Origin and Evolution of the Regulatory Gene *male-specific lethal-3*

Ignacio Marín[1] *and Bruce S. Baker*

Department of Biological Sciences, Stanford University

Dosage compensation in *Drosophila* is mediated by genes known as ''*male-specific lethals*'' (*msl*s). Several *msl*s, including *male-specific lethal-3* (*msl-3*), encode proteins of unknown function. We cloned the *Drosophila virilis msl-3* gene. Using the information provided by the sequences of the *Drosophila melanogaster* and *D. virilis* genes, we found that sequences of other species can be aligned along their entire lengths with *msl-3.* Among them, there are genes in yeasts (the *Schizosaccharomyces pombe Alp13* gene, as well as a putative *Alp13* homolog, found in *Saccharomyces cerevisae*) and in mammals (*MRG15* and *MSL3L1* and their relatives) plus uncharacterized sequences of the nematode *Caenorhabditis elegans* and the plants *Arabidopsis thaliana, Lycopersicon esculentum,* and *Zea mays.* A second *Drosophila* gene of this family has also been found. It is thus likely that *msl-3*-like genes are present in all eukaryotes. Phylogenetic analyses suggest that *msl-3* is orthologous to the mammalian *MSL3L1* genes, while the second *Drosophila melanogaster* gene (which we have called *Dm MRG15*) is orthologous to mammalian *MRG15.* These analyses also suggest that the *msl-3*/*MRG15* duplication occurred after the fungus/animal split, while an independent duplication occurred in plants. The proteins encoded by these genes have similar structures, including a putative chromodomain close to their N-terminal end and a putative leucine zipper at their C-terminus. The possible functional roles of these proteins are discussed.

Introduction

Dosage compensation in *Drosophila* occurs by hypertranscription of the single male X chromosome to achieve a level of transcripts similar to that produced by the two female X chromosomes. Hypertranscription is mediated by the action of a complex, termed the compensasome (Franke and Baker 1999), formed by at least five proteins and two RNA molecules that bind together along the male X chromosome. These proteins are encoded by the genes collectively known, because of their characteristic phenotypes, as ''*male-specific lethals*'' (*msl*s) (reviewed in Baker, Gorman, and Marín 1994; Bashaw and Baker 1996; Cline and Meyer 1996; Lucchesi 1998). The five *msl*s are called, respectively, *msl-1, msl-2, msl-3, maleless* (*mle*), and *males-absent-on-the-first* (*mof)* (see review by Lucchesi 1998). Recent data (Franke and Baker 1999; Smith et al. 2000) show that two noncoding RNAs (encoded by the *roX1* and *roX2* genes; Amrein and Axel 1997; Meller et al. 1997) are also part of the complex (see reviews by Lucchesi 1999; Stuckenholz, Kageyama, and Kuroda 1999). The presence of compensasomes along the male X chromosome results in this chromosome having a more relaxed chromatin structure than the X chromosomes of the females (see details in Baker, Gorman, and Marín 1994). Although the precise molecular mechanism of action of the compensasome is unknown, it has been demonstrated that the male X chromosome is enriched in histone H4 acetylated at the lysine-16 residue (H4Ac16; Turner, Birley, and Lavender 1992) and that the distribution of H4Ac16 along the male X chromosome is MSL-dependent (Bone et al. 1994; Smith et al. 2000). Because MOF has sequence similarity to histone acetyltransferases (Hilfiker et al. 1997), it is likely that at least part of the function of the compensasome is to carry out such histone modification. The complex may have other biochemical functions. For example, MLE is homologous to a mammalian RNA/DNA helicase (Kuroda et al. 1991; Lee and Hurwitz 1993) and ATPase and helicase activities have been demonstrated for the *Drosophila* protein (Lee et al. 1997). However, what role, if any, these activities have in dosage compensation is unknown.

How the MSLs recognize their targets on the X chromosome and how they bind to each other is also unclear, although certain basic properties of their association are known. Thus, genetic data indicate that each of the MSL proteins (reviewed in Baker, Gorman, and Marín 1994) and either one of the ROX RNAs (Franke and Baker 1999) is needed for the assembly of compensasomes along the male X chromosome. There are also recent data suggesting that certain places along the X chromosome may serve as entry sites for the compensasomes to start binding to the chromosome (Kelley et al. 1999). Moreover, immunoprecipitation and yeast two-hybrid assays suggest that MSL-1, MSL-2, MSL-3, and MOF are strongly interacting, while MLE is less tightly associated with the complex (Kelley et al. 1995; Copps et al. 1998; Smith et al. 2000).

It has been shown that the MSL complex is acting male X chromosome–specifically in many drosophilid species (Marín et al. 1996; Bone and Kuroda 1996). Thus, comparisons of the MSL protein sequences in different *Drosophila* species could provide insight into their functions. Generally, in this type of study the species taken as a model for comparison is *Drosophila virilis,* which belongs to a different subgenus of the *Drosophila* genus and diverged from *D. melanogaster* 40–60 MYA (Kwiatowski et al. 1994; Russo, Takezaki, and Nei 1995 and references therein). For example, in a recent work, Copps et al. (1998) described the sequence of the *D. virilis msl-2* gene, showing that both the func-

[1] Present address: Departamento de Genética and Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, Spain.

Key words: dosage compensation, duplicated genes, *male-specific lethals,* human, yeast.

Address for correspondence and reprints: Ignacio Marín, Departamento de Genética, Universidad de Valencia, Doctor Moliner, 50, Burjassot 46100 Valencia, Spain. E-mail: ignacio.marin@uv.es.

Mol. Biol. Evol. 17(8):1240–1250. 2000

tionally important ring finger and a second characteristic cysteine-rich region are conserved in both species.

We present here the cloning of the *D. virilis* homolog of the *msl-3* gene. MSL-3 was initially described as a novel protein with only a very weak similarity to some human and rice sequences of unknown function (Gorman, Franke, and Baker 1995). Subsequently, Koonin, Zhou, and Lucchesi (1995) described two regions of MSL-3 that, according to their analyses, had some similarity with a ''chromodomain,'' a motif that has been found in many proteins involved in chromatin regulation, such as POLYCOMB and HP1 (Aasland and Stewart 1995). Quite recently, Bertram et al. (1999) and Prakash et al (1999) described several mammalian genes with similarities to *msl-3.* These authors also detected similarities with other sequences in yeasts and plants. However, the evolutionary history of *msl-3* and its relatives has never been studied in detail. In this study, by comparison of the *D. melanogaster* and *D. virilis* sequences as well as the mammalian genes, we determined the regions of these proteins that show a high degree of sequence conservation. When the conserved regions were used to perform database searches, we found that MSL-3 is actually a member of a class of proteins found in many other eukaryotic organisms, including not only animals, but also yeasts and plants.

