

TARGETS OF RAPID EVOLUTION IN THE *CULEX PIPPIENS* COMPLEX  
PROTEOME AND INSECT SEX DETERMINATION CASCADE

By

DANA C. PRICE

A Dissertation submitted to the  
Graduate School-New Brunswick  
Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Entomology

Written under the direction of

Dina M. Fonseca

And approved by

---

---

---

---

New Brunswick, New Jersey

May 2015

## ABSTRACT OF THE DISSERTATION

Targets of rapid evolution in the *Culex pipiens* complex proteome  
and insect sex determination cascade

By DANA C. PRICE

Dissertation Director:

Dina M Fonseca

The ecology of arthropod disease vectors can greatly influence their vectorial capacity and thus both duration and severity of arboviral disease outbreaks. Elucidating the genetic factors that influence these ecological characteristics is key in developing mitigation strategies.

In Chapter 1, I report the results of a transcriptomic comparison between feral and domestic forms of the northern house mosquito, *Culex pipiens* forms *pipiens* and *molestus*. By examining the rate of nonsynonymous amino acid substitution between orthologous protein pairs, I define fast and slowly evolving genes and gene families that highlight the genetic variability in these two mosquito taxa. The results implicate genes involved in olfaction, digestion and immunity as likely constituents of the genetic component driving the dramatic differences in behavior and physiology so important to understand.

In Chapter 2, I examine the mosaicism present within the genome of *Culex pipiens pallens*, a putative hybrid mosquito resulting from a cross between *Culex pipiens* and *Culex quinquefasciatus*. Using a phylogenomic analysis, I quantify shared ancestry

between *Cx. pipiens pallens* gene sequences and those of either putative parental genome. Additionally, I identify genes and gene ontologies that show evidence of evolving at accelerated evolutionary rates among East Asian *Culex* species by calculating per-gene rates of peptide evolution, and identifying lineages with differential rates of evolution to examine how *Cx. pipiens pallens* has utilized and modified parental genes to exploit its environment and persist as a species. My results show that *Cx. pipiens pallens* and *Cx. quinquefasciatus* share a greater degree of phylogenetic affiliation and lower protein divergence than either do with *Cx. pipiens form molestus*, and that the genetic component of the *Cx. pipiens pallens* proteome assigned to *Cx. pipiens* contains genes that function in energy metabolism, cell cycle / signaling, and redox reactions; the genes assigned to *Cx. quinquefasciatus* are enriched in lipid transport function and extracellular scavenging / innate immunity.

In Chapters 3 and 4, I select a fast-evolving gene of interest, *doublesex*, from the analysis performed in Chapter 1 and describe its evolution within the *Culex pipiens* complex, and in all hexapods. *Doublesex* controls the somatic sexual fate of *Drosophila melanogaster* and may function thusly in many metazoans, including the malaria mosquito *Anopheles gambiae* and the dengue and yellow fever vector *Aedes aegypti*. In these insects, upstream genetic signaling mechanisms regulate the splicing of the *dsx* transcript to produce sex-specific peptide isoforms that ultimately differentiate male and female insects. This conserved function makes *dsx* a prime target for sterile insect technique (SIT) research. Here I provide a full-length gene sequence, with sex-specific splicing, regulatory and evolutionary analyses of the *doublesex* gene from the southern house mosquito *Culex quinquefasciatus*. I show that *Cxqdsx* maintains characters

possibly derived in the Culicine mosquitoes and present in the *Aedes aegypti dsx* gene, and retains presumably ancestral qualities present in *Anopheles gambiae (Angdsx)*. Interestingly, the *cis*-regulated splicing of *Cxqdsx* does not appear to follow either currently described mosquito model; each of the three mosquito genera maintain unique regulatory mechanisms. Additionally, using public sequence databases, I show *doublesex* to be ubiquitous in the hexapods and likely to have been present in the last common ancestor (LCA) of the group as a sex-specifically spliced multiple-copy gene.



## **Acknowledgements**

I would like to thank my advisor, Dr. Dina Fonseca, for her guidance and contributions to my research, and for always being accommodating of an advisee with a full-time job; my committee, Dr. Peter Armbruster, Dr. Karl Kjer, and Dr. Frank Carle (who spent six years encouraging me to apply to graduate school); Linda McCuiston at the Center for Vector Biology for her unsurpassed expertise in rearing mosquitoes; Andrea, for being wonderful and keeping me sane; Bowser, our dissertation puppy; Don Simonds for showing me how to pin my first butterfly; and Mom & Dad for the trips to the brook as a child. I would also like to acknowledge the New Jersey Mosquito Control Association D.M Jobbins Scholarship.

## Acknowledgement of Publication

Citations for chapters 1-4 are as follows:

Price D and DM Fonseca. 2015. Genetic Divergence between Populations of Feral and Domestic forms of a Mosquito Disease Vector assessed by transcriptomics.

*PeerJ*. 3:e807 <http://dx.doi.org/10.7717/peerj.807>

Price D and DM Fonseca. 2015. Genome mosaicism and evolutionary divergence within the Asian *Culex pipiens* complex. In preparation.

Price D, Egizi A and DM Fonseca. 2015. Characterization of the *doublesex* gene within the *Culex pipiens* complex suggests regulatory plasticity at the base of the mosquito sex determination cascade. In revision. *BMC Evolutionary Biology*.

Price D, Egizi A and DM Fonseca. 2015. *Doublesex* is ubiquitous and ancestral in the insects. In revision. *Scientific Reports*.

## Table of Contents

Abstract.....	ii
Acknowledgements.....	v
Acknowledgement of Publication.....	vi
Table of Contents.....	vii
List of Tables.....	viii
List of Figures.....	xii
Introduction.....	1
Chapter 1 – Genetic Divergence between Populations of Feral and Domestic forms of a Mosquito Disease Vector assessed by transcriptomics.....	8
Chapter 2 – Genome mosaicism and evolutionary divergence within the Asian <i>Culex</i> <i>pipiens</i> complex.....	50
Chapter 3 – Characterization of the doublesex gene within the <i>Culex pipiens</i> complex underscores plasticity at the base of the mosquito sex determination cascade.....	80
Chapter 4 – <i>Doublesex</i> is ubiquitous and ancestral in the insects .....	134
Literature Cited.....	160

## List of Tables

Table 1.S1	Observed and estimated Ka calculations, annotation and top-scoring Pfam IDs corresponding with 11,931 pairwise <i>Culex pipiens</i> forms pipiens and molestus homologous codon sequence alignments (ordered by decreasing Ka). Columns two and three denote genes present in the 95th percentile as ranked by Ka calculated using observed and likelihood estimated non-synonymous substitutions, respectively.....	35
Table 1.1	Gene ontology terms enriched in the upper 95th percentile of pairwise dN values calculated using <i>Culex pipiens</i> forms pipiens and molestus homologous codon sequence alignments.....	36
Table 1.S11	BLASTN output detailing the 3,687 <i>Culex quinquefasciatus</i> CDS sequences with at least one BLASTN alignment $\geq 200$ bp at $\geq 95\%$ similarity to another CDS in the genome.....	37
Table 1.S2	Gene set composing the serine-type endopeptidase ontology, found to be enriched in the 95th percentile of top-scoring <i>Culex pipiens</i> forms pipiens and molestus homologous codon sequence alignments as ranked by Ka value.....	38
Table 1.S3	Gene set composing the proteolysis ontology, found to be enriched in the 95th percentile of top-scoring <i>Culex pipiens</i> forms pipiens and molestus homologous codon sequence alignments as ranked by Ka value.....	39
Table 1.S4	Gene set composing the receptor binding ontology, found to be enriched in the 95th percentile of top-scoring <i>Culex pipiens</i> forms pipiens and molestus homologous codon sequence alignments as ranked by Ka value.....	40
Table 1.S5	Gene set composing the odorant binding ontology, found to be enriched in the 95th percentile of top-scoring <i>Culex pipiens</i> forms pipiens and molestus homologous codon sequence alignments as ranked by Ka value.....	41
Table 1.S6	Gene set composing the extracellular space ontology, found to be enriched in the 95th percentile of top-scoring <i>Culex pipiens</i> forms pipiens and molestus homologous codon sequence alignments as ranked by Ka value.....	42
Table 1.S7	Gene set composing the chitin binding ontology, found to be enriched in the 95th percentile of top-scoring <i>Culex pipiens</i> forms pipiens and	

	molestus homologous codon sequence alignments as ranked by Ka value.....	43
Table 1.S8	Gene set composing the chitin metabolic process ontology, found to be enriched in the 95th percentile of top-scoring <i>Culex pipiens</i> forms pipiens and molestus homologous codon sequence alignments as ranked by Ka value.....	44
Table 1.2	Gene ontology terms enriched in the set of 4,575 pairwise <i>Culex pipiens</i> forms pipiens and molestus homologous codon alignments devoid of non-synonymous substitutions. Asterisks indicate terms for which all members were present only in the test set.....	45
Table 1.S13	Extended analysis for all genes belonging to the GO terms from the highly conserved set (Table 2) for which all members were present only in the test set.....	46
Table 1.S9	Ka calculations corresponding with 13,587 pairwise <i>Culex quinquefasciatus</i> strain HAmCq and CpipJ1.3 homologous codon sequence alignments. Column two denotes genes present in the 95th percentile as ranked by Ka calculated using observed non-synonymous substitutions.....	47
Table 1.S10	Gene ontology terms enriched in the upper 95th percentile of pairwise Ka values calculated using <i>Culex quinquefasciatus</i> strains HAmCq and CpipJ1.3 homologous codon sequence alignments.....	48
Table 1.S12	Gene ontology terms enriched in the set of 3,687 <i>Culex quinquefasciatus</i> CDS sequences with at least one BLASTN alignment > 200bp at > 95% homology to another CDS in the genome.....	49
Table 2.S1	Genes for which phylogenetic trees were constructed, with AU and SH test results. Filtered for alignments > 200nt or > 50% recovery of the <i>Cx. quinquefasciatus</i> homolog.....	69
Table 2.S2	Gene set, with annotations and Ka/Ks, values for which significant topology was reported by CONSEL.....	70
Table 2.1	GO terms enriched in the gene set with ( <i>Cx. p. pipiens</i> form molestus + <i>Cx. p. pallens</i> ) tree topology.....	71
Table 2.2	GO terms enriched in the gene set with ( <i>Cx. p. pallens</i> + <i>Cx. quinquefasciatus</i> ) tree topology.....	71

Table 2.3	GO terms enriched in the gene set with ( <i>Cx. p. pipiens form molestus</i> + <i>Cx. quinquefasciatus</i> ) tree topology.....	71
Table 2.4	Set of 83 genes common to diverged (top 95 <sup>th</sup> percentile) sets of all three pairwise <i>Culex</i> species Ka calculations.....	72
Table 2.5	GO terms enriched in the set of 83 genes common to diverged (top 95 <sup>th</sup> percentile) sets of all three pairwise <i>Culex</i> species Ka calculations.....	74
Table 2.6	GO terms enriched in the set of 225 genes unique to the to diverged (top 95 <sup>th</sup> percentile) set from the <i>Cx. pipiens pallens</i> – <i>Cx. quinquefasciatus</i> pairwise Ka calculation.....	74
Table 2.7	GO terms enriched in the set of 222 genes shared uniquely between the to diverged (top 95 <sup>th</sup> percentile) sets of the <i>Cx. pipiens form molestus</i> – <i>Cx. quinquefasciatus</i> and <i>Cx. pipiens form molestus</i> – <i>Cx. pipiens pallens</i> pairwise Ka calculation.....	75
Table 2.S3	Annotations for genes enriched in the set of 225 genes unique to the to diverged (top 95 <sup>th</sup> percentile) set from the <i>Cx. pipiens pallens</i> – <i>Cx. quinquefasciatus</i> pairwise Ka calculation.....	76
Table 2.S4	Annotations for genes enriched in the set of 222 genes shared uniquely between the to diverged (top 95 <sup>th</sup> percentile) sets of the <i>Cx. pipiens form molestus</i> – <i>Cx. quinquefasciatus</i> and <i>Cx. pipiens form molestus</i> – <i>Cx. pipiens pallens</i> pairwise Ka calculation.....	77
Table 2.8	Mutiple comparison test after Kruskal-Wallis one-way analysis of variance.....	79
Table 3.S1	Primer sequences used for 5' RACE-PCR and to amplify RT-PCR products of <i>Cxqdsx</i> . ....	105
Table 3.S2	CENSOR tabular output with heat-map diagram for <i>Cx. quinquefasciatus doublesex</i> introns 1 though .....	106
Table 3.S3	Comparison of mobile genetic elements derived from <i>Ae. aegypti doublesex</i> introns 2-8 (Salvemini et al. [15]) and <i>Cx. quinquefasciatus dsx</i> introns 2-7 ( <i>Cxqdsx</i> lacks exon 5a and associated intron).....	107
Table 3.1	Splice donors and acceptors of the <i>Cxqdsx</i> gene. Coding (exon) sequences are in uppercase text, while the splice donor/acceptor and succeeding/preceding 12 nucleotides, respectively, are in lowercase. “Exon 4ex” denotes the alternate downstream splice donor of exon 4. Asterisk indicates the splice acceptor site deviates significantly from the genomic mean of 8.58 +/- 1.39 SE.....	108

Table 3.S4	Putative <i>cis</i> -element motifs of <i>D. melanogaster</i> , <i>An. gambiae</i> , <i>Ae. aegypti</i> and <i>Cx. quinquefasciatus</i> . Genomic location of <i>Cx. quinquefasciatus</i> elements are listed.....	109
Table 3.S5	RBP1 type-b motif enrichment scan results. <i>Culex quinquefasciatus</i> transcript id, with maximum number of RBP1b motifs per 547bp window, unique RBP1b motif sequence permutations present in the window, and nucleotide gene sequence are shown.....	110
Table 3.S6	Sliding window coordinates, Ka, Ks and Ka/Ks values calculated for each 30bp window of <i>Cx. quinquefasciatus</i> and <i>Cx. pipiens</i> form <i>pipiens dsx</i> CDS nucleotide alignment of male isoform.....	111

## List of Figures

Figure 1.1	Illustration of codon alignment generation process. 1. Illumina short read data are aligned to <i>Cx. quinquefasciatus</i> reference CDS sequence and used to build consensus sequences for both <i>Cx. pipiens</i> forms <i>pipiens</i> and <i>molestus</i> . 2. Consensus sequences for each gene are aligned, homologous positions free of Ns are removed and spliced. 3. GeneWise is used along with the corresponding full length <i>Cx. quinq.</i> peptide to create in-frame <i>f. pipiens/f. molestus</i> EST sequences from spliced alignments. 4. Codon alignments are created from EST sequences using TranslatorX. Ns denote unknown and/or unrecovered nucleotide data. ....33
Figure 1.S1	Maximum-likelihood phylogenetic tree showing monophyly of peritrophin-A domains reported here with peritrophic matrix proteins (labeled PMP), exclusive of the cuticular proteins analogous to peritrophins (labeled CPAP) of Jasrapuria et al. (2010). NCBI GI numbers are appended to <i>Tribolium castaneum</i> sequence IDs; all sequences are suffixed with "_subseq_[coordinate of first amino acid extracted]-[length of extracted peptide window]".....34
Figure 2.1	Venn diagram illustrating shared members of the 432-gene diverged sets (upper 95 <sup>th</sup> percentile of dataset as ranked by descending Ka) from the three pairwise comparisons.....67
Figure 2.2	Kruskal-Wallis one-way analysis of variance illustrating relationships among the Ka values obtained for each pairwise species comparison (chi-squared = 537.1547, df = 2, p < 2.2e-16).....68
Figure 3.1	Organization and splicing of <i>D. melanogaster</i> , <i>An. gambiae</i> , <i>Cx. quinquefasciatus</i> and <i>Ae. aegypti doublesex</i> genes. Homologous exons (colored boxes) are aligned vertically, with numbers placed above. Numbers within boxes represent the exon size, while those on a diagonal below represent intron size. Black arrows denote TRA/TRA2 binding sites, and asterisks are placed above weak splice acceptors. Yellow bars (not placed to scale) represent stop codons. Solid splice guides follow the female-specific spliceform, while dashed guides represent splicing in the male-specific form. The green/white stippled box adjacent to exon4 denotes the extension of the reading frame to the alternate splice donor.....112
Figure 3.2	Location of <i>Cxqdsx</i> RT-PCR primers (not to scale). Common exons are shown in green, the female-specific exon 5 in dark red, and male-specific (UTR in female) exons in blue. The exon4ex extension is represented with a green/white hatched box. The DBD/OD1 domain is indicated with



