**D. melanogaster Dumpy**

**Figure S1**  Three large multi-exon MPOs in *D. melanogaster dumpy*. The *D. melanogaster* dp gene (top) is shown with the MPOs (orange boxes) shown above the known coding exons (blue arrows). The PigFeast repeat region and the exon 37 to 70 region are indicated by bars. An expanded view of the exon 37 to 70 region (bottom), which contains 34 coding exons in 13 MPOs or reading frames. Incomplete sequences in the other species made a complete analysis impossible but a similar ratio of exons to MPOs was found in *D. virilis*, although the arrangement of corresponding exons is different between species.
**Figure S2**  Annotation of *Drosophila* MUD-PL. The manual annotation of the MUD-PL splice variant from exon 2 to 11 is shown for eight species with the MPOs (orange boxes) above the coding exons (blue arrows) below. The first coding exon could not be assigned with high confidence in at least one species. The numbers on the right indicate the gene’s genomic extent (top) and the translated protein size (bottom) for each model, the species is on the left. The gene models are to scale and are aligned on the left at the 3’ terminal stop codon. The target MPO identified in our screen for long MPOs in *D. melanogaster* is shaded in gray. The asterisk downstream of the target MPO is a clear intron loss in *D. melanogaster* and *D. erecta*. Interestingly this exon in *D. melanogaster* encodes a conserved functional motif in the protein.

Abbreviations: (Dmel: *D. melanogaster*; Dere: *D. erecta*; Dyak: *D. yakuba*; Dana: *D. ananassae*; Dpse: *D. psuedoobscura*; Dwil: *D. willistoni*; Dmoj: *D. mojavensis*; Dvir: *D. virilis*; Dgri: *D. grimshawi*)
Figure S3  Annotation of Drosophila Rno-PB. The manual annotation of the entire RNO-PB splice variant is shown, labeled as in Supplemental Figure 2. All the variation in the protein size between species is due to many indels within the large terminal coding exon identified in our screen.
Figure S4  Annotation of Drosophila Lva-PC. The manual annotation of exons 2 to 8 of the LVA-PC splice variant is shown, labeled as in Supplemental Figure 2. The first coding exon could not be assigned with high confidence in at least one species. In D. pseudobscura the two coding exons 5' of the terminal exon are interrupted by a single stop codon created by a non-triplet indel, which may be due to a sequencing error. All other splice sites were determined by their alignment with the known D. melanogaster cDNA and will require future RNA sequencing to determine the actual transcripts present in each species.
**Figure S5** Annotation of *Drosophila* Futsch-PC. The manual annotation of exons 3 to 8 of the Futsch-PC splice variant is shown, labeled as in Supplemental Figure 2. The first two coding exons could not be assigned with high confidence in at least one species. As in RNO-PB, the protein size variation is due to frequent indels in the exons corresponding to the single large coding exon associated with our target MPO in *D. melanogaster*. The single breaks in the large exons in *D. mojavensis* and *D. grimshawi* are due to a single base substitution that introduces an in-frame stop codon, probably due to sequencing errors. The next to last exon in species below *D. ananassae* are present but we could not predict where the potential 5′ intron would be in actual transcripts.
Supplemental Text

Our screen for extensive MPOs as compared to coding exon length identified several genes in *D. melanogaster*. Although we limited our analyses to genes with MPOs >3000 nt longer than an associated coding exon, it is likely the evolution of genes with less extensive MPOs will be interesting as well. For example, the largest MPO in *cnn* is only 1000 nt longer than one of the associated coding exons. Additionally, genes with significant variation in their MPO maps across some phylogenetic range should be good candidates for genomic and gene/protein evolutionary studies. Below is a very brief description of the seven genes we have manually annotated in nine *Drosophila* species.

