J Mol Evol (1993) 37:483–495

Journal of Molecular Evolution

# Homopolymer Length Variation in the *Drosophila* Gene *mastermind*

**Stuart J. Newfeld,* Aloisia T. Schmid, Barry Yedvobnick**

Department of Biology, 1510 Clifton Road, Emory University, Atlanta, GA 30322, USA

Received: 18 December 1992 / Accepted: 19 March 1993

**Abstract.** Runs of identical amino acids encoded by triplet repeats (homopolymers) are components of numerous proteins, yet their role is poorly understood. Large numbers of homopolymers are present in the *Drosophila melanogaster mastermind* (*mam*) protein surrounding several unique charged amino acid clusters. Comparison of *mam* sequences from *D. virilis* and *D. melanogaster* reveals a high level of amino acid conservation in the charged clusters. In contrast, significant divergence is found in repetitive regions resulting from numerous amino acid replacements and large insertions and deletions. It appears that repetitive regions are under less selective pressure than unique regions, consistent with the idea that homopolymers act as flexible spacers separating functional domains in proteins. Notwithstanding extensive length variation in intervening homopolymers, there is extreme conservation of the amino acid spacing of specific charge clusters. The results support a model where homopolymer length variability is constrained by natural selection.

**Key words:** *Drosophila* — *mastermind* — Gene comparison — Triplet repeat — Homopolymer — Protein evolution — Repeat length variation

The locus *mastermind* (*mam*) of *Drosophila melanogaster* comprises a component of a developmental pathway that mediates intercellular communication during several stages of the life cycle. *mam* encodes a nuclear protein containing an abundance of amino acid homopolymers and several unique charge clusters (Yedvobnick et al. 1988; Smoller et al. 1990). The arrangement of basic and acidic charge clusters in *mam* resembles a number of regulatory proteins that contain functional charged areas involved in DNA binding and transcriptional activation (Brendel and Karlin 1989). For example, basic charge clusters are found in GAL4, a yeast transcription factor (Fischer et al. 1988) and in the *Drosophila* DNA-binding protein *zeste* (Chen et al. 1992). In addition, the acidic charge cluster at the carboxy terminus of *mam* matches a consensus derived from transcriptional activation domains (Zhu et al. 1990). However, the role of such regions in *mam* has not been established. Comparison of *mam* to sequences deposited with the databases reveals no significant similarities in nonrepetitive regions, preventing the identification of functional domains.

Trinucleotide repeats occur in the translated regions of many genes encoding runs of amino acid homopolymers (Wharton et al. 1985; Duboule et al. 1987) yet their function remains unclear. Interspecific comparisons of homopolymer-containing genes of *Drosophila* (Kassis et al. 1986; Treier et al. 1989; Heberlein and Rubin 1990; Jones et al. 1991; Peixoto et al. 1992) as well as mammalian genes (Peterson et al. 1990; Danielson et al. 1986; Tseng and Green 1988) have demonstrated that over time these sequences are expanded or deleted in-frame, leading to repeat length variation. Studies of length polymorphism in repetitive regions from *D. melanogaster* indicate that these sequences are changing rapidly (Tautz 1989; Costa et al. 1991). The length

*Present address: Department of Cellular and Developmental Biology, 16 Divinity Ave., Harvard University, Cambridge, MA 02138, USA
*Correspondence to:* B. Yedvobnick

variability of homopolymers suggests that they act as a flexible connection in proteins, separating functional regions (Beachy et al. 1985). Nucleotide misalignment in the trinucleotide repeats that encode homopolymers could result in unequal crossover or slippage during replication, producing length variation (Treier et al. 1989), modifying the spacing of functional regions.

It has been postulated that length variation in a homopolymer, particularly within a regulatory gene, could result in altered protein activity and lead to a phenotypic change (Laughon et al. 1985). Recently, significant polymorphism and excessive amplification of trinucleotide repeats have been associated with the human inherited diseases X-linked spinal and bulbar muscular atrophy (Kennedy disease; LaSpada et al. 1991), myotonic dystrophy (Harley et al. 1992), and Fragile-X syndrome (Fu et al. 1991). Kennedy disease results from expansion of a glutamine homopolymer in the androgen receptor, consistent with ideas that such alterations can effect a phenotypic change.

An interspecific comparison of *mam* should help identify important functional residues. Further, the unusual concentration of homopolymers in *mam* suggests that an interspecific comparison could provide insight into their role and evolutionary instability. A preliminary sequence comparison between *D. melanogaster* and *D. virilis* (estimated divergence 60 million years; Beverley and Wilson 1984) demonstrated that unique and repetitive areas of *mam* are undergoing distinct patterns of evolutionary change and that one of the acidic clusters is highly conserved (Newfeld et al. 1991). Here we compare the complete amino acid sequence of *mam* from these species and also characterize the embryonic expression of *D. virilis mam.*

## Materials and Methods

*Genomic Analysis.* DNA purification, cloning, Southern blot analysis, genomic library construction, plaque hybridization, and chromosome walking were performed as previously described (Yedvobnick et al. 1988; Smoller et al. 1990). Unique subclones from *D. melanogaster mam* cDNA B4 (Smoller et al. 1990) were constructed for use as probes. B4J4 encodes acid domain 1 (exon 4). DM115 encodes most of acid domain 2 (exon 7). B1K spans exons 6 and 7. DRBPCR is a unique polymerase chain reaction product amplified from the basic domain (exon 3; provided by D. Bettler). The location of these probes in *D. melanogaster mam* can be visualized from Fig. 1A since the exon organization of *D. virilis mam* is identical in translated regions. A *D. virilis* (Bowling Green #15010-1051.0) *Sau*3A partial genomic library was constructed in lambda EMBL3 and propagated in the *recD* host TAP 90 (Patterson and Dean 1987). A screen (200,000 phage) with B4J4 was washed at standard stringency (0.1 × SSC, 0.1% SDS at 50°C) and a positive phage was identified. Phage 58 was restriction mapped and the smallest cross-hybridizing fragment was sequenced. This sequence is homologous to *D. melanogaster mam* exon 4 as reported in Newfeld et al. (1991). A chromosome walk in both directions, involving an additional 400,000 phage, using high-stringency washes at 70°C, identified 28 phage encompassing 97 kb of contiguous genomic DNA. DRBPCR, DM115, and B1K were hybridized, under standard conditions, to filters containing these phage to locate other *D. melanogaster mam* protein-coding exons (exons 3, 6, and 7). Exon 5, for which no cross-hybridizing region could be identified, was located in *D. virilis* by sequencing a 2 kb region upstream of exon 6. Exon numbers for *D. virilis mam* correspond to the number of the homologous exon in *D. melanogaster* (Smoller et al. 1990).

*RNA Localization.* Embryo preparation and hybridization with digoxigenin-labeled RNA probes transcribed from B4J4 were performed as described in Bettler et al. (1991).

*Characterization of Bacteriophage P1 Clones.* DVBasic is a unique subclone encoding the basic domain from exon 3 of *D. virilis mam.* This subclone was used to screen a bacteriophage P1 library (Smoller et al. 1991) containing *D. virilis* genomic DNA. Positive clones were amplified and DNA isolated according to D. Smoller (personal communication). Subsequent digestion, transfer, and hybridization with DVBasic, B1K, and several restriction fragments from *D. virilis mam* genomic phage provided a rough estimate of the extent of these clones. To determine the exact size and ascertain the integrity of the P1 clones, each was labeled and hybridized to filters containing phage from the *D. virilis mam* chromosome walk. In situ hybridization of DNA from the P1 clones to *D. virilis* polytene salivary gland chromosomes was completed as described in Smoller et al. (1991).

*DNA Sequencing Strategy.* Exon 4 was sequenced as described previously (Newfeld et al. 1991). Dideoxy sequencing reactions utilizing the 7-deaza dGTP kit with Sequenase 2.0 (U.S.B., Cleveland) provided the remaining *D. virilis* nucleotide sequence. Sequence data derived from single-strand (Dente et al. 1985) or double-strand (Chen and Seeburg 1985) reactions on templates cloned in Bluescript using the *E. coli* strain *Sure* (Stratagene) as a host. Reactions with the M13 Universal or Reverse primers or synthetic oligonucleotides (Emory Microchemical Facility) were completed as recommended with several modifications. At least 5 μg of plasmid DNA and threefold-more Sequenase were utilized and the termination step was completed at 45°C. The nucleotide sequence of both strands of genomic DNA for the protein-coding region of *D. virilis mam* (exons 3–7) was obtained. In addition, both strands of DNA sequence were obtained for (1) exon boundaries in the protein-coding region, (2) the complete introns between exons 5 and 6 and between exons 6 and 7, (3) the region immediately upstream of the translation initiation site, and (4) 200 bp upstream of the exon 4 splice acceptor. One strand of genomic DNA sequence was obtained from (1) the remaining 5′ untranslated region of exon 3, (2) the 3′ untranslated region of exon 7, and (3) roughly 2 kb of intron upstream of exon 5. 5,650 bp of transcribed DNA sequence (4,968 bp of coding region plus 360 bp of 5′ untranslated and 322 bp of 3′ untranslated sequence) was obtained.

