

Exon-Intron Structure in the Dipterans

The abundance of SDMA homologs in *A. aegypti* allows a close examination of the gene structure within this species. The most common exon-intron structures are shown in Supplemental Fig. 1a. Within the three exon/two intron class of SDMA genes, based on amino acid sequence conservation and position, there are three different conserved groups of 1st-2nd exon, and 2nd-3rd exon splice junctions (Supplemental Fig. 1b,c) suggesting three groups of paralogs. One gene of this class (AAEL013584) has a unique structure. The group 1 1st-2nd exon and 2nd-3rd exon junctions are also conserved between four *A. aegypti* and two *C. quinquefasciatus* and the group 1 2nd-3rd exon junction is, in addition to these, conserved in the *C. quinquefasciatus* CPIJ012846 and *A. gambia* AGAP006187 SDMA genes, suggesting an orthologous relationship of this group within Dipterans. This relationship is also suggested by the phylogenetic tree shown in Supplemental Fig. 13. Interestingly, two of the *C. quinquefasciatus* genes within this node are not of the three exon two-intron group 1 type. CPIJ012846-RA is of the two-exon one-intron IIa type (Supplemental Fig. 1a) and could be derived from a loss of the first intron of the group 1 three exon-two intron type based on the amino acid conservation between the 2nd-3rd exon splice site of CPIJ012844-RA and CPIJ012845-RA and the single splice junction of CPIJ012846-RA (Supplemental Fig. 1c). The other *C. quinquefasciatus* SDMA gene in this node, CPIJ012848-RA, is intronless suggested a loss of both introns. The other two groups of three-exon two-intron SDMA genes appear to be exclusive to *A. aegypti*. The β M monomer, AAEL010436-RA, clearly falls within the group 2 set of three-exon two intron genes with the 2nd-3rd exon junction showing complete identity to the others in this group although its 1st-2nd exon junction is entirely novel.

Two types of two-exon one-intron SDMA gene structures are seen repeatedly in Dipterans (types IIa and IIb in Supplemental Fig. 1a). Type IIa exists in three Dipteran species, *A. aegypti* alone has five, with two (AAEL001611-RA and AAEL001621-RA) being adjacent on the same strand and within 10 kb and as discussed in the main text, likely the result of a tandem duplication. *C. quinquefasciatus* has two of these (CPIJ005176-RA and CPIJ012846-RA) while *D. plexippus* has one (EHJ67085-RA). Supplemental Fig. 13 suggests that AAEL001611, AAEL001621 and CPIJ005176 are orthologs and that the type IIa structure arose earlier in the Dipteran lineage. The tree also suggests some orthologous relationships among Dipteran species that were not as well supported by the type of exon-intron structures discussed above. One relationship not suggested by sequence conservation is for AAEL013585, AAEL010438, and CPIJ012843, all of the IIb type, with these three having a sole splice junction immediately after the first methionine of each gene. *A. mellifera* has two of these and *C. quinquefasciatus* and *T. castaneum* one each. Although sequence conservation does not suggest orthology, Supplemental Fig. 13 suggests that AAEL001418, AAEL005340, and CPIJ011431 are orthologs also. Several Dipteran genomes contain intronless copies of SDMA. *C. quinquefasciatus* (CPIJ012848) each contain one while *D. melanogaster* (FBpp0072973 and FBpp0288411) and *A. gambia* (AGAP007557 and AGAP007414) have two. These show no evidence of being retrocopied

mRNAs. The SDMA genes with an N terminal extension exclusively occur in Dipterans, all are 4 exon/3 intron types, but there is little apparent conservation between the species at the splice junctions. Within the Lepidopteran species examined there is also conservation in splice site junctions (Fig. 1d), but the locations of the introns is distinct from that in the Dipterans.

The phylogenetic tree in Supplemental Fig 13 of all of the SDMA genes in our collection recapitulates the Hymenopteran-specific tree (box A) and suggests that SDMA genes are largely grouped by specific Orders. Exceptions to this do occur, as one *G. mortisans* and three *D. melanogaster* SDMA genes group closer to Lepidopteran SDMA genes (box B) and with two outliers from the Hymenoptera, GB10225 of *A. mellifera* and HSAL21866 of *H. saltator*. Only three of the *D. melanogaster* genes group with the Dipterans (box C). The organization of the Dipteran part of the tree suggests several orthologous groups containing one or two SDMA genes from any one given species that have not been discussed above. In only one of these Dipteran nodes (highlighted in red) does there seem to be a significant expansion of SDMA orthologs, this node accounts for all of the unusual expansion of SDMA genes seen in *A. aegypti* and a smaller expansion that is seen in *C. quinquefasciatus* relative to the other mosquitos.

