J Mol Evol (1993) 36:536–544

Journal of Molecular Evolution

Evolution of the *dec-1* Eggshell Locus in *Drosophila*. II. Intraspecific DNA Sequence Analysis Reveals Length Mutations in a Repetitive Region in *D. melanogaster*

Stefan Andersson[1] and Andrew Lambertsson[2]

[1] Department of Genetics, University of Umeå, S-901 87 Umeå, Sweden
[2] Department of Biology, Division of General Genetics, University of Oslo, P.O. Box 1031 Blindern, N-0315 Oslo 3, Norway

Summary. The *dec-1* eggshell gene in *Drosophila melanogaster* encodes follicle cell proteins required for proper eggshell assembly. As shown by Southern and Northern analyses the *dec-1* gene occurs in four alleles (*Fc1–4*) among wild-type strains. Its second exon has a distinct feature in the form of 12 repeats with 78–91 nucleotides; the first five show nearly 100% homology. DNA sequence comparison of the repeated region of the alleles revealed that the length polymorphisms are caused by changes in the numbers of the first five repeats. The results suggest that the alleles have been generated by unequal intragenic crossing-over and/or slippage during DNA replication and that the allelic length variants have arisen independently. The possibility that the most common allele, *FC1*, has a selective advantage over the other alleles is discussed.

Key words: Drosophila — *dec-1* eggshell gene — Wild-type variants — Repeated region — DNA sequencing

The X-linked *defective chorion-1* (*dec-1*) gene produces at least three transcripts by alternative RNA splicing during stages 9–12 of oogenesis (Waring et al. 1990). The developmental expression and genetic localization to region 7C1-9 on the X-chromosome of *dec-1* were first reported by Lineruth and Lambertsson (1985, 1986) and Lineruth et al. (1985), and later by Bauer and Waring (1987) and Komitopoulou et al. (1988). It has now been cytologically mapped to 7C3-4 (Hawley and Waring 1988; A. Lambertsson, T. Johansson, and S. Andersson, unpublished results). In *D. melanogaster,* several female-sterile mutants belonging to a large complementation group at 7C on the X-chromosome have been correlated to this locus (Lineruth and Lambertsson 1986; Komitopoulou et al. 1988). A screen of 130 *D. melanogaster* wild-type strains revealed four electrophoretic protein variants, *Fc1–Fc4* (Lineruth and Lindberg 1988), *Fc1* being by far the most common one. The protein products are posttranslationally modified (Lineruth and Lambertsson 1985; Bauer and Waring 1987; Lineruth 1987; Hawley and Waring 1988; Komitopoulou et al. 1988). However, the function(s) of the polypeptides encoded by the *dec-1* transcripts is as yet unknown.

The *dec-1* gene has been cloned and sequenced in several *D. melanogaster* wild-type and *dec-1* mutant strains and in other *Drosophila* species (Waring et al. 1990; Andersson and Lambertsson 1991; Andersson and Lambertsson, unpublished). The most conspicuous feature of the gene is a portion of the coding sequence that contains five identical copies of a repeated sequence 78 bp in length followed by seven repeats of varying length and homology (Waring et al. 1990). Sequence comparison between *D. melanogaster* and *D. erecta* showed that the amount of nucleotide and amino acid substitution in the repeated region is much larger than in the 5′ translated region (Andersson and Lambertsson 1991).

Offprint requests to: A. Lambertsson

Our initial molecular studies of the *dec-1* gene demonstrated that the protein variability was generated by a small deletion of varying length (65–200 bp) in a repeated portion of the central transcribed region in *D. melanogaster* (Andersson and Lambertsson 1991). Length variants were also found in *D. simulans* and *D. erecta* in a region of the *dec-1* gene corresponding to the repeated region in *D. melanogaster* (Andersson and Lambertsson 1991). Length polymorphism in coding DNA has also been observed in other internally repetitive genes (Muskavitch and Hogness 1982; Goodbourn et al. 1983; Swallow et al. 1987; Lyons et al. 1988; Teumer and Green 1989; Costa et al. 1991). All or part of this variability could be produced by unequal crossing-over (Smith 1976) or by slippage of DNA (Levinson and Gutman 1987; Dover 1989).

We have now extended these studies by analyzing at the DNA sequence level the repeated portion of the four variants in *D. melanogaster*. This analysis shows that the length differences are caused by changes in the numbers of the first five identical repeats.

Materials and Methods

Drosophila *Strains*. The wild-type strains Samarkand, Ghangry, Israel, Shahrinau, and Dilizhan were from the European Drosophila Stock Center in Umeå, Sweden.

D. melanogaster *Genomic Libraries*. High-molecular-weight genomic DNA (>100 kbp) was prepared from the Samarkand and Shahrinau wild-type strains by the procedure described in Lambertsson et al. (1989). DNA was partially digested with *Sau*3A to yield maximum molecules in the 14–20-kbp range (for details see Maniatis et al., 1982), dephosphorylated, ligated to *Bam*HI EMBL3 vector arms, and packaged into viral particles using Packagene extracts (Promega Biotech).

