The *Drosophila* Ribosomal Protein S6 Gene Includes a 3′ Triplication That Arose by Unequal Crossing-Over[1]

Mary J. Stewart[2] *and Rob Denell*

Center for Basic Cancer Research, Division of Biology, Kansas State University

Ribosomal protein S6 (rpS6) is the major phosphoprotein of the small ribosomal subunit of eukaryotes and is phosphorylated in response to treatment with mitogens and other stimuli. We have examined the organization of the *rpS6* gene of *Drosophila melanogaster*. Comparisons of a cDNA with genomic DNA identify a transcription unit including three exons. Two tandem repeats downstream of this transcription unit reiterate divergent copies of the third exon and flanking regions. Comparisons of these three repeats with respect to nucleotide base substitutions and deletions or insertions show clearly that they arose via a duplication and subsequent crossing-over between misaligned copies. Although no direct evidence exists that the downstream exons are transcribed, the maintenance of open reading frames in spite of extensive genetic changes is consistent with a protein-coding function.

Introduction

The importance of duplications in genome evolution has been widely emphasized (e.g., see Ohno 1970). Examples of the duplication of entire genes, often followed by divergence of function, have been extensively documented (Li and Graur 1991, pp. 143–147). In addition, duplication of internal portions of genes has been a frequent evolutionary event (Li and Graur 1991, pp. 140–143; States and Boguski 1991) and can result in proteins with repeated functional domains. In work presented elsewhere (Stewart and Denell 1993), we have shown that two transposon-induced mutations causing a loss of growth control of the *Drosophila* larval hematopoietic organs affect the gene encoding ribosomal protein S6 (rpS6). This protein is of interest because it is the major phosphoprotein of the small ribosomal subunit, and its phosphorylation in response to treatment by mitogens and other stimuli has been well characterized (Erikson 1991; Sturgill and Wu 1991). Here we present the sequence of the rpS6 region and show that there is an unusual tandem triplication of its 3′-most exon. These copies show considerable divergence, and sequence analysis strongly supports the view that they arose by unequal crossing-over.

Material and Methods

Nucleic Acid Preparations and Sequencing

Plasmid DNA was isolated by alkaline lysis (Maniatis et al. 1982) or with a Magic Mini Prep kit (Promega), by following the manufacturer's instructions. Lambda DNA was isolated according to the method of Maniatis et al. (1982, pp. 366–367). Fragments

1. Key words: ribosomal protein S6, exon duplication, unequal crossing-over, *Drosophila*.

2. Friedrich Miescher Institüt, Post Office Box 2543, CH-4002 Basel, Switzerland.

Address for correspondence and reprints: Rob Denell, Division of Biology, Ackert Hall, Kansas State University, Manhattan, Kansas 66506.

Mol. Biol. Evol. 10(5):1041–1047. 1993.

0737-4038/93/1005-0008$02.00

from genomic clones λ57 and λ64 (provided by Stephan Andersson and Dr. Andrew Lambertsson, Umea, Sweden) and an rpS6 cDNA (see Stewart and Denell 1993) were subcloned into pGEM-7Zf(+) or pGEM-3Z (Promega). Nested exonuclease III deletions of plasmid DNA were generated by using the Erase-a-Base system (Promega). Sequence determination of double-stranded plasmid DNA was performed by the dideoxynucleotide chain-termination method (Sanger et al. 1977) using the vector primers and Bst (BioRad), TaqTrack (Promega), or fmol (Promega) sequencing systems. For genomic and cDNA clones, both strands were sequenced, and the resulting data were submitted to the GenBank data base under accession numbers L0274 and L0275, respectively.

Sequence Data Manipulation

Nucleic acids were analyzed by using the SEQAIDII™ software package, available through BIONET, a National Computer Resource for Molecular Biology; some sequence alignments were adjusted manually. Predicted amino acid sequences were compared by using the FASTA program (Pearson and Lipman 1988).

Results

As described elsewhere (Stewart and Denell 1993), our studies of recessive lethal mutations resulting in hypertrophy of the larval hematopoietic organs and aberrancies of the immune response resulted in the molecular cloning and characterization of the *Drosophila* homologue encoding rpS6 (*rpS6*). The organization of the *rpS6* gene is shown schematically in figure 1, and the DNA sequence on which this interpretation is based is presented in figure 2. Comparison of an embryonic cDNA and genomic DNA defines a transcription unit including three exons, denoted "E1," "E2," and "E3A." Primer extension experiments show that this cDNA is full length and that two alternative transcriptional start sites just upstream of its 5′ terminus are used as well (Stewart and Denell 1993).

