Y   S   D   R
cct agt tct ggt ccg aga agt aga aga atg aag aag acg ggg cca agc gta gaa gag aag agg cct aga act gct
 P   S   S   G   P   R   S   R   R   M   K   K   T   G   P   S   V   E   E   K   R   P   R   T   A
ttc agc gct gca caa ctt gga aga cta aag cac gag ttt gcc gag aac cgc tac ctc acc gag cgt cga aga caa
 F   S   A   A   Q   L   G   R   L   K   H   E   F   A   E   N   R   Y   L   T   E   R   R   R   Q
gcc ttg gca gct gag cta ggc ctc gcc gaa gct cag atc aag atc tgg ttc cag aat aaa cga gct aag atc aag
 A   L   A   A   E   L   G   L   A   E   A   Q   I   K   I   W   F   Q   N   K   R   A   K   I   K
aag gcg act ggt cag agg aac ccc cta gca atg cag ctg atg gcc caa ggg ttg tac aat cac agc act gcc aac
 K   A   T   G   Q   R   N   P   L   A   M   Q   L   M   A   Q   G   L   Y   N   H   S   T   A   N
gag agt gat gag gaa gaa gag att aat gtt acg taa
 E   S   D   E   E   E   E   I   N   V   T   *
```

Fig. 5e

Operophtera_inv

```
atg gcg gcg gtc tcc acg cac cta caa gaa tct tct atc aaa ata cag gat gca agc gac gat gag cct tat tcc
 M   A   A   V   S   T   H   L   Q   E   S   S   I   K   I   Q   D   A   S   D   D   E   P   Y   S

cct aac aca aga gac acc acc tca cca gaa tac cat gaa gac gaa aag aca gaa gaa aga tcc ata cat tcg tct
 P   N   T   R   D   T   T   S   P   E   Y   H   E   D   E   K   T   E   E   R   S   I   H   S   S

tct ttt tct ata cat aat gtt ctg aag aag gaa aga gat agt aat agt cct gaa aac gtg ttt tca acg gag aag
 S   F   S   I   H   N   V   L   K   K   E   R   D   S   N   S   P   E   N   V   F   S   T   E   K

ttg ctg cgg aat aca ccc aaa ttt gaa gat agt aga att tta gaa agg aat ttt gaa gat tct aga aat tca gaa
 L   L   R   N   T   P   K   F   E   D   S   R   I   L   E   R   N   F   E   D   S   R   N   S   E

agg aat ttc gaa gat tct aga aat tca gaa agg aat ttc gaa gat tct agg aat tca gag agt gtt tct ccg tta
 R   N   F   E   D   S   R   N   S   E   R   N   F   E   D   S   R   N   S   E   S   V   S   P   L

aac gat gat gtt tcg agg aca gag atc agt gtt gac gat gat aat tct tgt tct agt gat gat act gtc ctg tca
 N   D   D   V   S   R   T   E   I   S   V   D   D   D   N   S   C   S   S   D   D   T   V   L   S

gtt ggc aac gag gct cct gtc agt ttt gac agt gac aag agt caa gac aat cca gga ctg aca tct ttt aag cat
 V   G   N   E   A   P   V   S   F   D   S   D   K   S   Q   D   N   P   G   L   T   S   F   K   H

atc cag acg cat tta aac gct atc tcg caa cta agt caa aac gtc atg aat caa ccc ttg ctt cta cga ccg agc
 I   Q   T   H   L   N   A   I   S   Q   L   S   Q   N   V   M   N   Q   P   L   L   L   R   P   S

cca atc acc ccc aac cct tta atg ttc cta aac cag ccc cta ctt ttc caa aac ccc tta atc aac caa gta gat
 P   I   T   P   N   P   L   M   F   L   N   Q   P   L   L   F   Q   N   P   L   I   N   Q   V   D

cta aaa aca gtt cct aga atg cct gta cct caa aat tta ccc aac caa ttt gga ttg aac ttt ggc ttt cgg aaa
 L   K   T   V   P   R   M   P   V   P   Q   N   L   P   N   Q   F   G   L   N   F   G   F   R   K

acc caa gaa cta aga cga aca gat gag aat cga aga cta tac agg cct aag tca cca gaa aat gag tca ggc aga
 T   Q   E   L   R   R   T   D   E   N   R   R   L   Y   R   P   K   S   P   E   N   E   S   G   R

gat ttt att aac cag aac tgc ctt aaa ttc agt ata gat aat ata ttg aag gca gat ttc gga aga agg atc aca
 D   F   I   N   Q   N   C   L   K   F   S   I   D   N   I   L   K   A   D   F   G   R   R   I   T

gat ccg att aag aga aag cag aag aga tat gag gct aaa gtg tct cct gtg aaa gag gtt cct gta tca aag gca
 D   P   I   K   R   K   Q   K   R   Y   E   A   K   V   S   P   V   K   E   V    P   V   S   K   A

gaa gaa gcc agg gtt cca gaa att aag gct gga ggt gga agt gat aag ggg gcg att gat ctc tct aaa tct gag
 E   E   A   R   V   P   E   I   K   A   G   G   G   S   D   K   G   A   I   D   L   S   K   S   E

gac agt ggg agc aac caa tca gga tca acg aat ggc gac ggc atg gtg tgg cca gcg tgg gtg tac tgt acg agg
 D   S   G   S   N   Q   S   G   S   T   N   G   D   G   M   V   W   P   A   W   V   Y   C   T   R

tac agc gac aga ccc agt tcc gga cga agt ccc cgg acg aga cgg ccg aag aag ccc ccc gga gag acg aac ccc
 Y   S   D   R   P   S   S   G   R   S   P   R   T   R   R   P   K   K   P   P   G   E   T   N   P

aac gat gag aaa cga cca aga acc gcc ttc tct ggg ccc caa ctt gca aga tta aag cac gag ttt gcc gag aac
 N   D   E   K   R   P   R   T   A   F   S   G   P   Q   L   A   R   L   K   H   E   F   A   E   N

cgc tat ctc acc gag cgt cga aga caa gcc ttg gca gct gag cta ggc ctc gcc gaa gct cag atc aag atc tgg
 R   Y   L   T   E   R   R   R   Q   A   L   A   A   E   L   G   L   A   E   A   Q   I   K   I   W

ttc cag aac aaa cga gct aag atc aag aag gcg tct gga cag agg aac ccc cta gct ttg cag ctg atg gcc cag
 F   Q   N   K   R   A   K   I   K   K   A   S   G   Q   R   N   P   L   A   L   Q   L   M   A   Q

ggg ttg tac aat cac agc act gtt cct ctc act aag gag gag gag gaa ttg gag atg aaa gct agg gag agg gag
 G   L   Y   N   H   S   T   V   P   L   T   K   E   E   E   E   L   E   M   K   A   R   E   R   E

gcg cag aat agg gta tag
 A   Q   N   R   V   *
```

Fig. 5f

Dendroctonus_inv

```
atg gac tcc agc agt gat cac ttc gaa cgc aat tca ccc cat att gat caa aac agt tgc tct agt gat gat acg
 M   D   S   S   S   D   H   F   E   R   N   S   P   H   I   D   Q   N   S   C   S   S   D   D   T
gtt ttg tca gtg ggc aat gaa aat gaa aac cag cct aac act gct ccg caa gta gca cca gct gag gag cct gaa
 V   L   S   V   G   N   E   N   E   N   Q   P   N   T   A   P   Q   V   A   P   A   E   E   P   E
tcc aca ttg agt ttc aaa aac att gaa aat cac ttg aac gcc tta tct caa atc aca aat agc act ctg cga aat
 S   T   L   S   F   K   N   I   E   N   H   L   N   A   L   S   Q   I   T   N   S   T   L   R   N
gaa cac act gac act gtt cga gtg tct tct agt cca gta tcc tct tca gcg cat cat cga ctg agt cca tca aac
 E   H   T   D   T   V   R   V   S   S   S   P   V   S   S   S   A   H   H   R   L   S   P   S   N
agc agc aca aag tcg ttt gga gga gac tgt ccg tcg cct caa aac aac tat ctg ttt aaa agt gaa atc ggg ggc
 S   S   T   K   S   F   G   G   D   C   P   S   P   Q   N   N   Y   L   F   K   S   E   I   G   G
ttt aaa agt gaa cat tta ggc tgt ttc ggg ttt cgg agt gaa cag ttc aat ttt agg aca gga gag acc agt tcg
 F   K   S   E   H   L   G   C   F   G   F   R   S   E   Q   F   N   F   R   T   G   E   T   S   S
agt ttg tgc tcg ccc aga tca gtg agg agc gat ggc ggt gac agt cct gga agt cct gga tcg aga tgt caa gca
 S   L   C   S   P   R   S   V   R   S   D   G   G   D   S   P   G   S   P   G   S   R   C   Q   A
gca agc cca ttc tat cac agc cat tcg tcg caa acg aac ccc tca gtc aat cca cca agc cct caa aac agc gac
 A   S   P   F   Y   H   S   H   S   S   Q   T   N   P   S   V   N   P   P   S   P   Q   N   S   D
cca aat atc tca aac gag tct gca gct tcc atc aat cag gaa agc ctg aaa ttc tcc ata gat aat att cta aaa
 P   N   I   S   N   E   S   A   A   S   I   N   Q   E   S   L   K   F   S   I   D   N   I   L   K
gct gat ttc ggc agg agt aaa att cta gac ccg att acc ata cgg aaa agc aaa cct cct tgt cgg aga acg tcg
 A   D   F   G   R   S   K   I   L   D   P   I   T   I   R   K   S   K   P   P   C   R   R   T   S
gcg gaa aag tta agt ggc tta gcg gct ctt cat gtt ggc gaa aag ttg tcg agt ttt ccg ggc gag tac aga ctg
 A   E   K   L   S   G   L   A   A   L   H   V   G   E   K   L   S   S   F   P   G   E   Y   R   L
caa gaa aaa ggg cga acc ctg tca acg tcc ccc ggt att tta aaa agc gac gga tgc gga acc gat gga gaa aaa
 Q   E   K   G   R   T   L   S   T   S   P   G   I   L   K   S   D   G   C   G   T   D   G   E   K
ggt cca gtc gat ctc agt cca aaa gga gat gga gtc tgt gac agt aag gga gag cga gga aaa gat ggg cag ccg
 G   P   V   D   L   S   P   K   G   D   G   V   C   D   S   K   G   E   R   G   K   D   G   Q   P
att ctg tgg cca gca tgg gtt tac tgc acc agg tac tcg gat cga cca agt tca gga aga agt ccg aga act cga
 I   L   W   P   A   W   V   Y   C   T   R   Y   S   D   R   P   S   S   G   R   S   P   R   T   R
cga atc aag aaa cct ggc acg aaa gca gca gtt ccg gaa gag aag cga cct cgt acc gct ttt tct ggg gcg caa
 R   I   K   K   P   G   T   K   A   A   V   P   E   E   K   R   P   R   T   A   F   S   G   A   Q
ctg gcg cga ctt aag aac gag ttt gct gag aac cga tac ctg acg gaa cgt aga cga cag caa ctg agc gct gaa
 L   A   R   L   K   N   E   F   A   E   N   R   Y   L   T   E   R   R   R   Q   Q   L   S   A   E
ttg ggc ctc aat gag gct cag att aaa atc tgg ttc cag aac aaa cga gcc aaa att aag aaa gct tcg ggc cag
 L   G   L   N   E   A   Q   I   K   I   W   F   Q   N   K   R   A   K   I   K   K   A   S   G   Q
aaa aat ccg ttg gct ttg caa tta atg gca caa gga ctg tac aat cac agt acc ata ccg ttg acg aaa gaa gaa
 K   N   P   L   A   L   Q   L   M   A   Q   G   L   Y   N   H   S   T   I   P   L   T   K   E   E
gag gag ttg gag aaa ttg cag tca caa ggg aaa att tcg tag
 E   E   L   E   K   L   Q   S   Q   G   K   I   S   *
```

Fig. 5g

Harpegnathos_inv