The *dec-1* locus was cloned in a chromosome walk from the *singed* locus at 7D1-2 (starting with the *sn9* clone, kindly provided by Dr. K. O'Hare) to 7C2-3 using an EMBL3 lambda phage genomic library from the Shahrinau wild-type strain.

Subcloning. When necessary, fragments from the EMBL3 clones were subcloned in the pUC19 plasmid.

Nucleic Acid Techniques. For restriction analysis adult *Drosophila* DNA was prepared according to the method described by Jowett (1986). Both small- and large-scale preparations of phage and plasmid DNA were according to Maniatis et al. (1982).

Total nucleic acids for Northern analysis were isolated from ovaries as described before (Hansson and Lambertsson, 1983).

DNA probes were obtained by restriction enzyme digestion, electrophoresis in low-melting-temperature agarose gels, excision of the desired fragment, and labeling of the DNA with [^{32}P]dCTP (sp act 1–2 × 10^8 cpm/μg) using the Promega Prime-a-Gene system (Promega Biotech).

Electrophoresis and Nucleic Acid Blots. After digestion the genomic DNA was fractionated on 0.8% agarose gels. Cloned *Drosophila* DNA was separated on either 0.4 or 0.8% agarose gels. RNA fractionation was on 1% agarose-formaldehyde gels as described before (Hansson and Lambertsson 1983). All blotting onto GeneScreen Plus filters (Dupont, NEN Research Products) was performed by the VacuGene Vacuum Blotting System (Pharmacia LKB Biotechnology AB). The filters were treated, hybridized, and washed according to the supplier's instructions.

Sequence Analysis. The *dec-1 Bam*HI-*Bam*HI fragment (Fig. 1, position +0.8 to +2.5) from Samarkand and Shahrinau was subcloned into the pUC19 vector. To obtain the homologous fragment from the Israel, Ghangry, and Dilizhan wild-type strains genomic DNA (1 μg) was subjected to polymerase chain reaction (PCR) amplification with *Taq* DNA polymerase (Promega) using either primers Dm1–Dm6 or Dm3–Dm6 (Dm1 = 5′[bp + 1]-GAGCTCCGGCGAACACAGATC-3′[bp + 21]; Dm3 = 5′[bp + 731]-GAATTCTGCAGATCCTCCGGCAG-3′[bp + 752]; Dm6 = 5′[bp + 2,168]-GTCGACGGATCCTCT-GTCCACTGC-3′[pb + 2,151]; this primer contains a *Bam*HI site-GGATCC). Exponential amplification with *Taq* DNA polymerase was performed in a 50-μl reaction mix containing 50 mM KCl, 10 mM Tris-HCl (pH 9.0 at 25°C), 1.5 mM $MgCl_2$, 0.01% gelatin (w/v), 0.1% Triton X-100, 0.2 mM for each of the dNTPs, 10 pmol of each primer, 1 μg of genomic DNA (denaturated by boiling for 5 min), and 2 units of *Taq* DNA polymerase (Promega). The reaction was overlaid with mineral oil. Thirty-five cycles (1 min at 94°C, 1 min at 55°C, 2.5 min at 72°C; the mix was finally left for 10 min at 72°C) were performed in a Temp-Cycler (model 60, Coy Laboratory Products, Inc.). The PCR products were then digested with *Bam*HI and the *Bam*HI-*Bam*HI fragment was cloned into the pUC19 vector.

Sequencing was performed using the pUC/M13 forward and reverse primers, [^{35}S]dATP, the chain termination technique (Sanger et al. 1977), and the TaqTrackSequencing system (Promega) or Sequenase 2.0 (USB) following the instructions of the suppliers. Gels (5%) of 38 × 50 cm were used, and to maximize the number of sequenced nucleotides, samples from each reaction were usually applied at intervals of 2 or 3 h. All sequences have been read at least twice.

It should be mentioned that only one amplified and subcloned PCR product each from Israel/*Fc2*, Ghanghry/*Fc2*, and Dilizhan/*Fc4* was sequenced. Since *Taq* polymerase does not contain any measurable 3′-to 5′ exonuclease "proofreading" activity, one should be aware of that nucleotides may be misincorporated at an error rate of 1×10^{-4} nt/cycle (Erlich 1992 and references therein). However, our sequence analysis (see Results and Fig. 4) of the cloned PCR products from the three strains mentioned above seems to show that undirected mutations have not occurred during the PCR.

Results

Southern and Northern Blot Analyses

The *Fc*-proteins were found in four variant forms among 130 wild-type strains (Lineruth 1987; Lineruth and Lindberg 1988), all producing normal and fertile eggs. Genomic DNA from these four variants was digested with *Bam*HI-*Hind*II and separated on 0.8% agarose gels. The DNA was blotted onto GeneScreen Plus filters and probed with the *Sac*I-*Sac*I 5.1-kb fragment (Fig. 1, position +0.0 to +5.1) from the transcribed region of *dec-1*. Figure 2

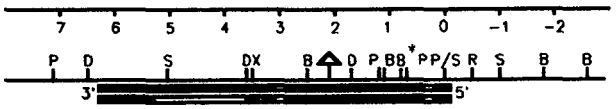

Fig. 1. Restriction map of the *dec-1* region of *D. melanogaster*. (Δ) denotes deletions in the repeated region, *filled and open boxes* represent exons and introns, respectively, in the two *dec-1* transcripts—the upper is 5.7 kb and the lower is 3.7 kb. Restriction enzymes: *B* = *Bam*HI, *D* = *Hin*dII, *P* = *Pst*I, *R* = *Eco*RI, *S* = *Sac*I, *X* = *Xba*I.