Southern blotting experiments (Stewart and Denell 1993) and additional sequencing indicated that the *rpS6* gene is in single copy but that the 3′-most portion of this transcription unit (region A in fig. 1) is tandemly repeated two additional times (copies B and C). The triplicated region includes the third exon and flanking regions (intron 5′ and nontranscribed region 3′). The downstream putative exons, referred to as "E3B" and "E3C," are identical to each other in protein-coding capacity and are diverged with respect to E3A. As discussed below, we will proceed as if E3A, E3B, and E3C are present in processed mRNA as alternative 3′-most exons, although direct evidence for this premise is presently lacking.

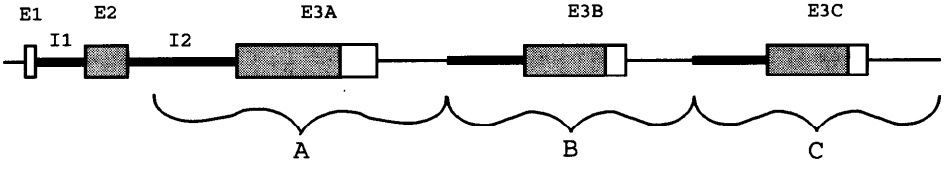


FIG. 1.—Map of the organization of a known *rpS6* transcription unit and downstream repeats with strong sequence similarity. A cDNA defines a transcription unit including three exons (E1, E2, and E3A) and two introns (I1 and I2). The indicated region is repeated two times, with each repeat including a putative alternative third exon (E3B and E3C). The protein-coding portions of known or suspected exons are shaded; E1 includes two codons that are not apparent at this scale.

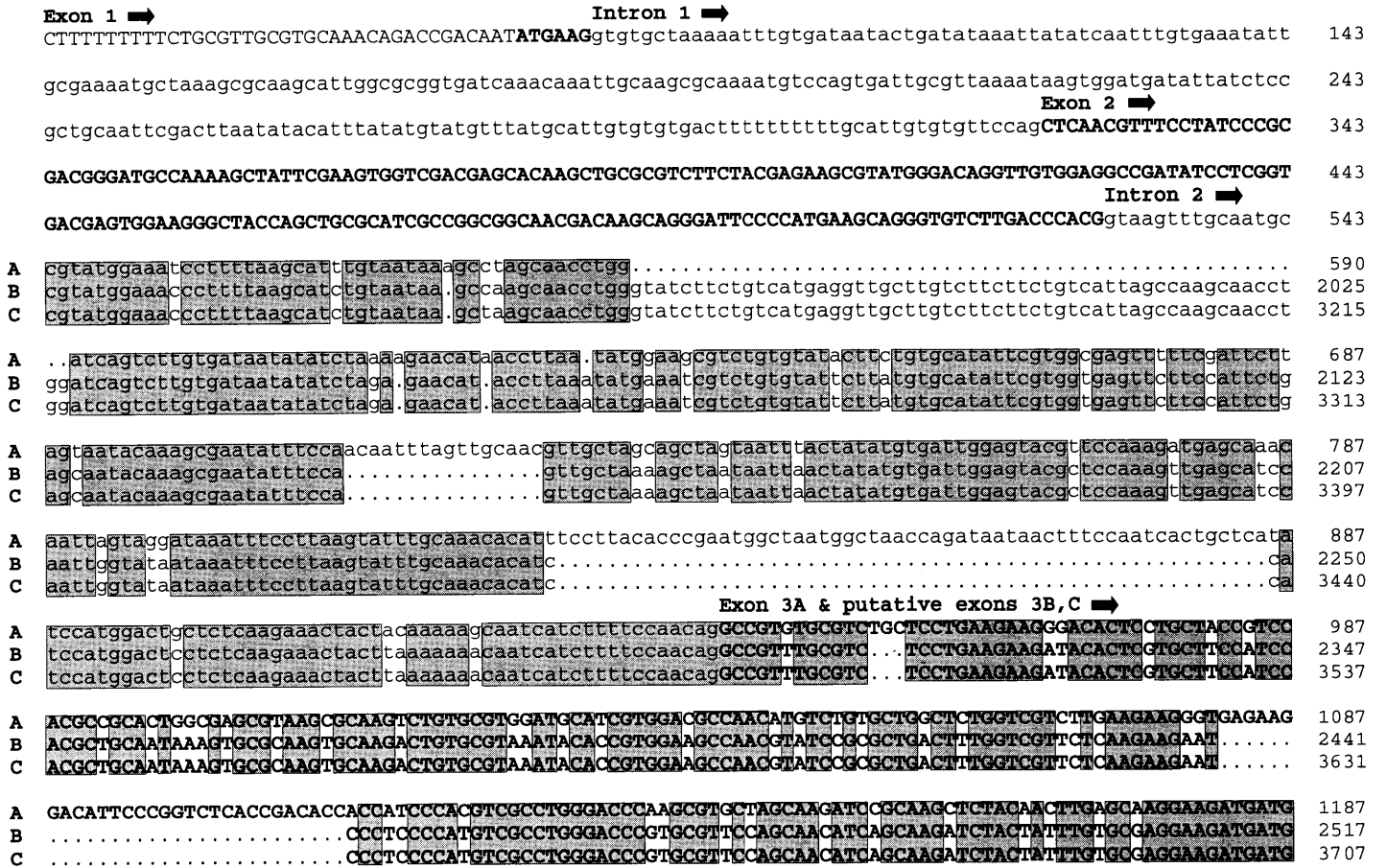
Figure 2 has been organized to align the three regions showing sequence similarity. Although nucleotide base substitutions have occurred, insertions/deletions (indels) represent the most dramatic differences. Immediately upstream of the alternative third exons, 50 of 54 bases are identical for all three copies, including a conventional splice-acceptor site and a variant branch-point signal (Rice et al. 1991). Each of the alternative exons shares an identical polyadenylation signal with a very good match to the consensus sequence (Rice et al. 1991). Slightly upstream, exon 3A has a second putative signal, which is less consistent with respect to the consensus; this region is not present in the two alternative exons. Although many gaps are necessary to align all three sequences in the 3′ flanking region, considerable sequence similarity is still apparent.