Materials and Methods

Cloning of *D. virilis msl-3*

A *D. virilis* genomic library generously provided by Paul McDonald (Department of Biological Sciences, Stanford University) was screened using a cDNA probe that contained the whole *D. melanogaster msl-3* gene (Gorman, Franke, and Baker 1995), labeled nonradioactively with biotin-dUTP. A total of 15 hybridizing phages were isolated from approximately 3.5 genome equivalents screened. Restriction analysis showed that these clones fell into six different classes, with one to four overlapping phages per class. Representatives of each class were then labeled with biotin-dUTP and used as probes for in situ hybridizations on polytene chromosomes of *D. virilis.* Only one of the phages mapped to chromosome 3 of *D. virilis,* which is known to be homologous to chromosome 3L in *D. melanogaster,* where *msl-3* is found. Because it is very rare to find homologous genes located on non-homologous chromosomes in *D. melanogaster* and *D. virilis* (Whiting et al. 1989), we centered our analysis on the four overlapping phages corresponding to this class. The hybridization signal was mapped to region 34B, according to the map by Gubenko and Evgenev (1984). These phages were then analyzed by restriction digests and Southern blots to determine the regions that hybridized with *msl-3.* Hybridization was restricted to a 3.0-kb *Spe*I-*Spe*I fragment found in three of the four phages, with the fourth having only part of this *Spe*I fragment. The 3.0-kb *Spe*I fragment was subcloned into pBluescript, and a new round of restriction digests and Southern blot analyses were performed in order to more precisely locate the position of the gene. Sequencing of both strands of all the hybridizing regions was performed using the dideoxy method. By analyzing the sequences obtained and by comparing them with the *D. melanogaster* gene, we were able to determine the likely 5′ and 3′ ends of the *D. virilis* gene, as well as the presence of five short putative introns, that were situated in exactly the same positions as the *msl-3* introns in the *D. melanogaster* gene.

Comparative Sequence Analysis

BLASTN, TBLASTN, and TBLASTX analyses (Altschul et al. 1997) were performed against the nonredundant, month, and dBEST databases at NCBI (http://www.ncbi.nlm.nih.gov/; see details below). The analyses were finished in early February 2000. Multiple-sequence alignments were obtained using CLUSTAL X (Thompson et al. 1997). The relationships among the sequences were determined using the neighbor-joining algorithm (Saitou and Nei 1987). The reliability of the tree topology obtained was assessed using the bootstrap routine also present in the CLUSTAL X program. A total of 1,000 bootstrap replicates were performed. Figures 2 and 3 were drawn using the programs GENEDOC (Nicholas and Nicholas 1997) and TREEVIEW 1.5 (Page 1996), respectively. To determine whether other proteins with similarity to the five conserved blocks were present in the databases, profiles of such blocks were obtained and the databases were searched with the highly sensitive program PSI-BLAST (Altschul et al. 1997) or, alternatively, with PROFILEMAKE and PROFILESEARCH (Gribskov, McLachlan, and Eisenberg 1987).

Results

Comparative Analysis Shows that Many Eukaryotic Species Have *msl-3*-Related Genes

Figure 1 shows the sequence of the *msl-3* gene of *D. virilis* (GenBank accession number AF247726). The congruent chromosomal localization, the apparently identical exon/intron structure, and the high similarity in amino acid sequence (62% identity and 74% similarity according to BLAST alignments after excluding some short duplicated regions of the *D. virilis* protein that will be detailed in the next section) established that we had indeed cloned the *D. virilis* ortholog of *msl-3,* and not a related gene. The degree of sequence similarity we found was comparable to that obtained for the orthologs of other developmentally important genes between these two species (see O'Neil and Belote 1992). After obtaining an alignment of the *D. melanogaster* and *D. virilis* genes, we used the regions with the highest degree of conservation (corresponding to amino acids 10–101, 194–352, and 401–490 in *D. melanogaster*) for TBLASTN searches against the nonredundant database at NCBI. We found highly significant scores for two yeast sequences, one from *Saccharomyces cerevisiae* (accession number Z49274) and the second from *Schizosaccharomyces pombe* (accession number Z88977; the ''expected values'' [E values], i.e., the number of sequences expected by chance to have such similarity in

```
   1 AGTGATAAGT TGATAGTTGG TGCCATATAT GCAAAAGTGA TGACGGCACA TGAAAACGAG
                                              M  T  A  H     E  N  E

  61 GTCCAGCTTT TCAACCGGGG TGAGAAAGTC CTGTGCTACG AGCCCGACGA ATCGAAGGCT
     V  Q  L   F  N  R  G   E  K  V   L  C  Y   E  P  D  E   S  K  A

 121 AGGGTCTTGT ACGACAGCAA GGTAACCCAC GCACACACAC AGACAGACAT GCATACACGC
      R  V  L   Y  D  S  K

 181 ACACAGCGAC AGCGTGCCGA AACGTCAAAG AAAAATGCCG AACACGAGGT GGCTGTTTCG
 241 ATTTTTCATC ACGAAAAACT GCTCTTTTCT ATAGGTTTTG GCCGTGTACG AGCGAAAGGA
                                        V  L   A  V  Y   E  R  K  D

 301 CGCTGCAAAT CTACGCTACT TTGATTACAA AATCCATTTT CAAGGCTGGA ACTCATCCTG
      A  A  N   L  R  Y   F  D  Y  K   I  H  F   Q  G  W   N  S  S  W

 361 GGATCGAAAT GTTCGTGCCG CCAGTCTGTT AAAAGACAAC GAAGAAAATC GCAAATTGCA
      D  R  N   V  R  A   A  S  L  L   K  D  N   E  E  N   R  K  L  Q

 421 GCGTGAGCTG GCAGAAGCCG CACAGCTACA AAAGTGAGTA AAATACGCGT GTGTGTGTGC
      R  E  L   A  E  A   A  Q  L  Q   K

 481 TCGTGTGTGT GCATGTGTGT GACCGTTTGT TTACAATTTT CAGAACTGGC GGCTACTCGT
                                                 T  G   G  Y  S

 541 ACAAAGACTC TAAGACTCCG ACGTTGCCAT CCAGCAAGAA GAAGCGCTTG GCCAGAGGTG
     Y  K  D  S   K  T  P   T  L  P   S  S  K  K   K  R  L   A  R  G

 601 GACACGTAGA AGATCCCACA GCCGATCCCC TAGACATATC GTTGCCATCC AGCAAGAAGA
     G  H  V  E   D  P  T   A  D  P   L  D  I  S   L  P  S   S  K  K

 661 AGCGCTTGGC CAGAGGTGGA CACGTAGAAG ATCCCACAGC CGATCCCCTA GACATATCGC
     K  R  L  A   R  G  G   H  V  E   D  P  T  A   D  P  L   D  I  S

 721 TGGCACATTT GCCGGCGACA CCAAAACCCG AACTGCCGGC ACAACAGAAG AGAGGCGCAC
     L  A  H  L   P  A  T   P  K  P   E  L  P  A   Q  Q  K   R  G  A

 781 GCTCGCGGGA TGGCAGTGGC AATCGGTCGC GGGATGGCAG TGGCAACCGG TCGCGGGACA
     R  S  R  D   G  S  G   N  R  S   R  D  G  S   G  N  R   S  R  D

 841 ACAGTAGCGG CCGCAAAAAG CAGCGCGACA AGTCGAAAGG CGGTGATAAA AACGATGATG
     N  S  S  G   R  K  K   Q  R  D   K  S  K  G   G  D  K   N  D  D

 901 GTGAGCGAAG ATCTGCTAGA TCTCAGTTTA GCCTGCATGT CTCACCTAAA GATACGCACA
     G  E  R  R   S  A  R   S  Q  F   S  L  H  V   S  P  K   D  T  H

 961 CAACTGATGC AGAAAAACGG ATCCATCAGG AGGACCGTGT GATGCTGCGC ATCAGCGAGC
     T  T  D  A   E  K  R   I  H  Q   E  D  R  V   M  L  R   I  S  E

1021 GGCTGCGCGA ATACATGGAA TACGACTACA ACATGGTCTG CAAGCTTGAA AAGCAGCACG
      R  L  R  E   Y  M  E   Y  D  Y   N  M  V  C   K  L  E   K  Q  H

1081 CATTGCCCGC TCGCACACCC ATTGTGACCA TATTGGAGAA CTTTGTGAAG CAGCGCGCCG
      A  L  P  A   R  T  P   I  V  T   I  L  E  N   F  V  K   Q  R  A

1141 TCGAGCTGGC CATTGGCATC AAGCAGGACA GTTCTCGAGC CCGCAACACT CTGAGCCGTA
     V  E  L  A   I  G  I   K  Q  D   S  S  R  A   R  N  T   L  S  R

1201 ATGCGCGTAT GGAGCGTGAG TACGATCGTG TGATGTCCAT GTGAGTATTC GGTGTTACAT

     N  A  R  M   E  R  E   Y  D  R   V  M  S  I
```

FIG. 1.—Nucleotide sequence of the *msl-3* gene of *Drosophila virilis* and its conceptual translation. The initial methionine is indicated in bold and italics. Bold letters correspond to the five conserved motifs of amino acids found in all of the MSL-3-related proteins (see text). Italics mark the five putative introns, located in the same positions as in *D. melanogaster*.