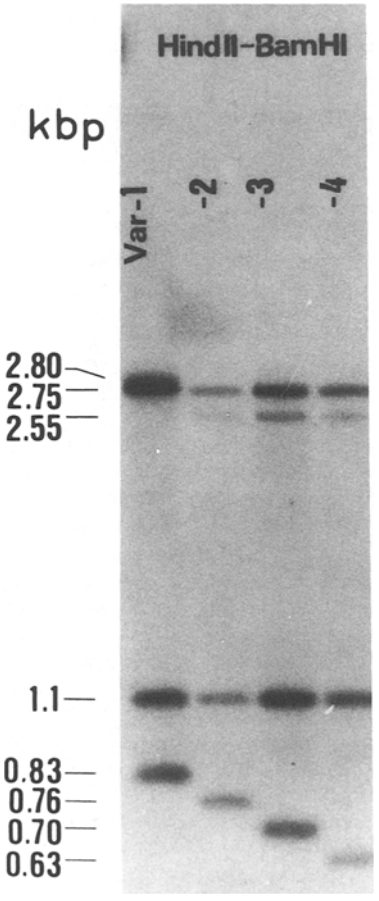

Fig. 2. Southern blot analysis of genomic *D. melanogaster* DNA. The DNAs were double digested with *Bam*HI and *Hin*dII. The 5.0-kb *Sac*I-*Sac*I fragment was used as probe (see Fig. 1; coordinates 0.0 to 5.0) at 4–6 × 10^5 cpm/ml in the hybridization mixture. Var-1, -2, -3, and -4 (corresponding to *Fc1*, *Fc2*, *Fc3*, and *Fc4*) denote wild-type strains Samarkand, Israel, Shahrinau, and Dilizhan, respectively; 4 μg DNA was loaded in each lane.

shows that Samarkand/*Fc1* lacks the *Bam*HI site at position +0.8, which makes the 2.55-kb *Bam*HI-*Bam*HI (Fig. 1, position −1.8 to +0.8), present in *Fc2–4*, 2.8 kb instead. However, the most striking observation is that variant forms *Fc2–4* are associated with a small deletion of approximately 70, 130, and 200 bp, respectively (Fig. 2), within the 0.83-bp *Hin*dII-*Bam*HI fragment (Fig. 1, position +1.7 to +2.5). This fragment contains most of the repeated region (Waring et al. 1990; see below).

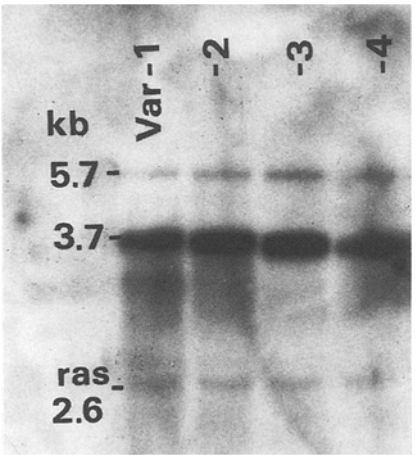

Fig. 3. Northern blot analysis of total ovarian nucleic acids. Nucleic acids were extracted from 15–20 pairs of ovaries; 8 μg of total nucleic acids was electrophorized in each lane. Hybridization was with the [^{32}P]labeled *Sac*I-*Sac*I probe used in Fig. 1. After stripping (and reexposure to confirm that the *Sac*I-*Sac*I probe had been removed), the filter was rehybridized with the *Drosophila ras* gene (internal standard). Var-1, -2, -3, and -4 (corresponding to *Fc1*, *Fc2*, *Fc3*, and *Fc4*) denote wild-type strains Samarkand, Israel, Shahrinau, and Dilizhan, respectively.

Total ovarian RNA from the four *Fc*-forms was extracted, separated on agarose-formaldehyde gels, and blotted onto GeneScreen Plus filters. The filters were probed with the *Sac*I-*Sac*I 5.1-kb fragment (see above), using the *ras* oncogene as internal standard. Figure 3 shows two transcripts of 3.7 and 5.7 kb in size in variant-1. In fact, there are two 3.7-kb transcripts, a and b, which accumulate in stages 9–10 in a 1:10 ratio, and the 5.7-kb transcript accumulates in stages 11–12 of oogenesis. (For further details on these transcripts see Waring et al. 1990.) Figure 3 also shows the reduced sizes of the *Fc2–4* mRNAs due to the deletion described above.

DNA Sequence of the Variable-Length Region

As shown by Waring et al. (1990) the *dec-1* gene contains a variable part composed of a tandemly repeating sequence. The basic repeat is 78 bp long, and there are 12 copies in the longer allele, *Fc1*. The first five repeats and the first two-thirds of repeat 6 show a nearly 100% homology and make up a somewhat different unit of the repeated region; the next seven repeats are of varying length and homology. The repeating unit seems to be able to tolerate extensive deletions, since females carrying deletions in this part of the gene seem to be as fertile as females with longer protein forms (K. Lineruth, unpublished results).