```
ATG TCG GGC ACA CCT CCG GAG CCG GTG AGT CTG GAC TCT GAC TTG CCG GAG GCC AGA TCG AGC CAT CTG GCT GGG
 M   S   G   T   P   P   E   P   V   S   L   D   S   D   L   P   E   A   R   S   S   H   L   A   G
TCG CGA TTG GAG CGG GAC GCG GAG GAC GCG GAG AAA CGC GAT GAA ACG CGA GCG GGG GCC GAG TGC AAC GCC GAC
 S   R   L   E   R   D   A   E   D   A   E   K   R   D   E   T   R   A   G   A   E   C   N   A   D
AGC GAC TGC GAG AGC GAC ACG AGC GAG GTG CTG AGC GTC GGT AGC GAG CCA ACG CCG AGC AGC GTG GTT GGG GTG
 S   D   C   E   S   D   T   S   E   V   L   S   V   G   S   E   P   T   P   S   S   V   V   G   V
CGC GTA TGC GGC TCC GTC AGG TAC GAC CGG GAG TCC GCG AGC AGC CGA GGT GAG TTT CGC GAT GAG GAG AAC ACG
 R   V   C   G   S   V   R   Y   D   R   E   S   A   S   S   R   G   E   F   R   D   E   E   N   T
AGA ACA TCG GCG ACG CCG TCG CCG TCC AGC TCC ACC AGC TGC AAC GAC CTC TAC TAC CAG CGT GCA GCC GGC AAT
 R   T   S   A   T   P   S   P   S   S   S   T   S   C   N   D   L   Y   Y   Q   R   A   A   G   N
CAC GTA CCG GCA CCG AGT CCA CCC GGC TAC AGC CAA CCA CCG TCG TCC CCT ACG GCG TCT TCC GTC AGA TCC CCA
 H   V   P   A   P   S   P   P   G   Y   S   Q   P   P   S   S   P   T   A   S   S   V   R   S   P
GCT TCC TCA TCG GCC ACT GCC TCG TCT CCG GGT CGA CCG GAA GCC ACC CAA GAG GCT ATC GTA CCC AGG TAC CAA
 A   S   S   S   A   T   A   S   S   P   G   R   P   E   A   T   Q   E   A   I   V   P   R   Y   Q
TCC AGT CAT CAG CAC CTA CTC CCT CGA TAC TCC TCC GGC AGG GAA GTC GAC ATT TCC GAC TAT CCG GCG CGC GTG
 S   S   H   Q   H   L   L   P   R   Y   S   S   G   R   E   V   D   I   S   D   Y   P   A   R   V
CAA CAC GGT CTG AGC CAC GAG ATC GTT TAC CCG ACC GTC GAG AGG CTG CAC CGT ACA CCC ATC AGC GTA CCC CTA
 Q   H   G   L   S   H   E   I   V   Y   P   T   V   E   R   L   H   R   T   P   I   S   V   P   L
GTG ACG AGA CTC TCG TTG TCG CCG CCG TCG GCC ATG ACG GTC ACC GGG CTC CAG GCG ACG ACG CCC GTT CTC CAT
 V   T   R   L   S   L   S   P   P   S   A   M   T   V   T   G   L   Q   A   T   T   P   V   L   H
CCC GCC GCA TGC AGG GAC CCG CGG GAC AGC ACC ACC TCG TTG ATG CCG CAC CAC AGC AGC CAA CAC CTG CAC CAT
 P   A   A   C   R   D   P   R   D   S   T   T   S   L   M   P   H   H   S   S   Q   H   L   H   H
GCG AAC GCC CAC ACC ACG CAT CTG CAC CAG GGC TCG CAG GTG CAG CAT GTG CCG GCC AAT TCG GTG CAC CAG CAT
 A   N   A   H   T   T   H   L   H   Q   G   S   Q   V   Q   H   V   P   A   N   S   V   H   Q   H
CGG CTG TCG GTC AGC AAG CTG TTG CAA CGG GAG CCC GGC AGC CCG GCG TCA CCC GGC GGC ACC GGG AGG GAG GAG
 R   L   S   V   S   K   L   L   Q   R   E   P   G   S   P   A   S   P   G   G   T   G   R   E   E
AAC GGA CGC ACC TTG GCC GCC GCG AAC GGC CTG CAG AAC AGC ATC GGC CAC AAC AAC GCC AAT CAT CAT CAC CAC
 N   G   R   T   L   A   A   A   N   G   L   Q   N   S   I   G   H   N   N   A   N   H   H   H   H
CAC CAC CAC CAC CAT CAC CAC AAC CAC CAC AAC CAC CAC AAC AAC AAA AAC AAC AAC AAC AAC AAC AAC AAT AAT
 H   H   H   H   H   H   H   N   H   H   N   H   H   N   N   K   N   N   N   N   N   N   N   N   N
CTG CAG CAC CAG GCG GGT CTC AAG TTC AGT ATA GAC AAT ATT CTC AAG GCG GAT TTC GGC CGG AGG ATC ACG GAC
 L   Q   H   Q   A   G   L   K   F   S   I   D   N   I   L   K   A   D   F   G   R   R   I   T   D
CCA ATC TCC CTG AAG AAA TCG CGC CCC AAG AAG GTG GCC TCG CGG CCG ATC GAC CTG ACC AAG GAC TTC CTC GAA
 P   I   S   L   K   K   S   R   P   K   K   V   A   S   R   P   I   D   L   T   K   D   F   L   E
TCG TCC TCC GAC ACT TCC GAG AGG AAC GGC ACG GAA ACG ACG ACG ACG ACG ACG ACG ACG ACG ACC GCG ACG ACC
 S   S   S   D   T   S   E   R   N   G   T   E   T   T   T   T   T   T   T   T   T   T   A   T   T
AAC GCG TCT CCC ACC GGC GTT TCG GCC GGA AAC CCG CCG CCC AAC CCT ACC GGA TCG ACC GGT ACC GAC CCT GGC
 N   A   S   P   T   G   V   S   A   G   N   P   P   P   N   P   T   G   S   T   G   T   D   P   G
AAG ATG TTG TGG CCG GCG TGG GTC TAC TGC ACC AGA TAC TCG GAC AGG CCC TCC TCG GGA CGA AGT CCG CGC ACG
 K   M   L   W   P   A   W   V   Y   C   T   R   Y   S   D   R   P   S   S   G   R   S   P   R   T
AGA CGA GTG AAG AGA ACG GCC GAC GGA CGC GGC GGT GGC ACC CCC GAG GAG AAA CGT CCC CGG ACG GCG TTC AGC
 R   R   V   K   R   T   A   D   G   R   G   G   G   T   P   E   E   K   R   P   R   T   A   F   S
GGC GAG CAA CTC GCG AGG CTC AAG CGG GAG TTC GCG GAG AAC CGA TAC CTG ACG GAG CGA CGG CGG CAG CAG CTC
 G   E   Q   L   A   R   L   K   R   E   F   A   E   N   R   Y   L   T   E   R   R   R   Q   Q   L
TCG AGG GAT CTC GGT CTG AAC GAG GCG CAG ATC AAG ATC TGG TTT CAG AAC AAG AGG GCG AAA ATC AAG AAG GCC
 S   R   D   L   G   L   N   E   A   Q   I   K   I   W   F   Q   N   K   R   A   K   I   K   K   A
AGC GGT CAG AAG AAT CCG CTG GCG CTT CAG CTG ATG GCG CAG GGG CTC TAC AAT CAT TCG ACG GTA CCG CTT ACG
 S   G   Q   K   N   P   L   A   L   Q   L   M   A   Q   G   L   Y   N   H   S   T   V   P   L   T
AAG GAG GAA GAG GAG CAG GCC GCG GAG CTC CAA GCG AAA TGA
 K   E   E   E   E   Q   A   A   E   L   Q   A   K   *
```

Fig. 5h

Pediculus_en1

```
ATG GCA TTA GAA GAT AGA TGC AGT CCT TCG AGC GCG TCA ACT CCA GGT CCA AAA GCT TCG AGC GAT CGT CCT GGA
 M   A   L   E   D   R   C   S   P   S   S   A   S   T   P   G   P   K   A   S   S   D   R   P   G
TCC GAC GGA AAT GCG GTT CGG GTA ACG TCT CCT CCT ACT CCA TCA TCG GTA AAG GAT AAT GAT AAT CGT CAA ACG
 S   D   G   N   A   V   R   V   T   S   P   P   T   P   S   S   V   K   D   N   D   N   R   Q   T
AAA TTT GAC GAG GGT TAC GAA TTA AAG AGA AAA ATT AAA ATG GAA CCC GAT GTC GAA GTA GAT AAT TAC GAA GAA
 K   F   D   E   G   Y   E   L   K   R   K   I   K   M   E   P   D   V   E   V   D   N   Y   E   E
AAA AAA ATT AGA ATT TGG AGA AAA ATA TCG GAT GAG GAA CAT GAT GAT TCG AAA TCA TCA AAT TCC GAT TTG GAA
 K   K   I   R   I   W   R   K   I   S   D   E   E   H   D   D   S   K   S   S   N   S   D   L   E
AAA CGT CTT TCT CCA GGA GGT TAT GTT AAA ACG ACG TCT TTT TCT CCG GAA ATA ATT CAA AGG TTT TCA TCG TAT
 K   R   L   S   P   G   G   Y   V   K   T   T   S   F   S   P   E   I   I   Q   R   F   S   S   Y
TCG ATC GAA AGG TTC GTC AAT TCG GAT ATC GGT CGT TGT GGC GGT GGT GAA AAC AGT TCG GGA AAT GGA ATT ATA
 S   I   E   R   F   V   N   S   D   I   G   R   C   G   G   G   E   N   S   S   G   N   G   I   I
CAA TCT AAC GGA CAA TAT CCG ATG GAT GTT AAT AAT GGT AAT TAT TTA AAT GTT CCT CTT TTC ATG TCG AAA ACG
 Q   S   N   G   Q   Y   P   M   D   V   N   N   G   N   Y   L   N   V   P   L   F   M   S   K   T
CAT AAA AAT GTA ACA ATG TTG AAT TCG TCG CAT CAA AAA GTT AGC AGG ACA CCA GAA AAC GGT AGG CTA TCA TCG
 H   K   N   V   T   M   L   N   S   S   H   Q   K   V   S   R   T   P   E   N   G   R   L   S   S
TCA GAT AAA AAA TCA CAA AAG CGA ACC GAT TCG ATC GAT GAG GAA CCT ACG GAA ACG AAT AAT AAT AAC GGG AGT
 S   D   K   K   S   Q   K   R   T   D   S   I   D   E   E   P   T   E   T   N   N   N   N   G   S
AAT ACA AAT AAC AAC AGC AAT AAT AAT AAT AAT AAC AAT AAT AAT AGC AAT AAC AAC AAC AAT AAC GAA ACA AGA
 N   T   N   N   N   S   N   N   N   N   N   N   N   N   N   S   N   N   N   N   N   N   E   T   R
ATA AAA TTT TCG GTT GAA GAT ATA TTA AAA CCG GAT TTC GGT TCA AAA TAC ATT CAG AAA AAT TGT GAA ATA TGG
 I   K   F   S   V   E   D   I   L   K   P   D   F   G   S   K   Y   I   Q   K   N   C   E   I   W
AAT CCG TTA AAG CGA ACG GTA ACA CCG ATC ATA AGA ACA AGT GAC GAA CTA AGA AAA CCT AAT AGA ACT GTA CAA
 N   P   L   K   R   T   V   T   P   I   I   R   T   S   D   E   L   R   K   P   N   R   T   V   Q
AAT AGA ACA AAT CGA AGT TTC GAT ATA GCA AGA TTA ACG GAA ACA GAT TCG GTA AAA AAT TCA TCT CAG GAA GAA
 N   R   T   N   R   S   F   D   I   A   R   L   T   E   T   D   S   V   K   N   S   S   Q   E   E
TAC GGT AAA AGA AGG CAT AGC GAT GCG GTA CCC ATA GAG CCA CCA AAA TTA CAA GAA AGA CGA AAA ACG ACA GAA
 Y   G   K   R   R   H   S   D   A   V   P   I   E   P   P   K   L   Q   E   R   R   K   T   T   E
TCG GAT GTC CCG AGA TTG TTA CCG GAA GTA ACG AAA GAA AGT CCA ACG AGG ATA CCA TTA CCA TCA CCG GCA GCT
 S   D   V   P   R   L   L   P   E   V   T   K   E   S   P   T   R   I   P   L   P   S   P   A   A
AGT TCA TCG TCA GAA GGA GAT CAA AGC ATA GGT GGT AAA GGT ACC GAA TTA TGG CCA GCA TGG GTT TAT TGT ACC
 S   S   S   S   E   G   D   Q   S   I   G   G   K   G   T   E   L   W   P   A   W   V   Y   C   T
AGA TAC TCG GAT CGT CCA TCT TCA GGT CCG AGA TCG AGA AGG ATA AAA CGT AAG GAT AAA AAA CCA GAA GAA AAA
 R   Y   S   D   R   P   S   S   G   P   R   S   R   R   I   K   R   K   D   K   K   P   E   E   K
CGA CCT AGA ACG GCA TTT TCC GGA GAT CAA TTG TCG AGG CTT AAA CAC GAA TTC GCG GAA AAT AGG TAC CTG ACG
 R   P   R   T   A   F   S   G   D   Q   L   S   R   L   K   H   E   F   A   E   N   R   Y   L   T
GAA AGG AGG AGA CAA GAT CTC GCT AAA GAA TTA GGA CTC AAT GAA GCT CAA ATA AAA ATA TGG TTT CAA AAC AAA
 E   R   R   R   Q   D   L   A   K   E   L   G   L   N   E   A   Q   I   K   I   W   F   Q   N   K
AGG GCG AAA ATG AAA AAA GCA AAA GGG GAA AAA AAT CCA CTA GCC CTT CAA TTA ATG GCT CAG GGG TTG TAT AAT
 R   A   K   M   K   K   A   K   G   E   K   N   P   L   A   L   Q   L   M   A   Q   G   L   Y   N
CAC AGT ACG ATA CCC GTT GAC GAA GAT GAA TAC CTG GAA GAA ATG GCG GCG GCT TCA AAT CAA TCA AAT CCC GTT
 H   S   T   I   P   V   D   E   D   E   Y   L   E   E   M   A   A   A   S   N   Q   S   N   P   V
TGA
 *
```

Fig. 5i

Pediculus_en2