*DNA Sequence Analysis.* Exon 4 sequence was compiled as described previously (Newfeld et al. 1991). All other nucleotide sequences were compiled using XTreePro (Executive Systems) and GenePro (Riverside Scientific). The analyses of *mam* amino acid composition and codon usage were completed using data supplied by M. Ashburner, formatted according to Grantham et al. (1981). Sequences were aligned by inspection using MASE (Faulkner and Jurka 1988). First, maximum identity between the

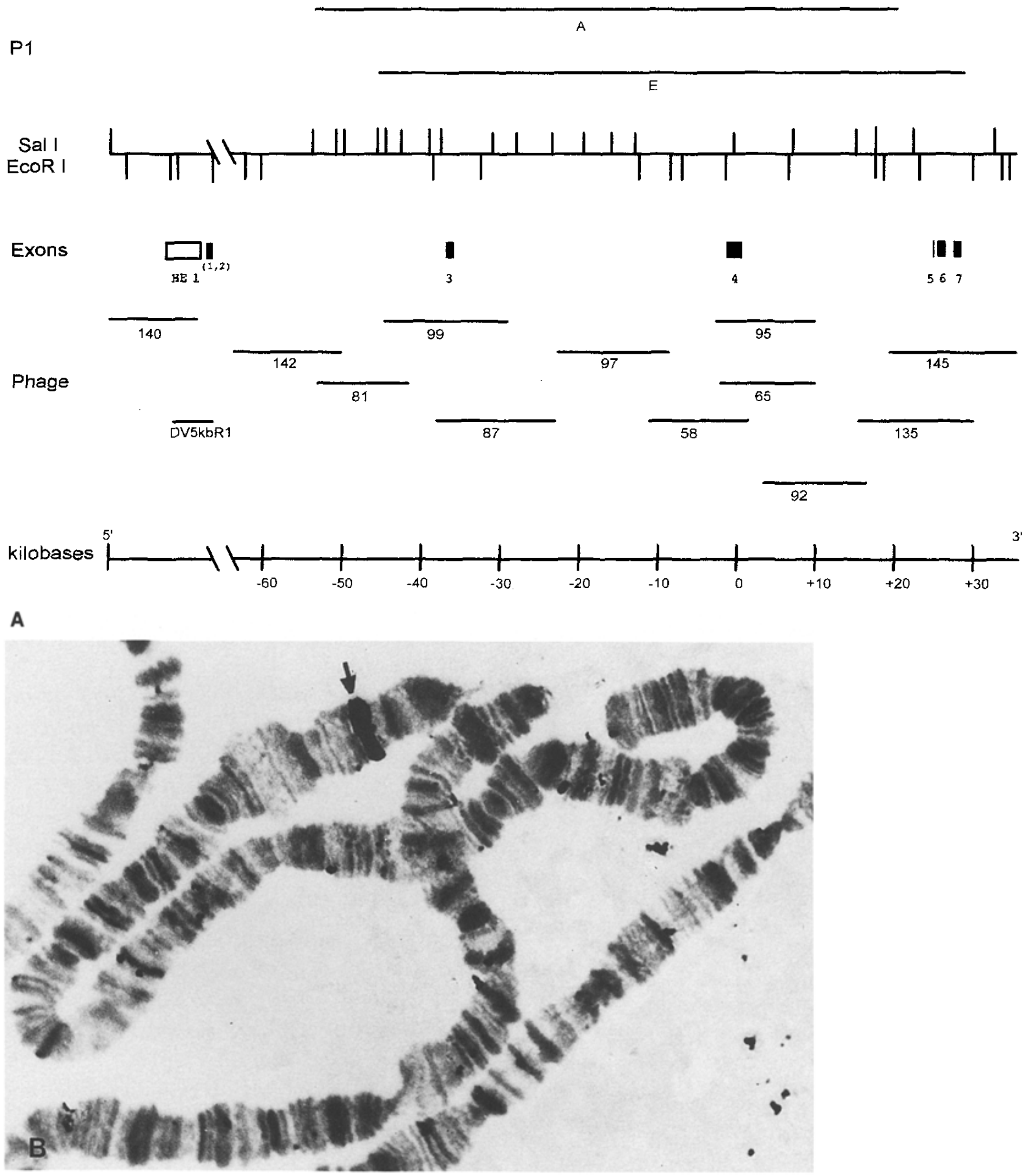


**Fig. 1.** Structural map of *D. virilis mastermind*. **A** The *horizontal line* depicts approximately 115 kb from the *D. virilis mam* chromosomal region. *Vertical lines* below the horizontal line indicate *Eco*RI restriction sites and above the horizontal line *Sal*I sites. The *coordinate scale* below the restriction map is indicated in kilobases and the zero coordinate marks the position of exon 4, following the numbering of *D. melanogaster* exons in Smoller et al. (1990). The position of bacteriophage P1 clones is depicted above the restriction map and the position of lambda EMBL3 clones is depicted below the restriction map. The location of *mam* exons is indicated by *black boxes* and corrects the exon locations reported in Newfeld et al. (1991). The locations of the 5-kb and 0.9-kb restriction fragments which cross-hybridized to cDNA HE1 of *D. melanogaster* are indicated by an *open box*. The 5-kb *Eco*RI fragment is contained in lambda gt10 clone DV5kbR1. DV5kbR1 is derived from a *D. virilis* subgenomic library constructed according to the pattern of cross-hybridization seen for HE1 on genomic Southerns (data not shown). HE1 derives from a gene whose transcription is divergent from *mam* and which begins within 100 bp of the most proximal *mam* transcription start site (Smoller et al. 1990). Friedel (1990) demonstrated that the restriction fragment in *D. melanogaster* which corresponds to DV5kbR1 contains the 5′ end of HE1 and the 5′ noncoding exons of *mam*. In order to identify the 5′ noncoding exons from *D. virilis mam*, which were predicted to exist by analogy to *D. melanogaster* (exons 1 and 2), but cannot be detected by cross-hybridization on a genomic Southern blot, a cDNA library was screened by polymerase chain reaction (PCR). The cDNA library was constructed in lambda gt10 from polyA$^+$ RNA isolated from *D. virilis* embryos. The library was screened in PCR reactions containing a 5′ oriented primer complementary to *mam* exon 3 and primers specific for each of the phage arms. A PCR product was identified that is likely to contain the 5′ noncoding exons of *D. virilis mam*. The PCR product hybridized at high stringency to a probe from *D. virilis mam* exon 3 and the insert of DV5kbR1 (predicted to contain *mam* 5′ noncoding exons by analogy to *D. melanogaster*). The exact distance between the cDNA HE1 cross-hybridizing region and the location of the 5′ noncoding exons of *mam* in DV5kbR1 has not been determined. The existence of two 5′ noncoding exons in *D. virilis* (indicated in *parentheses* in the figure) has not been determined. Numerous attempts to subclone DV5kbR1 and the PCR product into several plasmid vectors, utilizing a variety of recombination-deficient hosts, resulted in rearrangements. The distance between the nonoverlapping phage DV5kbR1 and 142 is not known. **B** In situ hybridization of P1 clone E to polytene band 59D of *D. virilis* chromosome 5.

amino acid sequences was obtained without regard to the number or placement of gaps in either species. Then this alignment was modified so required gaps were organized in the most parsimonious manner. This minimized the number of mutations required to explain all differences between the sequences. For example, a single large gap in a homopolymer was preferred to several smaller intermittent gaps even though a slight reduction in maximum identity is incurred. Nucleotide differences and their codon positions were identified using the find-diffs program of the PIMA software package (Smith and Smith 1992). Conservative amino acid substitutions were scored on biochemical similarity: D/E, K/R/H, N/Q, S/T, I/L/V, F/W/Y, and A/G (Smith and Smith 1990). Repetitive regions are defined as homopolymers (five or more consecutive identical amino acids) or areas where a single amino acid represents at least 50% of the residues (e.g., *D. virilis* 1336–1369; 70% glycine). In addition, the extent of the underlying triplet repeat was occasionally used in specifying the boundaries of repetitive regions. The repetitive regions are listed in the legend to Fig. 3.

## Results

### *Genomic Analysis of* D. virilis mam

The protein-coding exons of *D. virilis mam* are displayed within the chromosome walk in Fig. 1A. The *D. virilis* and *D. melanogaster mam* loci are identical in exon organization in translated regions (Smoller et al. 1990); the positions of *D. virilis* 5′ noncoding exons have not been precisely determined. Initial characterization of *D. virilis* genomic P1 clones demonstrated hybridization of the exon 3–specific clone (DVBasic) and the exon 6/exon 7–specific clone (B1K) to P1 clone E (P1E). The B1K homologous region in *D. virilis mam* is approximately 65 kb downstream of the location of DVBasic. Subsequently P1E was labeled and hybridized to filters containing phage derived from the chromosome walk through *D. virilis mam*. Comparison of the pattern of hybridization to the pattern of restriction fragments visible in an ethidium-bromide-stained agarose gel (data not shown) indicates that all restriction fragments between phage 81 and 145 hybridized to the probe. This suggests that P1E contains no large internal deletions and represents roughly 75 kb from *D. virilis mam* including the entire coding region. Similar analyses demonstrated that the genomic material in P1 clone A (P1A) is shifted upstream roughly 10 kb, so as not to include exons 5, 6, and 7. Together clones P1A and P1E span 85 kb from *D. virilis mam*. In situ hybridization revealed that P1A and P1E are located in band 59D (chromosome 5) of *D. virilis*, as shown in Fig. 1B for clone P1E.