SDMA homologs in the Drosophila Genus

The predicted proteins from the twelve Drosophila genomes available at Flybase (www.flybase.org) were downloaded and searched as described in Materials and Methods for proteins containing the SDMA domain. Overall there are 87 SDMA genes within these twelve species and none are NDMA proteins. As in other Dipteran genomes, some of these have an N terminal extension (at least one in each species) and only eight do not have a predicted signal peptide (Supplemental Table). All these species have at core of 7 SDMA genes which clear orthologous relationships, the only outliers being *Drosophila yakuba* with 8 and *Drosophila grimshawi* with 9 SDMA genes. The phylogenetic tree in Figure 14 shows the orthologous relationships between the individual SDMA genes in each species. There are seven well supported groups of orthologs, A-G. The two additional SDMA genes in *D. grimshawi* fall in groups A and F and the additional SDMA gene in *D. yakuba* is within Group G, but very divergent from the other SDMA genes in this group. Figure 15 notes the location of these orthologous groups with the *D. melanogaster* chromosomes as a reference. Fig. 16 a-c shows the syntenic relationships of three of these ortholog groups (A,B, and C) that are closely linked to each other on chr2R within 40-50 kb in each species.

Supplemental Figure Legends

Figure 1. Clustering of SDMA genes of *A. aegypti*. The figure depicts the relative location and orientation of the SDMA genes on each of the 6 supercontigs in the *A. aegypti* genome assembly. The color scheme indicates the exon-intron structure. Numbering within arrows indicates which group each of the three exon-two intron genes belong to (see Figure 2 in the main text).

Figure 2. Conserved exon-intron junctions of multi-exon SDMA genes. a) Five exon-intron structures are found in multiple species. Three different 3 exon-2 intron structures, designated Group 1-3 and two differing 2 exon-1 intron structures designated IIa and b are shown. The 20 amino acids surrounding conserved exon-intron structures are shown in b-e. b) 1st and 2nd exon junctions are shown, second exons are colored blue and the red residue is an amino acid whose codon spans the intronic junction. Three groups of 3 exon-2 intron structures, based on conserved amino acid similarity are seen. c) 2nd and 3rd exon junctions of the three groups are shown, again, the 2nd exon colored in blue. d) The two conserved two exon structures. The exon junctions are colored black/blue (instead of blue/black) represent the single junction. e) A single type of 3 exon-2 intron junction is conserved in several Lepidopterans, again, the 2nd exon is colored in blue and the 20 aa around each exon junction is shown.

Figure 3. AAEL010429-AAEL010431 region alignments. a) A clustalw alignment of AAEL010429 and AAEL010431 and adjacent sequence 1 kb upstream and downstream. Yellow highlighting indicates sequence identity. Location of the open reading frame for each gene is denoted by the black, underlined text. b) A clustalw alignment of 1 kb of sequence immediately downstream of the stop codon of AAEL010429, AAEL010431 and AAEL010436. Yellow highlighting indicates sequence identity in all three genes. Blue highlighting indicates conservation between 2/3 of the genes at each residue.

Figure 4. AAEL001611-AAEL001621 alignment. A clustalw alignment of AAEL001611 and AAEL001621 and adjacent sequence 1 kb upstream and downstream. Yellow highlighting indicates sequence identity. Location of the open reading frame for each gene is denoted by the black, underlined text.

Figure 5. AAEL013577-PA-AAEL013577-PB alignment. A clustalw alignment of AAEL013577-PA and AAEL013577-PB and adjacent sequence (only 21 bp upstream due to a gap in the genome) and 1 kb downstream. Yellow highlighting indicates sequence identity. Location of the open reading frame for each gene is denoted by the black, underlined text.

Figure 6. Blow up of the duplications in Figures 3 & 4. a) The 61 kb region common to c1.374 and c1.793 is shown. Highlighted in red is the 16 kb region containing the SDMA homologs. b) The 178 kb region common to c1.477 and c1.875 is shown. Highlighted in red is the 111 kb region containing the SDMA homologs.

Figure 7. Full dot plot analysis of c1.374 and c1.793. Dotplot is of the entire supercontig with the coordinates of c1.374 on the x axis and c1.793 on the y axis. Blue lines are regions of identity on the same strand and red lines are regions of identity on the opposite strands (inversions). The red box is the duplication involving the SDMA genes, a blow up of this is in Figure 5a.

Figure 8. Full dot plot analysis of c1.477 and c1.875. Dotplot is of the entire supercontig with the coordinates of c1.477 on the x axis and c1.875 on the y axis. Blue lines are regions of identity on

the same strand and red lines are regions of identity on the opposite strands (inversions). The red box is the duplication involving the SDMA genes, a blow up of this is in Figure 5b.