```
ATG TCT CCG GTG TTG AAT TGT ATT CCG TCT AGT CCT CCT TCG GAT CAA GAC AAT TAT CCA AAA GTC GGA AGC ATA
 M   S   P   V   L   N   C   I   P   S   S   P   P   S   D   Q   D   N   Y   P   K   V   G   S   I
CCG TCA GAT AGA TCC CAT CGG TCT CCT TTA GAG GTT AGT CCG ACT TCG CAA ACA AAT GGA GAT TTG ACG GAG TCG
 P   S   D   R   S   H   R   S   P   L   E   V   S   P   T   S   Q   T   N   G   D   L   T   E   S
GAA AAT TCA TGC TGT GAT GAT GTC AGA ACG GAA ACG GAG GAT TTG AAA CGT TCA AAC ATA ACT AAT AAT AAT AAT
 E   N   S   C   C   D   D   V   R   T   E   T   E   D   L   K   R   S   N   I   T   N   N   N   N
AAT AAT AAT ATA CCG ATA AAT TAT AAA AAA TCT CCT AGT CCG AAT AAA AAT TTT ACT TTT GAA ACT CCC ATA AGG
 N   N   N   I   P   I   N   Y   K   K   S   P   S   P   N   K   N   F   T   F   E   T   P   I   R
CCC TGG AAT TCT AGT CCG AGT TAT AGA AAT GTC GAT AGA GGA CAA TCG GTG AGT CCG TCA AAT TCT AGC CAA CGT
 P   W   N   S   S   P   S   Y   R   N   V   D   R   G   Q   S   V   S   P   S   N   S   S   Q   R
TTA TCA CCC TGT CCG AGA AAT TCC CTT TCT CCG GAT GCA TCA TCG GAA AAT TCT AGA ATT CCG AGT CAG GCA TCG
 L   S   P   C   P   R   N   S   L   S   P   D   A   S   S   E   N   S   R   I   P   S   Q   A   S
TCA GAA GCT ACT CAC GAT GCC GCC GTT TTA CCA CCC GAT AAT TTT AGA TCT CCT AAA ATG TGC TCG ACG GAG CTA
 S   E   A   T   H   D   A   A   V   L   P   P   D   N   F   R   S   P   K   M   C   S   T   E   L
CCT TAC GAT GTT CCT TTT CCG TTT CAT GAC AAT CGT CTT TCA TCC GCA GAT AGT TTT AGG TTT CCA AAT TCT TAC
 P   Y   D   V   P   F   P   F   H   D   N   R   L   S   S   A   D   S   F   R   F   P   N   S   Y
ATA TGT GAT CCG TTA AAT CCT TAT TTC GGA TTA GGT AAT CAT CCT CTA AAT CCA TTT CAA CCT CCT CCT AAT TTT
 I   C   D   P   L   N   P   Y   F   G   L   G   N   H   P   L   N   P   F   Q   P   P   P   N   F
TTG GCA CAT CCG TTT TTG GGT ACG ACC GAT CCG AAA TTG TTA TTA AAC AAT TCG GGT TTC GGT TTA ATA TCT CAA
 L   A   H   P   F   L   G   T   T   D   P   K   L   L   L   N   N   S   G   F   G   L   I   S   Q
CAA CTT CAC AAT TTA AAC GGT TTT CCT CCA AAC GTG AAT AAT AAA CTT CCA CCG GCG GAT TTG CAC AGA ATG TCA
 Q   L   H   N   L   N   G   F   P   P   N   V   N   N   K   L   P   P   A   D   L   H   R   M   S
ACA GTC CAA CAA TTA CAA TTT ATA CAA AAT CAT CTT CAA AGT CTG CCA ACG TTT TTA CAA CCC TCG TCG CCG TAT
 T   V   Q   Q   L   Q   F   I   Q   N   H   L   Q   S   L   P   T   F   L   Q   P   S   S   P   Y
ACA AAA TTA ACA GAA TCA AAA AAT CAA TTG TCG ATG TTG ATA CAA GCG AGG AAA CAA CCG AAA AAT TCC GAT TCG
 T   K   L   T   E   S   K   N   Q   L   S   M   L   I   Q   A   R   K   Q   P   K   N   S   D   S
GAT GCG GAA AAT GTT AAA AAA GAC ATT TTC CCT TTT TCC GAT GTG AGA AAA ATA GAA GAA GTA CAT AAT AAA ACA
 D   A   E   N   V   K   K   D   I   F   P   F   S   D   V   R   K   I   E   E   V   H   N   K   T
AAA GAT GTT AAA ATC AAA ACA AAC GAA ACA TTA GAA AAA AAT AGA AAA ACA TTA ACA GAA GAA CAT GAA GAA GAA
 K   D   V   K   I   K   T   N   E   T   L   E   K   N   R   K   T   L   T   E   E   H   E   E   E
GAA GAA GAT GGA GGA GGG GGA GGA GGA GAA GAA GTA GAA GAA ATG TCC GTA GAC AAT TAC GAA GAA AAC ATT TCG
 E   E   D   G   G   G   G   G   G   E   E   V   E   E   M   S   V   D   N   Y   E   E   N   I   S
GAT TCC GAT GAA CTT TTA AGC GTC GGA AGT GTT TCA CCA TCT CCG AGT AGT TGT CAA AAC AAA CAA AAT AAA ATA
 D   S   D   E   L   L   S   V   G   S   V   S   P   S   P   S   S   C   Q   N   K   Q   N   K   I
AAA CAA AAC GAT GGA AGT TTA TGT TCA CCT TCT TTA AAT TCT TCT TCG TCT TTA CAT ATA CAA AAT TCA AAT AAT
 K   Q   N   D   G   S   L   C   S   P   S   L   N   S   S   S   S   L   H   I   Q   N   S   N   N
AAT AAT AAT AAT AAT AAT CCG ACG GGG GGA TTT AAA AAA TAT CAA AAT TCG ACG TCA CCC GTG AGA ACG GTT CAA
 N   N   N   N   N   N   P   T   G   G   F   K   K   Y   Q   N   S   T   S   P   V   R   T   V   Q
GAG ACA TTA ACA AAC ACG AAA AGT TCA AAC GCG ACA AAT AGA ACA TTA AAA TTT AGC ATA GAC AAC ATA CTT AAA
 E   T   L   T   N   T   K   S   S   N   A   T   N   R   T   L   K   F   S   I   D   N   I   L   K
TCC GAT TTC GGT GGT GGT AAC GAT GAT ATT TTC AAA GAA CCC AAA CAA ATA GAA AAT GGG AAA AAA TTA AAA GAA
 S   D   F   G   G   G   N   D   D   I   F   K   E   P   K   Q   I   E   N   G   K   K   L   K   E
CAA CCG GAA AAA ACA AAT GTT ATC GTC GTT AAT AAA AAA CCG GAA ACG GAA AAA GAA AAA CCC GTG GAT TTA AGT
 Q   P   E   K   T   N   V   I   V   V   N   K   K   P   E   T   E   K   E   K   P   V   D   L   S
CAA GAC GGT GGT ACG ACG AAT TCA CCA TCA TCG GGT ACG GAT GCA CCA ATG TTA TGG CCT GCT TGG GTA TAT TGT
 Q   D   G   G   T   T   N   S   P   S   S   G   T   D   A   P   M   L   W   P   A   W   V   Y   C
ACC AGA TAC AGC GAC AGA CCA TCA TCA GGA AGA AGC CCG AGA ACT AGA AGA TCT AAA AAT AAA GAT AAA AAT TCA
 T   R   Y   S   D   R   P   S   S   G   R   S   P   R   T   R   R   S   K   N   K   D   K   N   S
GAT GAA AAA CGA CCG AGA ACG GCA TTT TCC GGG GAT CAA TTA TCG AGG TTA AAA CAC GAA TTC GCG GAA AAT AGG
 D   E   K   R   P   R   T   A   F   S   G   D   Q   L   S   R   L   K   H   E   F   A   E   N   R
TAC CTG ACG GAA AGG AGG AGA CAA GAT TTG GCG AGA GAA TTG GGA CTG AAC GAA GCT CAA ATA AAA ATA TGG TTC
 Y   L   T   E   R   R   R   Q   D   L   A   R   E   L   G   L   N   E   A   Q   I   K   I   W   F
CAA AAT AAA AGG GCG AAA ATG AAA AAG GCA AGA GGA GAA AAA AAT CCA TTG GCG CTA CAG CTC ATG GCT CAA GGC
 Q   N   K   R   A   K   M   K   K   A   R   G   E   K   N   P   L   A   L   Q   L   M   A   Q   G
CTT TAT AAT CAT AGT ACC ATA CCT TTG ACT AAA GAA GAA GAA GAA GCG GCG GCG GCG GAA CTC AGC GAT TGA
 L   Y   N   H   S   T   I   P   L   T   K   E   E   E   E   A   A   A   A   E   L   S   D   *
```

Fig. 5j

Zootermopsis_en1

```
atg gct ctt gaa acg gat cgc tgt agc ccc agc agc gcc tcg agc cca ggc ccc aca tca agt acg agg ccg gga
 M   A   L   E   T   D   R   C   S   P   S   S   A   S   S   P   G   P   T   S   S   T   R   P   G
agc gat ggc tcg tct ccc gcc aac ggg tcg gga acc cca gac tct tgc acg agt ctg tgt tgc aac ggc aaa atc
 S   D   G   S   S   P   A   N   G   S   G   T   P   D   S   C   T   S   L   C   C   N   G   K   I
ccg cag cct cct tca gtt ctt agc tca ccg tct gtt tcc cac ggg aca cct aat ctt aca cac cac aca ccc cct
 P   Q   P   P   S   V   L   S   S   P   S   V   S   H   G   T   P   N   L   T   H   H   T   P   P
tcg cac acc acg aac cac caa gaa caa ccg tgt cat tgc tgt ggg ccg gtg tca ccc cct gcg tat gct gat aac
 S   H   T   T   N   H   Q   E   Q   P   C   H   C   C   G   P   V   S   P   P   A   Y   A   D   N
atc ctt cac aac aag ccc aca gcg tgc ata cca cga tcc cct gat tca acc cca tcg tcc cct agt cat aat cac
 I   L   H   N   K   P   T   A   C   I   P   R   S   P   D   S   T   P   S   S   P   S   H   N   H
ccg tct tta tgc aac ccg cca ttt tca tct ccg tcc cct agt gag cgg cat aac cag ccc ttg aaa ccc ctt ccg
 P   S   L   C   N   P   P   F   S   S   P   S   P   S   E   R   H   N   Q   P   L   K   P   L   P
tcg cca cca agt aat ggg tct agg acc tcg ttc tca tct gtg tcc ccc caa tcg tct ctt cct tca tcc cct agt
 S   P   P   S   N   G   S   R   T   S   F   S   S   V   S   P   Q   S   S   L   P   S   S   P   S
cat aac cgg cat tgc caa gct tct cca ccg caa caa cac ctg ggg tct tcc ccc agt ggc ccg tca ccc cgt ccc
 H   N   R   H   C   Q   A   S   P   P   Q   Q   H   L   G   S   S   P   S   G   P   S   P   R   P
gcg aac aat gaa acg gtg ccc caa ccc acg aca ttg cct ttc tcg gta gcc aac atc tta aga cct gat ttc ggt
 A   N   N   E   T   V   P   Q   P   T   T   L   P   F   S   V   A   N   I   L   R   P   D   F   G
cga cga gca gtg ata aca tca aag cag caa gaa cca tta ttt agg cct gga ggc atc cgt cgc act gta aca ccg
 R   R   A   V   I   T   S   K   Q   Q   E   P   L   F   R   P   G   G   I   R   R   T   V   T   P
ata tgt gat tac caa aga ctt tat cgg cct cat gaa cac tta ccg gga atc cag ccc ttg ccg gcg cca aag cca
 I   C   D   Y   Q   R   L   Y   R   P   H   E   H   L   P   G   I   Q   P   L   P   A   P   K   P
acg aag aag cag aca tcg gct tgg cca tca gtg gta aga aat cca acg ctt cct aca gaa gac gtt agt ttc aaa
 T   K   K   Q   T   S   A   W   P   S   V   V   R   N   P   T   L   P   T   E   D   V   S   F   K
gtc tct gcg cgg gag act tct aac cac cag cag gcg agg agc aaa ggg gtt tcg acg cca gtt ccg acg tcg ggc
 V   S   A   R   E   T   S   N   H   Q   Q   A   R   S   K   G   V   S   T   P   V   P   T   S   G
gcc gcg agc cct ccg ctg tca cct gcc agc agc aca gta tcg gct tcg tcg tca aac ccg gat gac aaa gtg ggc
 A   A   S   P   P   L   S   P   A   S   S   T   V   S   A   S   S   S   N   P   D   D   K   V   G
aac gcc gac tgt gcc aaa ggg tct cag ctg tgg ccg gct tgg gtc tac tgc acg cgc tat tcc gac aga cct tcg
 N   A   D   C   A   K   G   S   Q   L   W   P   A   W   V   Y   C   T   R   Y   S   D   R   P   S
tct ggt ccg aga tct cgt cga cta aag aga aag gag aag aag ccg gaa gag aag aga cca cga aca gcg ttt agc
 S   G   P   R   S   R   R   L   K   R   K   E   K   K   P   E   E   K   R   P   R   T   A   F   S
ggc gag cag ttg gca cgc ctg aaa cac gag ttc acc gag aat cgt tac ctg act gag aga cgc cga aca gaa ctg
 G   E   Q   L   A   R   L   K   H   E   F   T   E   N   R   Y   L   T   E   R   R   R   T   E   L
gcg cga gaa ctg ggc ctg aac gag gcc cag atc aag atc tgg ttc cag aac aaa cgc gcc aag atc aag aag gcg
 A   R   E   L   G   L   N   E   A   Q   I   K   I   W   F   Q   N   K   R   A   K   I   K   K   A
agc gga cag aag aat ccg ctg gcg ctg cag ctc atg gcc cag ggg ctg tac aac cac agc act gtg cct gta gac
 S   G   Q   K   N   P   L   A   L   Q   L   M   A   Q   G   L   Y   N   H   S   T   V   P   V   D
gag gaa gaa gaa gaa gca aac gca ctc ttg ttg gcc aat aat gca cgg caa gac tga
 E   E   E   E   E   A   N   A   L   L   L   A   N   N   A   R   Q   D   *
```

Fig. 5k

Zootermopsis_en2

```
 ga aga agt cct cgt tcg aga cgt ctt aaa agg aaa gat aag aag ccg gaa gag aag aga cca cga aca gcg ttt
     R   S   P   R   S   R   R   L   K   R   K   D   K   K   P   E   E   K   R   P   R   T   A   F
agc ggc gag cag ttg gca cgc ctg aaa cac gag ttc acc gag aat cgt tac ctg act gag aga cgc cga aca gaa
 S   G   E   Q   L   A   R   L   K   H   E   F   T   E   N   R   Y   L   T   E   R   R   R   T   E
ctg gcg cga gaa ctg ggc ctg aac gag gcc cag atc aag atc tgg ttc cag aac aaa cgc gcc aag atc aag aag
 L   A   R   E   L   G   L   N   E   A   Q   I   K   I   W   F   Q   N   K   R   A   K   I   K   K
gcg agc gga cag aag aat ccg ctg gcg ctg cag ctc atg gcc cag ggg ctg tac aac cac agc acc att ccg atg
 A   S   G   Q   K   N   P   L   A   L   Q   L   M   A   Q   G   L   Y   N   H   S   T   I   P   M
aca aga gaa gag gag gaa caa gct gcc gct gcg gag gca aac gca ggc aaa tag
 T   R   E   E   E   E   Q   A   A   A   A   E   A   N   A   G   K   *
```

Fig. 5l

Ixodes_en

```
ATG GCG CAG GAC ATC GAC AAG CAG CAG CAG CCG ACC AGC CCC GAC CGG GCG ACC CCT TCT AAC GGA CTC AAG GCG
 M   A   Q   D   I   D   K   Q   Q   Q   P   T   S   P   D   R   A   T   P   S   N   G   L   K   A
ACG GCC GCC GGG GCG GCG GCT TCG CCG CGC GGC GGC AAC GCG GCG GTG GTC AGC GTG GCG GCC CCG GCG CCG CCG
 T   A   A   G   A   A   A   S   P   R   G   G   N   A   A   V   V   S   V   A   A   P   A   P   P
TCG CCG CCG CCG ACG CAG CAG CAG CTC AAG TTC TCG ATC GAC AGG ATC CTC TCG CCC GAG TTC GGG CCC AGG GCC
 S   P   P   P   T   Q   Q   Q   L   K   F   S   I   D   R   I   L   S   P   E   F   G   P   R   A
AAC GGC AGG CAC CAG GGC AGG GCC CGC AAG GAG GCC GCG GCC GCC GCG AGA CGC GAG GCC GCC GCG GCC GCC GCC
 N   G   R   H   Q   G   R   A   R   K   E   A   A   A   A   A   R   R   E   A   A   A   A   A   A
GCT GCG ACG GAC TCG TCC GGT GAC GAT GCC AGA TCT GTG GGC AAG GCG GAC CTG CCG GTC AAC GGC GCC AAG GCC
 A   A   T   D   S   S   G   D   D   A   R   S   V   G   K   A   D   L   P   V   N   G   A   K   A
CCC GGC ACG CCG GGG CTC CTA TGG CCG GCC TGG GTG TAC TGC ACG CGA TAC TCG GAC AGG CCT TCT TCA GGG CCC
 P   G   T   P   G   L   L   W   P   A   W   V   Y   C   T   R   Y   S   D   R   P   S   S   G   P
CGT TCC CGG CGG ATG AAG AAG AAG GAG AAG AAG GCG GAC GAG AAG CGG CCC CGG ACG GCC TTC ACG GCA GAC CAG
 R   S   R   R   M   K   K   K   E   K   K   A   D   E   K   R   P   R   T   A   F   T   A   D   Q
CTG GCG CGG CTC AAG CAG GAG TTC ACG GAG AAC CGC TAC CTG ACC GAG AAG CGG CGC CAG GAC CTG GCC CGG GAG
 L   A   R   L   K   Q   E   F   T   E   N   R   Y   L   T   E   K   R   R   Q   D   L   A   R   E
CTG AAA CTC AAC GAG TCC CAG ATC AAG ATC TGG TTC CAG AAC AAG CGC GCC AAG ATC AAG AAG GCG AGC GGA CAG
 L   K   L   N   E   S   Q   I   K   I   W   F   Q   N   K   R   A   K   I   K   K   A   S   G   Q
CGC AAC CCG CTG GCC CTC CAG CTC ATG GCC CAG GGG CTC TAC AAT CAC ACC ACG GCC TCG CAG CAG GGC ATG GGG
 R   N   P   L   A   L   Q   L   M   A   Q   G   L   Y   N   H   T   T   A   S   Q   Q   G   M   G
GAC GAC GAT GAC AAC TCG TCC TCG TGA
 D   D   D   D   N   S   S   S   *
```

Fig. 6 Aligned amino acid sequences for five engrailed-homology regions of arthropod engrailed-family proteins. Engrailed-homology regions for EH1 to EH5 are designated by arrows above the alignment. These aligned sequence regions are continuous from EH2 to the C-terminal, but EH1 is positioned on the N-terminal side, apart from the EH2. The regions indicated by red bars above the alignment were used in the GENECONV and phylogenetic analyses.

Fig. 6