### *Spatial Accumulation of* D. virilis mam *Transcripts*

A nonrepetitive probe was used previously to identify *D. virilis* embryonic transcripts similar in size to products from the *D. melanogaster* locus (Newfeld et al. 1991). In *D. melanogaster mam* transcripts are expressed ubiquitously within the early embryo. Subsequently, transcripts are accumulated in ventral regions during gastrulation and early germ band extension. During later embryonic stages expression is widespread, but ultimately becomes restricted to the central nervous system (Smoller et al. 1990; Bettler et al. 1991). The spatial expression of *D. virilis mam* during these stages is very similar. Accumulation is initially ubiquitous, but enhanced ventrally during gastrulation (Fig. 2A,B). During germ band extension, expression is widespread within the germ layers (Fig. 2C) and later predominates in the central nervous system (Fig. 2D). The major features of *mam* expression are identical in these species, suggesting that *mam* function has been conserved.

### *Alignment of* D. virilis *and* D. melanogaster mam

An open reading frame of 4,968 bp (1,655 amino acids) was identified as *D. virilis mam. D. melanogaster mam* is 4,791 bp (1,596 amino acids). A comparison of the inferred amino acid sequences is shown in Fig. 3. Within the alignment there are highly conserved areas interspersed with divergent areas. The charge clusters are extremely similar. The 64-amino acid basic domain in exon 3 (*D. virilis* 173–237) contains 8 substitutions, of which 4 are conservative. The 79-residue acidic domain in exon 4 (acid 1; *D. virilis* 475–554) has 3 substitutions, of which 1 is conservative. The 34-residue acidic domain in exon 7 (acid 2; *D. virilis* 1,619–1,652) contains 4 substitutions, of which 1 is conservative, and a gap of 1 amino acid. Conservation is also seen in other nonrepetitive areas.

Numerous homopolymers (defined here as at least five consecutive identical amino acids) are found throughout both open reading frames. *D. virilis* contains 32 runs of polyglutamine, 6 runs of polyglycine, 2 runs of polyasparagine, and 2 runs of polyalanine. *D. melanogaster* has 21 runs of polyglutamine, 4 runs of polyglycine, 3 runs of polyasparagine, 1 run of polyalanine, and 1 run of polythreonine. Repetitive regions are defined here as homopolymers or segments where a single amino acid represents at least 50% of the residues. The most striking aspect of divergence in repetitive regions is large insertions and deletions. For example, *D. melanogaster* is deleted for an asparagine-rich region (*D. virilis* 122–146), a glutamine homopolymer (*D. virilis* 913–929), and an alanine-rich region (*D. virilis* 1,068–1,085), while *D. virilis* is deleted for an asparagine-rich region (*D. melanogaster* 274–297) and a glutamine-rich region (*D. melanogaster*

102–120). The gaps required in repetitive regions can be extensive. The largest gap in *D. virilis* is 28 amino acids (371–372), and in *D. melanogaster* 39 amino acids (406–407). The total size of gaps in each species is substantial. In *D. virilis* 8.6% and in *D. melanogaster* 12.7% of the alignment is occupied by gaps. Gaps required to maximize identity between the sequences occupy 20.5% of the alignment. Figure 4 shows a schematic representation of the alignment. Despite the variability of the homopolymers some repetitive features of *mam* are maintained in both species. For example, the central domain is populated exclusively with polyglutamine runs and polyglycine regions flank acid domain 1.

In spite of the length variation in intervening repetitive areas, conservation is seen in the amino acid spacing of the charge clusters. Most striking is the amino acid distance between the basic domain and acid domain 2. The distance between these clusters differs by roughly 1% between the species. These charge clusters are the most distant, separated by 84% of the amino acids in the alignment (*D. virilis* 1,382 residues; *D. melanogaster* 1,367 residues). The distance between the two acid domains is also conserved.

In *mam* homopolymers of glutamine are predominantly encoded by triplet repeats of CAG; asparagine by AAC; threonine by ACC; and alanine by GCA. Accounting for frameshifts, the CAG, GCA, ACC, and AAC repeats can be generalized to CAX (where X stands for any nucleotide). CAX repeats (opa; Wharton et al. 1985) can encode homopolymers of glutamine, asparagine, threonine and alanine. In *mam,* CAX repeats are not confined to homopolymers. One CAX repeat in the *mam* alignment occurs in *D. virilis* and encodes 54 amino acids of glutamine (CAA or CAG) and histidine (CAC or CAT; 364–418). A CAX repeat encoding 30 amino acids in *D. virilis* shifts four times between frame 1 and frame 3, encoding three stretches of glutamine interrupted by two runs of alanine (1,059–1,089). The longest CAX repeat is 285 bp (95 amino acids). This repeat (*D. virilis* 10–105) has degenerated at points along its length to encode other amino acids, many of which are also found in *D. melanogaster.*

The triplet repeat GGX encodes glycine homopolymers and is found in the hexanucleotide repeat GGXGTX that encodes glycine-valine. Several repetitive regions are dominated by GGX repeats, particularly where glycine homopolymers and glycine-valine runs are consecutive (*D. virilis* 1,234–1,257) or the 33-amino acid region that is 70% glycine (*D. virilis* 1,336–1,369). There are several differences between the species in the size and placement of glycine-valine regions. The region most similar between the species (*D. virilis* 1,273–1,295) shows only a slightly different pattern of valines. Just upstream (*D. virilis* 1,140–1,157) a regular glycine-valine run in *D. melanogaster* is aligned with a degenerate region in *D. virilis*. The glycine-valine run in *D. melanogaster* exon 5 (985–997) is completely absent in *D. virilis*. *D. virilis* has a glycine-valine run (1,580–1,596) aligned with a much shorter run of glycine in *D. melanogaster.*

### *Analysis of* mam *Alignment*

Table 1 demonstrates that the atypical amino acid composition of *D. melanogaster mam* reported by Smoller et al. (1990) is also evident in *D. virilis mam.* When compared to an average *Drosophila* protein, both sequences display an excess of the amino acids glutamine, glycine, and asparagine. Together these three amino acids comprise 47.6% of *D. melanogaster mam* and 48% of *D. virilis mam.* Charged amino acids are underrepresented in both species (11.5% *D. virilis;* 10.8% *D. melanogaster*), when compared to an average *Drosophila* protein. Table 2 shows that the similarity between the sequences of *D. virilis* and *D. melanogaster mam* extends to codon usage bias. For amino acids encoded by more than one triplet, in *D. melanogaster mam* the percentage of utilized codons with G and C in the third position is always greater than the percentage of codons ending in A and T. The single exception is isoleucine. This G + C > A + T pattern at the third position is also found in *D. virilis* including isoleucine. In *D. virilis* there are two exceptions, histidine and asparagine. Biases within synonymous codons are also similar between the species, such as the significant bias in glutamine (threefold-greater preference for CAG over CAA). For glycine the bias toward GGC is more pronounced in *D. virilis.* Both species match very well to a table of codon bias compiled from published *D. melanogaster* sequences (Smoller et al. 1990).

### *Analysis of Unique and Repetitive Domains of* mam

The *mam* open reading frames can be subdivided into repetitive and unique domains. The repetitive domain refers to a composite of all repetitive regions in the alignment, as identified in the legend to Fig. 3. The unique domain constitutes 38% and the repetitive domain 62% of the alignment. Homopolymers account for 39% of the repetitive domain (24% of the alignment). At the amino acid level, the unique domain is 89.7% identical and 92.4% similar between the species. Amino acid similarity includes identical amino acids and conservative substitutions. The repetitive domain is also well conserved