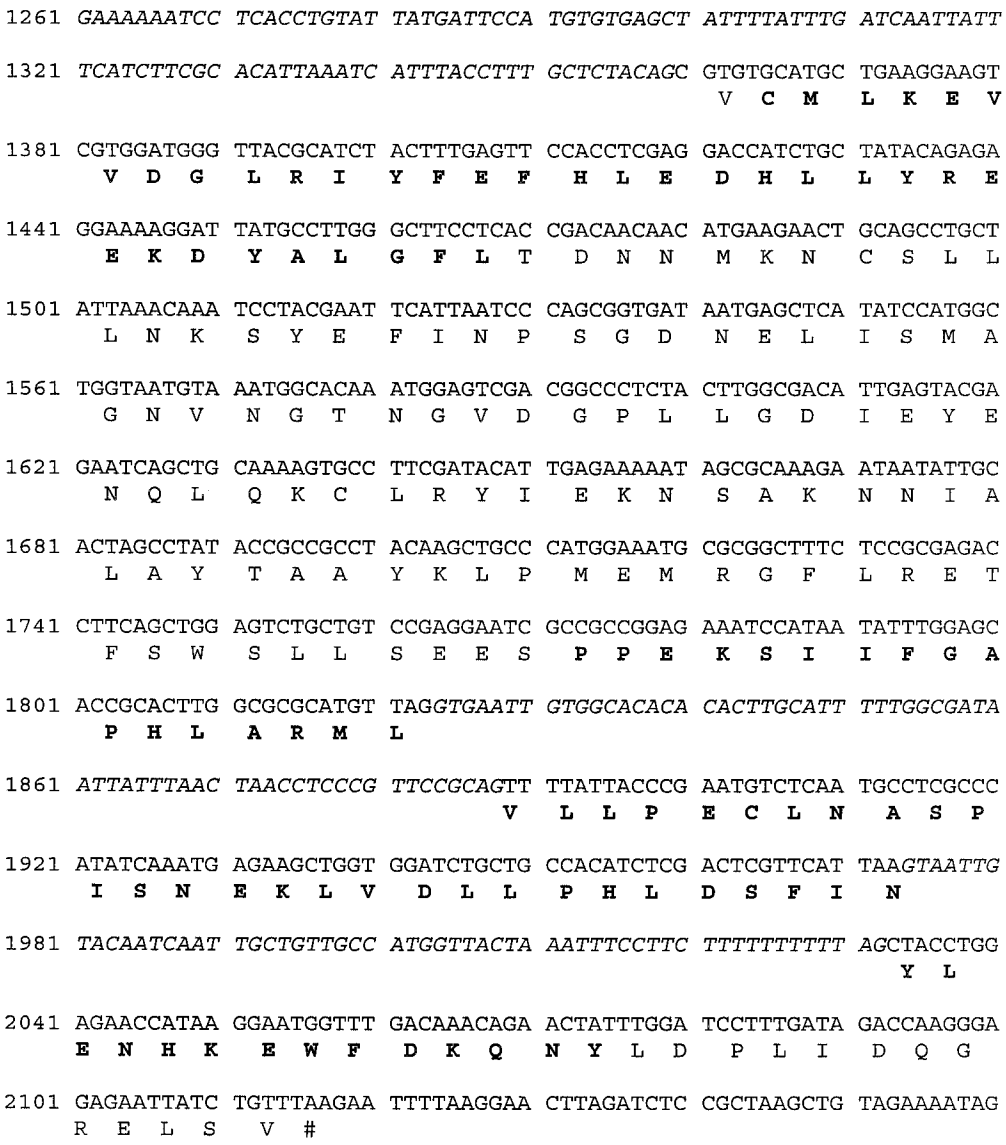

FIG. 1 *(Continued)*

a certain database, was 3×10^{-7} and 6×10^{-5}, respectively, for analyses using the nonredundant database). The *S. pombe* sequence corresponded to a gene known as *Alp13* (Radcliffe et al. 1998), while we call the *S. cerevisiae* sequence *Sc Alp13* (see below for a justification). We also found homology to a cosmid (accession number AC004554) containing human genomic sequences located in the chromosomal region Xp22 (E value = 9×10^{-4}). These sequences correspond to the human gene *MSL3L1* (*MSL-3 like 1*), described by Prakash et al. (1999; accession number NM_006800.1). Before the latter study was published, we had reconstructed the same gene by combining data from different ESTs present in the dBEST database and genomic sequences of the cosmid AC004554. Our independent characterization confirms the intron-exon structure and length of the gene described by Prakash et al. (1999). The gene has 13 exons. Prakash et al. (1999) also characterized a very similar gene in the mouse, which they also called *Msl3l1* (accession number NM_010832.1). After characterizing *MSL3L1*, we found another human clone, derived from chromosome 2 (accession number AC005538.2), that contained a sequence very similar (86% identity, 92% similarity) to *MSL3L1*. The most obvious difference was that the AC005538.2 sequence did not contain introns. Fragments of at least two cDNAs with homology to the sequence in AC05538.2 were present in the dBEST database (accession numbers AA476632 and AA355764). We discuss this sequence, which we have called *MSL3L1-b*, in detail below.

Additionally, we found that there were other overlapping *D. melanogaster* sequences (accession numbers AI238997, AA817270, AA951428, AA941284, and AA539729) present in the dBEST databases that showed clear protein similarities, but were not identical, to MSL-3 (E values = 2×10^{-6} to 3×10^{-7}). We used these *Drosophila* sequences to recheck the nonredundant and dBEST databases and found that they were very