Discussion

Unequal crossing-over has frequently been invoked as a mechanism for expanding and contracting the number of copies of duplicated genes or portions of genes. This mechanism, presented schematically in figure 3, was first proposed on the basis of genetic and cytogenetic studies of the *Bar* gene of *Drosophila* (Sturtevant 1925; Bridges 1936). If a tandem duplication is first generated, subsequent offset synapsis involving the duplicated regions and a precise exchange generate a triplication and an ancestral chromosome as reciprocal events. If intervening time is sufficiently long to allow the duplicated copies to diverge, then the central copy of the triplication will be a hybrid with respect to these homologous but dissimilar copies; that is, the point of exchange will separate regions derived from one or the other of the duplicated copies.

The structure of the *rpS6* gene argues strongly that it arose in a manner similar to that diagramed in figure 3; that is, it appears that a duplication of an ancestral third exon and flanking regions first occurred and that these copies then diverged. Subsequently, a triplication was generated by offset pairing and an exchange in the region near the end of the transcription unit. This model predicts that copy B will resemble copy C to the left of the point of exchange but that it will be similar to copy A to the right of the point of exchange. The alignment shown in figure 2 requires the introduction of nine gaps >1 bp to the left of an interval including the common polyadenylation signal and immediate downstream region. In all cases, copies B and C resemble each other and are dissimilar to copy A. Likewise, with respect to the five gaps that must be introduced between this interval and the 3′ end of the triplicated regions, copies A and B resemble one another in all cases. The same strong trend is apparent with respect to base-pair substitutions and indels of single base pairs. Upstream of the hypothesized point of exchange copy B resembles copy C in ~90% of such differences, whereas, downstream of this interval copy B resembles copy A in >80% of such examples.

What are the implications of these results for the possible function of the downstream alternative exons? Despite the presence of all motifs necessary for these exons to be processed into mature transcripts, we have yet to detect such mRNAs from staged whole-organism homogenates by northern blotting experiments. Nevertheless, aspects of the structure and sequence of putative exons 3B and 3C suggest that a protein-coding function has constrained their evolution. The alignment presented in figure 2 requires five gaps, presumably reflecting indels during the time that the original duplicate regions were diverging. Figure 4 shows the amino acid sequences predicted by the known cDNA including exon 3A and that which would be generated by putative alternative exons 3B and 3C, both compared with the predicted rat rpS6 protein. The alignment of *Drosophila* amino acids is based directly on that of the nucleic acid