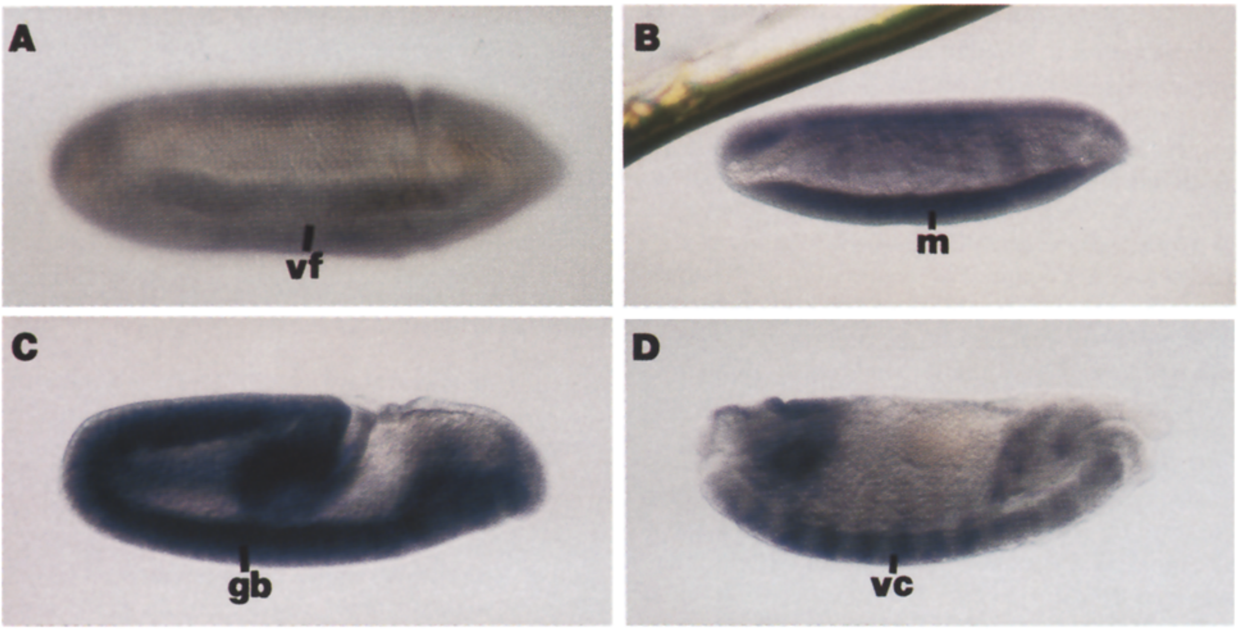
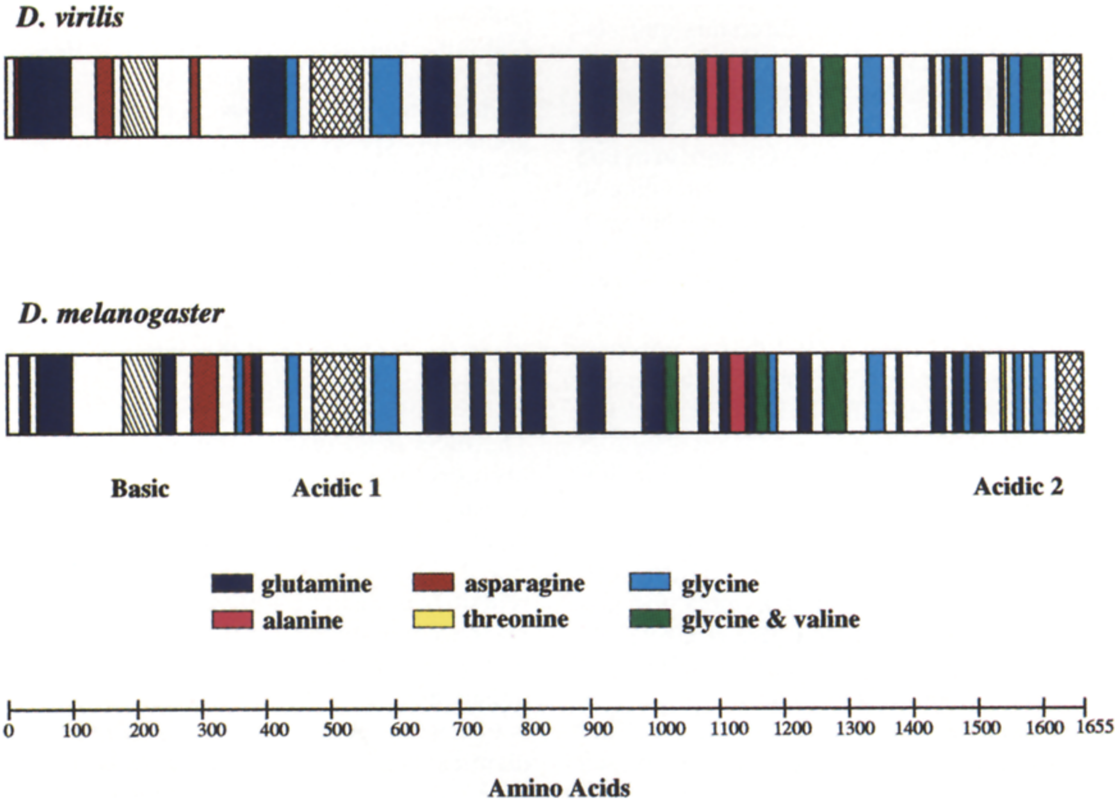


**Fig. 2.** Spatial accumulation of *mastermind* transcripts during *D. virilis* embryogenesis. Transcripts were detected using digoxigenin-labeled RNA probes. **A** Early gastrulation, expression is ubiquitous, but enhanced along the ventral furrow (*vf*). **B** Later gastrulation, high level of expression evident in mesoderm (*m*) and flanking endoderm. **C** Germ band extension, expression throughout ectodermal, neuroblast and mesodermal layers of germ band (*gb*). **D** Germ-band retraction, expression at high level in the central nervous system—i.e., the ventral cord (*vc*). Expression is still evident in the gut at this stage, but will soon become limited to the nervous system.



(74.1% identity; 76.6% similarity). The overall level of amino acid conservation is 81.5% identity and 84.1% similarity.

Table 3 summarizes the differences between the nucleotide and inferred amino acid sequences of *mam* for the total alignment as well as for the unique and repetitive domains. Chi-square analysis of the data in Table 3 shows that there are statistically significant differences between the domains in many categories. The repetitive domain exhibits more gaps and a larger proportion of the repetitive domain is devoted to gaps; 0.7% of the unique domain and 30.8% of the repetitive domain (a 43-fold difference) consist of gaps. There is no significant difference between the unique and repetitive domains in the total number of nucleotide substitutions. However, an examination of the distribution of nucleotide substitutions within a codon indicates that there are significant differences between the domains at all three positions. The repetitive domain shows more nucleotide substitutions in the first and second positions; the unique domain demonstrates more nucleotide substitutions in the third position. As a consequence, there are significantly more amino acid replacements in the repetitive domain. A comparison of silent and nonsilent substitutions in the unique and repetitive domains was statistically indistinguishable from the numbers reported in Table 3 as third-position substitutions and nonconservative amino acid replacements, respectively. The proportion of amino acid replacements per nucleotide substitution is twofold larger in the repetitive domain, a significantly faster rate of amino acid replacement. (See Table 3 for a description of the chi-square analysis.)

## Discussion

Beachy et al. (1985) proposed that a glycine homopolymer acts as a flexible connection between spatially distinct functional domains of *Ultrabithorax*. Laughon et al. (1985) suggested that genomic mechanisms such as unequal crossover, which would change homopolymer length, can generate variation in the physical spacing of functional regions and effect a phenotypic change. Treier et al. (1989) proposed that processes associated with molecular drive are important in the evolution of regulatory proteins. The unusual concentration of homopolymers in *mam* suggested that an interspecific comparison could provide insight into their role and evolutionary instability.

The analysis revealed that *D. virilis mam* is over 97 kb and identical, in exon organization of translated regions and major embryonic spatial transcription pattern, to the *D. melanogaster* locus (roughly 70 kb). This largely agrees with other investigators comparing genes between these species (e.g., Kassis et al. 1986; Treier et al. 1989; Michael et al. 1990), although Treier et al. (1989) did detect some differences in the expression of *D. virilis hunchback*. More detailed RNA and protein localization assays will be required to determine whether the interspecific expression patterns are identical in all tissues throughout the life cycle. The sequence comparison revealed that *D. virilis* and *D. melanogaster mam* share an unusual amino acid composition. Both loci are enriched in glutamine, glycine, and asparagine while containing fewer charged amino acids than an average *Drosophila* protein. Both genes demonstrate a typical codon bias for *Drosophila*. In addition, regions of alternating glycine and valine are found in both loci. No function has been demonstrated for this repeat that is found in several *Drosophila* genes (Finkelstein et al. 1990; Wilde and Akam 1987). Glycine-valine regions may serve the same spacer function as more hydrophilic homopolymer domains, with one important difference. The hydrophobicity of valine would sequester these regions from the surface of the protein. Thus, glycine-valine runs, which are predicted to form beta sheets, may act to maintain a specific conformation within the protein.