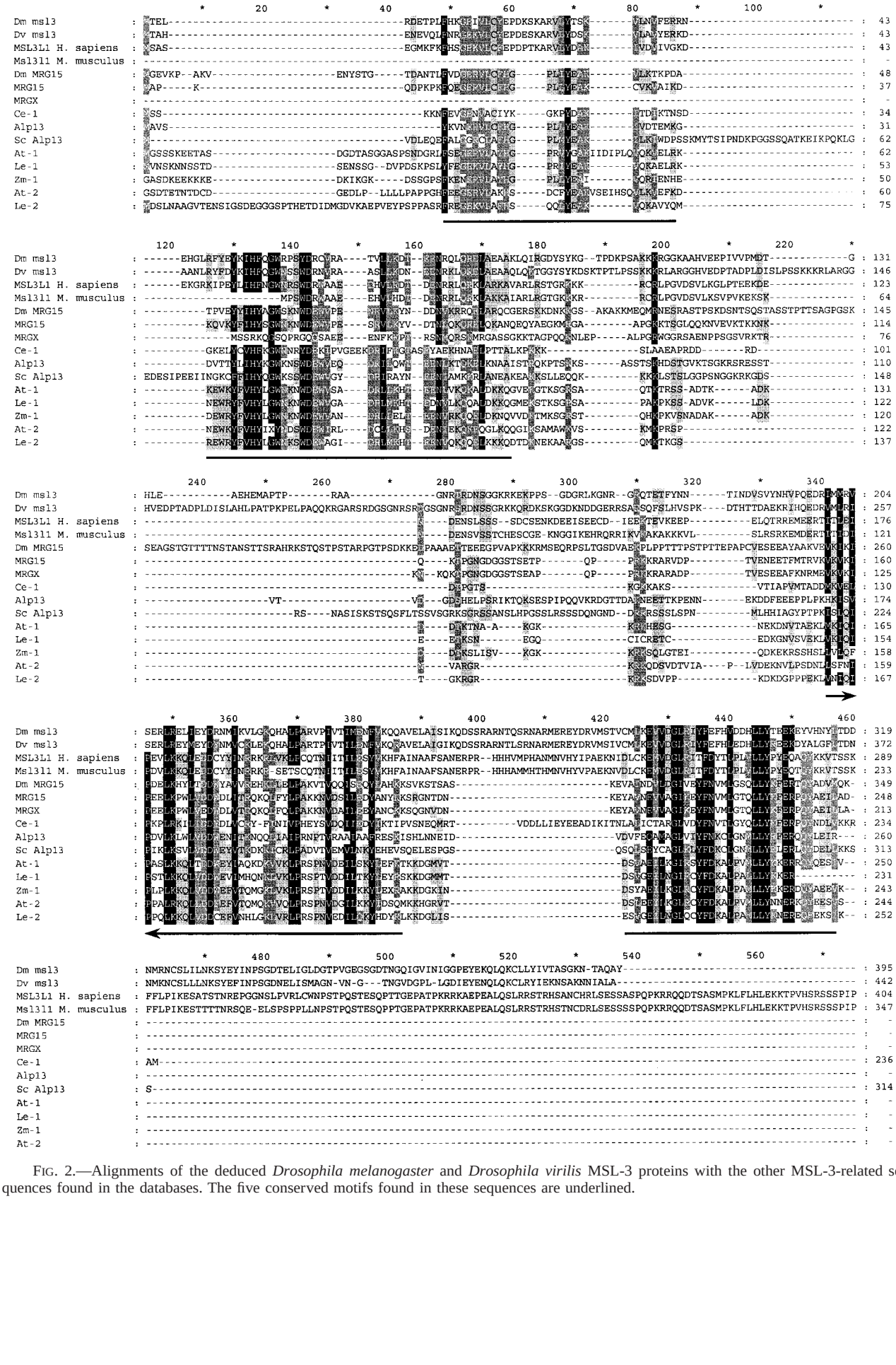

FIG. 2.—Alignments of the deduced *Drosophila melanogaster* and *Drosophila virilis* MSL-3 proteins with the other MSL-3-related sequences found in the databases. The five conserved motifs found in these sequences are underlined.

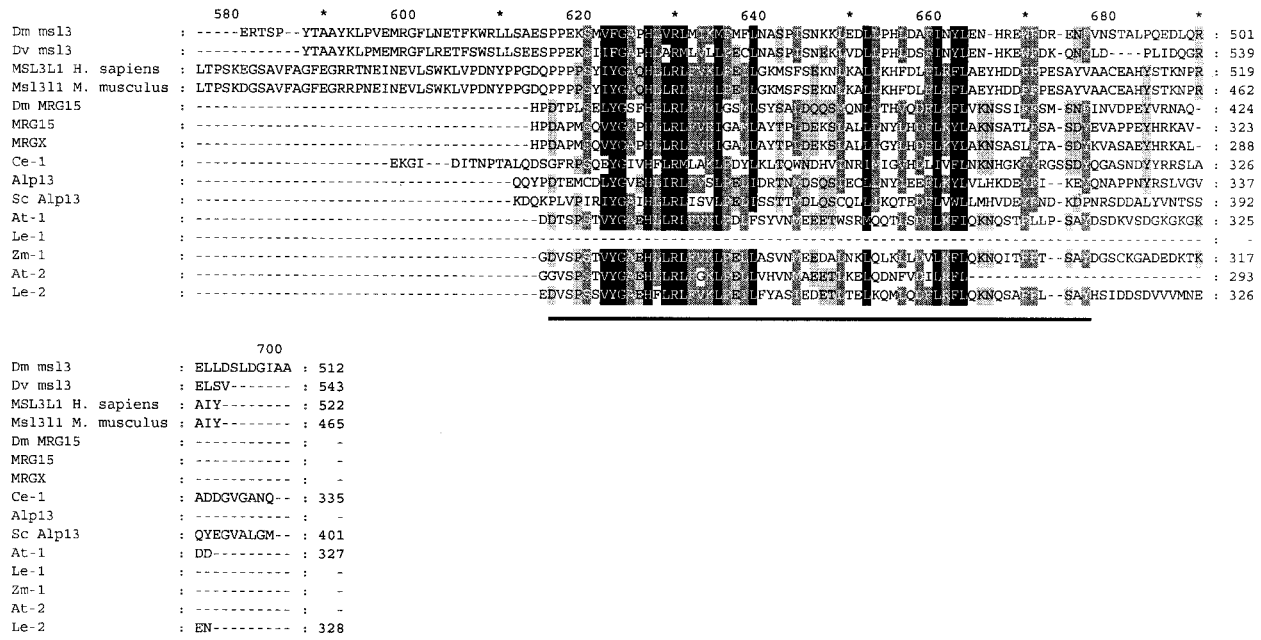

FIG. 2 *(Continued)*

similar (E value = 4×10^{-15}) to the human sequences (accession number D14812) already detected by Gorman, Franke, and Baker [1995] in their original report on *msl-3*, as well as to other, almost identical, sequences of other mammalian species. By combining all of the information on human sequences, we were also able to reconstruct the second human gene; it corresponds to a gene discovered by Bertram et al. (1999), who named it *MRG 15* (accession number NM_006791.1). These authors also discovered several *MRG 15*-related sequences (see below). More recently, a new *Drosophila* sequence was released (accession number AF152245; Bertram and Pereira-Smith, personal communication) that includes the ESTs described above and may correspond to the complete *Drosophila* gene. For reasons that will become clear later in this study, we call this *Drosophila msl-3*-related gene *Dm MRG15.*

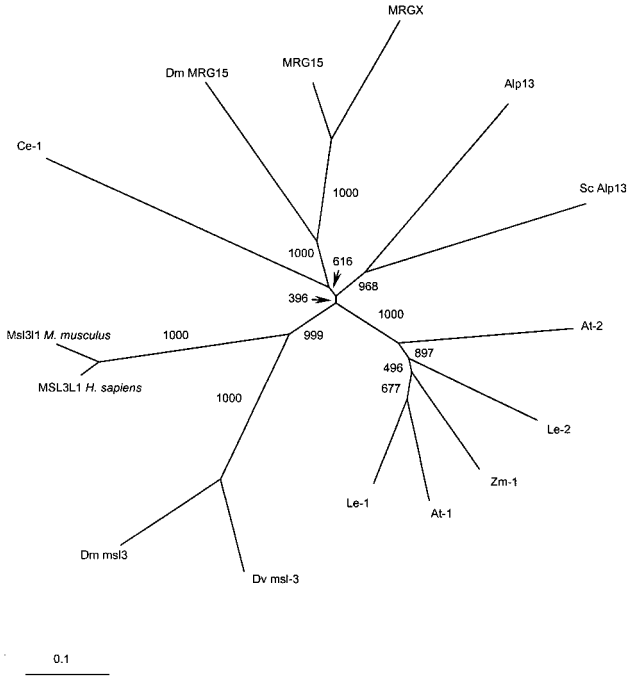

FIG. 3.—Phylogenetic tree for the *MSL-3*-related sequences. Bootstrapping results (out of 1,000 replicates) are indicated.

