*engrailed*-family gene RT-PCR primer sequences:

*Oncopeltus fasciatus* EH1 (forward non-degenerate) 5’-GCGATCCTCAGCCCATCCTT-3’;
Forward *invected*-specific domain 5’-GGATCGCTGYTNWSNGTNGG-3’;
Forward EH2 5’-GGCCCGCCTGGGTNTAYTG-3’;
Reverse EH2 5’-CGGAGTACGGGTCARWAAYCCA-3’;
Reverse EH5 (inner) 5’-GTGGAGTGGTTGTACAGGCCYTNGCCAT-3’;
Reverse EH5 (outer) 5’-CGGTCATGGGGCGGTGSRWTGRTTRTA-3’

RT-PCR was performed using a touchdown program:
Annealing; 65°C – 55°C dropping 0.5°C each cycle for 20 cycles, then 20 cycles at 55°C.

The *invected*-specific domain and the cloning of *O. fasciatus* 5' sequence

(a)  

<table>
<thead>
<tr>
<th>Species</th>
<th>Amino Acid Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. melanogaster</td>
<td>ENSVLSVGQE</td>
</tr>
<tr>
<td>A. gambiae</td>
<td>DD-E---------</td>
</tr>
<tr>
<td>B. mori</td>
<td>DDT------N-</td>
</tr>
<tr>
<td>A. mellifera</td>
<td>TSE------S-</td>
</tr>
<tr>
<td>T. castaneum</td>
<td>DDT------N-</td>
</tr>
<tr>
<td>S. gregaria</td>
<td>AD-L------S-</td>
</tr>
<tr>
<td>P. americana</td>
<td>-DEL------S-</td>
</tr>
</tbody>
</table>

(b)  

```
5'  Inv  | EH1 | EH2 | EH3 | EH4 | EH5 | 3' 
RT-PCR1 |     |     |     |     |     |     
RS micro-exon  
RT-PCR2   
```

(a) Amino acid alignment of the Invected-specific domain. (b) Using degenerate primers (blue) targeted against EH2 and sequence encoding the Invected-specific domain, we amplified 5’ sequence from an *O. fasciatus* engrailed-family gene (RT-PCR1). To identify which *O. fasciatus* gene this sequence was from, we designed a gene specific primer (red) targeting sequence within EH1, and used it in combination with an EH5 degenerate reverse primer to amplify 3’ sequence (RT-PCR2). Only clones encoding the RS motif were recovered, indicating that the 5’ sequence originates from *O.f.en-r2* (Peterson et al, 1998).
cDNA sequence of *Schistocerca gregaria* engrailed-1 (*Sgen-1*). The coding region is in larger font and accompanied by the translated amino acid sequence. The 5′ and 3′ untranslated regions are in small font. An in-frame upstream stop codon is underlined. The *Hind*III restriction site in the 3′UTR that was used in the production of the *in situ* hybridisation probe is highlighted in red. *Sgen-1* encodes a 287 amino acid protein. **EH1/Groucho-binding domain** is in red. **EH2/Extradenticle-binding domain** is in orange. **EH3** is in dark red. **EH4 or Homeobox/Homeodomain** is in blue. **EH5** is in green.
**Schistocerca gregaria engrailed-2 (Sgen-2).** The coding region is in larger font and accompanied by the translated amino acid sequence. The 3′ untranslated region is in small font. The SalI restriction site in the 3′UTR used in the production of the in situ hybridisation probe is highlighted in red. Sgen-2 encodes a protein at least 264 amino acid long. We are unsure whether the first methionine is the start codon due to the lack of an upstream stop codon and Kozak consensus sequence. The **invected-specific domain** is underlined. EH1/Groucho-binding domain is in red. EH2/Extradenticle-binding domain is in orange. The hexanucleotide micro-exon sequence and RS-motif is in violet.
CLUSTAL W (1.82) pairwise alignment of the Sgen-1 and Sgen-2 3'UTR in situ hybridisation probes. Note that the longest contiguous stretch of sequence identity (apart from the polyA tails) is only ten nucleotides in length. Gap penalties were set at the lowest possible levels, but alignments constructed with higher penalties for gaps had even shorter maximum stretches of contiguous sequence identity. The probes are unlikely to cross hybridise.