The comparison identified conservation of the basic and acidic charge clusters as well as other unique sequences. These regions demonstrate over 89% amino acid identity. Numerous interspecific sequence analyses, in *Drosophila* and other species, have shown that significant conservation is evident in functional domains. Even for genes which exhibit very low levels of overall conservation,

**Fig. 4.** Schematic representation of the *mastermind* alignment. The *wide rectangles* represent the inferred amino acid sequences of *D. virilis* and *D. melanogaster mam* as they are aligned in Fig. 3. The *coordinate scale* indicates amino acid positions as numbered for *D. virilis*. The areas depicted with *slashed lines* indicate the basic cluster. The areas depicted with *crossed lines* indicate the two acidic clusters. Repetitive regions are highlighted in *color. Purple* indicates polyglutamine; *magenta* indicates polyalanine; *red* indicates polyasparagine; *yellow* indicates polythreonine; *blue* indicates polyglycine- or glycine-rich regions and green indicates runs of alternating glycine and valine. The variability of the repetitive regions is evident. Note the following examples: (1) There is only one alanine homopolymer in *D. melanogaster* yet there are two alanine runs in *D. virilis*. (2) There are only two obvious glycine-valine regions in *D. virilis* and three in *D. melanogster*. (3) The number and composition of repetitive sequences surrounding the basic cluster are clearly distinct in each species. (4) There is no threonine homopolymer in *D. virilis*. The region homologous to the *D. melanogaster* threonine run contains only four consecutive threonine residues and we define five consecutive identical residues as the minimal homopolymer.

```
     exon3
vir  MDAGGLPVFQSASQAAAAAVAQQQQQQQQQQQQQQQQQQQQHLNLQLHQQHLQQQQSLGIHLQQQQQLQLQQQQHNAQAQQQQQLQVQQQQQQRQQQQQQ  100
mel  MDAGGLPVFQSASQAAA--VAQQQQQQQQQQ---------HLNLQLHQQHL------GLHLQQQQQLQLQQQ-HNAQAQQQQ-IQVQQQQQQQQQQQQQ   81
     *****************  ************         ***********      *.************** ********* .******** ******

vir  QQQHSLYNANLAAAGGIVGGLVPGGNGAGGVALQQVFGGPNGNNNSNNNNNSNNNSININNGNISPGDGLPTKRQPILDRLRRRMENYRRRQTDCVPRYE  200
mel  QQQHSPYNANLGATGGIAG--ITGGNGAGGPTNPGAVPTA-------------------------PGDTMPTKRMPVVDRLRRRMENYRRRQTDCVPRYE  154
     ***** *****.* *** *   . *******                                  ***  **** *..*********************

                                                                                                exon4
vir  QTFSTVCEQQNHETSALQKRFLESKNKRAAKKTEKKLPETQQQAQTQ-------------------MLAGQLQSSVHVQQKILKRPADDVDNGAENYEPP  281
mel  QAFNTVCEQQNQETTVLQKRFLESKNKRAAKKTDKKLPDPSQQHQQQQHQQQQQHQQHQQHQQAQTMLAGQLQSSVHVQQKFLKRPAEDVDNGPDSFEPP  254
     * * **********. *****************.****.  ** * *                   *************** *****.*****  . .***

vir  QKLPNNNNNNNNNNNNNNN------------------------SSSGVGGGSENLTKFSVEIVQQLEFTTSAANSQPQQISTNVTVKALTNTSVKSEPGV  357
mel  HKLPNNNNNSNSNNNNGNANANNGGNGSNTGNNTNNNGNSTNNNGGSNNNGSENLTKFSVEIVQQLEFTTSPANSQPQQISTNVTVKALTNTSVKSEPGV  354
      ******** * **** *                                *********************** ****************************

vir  GGGR----------------------------GRHQQQQQHQQHQQQQHQQQQHQQHQQHQQQQQHQQQQHQQQQHQQQQQQHHHQQQQQQGGGLGGLGN  429
mel  GGGGGGGGGGGSGNNNNNGGGGGGGNGNNNNNGGDHHQQQQQHQHQQQQQQQ---------------------------------------GGGLGGLGN  415
     ***                             *    *** *   *** ***                                       *********

vir  NGRGGGGPGGGGHMATGPGGV-----GVGMGPNMMSAQQKSALGNLANLVECKREPDHDFPDLGSLDKDGANGQFPGFPDLLGDDNSENNDTFKDLINNL  524
mel  NGRGGGPGG----MATGPGGVAGGLGGMGMPPNMMSAQQKSALGNLANLVECKREPDHDFPDLGSLDKDGGGGQFPGFPDLLGDDNSENNDTFKDLINNL  511
     ******  *    ********     * ** *************************************. ****************************

vir  HDFNPSFLDGFDEKPLLDIKTEDGIKVEPPNAQDLINSLNVKSETGLGHGFGGFGVGLGLDPQSMKMRPG-----VGFQNGPNGNANAGNGGPTAGGGGG  619
mel  QDFNPSFLDGFDEKPLLDIKTEDGIKVEPPNAQDLINSLNVKSEGGLGHGFGGFGLGL--DNPGMKMRGGNPGNQGGFPNGPNGGTGGAPNAGGNGGN--  607
      ******************************************** **********.**  *    **** *      ** *****   ..  .   **

vir  GNGPGGLMSEHSLAAQTLKQMAEQHQHKSAMGGMGGFHVPPHGM--QQQQPQQQQQAPQQQQQQHGQMMGGPGQGQQQQQQQQPRYNDYGGGFPNDFAMG  717
mel  ---SGNLMSEHPLAAQTLKQMAEQHQHKNAMGGMGGFPRPPHGMNPQQQQQQQQQQQQQQQAQQQHGQMMG---------QGQPGRYNDYGGGFPNDFGLG  695
        * ***** **************** ********  *****  **** *****  ** *********            * *  *************. *

vir  PNPTQQQQQ----------HLPPQFHQ-KAPGGGPGMNVQQNFLDIKQELFYSSPNDFDLKHLQQQQAMQQQQQQQQQQQQQQQHHAQQQQQHPNGPNMG  806
mel  PNGPQQQQQQAQQQQPQQQHLPPQFHQQKGPGPGAGMNVQQNFLDIKQELFYSSQNDFDLKRLQQQQAMQQQQQQQHHQQQQ------------------  777
     **  *****          ******** *.** * ******************* ******.************** ****

vir  VPMGGAGNFAKQQQQQVPTPQQQQQQQLQQQQQQ-----YSPFSNQNANA--NFLNCPPRGGPQGNQAPGNM------PQQQQQQPQQQQQPPRGPQSNP  893
mel  PKMGGVPNFNKQQQQQQVPQQQLQQQQQQQQQQQQQQQQYSPFSNQNPNAAANFLNCPPRGGPNGNQQPGNLAQQQQQPGAGPQQQQQRGNAGNGQQNNP  877
       ***  ** ******    ** **** ******      ******** **  *********** *** ***        *    ** **      * * **

                                                                                 exon5
vir  NAVPGGNAANATQQQQQQQQQQQQQQQQQQQQQQQQQQATTTTLQMKQTQQLHISQQGGGSHGIQVSAGQHLHLSSDMKSNVSVAAQQGVFFSQQQAAQQQQQ  993
mel  NTGPGGNTPNAPQQQQQQQ----------------STTTTLQMKQTQQLHISQQGGGAQGIQVSAGQHLHLSGDMKSNVSVAAQQGVFFSQQQAQQQQQQ  961
     *  ****  ** *******                 *********************  ************************************ *****

                                                                             exon6
vir  QQQQPGNA-GPNPQQQQQQPHGGNAGA-----------NGGGPNGPQQQQPNQNMNNSNVPSDGFSLSQSQSMNFTQQQQQQAAAAAAAAAAAQQQQAAA 1082
mel  QQ--PGGTNGPNPQQQQQQPHGGNAGGGVGVGVGVGVGNGGPNPGQQQQQPNQNMSNANVPSDGFSLSQSQSMNFNQQQQQQAAA--------------- 1044
     **  **   ****************.           ***  * ********* * ***************** *********

vir  AQQQQQQVPPNMRQRQTQAQAAAAAAAAAAAQAQAAANANGGPGGNVPLMQQQQQTPGGVPVGAGSGNASVGVPV----SAGGPNNGAMNQLGGPMGGMP 1177
mel  ---QQQQVQPNMRQRQTQAQ-AAAAAAAAAAQAQAAANASG---PNVPLMQQ-PQVGVGVGVGVGVGVGVGNGGVVGGPGSGGPNNGAMNQMGGPMGGMP 1136
        ***** *********** ***************** *     ******* *   ** ** * * .    *      ********** ********

vir  GMQMGGPGGVPINPMQMNPNGGAPNAQ-MMMGGNGGGPVPAAS--------QAKFLQQQQIMRAQAMQHQQQVQQHMAGARPPPPEYNATKAQLMQAQMM 1268
mel  GMQMGGP----MNPMQMNPNAAGPTAQQMMMGSGAGGPGQVPGPGQGPNPNQAKFLQQQQMMRAQAMQQQQQ---HMSGARPPPPEYNATKAQLMQAQMM 1229
     *******    ********...* ** ****  .***                ********* ******* ***   ** *********************

vir  QQTVGGGGGGGVGVGVGVGGGVGGGGGAGRFPNSAAQAAAMRRMTQQPIPPSGPMMRPQHAAMYMQQHGGAGGGPRGGMGGPYGGGGVGGAGGPMGGGGG 1368
mel  QQTVGGGGVGVGGVGVGVGVGGVGGANGGRFPNSAAQAAAMRRMTQQPIPPSGPMMRPQHA-MYMQQHGGAGGGPRTGMGVPYGGG----RGGPMGGP-- 1324
     ******** *  ******* *  **. .********************************* ************* *** *****     ******

                                                                         exon7
vir  GQQQQQRPPNVQVTPDGMPMGSQQEWRHMMMTQQQQQMGFG---PGGPMRQGPGGFNGGNFMPNGAPNAPGNGPNGGGGGGMMPGPNGPQMQLTPAQMQQ 1465
mel  --QQQQRPPNVQVTPDGMPMGSQQEWRHMMMTQQQTQMGFGGPGPGGPMRQGPGGFNGGNFMPNGAPNGAAGSGPNA--GGMMTGPNVPQMQLTPAQMQQ 1418
       ********************************** ****   ************************.  .    .  **** *** ************

vir  QHMRQQQQQQ------HMGPGGGGGGGGGNMQMQQLLQQQQNAAAGGGGGMMATQMQMTSIHMSQTQQQQQLTMQQQQ-FVQSTSTTTTHQQQQQLQLQM 1558
mel  QLMRQQQQQQQQQQQQHMGPGAANN-----MQMQQLLQQQQS---GGGGNMMASQMQMTS--MHMTQTQQQITMQQQQQFVQS-TTTTTHQQQQMMQMGP 1507
     * ********       *****..     ***********    **** ***.******  *  ** ***.****** ****  .*********  *

vir  QSQSGGPGGNGPSNNNGANQAGGVGVGVGVGVGVGVVGSSATIASASSISQTINSVVANSNDLCLEFLDNLP-DGNFSTQDLINSLDNDNFNIQDILQ/ 1655
mel  GGGGGGGGPGSANNNN----------GGGGGGAAGGGNSASTIASASSISQTINSVVANSNDFGLEFLDNLPVDSNFSTQDLINSLDNDNFNLQDFNMP/ 1596
         ** *    ***          * * * . *   *  ********************  ******** * ****************.**
```