Sequences of the variable region from the four variants (two *Fc2* variants) were subcloned into

Samarkand/*Fc1*

```
          5   10   15   20   25   30   35   40   45   50   55   60   65   70   75
          '    '    '    '    '    '    '    '    '    '    '    '    '    '    '
cons CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep1 CAGAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep2 CAAAATCCAATGATGATGCAGCAACGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep3 CAAAATCCAATGATGATGCAGCAACGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep4 CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCTAAGATCCAACAGAATCAACAGCAGATCCAG
rep5 CAAAATCCAATGATGGTGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep6 CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATTCAGCATGATCAACAGATGGCA
                                                                    CAACAGATGGCACAG
rep7 CAAGGTCTCATGATGACGGAGCAGAG
                              GCAAAGGCAGTGGTCAGAAGATCAGGCCAAAATTCAGCAGGCTCAACAGATGGCCCAA
rep8 CAGACACCCATGATGATGCCACA
                         GATGCAACAAAGGCAGTGGACAGAGGATCC
```

Israel/*Fc2*

```
          5   10   15   20   25   30   35   40   45   50   55   60   65   70   75
          '    '    '    '    '    '    '    '    '    '    '    '    '    '    '
cons CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep1 CAGAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep2 CAAAATCCAATGATGATGCAGCAACGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep4 CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCTAAGATCCAACAGAATCAACAGCAGATCCAG
rep5 CAAAATCCAATGATGGTGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep6 CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATTCAGCATGATCAACAGATGGCA
                                                                    CAACAGATGGCACAG
rep7 CAAGGTCTCATGATGACGGAGCAGAG
                              GCAAAGGCAGTGGTCAGAAGATCAGGCCAAAATTCAGCAGGCTCAACAGATGGCCCAA
rep8 CAGACACCCATGATGATGCCACA
                         GATGCAACAAAGGCAGTGGACAGAGGATCC
```

Ghanghry/*Fc2*

```
          5   10   15   20   25   30   35   40   45   50   55   60   65   70   75
          '    '    '    '    '    '    '    '    '    '    '    '    '    '    '
cons CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep1 CAGAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep2 CAAAATCCAATGATGATGCAGCAACGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep3 CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep4 CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCTAAGATCCAACAGAATCAACAGCAGATCCAG
rep6 CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATTCAGCATGATCAACAGATGGCA
                                                                    CAACAGATGGCACAG
rep7 CAAGGTCTCATGATGACGGAGCAGAG
                              GCAAAGGCAGTGGTCAGAAGATCAGGCCAAAATTCAGCAGGCTCAACAGATGGCCCAA
rep8 CAGACACCCATGATGATGCCACA
                         GATGCAACAAAGGCAGTGGACAGAGGATCC
```

Fig. 4. Nucleotide sequences of the *dec-1* repeated coding region from length variants *Fc1–Fc4* (two *Fc2*). *cons*, consensus repeat sequence; *rep1–rep8* represent the eight different repeats. Nucleotides in *bold type* indicate substitutions relative to the consensus sequence. The DNA sequences shown in *italic* are the internal tandem repeats; these are shown under each other to maintain alignment of the other repeats. Continued on page 540.

pUC19 either from genomic libraries or from PCR reactions. Figure 4 shows the sequences of repeats 1–8 from the four variants. All variants are compared with the Samarkand/*Fc1* repeated region and the repeat consensus sequence from this strain (Fig. 4). Starting with Samarkand/*Fc1*, the largest and most common *dec-1* allele, pairwise comparisons relative to the consensus repeat sequence reveal that repeat 1 has an A → G exchange at position 3; repeats 2 and 3 both have an A → C exchange at position 24; otherwise these two repeats are identical. Repeat 4 has a C → T exchange at position 48

Shahrinau/*Fc3*

```
          5   10   15   20   25   30   35   40   45   50   55   60   65   70   75
          '    '    '    '    '    '    '    '    '    '    '    '    '    '    '
cons CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep1 CAGAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep2 CAAAATCCAATGATGATGCAGCAACGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep3 CAAAATCCAATGATGATGCAGCAACGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep6 CAAAATCCAATGGTGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATTCAGCATGATCAACAGATGGCA
                                                                    CAACAGATGGCACAG
rep7 CAAGGTCTCATGATGACGGAGCAGAG
                              GCAAAGGCAGTGGTCAGAAGATCAGGCCAAAATTCAGCAGGCTCAACAGATGGCCCAA
rep8 CAGACACCCATGATGATGCCACA
                        GATGCATCAAAGGCAGTGGACAGAGGATCC
```