The comparison of the human and *Drosophila* sequences allowed us to establish the existence of five short conserved blocks of similarity between *msl-3* and all of its related sequences. We then rechecked the nonredundant, month, and dBEST NCBI databases using these blocks as query and found other sequences with the same characteristic blocks of conserved amino acids. These include one from *Caenorhabditis elegans* (Ce-1; accession number AL032626), two from *Arabidopsis thaliana* (At-1 [accession number Z99707.1] and At-2 [combination of sequences with accession numbers AAF02891 and AC022521.1]; At-2 lacks a few amino acids of the fifth conserved domain), one from *Lycopersicon esculentum* (Le-1; combination of sequences with accession numbers AW216739 and AW218605), and one from *Zea mays* (Zm-1; combination of sequences with accession numbers AW076378 and AW054199.1). Partial plant sequences with similarity to several of those blocks were also found. The most complete, which includes four of the five blocks, was observed in *L. esculentum* (Le-2; obtained by combining sequences AW223494.1 and AW224009.1), and it has also been included in the phylogenetic analyses below. Other related sequences from *Glycine max*, *Sorghum bicolor*, and *Oryza sativa* are quite short, and they have not been included in our analyses.

Bertram et al. (1999) found six other sequences related to *MRG 15*, but only two of them corresponded to active genes (four were processed pseudogenes). The

two other transcribed genes were called *MRGX* and *MORF4* by Bertram et al. (1999). The MORF4 protein is almost identical to MRG15, except that MORF4 lacks 88 amino acids at its N-terminal end. *MRGX* is less similar to *MRG15*, and the protein *MRGX* encodes also has an N-terminal truncation: MRGX lacks the 57 N-terminal amino acids of MRG15. When we carefully compared the *MSL3L1* sequence and its processed relative of the second chromosome (*MSL3L1-b*), we found a similar situation. In spite of the fact that the homology between these two sequences starts at amino acid 35 of *MSL3L1*, there are stop codons located 3′ of that position, and thus the putative start methionine in the *MSL3L1-b* sequence is located much more C-terminally, corresponding to position 167 in *MSL3L1*. Because both *MORF4* and *MSL3L1-b* are almost identical (but shorter) than *MRG15* and *MSL3L1*, they were not included in the phylogenetic comparisons below.

In summary, the rest of our analyses were based on a total of 15 related sequences, belonging to 10 different species. Three of these sequences were from insects, the *D. melanogaster msl-3* and *Dm MRG15* genes and the *D. virilis msl-3* gene. Four more were from mammalian species, the human genes *MSL3L1, MRGX*, and *MRG15* and the mouse gene *Msl3l1*. Two were from yeasts, the *S. pombe Alp13* gene and the *S. cerevisiae Sc Alp13* sequence. One was from the nematode *C. elegans* (Ce-1). Finally, five were from plants, namely, the *A. thaliana* At-1 and At-2 sequences, the *L. esculentum* Le-1 and Le-2 sequences, and the *Z. mays* Zm-1 sequence. Figure 2 shows the alignments of all of these sequences.

An important preliminary point to consider is whether we were able to precisely determine both the N- and the C-termini of all of these proteins. We think that was the case for 12 of the 15 proteins for the following reasons: (1) All of the characterized genes (Gorman, Franke, and Baker 1995; Bertram et al. 1999; Prakash et al. 1999) had start methionines and stop codons in similar positions. Moreover, a total of five conserved blocks of amino acids were found in these sequences, and in the known genes, the putative N- and C-terminal ends were situated similarly, very close to the two most extreme blocks of conserved amino acids (fig. 2). We expected the other proteins to have similar N- and C-termini, and that was what was observed. (2) We think the *D. virilis* sequence corresponds to the full-length gene because the structure described here was identical to the longest cDNA ever observed in *D. melanogaster*. Three proteins were, however, incomplete: the N-termini of the Zm-1 and At-2 sequences was unknown, and the Le-1 and At-2 sequences were missing part of their C-termini. While it is still possible that some alternative 5′ or 3′ ends of the genes exist, or even that some of the genes are longer than we were able to determine, these problems do not affect any of the conclusions presented below.

Figure 3 shows the neighbor-joining (Saitou and Nei 1987) tree corresponding to the alignment in figure 2. The topology of the tree was identical when the whole sequences were compared or when only the five conserved domains of the 15 sequences described above were selected. Figure 3 includes the bootstrap support for the latter analysis. These results suggest that the genes described by Prakash et al. (1999) for mammals are indeed the orthologs of *Drosophila msl-3*, while the second *Drosophila* gene, which we called *Dm MRG15*, is the ortholog of mammalian *MRG15* and its relatives, described by Bertram et al. (1999). Bertram et al. (1999) also suggest that the single *C. elegans* gene is an ortholog of the *MRG15* genes. The relationships among the plant genes are less clear, although the simpler explanation is that the two genes found in *Arabidopsis* and *Lycopersicon* are the product of a duplication that occurred before the separation of the lineages that gave rise to these species (see *Discussion*). Figure 3 also suggests that *Sc Alp13* and *Alp13* are orthologous.

Structural Analysis of MSL-3 and its Relatives

Koonin, Zhou, and Lucchesi (1995) described two chromodomains in MSL-3. The most N-terminal chromodomain they described would correspond to the second half of the first conserved block and roughly the first half of the second conserved block, as well as the nonconservative region between those two blocks. We searched the databases for other proteins containing similar sequences using a profile of this region and the PROFILESEARCH program. Our results confirm the finding by Koonin, Zhou, and Lucchesi (1995) that this region has similarity with previously described chromodomains. However, such similarity is low, and, in fact, the alignment is quite difficult outside of a region (positions 130–154 in fig. 2) that is much shorter than the accepted common definition for a chromodomain (at least 37 and up to more than 50 amino acids, depending on the authors; see Aasland and Stewart 1995; Koonin, Zhou, and Lucchesi 1995). Such difficulty is clearly shown by the fact that Koonin, Zhou, and Lucchesi (1995) already detected an EST that contained part of the human MRG15 sequence, but they failed to properly align such a sequence with MSL-3 and the other chromodomain-containing proteins. Thus, it is unclear whether this is a unique type of chromodomain, with features that are specific for MSL-3-related proteins, or an altogether different domain. Most interestingly, as we said before, the human sequences *MRG X, MORF 4* (Bertram et al. 1999), and *MSL3L1-b* do not have the same N-terminus as *MRG 15* or *MSL3L1*. In particular, they lack the just-described chromodomain-like sequence found in the *MRG 15* and *MSL3L1* genes (see *Discussion*).