such as *transformer* (36% amino acid identity between *D. virilis* and *D. melanogaster;* O'Neil and Belote 1992), small stretches of identity are functionally significant. In addition to the content of the charge clusters, conservation is seen in the amino acid spacing of these regions. The highest degree of conservation is seen in the amino acid distance between the basic domain (possibly DNA binding) and acid domain 2 (putative transcriptional activation), which differs by roughly 1% between the species. In *mam,* these charge clusters are the most distant, separated by 84% of the amino acids in the alignment (*D. virilis* 1,382 residues; *D. melanogaster* 1,367 residues). Given the length variability of the repetitive sequences which separate these regions, it is likely that natural selection is maintaining this spacing, although it cannot be formally ruled out that the similarity in spacing results simply from chance. Sequencing *mam* from a third *Drosophila* species or germline transformation assays with *mam* constructs containing modifications in homopolymer length will be required to distinguish these possibilities. However, it has been brought to our attention that the conservation in spacing between the basic domain and acid domain 2 can be quantitatively evaluated by the following argument. The distance between the basic domain and acid domain 1 is, averaged over the two species, 254 amino acids. For this segment, there is a difference in length between the species of 32 amino acids (12.6% of the average length). The average distance between the basic domain and acid domain 2 is 1,375 amino acids and the difference between the species is 15 (1% of the average length). On a purely random binomial basis, the standard error of the number of "successes" in *N* Bernoulli trials is proportional to the square root of *N*, and the standard error of the proportion of "successes" is proportional to 1/square root *N*. Thus, the expected number of added amino acids ("successes") in 1,375 amino acids is approximately sqrt (1,375/254) = 2.3 times as many as in 254 amino acids. The expected proportion of added amino acids in 1,375 is approximately 1/2.3 = 44% based on the proportion seen in 254 amino acids. The observed number of added amino acids in the long sequence is only 0.5 times as many (15/32), instead of the expected 2.3 times as many "successes". The observed proportion of additional amino acids in the long sequence is only 8% of the proportion of "successes" in the short sequence (1% vs 12.6%), whereas the expected proportion of additional amino acids in 1,375 is 44%. Thus, the expected length difference between the basic domain and acid 2 is 0.44/0.08 = 5.5 times larger than observed, using the short sequence as a standard. The apparent high level of conservation in sequence and spacing suggests that the charge clusters contain important functional residues and that their relative positions may be critical for *mam* function.

The conservation of the charged clusters is distinct from the variation evident within repetitive areas. The most striking aspect of divergence within the repetitive domain is the numerous large insertions and deletions. The proportion of gaps in the repetitive domain (30.8%) is 43-fold greater than the fraction of the unique domain occupied by gaps (0.7%), and 20.5% of the total alignment is devoted to insertions and deletions. Interspecific comparisons of other *Drosophila* genes that contain repetitive sequences (Kassis et al. 1986; Treier et al. 1989; Jones et al. 1991; Heberlein and Rubin 1990; Peixoto et al. 1992) demonstrated similar length alterations. Comparison of the human, mouse, and rat glucocorticoid receptors (Danielson et al. 1986) revealed that length variation in homopolymers is widespread. Intraspecific studies in *D. melanogaster* (Tautz 1989; Costa et al. 1991) indicate that homopolymer length can change rapidly.

In addition to length variation in the repetitive domain, a significant degree of divergence is evident in the amino acid content of this region. There

**Fig. 3.** Comparison of the inferred amino acid sequences from *D. virilis* and *D. melanogaster mastermind.* The amino acid sequences (in one-letter code) deduced from *D. virilis* genomic DNA sequence (4,968 bp) and *D. melanogaster* genomic and cDNA sequences (4,791 bp; Smoller et al. 1990) are aligned. Amino acid 1 corresponds to the first methionine in the open reading frames of both species. Amino acids are numbered consecutively for each species and indicated at the right margin. The position of exon boundaries is indicated above the *D. virilis* sequence and exons are numbered according to Smoller et al. (1990). The position of charge clusters is indicated by **bold type;** the basic cluster corresponds to *D. virilis* residues 173–236, acidic cluster 1 residues 475–552, and acidic cluster 2 residues 1619–1651. The four conserved cysteine residues are *underlined. Dashed lines* indicate gaps required in the alignment to demonstrate maximum identity between the sequences. *A slash* indicates the first stop codon encountered in the open reading frame of each species. *An asterisk* below the *D. melanogaster* sequence indicates an identical amino acid in both species. *A dot* below the *D. melanogaster* sequence indicates a conservative substitution, based on biochemical similarity of the amino acids: D/E, K/R/H, N/Q, S/T, I/L/V, F/W/Y, and A/G (Smith and Smith 1990). Repetitive regions are located at *D. virilis* residues 11–165, 241–247, 286–309, 358–449, 570–625, 664–700, 722–840, 872–933, 984–1031, 1059–1188, 1215–1243, 1269–1297, 1334–1374, 1401–1408, 1440–1599. The corresponding nucleotide sequences are available from GenBank; *D. virilis* M92914, *D. melanogaster* X54251. In addition, nucleotide conservation is evident for 60 nucleotides upstream of the predicted initiator codon. In this region, nearly 80% (47 of 60) of the nucleotides are identical. If this region were translated, in the same frame as the *mam* open reading frame, there would be 12 identical amino acids and one conservative substitution (65% amino acid similarity). Yet a comparison of these upstream sequences to Kozak's (1989) consensus for translation initiation shows a very poor match.

**Table 1.** Comparison of the inferred amino acid composition of *D. virilis* and *D. melanogaster mastermind* to an average *Drosophila* protein[a]

| | DV total | DM total | DV % | DM % | *Drosophila* % | Ratio[b] DV | Ratio[b] DM |
|---|---|---|---|---|---|---|---|
| Gln (Q) | 402 | 347 | 24.3 | 21.1 | 5.0 | 4.8 | 4.3 |
| Gly (G) | 255 | 269 | 15.4 | 16.8 | 7.8 | 2.0 | 2.2 |
| Asn (N) | 137 | 155 | 8.3 | 9.7 | 4.6 | 1.8 | 2.1 |
| Pro (P) | 118 | 123 | 7.1 | 7.7 | 5.7 | 1.2 | 1.4 |
| Ala (A) | 135 | 108 | 8.2 | 6.7 | 7.8 | 1.1 | 0.9 |
| Met (M) | 75 | 82 | 4.5 | 5.1 | 2.4 | 1.8 | 2.1 |
| Leu (L) | 79 | 71 | 4.8 | 4.4 | 8.2 | 0.6 | 0.5 |
| Ser (S) | 77 | 71 | 4.7 | 4.4 | 7.6 | 0.6 | 0.6 |
| Val (V) | 66 | 64 | 4.0 | 4.0 | 6.0 | 0.7 | 0.7 |
| Thr (T) | 50 | 58 | 3.0 | 3.6 | 5.5 | 0.6 | 0.7 |
| Asp (D) | 34 | 42 | 2.0 | 2.6 | 5.1 | 0.4 | 0.5 |
| Phe (F) | 39 | 42 | 2.4 | 2.6 | 3.4 | 0.6 | 0.8 |
| His (H) | 52 | 39 | 3.1 | 2.4 | 2.8 | 1.1 | 0.9 |
| Arg (R) | 34 | 34 | 2.1 | 2.1 | 5.2 | 0.4 | 0.4 |
| Lys (K) | 33 | 34 | 2.1 | 2.1 | 5.3 | 0.4 | 0.4 |
| Glu (E) | 26 | 24 | 1.6 | 1.5 | 6.2 | 0.3 | 0.2 |
| Ile (I) | 26 | 18 | 1.6 | 1.1 | 4.7 | 0.3 | 0.2 |
| Tyr (Y) | 11 | 10 | 0.7 | 0.6 | 3.2 | 0.2 | 0.2 |
| Cys (C) | 5 | 4 | 0.3 | 0.2 | 2.3 | 0.1 | 0.1 |
| Trp (W) | 1 | 1 | 0.06 | 0.06 | 1.1 | 0.06 | 0.06 |
| Total | 1,655 | 1,596 | | | | | |
| Positive (HRK) | 129 | 107 | 7.3 | 6.6 | 13.3 | 0.6 | 0.5 |
| Negative (DE) | 60 | 66 | 3.6 | 4.1 | 11.3 | 0.3 | 0.4 |