Exon 1
Intron 1
Exon 2
Intron 2
Exon 3A & putative exons 3B,C
A
B
C

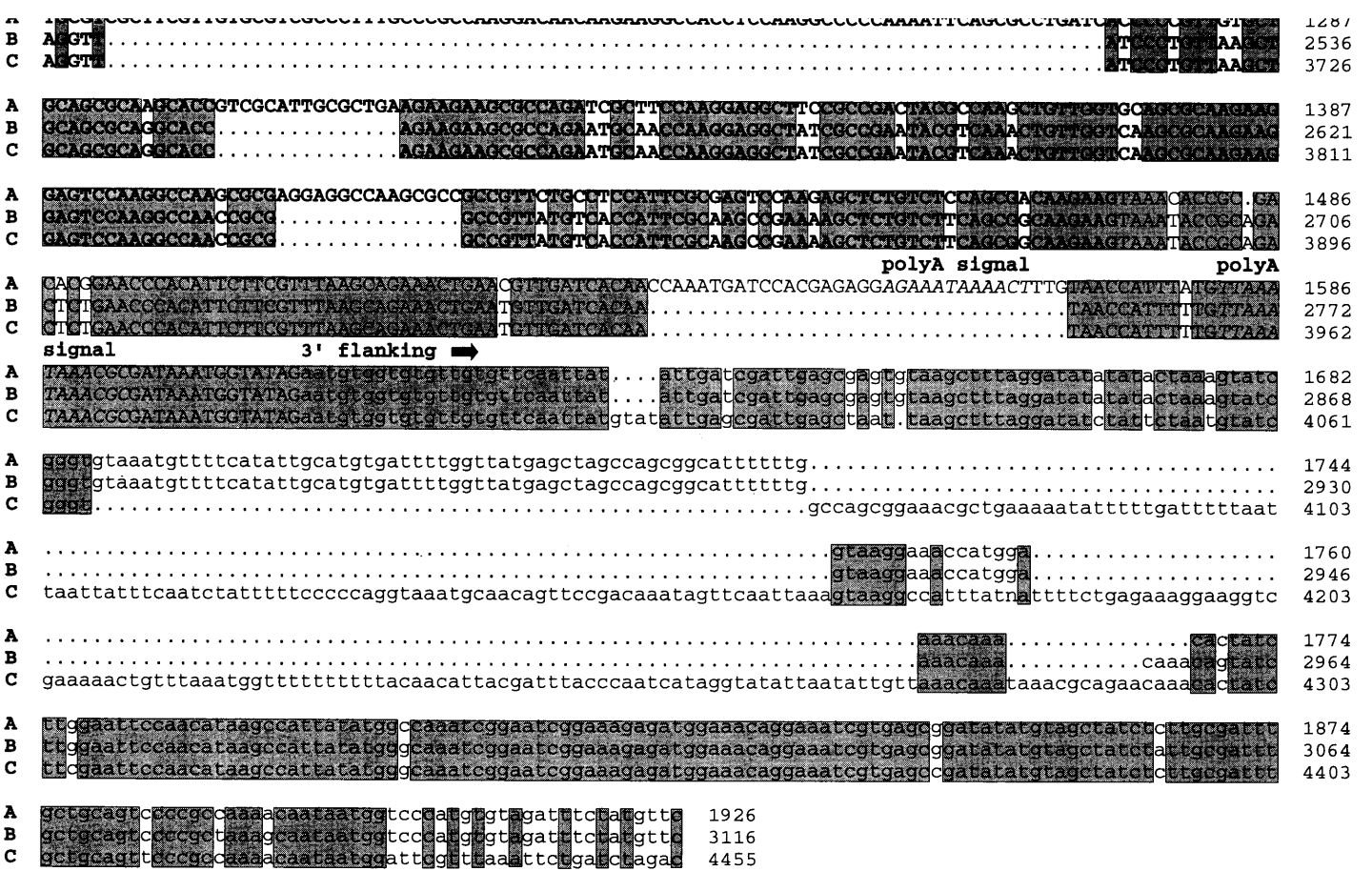


FIG. 2.—DNA sequence of the region depicted in fig. 1. The sequence begins with the first nucleotide of a cDNA, which is numbered 44 because an alternative upstream transcription start site is known to be used as well (Stewart and Denell 1993). The downstream repeats (B and C) have been aligned with A. Sequences of known or putative exons are presented in uppercase letters, and protein-coding regions are in boldface type; sequences of introns and flanking, nontranscribed regions are in lowercase letters. Gaps are indicated by periods. Putative signals for polyadenylation are indicated, and nucleotide bases common to all three repeats are boxed and shaded.