```
                           ← EH1 →              ←—— EH2 ——→                                     ← EH3
    Drosophila_en      SLAFSISNILSDRFGDV  ...  MWPAWVYCTRYSDRPSSG--PR------------------------YRRPKQP
   Drosophila_inv      ALKFSIDNILKADFGSR  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------ARKPKKP
     Ceratitis_en      SLAFSISNILSDRFGGN  ...  VWPAWVYCTRYSDRPSSG--PR------------------------YRRPKQP
    Ceratitis_inv      ALKFSIDNILKADFGRR  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------TRKPKKL
       Lucilia_en      TLAFSISNILSDRIGHQ  ...  MWPAWVFCTRYSDRPSSG--PR------------------------YRRPKLP
      Lucilia_inv      ALKFSIDNILKADFGRS  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------VRKPKAP
         Musca_en      TLAFSISNILSDRFGGV  ...  MWPAWVFCTRYSDRPSSG--PR------------------------YRRPKQP
        Musca_inv      ALKFSIDNILKGDFGRG  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------VRKPKKA
         Aedes_en      SITFSISNILSDTFGKT  ...  LWPAWVYCTRYSDRPSSG--PR------------------------YRRTKPP
     Anopheles_en      SISFSITNILSDRFGKA  ...  LWPAWVYCTRYSDRPSSG--PR------------------------YRRTKQP
    Anopheles_inv      SLKFSIDNILKADFGRR  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------TRKPKKS
       Papilio_en      PIAFSISNILHPEFGLS  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRMKKQ
      Papilio_inv      CLKFSIDNILKADFGRR  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------TRRPKKP
        Danaus_en      PIAFSISNILRPEFGVT  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRVKKK
       Danaus_inv      CLKFSIDNILKADFGRR  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------SRRVKKK
        Bombyx_en      PVAFSINNILHPEFGLN  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRVKKK
       Bombyx_inv      CLKFSIDNILKADFGRR  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------TRRPKKP
   Operophtera_en      PVAFSISNILHPEFGLN  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRMKKT
  Operophtera_inv      CLKFSIDNILKADFGRR  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------TRRPKKP
      Amyelois_en      QIGFSISNILHPEFGLN  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRMKKK
     Amyelois_inv      CLKFSIDNILKADFGRR  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------TRRPKKP
      Plutella_en      PITFSISNILHPEFGSG  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRSKKK
     Tribolium_en      TLKYSIRNILKPEFGKN  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRMKKP
    Tribolium_inv      NLKFSIDNILKADFGRR  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------TRRVKKP
  Dendroctonus_en      TLKYSITNILQPDFGKN  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRIKKP
 Dendroctonus_inv      SLKFSIDNILKADFGRS  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------TRRIKKP
          Apis_en      PLRFSVVNILRPDFGRE  ...  PWPAWVYCTRYSDRPSSG--PR------------------------TRRVKRS
         Apis_inv      GLKFSIDNILKADFGRR  ...  PWPAWVYCTRYSDRPSSGRSPR------------------------TRRVKRS
     Megachile_en      PLRFSVVNILRPDFGRE  ...  PWPAWVYCTRYSDRPSSG--PR------------------------TRRVKRT
    Megachile_inv      GLKFSIDNILKADFGRR  ...  PWPAWVYCTRYSDRPSSGRSPR------------------------TRRVKRS
   Ceratosolen_en      TLRFSIVNILRPDFGKE  ...  LWPAWIYCTRYSDRPSSG--PR------------------------TRRVKRS
  Ceratosolen_inv      NLKFSIDNILKADFGRR  ...  QWPAWVYCTRYSDRPSSGRSPR------------------------TRRTKVK
   Microplitis_en      PLRFSIVNILRPEFGRS  ...  VWPAWVYCTRYSDRPSSG--PR------------------------TRRVKRP
  Microplitis_inv      NLKFSIDNILKADFGRR  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------TRRVKRT
       Nasonia_en      PLRFSVVNILRPDFGRD  ...  LWPAWIYCTRYSDRPSSG--PR------------------------TRRVKRT
      Nasonia_inv      SLKFSIDNILKADFGRR  ...  QWPAWVYCTRYSDRPSSGRSPR------------------------TRRPKRT
    Camponotus_en      SLRFSVVNILKPDFGRE  ...  LWPAWVYCTRYSDRPSSG--PR------------------------TRRVKRS
   Camponotus_inv      GLKFSIDNILKGDFGRR  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------TRRVKRT
  Harpegnathos_en      PLRFSVSNILKPDFGLK  ...  QWPAWVYCTRYSDRPSSG--PR------------------------TRRVKRS
 Harpegnathos_inv      GLKFSIDNILKADFGRR  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------TRRVKRT
       Orussus_en      SLRFSVVNILRPDFGRE  ...  LWPAWVYCTRYSDRPSSG--PR------------------------TRRVKRS
      Orussus_inv      SLKFSIDNILKADFGRR  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------TRRVKRS
       Athalia_en      PLRFSVLNILRPDFGRK  ...  LWPAWVYCTRYSDRPSSG--PR------------------------TRRVKRS
      Athalia_inv      NLKFSIDNILKADFGRR  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------TRRVKRS
    Pediculus_en1      RIKFSVEDILKPDFGSK  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRIKRK
    Pediculus_en2      TLKFSIDNILKSDFGGG  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------TRRSKNK
 Schistocerca_en1      GLPFSVANILRPDFGRR  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRLKRK
 Schistocerca_en2      PLSFSIENILRPEFGKR  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------SRRLKRN
  Periplaneta_en1      SLPFSVANILKPDFGRR  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRLKRK
  Periplaneta_en2      ALKFSIDNILKPDFGRQ  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------SRRMKRK
 Zootermopsis_en1      TLPFSVANILRPDFGRR  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRLKRK
   Pedetontus_en1      SLPFSVENILKPEFGRR  ...  VWPAWVYCTRYSDRPSSG--PR------------------------SRRMKKR
   Pedetontus_en2      TLKFSIDNILKPDFGFT  ...  VWPAWVYCTRYSDRPSSGRSPR------------------------SRRTKKK
      Daphnia_en1      VLPFSVENILKPEFGRN  ...  KWPAWIFCTRYSDRPSSG--PR------------------------LRRTKKD
      Daphnia_en2      IIKFSIDNILNPEFGNR  ...  LWPAWVYCTRYSDRPSSGRSPR------------------------ARRIRTK
       Artemia_en      PLAFSIDSILRPDFGKE  ...  KWPAWVFCTRYSDRPSSGRSPR------------------------CRRMKKD
       Caligus_en      SLPFSIDNILKPSFGSA  ...  LWPAWVFCTRYSDRPSSG--PR------------------------ARKMKKK
       Argulus_en      SLNFSIDNILKPDFGLV  ...  LWPAWVFCTRYSDRPSSG--PR------------------------CRKIKRS
Sacculina_en-a.E9      HLNFSIDNILRPDFGRQ  ...  KWPAWVYCTRYSDRPSSG--PR------------------------SRRVSRK
Sacculina_en-a.E20     NLNFSIDNILRPDFGRQ  ...  KWPAWVYCTRYSDRPSSG--PR------------------------SRRVSRK
   Sacculina_en-b      PLNFSIDNILRPDFCLA  ...  KWPAWVYCTRYSDRPSSG--PR------------------------IRKIKKQ
   Cupiennius_en1      PLKFSIEKILSADFGRR  ...  LYPAWIYCSRISDRPSSG--PRRIRSK-AGKGSSSQDLSDDDQSPRARRIKKK
   Stegodyphus_en      SLKFSIEKILSPDFGRR  ...  MWPAWIFCTRYSDRPSSG--PRRIRSKHPKKDSGGQDFTDDNGSPRSRRLKKK
  Parasteatoda_en      TLAFSIEKILSPDFGKR  ...  SLPVWVFCTRYSARASSG--PR------------------------SRRMKKT
        Ixodes_en      QLKFSIDRILSPEFGPR  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SRRMKKK
   Metaseiulus_en      DLKFSIEKILSPAFGND  ...  LWPAWVYCTRYSDRPSSG--PR------------------------SKRTKRK
  Archegozetes_en      SLKFSIENILSPEFGKN  ...  VWPAWVYCTRYSDRPSSG--PR------------------------SRKMKKK
   Limulus_en-1Ba      SLKFSIEKILSPDFGRH  ...  VWPAWVYCTRYSDRPSSG--PR------------------------SRRMKKK
   Limulus_en-1Bb      ALKFSIEKILSPDFGRH  ...  VWPAWVYCTRYSDRPSSG--PR------------------------SRRIKKK
   Limulus_en-1Bc      SLKFSIENILSPDFGRA  ...  VWPAWVFCTRYSDRPSSG--PR------------------------MRKLKSK
   Limulus_en-1Bd      SLKFSIENILSPDFGKA  ...  PWPAWVFCTRYSDRPSSG--PR------------------------LRKKKTS
```

Fig. 6 continued

EH3 → ← EH4 = Homeodomain

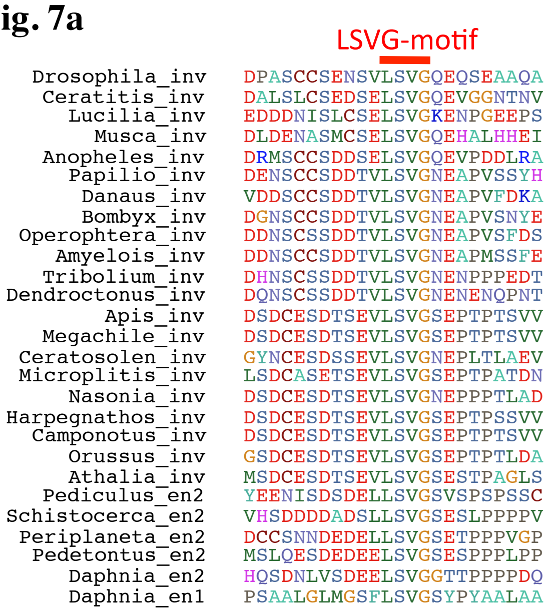
```
    Drosophila_en  ----------------------------------KDKTNDEKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKI
   Drosophila_inv  -------------ATSSSAAGGGGGGVEKGEAADGGGVPEDKRPRTAFSGTQLARLKHEFNENRYLTEKRRQQLSGELGLNEAQIKI
     Ceratitis_en  ----------------------------------KDKTTDEKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSAELGLNEAQIKI
    Ceratitis_inv  --------------TGEKSTGGSGVSSGSGSGGSSAATAEDKRPRTAFSGSQLARLKHEFTENRYLTEKRRQQLSSELGLNEAQIKI
       Lucilia_en  ----------------------------------KDKTTDEKRPRTAFSSEQLVRLKREFNENRYLTERRRQQLSSELGLNEAQIKI
      Lucilia_inv  ----KSTNSSSATAAASASSAAGVEKANSPSSTSSANNNEDKRPRTAFSGSQLARLKHEFNENRYLTEKRRQQLSSELGLNEAQIKI
         Musca_en  ---------------------------------KDPKATDEKRPRTAFSGEQLVRLKREFNENRYLTERRRQQLSAELGLNEAQIKI
        Musca_inv  TGPNDKANSPTGTSSSTSAGSGGSASAAAASAASSSSSSEDKRPRTAFSGSQLARLKHEFNENRYLTEKRRQQLSSELGLNEAQIKI
         Aedes_en  ---------------------------------KQKGETEEKRPRTAFTTAQLQRLKNEFNENRYLTEKRRQALSAELNLNESQIKI
     Anopheles_en  ---------------------------------KKRADSEEKRPRTAFSNAQLQRLKNEFNENRYLTEKRRQTLSAELGLNEAQIKI
    Anopheles_inv  -------------------PAEKSSSGPAQQSASAAALADDKRPRTAFSGPQLARLKHEFAENRYLTERRRQQLSAELGLNEAQIKI
       Papilio_en  ------------------------------------MNSEEKRPRTAFSASQLARLKHEFAENRYLTERRRQALAAELGLAEAQIKI
      Papilio_inv  --------------------------------PGEGPTSDEKRPRTAFSGPQLARLKHEFAENRYLTERRRQALAAELGLAEAQIKI
        Danaus_en  ------------------------------------ASPEEKRPRTAFSASQLTRLKHEFAENRYLTERRRQALAAELGLAEAQIKI
       Danaus_inv  ------------------------------------ASPEEKRPRTAFSASQLTRLKHEFAENRYLTERRRQALAAELGLAEAQIKI
        Bombyx_en  ------------------------------------AAPEEKRPRTAFSGAQLARLKHEFAENRYLTERRRQSLAAELGLAEAQIKI
       Bombyx_inv  --------------------------------PGDTASNDEKRPRTAFSGPQLARLKHEFAENRYLTERRRQSLAAELGLAEAQIKI
   Operophtera_en  -----------------------------------GPSVEEKRPRTAFSAAQLGRLKHEFAENRYLTERRRQALAAELGLAEAQIKI
  Operophtera_inv  --------------------------------PGETNPNDEKRPRTAFSGPQLARLKHEFAENRYLTERRRQALAAELGLAEAQIKI
      Amyelois_en  ------------------------------------INTEEKRPRTAFSAAQLARLKHEFTENRYLTERRRQALAAELGLAEAQIKI
     Amyelois_inv  -------------------------------PGETGATNDEKRPRTAFSGPQLARLKHEFAENRYLTERRRQALAAELGLAEAQIKI
      Plutella_en  ----------------------------------SPSAAEEKRPRTAFSAAQLNRLKHEFAENRYLTERRRQALAAELGLAEAQIKI
     Tribolium_en  ----------------------------------SKPNGEDKRPRTAFSSAQLARLKHEFNENRYLTERRRQQLSAELGLNEAQIKI
    Tribolium_inv  ------------------------------GAKQGAPTAEEKRPRTAFSGAQLARLKHEFAENRYLTERRRQQLSAELGLNEAQIKI
  Dendroctonus_en  ---------------------------------QSVSKTEDKRPRTAFSSAQLARLKTEFNENRYLTESRRQKLSTELGLNEAQIKI
 Dendroctonus_inv  --------------------------------GTKAAVPEEKRPRTAFSGAQLARLKNEFAENRYLTERRRQQLSAELGLNEAQIKI
          Apis_en  -------------------------------HNGKNGSPEEKRPRTAFSAEQLARLKREFAENRYLTERRRQQLSRDLGLNEAQIKI
         Apis_inv  ------------------------------DGRGNGGTPEEKRPRTAFSGEQLARLKREFAENRYLTERRRQQLSRDLGLNEAQIKI
     Megachile_en  ------------------------------ANGSKNGTPEEKRPRTAFSAEQLARLKREFAENRYLTERRRQQLSRDLGLNEAQIKI
    Megachile_inv  ------------------------------DNRGSTGTPEEKRPRTAFSGEQLARLKREFAENRYLTERRRQQLSRDLGLNEAQIKI
   Ceratosolen_en  -----------------------------KSEKPESATSEEKRPRTAFSAEQLNKLRLEFKENRYLTEFRRQKLSKELGLNEQQIKI
  Ceratosolen_inv  ------------------------------DELIRTAVPEEKRPRTAFSIEQLARLKREFAENKYLTEQRRQALSKELGLNEAQIKI
   Microplitis_en  --------------------------------VAEKQSADDKRPRTAFSAEQLARLKTEFTENRYLTERRRQQLSRELGLNEAQIKI
  Microplitis_inv  -----------------------------NEVRPTGNTPEEKRPRTAFSGEQLARLKREFAENRYLTERRRQQLSRDLGLNEAQIKI
       Nasonia_en  ----------------------------NSEKSQSVSTPEEKRPRTAFSAEQLAKLQLEFTENRYLNEQRRQKLSKELGLNEQQIKI
      Nasonia_inv  ------------------------------ADRPPAATPEEKRPRTAFSAEQLARLRDEFTENKYLTEQRRQTLSKELGLNEAQIKI
    Camponotus_en  -------------------------------QNSKNSLPEEKRPRTAFSAEQLSRLKREFNENRYLTEKRRQELSRELNLNEAQIKI
   Camponotus_inv  -------------------------------DGRGGGTPEEKRPRTAFSGEQLARLKREFAENRYLTERRRQQLSRDLGLNEAQIKI
  Harpegnathos_en  -------------------------------QNVKNSTPEEKRPRTAFSAEQLTRLKREFTENRYLTERRRQQLSQDLGLNEAQIKI
 Harpegnathos_inv  ------------------------------ADGRGGGTPEEKRPRTAFSGEQLARLKREFAENRYLTERRRQQLSRDLGLNEAQIKI
       Orussus_en  -----------------------------QPAGEKSSSSEEKRPRTAFNAEQLARLKREFAENRYLTERRRQQLSQDLGLNEAQIKI
      Orussus_inv  -----------------------------GDGRGSGATPEEKRPRTAFSGEQLARLKREFAENRYLTERRRQQLSKDLGLNEAQIKI
       Athalia_en  ------------------------------SGSDKNASSEEKRPRTAFSAEQLARLKQEFAENRYLTERRRQQLSRDLGLNEAQIKI
      Athalia_inv  ----------------------------NGDSRTTGSTPEEKRPRTAFSGEQLARLKREFAENRYLTERRRQQLSRDLGLNEAQIKI
    Pediculus_en1  -----------------------------------DKKPEEKRPRTAFSGDQLSRLKHEFAENRYLTERRRQDLAKELGLNEAQIKI
    Pediculus_en2  -----------------------------------DKNSDEKRPRTAFSGDQLSRLKHEFAENRYLTERRRQDLARELGLNEAQIKI
 Schistocerca_en1  -----------------------------------DKKPEEKRPRTAFSGEQLARLKHEFTENRYLTERRRQELARELGLNEAQIKI
 Schistocerca_en2  ------------------------------------KKPEEKRPRTAFSGEQLARLKHEFTENRYLTERRRQELARELGLNEAQIKI
  Periplaneta_en1  -----------------------------------EKKPEEKRPRTAFSGEQLARLKSEFTENRYLTERRRTELARELGLNEAQIKI
  Periplaneta_en2  -----------------------------------DKKPEEKRPRTAFSGEQLARLKHEFTENRYLTERRRTELARELGLNEAQIKI
 Zootermopsis_en1  -----------------------------------EKKPEEKRPRTAFSGEQLARLKHEFTENRYLTERRRTELARELGLNEAQIKI
   Pedetontus_en1  -----------------------------------ERRPEEKRPRTAFTQEQLARLRREFEENRYLTERRRQDLARELHLHENQIKI
   Pedetontus_en2  -----------------------------------ERRPEDKRPRTAFTQEQLARLRREFEENRYLTERRRQDLARDLNLHENQIKI
      Daphnia_en1  ------------------------------------RNAEEKRPRTAFTSEQLARLKNEFTENRYLNEKRRQELANELQLHENQIKI
      Daphnia_en2  ----------------------------------KDRKAEEKRPRTAFTSEQLARLKSEFTENRYLTEKRRQDLARELQLHENQIKI
       Artemia_en  ----------------------------------KAITPDEKRPRTAFTAEQLSRLKHEFNENRYLTERRRQDLARELGLHENQIKI
       Caligus_en  ----------------EKTILPSVASSACSSSPTSPESSEEKRPRTAFSSEQLARLKREFDENRYLNEERRRALSSELGLNETQIKI
       Argulus_en  -----------------------------------KSSSSEKRPRTAFTADQLSRLKREFQDNKYLTEKRRQDLARELQLNETQIKI
Sacculina_en-a.E9  --------------------------------------TDEKRPRTAFSSEQLQRLASEFTDNRYLSEERRQRLARQLGLNESQIKI
Sacculina_en-a.E20 --------------------------------------TDEKRPRTAFSSEQLQRLASEFTDNRYLSEERRQRLARQLGLNESQIKI
   Sacculina_en-b  ------------------------------------KTSDEKRPRTAFSSEQLARLKMEFQQNRYLTEKRRQDLAGELQLNESQIKI
   Cupiennius_en1  -----------------------------------DKKPDDKRPRTAFTADQLSRLKHEFQENRYLTERRRQDLAKDLQLNESQIKI
   Stegodyphus_en  -----------------------------------EKKPDEKRPRTAFSAEQLSRLKQEFQENRYLTEKRRMDLARDLKLNESQIKI
  Parasteatoda_en  -----------------------------------SDGKADKRPRTAFSSEQLNRLRQEFSENRYLTERRRQDMARDLKLNESQIKI
        Ixodes_en  -----------------------------------EKKADEKRPRTAFTADQLARLKQEFTENRYLTEKRRQDLARELKLNESQIKI
   Metaseiulus_en  ------------------------------------EKSDEKRPRTAFTADQLQRLKKEFQENKYLTEKRRQDLASELGLNESQIKI
  Archegozetes_en  -----------------------------------DKKADEKRPRTAFTAEQLARLKQEFQENRYLTEKRRQDLAKDLKLNESQIKI
   Limulus_en-1Ba  ------------------------------------KKPDEKRPRTAFTADQLARLKAEFQENRYLTEKRRQDLARELQLNESQIKI
   Limulus_en-1Bb  ------------------------------------LKPDEKRPRTAFTADQLARLKAEFQENRYLTEKRRQDLARELQLNESQIKI
   Limulus_en-1Bc  -----------------------------------KKAPDEKRPRTAFTADQLARLKQEFQENRYLTEKRRQELAQNLQLKESQIKI
   Limulus_en-1Bd  -----------------------------------RKTPEEKRPRTAFTTDQLARLKKEFHENRYLTEKRRQELARELQVNESQIKI
```

Fig. 6 continued