[a] The amino acid composition of an average *Drosophila* protein was calculated from a table compiled by M. Ashburner (unpublished). The amino acid composition of *D. melanogaster mam* was reported in Smoller *et al.* 1990

[b] Ratio = *mastermind%/Drosophila%*

**Table 2.** Comparison of *D. virilis* and *D. melanogaster* codon usage at *mastermind*

| | | DV% | DM% | | | DV% | DM% | | | DV% | DM% | | | DV% | DM% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TTT | Phe | 0.66 | 0.81 | TCT | Ser | 0.12 | 0.25 | TAT | Tyr | 0.18 | 0.25 | TGT | Cys | 0.06 | 0.06 |
| TTC | Phe | 1.69 | 1.82 | TCC | Ser | 0.91 | 0.88 | TAC | Tyr | 0.48 | 0.38 | TGC | Cys | 0.24 | 0.19 |
| TTA | Leu | 0.06 | 0.06 | TCA | Ser | 0.24 | 0.13 | TAA | *** | 0.00 | 0.00 | TGA | *** | 0.00 | 0.00 |
| TTG | Leu | 0.91 | 0.81 | TCG | Ser | 0.97 | 1.06 | TAG | *** | 0.06 | 0.06 | TGG | Trp | 0.06 | 0.06 |
| CTT | Leu | 0.18 | 0.06 | CCT | Pro | 0.66 | 0.63 | CAT | His | 1.75 | 1.19 | CGT | Arg | 0.42 | 0.19 |
| CTC | Leu | 1.09 | 1.19 | CCC | Pro | 2.48 | 2.82 | CAC | His | 1.39 | 1.31 | CGC | Arg | 0.85 | 0.88 |
| CTA | Leu | 0.18 | 0.19 | CCA | Pro | 1.09 | 1.44 | CAA | Gln | 5.92 | 4.76 | CGA | Arg | 0.30 | 0.38 |
| CTG | Leu | 2.36 | 2.13 | CCG | Pro | 2.90 | 2.76 | CAG | Gln | 18.30 | 16.91 | CGG | Arg | 0.30 | 0.38 |
| ATT | Ile | 0.54 | 0.44 | ACT | Thr | 0.18 | 0.25 | AAT | Asn | 4.47 | 4.51 | AGT | Ser | 0.30 | 0.56 |
| ATC | Ile | 0.91 | 0.56 | ACC | Thr | 1.03 | 1.88 | AAC | Asn | 3.74 | 5.13 | AGC | Ser | 2.11 | 1.57 |
| ATA | Ile | 0.12 | 0.13 | ACA | Thr | 0.79 | 0.50 | AAA | Lys | 0.54 | 0.44 | AGA | Arg | 0.12 | 0.00 |
| ATG | Met | 4.53 | 5.13 | ACG | Thr | 1.03 | 0.94 | AAG | Lys | 1.45 | 1.69 | AGG | Arg | 0.06 | 0.31 |
| GTT | Val | 0.72 | 0.75 | GCT | Ala | 0.79 | 1.00 | GAT | Asp | 0.91 | 1.50 | GGT | Gly | 2.29 | 4.82 |
| GTC | Val | 1.15 | 0.81 | GCC | Ala | 3.02 | 2.88 | GAC | Asp | 1.39 | 1.13 | GGC | Gly | 10.63 | 7.89 |
| GTA | Val | 0.48 | 0.31 | GCA | Ala | 2.29 | 1.06 | GAA | Glu | 0.60 | 0.38 | GGA | Gly | 1.63 | 3.38 |
| GTG | Val | 1.63 | 2.13 | GCG | Ala | 1.99 | 1.94 | GAG | Glu | 0.97 | 1.13 | GGG | Gly | 0.79 | 0.81 |

is no difference between the unique and repetitive domain of *mam* with regard to the total number of nucleotide substitutions; however, the distribution of nucleotide substitutions within a codon shows significant differences between the domains at all three positions. This distribution of nucleotide substitutions is reflected in significantly more amino acid replacements in the repetitive domain. As a result, the proportion of amino acid replacements per nucleotide substitution is twofold larger in the

**Table 3.** Differences between the aligned nucleotide and inferred amino acid sequences of *mastermind*[a]

| | | | | Nucleotide substitutions | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Gaps | | Codon position | | | | Nonconservative amino acid | Amino acid replacements/total |
| Domain | Size (bp) | No. | Size (bp) | 1 | 2 | 3 | Total | replacements | substitutions |
| Unique | 2,079 | 4 | 15 | 59 | 38 | 238* | 335 | 52 | 0.15 |
| Repetitive[b] | 3,315 | 44*** | 1020*** | 126*** | 146*** | 323 | 595 | 179*** | 0.30** |
| Total | 5,394[c] | 48 | 1035 | 185 | 184 | 561 | 930 | 231 | 0.25 |

[a] Asterisks indicate the degree of significant difference between the unique and repetitive domains; * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$. The location of the asterisk in each category indicates which domain significantly exceeded its expected value in a chi-square test of the null hypothesis: after correcting for size differences the unique and repetitive domains are equivalent. In *mam,* the repetitive domain is 60% larger than the unique domain. Thus for the test a correction factor of 1.6 must be applied to observed values in the unique domain to normalize differences in domain size. The normalized values for the unique domain were then tested for deviation from expected values derived from the repetitive domain

[b] Repetitive regions are defined as distinct homopolymers (at least five consecutive identical amino acids) and regions dominated by a single amino acid. The specific regions of the alignment described as the repetitive domain are listed in the legend of Figure 3.

[c] Includes 4,965 bp of *D. virilis* open reading frame plus 429 bp equivalent to gaps.

repetitive domain. Statistical analysis indicates that this represents a significantly faster rate of amino acid replacement.

The differences in the degree of conservation indicate that the unique and repetitive domains of *mam* are characterized by distinct patterns of evolutionary change. Unique regions are highly conserved, varying only occasionally through amino acid replacement and small insertions and deletions. In contrast, repetitive areas have diverged as a result of numerous large insertions and deletions and a significantly greater number of amino acid replacements. The data suggest that the repetitive domain of *mam* is under less selective constraint than the unique domain. This is consistent with the idea that homopolymers act as flexible spacers and consequently are better able to tolerate amino acid replacements than unique sequences (Beachy et al. 1985). However, repetitive regions do not change nearly as fast as noncoding segments of *mam* (Newfeld et al. 1991), indicating some selective constraint. One possibility, suggested by the flexible spacer hypothesis, is that the repetitive regions of *mam* must maintain an unstructured conformation. Amino acid stretches devoid of charge or hydrophobic residues (such as certain homopolymers) may form random coils that provide physical flexibility to tertiary conformation (Brendel and Karlin 1989). Thus, the conservation seen in the amino acid content of repetitive regions may reflect the level of selective constraint necessary to remain unstructured.

The length variability in the repetitive domain is likely due to the higher probability of nucleotide misalignment in the simple sequences which encode homopolymers. A misalignment may lead to slipped-strand mispairing during DNA replication or repair (Levinson and Gutman 1987), short-tract gene conversion (Wheeler et al. 1990), unequal crossover between alleles (Lyons et al. 1988), or unequal sister chromatid exchange (Jeffreys et al. 1988). These genomic mechanisms (associated with molecular drive; Dover 1986) would alter the number of residues in the encoded homopolymer. Studies in *E. coli* and yeast (Farabaugh et al. 1978; Rothstein et al. 1987) have demonstrated that repetitive regions are hotspots for insertions or deletions that change the number of repeats. Current models of molecular drive depict these genomic processes acting independent of selection.

However, a mathematical model for replication slippage in repetitive sequences which assumes selective neutrality with regard to repeat number was rejected by data from interspecific comparisons (Tachida and Iizuka 1992). In addition, evidence of selective constraint on repeat length variation was recently reported for the repetitive region of the *period* gene in *Drosophila* (Peixoto et al. 1992). Data from *mam* is consistent with these results. The extensive length variation in repetitive domains of *mam,* presumably generated by internal genomic processes, appears to be balanced by natural selection acting to maintain the distance between specific charge clusters.

If the interaction between drive and selection (as proposed for *mam*) is generalized to other homopolymer-containing proteins there may be advantages to the population. For example, a change in the selective force on a protein can result from mutations in an interacting molecule. Such mutations may be successfully accommodated more rapidly through the high frequency of misalignment-mediated events in repetitive sequences than through point mutation. The disparity in frequency between these two types of event is enormous. Jeffreys et al. (1988) report a spontaneous mutation rate to new length alleles of $5 \times 10^{-4}$/locus/gamete for human minisatellites (noncoding repetitive sequences) with length changes of 4–200 repeat units. Alternatively, the synonymous substitution rate for humans is $1.1 \times 10^{-9}$/site/year (Li and Tanimura 1987). The concept of a dynamic drive-selection interaction may partially explain the common occurrence of homopolymers in regulatory molecules. Viewed from this perspective, the potential physical flexibility and length variability (evolutionary flexibility) of homopolymers could provide a reservoir of adaptations for a population.