Dilizhan/*Fc4*

```
          5   10   15   20   25   30   35   40   45   50   55   60   65   70   75
          '    '    '    '    '    '    '    '    '    '    '    '    '    '    '
cons CAAAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep1 CAGAATCCAATGATGATGCAGCAAAGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep2 CAAAATCCAATGATGATGCAGCAACGACAATGGTCGGAGGAGCAGGCCAAGATCCAACAGAATCAACAGCAGATCCAG
rep6 CAAAATCCAATGATGATGCAGCAACGACAATGGTCGGAGGAGCAGGCCAAGATTCAGCATGATCATCAGATGGCA
                                                                    CAACAGATGGCACAG
rep7 CAAGGTCTCATGATGACGGAGCAGAG
                              GCAAAGGCAGTGGTCAGAAGATCAGGCCAAAATTCAGCAGGCTCAACAGATGGCCCAA
rep8 CAGACACCCATGATGATGCCACA
                        GATGCAACAAAGGCAGTGGACAGAGGATCC
```

Fig. 4. Continued from page 539.

and in repeat 5 there is an A → G exchange at position 16 (Fig. 4).

Repeat 6 is 12 nucleotides longer than repeats 1–5, and this is due to an internal tandem repeat of nucleotides 64–75 (Fig. 4). It is noteworthy that the exchanges in this part of the sequence are conserved between the internal tandem repeats of all variants except in Dilizhan/*Fc4*, where there is an A → T substitution at position 66. The exchanges at positions 70, 71, 73, and 74 are also present in repeat 7. Apart from the exchanges mentioned above, repeat 6 is identical with the consensus repeat through nucleotide 53. Then there are four exchanges among the eight following nucleotides: a C → T at position 54, an A → G at position 57, a G → T at position 60, and at position 61 there is an A → G exchange. From position 70 there is even less homology to the other repeats. (See above).

Imperfect internal tandem repeats are found in repeats 7 and 8 (Fig. 4). In repeat 7 it is six nucleotides between positions 21 and 26, and in repeat 8 it is nine nucleotides between positions 15 and 23; in both cases the internal tandem repeats are out of step with the amino acid frame. In repeat 7 the A → G exchange at position 24 in the first internal tandem repeat is reversed in the second. Similarly, in repeat 8 there is a reversion in the second internal tandem repeat of the A → C substitution at position 20. Note that the G → A exchange at position 21 in repeat 8 is conserved in all variants except in Shahrinau/*Fc3*, where there is an A → T substitution in the second internal tandem repeat. Besides the internal tandem repeat both repeats 7 and 8 show less homology to the consensus repeat, and repeat 8 is also considerably shorter. (See also Waring et al. 1990.).

These exchanges give each repeat a specific and useful "fingerprint"; note, however, that 2 and 3 are identical in Samarkand and in Shahrinau. (See below.) Altogether, there are 54 substitutions in the first 8 repeats (617 nucleotides) in Samarkand/*Fc1*; 32 are transitions and 22 are transversions. In the first 5 repeats and the first 2/3 of repeat 6 (443 nucleotides) there are only 3 transitions and 2 transversions; of these, the A → G transition at position 16 in repeat 5 (Fig. 4) leads to an amino acid exchange. (See below and Fig. 5.) The other 49 substitutions, 29 transitions, and 20 transversions are in the remaining 173 nucleotides and several of these result in amino acid substitutions (Fig. 5).

Sequence analysis of the repeating region of the wild-type Israel/*Fc2* shows that repeat 3 (or repeat 2) is deleted. Alternatively, this deletion could be explained by taking away any 78 consecutive nucleotides between nucleotide C at position 48 in repeat 1 and nucleotide A at position 25 in repeat 4. How-

Samarkand/*Fc1*

rep1 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep2 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep3 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep4 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep5 QNP MM**V** QQR QWS EEQ AKI QQN QQQ IQ
rep6 QNP MMM QQR QWS EEQ AKI Q**H**N QQ**M A**
QQ**M** A**Q**
rep7 Q**GL** MM**T E**QR
QR QWS E**D**Q AKI QQ**A** QQ**M** A**Q**
rep8 Q**T**P MMM **P**Q
M QQR Q**W**T E**D**

Israel/*Fc2*

rep1 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep2 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep4 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep5 QNP MM**V** QQR QWS EEQ AKI QQN QQQ IQ
rep6 QNP MMM QQR QWS EEQ AKI Q**H**N QQ**M A**
QQ**M** A**Q**
rep7 Q**GL** MM**T E**QR
QR QWS E**D**Q AKI QQ**A** QQ**M** A**Q**
rep8 Q**T**P MMM **P**Q
M QQR Q**W**T E**D**

Ghanghry/*Fc2*

rep1 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep2 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep3 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep4 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep6 QNP MMM QQR QWS EEQ AKI Q**H**N QQ**M A**
QQ**M** A**Q**
rep7 Q**GL** MM**T E**QR
QR QWS E**D**Q AKI QQ**A** QQ**M** A**Q**
rep8 Q**T**P MMM **P**Q
M QQR Q**W**T E**D**

Shahrinau/*Fc3*

rep1 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep2 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep3 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep6 QNP M**V**M QQR QWS EEQ AKI Q**H**N QQ**M A**
QQ**M** A**Q**
rep7 Q**GL** MM**T E**QR
QR QWS E**D**Q AKI QQ**A** QQ**M** A**Q**
rep8 Q**T**P MMM **P**Q
M QQR Q**W**T E**D**