```
                     — EH4 →   ← EH5 →
      Drosophila_en  WFQNKRAKIKKSTGSKNPLALQLMAQGLYNHTTV---PLTKEEEELEMRMNGQIP*
     Drosophila_inv  WFQNKRAKLKKSSGTKNPLALQLMAQGLYNHSTI---PLTREEEELQELQEAASARAAKEPC*
       Ceratitis_en  WFQNKRAKIKKSSGSKNPLALQLMAQGLYNHTTV---PLTKEEEELEMRMNGQIP*
      Ceratitis_inv  WFQNKRAKLKKSSGTKNPLALQLMAQGLYNHSTV---PLTREEEELQELQERENAAAAAAAAAAPEARSGGAGTTTVTTVST*
         Lucilia_en  WFQNKRAKIKKSSGSKNPLALQLMAQGLYNHTTV---PLTKEEEELEMRMNGQIP*
        Lucilia_inv  WFQNKRAKLKKSSGVKNPLALQLMAQGLYNHSTI---PLTREEEELQELQEREKNNSNNNTNSLQQQQQAASAVTS*
           Musca_en  WFQNKRAKIKKSTGTKNPLALQLMAQGLYNHTTV---PLTKEEEELEMRMNGQIP*
          Musca_inv  WFQNKRAKLKKSSGVKNPLALQLMAQGLYNHSTI---PLTREEEELQELQEREKSANNNNLTQPTASAVSS*
           Aedes_en  WFQNKRAKIKKTSSEKNPLALQLMAQGLYNHSTV---PLTKEEEELEMRMNGQIP*
       Anopheles_en  WFQNKRAKIKKSSSEKNPLALQLMAQGLYNHSTV---PLTKEEEELEMRMNGQIP*
      Anopheles_inv  WFQNKRAKIKKSSGQKNPLALQLMAQGLYNHSTV---PLTREEEELQEMQAAAAAAAESRTGSAGPAASATVANA*
         Papilio_en  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHSTA---TESDEDDEEISVT*
        Papilio_inv  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHSTV---PLTKEEEELEMKAREREKELQNRH*
          Danaus_en  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHATV---TESEDEEISVT*
         Danaus_inv  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHSTI---PLTKEEEELEMKAREREQRN*
          Bombyx_en  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHSTV---TESDDEEEINVT*
         Bombyx_inv  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHSTV---PLTKEEEELEMKARERERELKNRC*
     Operophtera_en  WFQNKRAKIKKATGQRNPLAMQLMAQGLYNHSTA---NESDEEEEINVT*
    Operophtera_inv  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHSTV---PLTKEEEELEMKAREREAQNRV*
        Amyelois_en  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHSTA---TESDEDEEINII*
       Amyelois_inv  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHSTV---PLTREEEELEMKAREREQQNRC*
        Plutella_en  WFQNKRAKIKKANGQRNPLALQLMAQGLYNHTTA---TESDEEEEISVT*
       Tribolium_en  WFQNKRAKIKKSSGQKNPLALQLMAQGLYNHSTV---ACDEDDLPLSS*
      Tribolium_inv  WFQNKRAKIKKASGTKNPLALQLMAQGLYNHSTI---PLTKEEEELQEMQGTKSPA*
     Dendroctonus_en  WFQNKRAKIKKSTGQKNPLALQLMAQGLYNHSTV---ACDEDDMPLSA*
    Dendroctonus_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTI---PLTKEEEELEKLQSQGKIS*
            Apis_en  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PVDEDGEEIVTGNNHSH*
           Apis_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLTKEEEEQAAELQAK*
       Megachile_en  WFQNKRAKIKKASGQKNPLALQLMAEGLYNHSTV---PVDEDGEEVGTGNNHPH*
      Megachile_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLTKEEEEQAAELQAK*
     Ceratosolen_en  WFQNKRAKMKKASGQKNPLALQLMAQGLYNHSTM---PVDESNEEITICEI*
    Ceratosolen_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLSKDELEQAAELQANGI*
     Microplitis_en  WFQNKRAKIKKSTGQKNPLALQLMAQGLYNHSTV---PVDEDGEEIIPH*
    Microplitis_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLSKEEEEQAAELQAK*
         Nasonia_en  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PVDEDGEEIPE*
        Nasonia_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLSKEEMEQAAELQAK*
      Camponotus_en  WFQNKRAKLKKASGQKNPLALQLMAQGLYNHSTV---PVDEDDEEMVTERNH*
     Camponotus_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLTKEEEEQAAELQAK*
     Harpegnathos_en  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PVDEDGEEIVTGRNP*
    Harpegnathos_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLTKEEEEQAAELQAK*
         Orussus_en  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PVDEDGEEIVPGKA*
        Orussus_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLTKEEEEQAAELQAK*
         Athalia_en  WFQNKRAKIKKSTGQKNPLALQLMAQGLYNHSTV---PVDEDGEEIPPEENRS*
        Athalia_inv  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PLTREEEEQAAELQAK*
      Pediculus_en1  WFQNKRAKMKKAKGEKNPLALQLMAQGLYNHSTI---PVDEDEYLEEMAAASNQSNPV*
      Pediculus_en2  WFQNKRAKMKKARGEKNPLALQLMAQGLYNHSTI---PLTKEEEEAAAAELSD*
   Schistocerca_en1  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PIDEDDEETTAPPPQQQQPPTAPQTVSGVLAAPTTP*
   Schistocerca_en2  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PMTREEEEQAIASDK*
    Periplaneta_en1  WFQNKRAKIKKASGRKNPLALQLMAQGLYNHSTV---PIDEEEEEANALLLANNARQD*
    Periplaneta_en2  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTI---PMTREEEEQAAAAEANAKKT*
   Zootermopsis_en1  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTV---PVDEEEEEANALLLANNARQD*
     Pedetontus_en1  WFQNKRAKIKKSTGQKGGLALQLMAQGLYNHSTV---SVDEDESNPMPLSPTSVGHQSDI*
     Pedetontus_en2  WFQNKRAKIKKATGQKGGLALQLMAQGLYNHSTI---PLRDGEEDDSTCSPPPT*
        Daphnia_en1  WFQNKRAKIKKTTGQKNPLALQLMAQGLYNHSTV---PVGEEEEDEEFYASHQQHQQQQREQHSE*
        Daphnia_en2  WFQNKRAKIKKASGQKNPLALQLMAQGLYNHSTM---AMDEDEDDDDEMMQQE
         Artemia_en  WFQNNRAKLKKSSGQKNPLALQLMAQGLYNHSTI---PTEDDEDDEISSTSLQARIE*
         Caligus_en  WFQNKRAKLKKTTGGKGDLAKMLEAQGLYNHATV---SLEEEEELLSAI*
         Argulus_en  WFQNKRAKIKKSTGSRNPLAMQLMAQGLYNHSTV---PLNEDGTDEQ*
  Sacculina_en-a.E9  WFQNKRAKLKKTIPDKPSLAKKLMEQGLYNHTTI---LPEDEEKLMQLYKQSAPHISA*
 Sacculina_en-a.E20  WFQNKRAKLKKTIPDKPSLAKKLMEQGLYNHTTI---LPEDEEKLMQLYKQSAPHISV*
     Sacculina_en-b  WFQNKRAKLKKTTGNRNPLALSLMTEGLYNHSTM---TVDEDE*
     Cupiennius_en1  WFQNRRAKLKKASGQRSALALQLMAQGLYNHSTI---PIRGDEDDDERPKSSSSS*
     Stegodyphus_en  WFQNRRAKLKKNNPTRNPLALQLMAQGLYNHSTI---PLRDDDDDDERPKSSSSS
     Parasteatoda_en  WFQNRRAKLKKINPRRSPLALQLMAEGLYDHRTL---PVKDDDDDERPKSSSSS*
          Ixodes_en  WFQNKRAKIKKASGQRNPLALQLMAQGLYNHTTASQQGMGDDDDNSSS*
      Metaseiulus_en  WFQNKRAKLKKSTGRPNPLALELMAQGLYNHSTV---GPDGELVHDDMDDGR*
     Archegozetes_en  WFQNKRAKIKKSTGSRNPLAMHLMAQGLCNHSTI---AVDDEDDDDDDDYLEERESLCTSPPKHSDLKNNKS*
     Limulus_en-1Ba  WFQNKRAKIKKAVGHPNTLAIQLMAQGLYNHTTV---PIRDDMDDKDIDSS*
     Limulus_en-1Bb  WFQNKRAKIKKATGHTNSLAIQLMAQGLYNHSTI---PVRDDLDDKDIDSS*
     Limulus_en-1Bc  WFQNKRAKIKKASCEKNPLALQLMAQGLYNHSTV---PVHNEDDEDGDSSYSA*
     Limulus_en-1Bd  WFQNKRAKIKKSTGEKNPLALQLMAQGLYNHTTL---PITHGIDDDGDSNIST*
```

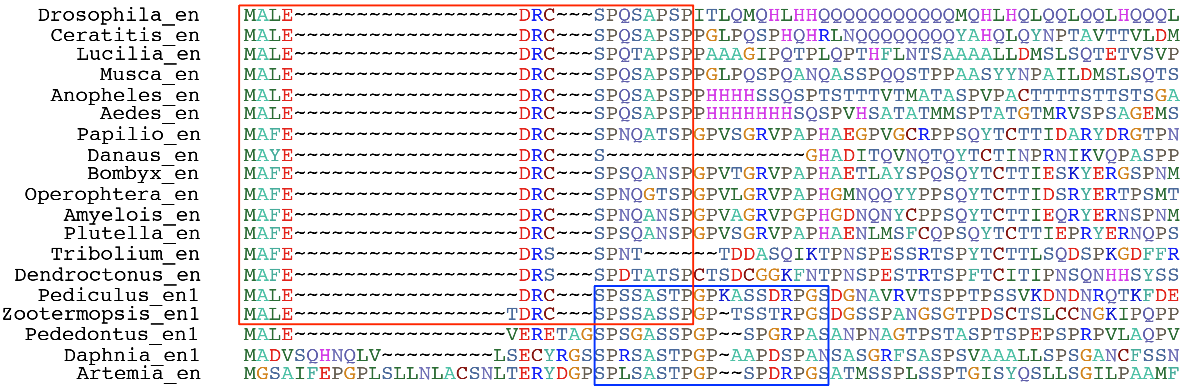
Fig. 7 a: Amino acid sequences around the LSVG-motif in invected in holometabolans, engrailed2 in the other insects, *Daphnia*_en2 and *Daphnia*_en1. **b**: Amino acid sequences of engrailed-specific domains in engrailed in dipterans, lepidopterans and coleopterans, and *Pediculus*_en1, *Zootermopsis*_en1, *Pedetontus*_en1, *Daphnia*_en1 and *Artemia*_en. The methyonine residue on the N-terminal side of each sequence is a start methyonine. The red box indicates the conserved region among a portion of the holometabolous species, *Pediculus* and *Zootermopsis*, and blue box indicates the conserved region among *Pediculus*, *Zootermopsis*, *Pedetontus* and branchiopods.

Fig. 7a