In conclusion, the distinct pattern of divergence in repetitive areas of *mam* is consistent with the hypothesis that homopolymers are flexible spacers within proteins. The conservation of the charge clusters implies an important functional role. The juxtaposition of length variability in repetitive areas and conserved spacing for charge clusters is proposed to reflect an interaction between molecular drive and natural selection. This may be a factor in the evolution of homopolymer-containing proteins.

*Acknowledgments.* For valuable discussions we thank Antonio Pannuti, Temple Smith, and Norma Takaesu. Helpful comments on drafts of this manuscript from Bill Gelbart and Charles Aquadro are acknowledged. We thank an anonymous reviewer for suggesting the mathematical evaluation of length conservation. Elena Lozovskaya, Dimitri Petrov, and Dave Smoller (Hartl laboratory, Washington University) screened the *D. virilis* P1 library and Elena Lozovskaya performed the salivary in situ hybridization. This work was supported by the National Science Foundation. S.J.N. received a Biomolecular Engineering Research Center (Boston University) visiting scientist fellowship.

## References

Beachy P, Helfand S, Hogness D (1985) Segmental distribution of *bithorax* complex proteins during *Drosophila* development. Nature 313:545–551

Bettler D, Schmid A, Yedvobnick B (1991) Early ventral expression of the *Drosophila* neurogenic locus *mastermind.* Dev Biol 144:436–439

Beverley S, Wilson A (1984) Molecular evolution in *Drosophila* and higher *Diptera*. II. A time scale for fly evolution. J Mol Evol 21:1–13

Brendel V, Karlin S (1989) Association of charge clusters with functional domains of cellular transcription factors. Proc Natl Acad Sci USA 86:5696–5702

Chen DJ, Chan CS, Pirrotta V (1992) Conserved DNA binding and self-association domains of the *Drosophila zeste* protein. Mol Cel Biol 12:598–608

Chen E, Seeburg P (1985) Supercoil sequencing: a fast and simple method for sequencing plasmid DNA. DNA 4:165–170

Costa R, Peixoto A, Thackery J, Dalgleish R, Kyriacou C (1991) Length polymorphism in the threonine-glycine encoding repeat region of the *period* gene in *Drosophila.* J Mol Evol 32:238–246

Danielson M, Northrop J, Ringold G (1986) The mouse glucocorticoid receptor: Mapping of functional domains by cloning, sequencing and expression of wild-type and mutant receptor proteins. EMBO J 5:2513–2522

Dente L, Sollazzo M, Baldari C, Cesareni G, Cortese R (1985) The pEMBL family of single stranded vectors. In: Glover D (ed) DNA cloning: A practical approach, volume I. IRL Press, Oxford, England, pp 101–108

Dover G (1986) Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet 2:159–165

Duboule D, Haenlin M, Galliot B, Mohier E (1987) DNA sequences homologous to the *Drosophila* opa repeat are present in murine mRNAs that are differentially expressed in fetuses and adult. Mol Cell Biol 7:2003–2006

Farabaugh P, Schmeissner U, Hoter M, Miller J (1979) On the molecular nature of spontaneous hotspots in the *lacI* gene of *Escherichia coli.* J Mol Biol 126:847–863

Faulkner DV, Jurka J (1988) Multiple aligned sequence editor (MASE). Trends Biochem Sci 13:321–322

Finkelstein R, Smouse D, Capaci T, Spradling A, Perrimon N (1990) The *orthodenticle* gene encodes a novel homeodomain protein involved in the development of the *Drosophila* nervous system and ocellar visual structure. Genes Dev 4:516–527

Fischer J, Giniger E, Maniatis T, Ptashne M (1988) GAL4 activates transcription in *Drosophila.* Nature 332:853–856

Friedel C (1990) An analysis of the transcriptional activity at the neurogenic locus *mastermind* in *Drosophila melanogaster.* MS thesis, Emory University, Atlanta, GA

Fu YH, Kuhl DPA, Pizzuti A, Pieretti M, Sutcliffe J, Richards S, Verkerk AJMH, Holden JJA, Fenwick Jr RG, Warren ST, Oostra BA, Nelson DL, Caskey CT (1991) Variation of the CGG repeat at the Fragile X site results in genetic instability: resolution of the Sherman paradox. Cell 67:1047–1058

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r73

Harley H, Brook J, Rundle S, Crows S, Reardon W, Buckler A, Harper P, Houseman D, Shaw D (1992) Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy. Nature 355:545–546

Heberlein U, Rubin G (1990) Structural and functional comparisons of the *Drosophila virilis* and *Drosophila melanogaster* rough genes. Proc Natl Acad Sci USA 87:5916–5920

Jeffreys A, Royle N, Wilson V, Wong Z (1988) Spontaneous mutation rates to new length alleles at tandem repetitive hypervariable loci in human DNA. Nature 322:278–281

Jones C, Dalton M, Townley L (1991) Interspecific comparisons of the structure and regulation of the *Drosophila* ecdysone inducible gene E74. Genetics 127:535–543

Kassis J, Poole S, Wright D, O'Farrell P (1986) Sequence conservation in the protein coding and intron regions of the *engrailed* transcription unit. EMBO J 5:3583–3589

Kozak M (1989) The scanning model for translation: An update. J Cell Biol 108:262–273

LaSpada A, Wilson E, Lebahn D, Harding A, Fischback K (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352:77–79

Laughon A, Carrol S, Storter F, Riley P, Scott M (1985) Common properties of proteins encoded by the *Antennapedia* complex genes of *Drosophila melanogaster.* Cold Spring Harb Symp Quant Biol 50:253–262

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

Li WH, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93–96

Lyons K, Stein J, Smithies O (1988) Length polymorphism in human proline rich proteins generated by intragenic unequal crossingover. Genetics 120:267–278

Michael W, Bowtell D, Rubin G (1990) Comparison of the *sevenless* genes of *Drosophila virilis* and *Drosophila melanogaster.* Proc Natl Acad Sci USA 87:5351–5353

Newfeld S, Smoller D, Yedvobnick B (1991) Interspecific comparison of the unusually repetitive *Drosophila* locus *mastermind.* J Mol Evol 32:415–420

O'Neil M, Belote J (1992) Interspecific comparison of the *transformer* gene of *Drosophila* reveals an unusually high degree of evolutionary divergence. Genetics 131:113–128.

Patterson T, Dean M (1987) Preparation of high titre lambda phage lysates. Nucleic Acids Res 15:6298

Peterson M, Tanese N, Pugh B, Tjian R (1990) Functional domains and upstream activation properties of cloned human TATA binding protein. Science 248:1625–1630

Peixoto AA, Costa R, Wheeler DA, Hall JC, Kyriacou CP (1992) Evolution of the threonine-glycine repeat region of the period gene in the *melanogaster* species subgroup of *Drosophila.* J Mol Evol 35:411–419

Rothstein R, Helsm C, Rosenberg N (1987) Concerted deletions and inversions are caused by mitotic recombination between delta sequences in *Saccharomyces cerevisiae.* Mol Cell Biol 7:1198–1207

Smith RF, Smith TF (1990) Automatic generation of primary sequence patterns from sets of related protein sequences. Proc Natl Acad Sci USA 87:118–122

Smith RF, Smith TF (1992) Pattern Induced Multi-sequence Alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. Protein Engineering 5:35–41

Smoller D, Friedel C, Schmid A, Bettler D, Lam L, Yedvobnick B (1990) The *Drosophila* neurogenic locus *mastermind* encodes a nuclear protein unusually rich in amino acid homopolymers. Genes Dev 4:1688–1700

Smoller D, Petrov D, Hartl D (1991) Characterization of bacteriophage P1 library containing inserts of *Drosophila* DNA of 75–100 kilobase pairs. Chromosoma 100:487–494

Tachida H, Iizuka M (1992) Persistence of repeated sequences that evolve by replication slippage. Genetics 131:471–478

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17:6463–6471

Treier M, Pfeifle C, Tautz D (1989) Comparison of the gap segmentation gene *hunchback* between *Drosophila melanogaster* and *Drosophila virilis* reveals novel modes of evolutionary change. EMBO J 8:1517–1525

Tseng H, Green H (1988) Remodeling of the involucrin gene during primate evolution. Cell 54:411–496

Wharton K, Yedvobnick B, Finnerty V, Artavanis-Tsakonas S (1985) Opa: A novel family of transcribed repeats shared by the *Notch* locus and other developmentally regulated loci in *D. melanogaster.* Cell 40:55–62

Wheeler C, Maloney D, Fogel S, Goodenow R (1990) Microconversion between murine H-2 genes integrated into yeast. Nature 347:192–194

Wilde CD, Akam M (1987) Conserved sequence elements in the 5′ region of the *Ultrabithorax* transcription unit. EMBO J 5:1393–1401

Yedvobnick B, Smoller D, Young P, Mills D (1988) Molecular analysis of the neurogenic locus *mastermind* of *Drosophila melanogaster.* Genetics 118:483–497

Zhu Q, Smith T, Lathrop R, Figge J (1990) Acid helix-turn activator motif. Proteins 8:156–163