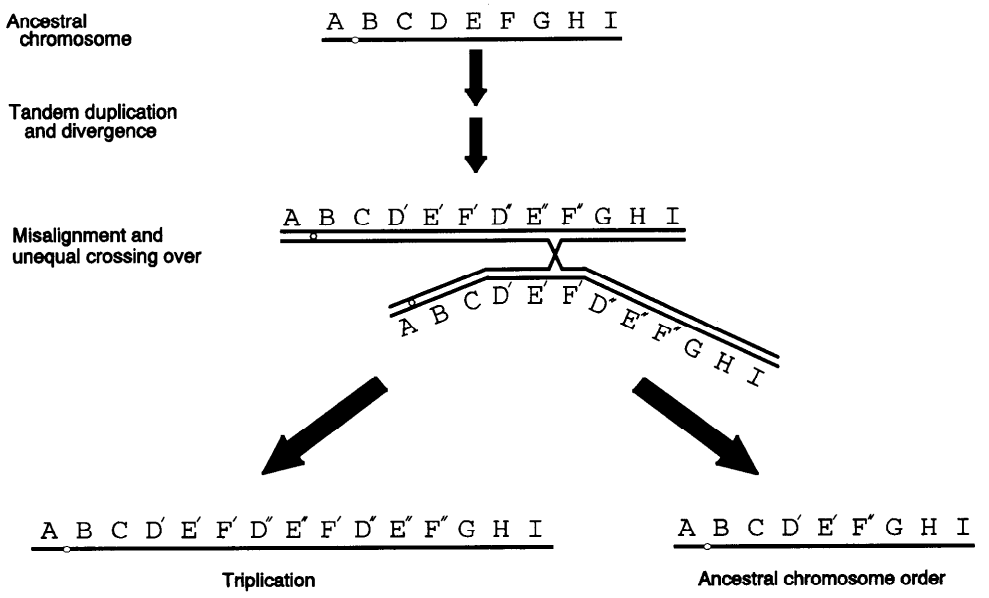


FIG. 3.—Schematic representation showing how a triplication can arise from a previous duplication by unequal crossing-over. If the duplicated copies have diverged from one another, the central copy of the triplication is expected to include portions similar to each, as diagramed. A chromosome with the ancestral organization is generated as a reciprocal product.

sequences given in figure 2. In each of the five cases where indels have occurred, the amino acid sequence predicted by exon 3A resembles the rat sequence, implying that it is ancestral. Thus, the observation that these indels occurred in the downstream copy of the original duplication, without introducing a frameshift, suggests that this repeat includes a functional exon. This interpretation is also consistent with the observation that exons 3A and 3B/C differ by base-pair substitutions at 76 positions,

```
Rat          MKLNISFPATGCQKLIEVDDERKLRTFYEKRMATEVAADALGEEWKGYVVRISGGNDKQG
D. mel       MKLNVSYPATGCQKLFEVVDEHKLRVFYEKRMGQVVEADILGDEWKGYQLRIAGGNDKQG

Rat          FPMKQGVLTHGRVRLLLSKGHSCYRPKRTGERKRKSVRGCIVDANLSVLNLVIVKKGEKD
D. mel       FPMKQGVLTHGRVRLLLKKGHSCYRPRRTGERKRKSVRGCIVDANMSVLALVVLKKGEKD
D. mel B/C                RLR..LLKKIHSCFHPRCNKVRKCKIVRKYTVEANVSALTLVVLKKN...

Rat          IPGLTDTTVPRRLGPKEASRIRKLFNLSKEDDVRQYVVRKPL.NKEGKKPRTKAPKIQRL
D. mel       IPGLTDTTIPRRLGPKEASKIRKLYNLSKEDDVRRFVVRRPLPAKDNKKATSKAPKIQRL
D. mel B/C   .......PSPCRLGPVESSNLSKIYYLCEEDDEVI.....................

Rat          VTPRVLQHKRRRIALKKQRTKKNKEEAAEYAKLLAKRMKEAKEKRQEQIAKRRLSSLRA
D. mel       ITPVVLQRKHRRIALKKKRQIASKGASADYAKLLVQRKKESKAKEE..AKRRGSASIRE
D. mel B/C   ..PVKLQRRHQ.....KKRQNAIKEAIAEYVKLLVKRKKESKANRG.......EYVTIRK

Rat          STSKSESSQK
D. mel       SKSSVSSDKK
D. mel B/C   PKSSVFSGKK
```

FIG. 4.—Predicted amino acid sequence of the rat rpS6 protein (Chan and Wool 1987) and that predicted by the known *Drosophila* cDNA (including exon 3A), compared with one another and with the protein potentially encoded by putative alternative exons 3B and 3C. The *Drosophila* amino acid sequences are aligned on the basis of the nucleotide base alignment presented in fig. 2; gaps are indicated by periods, and residues common to all three sequences are boxed and shaded.