```
  Drosophila_inv    DPASCCSENSVLSVGQEQSEAAQA
   Ceratitis_inv    DALSLCSEDSELSVGQEVGGNTNV
     Lucilia_inv    EDDDNISLCSELSVGKENPGEEPS
       Musca_inv    DLDENASMCSELSVGQEHALHHEI
   Anopheles_inv    DRMSCCSDDSELSVGQEVPDDLRA
     Papilio_inv    DENSCCSDDTVLSVGNEAPVSSYH
      Danaus_inv    VDDSCCSDDTVLSVGNEAPVFDKA
      Bombyx_inv    DGNSCCSDDTVLSVGNEAPVSNYE
 Operophtera_inv    DDNSCSSDDTVLSVGNEAPVSFDS
    Amyelois_inv    DDNSCCSDDTVLSVGNEAPMSSFE
   Tribolium_inv    DHNSCSSDDTVLSVGNENPPPEDT
Dendroctonus_inv    DQNSCSSDDTVLSVGNENENQPNT
        Apis_inv    DSDCESDTSEVLSVGSEPTPTSVV
   Megachile_inv    DSDCESDTSEVLSVGSEPTPTSVV
 Ceratosolen_inv    GYNCESDSSEVLSVGNEPLTLAEV
 Microplitis_inv    LSDCASETSEVLSVGSEPTPATDN
     Nasonia_inv    DSDCESDTSEVLSVGNEPPPTLAD
Harpegnathos_inv    DSDCESDTSEVLSVGSEPTPSSVV
  Camponotus_inv    DSDCESDTSEVLSVGSEPTPTSVV
     Orussus_inv    GSDCESDTSEVLSVGSEPTPTLDA
     Athalia_inv    MSDCESDTSEVLSVGSESTPAGLS
   Pediculus_en2    YEENISDSDELLSVGSVSPSPSSC
Schistocerca_en2    VHSDDDDADSLLSVGSESLPPPPV
 Periplaneta_en2    DCCSNNDEDELLSVGSETPPPVGP
  Pedetontus_en2    MSLQESDEDEELSVGSESPPPLPP
     Daphnia_en2    HQSDNLVSDEELSVGGTTPPPPDQ
     Daphnia_en1    PSAALGLMGSFLSVGSYPYAALAA
```

Fig. 7b