<table>
<thead>
<tr>
<th>Sgen-1</th>
<th>Sgen-2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>-AGCTTGTAGCTTA-AT-TATA-AGT-TTATG-CTG-GT-ATCTTTGTGTTCTG 46</strong></td>
<td></td>
</tr>
<tr>
<td><strong>TCGACT-GTA-TTATGTTAGAGGCGGCAAACCT-GCGATTGTGTTTATC-CTG-CTC 54</strong></td>
<td></td>
</tr>
<tr>
<td><strong>-TTCAC-GA-GAATAA-TGCTG-ATAATG-AGTGATAGGAATTC-CAGCAGTTAA-96</strong></td>
<td></td>
</tr>
<tr>
<td><strong>TTAATCT-GAACAATGACTG-TGAATAATTTAGT-ACC-CATTTCATTCA-CA-CTGAC 106</strong></td>
<td></td>
</tr>
<tr>
<td><strong>AAATG-AGCATTTCTCCCAAGGATCTACCGACATTAGT-CTTATACGTATCA-G-ACA 162</strong></td>
<td></td>
</tr>
<tr>
<td><strong>-CTGCTGTATTTTGC-CATTAGC-TAAATATCATTATTTATTTTGG-TAT-G-C 192</strong></td>
<td></td>
</tr>
<tr>
<td><strong>ACTG-C-AG-A-ACACAA-AGTGCAAAT-AGATATTATA-GGAT-GAGA-AGA 208</strong></td>
<td></td>
</tr>
<tr>
<td><strong>CA-GAATACGTGCTGAAA-ATGCTCCAAGAAAAGATA-GTA-GAACAGATT-AAAAAA 241</strong></td>
<td></td>
</tr>
<tr>
<td><strong>CAAGGAGCT-TTATCACAATCTGCA-AATAAGCTAAAGCAAAACAATTTTTTAAA 265</strong></td>
<td></td>
</tr>
<tr>
<td><strong>-CAAAATATGCGTGTTGAT-CGAAATAAATGGGTACGATTTTTCAT-TCGAA 312</strong></td>
<td></td>
</tr>
<tr>
<td><strong>T-TATTATTA-TTTATATTTATTTATTTATTTGGAGATGTGATGATTATTTTACAAC 379</strong></td>
<td></td>
</tr>
<tr>
<td><strong>GC-CTG----TACATATCAGTAAACACT-TAAGA-AAAATCA-G-A-GACACGATGT 373</strong></td>
<td></td>
</tr>
<tr>
<td><strong>GCTGTGAAGGTATATATT-ATA-TACTGTAGTGATCAAGCTAAACGTCGGACACAT 434</strong></td>
<td></td>
</tr>
<tr>
<td><strong>TTACT-CTTATCCTTCTG-CCAAAT-ATGACGTGCAATCAA-GCCGTTTACAT 487</strong></td>
<td></td>
</tr>
<tr>
<td><strong>-CAGTATAGT-CAATATTTTCCTCAAGTAAAGGTCCCTTGGG-GGT-GGTG-ATCCT 471</strong></td>
<td></td>
</tr>
<tr>
<td><strong>TTTCGATGTACAACTGGTGTGTCT-CTTGATGAACTC-TTTGATGCAAGTACATG 545</strong></td>
<td></td>
</tr>
<tr>
<td><strong>TCATTGAGA-AT-GAT-GT-CTG-CTG-A-C-AGGAATT-GTATTAAA-AA-TA 514</strong></td>
<td></td>
</tr>
<tr>
<td><strong>-CAATAGGAAGTATATCTTGTCAATCGGAAATGCTCGGTGTTGAAACGAT 603</strong></td>
<td></td>
</tr>
<tr>
<td><strong>AATAAACATGTAT-A-TTATTTTCATGGCAACATGTTAATAACTCAGATGTCGACAAGAT 572</strong></td>
<td></td>
</tr>
<tr>
<td><strong>GATAAGTGGTTTGTGTGTTTTTCTC-T-CTC-TCCTG-CT-GAGAGAAGA-653</strong></td>
<td></td>
</tr>
</tbody>
</table>
Sgen-1/Sgen-2 3'UTR alignment (cont.)