The putative second chromodomain would correspond, according to Koonin, Zhou, and Lucchesi (1995), to positions 639–677 in figure 2, that is, approximately to the region where the fifth conserved block is found in MSL-3 and its relatives. However, in this case, it is clear that the conserved region is much larger than the chromodomain as defined by Koonin, Zhou, and Lucchesi (1995). Moreover, PROFILESEARCH and PSI-BLAST analyses using the MSL-3-related sequences from this region failed to show any homology with other chromodomain-containing proteins. Programs that pre-

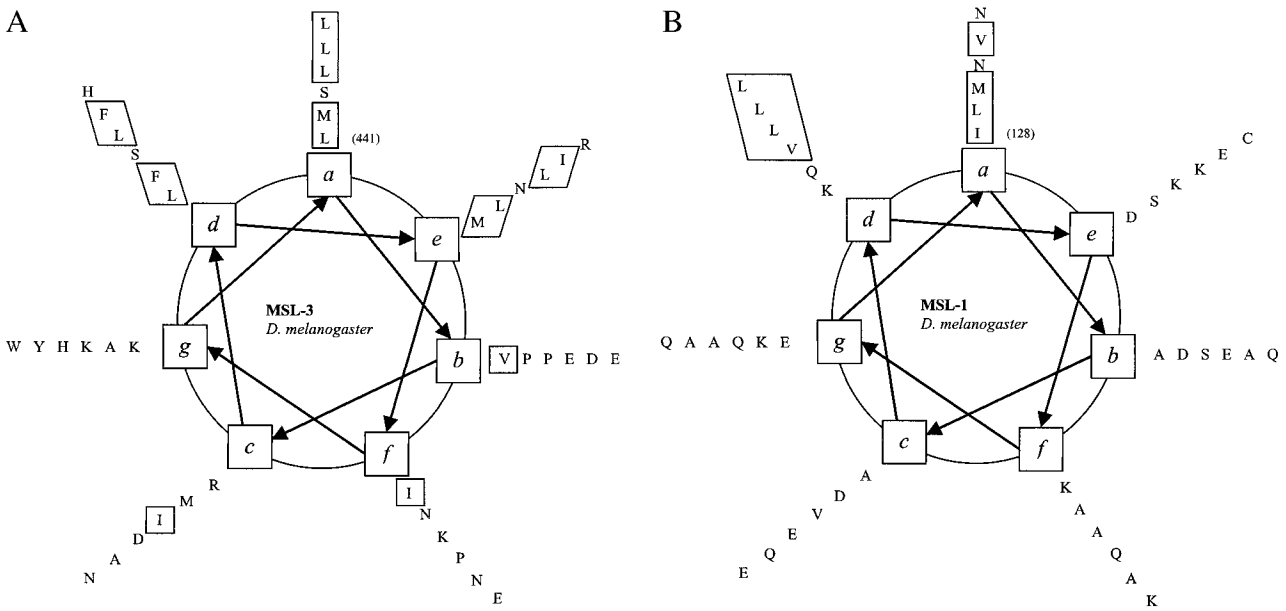

FIG. 4.—Helical wheel representation of the putative alpha helical domains with hydrophobic surfaces present in MSL-3 (*A*) and MSL-1 (*B*). For MSL-3, the conserved hydrophobic residues are boxed, while the boxed residues in MSL-1 correspond to the hydrophobic amino acids in positions *a* and *d*. The numbers in parentheses indicate the starting positions for these regions.

dict secondary structure, like nnpredict (Kneller, Cohen, and Langridge 1990) or PHD (Rost 1996), suggest that this region would form an alpha helix or two alpha helices separated by a short loop. Figure 4*A* shows the helical wheel representation of the corresponding region of *D. melanogaster* MSL-3. Similar plots can be obtained for the other MSL-3-related proteins. All of them show that similar, conserved hydrophobic residues are found in positions *a* and *d*, as expected for regions of a protein that form an alpha helix with a hydrophobic surface. This type of leucine zipper–like structure may be involved in protein-protein interactions. Although one possibility is that MSL-3 and related proteins form homodimers, we searched using PAIRCOIL (Berger et al. 1995) for putative leucine zippers in the other MSLs, finding a possible one in MSL-1 (fig. 4*B*). A short coiled coil structure was also detected in MSL-2, but the PAIRCOIL scores were lower than for the MSL-1 region.

The alignments of all of these proteins revealed two other interesting characteristics. First, the nonconserved region between conserved blocks 2 and 3 is highly charged in all of the MSL-3-related proteins. The second characteristic is the long distance between conserved blocks 4 and 5 in the MSL-3 proteins of flies and mammals. Significant similarity among these proteins can be detected in this region when they are aligned separately (not shown). This qualitative result again suggests a close relationship between the mammalian and insect genes, confirming the results obtained in the phylogenetic analysis presented in figure 3.

To investigate whether any of these proteins might be interacting with nucleic acids, and especially with the roX RNAs, we searched specifically for previously defined RNA-binding motifs. We inspected the MSL-3-related sequences for the motifs described by Burd and Dreyfuss (1994), and a similarity to KH domain-containing proteins (Gibson, Thompson, and Heringa 1993) was observed in the *D. melanogaster* MSL-3 sequence. The region that could correspond to a KH domain would start in position 470 in figure 2; that is, it would be precisely in the only region that is specific for MSL-3 and its close relatives. The highly conserved hhGXXGXXh motif typical of the KH domain, with h meaning any hydrophobic amino acid, would correspond to positions 484–492 in figure 2. However, no apparent KH domain was detected either in *D. virilis* MSL-3 or in the mammalian MSL-3 relatives.

Discussion

We found that the *msl-3* gene was part of a small gene family that was widespread in eukaryotes. As detailed in the previous section, both in mammals and in *Drosophila* there are two related genes, *msl-3* and *MRG15*. Considering their high similarities, the origin of the sequences *MORF 4* on one hand and *MSL3L1-b* on the other hand must have been relatively recent retrotranscriptions and insertions of processed mRNAs derived from *MRG 15* and *MSL3L1*, respectively (see Bertram et al. [1999] for a discussion). In spite of the fact that we found ESTs corresponding to *MSL3L1-b*, it is unclear at present whether it corresponds to a functional gene (such as *MORF4*) or to a processed pseudogene with some residual transcriptional activity. *MRGX* contains introns, and thus it can be interpreted as a conventional duplication of *MRG15*, which arose quite recently and with a truncated N-terminal end.

There are two hypotheses to explain why both plants and animals have duplicated genes, while yeasts do not. First, it is possible that there was a single duplication preceding the plants/animals/fungi split and that yeast have subsequently lost one of the two dupli-

cates. Alternatively, the duplications in plants and animals may be independent, with the plant duplication occurring after the plants-versus-fungi/metazoans split and the duplication found in animals occurring after the fungi/metazoans split. We favor this second hypothesis, considering the high similarity of the plant paralogs. If they were produced by a duplication before the split of plants with fungi and animals, their degree of differentiation should be very high, branching in the inner part of the tree. Thus, we conclude that the simplest explanation for our data is that the plant paralogs have been duplicated relatively recently, while *msl-3* and *MRG15* are older duplicates, probably produced soon after the fungi/animals split. Figure 3 also suggests that *Sc Alp13* is the *Alp13* ortholog in *Saccharomyces.* This is likely, considering that the *S. cerevisiae* genome has been completely sequenced and no other gene has shown any similarity to the *S. pombe Alp13* gene.

We have data to suggest that *msl-3* evolved at a higher rate than *MRG15* after the *msl-3*/*MRG15* duplication. This hypothesis is supported by two independent lines of evidence. First, the tree presented in figure 3, which is based only on the five conserved motifs, shows that the distance between the *Drosophila* and mammalian *MRG 15* genes is shorter than the distance between the mammalian and *Drosophila msl-3* genes. Second, there are qualitative features, in regions not included in the analysis that provided the phylogenetic tree in figure 3, that are characteristic of the *msl-3* gene in *Drosophila* and mammals: the very long stretch of amino acids between the conserved regions 4 and 5 (see fig. 2), as well as the characteristic five additional amino acids in the middle of conserved motif 1. These changes might be related to the acquisition of new functional roles in the *msl-3* genes. Recent data suggest that, contrary to the theoretical expectations that most duplicated genes should accumulate mutations and be lost, acquisition of new functions may be frequent after gene duplication (Nadeau and Sankoff 1997).