```
    Drosophila_en    MALE~~~~~~~~~~~~~~~~~~DRC~~~SPQSAPSPITLQMQHLHHQQQQQQQQQQQMQHLHQLQQLQQLHQQQL
     Ceratitis_en    MALE~~~~~~~~~~~~~~~~~~DRC~~~SPQSAPSPPGLPQSPHQHRLNQQQQQQQQYAHQLQYNPTAVTTVLDM
       Lucilia_en    MALE~~~~~~~~~~~~~~~~~~DRC~~~SPQTAPSPPAAAGIPQTPLQPTHFLNTSAAAALLDMSLSQTETVSVP
         Musca_en    MALE~~~~~~~~~~~~~~~~~~DRC~~~SPQSAPSPPGLPQSPQANQASSPQQSTPPAASYYNPAILDMSLSQTS
     Anopheles_en    MALE~~~~~~~~~~~~~~~~~~DRC~~~SPQSAPSPPHHHHSSQSPTSTTTVTMATASPVPACTTTTSTTSTSGA
         Aedes_en    MALE~~~~~~~~~~~~~~~~~~DRC~~~SPQSAPSPPHHHHHHHSQSPVHSATATMMSPTATGTMRVSPSAGEMS
       Papilio_en    MAFE~~~~~~~~~~~~~~~~~~DRC~~~SPNQATSPGPVSGRVPAPHAEGPVGCRPPSQYTCTTIDARYDRGTPN
        Danaus_en    MAYE~~~~~~~~~~~~~~~~~~DRC~~~S~~~~~~~~~~~~~~~~GHADITQVNQTQYTCTINPRNIKVQPASPP
        Bombyx_en    MAFE~~~~~~~~~~~~~~~~~~DRC~~~SPSQANSPGPVTGRVPAPHAETLAYSPQSQYTCTTIESKYERGSPNM
   Operophtera_en    MAFE~~~~~~~~~~~~~~~~~~DRC~~~SPNQGTSPGPVLGRVPAPHGMNQQYYPPSQYTCTTIDSRYERTPSMT
      Amyelois_en    MAFE~~~~~~~~~~~~~~~~~~DRC~~~SPNQANSPGPVAGRVPGPHGDNQNYCPPSQYTCTTIEQRYERNSPNM
      Plutella_en    MAFE~~~~~~~~~~~~~~~~~~DRC~~~SPSQANSPGPVSGRVPAPHAENLMSFCQPSQYTCTTIEPRYERNQPS
     Tribolium_en    MAFE~~~~~~~~~~~~~~~~~~DRS~~~SPNT~~~~~~TDDASQIKTPNSPESSRTSPYTCTTLSQDSPKGDFFR
  Dendroctonus_en    MAFE~~~~~~~~~~~~~~~~~~DRS~~~SPDTATSPCTSDCGGKFNTPNSPESTRTSPFTCITIPNSQNHHSYSS
    Pediculus_en1    MALE~~~~~~~~~~~~~~~~~~DRC~~~SPSSASTPGPKASSDRPGSDGNAVRVTSPPTPSSVKDNDNRQTKFDE
Zootermopsis_en1    MALE~~~~~~~~~~~~~~~~~TDRC~~~SPSSASSPGP~TSSTRPGSDGSSPANGSGTPDSCTSLCCNGKIPQPP
  Pededontus_en1    MALE~~~~~~~~~~~~~~~~~VERETAGSPSGASSPGP~~SPGRPASANPNAGTPSTASPTSPEPSPRPVLAQPV
     Daphnia_en1    MADVSQHNQLV~~~~~~~~~LSECYRGSSPRSASTPGP~AAPDSPANSASGRFSASPSVAAALLSPSGANCFSSN
      Artemia_en    MGSAIFEPGPLSLLNLACSNLTERYDGPSPLSASTPGP~~SPDRPGSATMSSPLSSPTGISYQSLLSGILPAAMF
```

Fig. 8 Distribution of fragments detected in GENECONV analysis and their corresponding positions in the exons of the engrailed-family genes. Only fragments detected in paired paralogs in single species with the condition of gscale = 2 to 5 and 0 are shown. In the case that fragments of different lengths were detected from a single species with a different gscale, the longest and shortest are shown. The exon/intron structure shown at the top of the figure is an illustrative example. Some of the genes possess extra introns not indicated in this figure.

Fig. 8

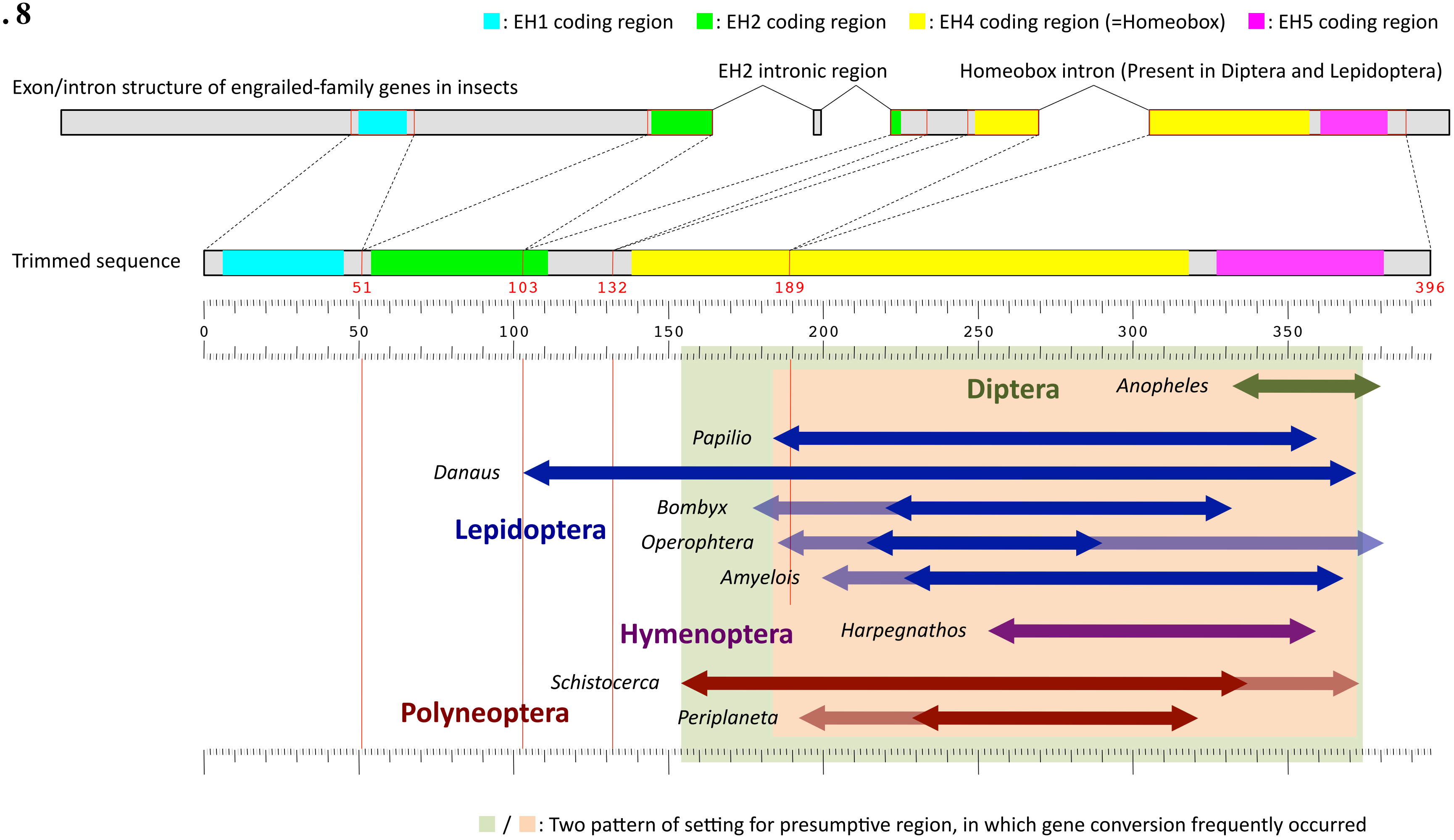

: EH1 coding region
: EH2 coding region
: EH4 coding region (=Homeobox)
: EH5 coding region
EH2 intronic region
Homeobox intron (Present in Diptera and Lepidoptera)
Exon/intron structure of engrailed-family genes in insects
Trimmed sequence
51
103
132
189
396
0
50
100
150
200
250
300
350
Diptera
Anopheles
Papilio
Danaus
Bombyx
Lepidoptera
Operophtera
Amyelois
Hymenoptera
Harpegnathos
Schistocerca
Polyneoptera
Periplaneta
/ : Two pattern of setting for presumptive region, in which gene conversion frequently occurred

Fig. 9 Molecular phylogenetic trees of arthropod engrailed-family genes determined by the maximum likelihood method based on the Tamura-Nei model. An alignment of 71 DNA sequences was used for the analysis (see Fig. 6) but with two different regions: 189 nucleotides from position 184 to 372 (**a**) and the other 207 positions (**b**). The trees with the highest log likelihood (-6484.5271 for former and -9331.8732 for latter) are shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying neighbor-joining and BioNJ algorithms to a matrix of pairwise distances estimated using the maximum composite likelihood (MCL) approach, and the topology with the superior log likelihood value was selected. A discrete gamma distribution was used to model the evolutionary rate differences among sites (5 categories; +G, parameter = 0.8346 for the positions from 184 to 372 and 0.8552 for the other). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 29.3187% sites for positions from 184 to 372 and 0.8552, 21.1125% for the other). The trees are drawn to scale with branch lengths representative of the number of substitutions per site. Codon positions included were 1st + 2nd + 3rd. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA6.06. *Engrailed* genes in holometabolans and *engrailed1* genes in non-holometabolan insects and *Daphnia* are in blue font. *Invected* genes in holometabolans and *engrailed2* genes in non-holometabolan insects and *Daphnia* are in red font. Each clade consisted of *engrailed*/*engrailed1* and

invected/*engrailed2* in a species is indicated by a rounded square. Orange rounded squares indicate the gene pairs detected in the GENECONV analysis and gray ones indicate the gene pairs not detected.

Fig. 9a

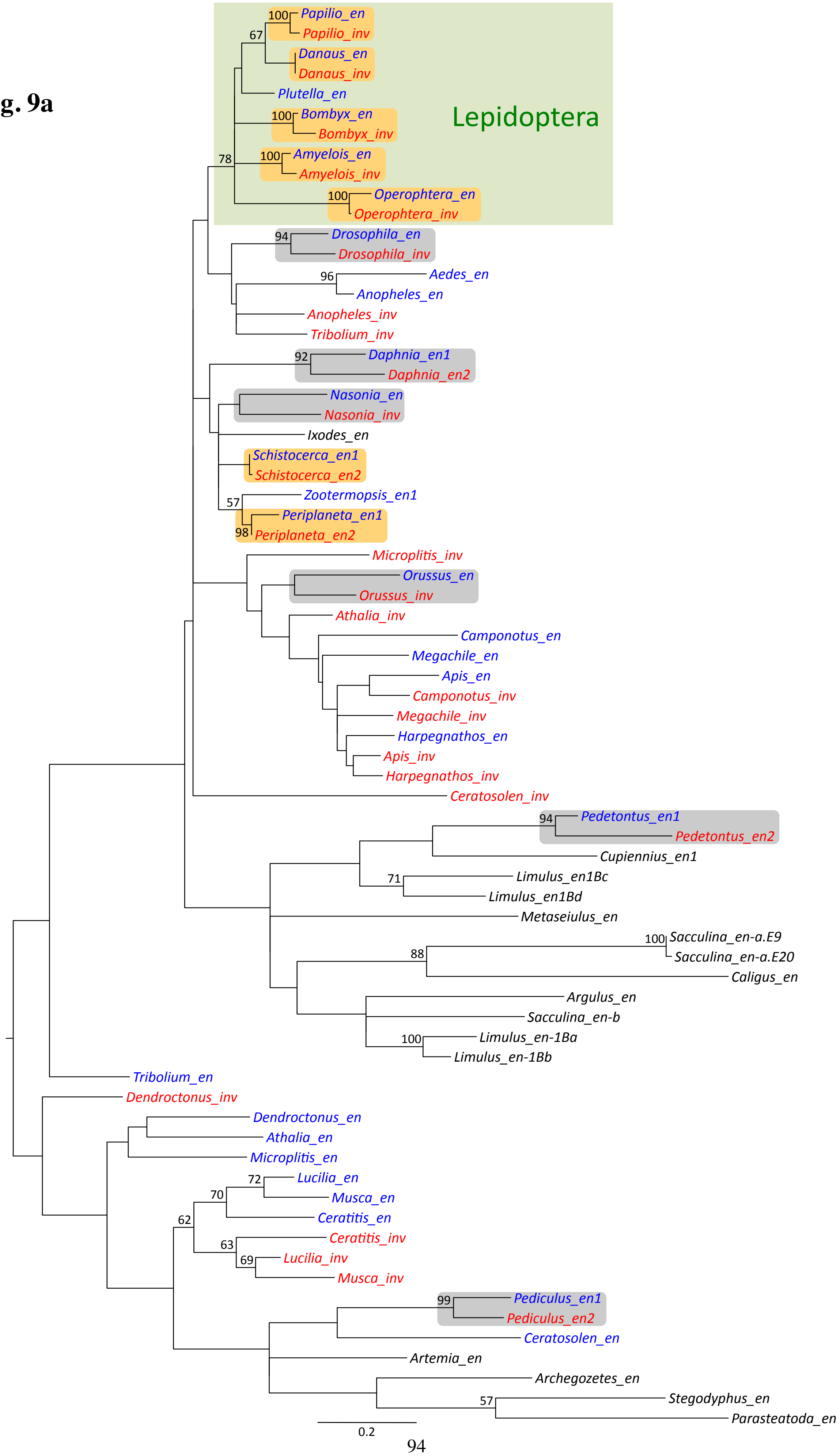

Lepidoptera
100
Papilio_en
Papilio_inv
67
Danaus_en
Danaus_inv
Plutella_en
100
Bombyx_en
Bombyx_inv
78
100
Amyelois_en
Amyelois_inv
100
Operophtera_en
Operophtera_inv
94
Drosophila_en
Drosophila_inv
96
Aedes_en
Anopheles_en
Anopheles_inv
Tribolium_inv
92
Daphnia_en1
Daphnia_en2
Nasonia_en
Nasonia_inv
Ixodes_en
Schistocerca_en1
Schistocerca_en2
57
Zootermopsis_en1
Periplaneta_en1
98
Periplaneta_en2
Microplitis_inv
Orussus_en
Orussus_inv
Athalia_inv
Camponotus_en
Megachile_en
Apis_en
Camponotus_inv
Megachile_inv
Harpegnathos_en
Apis_inv
Harpegnathos_inv
Ceratosolen_inv
94
Pedetontus_en1
Pedetontus_en2
Cupiennius_en1
71
Limulus_en1Bc
Limulus_en1Bd
Metaseiulus_en
100
Sacculina_en-a.E9
88
Sacculina_en-a.E20
Caligus_en
Argulus_en
Sacculina_en-b
100
Limulus_en-1Ba
Limulus_en-1Bb
Tribolium_en
Dendroctonus_inv
Dendroctonus_en
Athalia_en
Microplitis_en
72
Lucilia_en
70
Musca_en
62
Ceratitis_en
Ceratitis_inv
63
69
Lucilia_inv
Musca_inv
99
Pediculus_en1
Pediculus_en2
Ceratosolen_en
Artemia_en
Archegozetes_en
57
Stegodyphus_en
Parasteatoda_en
0.2

Fig. 9b

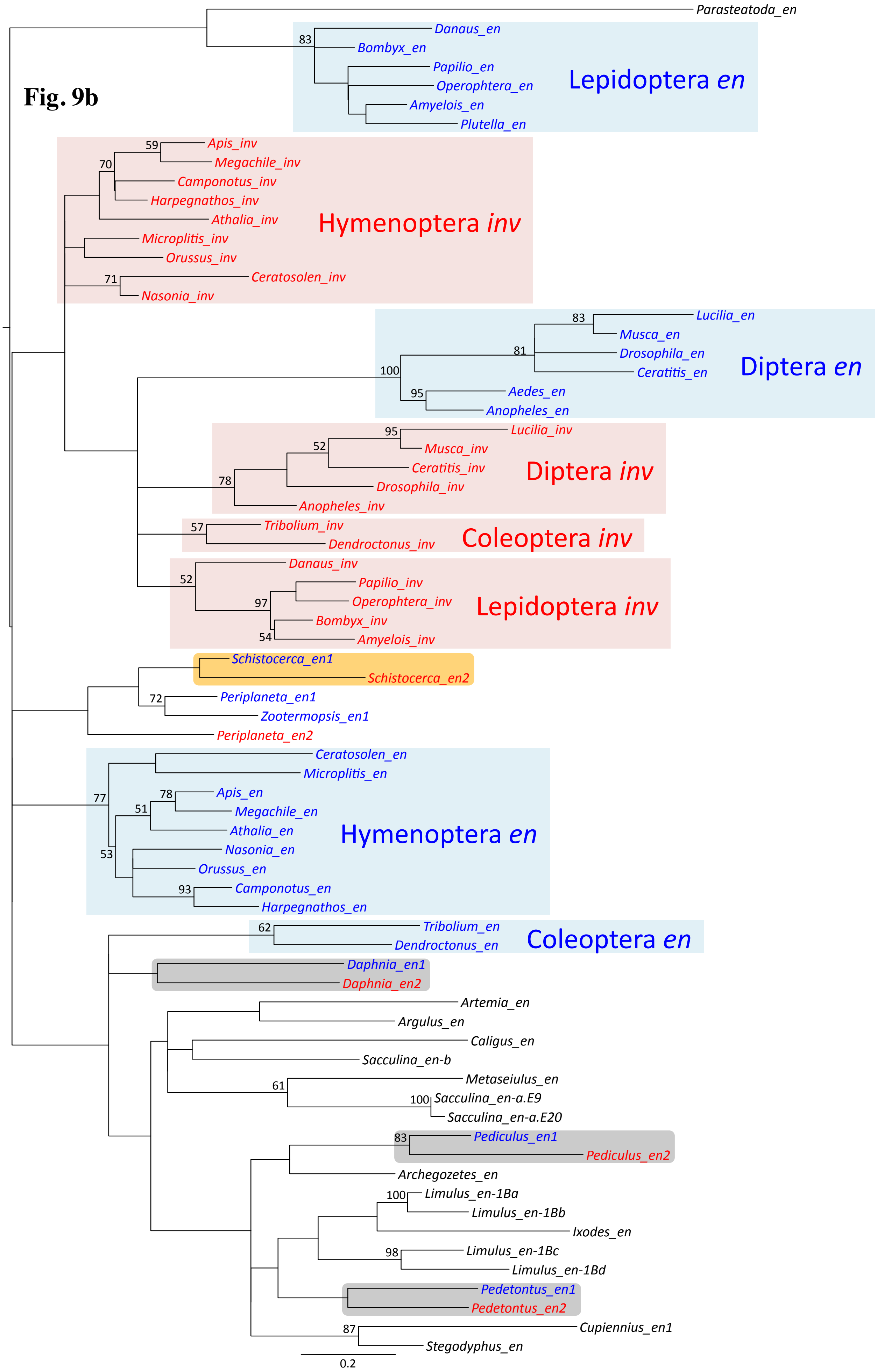

Parasteatoda_en
83
Danaus_en
Bombyx_en
Papilio_en
Operophtera_en
Amyelois_en
Plutella_en
Lepidoptera en
59
Apis_inv
70
Megachile_inv
Camponotus_inv
Harpegnathos_inv
Athalia_inv
Hymenoptera inv
Microplitis_inv
Orussus_inv
71
Ceratosolen_inv
Nasonia_inv
83
Lucilia_en
Musca_en
81
Drosophila_en
Ceratitis_en
Diptera en
100
95
Aedes_en
Anopheles_en
95
Lucilia_inv
52
Musca_inv
Ceratitis_inv
Diptera inv
78
Drosophila_inv
Anopheles_inv
57
Tribolium_inv
Dendroctonus_inv
Coleoptera inv
Danaus_inv
52
Papilio_inv
97
Operophtera_inv
Lepidoptera inv
Bombyx_inv
54
Amyelois_inv
Schistocerca_en1
Schistocerca_en2
72
Periplaneta_en1
Zootermopsis_en1
Periplaneta_en2
Ceratosolen_en
Microplitis_en
78
Apis_en
77
51
Megachile_en
Athalia_en
Hymenoptera en
53
Nasonia_en
Orussus_en
93
Camponotus_en
Harpegnathos_en
62
Tribolium_en
Dendroctonus_en
Coleoptera en
Daphnia_en1
Daphnia_en2
Artemia_en
Argulus_en
Caligus_en
Sacculina_en-b
61
Metaseiulus_en
100
Sacculina_en-a.E9
Sacculina_en-a.E20
83
Pediculus_en1
Pediculus_en2
Archegozetes_en
100
Limulus_en-1Ba
Limulus_en-1Bb
Ixodes_en
98
Limulus_en-1Bc
Limulus_en-1Bd
Pedetontus_en1
Pedetontus_en2
87
Cupiennius_en1
Stegodyphus_en
0.2

Fig. 10 Molecular phylogenetic trees of arthropod engrailed-family genes determined by the maximum likelihood method based on the Tamura-Nei model and the Hasegawa-Kishino-Yano model. Because the branching patterns were identical between these models, only topologies produced from the Tamura-Nei model are shown. An alignment of 71 DNA sequences was used for the analyses (see Fig. 6). The two trees are produced with different regions in the sequence alignment: 221 positions from position 154 to 374 (**a**) and the remaining 175 positions (**b**). Trees with the highest log likelihood (-7825.9957 for the former and -8033.8014 for the latter) are shown. The percentage of trees in which the associated taxa clustered together are shown next to the branches: percentage in the Tamura-Nei model is shown above and that for the Hasegawa-Kishino-Yano model is shown below. Initial tree(s) for the heuristic search were obtained automatically by applying the neighbor-join and BioNJ algorithms to a matrix of pairwise distances estimated using the maximum composite likelihood (MCL) approach, and then the topology with the superior log likelihood value was selected. A discrete gamma distribution was used to model the evolutionary rate differences among sites (5 categories; +G, parameter = 0.7831 for the sequences from position 154 to 374 and 0.9612 for the other). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 19.0894% sites for the sequences from position 154 to 374 and 0.8552, 29.9485% for the other). The trees are drawn to scale with branch lengths representative of the number of substitutions per site. Codon positions included were 1st +

2nd + 3rd. All positions containing gaps and missing data were eliminated. Evolutionary analyses were conducted in MEGA6.06. *Engrailed* genes in holometabolans and *engrailed1* genes in non-holometabolan insects and *Daphnia* are in blue font. *Invected* genes in holometabolans and *engrailed2* genes in non-holometabolan insects and *Daphnia* are in red font. Each clade consisted of *engrailed*/*engrailed1* and *invected*/*engrailed2* in a species is indicated by a rounded square. Orange rounded squares indicate the gene pairs detected in the GENECONV analysis and gray ones indicate the gene pairs not detected.

Fig. 10a

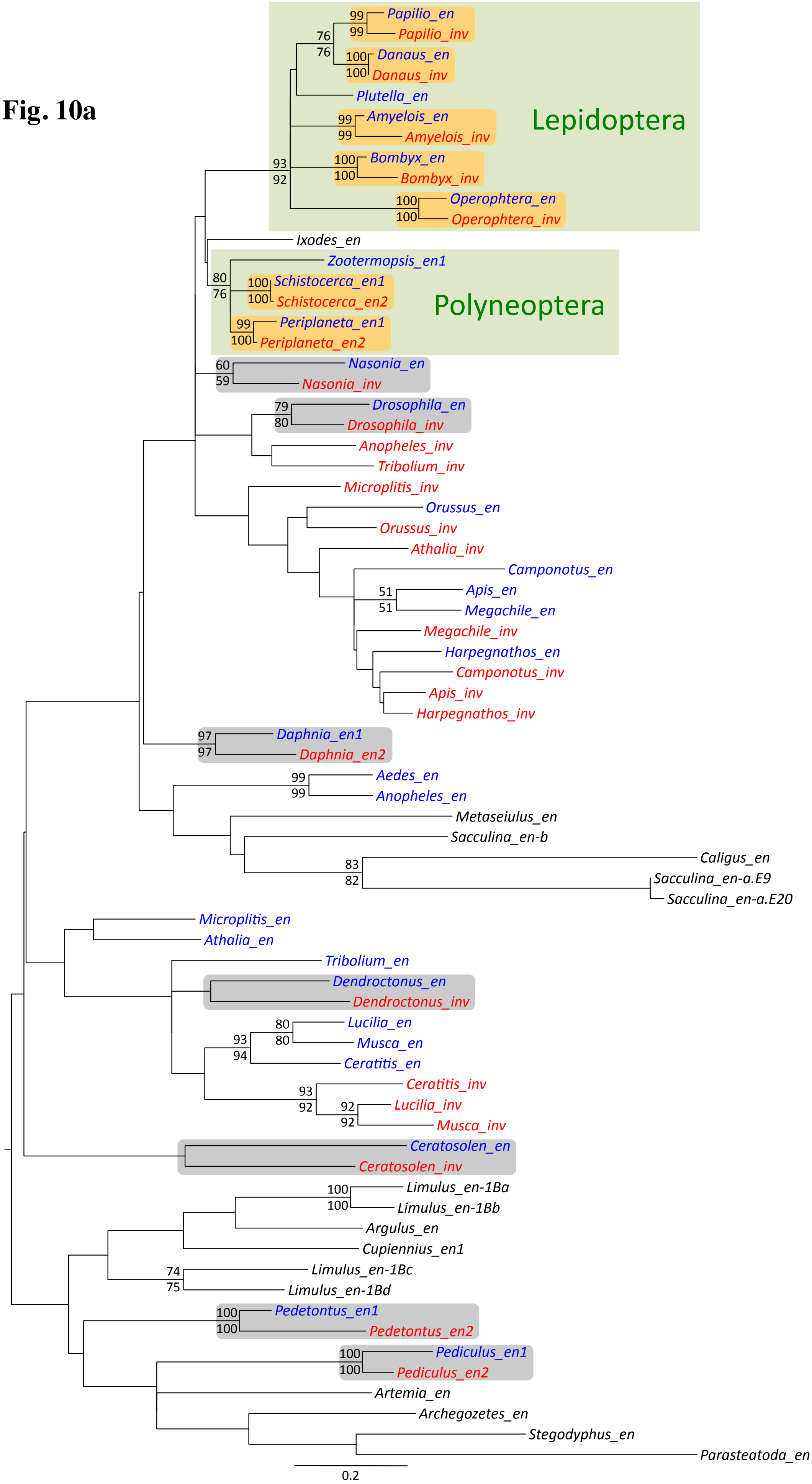

Lepidoptera
Polyneoptera
99
99
Papilio_en
Papilio_inv
76
76
100
100
Danaus_en
Danaus_inv
Plutella_en
99
99
Amyelois_en
Amyelois_inv
93
92
100
100
Bombyx_en
Bombyx_inv
100
100
Operophtera_en
Operophtera_inv
Ixodes_en
Zootermopsis_en1
80
76
100
100
Schistocerca_en1
Schistocerca_en2
99
100
Periplaneta_en1
Periplaneta_en2
60
59
Nasonia_en
Nasonia_inv
79
80
Drosophila_en
Drosophila_inv
Anopheles_inv
Tribolium_inv
Microplitis_inv
Orussus_en
Orussus_inv
Athalia_inv
Camponotus_en
51
51
Apis_en
Megachile_en
Megachile_inv
Harpegnathos_en
Camponotus_inv
Apis_inv
Harpegnathos_inv
97
97
Daphnia_en1
Daphnia_en2
99
99
Aedes_en
Anopheles_en
Metaseiulus_en
Sacculina_en-b
83
82
Caligus_en
Sacculina_en-a.E9
Sacculina_en-a.E20
Microplitis_en
Athalia_en
Tribolium_en
Dendroctonus_en
Dendroctonus_inv
80
80
Lucilia_en
93
94
Musca_en
Ceratitis_en
93
92
Ceratitis_inv
92
92
Lucilia_inv
Musca_inv
Ceratosolen_en
Ceratosolen_inv
100
100
Limulus_en-1Ba
Limulus_en-1Bb
Argulus_en
Cupiennius_en1
74
75
Limulus_en-1Bc
Limulus_en-1Bd
100
100
Pedetontus_en1
Pedetontus_en2
100
100
Pediculus_en1
Pediculus_en2
Artemia_en
Archegozetes_en
Stegodyphus_en
Parasteatoda_en
0.2

Fig. 10b

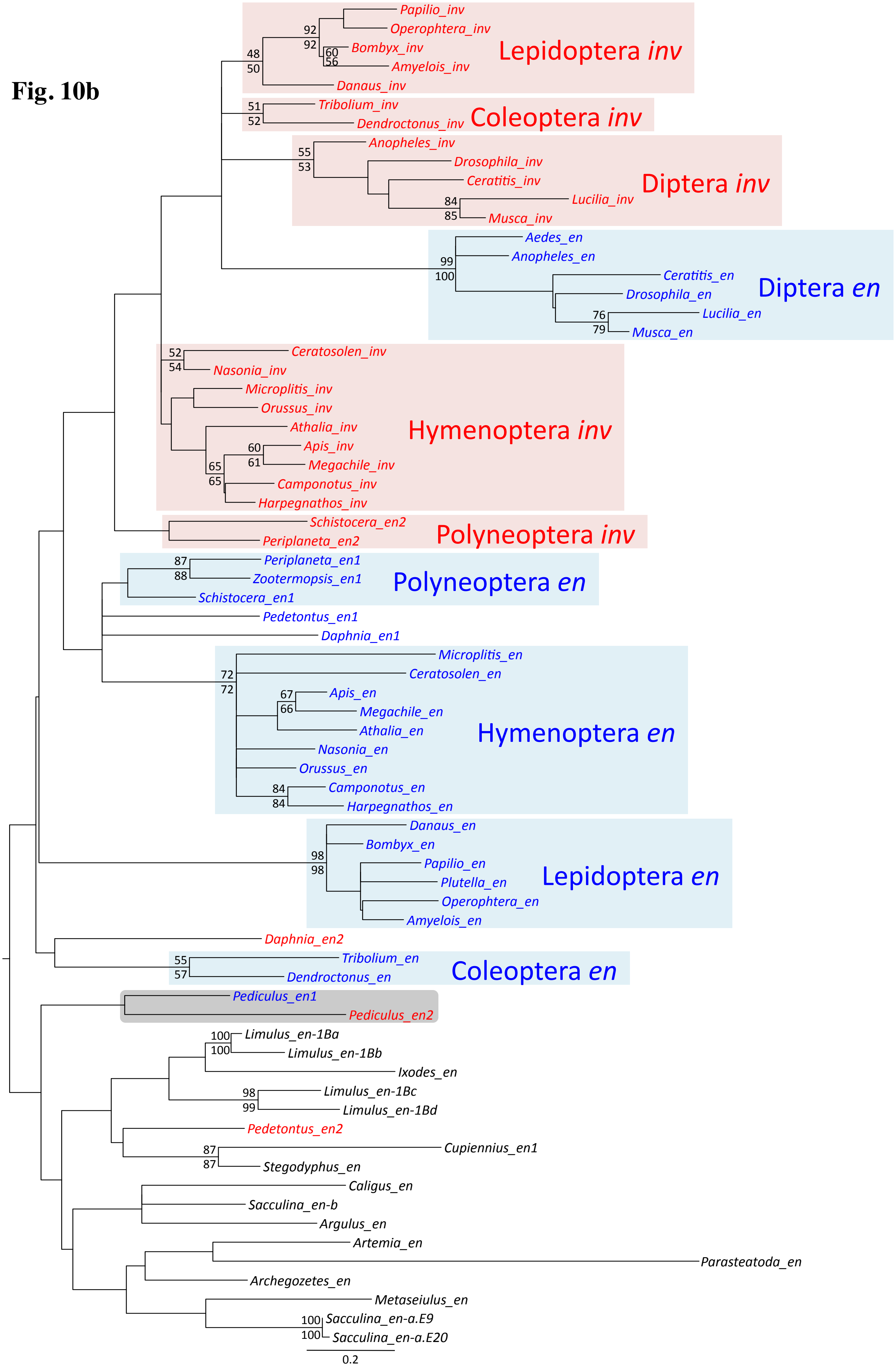

Papilio_inv
Operophtera_inv
Bombyx_inv
Amyelois_inv
Danaus_inv
92
92
60
56
48
50
Lepidoptera inv
Tribolium_inv
Dendroctonus_inv
51
52
Coleoptera inv
Anopheles_inv
Drosophila_inv
Ceratitis_inv
Lucilia_inv
Musca_inv
55
53
84
85
Diptera inv
Aedes_en
Anopheles_en
Ceratitis_en
Drosophila_en
Lucilia_en
Musca_en
99
100
76
79
Diptera en
Ceratosolen_inv
Nasonia_inv
Microplitis_inv
Orussus_inv
Athalia_inv
Apis_inv
Megachile_inv
Camponotus_inv
Harpegnathos_inv
52
54
60
61
65
65
Hymenoptera inv
Schistocera_en2
Periplaneta_en2
Polyneoptera inv
Periplaneta_en1
Zootermopsis_en1
Schistocera_en1
87
88
Polyneoptera en
Pedetontus_en1
Daphnia_en1
Microplitis_en
Ceratosolen_en
Apis_en
Megachile_en
Athalia_en
Nasonia_en
Orussus_en
Camponotus_en
Harpegnathos_en
72
72
67
66
84
84
Hymenoptera en
Danaus_en
Bombyx_en
Papilio_en
Plutella_en
Operophtera_en
Amyelois_en
98
98
Lepidoptera en
Daphnia_en2
Tribolium_en
Dendroctonus_en
55
57
Coleoptera en
Pediculus_en1
Pediculus_en2
Limulus_en-1Ba
Limulus_en-1Bb
100
100
Ixodes_en
Limulus_en-1Bc
Limulus_en-1Bd
98
99
Pedetontus_en2
Cupiennius_en1
Stegodyphus_en
87
87
Caligus_en
Sacculina_en-b
Argulus_en
Artemia_en
Parasteatoda_en
Archegozetes_en
Metaseiulus_en
Sacculina_en-a.E9
Sacculina_en-a.E20
100
100
0.2

Fig. 11 Comparison of genomic sequence from 3' side region of EH2 intronic region to 3' UTR in *Danaus_en* and *Danaus_inv*. Asterisks indicate the positions identical between *Danaus_en* and *Danaus_inv*. The continuous 807 positions shown with shading are identical except for three positions: two of nucleobase minmatchs and one of gap insertion.

Fig. 11