Sgen-1
GTCTGTTGCTCCA-G-G-CTTCTAGTCCATT---TC-A-T---AAAC--AATACGTTACC 619
ACCCTG-TC-CCATATATCTTCT---TTTATGATCTAATACGGAAGACAATCCACACACAC 708
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-2
ACGTTCATCAGGCTGTGAACTTGACTGGAGAATCATT-GGTTTCCCT-CAG-ACCTGT 671
AGGTG-CAGC-AGTAACCTTTTTTTAATACTTCATCTGCTGCTTTTACCATGACT-C 764
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
TG---GAG---GA------------AATTCTTAA---GAA-AT---TA-AA---AAA-AAA-GC 704
Sgen-2
TGAAATTAGAAGTACATAGAAGTTATTTATCTGCTGCTTATATACATATATATATTATATAT 824
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
TCAAGT-TGC---TGC-AAGTTTGAT-TCTT-CTT---CA-GTTATTTTTGAGA-AGTT- 751
TTAAATTATGCAATCTGCTGAGAGATTTATCTATGCAATGAGATCAAGATATCAGGCTTCAACA 880
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-2
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
TTGTAGTTAAAA---AAT-ATCTATTATCTTTGTGAATAAAAGGAAACATATTACC 806
Sgen-2
TTACA-CCCAATGGTCTTATTTATTATT-C-TCCA-GTGTTAGAAGA-GTAA---TG-- 928
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
AAAAACATT---TCG-TTT-ATAACT---A-ATA---ACA-AT--CAT---TTTGTG 845
GAAACATTGCTGCTTATGGTGATGCTGCGATCTGGATGAGGCTGAGTACTTCCACA 988
*** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-2
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
-TTGCT-TATC-C-G-------------AAAT-ATAGA-C-A-AATATTTAG-TGATT-TA 886
Sgen-2
TGCTGCTTCCTCTCTGACTGACAAAAATATGAAAAAGCAGGATTTGATG-TATAAAGGA 1047
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
ACACAG---TG---ATA-AACTTGAAGATG---ACTCGTGTTAA-CAG-CTAG--- 928
Sgen-2
AGACAAAAATGTACATATAATTTTCAAAATATGCAGAGGACATATTATATACATTGATGTCATCCTCAATGAATGT 1106
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-2
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
-GA----AAAT---G-T---TTAA---GA---AAGTG-A-ATGACG-GTCAAGGCCT 965
Sgen-2
ACAGAAAAATATCTGATATAATTAAATATGTTGATTTATATACATTGATGTCCTCCTCAATGAATGT 1163
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
GTAGATT---TATATGT-------AGCA---G---AGCA-A-ATAATATATGTTG-ATT-AG 1007
Sgen-2
ACA-TATTTCTACATGTTTCTCTGTGATGCTAGAGACTAAT-TAC---TGATTTA- 1217
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
TACA-CA-ACA------TC-CTGTTGA-GA-CTGCG--TAC---GAA-A-TGGGCTGTTCTCAACA 1050
Sgen-2
TGCGA-TGAGCAAGGTTTCTCTTGCTGAAACAGCAATTTTCAATGGAAGATTATGCCC---GA 1274
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
Sgen-2
TTGATT-GTCACTGTTAAGATGATGTATTGATATTTATATATACATCTGCAAGATTACCTCTTCTTTT 1333
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
GTGATATCCTTC-TATAA-------TGAA-G-AAT-AATTAAAGACATATGAA---- 1137
Sgen-2
CTCAT---CT-CATCGAAGAATGATCTGGATTTATATATATTTATGTGATGAGCTCAGAATGAACTCT 1389
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Sgen-1
AAAAACATT---CCTAAAAAAAAAAAAAAA 1169
Sgen-2
TACTTCACTTTAATTTTTTTTTTTTTTTT 1420
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Whole mount in situ hybridisation for Sgen-1 (A-B, E, G-H, K-L, O) and Sgen-2 (C-D, F, I-J, M-N, P), on embryos from five egg pods (boxes) at different stages in development. Embryos from the same egg pod (boxes) are synchronised in their development to within about 1%. At 15% of development (A-D), Sgen-2 is typically expressed as two chevrons in the head (C), while Sgen-1 is typically expressed as only one chevron (A) (B & D: The same embryos stained with methylene blue). By 17% of development (E & F) expression of both genes has appeared in seven stripes, corresponding to the antennal, three gnathal and three thoracic segments. Note that expression in the maxillary segment (*) is particularly weak, perhaps suggesting it has only just appeared. Between 17% and 30% development abdominal stripes appear in an anterior to posterior progression (arrows) (G-N). The within pod variation in expression patterns is the same for Sgen-1 as for Sgen-2 (compare G-H with I-J, and K-L with M-N). By 30% of development expression of both genes has also appeared in the intercalary segment (red arrows), dorsal ridge (blue arrows) and cerci (green arrows) (O & P).
The genomic DNA sequence (minus introns) of *Tribolium castaneum* engrailed.
This second engrailed-family gene was not identified by Brown *et al.*, 1994. The coding region is in larger font and accompanied by the transcribed amino acid sequence. The 5' and 3' untranslated regions are in small font. An in-frame upstream stop codon is underlined. *Tcen-1* encodes a 284 amino acid protein. **EH1/Groucho-binding domain is in red.**
**EH2/Extradenticle-binding domain is in orange.** **EH3** is in dark red. **EH4** or **Homeobox/Homeomain** is in blue. **EH5** is in green. The position of the EN2 intron is shown with an arrow.
The genomic sequence (minus introns) of *Tribrobilum castaneum invected*. The coding region is in larger font and is underlined by the translated amino acid sequence. The 5' & 3' untranslated regions are in small font. In-frame upstream stop codons are underlined. *Tel2* encodes a protein 327 amino acid long. Bases and amino acids that differ from the sequence published by Brown et al., 1994 are boxed. The *invected*-specific domain is underlined. EH1/Groucho-binding domain is in red. EH2/Extradenticle-binding domain is in orange. The hexanucleotide micro-exon sequence and RS-motif is in violet. EH3 is in dark red. EH4 or Homeobox/Homeodomain is in blue. EH5 is in green. The positions of the EH2 introns are shown with arrows.