Figure 9. Gene clusters in the Formicidae. Linkage of SDMA and 2DMA genes in the seven Formicidae species examined is depicted, genes colored similarly are likely orthologs based on the phylogenetic analysis of these genes. The 2DMA orthologs are shaded in blue.

Figure 10. Phylogenetic analysis of the Formicidae SDMA and 2DMA genes. Colored dots in the tree correspond to the color of the genes depicted in Fig. 7. Numbers at each of the nodes represent posterior probabilities from the two separate runs of the tree. Group A and B represent the likely duplication history of these genes (see Fig. 10, below for a model). Gene ID abbreviations are as follows: ACEP (*Atta cephalotes*), AECH (*Acromyrmex echinator*), CFLO (*Camponotus floridanus*), HSAL (*Harpegnathos saltator*), LH (*Linepithema humile*), PB (*Pogomyrmex barbatus*), SINV (*Solenopsis invicta*).

Figure 11. A model for the evolution of the SDMA gene cluster in the Formicidae. Shown is one possible model based on the phylogenetic history of these homologs. Step one is an initial duplication and inversion of an ancestral SDMA gene. The order of the next steps is unknown but in this scenario, step two is a duplication of the yellow homolog to a distantly linked location in the genome. The generation of the 2DMA gene necessitates step 3, a duplication of the yellow gene followed by step 4, an intragenic tandem duplication of this SDMA gene to generate the 2DMA gene.

Figure 12. RNAseq analyses of SDMA gene expression in *A. aegypti*. a) A time course of gene expression at various life stages. FPKM (fragments per kb per million reads) values for each SDMA homolog are shown on the y axis. SDMA genes are organized by evolutionary relatedness described in Fig. 4. AAEL010436-RA (β monomer) and AAEL010429-RA (AEG12) are highlighted. Control genes (ribosomal S13 and AAEL017558-RA) are shown at right.

Figure 13. Phylogeny of SDMA genes In Insects. The tree was built by Bayesian inference, the numbers at each node indicate the posterior probability at that node. The shaded boxes represent the three main Orders represented; A (Hymenoptera), B (Lepidoptera), C (Diptera). Gene ID abbreviations are as follows: AAEL (*Aedes aegypti*), ACEP (*Atta cephalotes*), ADAR (*Anopheles darlingi*), AECH (*Acromyrmex echinator*), AGAP (*Anopheles gambia*), ASTM (*Anopheles stephensi*), BGIMB (*Bombyx mori*), CFLO (*Camponotus floridanus*), CPIJ (*Culex quinquefasciatus*), EHJ (*Danaus plexippus*), FB (*Drosophila melanogaster*), GB (*Apis mellifera*), HMEL (*Heliconius melepomene*), HSAL (*Harpegnathos saltator*), LH (*Linepithema humile*), PB (*Pogomyrmex barbatus*), SINV (*Solenopsis invicta*), TCOGS (*Tribolium castaneum*), TMP (*Glossina mortisans*).

Figure 14. Phylogeny of SDMA genes in Drosophilae. The tree was built by Bayesian inference, the numbers at each node indicate the posterior probability at that node. The seven orthologous groups of SDMA genes (A-G) are noted. Circled in red are the two likely duplications of SDMA

genes in *D. grimshawi* and the one duplication in *D. yakuba*. For clarity, the FBpp prefix for the IDs for each gene have been replaced with a species specific notation, the number does correspond to the original Flybase protein ID. Abbreviations used are *D. ananassae* (Dana), *D. erecta* (Dere), *D. grimshawi* (Dgri), *D. melanogaster* (Dmel), *D. mojavensis* (Dmoj), *D. persimilis* (Dper), *D. pseudoobscura* (Dpse), *D. sechellia* (Dsec), *D. simulans* (Dsim), *D. virilis* (Dvir), and *D. willistoni* (Dwil) *D. yakuba* (Dyak).

Figure 15. Genome Level Organization of the SDMA Genes in Drosophilae. SDMA genes are on three of the chromosome arms in this Genus. Groups A,B, and C are closely linked within the region of 12.3 Mb and 12.7 Mb on chr2R (*D. melanogaster* coordinates are used as reference). An expansion of this region is shown in Fig. 12a-c.

Figure 16. chr2R Synteny and Clustering of Group A,B, and C SDMA Genes. a,b, and c show screenshots from Flybase comparing the chr2R region of *D. melanogaster* to the eleven other species of Drosophila. Blue lines connect the orthologous genes identified in the phylogenetic analysis of these genes. Abbreviations of species are as in Figure 14.