The tree presented in figure 3, as well as the fact that the *C. elegans* sequence Ce-1 lacks the long region between motifs 4 and 5 characteristic of *msl-3* in *Drosophila* and mammals, suggests that Ce-1 corresponds to an ortholog of *MRG15.* Considering the data supporting the hypothesis that nematodes are close relatives of insects (Aguinaldo et al. 1997; de Rosa et al. 1999), we hypothesize that the *msl-3* ortholog has been lost in the nematode lineage. Results for plant genes are more ambiguous. Although the simplest hypothesis is that a duplication occurred relatively recently in plants, the topology of the tree shown in figure 3 in fact does not correspond to that expected for two paralogs (the At-1, Zm-1, and Le-1 sequences appear to be derived from the At-2 sequence). However, the fact that two of the sequences that we analyzed in plants are incomplete (Le-1 is missing the last domain, while Le-2 lacks part of it) may significantly alter the relationships among the plant sequences shown in figure 3. The relationships among these plant genes should be reconsidered when more sequences are available.

It is interesting to speculate about the possible function of the *msl-3*-related genes in other species, as well as that of *MRG 15* in *Drosophila.* The function of *MRG 15* in mammals is unknown, although it has been suggested that the related chromodomain-lacking gene *MORF 4* may be involved in cellular senescence (Bertram et al. 1999). One possibility to consider is whether *MRG 15* may also be acting in dosage compensation in fly species. However, finding *msl-3*-related genes in many species, including yeasts, that do not have a dosage compensation system, raises the possibility that these genes may perform some ancestral gene function from which the *msl-3* dosage compensation function evolved. Although the precise function of yeast gene *Alp13* is also currently unknown, it is known that ALP13 is a nuclear protein and that *Alp13* mutants are viable but sterile and lead to alterations of cell shape (Radcliffe et al. 1998; Radcliffe and Toda, personal communication). At another level, the description of *msl-3*-related genes in mammals opens up the possibility of studying their possible interactions with the previously described mammalian *maleless* relative (*Helicase A*) to determine whether a similar MSL complex is present in mammalian species.

The comparison of all the *msl-3*-related sequences has allowed us to reevaluate the proposals of Koonin, Zhou, and Lucchesi (1995) as to the existence of two chromodomains in MSL-3. We conclude that only the more N-terminal chromodomain-containing regions hypothesized by Koonin, Zhou, and Lucchesi (1995) have a pattern of conserved residues in the members of this gene family that is compatible with being a chromodomain-related structure. However, the reason why this pattern was not originally detected in *D. melanogaster* MSL-3 is because that region has a very low similarity with the canonical chromodomain-containing proteins, such as POLYCOMB or HP1 (see Aasland and Stewart 1995; Singh and James 1995). Most data suggest that the chromodomain is a protein-protein interaction domain (Cavalli and Paro 1998), but it is an open question whether the chromodomain and other related sequences, like the chromo-shadow domain (Aasland and Stewart 1995) or the chromodomain-like sequences in MSL-3, actually perform the same role. On the other hand, the most C-terminal putative chromodomain-containing region hypothesized by Koonin, Zhou, and Lucchesi (1995) actually does not show the expected pattern of conserved residues in the different MSL-3-related proteins, suggesting that this region does not correspond to a chromodomain. The regular disposition of some conserved hydrophobic amino acids suggests that it may instead correspond to a leucine zipper–like structure. Thus, MSL-3 and its relatives might be able to form dimers or interact using such a surface with MSL-1 or MSL-2 (see *Results*). All of these structural features could be related to MSL-3 being a structural component of the compensasome, important for binding together proteins that provide the different enzymatic activities (e.g., helicase, acetyltransferase) required for the action of the complex. MSL-3 relatives could be performing

similar mediator functions in other, related or unrelated, protein complexes.

After all these analyses, it is still unclear whether MSL-3 and its relatives are able to interact in some way with nucleic acids or other structural components of chromatin, such as histones. Two interesting features that we have observed in these proteins which may be significant for their function in regulating chromatin structure are the highly polar region between motifs 2 and 3 and the putative KH domain in *D. melanogaster msl-3*. Many other chromatin-associated proteins, such as topoisomerase I, nucleoplasmin, and the chromodomain-containing protein SWI6, also have characteristic charged regions (Burglin et al. 1987; Lorentz et al. 1994; Shaiu and Hsieh 1998). It is possible that the differences in disposition of charges among MSL-3-related proteins detected in our analyses, for example, the longer polar domain in *D. virilis* MSL-3 than in *D. melanogaster* MSL-3, may be related to the fine tuning of the functions of those proteins in the different species. Similarly, the fact that *D. melanogaster msl-3*, but apparently not its *D. virilis* homolog or other MSL3-related proteins, has a region with some similarity to the KH domain, a type of RNA-binding domain (Gibson, Thompson, and Heringa 1993), may be a demonstration of slightly different roles of the two homologous proteins in their respective complexes. Finally, the finding by Prakash et al. (1999) of alternative splicing for mammalian *MSL3L1*, with one of the isoforms produced lacking part of the chromodomain, is interesting. Additionally, Bertram et al. (1999) described chromodomain-lacking duplicates of *MRG15*. It may be significant, then, that Gorman, Franke, and Baker (1995) found that in *Drosophila melanogaster*, there are two types of *msl-3* RNAs, and one of them would lack part of the chromodomain-like region. These results suggest that there may be, in both mammals and *Drosophila*, significant roles for the truncated versions of the MSL-3-type proteins.

Acknowledgments

The authors thank Mark Siegal and two anonymous reviewers for their comments on the manuscript. This work was supported by an NIH grant awarded to B.S.B.

LITERATURE CITED

AASLAND, R., and A. F. STEWART. 1995. The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. Nucleic Acids Res. **23**:3168–3173.

AGUINALDO, A. M., J. M. TURBEVILLE, L. S. LINFORD, M. C. RIVERA, J. R. GAREY, R. A. RAFF, and J. A. LAKE. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature **387**:489–493.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

AMREIN, H., and R. AXEL. 1997. Genes expressed in neurons of adult male *Drosophila*. Cell **88**:459–469.

BAKER, B. S., M. GORMAN, and I. MARÍN. 1994. Dosage compensation in *Drosophila*. Annu. Rev. Genet. **28**:491–521.

BASHAW, G. J., and B. S. BAKER. 1996. Dosage compensation and chromatin structure in *Drosophila*. Curr. Opin. Genet. Dev. **6**:496–501.

BERGER, B., D. B. WILSON, E. WOLF, T. TONCHEV, M. MILLA, and P. S. KIM. 1995. Predicting coiled coils by use of pairwise residue correlations. Proc. Natl. Acad. Sci. USA **92**: 8259–8263.

BERTRAM, M. J., N. G. BÉRUBÉ, X. HANG-SWANSON et al. (15 co-authors). 1999. Identification of a gene that reverses the immortal phenotype of a subset of cells and is a member of a novel family of transcription factor-like genes. Mol. Cell. Biol. **19**:1479–1485.

BONE, J. R., and M. I. KURODA. 1996. Dosage compensation regulatory proteins and the evolution of sex chromosomes in *Drosophila*. Genetics **144**:705–713.

BONE, J. R., J. LAVENDER, R. RICHMAN, M. J. PALMER, B. M. TURNER, and M. I. KURODA. 1994. Acetylated histone H4 on the male X chromosome is associated with dosage compensation in *Drosophila*. Genes Dev. **8**:96–104.

BURD, C. G., and G. DREYFUSS. 1994. Conserved structures and diversity of functions of RNA-binding proteins. Science **265**:615–621.

BURGLIN, T. R., I. W. MATTAJ, D. D. NEWMEYER, R. ZELLER, and E. M. DEROBERTIS. 1987. Cloning of nucleoplasmin from *Xenopus laevis* oocytes and analysis of its developmental expression. Genes Dev. **1**:97–107.

CAVALLI, G., and R. PARO. 1998. Chromo-domain proteins: linking chromatin structure to epigenetic regulation. Curr. Opin. Cell Biol. **10**:354–360.