Dilizhan/*Fc4*

rep1 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep2 QNP MMM QQR QWS EEQ AKI QQN QQQ IQ
rep6 QNP MMM QQR QWS EEQ AKI Q**H**N QQ**M A**
QQ**M** A**Q**
rep7 Q**GL** MM**T E**QR
QR QWS E**D**Q AKI QQ**A** QQ**M** A**Q**
rep8 Q**T**P MMM **P**Q
M QQR Q**W**T E**D**

ever, for convenience we favor the idea that the length mutations are caused by deletion of one or more intact repeats (from repeats 1–5), as is indicated in Fig. 4. Otherwise, the repeated sequence of Israel/*Fc2* is identical with that of Samarkand/*Fc1*.

Wild-type strain Ghanghry also represents *Fc2* but sequence comparisons with Israel/*Fc2* reveal that Ghanghry most likely has evolved through a deletion of repeat 5. In addition, repeat 3 has a C → A exchange at position 25, which makes it identical with the consensus sequence. Wild-type Shahrinau/*Fc3* lacks repeats 4 and 5, i.e., 156 bp, and has an A → G exchange at position 13 in repeat 6.

The largest deletion is found in wild-type strain Dilizhan/*Fc4*. This deletion comprises repeats 3, 4, and 5, i.e., 234 bp. Furthermore, repeat 6 has an A → C exchange at position 25 and an A → T exchange at position 66, which is in the internal tandem repeat. (See above.)

Finally, it is interesting to note that repeats 1, 2, 7, and 8 are identical in the variants sequenced in this paper (Fig. 4).

Amino Acid Sequence of the Variable-Length Region

Figure 5 shows the deduced amino acid sequence of the repeated region in the variants. In Samarkand/*Fc1* the amino acid sequences of the first four repeats are identical (consensus sequence) whereas repeat 5 has an exchange at position 6, Met → Val (M → V), due to the A → G transition referred to above. Repeat 6 is identical with the consensus repeat through position 19; then, at position 20 there is a Glu → His exchange (Q → H), a Glu → Met (Q → M) exchange at position 24, and an Ile → Ala (I → A) substitution at position 25. The sequence QQMA (from 22 to 25) is repeated once (see above) making this repeat four amino acids longer; the same sequence is also present in repeat 7. Similarly, the internal tandem repeat QR at position 8–9 makes repeat 7 two amino acids longer. As mentioned above, both repeats 7 and 8 have many nucleotide exchanges, and several of these result in amino acid substitutions in the regions flanking the internal tandem repeats (Fig. 5). Altogether, there are 18 exchanges: 1 in the first 5 repeats and the first

Fig. 5. The amino acid sequence of the repeating motifs deduced from the sequences shown in Fig. 4. *rep1–rep8*, the eight different repeats. The sequence of either one of *rep1–4* may represent the consensus repeat sequence. Amino acids in *bold type* indicate exchanges relative to the consensus sequence. The internal tandem repeats are shown under each other to maintain alignment of the other repeats.

two-thirds of repeat 6 (149 amino acids) and 17 in the following 56 amino acids.

Apart from one or more deleted repeats the other variant forms are identical with Samarkand/*Fc1* except for Shahrinau/*Fc3*, which has a Met → Val (M → V) exchange at position 5 in repeat 6, and Dilizhan/*Fc4*, which has a Glu → His (Q → H) exchange at position 22 in repeat 6 (Fig. 5).

The central, repeated region of the *dec-1* products has an extraordinarily high content of glutamine and methionine, 39 and 16%, respectively (Waring et al., 1990). It is noteworthy that the repeating region we have sequenced here contains ≈40% glutamine and ≈13% methionine regardless of variant.

Discussion

The present study demonstrates that the length polymorphism in the *dec-1* gene found among wild-type strains is caused by differing copy numbers of a tandemly repeating sequence. Similar length polymorphism in coding DNA also occurs in several other internally repetitive genes; e.g., the *period* and salivary gland secretion genes in *Drosophila* (Costa et al., 1991; Muskavitch and Hogness 1982), the human ζ-globin gene complex (Goodbourn et al., 1983), the human proline-rich protein (PRP) genes (Lyons et al., 1988), and the human and gorilla involucrin gene (Teumer and Green 1989). It is interesting that the *dec-1* product, and in particular its repeated region, is similar to involucrin in having an extraordinarily high content of glutamine (Waring et al. 1990; Eckert and Green 1986). Involucrin is believed to play a central role in the cross-linking and stabilization of the kerotinocyte envelope (Eckert and Green 1986), and Waring et al. (1990) suggested that the *dec-1* protein(s), as involucrin, aligns and cross-links proteins within the eggshell. Cross-linking between proteins requires domains of amino acids with specific sequence and composition, and it is possible that the repeated region, with its extraordinarily high content of glutamine (in particular repeats 1–6 with 11 out of 26 amino acids per repeat being glutamine), is involved in these interactions. If so, it cannot be ruled out that there may be constraints on the maximum and minimum number of repeats present in the *dec-1* gene products. That not more than four variant forms (with maximum 5 and minimum 2 of the first 5 identical repeats present) were found among 130 wild-type strains may support this notion (Lineruth and Lindberg 1988). Furthermore, there is a precedent for an upper, functional size limit of the *dec-1* proteins in that the enlarged protein produced by the *dec-1* allele *fs(1)1501* result in female sterility (Lineruth and Lambertsson 1986). In this case the enlarged proteins are the result of an efficient utilization of an alternative splice site in the *fs(1)1501* mutant egg chambers (Waring et al. 1990).