The *dump* (*dp*) gene makes maximum use of multi-exon MPOs and contained the longest excess beyond an annotated exon end (14,000 nt). The *dp* locus spans over 100 kb, encoding a 2.5 MDa modular extracellular matrix protein with 308 EGF modules and 185 DPY modules required for the anchoring of epidermal tissue to the fly cuticle (Wilson *et al.* 2000). The 79 *dp* coding exons are in 53 MPOs, and the region from exon 37 to 70 has 34 coding exons in 13 MPOs (Supplemental Figure 1). Due to extensive repeats within the gene the exact changes in coding exon to MPO organization within the genus *Drosophila* are difficult to determine precisely, all species have a similar multi-exon MPO arrangement. A recent study of *dp* found the area containing the greatest concentration of multi-exon MPOs is alternatively spliced, although the actual splicing is unknown (Carmon *et al.* 2010). Because many *dp* genes in the genus *Drosophila* appear to have assembly errors creating large deletions, possibly due to the highly repetitive nature of *dp* protein modules, we were not able to do a complete analysis of the divergence of target MPOs.

The target MPOs in four of the genes identified, *mushroom body defective* (*mud*) (Supplemental Figure 2), *rhinoceros* (*rno*) (Supplemental Figure 3), *java lamp* (*lva*) (Supplemental Figure 4), and *Futsch* (Supplemental Figure 5), span a small and large coding exon in *D. melanogaster*. Interestingly, the ab initio models for these four genes are variable in the genus *Drosophila*, including a split of the gene into two separate transcription units and the elimination of relatively large regions of coding sequence. Most of the variability in these models is associated with the region covered by our target MPOs, suggesting these coding regions confound modeling algorithms, possibly due to a lack of strong splicing signals. However, the manual annotation of these genes finds little evidence to support significant changes relative to known transcripts from *D. melanogaster*.

The *rno* locus encodes a transcription factor predicted to be a chromatin-remodeling protein required to restrict the Ras pathway during eye development in *D. melanogaster* (Voas & Rebay, 2003). The target MPO encodes a long (3929 aa) coiled-coil peptide and is present in all *Drosophila* species used in this study, as are the five amino terminal coding exons. Based on these results there is no clear explanation for the significant variations in gene models.

The *lva* gene encodes a rapidly evolving coiled-coil golgin protein required for Golgi vesicle transport and cellularization during embryogenesis in *D. melanogaster* (Sisson *et al.* 2000). The *lva* target MPO spans exon 4 and exon 5 in the Lva-PC isoform but the intron between these exons is retained in an alternatively spliced variant (Sisson *et al.*, 2000). The only evidence we see for model changes in the genus involve fusions with the downstream adjacent exon and the possible loss of an intron. However, based on the exon6/7 splicing in *cnn* (described in the main text) accurate models will require sequencing transcripts or an analysis of RNA-Seq data. Nevertheless even in the absence of cDNA or RNA-Seq data all *Drosophila lva* gene models appear to be similar to *D. melanogaster*. 

R. C. Eisman and T. C. Kaufman
The *futsch* locus encodes a cytoskeletal protein required for the development of axons and dendrites in *D. melanogaster* (HUMMEL et al. 2000), and organizes the microtubule cytoskeleton during synaptic growth (ROOS et al. 2000). The Futsch protein appears to be a chimera of vertebrate genes, comprised of vertebrate MAP1B-like amino- and carboxy-termini separated by a repetitive central region similar to vertebrate neurofilament proteins (HUMMEL et al. 2000). The central region contains 60 repeats of approximately 37 amino acids each that may be phosphorylation targets (HUMMEL et al. 2000) and are encoded by Futsch-PB exon 6 which is in our target MPO. Exon 6 is a continuous exon in all species except *D. mojavensis* and *D. grimshawi*, which both have single base changes creating a stop codon, and *D. yakuba* appears to have a small insertion and stop codon. While these changes may be real, since they occur in the highly repetitive sequence it is also possible they are sequencing or assembly errors. Surprisingly, even though this exon is similar in size to *D. melanogaster* in the other species, the gene models for several species remove approximately half of the exon from the 5' end.