```
Danaus_en    ggtgagagatttattacagagtacagtcgcacgctaaactggtcagatccgcgccacagattatacagaattatattacgttaagccgct
               * *    **   *      *    * **  *         * *      ***************************************
Danaus_inv   tcttaatatttatatgattgttttgcttgcgtgaacccacatttaatgatacgccacagattatacagaattatattacgttaagccgct

Danaus_en    attacacgcatgggttgaaaaaaaaaacacttaagccgctagtacatggtaatatttacatacatatatataatatattttcatttaaaa
             ******************************************************************************************
Danaus_inv   attacacgcatgggttgaaaaaaaaaacacttaagccgctagtacatggtaatatttacatacatatatataatatattttcatttaaaa

Danaus_en    atgttaatcataaataaaaggtttaattaattcgcgataacaatccgtagcgattaatataaacgtcatgttattttatataatatttat
             ******************************************************************************************
Danaus_inv   atgttaatcataaataaaaggtttaattaattcgcgataacaatccgtagcgattaatataaacgtcatgttattttatataatatttat

Danaus_en    ttgacagtcgcgtgtttacatagtttttaaggcgataaacccgctggtacaactttattgtgaaattaattattactaaggaggttttat
             ******************************************************************************************
Danaus_inv   ttgacagtcgcgtgtttacatagtttttaaggcgataaacccgctggtacaactttattgtgaaattaattattactaaggaggttttat

                                                                EH2 intronic region <-|
Danaus_en    taaatgagaaatttaaattgttcatttgtaat-tttgttcaatgtatacaattagtttacgtttcgttaccaggtcccaggagtagacgg
             ******************************** *********************************************************
Danaus_inv   taaatgagaaatttaaattgttcatttgtaatatttgttcaatgtatacaattagtttacgtttcgttaccaggtcccaggagtagacgg

                                                                                        |-> Homeobox
                                                                                            intron
Danaus_en    gtgaagaagaaggcgagccctgaggagaagagaccgaggactgccttcagcgcctcgcagctaacaagattaaaggtactatataaaata
             ******************************************************************************************
Danaus_inv   gtgaagaagaaggcgagccctgaggagaagagaccgaggactgccttcagcgcctcgcagctaacaagattaaaggtactatataaaata

Danaus_en    tatatttatataagagaataaataattaaagccaggatttgaacctgctgtccttcgaacatagagtcttagttgcttgtgaagtcgagc
             ************************************************************************************* ****
Danaus_inv   tatatttatataagagaataaataattaaagccaggatttgaacctgctgtccttcgaacatagagtcttagttgcttgtgaagttgagc

                                        Homeobox intron <-|
Danaus_en    catcgtgtccctgtgagtgttattaaggtagctatgtttcggcagcacgagttcgcggagaaccgctacctgacggagaggaggaggcag
             ************* ****************************************************************************
Danaus_inv   catcgtgtccctgagagtgttattaaggtagctatgtttcggcagcacgagttcgcggagaaccgctacctgacggagaggaggaggcag

Danaus_en    gcgctggccgcggagctggggctggcggaggctcagatcaagatctggttccagaacaagagggccaagatcaagaaggcctcgggccag
             ******************************************************************************************
Danaus_inv   gcgctggccgcggagctggggctggcggaggctcagatcaagatctggttccagaacaagagggccaagatcaagaaggcctcgggccag

Danaus_en    aggaacccgctggcgctgcagctcatggcgcaggggctgtacaaccacgccacagtcaccgagagcgaggacgaggagatcagcgtcacg
             ************************************************  ***  *  *   **   **** ******      *  * *
Danaus_inv   aggaacccgctggcgctgcagctcatggcgcaggggctgtacaaccacagcaccataccgttgacgaaggaggaggaggagttagagatg

             |-> 3' UTR
Danaus_en    tagatgtatgatagtccgccattagtattaataagaaagttcctgtctattataataaat
              **      ** **   **        * *     *                 *  **
Danaus_inv   aaggccagggagagggagcagaggaattgagaggaatccggaggaagatcataacaaaga
                                        |-> 3' UTR
```

Fig. 12 Comparison for genomic sequences of each pair of engrailed-family genes in three lepidoterans, *Papilio*, *Bombyx* and *Operophtera*, in the sequence regions around the 5' and 3' ends of the homeobox intron. Asterisks indicate positions identical between the pair.

Fig. 12

```
                                                               →Homeobox intron
Papilio_en       agaccgagaaccgccttcagtgcttctcaactagcaaggctaaaggtataaatagtttgctttaacaactttaattagagcgacagt···
                  **** ** ** ******   *   **** ***** ** *********   ** *  *    *         *  *   **   * *
Papilio_inv      cgacctaggactgccttctccggacctcagctagctagactaaaggtaggcatggagtagacttcttctagcatctgctgcttgact···
```

```
                                             Homeobox intron←
Papilio_en       ···aatgttcctaacttcagactaatgtagttcttttgttcttagcacgagttcgctgagaaccgttacctgacggagcggcggcggcag
                    **       * *            ********** **** ********** ************************************
Papilio_inv      ···aagacataaatcaagcatacttattagttcttttattctcagcacgagtttgctgagaaccgttacctgacggagcggcggcggcag
```

```
                                                               →Homeobox intron
Bombyx_en        agaccacggaccgctttcagcggggcacaactcgcgagattgaaggtaggtaaaaaacgttatttttataggttcttcaatattact···
                 ****** ******* ***  **** *************  * ****** *** *  **       * *    * *      *
Bombyx_inv       agaccaaggaccgcattctccgggccacaactcgcgaggctaaaggtatgtatagtacccggcgataaatattgcacatcgaccgcg···
```

```
                                             Homeobox intron←
Bombyx_en        ···cggatgtcgtgacgaaacgctaatatatcgctttgtcctcagcacgagttcgccgagaaccgctacctgactgagcggcggcgtcag
                       *  *       *      ** *     ******************** ** ******** ** ***** *********** ***
Bombyx_inv       ···aaaacctttgacgcatcataaaaaaaggtactttgtcctcagcacgagtttgcagagaaccggtatctgacagagcggcggcggcag
```

```
                                                               →Homeobox intron
Operophtera_en   aggcctagaactgctttcagcgctgcacaacttggaagactaaaggtaaacatatataatatcacctttctactccgtaagctaagg···
                  * ** ***** ** ***   *   * ******* **** *********  * **   *   *         * * *     **
Operophtera_inv  cgaccaagaaccgccttctctgggccccaacttgcaagattaaaggtaattacatcataggacgagcagacgcacggctcagtacct···
```

```
                                             Homeobox intron←
Operophtera_en   ···gatctatacaaaaactattctaatttttttctttgttcatagcacgagtttgccgagaaccgctacctcaccgagcgtcgaagacaa
                        ***     *  ** **  **** *   * * *    ************************* *********************
Operophtera_inv  ···ttattatttttcatgtagtcacatttgtcattctctctccagcacgagtttgccgagaaccgctatctcaccgagcgtcgaagacaa
```