Our data do not allow any conclusion as to the evolutionary relationships among the four variants analyzed in this study. However, several of our results indicate that unequal crossing-over and/or slippage during DNA replication (Smith 1976; Dover 1989) has played a significant role in the evolution of the locus. First, the extreme similarity among repeats 1–6 (unequal crossing-over maintains identical repeats; Smith 1976); second, the extensive variation in repeat number; third, of 130 wild-type strains, 23 were dimorphic and two were trimorphic for the *dec-1* locus (Lineruth 1987); fourth, our sequence analysis also shows (Fig. 4) that the two *Fc2* variants analyzed here were, in all probability, generated by deletion of two different repeats; and fifth, the same base substitution is found in repeats 2 and 3 in Samarkand/*Fc1* and Shahrinau/*Fc3*, and in repeats 2 and 6 in Dilizhan/*Fc4* as well as in all the internal tandem repeats. These exchanges may be fortuitous, but they may, on the other hand, support the notion that base exchanges spread to multiple copies of a tandem repeat because of intragenic unequal crossing-over (Smith 1974). These results also imply that while the frequency of unequal crossing-over is high in the first half of the repeated region, the lesser homology between repeats and the lack of allelic variation indicate that it occurs less frequently in the second half of the region. In the Hawaiian *Drosophila* the vast majority of length mutations in the chorion gene cluster are associated with tandem direct repeats generated by unequal crossing-over or slippage and mispairing of DNA strands during replication (Martìnez-Cruzado 1990).

The sequence of the first five repeats and the first two-thirds of repeat 6 is strikingly conserved among the five strains analyzed here. It can be argued that the selection of wild-type strains is biased toward strains from the (former) Soviet Union—four out of five. However, each of these strains represents a different variant form, and they were collected in four different regions of that gigantic country, at different times, and before 1980—e.g., Samarkand/*Fc1* as early as in 1936 in Uzbekistan and Shahrinau/*Fc3* in 1978 in Tadzhikistan (Lineruth and Lindberg 1988). Furthermore, a comparison with both the Israel/*Fc2* (collected in Israel 1970; this study) and the Canton S strains (Waring et al. 1990; this strain was selected by C. Bridges from wild-type flies collected in Ohio, USA, probably in the 1930s; Lindsley and Grell 1968) shows that the sequence of the first five repeats and the first two-thirds of repeat 6 is identical.

The *D. melanogaster dec-1* locus occurs in four protein variants in wild-type strains, *Fc1–Fc4*, with *Fc1* by far by the most common one (Lineruth and Lindberg 1988). The other variants are also found in natural populations in different parts of the world, and in the order of frequency *Fc2* > *Fc3* > *Fc4*. That *Fc1* is the most frequent and *Fc4* the least frequent variant form suggests that Fc1 may have a selection advantage over *Fc2–4* because of its higher number of repeats. Results from a recent selection experiment involving *Fc1* and *Fc3* suggest that this may be the case (Lineruth and Lindberg 1988). In two selection lines, A and B, the initial frequency of *Fc1* was 0.67 and 033, respectively; after 32 weeks (≈20 generations) the frequency of *Fc1* in lines A and B had increased to 0.82 and 0.53, respectively. This observation indicates that the *Fc1* form has a selective advantage that may be due to the presence of extra repeats. However, the two selection lines were initially set up from two wild-type strains, Samarkand/*Fc1* and Shahrinau/*Fc3*, and there may be other factors of selective value other than the number of repeats in *dec-1*. On the other hand, if the *Fc* alleles are selectively neutral their frequencies do not have to be equal. However, results from Ewer et al. (1990) revealed that in flies transformed with a *per* gene from which the Thr-Gly repeated region was deleted, the circadian period in locomotor activity became temperature dependent. The result suggests the Thr-Gly repeat may play a role in the thermostability of the *per* protein and thus have a selective value.

Further selection experiments coupled with PCR and sequence analysis of single flies to check for unequal crossing or slippage events will shed more light on the evolution of the *dec-1* eggshell gene in *Drosophila*. In addition, sequencing of the repeated region of several *Fc1* (or *Fc2*, *Fc3*, and *Fc4*) variants will reveal more about how frequent these events are. It will also be interesting to compare the repeated region from other species both in the *melanogaster* species subgroup and in more distantly related species—e.g., *D. pseudoobscura* (A. Zomorodipour, S. Andersson, A. Lambertsson, work in progress).