Of the four genes analyzed with multi-exon MPOs, *mud* has the most complex and potentially exhibits real changes in splicing. The *mud* gene encodes a microtubule binding protein required at the meiosis II spindle in *Drosophila* embryos (YU et al. 2006), and is required for the proliferation (GUAN et al. 2000) and asymmetric division of neuroblast cells (IZUMI et al. 2006; SILLER et al. 2006). Mud protein is predicted to be functional ortholog of mammalian NuMA based on conserved domains in the carboxy-terminus of the *D. melanogaster* protein (SILLER et al. 2006). Our target MPO spans the Mud-PB large exons 6 and 7, encoding a coiled-coil domain (GUAN et al. 2000). Based on protein alignments it appears the coding portions of the exons are fused in *D. yakuba*, *D. erecta*, *D. pseudoobscura* and *D. grimshawi* as removal of an intron would result in a significantly smaller protein compared to other species in this analysis. Additionally, these proteins are several hundred amino acids shorter than species with the intron. In *D. mojavensis* there are two exons but the 3’ end of exon 6 is fused to exon 7, producing a much smaller peptide. Finally, in *D. virilis* there is a short fused single coding exon and an upstream insertion of approximately 5 kb, bearing no sequence similarity to any of the *mud* genes in this study. Our *mud* target MPO identified the region of Mud protein that led others to identify *D. melanogaster* as a functional ortholog of NuMA rather than an orthologous gene. In their study they showed conserved domains in Mud retained NuMA function but the coiled-coil region had no similarity to vertebrate NuMA (GUAN et al. 2000), which are 80% identical between human and mouse (WHITE and ERICKSON 2006). While our results neither strengthen nor weaken orthology arguments they do suggest the evolution of coiled-coil domains may be more plastic in *Drosophila* compared to vertebrates.

The last two genes in our screen were identified because the target MPOs spanned exons known to be alternatively spliced by an intron retention/exclusion mechanism. The *prospero* (*pros*) gene encodes a transcription factor required for cell fate specification during central nervous system development in *D. melanogaster* (CHU-LAGRAFF et al. 1991). Orthologous genes of the *Pros/Prox1* family all contain a conserved COOH terminus which contains the structurally-unique Homeo-Prospero domain (RYTER et al. 2002), and motifs required for the complex regulation of the protein (BI et al. 2003). The amino half of *D. melanogaster* Pros protein contains two small conserved motifs present in chicken Prox1, but shares no other similarities upstream of the homeodomain (TOMAREV et al. 1996). The *pros* target MPO spans the exon encoding the amino half of Pros and accounts for the majority of the changes between *Drosophila* species. Protein alignments between *Drosophila* species show the encoded peptide is a mix of high conservation interspersed with regions of nearly 100% divergence and gaps accumulate with phylogenetic distance, similar to all MPOs identified in the screen.

The last gene in this study is *shortstop* (*shot*), a transcriptionally complex gene, which is a member of the Plakin protein family. Shot proteins link the actin and microtubule cytoskeleton to a variety of junctional complexes in different cell types (SONNENBERG and LIEM 2007) required for cytoskeleton organization during neurogenesis and in muscles and tendon cells.
SUBRAMANIAN et al. 2003). The central region of Shot long isoforms form a long coiled coil (LEE et al. 2000), a region partially covered by the shot MPO. Alternative 3’ splicing of our target MPO in shot produces eighteen mRNAs that include the 5’ terminus of the exon encoding a conserved peptide (154 aa), whereas the Shot-RH mRNA includes the entire exon encoding a large (3499 aa in D. melanogaster) divergent peptide. Three Shot splice variants skip this exon entirely. Interestingly, 27 exons common to all isoforms are highly conserved in Drosophila and associated MPOs are similar to coding sequences in length. This includes an adjacent exon immediately downstream of the target MPO encoding a coiled-coil domain (2493 aa) that is 94% identical to D. virilis Shot and contains no gaps in alignments with other Drosophila species. This is in stark contrast to the long target MPO peptide, which is 78% identical in D. virilis and randomly varies in length from 3070 to 3498 amino acids in the genus Drosophila. This result shows the rapid divergence typical of the MPOs identified in our screen can occur independent of the surrounding gene sequence.