	a yellow box and OD2 with an orange box. Red triangles denote stop codons.....	113
Figure 3.3	Peptide alignment of <i>D. melanogaster</i> , <i>An. gambiae</i> , <i>Cx. quinquefasciatus</i> and <i>Ae. aegypti doublesex</i> genes; divided into 1. Common region (present in both male and female peptide), 2. Female-specific region 1 (C-terminus of female-specific protein), 3. Female-specific peptide C-terminus generated by use of alternate exon 4 splice donor, and 4. Male-specific region (C-terminus of male-specific protein). NCBI identification numbers are appended to the sequence ID. The in-frame intronic sequence is in bold/underline. The DNA-binding oligomerization domain (DBD/OD1) is boxed in yellow, while the common and female-specific portions of OD2 are boxed in orange.....	114
Figure 3.4	RT-PCR gel visualizations. (A). RT-PCR products derived from primers quinqOD12F/quinqOD12Rfem used to amplify female (left) and male (right) cDNA. (B). RT-PCR products derived from primers dsx8F/dsx6R used to amplify male (left) and female (right) cDNA. (C.) RT-PCR products derived from 5'RACE-PCR reaction after final amplification with primer DSX-5RACEGSP2 on female (left) and male (right) cDNA.....	116
Figure 3.5	Pairwise nucleotide alignment of putative <i>doublesex</i> promoter regions of <i>Culex quinquefasciatus</i> (top) and <i>Aedes aegypti</i> (bottom). Initiator box (Inr) and downstream promoter element (DPE) are boxed in yellow. The transcription start site (TSS) at position +1 and DPE at position +28 are marked. Exon 1 (spliced as UTR in both male and female mRNA) is outlined in black. The sequence logo plot below the alignment illustrates conservation.....	117
Figure 3.S1	Short-read mapping of <i>Cx. quinquefasciatus</i> RNAseq data (below; paired-end reads in blue, single-end reads in red/green) generated by Leal et al. [42] to the derived location of <i>Cxqdsx</i> exon 7 (green arrow). Reads spanning the splice junction to exon 6 are indicated with dashes at left. Data are as visualized in the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark).....	118
Figure 3.S2	Short-read mapping of <i>Cx. quinquefasciatus</i> RNAseq data (below; paired-end reads in blue, single-end reads in red/green) generated by Leal et al. [42] illustrating alternate exon 4 splice donor (boxed). Reads spanning the splice junction to exon 5 are indicated with dashes. Data are as visualized in the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark).....	119
Figure 3.6	Female RT-PCR prodcuts. RT-PCR products derived from amplification of female cDNA with primers quinqOD12F/quinqOD12Rfem illustrating the four female-specific amplicons obtained by splicing of the exon 4	

	extension and/or the exon 2 in-frame intron (ca. 905, 830, 745 and 670bp).....	120
Figure 3.S3	Short-read mapping of <i>Ae. aegypti</i> RNAseq data (below; paired-end reads in blue, single-end reads in red/green) from NCBI SRA accession SRR789758 illustrating the derived location of <i>Aeadsx</i> exon 1 (green arrow). Reads spanning the splice junction to exon 2 are indicated with dashes at right. The exon 1 annotation begins at the transcription start site (TSS), or the first adenine nucleotide of the initiator (Inr) sequence. Data are as visualized in the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark).....	121
Figure 3.S4	Short-read mapping of <i>Ae. aegypti</i> RNAseq data (below; paired-end reads in blue, single-end reads in red/green) from NCBI SRA accession SRR789758 illustrating the derived location of <i>Aeadsx</i> exon 7 (green arrow). Reads spanning the splice junction to exon 6 are indicated with dashes at left. Data are as visualized in the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark).....	122
Figure 3.S5	Splicing and alignment of RNAseq reads (numbered 1 through 14) from NCBI SRA accession SRR789758 to the <i>Aeadsx</i> exon5a/5b junction (exon 5a in yellow, 5b in green) illustrating canonical gt/ag splice donor/acceptor.....	123
Figure 3.7	<i>Cis</i> -element distribution. Graphical representation of <i>cis</i> -element distribution within the exon 4 extension, intron 4, exon 5 (female-specific), intron 5 and exon 6 (male-specific/common UTR). Transformer/transformer 2 complex (TRA/TRA2) binding sites are colored light blue, purine-rich elements (PRE) are colored red, Nasonia-like TRA/TRA2 sites in yellow, TRA-2-ISS elements in orange and RBP2b elements in dark blue. Exons are represented as green boxes.....	124
Figure 3.S6	Nucleotide sequence of exon 4 extension, intron 4, exon 5, intron 5 and exon 6 with putative <i>cis</i> -element binding sites annotated. See Fig. 3.7 for graphical representation.....	125
Figure 3.S7	Amino acid (above) and nucleotide (below) alignment of <i>Cx. quinquefasciatus</i> and <i>Cx. pipiens</i> form <i>pipiens</i> female <i>doublesex</i> isoforms. Bolded text denotes female-specific portion of protein.....	130
Figure 3.S8	Amino acid (above) and nucleotide (below) alignment of <i>Cx. quinquefasciatus</i> and <i>Cx. pipiens</i> form <i>pipiens</i> male <i>doublesex</i> isoforms. Bolded text denotes male-specific portion of protein.....	131

Figure 3.8	Ka/Ks graph. Graphical representation of Ka (top), Ks (middle) and Ka/Ks (bottom) values calculated for male-specific <i>Cx. quinquefasciatus</i> / <i>Cx. pipiens</i> form <i>pipiens doublesex</i> pairwise codon alignment. Values are recalculated for each 30 nucleotide (10 amino acid) window, which slides 3nt (1 AA) at a time. Numbers on the x-axis denote the coordinate of the central nucleotide in the window. Ka/Ks values were truncated at a maximal value of six for display purposes. The common portion of the transcript ends and male-specific sequence begins with the window centered at nucleotide position 735.....	133
Figure 4.1	Summary of evidence for hexapod <i>doublesex</i> recovered in this study. Orders for which we identified a high-confidence <i>dsx</i> EST encoding both OD1 and OD2 domains are marked ‘SHARED’; those for which only a Type-B OD1 domain was recovered (with Type-A absent) are annotated as such. Phylogeny as per Misof et al. 2014.....	149
Figure 4.S1	High-confidence <i>doublesex</i> transcripts reported in this study. Highly conserved amino acid positions are in red text. The conserved Lysine residue characteristic of the <i>dsx</i> DM domain (OD1, see Results and Discussion) is highlighted in yellow. Also included are OD2-encoding singleton domains recovered for taxa that possessed a high-confidence EST, to illustrate evidence for multi-copy <i>dsx</i> .....	150
Figure 4.S2	Type-A OD1-encoding singletons recovered in our analyses. Highly conserved amino acid positions are in red text. The conserved Lysine residue characteristic of the Type-A <i>dsx</i> DM domain (OD1, see Results and Discussion) is highlighted in yellow. E-values were calculated against our profile HMM using HMMer v3.1; asterisks indicate e-value was calculated via NCBI Conserved Domain Database search.....	152
Figure 4.S3	Type-B OD1-encoding singletons recovered in our analyses. Highly conserved amino acid positions are in red text. The conserved Isoleucine residue characteristic of the Type-B <i>dsx</i> -like DM domain (OD1, see Results and Discussion) is highlighted in yellow. E-values were calculated against our profile HMM using HMMer v3.1; asterisks indicate e-value was calculated via NCBI Conserved Domain Database search.....	153
Figure 4.S4	Dimerization (OD2) domain-encoding singletons recovered in our analyses. Highly conserved amino acid positions are in red text. E-values were calculated against our profile HMM using HMMer v3.1; asterisks indicate e-value was calculated via NCBI Conserved Domain Database search.....	155

Figure 4.S5	Alignment of both putative <i>doublesex</i> transcripts from each of the Zygentoma and Ephemeroptera species reported in this study. The high-confidence ESTs (black text) share similarity with each other, while the alternate motif (red text) segregates.....	156
Figure 4.2	Alternative splicing of <i>doublesex</i> transcripts from <i>Anurida maritime</i> (Collembola) and <i>Lepismachilis y-signata</i> (Archaeognatha). The consensus EST contig from each species (top, bold text), with short-read data mapped below illustrates the diverging transcript isoforms. The common 5' OD2 sequence is boxed in each species, while the diverging 3' ends are in red and orange text.....	157
Figure 4.3	Distance in nucleotides (grey) between the OD1 (red) and OD2 domains (blue), for each high-confidence <i>doublesex</i> transcript reported in this study. Total EST length (may include untranslated regions) is to the right of each bar.....	158
Figure 4.S6	De-novo assembled contigs generated in this study.....	159

## Introduction

The mosquitoes (Diptera: Culicidae) are a globally distributed family of arthropod vectors that have inarguably shaped the evolution of mankind. Mosquito-borne diseases affect over 475 million humans annually (WHO 2014

[<http://www.who.int/heca/infomaterials/en/vector-borne.pdf>, accessed Feb. 2015],

Fonseca et al. 2004a, Garske et al. 2014, Bailey et al. 2015) and confer significant morbidity and mortality. Developing effective vector and pathogen control strategies is thus of utmost importance to human health. These objectives require knowledge of the ecological factors that influence pathogen transmission and contribute to vectorial capacity. As many of these factors are influenced by or under the control of host genetics, it follows that the genomic age stands to yield powerful insights into vector biology and control.

For example, the deadliest human malaria agent (*Plasmodium falciparum*), the agent of lymphatic filariasis (*Wuchereria bancrofti*) and both dengue and yellow fever viruses are vectored by mosquitoes living in close association and feeding primarily on humans, thus maximizing transmission rate, and allowing high parasite virulence (Dieckmann et al. 2002). Emerging diseases such as West Nile Virus (WNV) or eastern equine encephalitis however require a bridge vector (with a broader host range) to generate epidemic events (Kilpatrick et al. 2006, Farajollahi et al. 2011). Multiple studies have shown divergence of host preference between *Culex* species and/or species forms (Fonseca et al. 2004a, Gomes et al. 2012, Fritz et al. 2015), however genetic factors influencing host preference in mosquitoes remain largely unknown and only recently has their identification begun (i.e Hodges et al. 2014, McBride et al. 2014).

Transmission and pathogen virulence of WNV, Saint Louis encephalitis (SLE) virus, avian malaria (*Plasmodium spp.*) and *W. bancrofti* have all been shown to modulate in accordance with mosquito genotype and/or gene expression (James 2002, Kumar and Paily 2008, Shin et al. 2014), and each of these pathogens are often vectored primarily by members of the *Culex pipiens* mosquito complex. The *Cx. pipiens* complex contains four species that are evolutionarily closely related and often difficult to identify: *Cx. pipiens* Linnaeus, *Culex quinquefasciatus* Say, *Culex australicus* Dobrotworsky & Drummond and *Culex globocoxitus* Dobrotworsky (Farajollahi et al. 2011; see <http://www.wrbu.si.edu> for current taxonomy). *Cx. quinquefasciatus* inhabits primarily tropical regions of North and South America, the African lowlands, Southern Asia and Australia (Fonseca et al. 2006) while *Cx. pipiens* is temperate and originally restricted to Europe and Asia (Harbach et al. 1984). However, as a result of human migration and commercial traffic, the two species overlap in sub-tropical areas leading to expansive hybrid zones in North America, Argentina, Madagascar, Japan and South Korea (Barr 1957, Urbanelli et al. 1997, Wang et al. 2000, Cornel et al. 2003, Fonseca et al. 2009). Hybridization between *Cx. pipiens* and *Cx. quinquefasciatus* in areas where both have been introduced has often led to taxonomic confusion, particularly by those still embracing old views of the biological species concept.

*Culex pipiens* is currently classified into two subspecies: *Culex pipiens pipiens*, a cosmopolitan mosquito distributed throughout temperate Europe with a Southern reach to the S. African highlands, and *Culex pipiens pallens*, distributed throughout temperate Eastern Asia as far as Japan. Additionally, *Culex p. pipiens* has two recognized forms, or biotypes: “pipiens” and “molestus”. Despite being morphologically identical, these two

forms exhibit stark behavioral and ecological contrasts. Form *pipiens* is a feral mosquito that requires a vertebrate bloodmeal for egg development, enters winter diapause in response to ambient light levels, swarms prior to mating and is primarily ornithophilic (Spielman 2001, Fonseca et al. 2004a). Form *molestus* however can forego a bloodmeal prior to its first gonotrophic cycle, remains gonoactive during winter months while occupying warm subterranean environments (e.g. subways, sewers), and mates in very confined spaces. Female *f. molestus* mosquitoes will readily feed on mammals, including humans (Fonseca et al. 2004a, Gomes et al. 2012, Fritz et al. 2015). This “domestic” form of *Culex pipiens* thus maintains various phenotypic characteristics that allow it to associate almost entirely with human ecosystems and consequently spread to all continents except Antarctica (Farajollahi et al. 2011). Microsatellite analysis indicates that evolutionary divergence of forms *pipiens* and *molestus* occurred as recently as 10,000 years ago (Fonseca et al. 2004b).

The East Asian temperate subspecies of *Cx. pipiens*, *Cx. pipiens pallens*, exhibits an intermediate or hybrid male phallosome that follows a morphological cline from “*pipiens*-like” in the northern latitudes to “*quinquefasciatus*-like” in the south (Bekku 1956). *Cx. pipiens pallens* has long been hypothesized to be a hybrid of *Cx. pipiens* and *Cx. quinquefasciatus* (Barr 1957, Tanaka et al. 1979). Fonseca et al. (2009) however found that Japanese and Korean *pallens* mosquitoes have a distinctive genotypic signature at both microsatellite and other nuclear loci even in the absence of hybridization between *Cx. p. pallens* and *Cx. quinquefasciatus*. Additionally, they found that all male *Cx. p. pallens*, even in northern populations, contain both the diagnostic *pallens ace-2* allele and a second copy identical to that of *Cx. quinquefasciatus*. These data point to an

ancient and independent origin of *Cx. pipiens pallens* resulting from a hybridization event between *Cx. quinquefasciatus* (as evidenced by the *ace-2* “relic”) and Eastern European *Cx. pipiens* (evidenced by external genitalia morphology and similarity of ecology and behavior [Fonseca et al. 2009]).

The results of evolutionary processes that lead to cryptic and incipient speciation as evidenced in the *Culex pipiens* complex are intriguing; the “domestication” or adaptation to (and reliance on) the human environment by *Culex pipiens* resulting in form *molestus* has placed a global vector mosquito in close association with man, yet apparent morphological identity persists. Conversely, the natural hybridization (Arnold 1997) that resulted in a homoploid yet morphologically distinct *Cx. pipiens pallens* presumably conferred a fitness advantage, as witnessed by its ability to secure resources and persist. Hybrid speciation is relatively rare among animals (see Mallet [2007] and Mavárez and Linares [2008]), and these events afford us a unique opportunity to examine the genomic consequences of both adaptive evolution and hybrid speciation in a single species complex of medically important and globally distributed mosquito. Elucidating how the genomes of these mosquitoes change concomitant with their phenotypes will highlight the genetic toolkit that comprises a domestic mosquito (as in the form *molestus* system) or the retention of parental alleles central to niche adaptation (in the *Cx. pipiens pallens* system).

The Culicidae comprise the world’s deadliest animal(s), and thus evolutionary studies dealing with them rarely neglect to address their control. One promising strategy for population reduction and thus vector-borne disease mitigation is sterile insect technique (SIT; Gilles et al. [2014]), which relies on the release of sterile insects (usually



male) to facilitate unsuccessful matings. Manipulation of sex ratio is thus central to SIT, and it follows that molecular mechanisms of sex determination in mosquitoes and other insects are a major focus of SIT research and development (Dafa'alla et al. 2010). The mosquitoes (as do other Dipteran families) regulate somatic sexual differentiation via a molecular cascade that ultimately results in the sex-specific splicing of the *doublesex* gene.

*Doublesex (dsx)* is a member of the *Doublesex/Mab-3* Related Transcription factor (DMRT) family of zinc-finger proteins (Kopp 2012). These transcription factors are widespread in eukaryotes and regulate a variety of downstream target genes (Clough et al. 2014) to ultimately differentiate gender-specific somatic cell fate, gonadal and neural tissues. Within the insects, this process involves a genetic cascade first elucidated in the model fly *Drosophila melanogaster* (Baker and Wolfner 1988) whereby a primary signal triggers sex-specific splicing of one or more regulatory factors which subsequently bind pre-mRNA of the conserved DMRT “major switch” gene, *doublesex*, and direct its sex-specific splicing, thus initiating development of male or female forms (Sánchez 2008). The primary upstream signals that initiate the cascade and splice sex-specific *dsx* transcripts are diverse (see 1), however the gene appears conserved as the final major switch (Geuverink and Beukeboom 2014, Wexler et al. 2014). Partial *dsx* gene models have proposed for the mosquitoes *Anopheles gambiae* (*Angdsx* [Scali et al. 2005]) and *Aedes aegypti* (*Aeadsx* [Salvemini et al. 2011]). Both genes show evidence of sex-specific splicing, yet differ in several evolutionary aspects: *Aeadsx* exhibits instances of exon gain, contains *cis*-regulatory binding sites not present in *Angdsx*, and presents alternative signals to the splicesosomal machinery to facilitate alternative splicing. These

observations led Salvemini et al. (2011) to conclude that *Aeadsx* and *Angdsx* may be under the influence of different upstream factors. This hypothesis is bolstered by the fact that Anopheline mosquitoes possess heteromorphic (XY) sex chromosomes, while sex in *Aedes* (and other Culicine mosquitoes, e.g. *Culex*) is determined at a locus (Newton et al. 1974) and thus the top-level or primary signals may diverge.

Recently, Whyard et al. (2015) found that RNAi-mediated knockdown of the female-specific *dsx* transcript in *Aedes aegypti* resulted in male-biased adult development via female mortality (and not gender reversal). Should SIT continue to pursue *dsx* as a research target, it will be imperative to fully elucidate the breadth of its variation among the vectors targeted; the striking variability that has been shown in the initial (and only) two mosquito *dsx* genes described (Scali et al. 2005, Salvemini et al. 2011) attest to the need for more research.

The evolutionary impact of *doublesex* extends far beyond vector control, however. The gene has been shown to influence both behavioral and morphological secondary sex characteristics (Siwicki and Kravitz 2009, Kijimoto et al. 2012, Devi and Shyamala 2013) in addition to gender, and therefore may have extensive influence on mating success and reproductive isolation. Additionally, *dsx* has been shown to be under positive selection in multiple orders (Ruiz et al. 2007, Hughes 2011, Sobrinho and de Brito 2012) and implicated in a system of “runaway evolution” due to its developmental influence on secondary sex characteristics and the response of female preference to genetic drift (Hughes 2011). These data inextricably link *doublesex* which mechanisms of incipient speciation, but does this paradigm apply to all insects? Thus far *dsx* homologs have been recovered from genome and/or transcriptome data of only seven

insect orders (Geuverink and Beukeboom 2014), yet are present in the crustacean *Daphnia magna* (Kato et al. 2011) as well as the arachnid *Metaseiulus occidentalis* (Pomerantz and Hoy 2015), both of which are hexapod outgroups within the Arthropoda. In these organisms, *dsx* pre-mRNA shows no sign of alternative splicing and directs sexual fate via sex-biased expression in males.

The *Culex pipiens* species complex presents an excellent model for the study of genome evolution and phenotypic response in a global disease vector. To understand how *Culex pipiens* forms *pipiens* and *molestus* have diverged on a molecular level, and to associate this divergence with phenotype (i.e the “domestication” of form *molestus*), I will examine evolution of the mosquito proteome and identify loci that are likely targets of adaptive evolution in Chapter 1. In Chapter 2, I will examine the hybrid speciation evidenced in *Cx. pipiens pallens* by defining genome-wide mosaicism as evidenced by phylogenetic affinity to coding sequences of either donor (*Cx. pipiens pipiens* and *Cx. quinquefasciatus*). Additionally, I examine peptide evolution between these three mosquitoes (*Cx. pipiens pallens*, *Cx. pipiens pipiens* and *Cx. quinquefasciatus*), and define genes and gene ontologies that may be central to the evolution and adaptation of each taxon. In Chapter 3, I will characterize the *doublesex* gene from *Culex quinquefasciatus* and perform comparative analyses that accentuate the rapid evolution at the base of the mosquito sex determination cascade. Finally, in Chapter 4, I will illustrate both the ubiquity and rapid evolution of *dsx* throughout all hexapods, and provide evidence that the alternative splicing of sex-specific *dsx* transcripts was present in the last common ancestor.

## Chapter 1:

### Genetic Divergence between Populations of Feral and Domestic forms of a Mosquito Disease Vector assessed by transcriptomics

#### Abstract

*Culex pipiens*, an invasive mosquito and vector of West Nile virus in the US, has two morphologically indistinguishable forms that differ dramatically in behavior and physiology. *Cx. pipiens* form *pipiens* is primarily a bird-feeding temperate mosquito, while the sub-tropical *Cx. pipiens* form *molestus* thrives in sewers and feeds on mammals. Because the feral form can diapause during the cold winters but the domestic form cannot, the two *Cx. pipiens* forms are allopatric in northern Europe and, although viable, hybrids are rare. *Cx. pipiens* form *molestus* has spread across all inhabited continents and hybrids of the two forms are common in the US. Here we elucidate the genes and gene families with the greatest divergence rates between these phenotypically diverged mosquito populations, and discuss them in light of their potential biological and ecological effects. After generating and assembling novel transcriptome data for each population, we performed pairwise tests for nonsynonymous divergence ( $K_a$ ) of homologous coding sequences and examined gene ontology terms that were statistically over-represented in those sequences with the greatest divergence rates. We identified genes involved in digestion (serine endopeptidases), innate immunity (fibrinogens and  $\alpha$ -macroglobulins), hemostasis (D7 salivary proteins), olfaction (odorant binding proteins) and chitin binding (peritrophic matrix proteins). By examining molecular divergence between closely related yet phenotypically divergent forms of the same species, our

results provide insights into the identity of rapidly-evolving genes between incipient species. Additionally, we found that families of signal transducers, ATP synthases and transcription regulators remained identical at the amino acid level, thus constituting conserved components of the *Cx. pipiens* proteome. We provide a reference with which to gauge the divergence reported in this analysis by performing a comparison of transcriptome sequences from conspecific (yet allopatric) populations of another member of the *Cx. pipiens* complex, *Cx. quinquefasciatus*.

## Introduction

Specific life-history traits of arthropod disease vectors can determine the duration and severity of outbreaks by influencing vectorial capacity (NAS 2008). *Plasmodium falciparum*, the deadliest of human malaria agents, *Wuchereria bancrofti*, the widespread causative agent of lymphatic filariasis, and both dengue and yellow fever viruses are transmitted by mosquito vectors that live in close association with and feed near-exclusively on humans. Anthropophilic mosquito phenotypes maximize transmission rates and promote high pathogen virulence of these diseases (Dieckmann et al. 2002). In contrast, zoonotic diseases requiring amplification cycles in non-human vertebrate hosts such as West Nile virus or eastern equine encephalitis will only spill over to humans (often to the detriment of the parasite and the human) if a vector with a broader range of hosts becomes involved (Kilpatrick et al. 2006, Farajollahi et al. 2011). Although blood meal analyses have demonstrated strong associations between vector species and suites of vertebrate hosts, the mechanisms underlying host-choice are still broadly unknown and are often ascribed to environmental instead of genetic causes (Chaves et al. 2010).

The northern house mosquito, *Culex pipiens*, is comprised of two morphologically indistinguishable forms (eco/biotypes), *Cx. pipiens* form *pipiens* L. and *Cx. pipiens* form *molestus* Forskål (herein f. *pipiens* and f. *molestus*, respectively). Despite their morphological identity and very close phylogenetic history (Fonseca et al. 2004b), the two forms exhibit notable ecological and behavioral differences that make their identification possible. The feral form, f. *pipiens* requires a vertebrate bloodmeal for all egg development (anautogeny), enters winter diapause when ambient light levels decrease below a locally pre-established threshold in the fall (heterodynamous), swarms as a prelude to mating (eurygamous), and is primarily ornithophilic. In contrast, f. *molestus* can forego a bloodmeal for its first gonotrophic cycle (autogeny), adults remain gonoactive during winter months (homodynamous), which means they are often restricted to subterranean environments with standing water such as subways and sewers (hypogeous) that remain warm. Males of f. *molestus* will mate in very confined spaces (stenogamous) and females frequently feed on mammals, including humans (references summarized in Fonseca et al. [2004a]; see Gomes et al. [2012] for latest blood meal studies). *Cx. pipiens* f. *molestus* is a worldwide invasive species, spread by humans to all continents except Antarctica (Farajollahi et al. 2011) while f. *pipiens* has remained restricted to Northern Europe. *Cx. pipiens* populations within the United States are hybrids of the two forms (Fonseca et al. 2004b), and are implicated in the maintenance and transmission of epizootic arboviruses such as West Nile Virus (WNV) to humans resulting in illness and occasionally death (Kramer et al. 2008).

The two forms of *Cx. pipiens* are very closely related, as is evident from their identical morphology and genetic similarity (Fonseca et al. 2004b). This has led to

controversy over their taxonomic standing (Harbach et al. 1984, Spielman et al. 2004). However, they are differentiated at hyper-variable loci such as the flanks of microsatellites (Bahnck and Fonseca 2006) indicating recent separate evolutionary histories. The genetic similarity despite striking differences in ecology, behavior and physiology indicate that *f. molestus* may have diverged from *f. pipiens* and evolved its association with humans as recently as 10,000 years ago (Fonseca et al. 2004b). This recent split represents an exceptional opportunity to test whether targets of molecular evolution in *Cx. pipiens* mosquitoes can be elucidated using two phenotypically diverged populations. Additionally, by framing the results in context of phenotype, the data generated would serve as a first look at the molecular basis for domestication.

To start testing this hypothesis, we generated and compared de-novo whole-transcriptomes from one representative population each of *Cx. pipiens f. pipiens* and *f. molestus* using the *Cx. quinquefasciatus* genome (CpipJ1.3 Johannesburg, South Africa, (Arensburger et al. 2010)) as a reference. *Cx. quinquefasciatus* is a closely related sibling species of *Cx. pipiens* (Farajollahi et al. 2011), and is the only available annotated *Culex* genome assembly. We performed pairwise comparisons of orthologous coding (CDS) nucleotide sequences to identify genes and gene ontologies that show evidence of evolving at accelerated evolutionary rates between *f. pipiens* and *f. molestus* by calculating per-gene rates of non-synonymous substitution per non-synonymous site ( $K_a$ , or  $dN$ ). Wang et al. (2011) show that commonly used tests for natural selection that normalize  $K_a$  by a ‘background mutation rate’, or  $K_s$  (synonymous substitutions per synonymous site) often produce non-uniform results among closely related genomes, yet find that  $K_a$  alone remains stable and an adequate gauge for rate of “uncorrected” peptide

evolution. This is primarily due to the varying manner in which  $K_s$  is calculated in a likelihood framework by different algorithms, and can also be influenced by sequence composition (Parmley and Hurst 2007, Wang et al. 2011). Additionally,  $K_a/K_s$  calculations are often incorrectly elevated among isolated populations and closely related lineages due to segregating polymorphisms (both neutral and slightly deleterious) present at the time of divergence (Kryazhimskiy and Plotkin 2008, Peterson and Masel 2009, Mugal et al. 2014). Since there is minimal phylogenetic distance between the two forms we sequenced, synonymous substitutions would be expected to far outnumber those that are non-synonymous. This scenario is particularly susceptible to the aforementioned biases, as even small stochastic variation in synonymous substitution rates coupled with artifacts in  $K_s$  calculation can exert disproportionately large influence on the selection signature (Koonin and Rogozin 2003, Parmley and Hurst 2007, Wang et al. 2009). For these reasons, we elected to use  $K_a$  as the primary metric for presentation of our data. As the software we selected for our calculations implements the test in a likelihood framework which corrects for multiple substitutions at sites, a process less likely to have occurred in such closely related taxa, we performed primary calculations using also observed substitutions in addition to those derived from the model and discuss congruence between the two approaches. Although our primary objective was to elucidate components of the mosquito genome evolving at accelerated rate, we also report here ontologies enriched in the set of genes devoid of non-synonymous substitutions as they provide candidates for targets of negative or purifying selection and define critical biological processes and cellular components in the *Cx. pipiens* genome.



To contrast the amount of genetic variation uncovered in the comparison of *Cx. pipiens* forms with another geographically isolated yet conspecific population, we repeated the analysis with publicly available transcriptome data from two strains of *Cx. quinquefasciatus*: a North American strain (Reid et al. 2012) and the Johannesburg reference (Arensburger et al. 2010). We hypothesized that a greater amount of divergence would be witnessed between the two *Cx. pipiens* populations, which exhibit qualifiable phenotypic differences characteristic of the taxonomic forms, rather than between conspecific *Cx. quinquefasciatus* populations. In addition, we examined whether particular GO terms present in our results may be derived from ambiguous placement of read data from paralogous or multiple-copy genes by testing for their presence within an enriched ontology list derived from genes which share significant DNA similarity with others in the genome.

## Materials and Methods

Because only *Cx. pipiens* f. *molestus* or hybrids of the two *Cx. pipiens* forms occur in the U.S, we obtained egg rafts of f. *pipiens* from Baden-Württemberg in southwestern Germany. Multiple individual egg rafts were isolated, hatched and DNA was extracted from ca. 10 larvae from each using a Qiagen DNEasy Blood & Tissue kit (Qiagen, Valencia CA). PCR-based positive species identification of *Cx. pipiens* was performed via the acetylcholinesterase-2 assay developed by Smith and Fonseca (2004), and further to f. *pipiens* using the CQ11 assay of Bahnck and Fonseca (2006). Field populations of pure f. *molestus* are difficult to obtain since they are strictly subterranean and mostly found by chance (Fonseca DM personal experience). Therefore, egg rafts of f.

*f. molestus* were obtained from a young colony, initiated from a large subterranean swarm of females detected in a New York, NY residential basement in December 2010.

Blooded females that had been biting local residents were allowed to lay egg rafts in the laboratory and henceforth the colony has been maintained without access to blood.

Representative specimens of the NYC colony of *f. molestus* have been genotyped with a panel of 8 microsatellite loci and have a genetic signature that matches that of populations of *f. molestus* from southwestern Germany, as do other *f. molestus* specimens obtained from multiple locations around the world (Fonseca et al. 2004b, Micieli et al. 2013, Turell et al. 2014).

Once eggs hatched, larvae of both forms were reared in ceramic pans under a 16:8 L:D cycle on a diet of ground rat chow prior to emergence. Four specimen groups were created: thirty 1<sup>st</sup>/2<sup>nd</sup> instar, eight 3<sup>rd</sup>/4<sup>th</sup> instar, eight pupae and eight non-blood fed adult (4 male, 4 female) mosquitoes. Each group was placed in a separate plastic 2ml microcentrifuge tube containing a 5mm sterile stainless steel bead and 900ul QIAzol lysis reagent prior to disruption with a TissueLyser II (Qiagen, Valencia CA) for 2 minutes at 20Hz. Total RNA extraction was then carried out on each group using the RNeasy Plus Universal kit (Qiagen, Valencia CA) per manufacturer protocol and quantified on a Qubit 2.0 fluorometer (Life Technologies) using the RNA Broad-range buffer. One ug of RNA from each group was combined and used to prepare an Illumina sequencing library using the TruSeq RNA Sample Prep kit v2 (Illumina, Inc. San Diego, CA) per manufacturer protocol. The *Cx. pipiens f. molestus* library was sequenced twice on an Illumina MiSeq (Illumina, Inc), once using a 500-cycle (2x250bp paired-end) MiSeq Reagent Kit v2, and once using 1/3 of a multiplexed 600-cycle (2x300bp paired-end)

MiSeq Reagent Kit v3. *Culex pipiens* f. *pipiens* was sequenced once using 1/3 of a multiplexed 600-cycle (2x300bp paired-end) MiSeq Reagent Kit v3. Raw sequence data were quality trimmed using the CLC Genomics Workbench (Limit score cutoff = 0.05, CLC Bio, Aarhus, DK).

To assemble EST sequences for each mosquito taxon (illustrated in Fig. 1.1), we used the sequenced genome of another recognized member of the *Cx. pipiens* complex, *Culex quinquefasciatus* Say (Arensburger et al. 2010) (for current taxonomy see <http://wrbu.si.edu>) as a reference. We mapped raw read data for each form individually to the *Cx. quinquefasciatus* genome CDS sequence, extracted from the CpipJ1.3 genome assembly (<http://www.vectobase.org/organisms/Culex-quinquefasciatus>, [Megy et al. 2012]) using the CLC Genomics Workbench (CLC Bio, Aarhus, DK) at a nucleotide similarity of 95% over a required length fraction of 95% of the read. Reads that had more than one best alignment (i.e potentially paralogous DNA) were ignored. Consensus sequences for each CDS were then generated from the alignment, with conflicts resolved by choosing the base with the highest additive quality score and a minimum coverage of 2x. Areas of < 2x coverage were filled with Ns from the reference. The f. *pipiens* and f. *molestus* CDS sequences were aligned with each other, and sites with Ns in either or both forms were removed. Genewise (Birney et al. 2004) was used to create in-frame CDS sequences using the homologous peptide sequence of the *Cx. quinquefasciatus* as a guide, and any sequences that had stop codons introduced after this process were removed. Codon alignments were created with TranslatorX (Abascal et al. 2010), guided by a peptide alignment of their translations generated via MAFFT v.6.9 (Katoh and Toh 2010). This codon alignment was used to calculate Ka values using the KaKs Calculator

v.2 (Wang et al. 2010) using both observed non-synonymous substitutions and those estimated via maximum-likelihood estimation under likelihood model averaging (MA). We retained Ka values for CDS codon alignments greater than 200bp, or for alignments < 200bp for which >50% of the sequence length (as calculated from the *Cx. quinquefasciatus* homolog) was recovered in the *f. molestus* – *f. pipiens* comparison. As this test compares single haploid gene sequences, and we reduced allelic variation within and among individuals sequenced from the population by generating haploid consensus gene sequences (above), it is likely that our Ka calculations underestimate the true amount of non-synonymous variation within the populations sequenced. Additionally, the alignment stringency (95%) of the mapping will exclude genes that have diverged significantly between the subject and the reference, however we find it a conservative value with which to avoid false positives generated from gene paralogs. Enrichment tests were performed using Blast2GO (Conesa et al. 2005) with a reference set consisting of 11,930 genes (Table 1.S1) that met the length criteria above (GO Term Filter Value = .05, Term Filter Mode = FDR, single-tailed test) and a test set composed of the 95<sup>th</sup> percentile of CDS sequences with highest calculated Ka. Additionally, to discern possible candidates of purifying selection, a test set of genes lacking non-synonymous substitutions from the *f. pipiens* – *f. molestus* comparison was created by selecting 4,575 CDS alignments (generated above, Table 1.S1) from our data with 100% amino acid identity and used in a separate enrichment test coupled with the reference set above.

For the intra-specific *Cx. quinquefasciatus* comparison, data generated by Reid et al. (2012) from colonies started from an Alabama, USA population (strain HAmCq1 and HAmCq8) were compared to the CpipJ1.3 reference as above; briefly, reads from NCBI

SRA libraries SRR364515 and SRR364516 were combined and mapped to the CpipJ1.3 CDS sequence, consensus sequences were built using the same protocol and parameters as above, and genewise / translatorX were used to construct the codon alignment prior to Ka calculation. From this, we constructed a reference set containing 13,281 genes which met the *f. pipiens* – *f. molestus* length cutoff above. As this was a conspecific comparison (assuming minimal evolution), we used only observed substitutions as opposed to those derived via maximum likelihood estimation (MLE) for the dN calculation.

To examine whether particular gene ontologies present in our results may be derived via ambiguous placement of read data from paralogous or multiple-copy genes, we tested for their presence within an enriched ontology list derived from genes that share significant DNA similarity with others in the genome. This was accomplished by blasting the *Cx. quinquefasciatus* CpipJ1.3 CDS sequence data used above into itself via BLASTN (Altschul et al. 1990) with an e-value cutoff of  $1 \times 10^{-5}$  and saving all ‘non-self’ hits for genes which had a 95% similarity over a local alignment of 200nt (a value we chose as our average read length after trim was 211nt). This returned 3,687 (Table 1.S11) sequences that were used as a test set in a Blast2GO enrichment test against a reference consisting of all CDS sequences.

In all tests, we retained GO terms with an False Discovery Rate (FDR) corrected (Benjamini and Hochberg 1995) p-value of  $p \leq .05$ . Gene names reported are retained from their *Cx. quinquefasciatus* reference used to construct the consensus. Annotations were performed against the NCBI nr database and via InterProScan v.5 (Apweiler et al. 2000). Phylogenetic analysis of the Peritrophin-A domain-containing proteins was

performed by extracting the peptide sequence for each chitin-binding domain from the *Cx. quinquefasciatus* homolog corresponding to each of our candidate genes based on coordinates returned via InterproScan v.5 prior to alignment with a selection of peritrophic matrix protein (PMP) and cuticular proteins analogous to peritrophin (CPAP) domains of Jasrapuria et al. (Jasrapuria et al. 2010) extracted in the same manner. Sequences were aligned using T-COFFEE v.10.00.r1613 (Notredame et al. 2000) and tree reconstruction under automatic model selection and 1500 bootstrap replicates was performed using IQTREE v. 0.9.6 (Minh et al. 2013).

## Results and Discussion

### *Transcriptome sequencing and Ka calculation*

Transcriptome sequencing generated 58.7 million (11.2Gbp) and 24.7 million (5.3Gbp) of short-read data for *f. molestus* and *f. pipiens*, respectively. The *f. molestus* data mapped to 18.4 Mbp (74%) of the 25.0 Mbp *Cx. quinquefasciatus* CDS sequence reference by length (15,624 of 19,019 transcripts had at least one mapped read), with an average coverage of 71x and median coverage (50<sup>th</sup> percentile) value of 17x. The *f. pipiens* RNAseq data mapped to 17.2 Mbp (70%) of the *Cx. quinquefasciatus* reference by length (14,537 transcripts had at least 1 mapped read) with an average coverage of 45.5x and median of 8x at our alignment stringency (95% nt similarity over 95% of the read length, see Methods). After refinement by length and coverage (see Methods), the short read alignments were used to create 11,930 pairs of putative ortholog consensus sequences (one pair for each of 11,930 genes). Each taxon contributed 14.15 Mbp of sequence data. After codon alignment, the gene set was ranked by pairwise Ka value

calculated via both the maximum-likelihood estimation and by observed count, and the top 5% (n=597, Table 1.S1) of genes from each were selected to create two Blast2GO test sets for Enrichment Analysis (Fisher's Exact Test).

### *Enrichment within the fast-evolving genes*

When reduced to most-specific terms (i.e parent terms removed), the analysis identified the same seven Gene Ontology (GO) terms as enriched for both the observed and log-likelihood test sets (Table 1.1): serine-type endopeptidase activity (GO0004252), proteolysis (GO0006508), receptor binding (GO0005102), odorant binding (GO0005549), extracellular space (GO0005615), chitin metabolic process (GO0006030) and chitin binding (GO0008061). As both test sets converged on the same terms, we will present all further results and data tables corresponding to output from the observed count analysis.

The *Serine-type endopeptidase activity* (GO:0004252) ontology comprises a family of enzymes that utilize a nucleophilic serine at the active site to cleave peptide bonds in proteins. These enzymes are widely distributed throughout both pro- and eukaryotes and classified into 16 superfamilies. Most eukaryotic serine endopeptidases belong to the Chymotrypsin serine protease S1 family, where both chymotrypsin-like and trypsin-like proteases function as digestive enzymes in hydrolyzing proteins to smaller peptides and amino acids for further digestion (Rawlings and Barrett 1994, Madala et al. 2010). Annotation of the serine endopeptidases within our enriched set (Table 1.S2) shows 45 of the 50 proteins carry a trypsin domain (Pfam PF00089). Mosquito trypsins, secreted by gut epithelium, function in digestion of protein-rich bloodmeals within the

female after encapsulation by a peritrophic matrix (Borovsky and Schlein 1987, Borovsky 2003). In a process currently considered unique to mosquitoes (Diptera: Culicidae), two forms of trypsin are critical for complete bloodmeal digestion (Felix et al. 1991). Within 1 hour following ingestion, early trypsin protein is translated from mRNA stored in the gut epithelium. This early trypsin protein functions to partially digest the bloodmeal, creating smaller peptides that in turn trigger and regulate late trypsin transcription and translation (Noriega et al. 1999, Borovsky 2003). Late trypsins then further digest the bloodmeal to free amino acids sourced for egg development. This feedback mechanism ensures that digestive proteases are produced only in response to blood (as opposed to carbohydrate/sugar) and in quantities commensurate with “pre-assessment” of bloodmeal protein content by early trypsin digestion. In addition to digestion, Valenzuela et al. (2002) found several secreted salivary serine proteases with homology to *Manduca* prophenoloxidase-activating enzymes that are likely involved in the innate melanotic immune response.

The presence of such elevated levels of trypsin variation between populations may indicate that differences in the source of bloodmeal necessitated adaptive changes in digestive enzymes to hydrolyze differentially abundant proteins. Further study will be required to determine whether the proteins highlighted in our analysis represent early and/or late trypsins, as two proteins carried an annotation of late trypsin and only four trypsins have been annotated as early or late to date within the *Cx. quinquefasciatus* genome project (via Vectorbase; <https://www.vectorbase.org/organisms/culex-quinquefasciatus> retrieved Jun 2014). Five proteins in our set were annotated as coagulation factors, however an NCBI Conserved Domain analysis



(<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>, results not shown) fails to return evidence for canonical Gla and/or EGF domains within these peptides, indicative of the coagulation factors (Stavrou and Schmaier 2010).

The proteolytic enzymes within the *Proteolysis* (GO:0006508) ontology hydrolyze proteins to smaller peptides and/or amino acids. This gene ontology contained primarily the serine endopeptidase enzymes discussed above, with the addition of several serine protease inhibitors, metallopeptidases and apoptotic caspases (Table 1.S3).

*Receptor binding* (GO:0005102) protein molecules interact selectively with specific cellular receptors to initiate changes in cell function. Eighteen such proteins were present in the enriched set, of which all were found to carry a fibrinogen beta and gamma chain Pfam (PF00147, Table 1.S4) annotation. In the invertebrates, including mosquitoes, fibrinogen-related proteins (FREPs) are restricted to the innate immune response, functioning in pathogen recognition and agglutination (Dong and Dimopoulos 2009, Hanington and Zhang 2011). Many *Anopheles gambiae* FREP genes display immune-responsive transcription after being challenged with bacteria, fungi or both rodent and human malaria protozoa (Dong and Dimopoulos 2009) indicating that they play a pivotal role in mosquito vectorial capacity. This gene family has undergone lineage-specific duplications with relaxed selective constraints, as the *An. gambiae* genome contains 59 FREP members, with 32 and 87 members currently annotated in the genomes of *Ae. aegypti* and *Cx. quinquefasciatus*, respectively (Arensburger et al. 2010), while the *Drosophila melanogaster* genome contains twenty (Wang et al. 2005). This likely reflects the diverse pathogen load faced by each particular dipteran species during its life cycle. Further annotation reveals four putative ficolins in our set, a particular

oligomeric lectin containing a C-terminal fibrinogen-like domain able to bind N-acetylglucosamine, a chitin monomer, as part of immune response (Krarup et al. 2004). It is likely that the two populations of *Cx. pipiens* sequenced here are challenged by different bacterial communities within their respective environments, and experience both varying larval habitat (subterranean sewers and subway systems [form molestus] vs. stagnant, above-ground pools [form pipiens]) and bloodmeal hosts (with associated food-borne pathogens; see Serine endopeptidases above). The rate of peptide evolution seen in this component of the innate immune system may be a result of adaptation to these ecological stressors.

Members of the *odorant binding* (GO:0005549) ontology compose a large multi-gene family of water-soluble proteins secreted by support cells into sensillum lymph of the female mosquito antennal hairs (Schultze et al. 2013). These proteins bind various odorant molecules, thus triggering chemosensory mechanisms such as host-seeking and oviposition site recognition (Pelosi and Maïda 1995). Characterized by a six alpha-helical domain and the disulphide bonds created by six conserved cysteine residues, the mosquito odorant binding proteins (OBPs) have been studied extensively in the available mosquito genomes. Like the fibrinogens, the OBP protein family has been found to be very divergent within the Culicidae, with low sequence identity between interspecific homologs (Vieira and Rozas 2011) and can be further divided into four subfamilies: (1) *Classic OBPs*, which conform to the domain characterization above, (2) *PlusC* and *MinusC* OBPs, which contain six additional disulfide-bonded cysteine residues or lack two, respectively (Hekmat-Safe et al. 2002), and (3) *Atypical* OBPs, which contain two complete Classic OBP domains (e.g. “dimer OBPs”, [Vieira and Rozas 2011]). In a

recent study, Manoharan et al. (2013) expanded the number of known OBPs from the three published mosquito genomes by 110 members to a total of 289, while classifying each by subfamily. Ascribing function to peptides based on sequence homology to known OBPs can prove difficult. Leal et al. (2005) note that several gene families with OBP-like domain structure show no evidence of involvement in olfactory or pheromone-mediated responses, and suggests the term “encapsulins” supersede “odorant-binding proteins” to more accurately describe the common function (ligand encapsulation) performed by the peptide.

An additional protein family often included in evolutionary analyses of mosquito OBPs is the D7 salivary protein family, which exhibits domain structure similar to that of the OBPs with the addition of a seventh helix (Kalume et al. 2005). Classified into *short* (15-20 kDa) and *long* (30-36kDa) subfamilies, the long-form D7 salivary proteins contain a second OBP-like domain in an N-terminal extension (Valenzuela et al. 2002, Calvo et al. 2006). The singular domain in the short-form and C-terminus of the long-form salivary D7 protein has been shown to bind biogenic amines (serotonin, histamines and norepinephrine) with high affinity, while the N-terminal domain of the long-form protein binds leukotriene inflammatory mediators, thus inhibiting platelet aggregation, vasoconstriction and inflammation (collectively hemostasis) during blood-feeding (Calvo et al. 2006, Calvo et al. 2009a).

Our analysis identified sixteen proteins with an odorant binding cellular function (Table 1.S5), of which fourteen carried a Pfam ID of PF01395 (PBP/GOBP Family). Annotation of these proteins via Vectorbase reveals the list is comprised of six D7 salivary peptides, representing 60% of the known D7 proteins in the *Cx. quinquefasciatus*

genome (n=10, <https://www.vectorbase.org/organisms/culex-quinquefasciatus>) and eight odorant-binding proteins. The *Cx. quinquefasciatus* homologs of all OBPs in our set were recently classified by Manoharan et al. (2013), which allowed us to further assign our representatives to subfamily and cluster. Seven of the eight proteins were of the Classic OBP subfamily, i.e containing a singular OBP domain, with four of these being minus-C type and lacking two of the canonical cysteines.

These results indicate that the transcriptome of the two representative *Cx. pipiens* populations sequenced were most divergent within their odorant-binding domain-containing proteome at the D7 salivary proteins, and predominantly among the minus-C forms of the Classic Odorant-binding protein subfamily. Since the two forms differ in their propensity for taking mammalian (including human) vs. bird bloodmeals (Huang et al. 2008, Osório et al. 2014) the particular OBP subset highlighted here may contribute to the olfactory response to differing host cues. Additionally, the oviposition habitat available to subterranean mosquitoes (i.e. sewers) likely presents olfactory cues that differ from those above ground. The concomitant chemosensory response may necessitate evolution of OBP-encoding genes. As all but one OBP in our set were newly described by Manoharan et al. (2013) and were not included in the tissue-specific expression analysis of Leal et al. (2013), it is unknown whether they may be localized to antennae, palps or other somatic tissues. However, the representation of D7 salivary proteins in the enriched set may indicate that the immunosuppressive complement of mosquito saliva has diverged in accordance with local environment. The mosquito sialome has previously been shown to exhibit accelerated evolutionary pressures at the interspecific level; in a comparative analysis of New World (*An. darlingi*) and Old World

(*An. gambiae*) Anopheline sialotranscriptomes, Calvo et al. (2009b) found that on average, salivary proteins were only 53% identical at the amino acid level as opposed to 86% identity among housekeeping genes.

Components of the *extracellular space* (GO:0005615) gene ontology exist outside the cell plasma membrane within interstitial fluids. Our test set contained ten such proteins (Table 1.S6), with seven fibrinogens discussed above (and annotated as having extracellular localization) being re-listed here. The remaining three proteins were of the macroglobulin complement family, which carry alpha-2 macroglobulin family N-terminal (Pfam PF07703) and alpha-macroglobulin receptor (Pfam PF07677) domains. Alpha-2 macroglobulins ( $\alpha 2M$ ) are proteinase-binding and inhibiting glycoproteins commonly secreted by hemocytes within insect hemolymph (Sottrup-Jensen et al. 1989), which have been found recently to play integral roles in complement-like pathways that bind parasite surface targets (Blandin et al. 2008). The full-length protein exposes a “baited” peptide stretch, which when cleaved by proteinases present with septic injury will change protein conformation to an active state that covalently binds the activating proteinase. This conformational change also exposes binding sites with high affinity for both gram-positive and negative bacteria (Sottrup-Jensen et al. 1989, Blandin et al. 2008). The complex is then targeted for phagocytosis. Like the fibrinogens, the presence of these proteins in the most diverged set indicates that the two populations of *Cx. pipiens* may experience very different microbiome challenges, consistent with the differences between forms (e.g. utilization of sewers) in larval habitat (Harbach et al. 1984). Furthermore, as  $\alpha 2M$  inhibits the coagulation proteinases thrombin and factor Xa, it serves to inhibit the

coagulation cascade and thus may function in blood-feeding hemostasis as well (de Boer et al. 1993).

The *Chitin metabolic process* (GO:0006030) ontology (inclusive of all genes composing the *Chitin binding* (GO:0008061) ontology, Table 1.S7, 1.S8) is composed of reactions and pathways involving chitin, a linear polysaccharide polymer that consists of linked glucosamine residues and forms the main component of arthropod exoskeleton, tracheae and peritrophic membrane (PM). Seventeen proteins in the test set were annotated with this term; eleven with a Pfam Chitin binding Peritrophin-A domain (PF01607). The additional two peptides were annotated with a chitinase molecular function, each with two Pfam Chitinase class I domains (PF00182). Peritrophins are structural proteins consisting of one to many chitin-binding domains responsible for cross-linking chitin fibrils (Wang and Granados 2001). The semi-permeable lattice created, known as the peritrophic membrane, surrounds the insect food bolus and separates it from the midgut epithelial cells. This serves to protect the gut (and insect) from physical damage, pathogens and toxins. There is evidence that the PM plays a central role in binding toxic free heme via the chitin-binding domain (CBD) (Pascoa et al. 2002, Devenport et al. 2006) during bloodmeal digestion, indicating free CBDs on bound peritrophins of the PM serve additional purposes. Pascoa et al. (2002) found an amount of free heme bound to the *Aedes aegypti* PM equivalent to hydrolysis of 2ul of a typical 3ul bloodmeal. To determine whether our peptides were in fact peritrophins associated with a midgut PM, as opposed to non-specific cuticular proteins analogous to peritrophins (CPAPs, see Jasrapuria et al. [2010]) which also exhibit Peritrophin-A domain homologs, we aligned our nine candidate peptide domains to a selection of those

from the classification of Jasrapuria et al. (2010) and produced a maximum-likelihood tree which grouped all 21 of our sequences in a Peritrophic Matrix Protein (PMP) clade at a bootstrap support of 99% (Fig. 1.S1). This indicates our candidates are in fact likely associated with the midgut PM and involved in digestion.

Chitinases are integral enzymes in the creation and destruction of the adult mosquito PM. Initially synthesized as a zymogen upon ingestion of a blood meal, it is later activated by removal of a propeptide from the N-terminus (Bhatnagar et al. 2003) and begins to hydrolyze the glycosidic linkages of the PM chitin matrix to chitobiose (a glucosamine dimer) as the blood meal is digested. Like the PM itself, chitinase enzymes are important research targets for pathogen defense. The *Plasmodium* parasite ookinete expresses a mosquito chitinase ortholog able to accelerate PM degradation and facilitate escape (Langer and Vinetz 2001). Bhatnagar et al. (2003) were able to utilize the inhibitory effect of the propeptide on its cognate enzyme to block chitinase activity in both *Anopheles gambiae* and *Ae. aegypti*, thus suppressing development of human and avian *Plasmodium*, respectively, in the two mosquito species. Initial blood meal digestion within the female midgut requires transit of trypsins across the PM, and later, diffusion back to the ectoperitrophic space (Terra and Ferreira 1981). The peritrophic membrane has important dual-responsibilities in digestion and immunity, two systems we have associated with other enriched GO terms, further implicating this structure as a driving force in the differentiation of the two *Cx. Papiens* populations.

The insect immune system has been shown to be a common target of positive selection (Bulmer 2010, Roux et al. 2014), and the role it plays in differentiation of these two mosquitoes is further exemplified by examining the gene with the largest calculated

Ka in our comparison (Table 1.S1), a homolog of CPIJ006559 representing a peptidoglycan recognition protein (PGRP) containing a N-acetylmuramoyl-L-alanine amidase (Pfam PF01510). This particular PGRP (PGRP-LC) is a transmembrane molecule that, upon binding bacterial peptidoglycan, triggers the immune deficiency (Imd) pathway in *Drosophila* (Choe et al. 2005). A manual scan of our test set for other immune-related peptides that may bind peptidoglycan and/or carbohydrate yields eight proteins with a carbohydrate binding cellular function (GO:0030246) of which seven are lectins, with 5 annotated as salivary C-type lectins. These likely serve in food-borne pathogen identification (Valenzuela et al. 2002, Ribeiro et al. 2004) however the possibility exists that these proteins function instead as anti-clotting agents as has been reported in snake venom (Koo et al. 2002) and in the phlebotomine sand fly *Lutzomyia longipalpis* (Charlab et al. 1999).

#### *Highly conserved genes*

An enrichment test using the gene set devoid of non-synonymous substitutions from the f. pipiens – f. molestus comparison retained 19 significantly enriched GO terms (Table 1.2). These included primarily transcription and translational machinery (Structural constituent of ribosome, rRNA binding, Transcription regulatory region sequence-specific DNA binding), cell signaling components (GTP binding, GTPase mediated signal transduction, postsynaptic membrane, cell junction, G-protein coupled receptor signaling, outer membrane-bound periplasmic space, MAPK cascade, regulation of ion transmembrane transport) and ATP coupled proton transport (ATP hydrolysis coupled proton transport, ATP synthesis coupled proton transport, proton-transporting V-



type ATPase). Of particular interest were the four GO terms for which all members were present in the enriched set only (i.e the GO term constituents contained only synonymous substitutions; Table 1.S13): (1) outer-membrane bound periplasmic space (GO0030288) contained glutamate receptors responsible for postsynaptic excitation of insect neuronal and muscle cells (Briley et al. 1981), (2) the MAPK cascade (GO0000165) that communicates biotic and abiotic signals from extracellular ligands to the nucleus, initiating a response (e.g division, apoptosis, etc.) from the cell (McKay and Morrison 2007), proton-transporting V-type ATPases (GO0033180) that are a diverse and highly conserved membrane-spanning enzyme coupling ATP hydrolysis to proton transport (Nelson et al. 2000), and (4) the transcription regulatory region sequence-specific DNA binding ontology (GO0000976) that contains several homeobox domains encoding transcription factors which activate and regulate patterns of morphogenesis (Gehring 1992). Several of these pathways have been previously described as highly conserved in eukaryotes (Bejerano et al. 2004, Li et al. 2011), and when taken together define a genetic “core” in *Cx. pipiens* that confer critical phenotypes and cellular processes refractory to amino acid substitutions and are the strongest candidates for negative or purifying selective pressures.

#### *Comparison between geographically isolated populations*

The populations of *Cx. pipiens* forms *pipiens* and *molestus* mosquitoes sequenced in this study were geographically isolated. To assess how the amount of variation between *Cx. pipiens* forms (defined by number and IDs of enriched GO terms) reported in our analyses compared to conspecific isolated populations, we repeated our analysis

using publicly available data from a recently colonized population of *Cx. quinquefasciatus* isolated from Alabama, USA (Reid et al. 2012) and the CpipJ1.3 Johannesburg reference CDS sequences. Short-read mapping produced alignments covering 17,410 of 19,019 CDS sequences with > 1 read and covered 19.8 Mbp (79%) of the reference, with average coverage of 133x and median coverage of 14.7x. After applying the length and 2x coverage cutoff (see Methods), we retained 13,586 CDS codon alignments for analysis with the 95<sup>th</sup> percentile test set composed of 679 sequences (Table 1.S9). Blast2GO analysis retained only two significant GO terms when reduced to the most-specific set (Table 1.S10). Neither of these terms (both composed of the same seven genes encoding reverse transcriptase enzymes and retrotransposons) are present in our *Cx. pipiens* comparison, indicating that the *f. pipiens* – *f. molestus* populations sampled here maintain a greater degree of evolutionary protein divergence than the isolated yet conspecific *Cx. quinquefasciatus* populations.

#### *Assessing effects of paralogy and sequence identity*

Some of the gene families and protein domains reported in this study are among the most abundant in the mosquito genome. For example, Interproscan5 analysis of the CpipJ1.3 transcripts (not shown) uncovers 477 trypsin and 283 peritrophin-A domains. Even though we discarded sequencing reads with multiple top-scoring genome alignments, to ensure our results were not reflective of incorrect short read placement among multiple paralogous genes, we tested the propensity for our reported GO terms to be enriched among those genes that share significant sequence identity to others in the genome. Using all CpipJ1.3 CDS sequences with BLASTN alignments  $\geq 200$ bp at  $\geq 95\%$  similarity

to another CDS in the genome (Table 1.S11), we derived a test set which contained 41 enriched terms (Table 1.S12). This list contained no GO terms previously reported here, thus we find no evidence that the resultant terms from our *f. pipiens* – *f. molestus* comparison originate from gene families with biased sequence identity.

## Conclusions

These are the first insights into the genome-wide molecular differentiation of two closely related yet phenotypically divergent populations of an important disease vector, *Cx. pipiens*. Analysis of over-represented gene ontology terms within the fastest evolving peptides elucidates the biological systems that are targets of local adaptation. Although further analyses with additional representative populations of the two forms are necessary, our results likely hold clues as to the molecular mechanisms responsible for phenotypic divergence between the two taxonomic forms, and subsequently confer *Culex pipiens* form *molestus* the ability to exploit human environments. The recurring localizations within our data to gene families functioning in odorant binding, hemostasis, digestion, and innate immunity can all be linked to a differential propensity of these forms to seek a mammalian host, ability to obtain and process a bloodmeal, and to thrive as larvae and adults in subterranean sewers rich with organic wastes and associated bacteria. In addition, we provide candidate loci for future functional in-vivo assays to qualify effects on phenotype. Interestingly, of the seven GO terms reported in this study, five terms (chitin metabolic process, chitin binding, serine-type endopeptidase activity, proteolysis and odorant binding) were enriched along the ‘fly’ branch (represented by the *Drosophila melanogaster* genome (Adams et al. 2000)) of the branch-site selection tests

conducted by Roux et al. (2014), indicating they may represent a genetic ‘core’ remaining under selection and responsible for adaptive evolution within the Diptera. Further sequencing of members of the *Culex pipiens* complex (Farajollahi et al. 2011) will enable additional tests involving lineage-specific estimates of evolutionary rates (e.g. Mensch et al. [2013]) and definition of functional classes of genes with significantly elevated selection coefficients as compared to ancestral states in the phylogeny (Serra et al. 2011), as well as defining the role of differential gene expression in the divergence of a global mosquito vector.

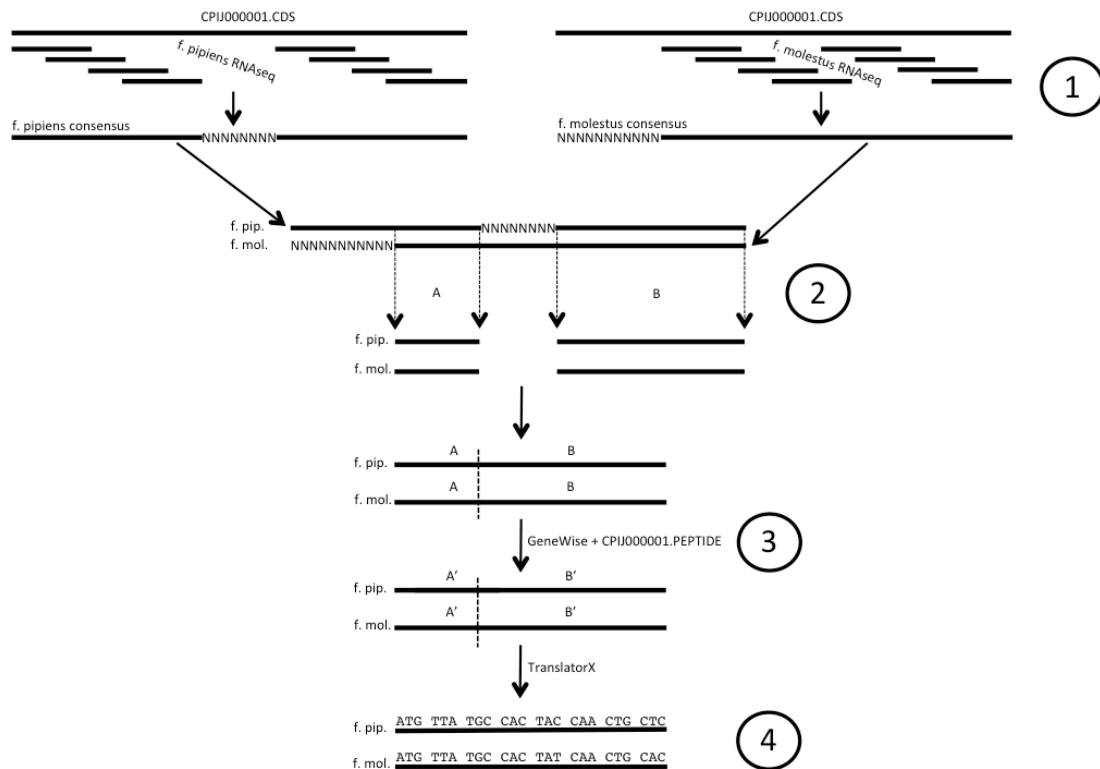
#### *Availability of supporting data*

The Illumina sequencing libraries for *Cx. pipiens* forms *pipiens* and *molestus* are available via NCBI BioProject PRJNA275017. The pairwise codon alignments can be downloaded from the PeerJ website via <http://dx.doi.org/10.7717/peerj.807/supp-15>

#### **Acknowledgements**

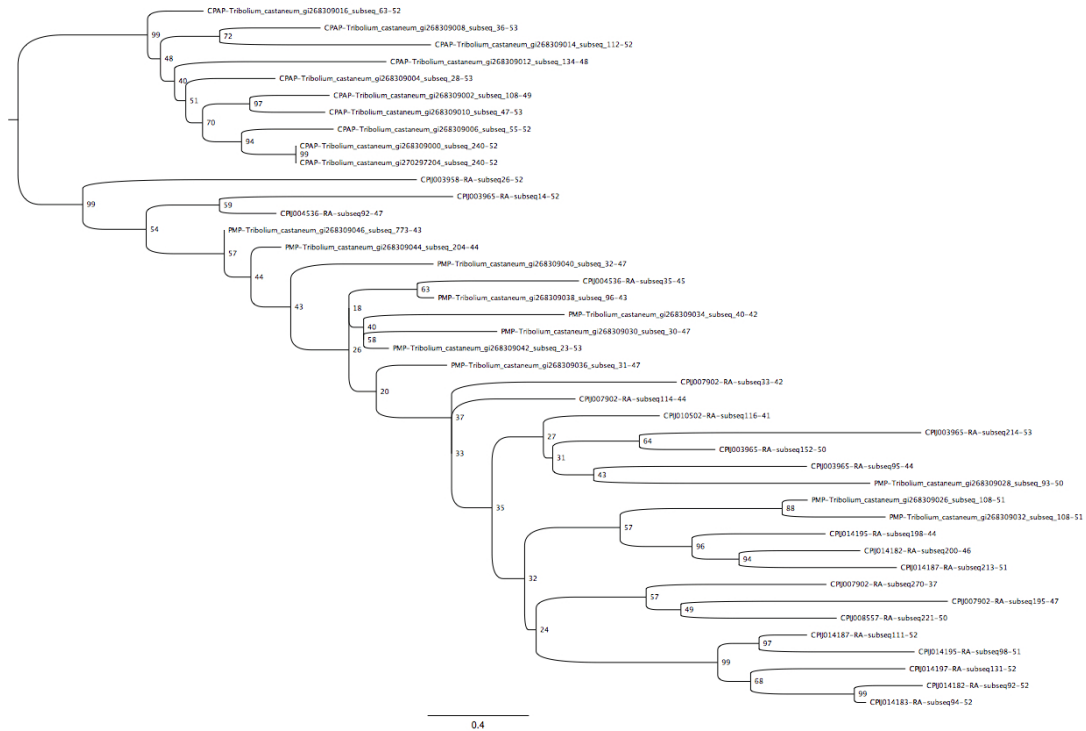
We are grateful to Linda McCuiston for her unsurpassed expertise in rearing and colonizing the mosquitoes used in our study, to Nicole Wagner at the Rutgers University School of Environmental and Biological Sciences Genome Cooperative for performing our Illumina Sequencing, and to Peter Armbruster for comments on the manuscript.

**Figure 1.1** Illustration of codon alignment generation process. 1. Illumina short read data are aligned to *Cx. quinquefasciatus* reference CDS sequence and used to build consensus sequences for both *Cx. pipiens* forms *pipiens* and *molestus*. 2. Consensus sequences for each gene are aligned, homologous positions free of Ns are removed and spliced. 3. GeneWise is used along with the corresponding full length *Cx. quinquefasciatus* peptide to create in-frame *f. pipiens*/*f. molestus* EST sequences from spliced alignments. 4. Codon alignments are created from EST sequences using TranslatorX. Ns denote unknown and/or unrecovered nucleotide data.



**Figure 1.S1** Maximum-likelihood phylogenetic tree showing monophyly of peritrophin-A domains reported here with peritrophic matrix proteins (labeled PMP), exclusive of the cuticular proteins analogous to peritrophins (labeled CPAP) of Jasrapuria et al. (2010). NCBI GI numbers are appended to *Tribolium castaneum* sequence IDs; all sequences are suffixed with "\_subseq\_[coordinate of first amino acid extracted]-[length of extracted peptide window]".

NOTE - An enlarged version of this figure is available in the dissertation supplement.



**Table 1.S1** Observed and estimated Ka calculations, annotation and top-scoring Pfam IDs corresponding with 11,931 pairwise *Culex pipiens* forms *pipiens* and *molestus* homologous codon sequence alignments (ordered by decreasing Ka). Columns two and three denote genes present in the 95th percentile as ranked by Ka calculated using observed and likelihood estimated non-synonymous substitutions, respectively.

Table 1.S1 is located in the online supplementary material.

**Table 1.1** Gene ontology terms enriched in the upper 95th percentile of pairwise dN values calculated using *Culex pipiens* forms *pipiens* and *molestus* homologous codon sequence alignments

GO ID	Go Term	FDR	p	# in test set	# in ref. set	# unannotated test set	# unannotated reference set
GO:0004252	serine-type endopeptidase activity	1.20E-13	7.60E-17	51	232	364	7988
GO:0006508	proteolysis	1.40E-09	1.80E-12	71	546	344	7674
GO:0005102	receptor binding	7.50E-09	1.50E-11	25	80	390	8140
GO:0005549	odorant binding	1.40E-06	3.20E-09	16	39	399	8181
GO:0005615	extracellular space	7.30E-04	2.00E-06	10	23	405	8197
GO:0006030	chitin metabolic process	5.80E-03	1.70E-05	17	93	398	8127
GO:0008061	chitin binding	1.20E-02	4.80E-05	15	81	400	8139



**Table 1.S11** BLASTN output detailing the 3,687 *Culex quinquefasciatus* CDS sequences with at least one BLASTN alignment  $\geq 200$ bp at  $\geq 95\%$  similarity to another CDS in the genome

Table 1.S11 is located in the online supplementary material.

**Table 1.S2** Gene set composing the serine-type endopeptidase ontology, found to be enriched in the 95th percentile of top-scoring *Culex pipiens* forms *pipiens* and *molestus* homologous codon sequence alignments as ranked by Ka value.

Table 1.S2 is located in the online supplementary material.

**Table 1.S3** Gene set composing the proteolysis ontology, found to be enriched in the 95th percentile of top-scoring *Culex pipiens* forms *pipiens* and *molestus* homologous codon sequence alignments as ranked by Ka value

Table 1.S3 is located in the online supplementary material.

**Table 1.S4** Gene set composing the receptor binding ontology, found to be enriched in the 95th percentile of top-scoring *Culex pipiens* forms *pipiens* and *molestus* homologous codon sequence alignments as ranked by Ka value.

Table 1.S4 is located in the online supplementary material.

**Table 1.S5** Gene set composing the odorant binding ontology, found to be enriched in the 95th percentile of top-scoring *Culex pipiens* forms *pipiens* and *molestus* homologous codon sequence alignments as ranked by Ka value. Column ‘MEA’ denotes Manoharan et al. 2014 classification.

Table 1.S5 is located in the online supplementary material.

**Table 1.S6** Gene set composing the extracellular space ontology, found to be enriched in the 95th percentile of top-scoring *Culex pipiens* forms *pipiens* and *molestus* homologous codon sequence alignments as ranked by Ka value.

Table 1.S6 is located in the online supplementary material.

**Table 1.S7** Gene set composing the chitin binding ontology, found to be enriched in the 95th percentile of top-scoring *Culex pipiens* forms *pipiens* and *molestus* homologous codon sequence alignments as ranked by Ka value.

Table 1.S7 is located in the online supplementary material.

**Table 1.S8** Gene set composing the chitin metabolic process ontology, found to be enriched in the 95th percentile of top-scoring *Culex pipiens* forms pipiens and molestus homologous codon sequence alignments as ranked by Ka value.

Table 1.S8 is located in the online supplementary material.



**Table 1.2** Gene ontology terms enriched in the set of 4,575 pairwise *Culex pipiens* forms *pipiens* and *molestus* homologous codon

alignments devoid of non-synonymous substitutions. Asterisks indicate terms for which all members were present only in the test set.

GO ID	Go Term	FDR	p	# in test set	# in ref. set	# unannotated, test set	# unannotated, reference set
GO:0003735	structural constituent of ribosome	1.00E-15	7.60E-19	98	30	3209	5298
GO:0005525	GTP binding	2.70E-09	1.70E-11	113	67	3194	5261
GO:0007264	small GTPase mediated signal transduction	2.90E-04	6.40E-06	104	89	3203	5239
GO:0051301	cell division	2.50E-03	8.60E-05	27	12	3280	5316
GO:0007186	G-protein coupled receptor signaling pathway	4.10E-03	1.50E-04	76	66	3231	5262
GO:0003924	GTPase activity	6.60E-03	2.50E-04	54	42	3253	5286
GO:0030288	outer membrane-bounded periplasmic space*	1.10E-02	4.60E-04	8	0	3299	5328
GO:0030054	cell junction	1.30E-02	5.70E-04	28	16	3279	5312
GO:0004930	G-protein coupled receptor activity	1.80E-02	8.10E-04	45	35	3262	5293
GO:0000165	MAPK cascade*	2.50E-02	1.20E-03	7	0	3300	5328
GO:0006334	nucleosome assembly	2.70E-02	1.40E-03	18	8	3289	5320
GO:0005509	calcium ion binding	3.10E-02	1.60E-03	103	110	3204	5218
GO:0015991	ATP hydrolysis coupled proton transport	3.50E-02	1.90E-03	14	5	3293	5323
GO:0019843	rRNA binding	3.50E-02	1.90E-03	10	2	3297	5326
GO:0015986	ATP synthesis coupled proton transport	3.50E-02	1.90E-03	10	2	3297	5326
GO:0034765	regulation of ion transmembrane transport	3.80E-02	2.10E-03	15	6	3292	5322
GO:0045211	postsynaptic membrane	4.50E-02	2.70E-03	19	10	3288	5318
GO:0033180	proton-transporting V-type ATPase, V1 domain*	5.00E-02	3.10E-03	6	0	3301	5328
GO:0000976	transcription regulatory region sequence-specific DNA binding*	5.00E-02	3.10E-03	6	0	3301	5328

**Table 1.S13** Extended analysis for all genes belonging to the GO terms from the highly conserved set (Table 1.2) for which all members were present only in the test set.

Table 1.S13 is located in the online supplementary material.

**Table 1.S9** Ka calculations corresponding with 13,587 pairwise *Culex quinquefasciatus* strain HAmCq and CpipJ1.3 homologous codon sequence alignments. Column two denotes genes present in the 95th percentile as ranked by Ka calculated using observed non-synonymous substitutions.

Table 1.S9 is located in the online supplementary material.

**Table 1.S10** Gene ontology terms enriched in the upper 95th percentile of pairwise Ka values calculated using *Culex quinquefasciatus* strains HAmCq and CpipJ1.3 homologous codon sequence alignments.

Go Term	Adjusted p-val	p-val	# in test group	# in reference group
RNA-dependent DNA replication	1.50E-02	1.10E-05	7	11
RNA-directed DNA polymerase activity	1.50E-02	1.10E-05	7	11

**Table 1.S12** Gene ontology terms enriched in the set of 3,687 *Culex quinquefasciatus*

CDS sequences with at least one BLASTN alignment > 200bp at > 95% homology to another CDS in the genome.

GO ID	Go Term	Adjusted p-val	p-val	# in test group	# in reference group
GO:0000786	nucleosome	7.60E-38	3.70E-41	99	34
GO:0006334	nucleosome assembly	3.50E-37	9.40E-40	107	47
GO:0046982	protein heterodimerization activity	2.60E-28	1.30E-30	87	43
GO:0005811	lipid particle	3.10E-11	3.60E-13	18	0
GO:0006426	glycyl-tRNA aminoacylation	2.80E-07	4.60E-09	17	4
GO:0004820	glycine-tRNA ligase activity	2.80E-07	4.60E-09	17	4
GO:0006420	arginyl-tRNA aminoacylation	1.40E-06	2.50E-08	11	0
GO:0004814	arginine-tRNA ligase activity	1.40E-06	2.50E-08	11	0
GO:0003677	DNA binding	1.60E-06	2.90E-08	275	732
GO:0003735	structural constituent of ribosome	4.80E-05	9.40E-07	91	188
GO:0050567	glutamyl-tRNA synthase (glutamine-hydrolyzing) activity	2.30E-04	5.00E-06	11	3
GO:0030956	glutamyl-tRNA(Gln) amidotransferase complex	2.30E-04	5.00E-06	11	3
GO:0070681	glutamyl-tRNA(Gln) biosynthesis via transamidation	2.30E-04	5.00E-06	11	3
GO:0006310	DNA recombination	3.20E-04	7.20E-06	22	21
GO:0031409	pigment binding	5.00E-04	1.10E-05	12	5
GO:0003910	DNA ligase (ATP) activity	5.00E-04	1.10E-05	12	5
GO:0030060	L-malate dehydrogenase activity	5.50E-04	1.30E-05	14	8
GO:0016075	rRNA catabolic process	9.10E-04	2.20E-05	9	2
GO:0032543	mitochondrial translation	1.60E-03	3.90E-05	11	5
GO:0005643	nuclear pore	1.90E-03	5.20E-05	25	32
GO:0030301	cholesterol transport	2.60E-03	7.30E-05	9	3
GO:0042302	structural constituent of cuticle	2.80E-03	8.20E-05	63	134
GO:0006108	malate metabolic process	3.20E-03	9.50E-05	14	11
GO:0005524	ATP binding	5.80E-03	1.70E-04	308	955
GO:0004525	ribonuclease III activity	6.30E-03	1.90E-04	9	4
GO:0030127	COPII vesicle coat	1.10E-02	3.70E-04	11	8
GO:0003857	3-hydroxyacyl-CoA dehydrogenase activity	1.20E-02	4.20E-04	6	1
GO:0051103	DNA ligation involved in DNA repair	1.20E-02	4.20E-04	6	1
GO:0004527	exonuclease activity	1.60E-02	5.80E-04	21	30
GO:0008158	hedgehog receptor activity	4.10E-02	1.70E-03	9	7
GO:0006433	prolyl-tRNA aminoacylation	4.10E-02	1.70E-03	4	0
GO:0004827	proline-tRNA ligase activity	4.10E-02	1.70E-03	4	0
GO:0045028	G-protein coupled purinergic nucleotide receptor activity	4.10E-02	1.70E-03	4	0
GO:0035589	G-protein coupled purinergic nucleotide receptor signaling pathway	4.10E-02	1.70E-03	4	0
GO:0008113	peptide-methionine (S)-S-oxide reductase activity	4.10E-02	1.70E-03	4	0
GO:0004602	glutathione peroxidase activity	4.10E-02	1.80E-03	5	1
GO:0015876	acetyl-CoA transport	4.20E-02	1.90E-03	11	11
GO:0008521	acetyl-CoA transporter activity	4.20E-02	1.90E-03	11	11
GO:0005097	Rab GTPase activator activity	4.20E-02	1.90E-03	16	22
GO:0032851	positive regulation of Rab GTPase activity	4.20E-02	1.90E-03	16	22
GO:0070403	NAD+ binding	4.80E-02	2.20E-03	7	4

## Chapter 2:

### Genome mosaicism and evolutionary divergence within the Asian *Culex pipiens* complex

#### Abstract

Hybrid speciation provides evolutionary biology with a unique opportunity to study the role of adaptive evolution over short time scales. Given the current “postgenomic era” which has seen genome sequencing become commonplace, it then follows that hybridization phenomena allow research into molecular mechanisms responsible for adaptive traits within hybrid species that persist. We generated transcriptome data for *Culex pipiens pallens*, an East Asian member of the *Culex pipiens* complex that is hypothesized to be a hybrid between the Northern and Southern house mosquitoes, *Culex pipiens* and *Culex quinquefasciatus*. We use a phylogenomic analysis to identify shared ancestry between *Cx. pipiens pallens* gene sequences and those of either putative parental genome. Additionally we identify genes and gene ontologies that show evidence of evolving at accelerated evolutionary rates among East Asian *Culex* (*Cx. pipiens pipiens*, *Cx. pipiens pallens* and *Cx. quinquefasciatus*) by calculating per-gene rates of peptide evolution, and identifying lineages with differential rates of evolution to examine how *Cx. pipiens pallens* has utilized and modified parental genes to exploit its environment and persist as a species. We show the genetic component of the *Cx. pipiens pallens* proteome assigned to *Cx. pipiens* contains genes that function in energy metabolism, cell cycle / signaling, and redox reactions; the genes assigned to *Cx. quinquefasciatus* are enriched in lipid transport function and extracellular scavenging / innate immunity. Our analyses show that *Cx. pipiens pallens* and *Cx. quinquefasciatus*

share a greater degree of phylogenetic affiliation and lower protein divergence than either do with *Cx. pipiens form molestus*. Moreover, gene ontology enrichment shows a substantial amount of divergence at common loci between *Cx. pipiens pipiens* and (*Cx. pipiens pallens* + *Cx. quinquefasciatus*), yet that divergence between *Cx. pipiens pallens* and *Cx. quinquefasciatus* is unique to their comparison. These data are consistent with a close “degree of relatedness” between *Cx. pipiens pallens* and *Cx. quinquefasciatus*, as would be imparted by hybridization.

## Introduction

Hybrid speciation, or the formation of a new species via hybridization of two different parental species (Mallet 2007) is believed to be infrequent in the Metazoa (Mallet 2007, Mavárez and Linares 2008). Referred to by R.A Fisher (1930) as “the grossest blunder in sexual preference which we can conceive of an animal making”, the phenomenon has recently lead to fascinating studies on ecological adaptation and accelerated adaptive radiations in well studied populations such as Darwin’s finches (Grant and Grant 2002) and cichlid fish (Joyce et al. 2011). Incipient homoploid (i.e maintaining parental ploidy) species formed as a result of hybridization are likely to be at an immediate disadvantage due to competitive exclusion and/or species fusion (Abbott and Rieseberg 2012) with the parental lineage(s). Those hybrids that do persist and partition resources are of clear scientific interest, as the selective advantage imparted over a brief evolutionary time may hold keys to rapid radiation and niche adaptation.

The *Cx. pipiens* complex contains four species that are evolutionarily closely related and often difficult to identify: *Cx. pipiens* Linnaeus, *Culex quinquefasciatus* Say,

*Culex australicus* Dobrotworsky & Drummond and *Culex globocoxitus* Dobrotworsky (Farajollahi et al. 2011) (see <http://www.wrbu.si.edu> for current taxonomy). Male mosquitoes within the complex can be identified to species via external genitalia (Dobrotworsky 1967), however females cannot. Hybridization in North America is pervasive between *Cx. pipiens* and *Cx. quinquefasciatus* in areas where both have been introduced (Kothera et al. 2013, Silverbush and Sharan 2014). In East Asia, a temperate subspecies of *Cx. pipiens*, *Cx. pipiens pallens*, exhibits an intermediate or hybrid male phallosome that follows a morphological cline from “*pipiens*-like” in the northern latitudes to “*quinquefasciatus*-like” in the south (Bekku 1956). For this reason, *Cx. pipiens pallens* has long been hypothesized to be a hybrid of *Cx. pipiens* and *Cx. quinquefasciatus* (Barr 1957, Tanaka et al. 1979). Fonseca et al. (2009) however found that *pallens* mosquitoes have a distinctive genotypic signature at both microsatellite and nuclear loci even if there is evidence of ongoing hybridization with *Cx. quinquefasciatus* in southern regions of Japan and the Korean peninsula. Additionally, they found that all male *Cx. p. pallens*, even in northern populations, contain both the diagnostic *pallens* *ace-2* allele and a second copy identical to that of *Cx. quinquefasciatus*. These data point to an ancient and independent origin of *Cx. pipiens pallens* resulting from a hybridization event between *Cx. quinquefasciatus* (as evidenced by the *ace-2* “relic”) and Eastern European *Cx. pipiens* (evidenced by external genitalia morphology and similarity of ecology and behavior [Fonseca et al. 2009]).

During the mid-twentieth century, *Cx. p. pipiens* form molestus was introduced to Tokyo, Japan (likely during World War II [Tanaka et al. 1979]), and subsequently spread to all of temperate Japan, eastern Russia, Korea and Beijing, China by 1992 (summarized



in Mogi [2012]). As the old world *Culex p. pipiens* form *pipiens* has not spread East of northwestern Xinjiang Province, China (Lu 1997), form *molestus* represents the sole *Cx. p. pipiens* biotype in East Asia. Therefore, three members of the *Culex pipiens* complex occur in Eastern Asia: *Culex p. pallens* in the north; *Culex quinquefasciatus* in the south and *Cx. p. pipiens* form *molestus* in urban areas.

The natural hybridization (Arnold 1997) that resulted in the homoploid *Cx. pipiens pallens* presumably conferred a fitness advantage, as evidenced by its ability to secure resources and persist. This event affords us a unique opportunity to elucidate the genomic basis for this beneficial phenotype. Here we define the genome mosaicism within *Cx. pipiens pallens* by constructing denovo transcriptome coding sequences and identifying genes that remain phylogenetically attributable to either donor species (*Cx. pipiens*, *Cx. quinquefasciatus*) via statistically supported monophyly with either taxon in a phylogenomic analysis. Additionally, we performed pairwise comparisons of orthologous coding nucleotide sequences to identify genes and gene ontologies that show evidence of evolving at accelerated evolutionary rates among East Asian *Culex* (*Cx. p. pipiens*, *Cx. p. pallens* and *Cx. quinquefasciatus*) by calculating per-gene rates of non-synonymous substitution per non-synonymous site (Ka, or dN). This method has been used previously to identify peptide evolution within and outside of the mosquitoes (Wang et al. 2011, Price and Fonseca 2015). By qualifying and quantifying shared ancestry and lineage-specific divergence within the *Culex pipiens* complex, our results highlight genes and corresponding biological functions that may be central to adaptation within the hybrid *Cx. p. pallens*, and evolutionary divergence within a complex of medically important and globally distributed mosquitoes.

## Methods

*Cx. p. pallens* and *Cx. pipiens* form molestus mosquitoes were obtained from young colonies initiated from individuals collected from Kyoto, Japan in Sept. of 2008, and from a large subterranean swarm of blooded females detected in a New York City, USA residential basement in December 2010, respectively. PCR-based positive species identification of *Cx. p. pallens* was performed via the acetylcholinesterase-2 assay developed by Smith and Fonseca (2004). RNA extraction and Illumina library preparation for each species were carried out as described in Price and Fonseca (2015): Larvae of both were reared in ceramic pans under a 16:8 L:D cycle on a diet of ground rat chow prior to emergence. Four specimen groups were created: thirty 1<sup>st</sup>/2<sup>nd</sup> instar, eight 3<sup>rd</sup>/4<sup>th</sup> instar, eight pupae and eight non-blood fed adult (4 male, 4 female) mosquitoes. Each group was placed in a separate plastic 2ml microcentrifuge tube containing a 5mm sterile stainless steel bead and 900ul QIAzol lysis reagent prior to disruption with a TissueLyser II (Qiagen, Valencia CA) for 2 minutes at 20Hz. Total RNA extraction was then carried out on each group using the RNeasy Plus Universal kit (Qiagen, Valencia CA) per manufacturer protocol and quantified on a Qubit 2.0 fluorometer (Life Technologies) using the RNA Broad-range buffer. One ug of RNA from each group was combined and used to prepare an Illumina sequencing library using the TruSeq RNA Sample Prep kit v2 (Illumina, Inc. San Diego, CA) per manufacturer protocol. Both libraries were sequenced twice on an Illumina MiSeq (Illumina, Inc): once using a 500-cycle (2x250bp paired-end) MiSeq Reagent Kit v2, and once using 1/3 of a multiplexed 600-cycle (2x300bp paired-end) MiSeq Reagent Kit v3. Raw sequence data were quality

trimmed using the CLC Genomics Workbench (Limit score cutoff = 0.05, CLC Bio, Aarhus, DK). EST sequences were assembled using the *Cx. quinquefasciatus* nucleotide transcripts as a reference, as described in Price and Fonseca (2015). Briefly, raw reads for each species were individually mapped to the *Cx. quinquefasciatus* genome CDS sequence, extracted from the CpipJ1.3 genome assembly available via VectorBase (<http://www.vectobase.org/organisms/Culex-quinquefasciatus>; Megy et al. [2012]) using the CLC Genomics Workbench (CLC Bio, Aarhus, DK) at a nucleotide similarity of 95% over a required length fraction of 95% of the read. Reads that had more than one best alignment (i.e., potentially paralogous DNA) were ignored. Consensus sequences for each CDS were then generated from the alignment, with conflicts resolved by choosing the base with the highest additive quality score and a minimum coverage of 2x. Areas of <2x coverage were filled with Ns from the reference. In-frame coding sequences were created for each EST using TranslatorX (Abascal et al. 2010) with the *Cx. quinquefasciatus* peptide homolog as a guide.

Our phylogenetic tree-based analyses require constraint trees for probability assignment of ingroup (*Culex* spp) monophyly. We therefore added a fourth proteome, that of *Aedes aegypti* (Nene et al. 2007) to achieve a 4-taxon bifurcating tree. *Ae. aegypti* was chosen as it is the closest relative to *Culex* with a fully sequenced and/or predicted genome. A reciprocal BLASTp (Altschul et al. 1990, Moreno-Hagelsieb and Latimer 2008) was used to determine *Cx. quinquefasciatus* – *Ae. aegypti* ortholog pairs. The *Cx. pipiens* form *molestus* and *Cx. p. pallens* coding sequences were translated to amino acids and aligned with the corresponding *Cx. quinquefasciatus* and *Ae. aegypti* homologous protein via MAFFT v.6.9 (Katoh and Toh 2010). This peptide alignment

was then used to produce a codon alignment of all four nucleotide CDS sequences via TranslatorX. All gaps were removed from the alignment, thus only sites with aligned codons present in all four species were used. Alignments less than 200nt were removed from the dataset unless they represented greater than 50% of the length of the corresponding *Cx. quinquefasciatus* reference gene.

A model of sequence evolution was chosen for each alignment using jModelTest v.2.1.6 (Darriba et al. 2012) and three maximum-likelihood phylogenetic trees were generated via PhyML v.20120412 (Guindon et al. 2010) using the model chosen and under each of three topological constraints:

1. (pallens,molestus),quinquefasciatus,aegypti 2.

(pallens,quinquefasciatus),molestus,aegypti and 3.

(molestus,quinquefasciatus),pallens,aegypti. Consel v0.20 (Shimodaira and Hasegawa 2001) was used to assess confidence and assign a probability value to each of the three trees using the PhyML-generated site log-likelihood values. Additionally, a neighbor-joining (F84 distance) tree was constructed using the PHYLIP package v.3.69

(Felsenstein 1989), and a majority-rule consensus was generated using 1000 bootstrap replicates. We retained gene trees with Consel-generated Approximately Unbiased (AU) test (Shimodaira 2002) p-values  $\geq 0.90$  that congrued with the neighbor-joining topology, and grouped them based on the topological constraints above. Each group was then used as a test set in a Blast2GO (Conesa et al. 2005) enrichment analysis (Fisher's Exact test), with a reference set consisting of the 8,961 proteins for which we were able to derive trees. All GO terms returned in the analysis failed the False Discovery Rate multiple test correction (Benjamini and Hochberg 1995), however the FDR correction is often

considered too conservative (i.e Huang et al. [2009]); given the close relationship of our ingroup taxa, some test sets contained a comparatively small number of elements (118 genes from the [molestus + quinquefasciatus] topology, 299 genes from the [molestus + pallens] topology). We therefore made the decision to use a single-test p-value cutoff of  $1.0 \times 10^{-2}$ .

To identify the subset(s) of genes and gene ontologies that were most highly diverged between each species, we created two-taxon (pairwise) codon alignments using the 4-taxon codon alignments above for all genes from each of the three species groupings (*Cx. p. pipiens* form molestus + *Cx. p. pallens*, *Cx. p. pallens* + *Cx. quinquefasciatus* and *Cx. p. pipiens* form molestus + *Cx. quinquefasciatus*). We then used the KaKs calculator (Wang et al. 2010) under model averaging (MA) to calculate Ka values for each pairwise alignment. The KaKs calculator implements the calculation in a likelihood framework that corrects for multiple substitutions at sites. As the *Culex* taxa examined are all very closely related, and multiple substitutions at any one site are unlikely to be common enough to bias our findings, we performed primary calculations using observed substitutions in the data. The gene sets for each of the three comparisons were then ranked by descending Ka value, and the top 5% (95<sup>th</sup> percentile, or ‘diverged set’) of each were retained for further analyses both alone and against each other. Those genes found to be present in the diverged set of all three comparisons (i.e under accelerated peptide evolution in all *Culex* comparisons) were used in a Blast2GO enrichment analysis as above. This was repeated using members of the diverged set found unique to the *pallens* – *quinquefasciatus* comparison and those shared uniquely between the molestus – *pallens* and molestus – *quinquefasciatus* comparisons, as they

constitute lineage-specific divergence. The reference gene set for these tests consisted of the 8,961 peptides for which alignments were created (above). GO terms with a single test p-value of  $1 \times 10^{-2}$  were retained.

To test for differences in the mean protein evolutionary rate (Ka) among the genes as grouped by phylogenetic affiliation above and by pairwise comparison group, we used a non-parametric Kruskal-Wallis test implemented in R (<http://www.R-project.org>) as the data were severely positive-skewed.

## Results and Discussion

Paired-end Illumina sequencing of *Cx. pipiens pallens* mRNA generated 46.9M 150x150 (300bp) and 25.9M 150x150 (300bp) and 300x300 (600bp) reads that mapped to 19.7Mbp (79% by length, average coverage = 137x) of the total *Cx. quinquefasciatus* reference coding (CDS) nucleotide sequence. Using the 10,585 *Ae. aegypti* – *Cx. quinquefasciatus* orthologs derived from the reciprocal BLASTp (not shown), and the consensus nucleotide CDS sequences from our reference-based assembly of *Cx. pipiens form molestus* and *Cx. pipiens pallens*, we constructed 9,043 multiple-sequence codon alignments. Each alignment contained only homologous codons from each of the four target taxa.

### *Phylogenetic assignment of Cx. pipiens pallens proteome*

Length filtering and maximum-likelihood (ML) phylogenetic tree construction produced 8,352 (Table 2.S1) trees, and the confidence of each was assessed by CONSEL. Trees that yielded the same topology under both ML and neighbor-joining and passed the

AU significance test at  $p > .90$  were retained. As our ingroup (*Culex spp.*) taxa are very closely related, many trees failed significance testing; our results produced 299 gene trees grouping *Cx. pipiens* form molestus + *Cx. pipiens pallens*, 1,011 trees grouping *Cx. pipiens pallens* + *Cx. quinquefasciatus* and 118 trees grouping *Cx. pipiens* form molestus + *Cx. quinquefasciatus* (see Table 2.S2). The genes composing each group were used as individual test sets in a Blast2GO enrichment test.

At a single test p-value cutoff of  $1 \times 10^{-2}$  and using the 8,352 gene set (above) as a reference, we retained seven GO terms enriched in the genes retained with (*Cx. p. pipiens* form molestus + *Cx. p. pallens*) tree topologies (Table 2.1): *protein phosphatase type 2A regulator activity*, which includes enzymes that regulate removal of phosphate from serine and threonine side chains and influence cellular apoptosis, proliferation and differentiation (Zolnierowicz 2000); *L-alanine:2-oxoglutarate aminotransferase activity*, involved in energy (protein, carbohydrate) metabolism (Weeda et al. 1980); *selenium binding*, which modulates redox homeostasis (Porat et al. 2000), *phosphatidylinositol N-acetylglucosaminyltransferase activity*, containing enzymes involved in synthesis of glycosylphosphatidylinositol (GPI) protein anchors (glycolipid anchors for plasma-membrane proteins [Low and Kincade 1985, Porat et al. 2000]); *phosphoglycolate phosphatase activity*, a haloacid dehydrogenase family of enzymes with varied functions including amino acid biosynthesis and detoxification (Koonin and Tatusov 1994); *NADP binding*, containing proteins that bind the coenzyme NADP during redox and biosynthetic reactions, and *regulation of cyclin-dependent protein serine/threonine kinase activity* that was comprised of three cyclins that regulate the cell cycle (Minshull et al. 1989).

Four enriched terms were retained from the gene set that maintained a *Cx. p. pallens* + *Cx. quinquefasciatus* tree topology (Table 2.2): *lipid transport*, containing alipophorins and phospholipid-transporting ATPases; *scavenger receptor activity*, used to describe proteins that bind lipoproteins or anionic ligands (e.g apoptotic cells, bacteria, glycosylated products) and deliver the product to the cell via endocytosis (Peiser et al. 2002); *calcium ion binding*, a very broad suite of proteins that interact selectively with calcium ions, and ATPase activity, containing phospholipid-transporting ATPases recovered under *lipid transport* above combined with copper and sodium ion transporters.

Two enriched terms were retained from the gene set that maintained a (*Cx. p. pipiens* form molestus + *Cx. quinquefasciatus*) tree topology (Table 2.3): *structural constituent of ribosome*, containing ribosomal proteins, and *aspartic-type endopeptidase activity*, a family of enzymes that hydrolyze peptide bonds via activated water molecules bound to Aspartic acid residues. The enrichment of ribosomal proteins in this dataset is curious, and may indicate that the ribosomal proteins of *Cx. p. pallens* have diverged from those of *Cx. pipiens* and *Cx. quinquefasciatus*.

These results show that despite the evolutionary time-since-hybridization, we are able to ascribe selected components of the *Cx. p. pallens* transcriptome via phylogenetics to either of the *Cx. quinquefasciatus* or *Cx. pipiens* genomes. The genetic component assigned to *Cx. pipiens* contains members that function in energy metabolism, cell cycle / signaling, and redox reactions; the genes assigned to *Cx. quinquefasciatus* are enriched in lipid transport function and extracellular scavenging / innate immunity.



*Common foci of peptide evolution among Culex complex members*

Ka values were calculated from two-taxon (pairwise) alignments created from the 4-taxon alignments used in the phylogenetic tests (above) for each of the three groups defined previously (*Cx. p. pipiens* form *molestus* + *Cx. p. pallens*, *Cx. p. pallens* + *Cx. quinquefasciatus* and *Cx. p. pipiens* form *molestus* + *Cx. quinquefasciatus*). The 95<sup>th</sup> percentile (n=434) of genes from each comparison (herein the “diverged set(s)”) as ranked by descending Ka were then compared and contrasted.

We derived a set of 83 genes (19.0%, Table 2.4, Fig. 2.1) common to the diverged sets of all three groups, and thus exhibiting rapid peptide divergence among all species in our analysis. Gene ontology enrichment tests using these genes as a test set and a single test significance cutoff of  $1.0 \times 10^{-2}$  returned 5 GO terms (Table 2.5). Interestingly, three of these terms were also identified in our earlier comparison of feral and domestic forms of *Cx. pipiens pipiens* (Chapter 1; Price and Fonseca [2015]): *serine-type endopeptidase activity, proteolysis* (which contained all genes comprising *serine-type endopeptidase activity*) and *odorant binding*. The single gene that comprises the *biotin-[acetyl-CoA-carboxylase] ligase activity* ontology was also returned in the Ch. 1 analysis (Price and Fonseca 2015), although it fell below the stringent significance cutoff used. This enzyme (BirA) catalyzes the first step of lipid biosynthesis by biotinylating Acetyl Coenzyme A Carboxylase, and is ubiquitous among eukaryotes; most encode only a single biotin protein ligase (Chapman-Smith and Cronan 1999). BirA may bind many other putative substrates (i.e pyruvate carboxylase, propionyl CoA carboxylase, etc), and co-evolution of the enzyme with these substrates of each organism may lead to accelerated evolution of the enzyme (our result shows it has maintained significant divergence among all

mosquito comparisons thus far). The fifth GO term was a *cysteine-type endopeptidase* ontology that contained two caspases (included in *proteolysis*) that play integral roles in programmed cell death (apoptosis).

### *Differential targets of peptide evolution*

We next compared the diverged sets from each pairwise comparison to the other two, and isolated the genes unique to each (summarized in Fig. 2.1). Surprisingly, the *Cx. p. pallens* – *Cx. quinquefasciatus* diverged set contained 225 unique genes of the 434 total, while the *Cx. p. pipiens* form *molestus* – *Cx. p. pallens* set contained 84 unique genes and the *Cx. p. pipiens* form *molestus* – *Cx. quinquefasciatus* set contained 48. These results indicate that the latter two comparisons share a substantial number of “most diverged” genes; indeed, they contain 305 (70.2%) common members (Fig. 2.1), with 222 being specific to the two datasets (and not present in the *pallens* – *quinquefasciatus* comparison). As the taxon common to both sets is *Cx. pipiens* form *molestus*, it is likely that our analysis is highlighting divergence between this mosquito and both *Cx. p. pallens* and *Cx. quinquefasciatus*, and that this divergence occurred at common loci.

Enrichment tests performed using the above set of 225 and 222 genes exhibiting branch-specific evolutionary rates (Tables 2.6, 2.7) show that two GO terms are enriched in both sets (*serine-type endopeptidase/proteolysis* and *chitin binding*) and thus that common ontologies, yet different genes, are targeted by evolution. These ontologies were also enriched in the set of rapidly evolving genes that describe divergence between feral *Cx. p. pipiens* form *pipiens* and “domestic” form *molestus* (Chapter 1; Price and Fonseca [2015]). Given these results, it is likely that they are central to molecular

evolution in the *Culex pipiens* complex. We find four additional ontologies enriched in the set of 225 diverged genes unique to the *pallens* – *quinquefasciatus* comparison (and thus “unique to their divergence”, Tables 2.S3): *tissue regeneration*, *cyclin-dependent protein serine/threonine kinase inhibitor*, *NADP binding* and *N,N dimethylaniline monooxygenase*. All members of the *N,N dimethylaniline monooxygenase* ontology were also contained in *NADP binding*, thus we combined the discussion of both. All three genes annotated with tissue regeneration processes carried Ninjurin domains (PF04923). Ninjurins are small transmembrane proteins that are upregulated in response to cellular injury and/or infection (Araki and Milbrandt 1996), and have recently been shown to induce non-apoptotic cell death in *Drosophila* after septic injury (Broderick et al. 2012). Both genes with *cyclin-dependent protein serine/threonine kinase (CDK) inhibitor* ontologies currently lack annotations within the *Cx. quinquefasciatus* genome. CDK inhibitors block transitions of the cell cycle at the G<sub>1</sub> phase (and are thus often discussed as putative treatment for cancer (Kawamata et al. 1995, Guha 2012) and malaria (Brinen and Stout 2003)). When considered together with the cellular-death responsive Ninjurins, a focus on the cell cycle for these gene sets becomes a possibility. The *NADP binding* and *N,N dimethylaniline monooxygenase* ontologies share two genes, both with Flavin-binding monooxygenase (FMO)-like domains (PF00743). Insect FMO genes, not unlike P450 monooxygenases, act to convert toxic and xenobiotic chemicals to excretable metabolites.

Four additional ontologies (aside from *serine-type endopeptidase/proteolysis* and *chitin binding*, discussed above) were found to be enriched in the 222 genes unique to the variable set of both *molestus* – *pallens* and *molestus* – *quinquefasciatus* (Tables 2.S4):

*tRNA-intron endonuclease activity, homophilic cell adhesion, chitinase activity and nucleic acid phosphodiester bond hydrolysis.* *tRNA-intron endonucleases* are responsible for intron removal prior to maturation of tRNA molecules. *Chitinases* (discussed in Chapter 1) function to degrade structures such as the midgut peritrophic membrane and may thus co-evolve with the chitin-binding peritrophins. The *nucleic acid phosphodiester bond hydrolysis* ontology contained both genes from the *tRNA-intron endonuclease ontology*, with the addition of an exonuclease. All three members of the homophilic cell adhesion ontology lacked annotation in the *Cx. quinquefasciatus* genome, however all were found to contain multiple cadherin\_repeat domains via NCBI conserved domain database search (Marchler-Bauer et al. 2005) search. The cadherin\_repeat domain is an extracellular calcium ion-binding domain found in functional cadherin proteins that facilitates cell-cell adhesion (Hulpiau and van Roy 2009); it is thus likely that these peptides are cadherins.

#### *Comparative evolutionary rates*

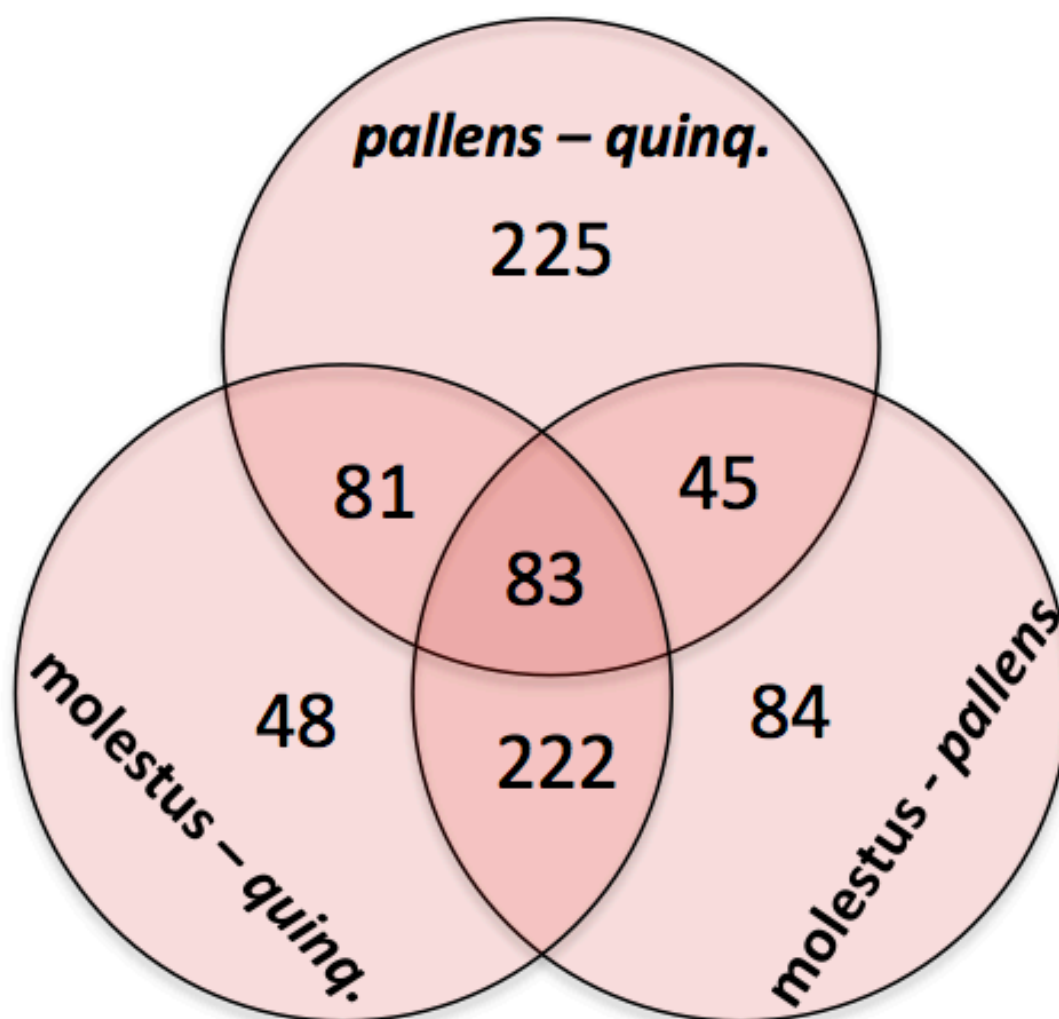
To examine not only which gene ontologies were evolving at differential rates but also how those rates compared between members of the pipiens complex, we next tested the variance of our  $K_a$  calculations between groups. A Kruskal-Wallis one-way analysis of variance conducted in R shows the mean  $K_a$  of the *Cx. pipiens pallens* – *Cx. quinquefasciatus* diverged (434 gene) set to be significantly lower than those of either remaining set (chi-squared = 537.1547, df = 2,  $p < 2.2e-16$ ; Fig. 2.2, Table 2.8). We have shown above that these two taxa share a greater discernible proportion of phylogenetic affinity at the transcript level, and our results here indicate that the rate of non-

synonymous change (even among the “fastest” evolving protein set) is less than that of either mosquito as tested with *Cx. p. pipiens* form *molestus*. These results both suggest that the “degree of relatedness” between *pallens* and *quinquefasciatus* is comparably high, as would be expected given the hybridization hypothesis regarding its origin.

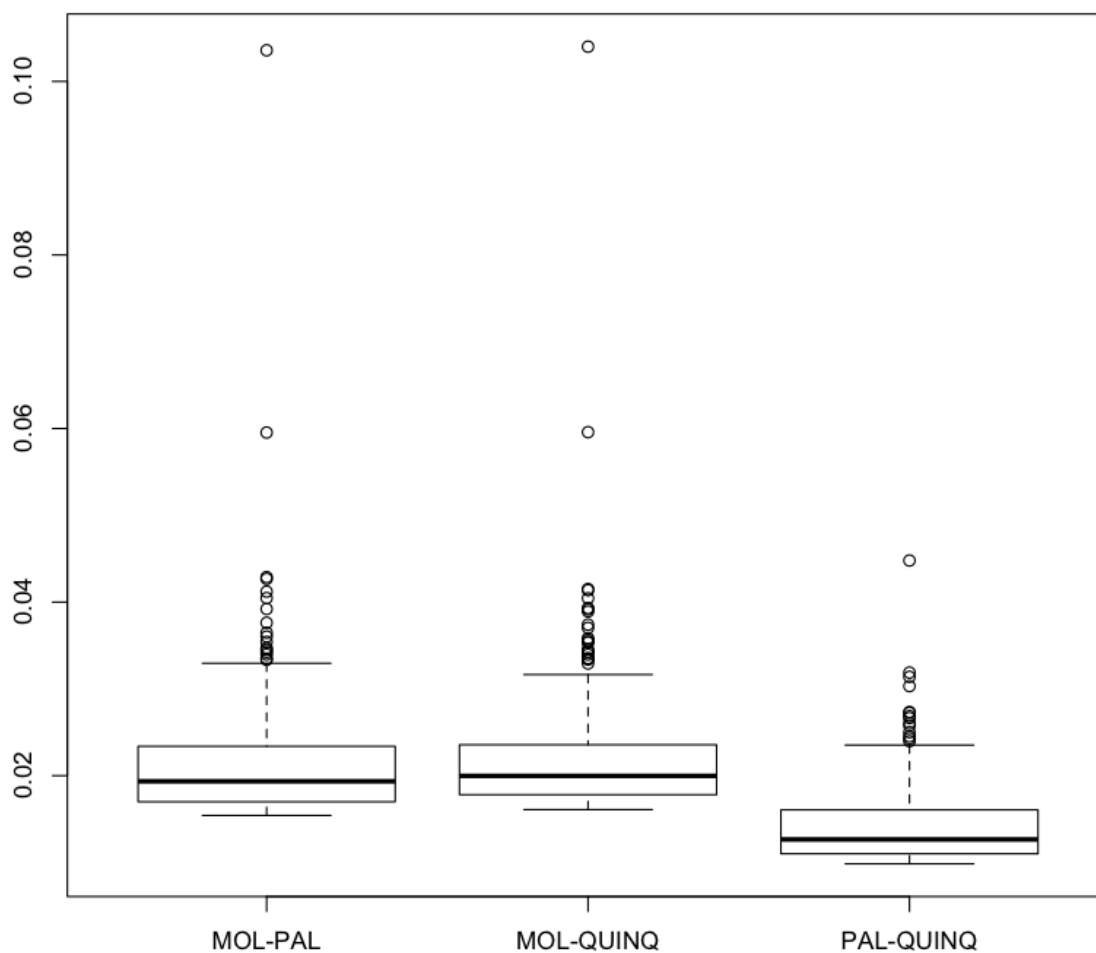
### *Conclusions*

Our results here represent a first attempt at qualifying and quantifying the genetic divergence between three Asian members of the *Culex pipiens* mosquito complex. We illustrate the mosaic nature of the *Cx. pipiens pallens* genome by assigning 1,052 and 323 genes to sister group relationships with the *Cx. quinquefasciatus* or *Cx. pipiens* genomes, respectively, which are hypothesized to have contributed to the hybridization that resulted in present day *pallens*. Further, we show the targets (i.e genes and gene ontologies) of peptide evolution within these three taxa to contain both shared and differential members; the diverged sets (made of fastest evolving proteins) from all pairwise comparisons shared 19% of their constituent genes, and these genes were enriched in digestive enzymes, odorant binding proteins, and curiously, a ubiquitous biotinylating enzyme. As these genes and/or ontologies have been suggested as also contributing divergence between *Cx. pipiens* forms *pipiens* and *molestus* (Price and Fonseca 2015), they appear to constitute a fulcrum of molecular divergence throughout the complex. By examining the differential occurrence of specific GO terms and their gene members across the three pairwise comparisons we find *Cx. p. pallens* diverges from *Cx. quinquefasciatus* at loci related to cellular regeneration and death (Ninjursins, CDK inhibitors) as well as detoxification (flavin-binding monooxygenases), while both *Cx. p. pallens* and *Cx.*

*quinquefasciatus* diverge from *Cx. p. pipiens* form molestus at tRNA intron endonucleases, chitinases, and a triplet of cadherin-like genes associated with homophilic cell adhesion. Finally, we show that the quantified protein divergence within the diverged sets as calculated by Ka values is significantly lower between *Cx. pipiens pallens* and *Cx. quinquefasciatus* than between either of these mosquitoes and *Cx. pipiens form molestus*, thus reinforcing the results of the phylogenetic tests and suggesting that *pallens* and *quinquefasciatus* share a greater degree of genomic similarity. Further sequencing of multiple populations of *Cx. pipiens* complex members from multiple populations will be required to create a “representative” set of candidates for evolutionary divergence and to further link genotype and phenotype in a manner that allows us to pinpoint the precise set of genes and thus functions that impart fitness advantage within a vector mosquito hybridization.



**Fig. 2.1.** Venn diagram illustrating shared members of the 432-gene diverged sets (upper 95<sup>th</sup> percentile of dataset as ranked by descending Ka) from the three pairwise comparisons.



**Fig. 2.2.** Kruskal-Wallis one-way analysis of variance illustrating relationships among the Ka values obtained for each pairwise species comparison (chi-squared = 537.1547, df = 2,  $p < 2.2e-16$ )



**Table 2.S1** Genes for which phylogenetic trees were constructed, with AU and SH test results. Filtered for alignments > 200nt or > 50% recovery of the *Cx. quinquefasciatus* homolog by length.

Table 2.S1 is located in the online supplementary material.

**Table 2.S2** Gene set, with annotations and Ka/Ks, values for which significant topology was reported by CONSEL.

Table 2.S2 is located in the online supplementary material.

**Table 2.1.** GO terms enriched in the gene set with (*Cx. p. pipiens* form molestus + *Cx. p. pallens*) tree topology

GO Term	Name	Type	FDR	single test p-Value	# in test group	# in reference	# non annot test	# non annot reference group	Over/Under
GO:0008601	protein phosphatase type 2A regulator activity	F	9.40E-01	1.20E-03	2	0	213	5939	over
GO:0004021	L-alanine:2-oxoglutarate aminotransferase activity	F	9.40E-01	1.20E-03	2	0	213	5939	over
GO:0008430	selenium binding	F	1.00E+00	3.60E-03	2	1	213	5938	over
GO:0017176	phosphatidylinositol N-acetylglucosaminyltransferase activity	F	1.00E+00	7.00E-03	2	2	213	5937	over
GO:0008967	phosphoglycolate phosphatase activity	F	1.00E+00	7.00E-03	2	2	213	5937	over
GO:0050661	NADP binding	F	1.00E+00	7.30E-03	3	9	212	5930	over
GO:0000079	regulation of cyclin-dependent protein serine/threonine kinase activity	P	1.00E+00	9.30E-03	3	10	212	5929	over

**Table 2.2.** GO terms enriched in the gene set with (*Cx. p. pallens* + *Cx. quinquefasciatus*) tree topology

GO Term	Name	Type	FDR	single test p-Value	# in test group	# in reference	# non annot test	# non annot reference group	Over/Under
GO:0006869	lipid transport	P	4.10E-01	2.10E-04	9	11	717	5417	over
GO:0005319	lipid transporter activity	F	6.50E-01	5.00E-04	7	7	719	5421	over
GO:0005044	scavenger receptor activity	F	1.00E+00	1.90E-03	5	4	721	5424	over
GO:0005509	calcium ion binding	F	1.00E+00	2.30E-03	28	109	698	5319	over
GO:0015662	ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	F	1.00E+00	7.40E-03	6	10	720	5418	over

**Table 2.3.** GO terms enriched in the gene set with (*Cx. p. pipiens* form molestus + *Cx. quinquefasciatus*) tree topology

GO Term	Name	Type	FDR	single test p-Value	# in test group	# in reference	# non annot test	# non annot reference	Over/Under
GO:0003735	structural constituent of ribosome	F	3.10E-01	1.60E-04	8	93	88	5965	over
GO:0004190	aspartic-type endopeptidase activity	F	7.70E-01	2.30E-03	2	3	94	6055	over

**Table 2.4.** Set of 83 genes common to diverged (top 95<sup>th</sup> percentile) sets of all three pairwise *Culex* species Ka calculations.

Cx. quinq homolog	Annotation	# GO terms	GO terms	Enzyme codes	Pfam ID	Pfam annotation
CPIJ016737-RA	scavenger receptor cysteine-rich protein	0	-	-	PF01826	Trypsin inhibitor like cysteine rich domain
CPIJ017838-RA	ketohekoxinase-like isoform 2	2	F:kinase activity; P:phosphorylation	-	PF00294	pkB family carboxylate kinase
CPIJ000315-RA	conserved hypothetical protein	0	-	-	PF13855	Leucine rich repeat
CPIJ017156-RA	ankyrin repeat and socs box protein 13	0	-	-	PF13637	Ankyrin repeats (many copies)
CPIJ004243-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ016965-RA	odorant-binding protein 56e	1	F:odorant binding	-	PF01395	PBP/GOBP family
CPIJ004946-RA	leucine-rich repeat-containing protein 15	7	F:kinase activity; F:ATP binding; F:protein kinase activity; P:phosphorylation; P:protein phosphorylation; F:protein tyrosine kinase activity; F:transferase activity; transferring phosphorus-containing groups	-	PF13855	Leucine rich repeat
CPIJ007828-RA	bile salt-activated lipase	1	F:hydrolase activity	-	PF00135	Carboxylesterase family
CPIJ011009-RA	kda secreted salivary protein	0	-	-	PF06477	Protein of unknown function (DUF1091)
CPIJ020192-RA	trypsin-like salivary secreted protein	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ008867-RA	conserved hypothetical protein	1	F:odorant binding	-	PF00962	Adenosine/AMP deaminase
CPIJ014249-RA	adenosine deaminase	2	F:deaminase activity; P:purine ribonucleoside monophosphate biosynthetic process	-	PF13855	Leucine rich repeat
CPIJ008561-RA	carboxypeptidase n subunit 2	1	F:carboxypeptidase activity	-	PF05444	Protein of unknown function (DUF753)
CPIJ002358-RA	conserved hypothetical protein	2	F:zinc ion binding; F:nucleic acid binding	-	-	-
CPIJ008249-RA	conserved hypothetical protein	0	-	-	PF01607	Chitin binding Peritrophin-A domain
CPIJ014194-RA	conserved hypothetical protein	3	C:extracellular region; F:chitin binding; P:chitin metabolic process	-	PF13465	Zinc-finger double domain
CPIJ010202-RA	zinc finger protein 436	3	F:zinc ion binding; F:nucleic acid binding; C:intracellular	-	PF07776	Zinc-finger associated domain (zf-AD)
CPIJ005072-RA	conserved hypothetical protein	3	F:nucleic acid binding; C:nucleus; F:zinc ion binding	-	PF00096	Zinc finger, C2H2 type
CPIJ011972-RA	zinc finger protein 774	3	F:nucleic acid binding; C:intracellular; F:zinc ion binding	-	PF13857	Ankyrin repeats (many copies)
CPIJ011920-RA	conserved hypothetical protein	3	Cubiquitin ligase complex; P:protein ubiquitination; F:ubiquitin-protein ligase activity	-	-	-
CPIJ003983-RA	conserved hypothetical protein	0	-	-	PF13855	Leucine rich repeat
CPIJ002478-RA	leucine rich protein	0	-	-	PF00089	Trypsin
CPIJ006014-RA	serine protease	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF13855	Leucine rich repeat
CPIJ000317-RA	conserved hypothetical protein	0	-	-	PF05444	Protein of unknown function (DUF753)
CPIJ005902-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ00266-RA	cytochrome p450 4c1	6	F:heme binding; F:oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen; F:monooxygenase activity; F:iron ion binding; F:electron carrier activity; P:oxidation-reduction process	-	PF00067	Cytochrome P450
CPIJ014551-RA	long form d7 salivary protein	2	F:odorant binding	-	PF01395	PBP/GOBP family
CPIJ01576-RA	biotin protein ligase	2	P:cellular protein modification process; F:biotin-acyl-CoA-carboxylase] ligase activity	EC:6.3.4.15	PF03099	Biotin/ligase A/B protein ligase family
CPIJ016503-RA	transcription factor grauzone	3	F:nucleic acid binding; C:intracellular; F:zinc ion binding	-	PF13465	Zinc-finger double domain
CPIJ012720-RA	odorant-binding protein	1	F:odorant binding	-	PF01395	PBP/GOBP family
CPIJ015252-RA	zinc carboxypeptidase	3	P:proteolysis; F:metallocarboxypeptidase activity; F:zinc ion binding	EC:3.4.17.0	PF02244	Carboxypeptidase activation peptide
CPIJ008109-RA	conserved hypothetical protein	1	F:transferase activity, transferring acyl groups other than amino-acyl groups	EC:3.1.1.0	PF01757	Acyltransferase family
CPIJ012580-RA	caspase-3 precursor	3	P:apoptotic process; P:proteolysis; F:cysteine-type endopeptidase activity	EC:3.4.22.0	PF00656	Caspase domain
CPIJ008637-RA	salivary mucin with chitin-binding domain	3	C:extracellular region; F:chitin binding; P:chitin metabolic process	-	PF01607	Chitin binding Peritrophin-A domain
CPIJ012010-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ018035-RA	trypsin li-p29	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ01860-RA	bcdin3 domain containing	2	F:methyltransferase activity; P:methylation	EC:2.1.1.0	PF13847	Methyltransferase domain
CPIJ00089-RA	conserved hypothetical protein	1	F:DNA binding	-	PF03184	DDE superfamily endonuclease
CPIJ015679-RA	conserved hypothetical protein	0	-	-	PF13855	Leucine rich repeat
CPIJ014651-RA	serine protease	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ006293-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ008620-RA	ionotropic glutamate receptor	4	F:extracellular-glutamate-gated ion channel activity; F:ionotropic glutamate receptor activity; P:ionotropic glutamate receptor signaling pathway; C:membrane	-	PF00060	Ligand-gated ion channel
CPIJ003579-RA	conserved hypothetical protein	0	-	-	-	-

Table 2.4 Continued

Cx. quinq homolog	Annotation	# GO terms	GO terms	Enzyme codes	Pfam ID	Pfam annotation
CPIJ016875-RA	conserved hypothetical protein	2	F:zinc ion binding; C:nucleus	-	PF1901	Protein of unknown function (DUF3421)
CPIJ006640-RA	conserved hypothetical protein	2	F:DNA binding; P:regulation of transcription, DNA-dependent	-	PF03615	GCM motif protein
CPIJ002300-RA	hypothetical protein CpijL_CPIJ002300	0	-	-	-	-
CPIJ002250-RA	conserved hypothetical protein	1	F:transferase activity, transferring phosphorus-containing groups	-	PF02958	Ecdysteroid kinase
CPIJ007028-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ003335-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ011842-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ007699-RA	conserved hypothetical protein	2	F:zinc ion binding; C:nucleus	-	PF04500	FLYWCH zinc finger domain
CPIJ005046-RA	conserved hypothetical protein	3	F:nucleic acid binding; C:nucleus; F:zinc ion binding	-	PF07776	Zinc-finger associated domain (zf-AD)
CPIJ009381-RA	transcriptional protein swt1	0	-	-	PF00397	WW domain
CPIJ012958-RA	conserved hypothetical protein	1	F:nucleic acid binding	-	PF05485	THAP domain
CPIJ012893-RA	breast cancer type 2 susceptibility	4	P:regulation of S phase of mitotic cell cycle; P:double-strand break repair via homologous recombination; F:single-stranded DNA binding; C:nucleus	-	PF09103	BRCA2, oligonucleotide/oligosaccharide-binding, domain 1
CPIJ017458-RA	4-coumarate- ligase 1	2	F:