CLINE, T. W., and B. J. MEYER. 1996. Vive la différence: males vs. females in flies vs. worms. Annu. Rev. Genet. **30**:637–702.

COPPS, K., R. RICHMAN, L. M. LYMAN, K. A. CHANG, J. RAMPERSAD-AMMONS, and M. I. KURODA. 1998. Complex formation by the *Drosophila* MSL proteins: role of the MSL2 RING finger in protein complex assembly. EMBO J. **17**: 5409–5417.

DE ROSA, R., J. K. GRENIER, T. ANDREEVA, C. E. COOK, A. ADOUTTE, M. AKAM, S. B. CARROLL, and G. BALAVOINE. 1999. *Hox* genes in brachiopods and priapulids and protostome evolution. Nature **399**:772–776.

FRANKE, A., and B. S. BAKER. 1999. The *rox1* and *rox2* RNAs are essential components of the ''compensasome'' that mediates dosage compensation in *Drosophila*. Mol. Cell **4**: 117–122.

GIBSON, T. J., J. D. THOMPSON, and J. HERINGA. 1993. The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding to nucleic acid. FEBS Lett. **324**:361–366.

GORMAN, M., A. FRANKE, and B. S. BAKER. 1995. Molecular characterization of the *male-specific lethal-3* gene and investigations of the regulation of dosage compensation in *Drosophila*. Development **122**:463–475.

GRIBSKOV, M., A. D. MCLACHLAN, and D. EISENBERG. 1987. Profile analysis: detection of distantly related proteins. Proc. Natl. Acad. Sci. USA **84**:4355–4358.

GUBENKO, J. S., and M. B. EVGENEV. 1984. Cytological and linkage maps of *Drosophila virilis* chromosomes. Genetica **65**:127–139.

HILFIKER, A., D. HILFIKER-KLEINER, A. PANNUTI, and J. C. LUCCHESI. 1997. *mof*, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in *Drosophila*. EMBO J. **16**:2054–2060.

KELLEY, R. L., V. H. MELLER, P. R. GORGADZE, G. ROMAN, R. L. DAVIS, and M. I. KURODA. 1999. Epigenetic spreading

of the *Drosophila* dosage compensation complex from *roX* RNA genes into flanking chromatin. Cell **98**:513–522.

KELLEY, R. L., I. SOLOVYEVA, L. M. LYMAN, R. RICHMAN, V. SOLOVYEV, and M. I. KURODA. 1995. Expression of *msl-2* causes assembly of dosage compensation regulators on the X chromosomes and female lethality in *Drosophila.* Cell **81**:867–877.

KNELLER, D. G., F. E. COHEN, and R. LANGRIDGE. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. **214**:171–182.

KOONIN, E. V., S. ZHOU, and J. C. LUCCHESI. 1995. The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. Nucleic Acids Res. **23**:4229–4233.

KURODA, M. I., M. J. KERNAN, R. KREBER, B. GANETZKY, and B. S. BAKER. 1991. The maleless protein associates with the X chromosome to regulate dosage compensation in *Drosophila.* Cell **66**:935–947.

KWIATOWSKI, J., D. SKARECKY, K. BAILEY, and F. J. AYALA. 1994. Phylogeny of Drosophila and related genera inferred from the nucleotide sequence of the Cu,Zn Sod gene. J. Mol. Evol. **38**:443–454.

LEE, C.-G., K. A. CHANG, M. I. KURODA, and J. HURWITZ. 1997. The NTPase/helicase activities of *Drosophila maleless,* an essential factor in dosage compensation. EMBO J. **16**:2671–2681.

LEE, C.-G., and J. HURWITZ. 1993. Human RNA helicase A is homologous to the maleless protein of *Drosophila.* J. Biol. Chem. **268**:16822–16830.

LORENTZ, A., K. OSTERMANN, O. FLECK, and H. SCHMIDT. 1994. Switching gene *swi6,* involved in repression of silent mating-type loci in fission yeast, encodes a homologue of chromatin-associated proteins from *Drosophila* and mammals. Gene **27**:139–143.

LUCCHESI, J. C. 1998. Dosage compensation in flies and worms: the ups and downs of X-chromosome regulation. Curr. Opin. Genet. Dev. **8**:179–184.

———. 1999. Dosage compensation: roX marks the spot. Curr. Biol. **9**:R807–R808.

MARÍN, I., A. FRANKE, G. J. BASHAW, and B. S. BAKER. 1996. The dosage compensation system of *Drosophila* is co-opted by newly evolved X chromosomes. Nature **383**:160–163.

MELLER, V. H., K. H. WU, G. ROMAN, M. I. KURODA, and R. L. DAVIS. 1997. *roX1* RNA paints the X chromosome of male *Drosophila* and is regulated by the dosage compensation system. Cell **88**:445–457.

NADEAU, J. H., and D. SANKOFF. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics **147**:1259–1266.

NICHOLAS, K. B., and H. B. NICHOLAS JR. 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the authors (www.cris.com/~ketchup/genedoc.shtml).

O'NEIL, M. T., and J. M. BELOTE. 1992. Interspecific comparison of the *transformer* gene of Drosophila reveals an unusually high degree of evolutionary divergence. Genetics **131**:113–128.

PAGE, R. D. M. 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12**:357–358.

PRAKASH, S. K., I. B. VAN DEN VEYVER, B. FRANCO, M. VOLTA, A. BALLABIO, and H. Y. ZOGHBI. 1999. Characterization of a novel chromo domain gene in Xp22.3 with homology to *Drosophila msl-3.* Genomics **59**:77–84.

RADCLIFFE, P., D. HIRATA, D. CHILDS, L. VARDY, and T. TODA. 1998. Identification of novel temperature-sensitive lethal alleles in essential β-tubulin and nonessential α2-tubulin genes as fission yeast polarity mutants. Mol. Biol. Cell **9**:1757–1771.

ROST, B. 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol. **266**:525–539.

RUSSO, C. A. M., N. TAKEZAKI, and M. NEI. 1995. Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. **12**:391–404.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SHAIU, W. L., and T. S. HSIEH. 1998. Targeting to transcriptionally active loci by the hydrophilic N-terminal domain of *Drosophila* DNA topoisomerase I. Mol. Cell. Biol. **18**:4358–4367.

SINGH, P. B., and T. C. JAMES. 1995. Chromobox genes and the molecular mechanisms of cellular determination. Pp.71–108 *in* R. OHLSSON, K. HALL, and M. RITZEN, eds. Genomic imprinting. Causes and consequences. Cambridge University Press, Cambridge, England.

SMITH, E. R., A. PANNUTI, W. GU, A. STEURNAGEL, R. G. COOK, C. D. ALLIS, and J. C. LUCCHESI. 2000. The Drosophila MSL complex acetylates histone H4 at lysine 16, a chromatin modification linked to dosage compensation. Mol. Cell. Biol. **20**:312–318.

STUCKENHOLZ, C., Y. KAGEYAMA, and M. I. KURODA. 1999. Guilt by association. Non-coding RNAs, chromosome-specific proteins and dosage compensation in Drosophila. Trends Genet. **15**:454–458.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, and D. G. HIGGINS. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research **24**:4876–4882.

TURNER, B. M., A. J. BIRLEY, and J. LAVENDER. 1992. Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in *Drosophila* polytene nuclei. Cell **69**:375–384.

WHITING, J. H., M. D. PLILEY, J. L. FARMER, and D. E. JEFFERY. 1989. *In situ* hybridization analysis of chromosomal homologies in *Drosophila melanogaster* and *Drosophila virilis.* Genetics **122**:99–109.

CHARLES AQUADRO, reviewing editor

Accepted April 18, 2000