Acknowledgments. We thank Mr. Thore Johansson for expert technical assistance, Dr. Katrin Lineruth for sharing unpublished results, and Ms. Karin Ekström for providing the *Drosophila* stocks. S. Andersson was supported by the Sven and Lilly Lawski foundation. This work was supported by the Swedish Natural Science Research Council.

References

Andersson S, Lambertsson A (1991) Evolution of the *dec-1* locus in *Drosophila*. I. Restriction site mapping and limited sequence comparison in the *melanogaster* species subgroup. J Mol Evol 33:321–331

Bauer BJ, Waring GL (1987) 7C female sterile mutants fail to accumulate early eggshell proteins necessary for later chorion morphogenesis in *Drosophila*. Dev Biol 121:349–358

Costa R, Peixoto AA, Thackeray JR, Dalgleish R, Kyriacou CP (1991) Length polymorphism in the Threonine-Glycine-encoding repeat region of the *period* gene in *Drosophila*. J Mol Evol 32:238–246

Dover GA (1989) Slips, strings and species. Trends Genet 5:100–102

Eckert RL, Green H (1986) Structure and evolution of the human involucrin gene. Cell 46:583–589

Erlich HA (1992) Basic methodology. In: Erlich HA (ed) PCR technology: principles and applications for DNA amplification. WH Freeman, New York, pp 1–5

Ewer J, Hamblen-Coyle M, Rosbash M, Hall JC (1990) Requirement for *period* gene expression in the adult and not during development for locomotor activity rhythms of imaginal *Drosophila melanogaster*. J Neurogenet 7:31–73

Goodbourn SEY, Higgs DR, Clegg JB, Weatherall DJ (1983) Molecular basis of length polymorphism in the human ζ-globin complex. Proc Natl Acad Sci USA 80:5022–5026

Hansson L, Lambertsson A (1983) The role of su(f) gene function and ecdysterone in transcription of glue polypeptide mRNAs in *Drosophila* melanogaster. Mol Gen Genet 192:395–401

Hawley RJ, Waring GL (1988) Cloning and analysis of the *dec-1* female-sterile locus, a gene required for proper assembly of the *Drosophila* eggshell. Genes & Dev 2:341–349

Jowett T (1986) Preparation of nucleic acids. In: Roberts DB (ed) *Drosophila*. A practical approach. IRL Press, Oxford, pp 275–286

Komitopoulou K, Margaritis LH, Kafatos FC (1988) Structural and biochemical studies on four sex-linked chorion mutants of *Drosophila melanogaster*. Dev Biol 9:37–48

Lambertsson A, Andersson S, Johansson T (1989) Cloning and characterization of variable-sized *gypsy* mobile elements in *Drosophila melanogaster*. Plasmid 22:22–31

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Lindsley DL, Grell EH (1968) Genetic variations of *Drosophila melanogaster*. Carnegie Inst Washington Publ, p 627

Lineruth K (1987) An ovarian protein locus and its developmentally regulated expression in *Drosophila*. Doctoral thesis, University of Umeå, Sweden

Lineruth K, Lambertsson A (1985) Stage specific synthesis of some follicle cell proteins in *Drosophila melanogaster*. Wilhelm Roux's Arch Dev Biol 194:436–439

Lineruth K, Lambertsson A (1986) Correlation between a female sterile mutation and a set of follicle cell proteins in *Drosophila melanogaster*. Mol Gen Genet 205:213–216

Lineruth K, Lambertsson A, Lindberg M (1985) Genetic localization of a follicle cell protein locus in *Drosophila melanogaster*. Mol Gen Genet 201:375–378

Lineruth K, Lindberg M (1988) Electrophoretic variant forms of a set of follicle cell proteins in *Drosophila melanogaster:* frequencies in wild-type strains and response in a selection experiment. Hereditas 108:59–64

Lyons KM, Stein JH, Smithies O (1988) Length polymorphism in human proline-rich protein genes generated by intragenic unequal crossing-over. Genetics 120:267–278

Maniatis T, Fritsch EF, Sambrook J (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp 368–369

Martìnez-Cruzado JC (1990) Evolution of the autosomal chorion cluster in *Drosophila*. IV. The Hawaiian *Drosophila:* Rapid protein evolution and constancy in the rate of DNA divergence. J Mol Evol 31:402–423

Muskavitch MAT, Hogness DS (1982) An expandable gene that

encodes a *Drosophila* glue protein is not expressed in variants lacking remote upstream sequences. Cell 29:1041–1051

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467

Smith GP (1974) Unequal crossover and the evolution of multigene families. Cold Spring Harbor Symp Quant Biol 38:507–513

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Swallow DM, Gendler S, Taylor-Papadimitriou J, Bramwell ME (1987) The human tumor-associated epithelial mucins are coded by an expressed hypervariable gene locus. Nature 328:82–84

Teumer J, Green H (1989) Divergent evolution of part of the involucrin gene in the hominoids: unique intragenic duplications in the gorilla and human. Proc Natl Acad Sci USA 86:1283–1286

Waring GL, Hawley RJ, Schoenfeld T (1990) Multiple proteins are produced from the *dec-1* eggshell gene in *Drosophila* by alternative RNA splicing and proteolytic cleavage events. Dev Biol 142:1–12

Received June 19, 1992/Accepted December 4, 1992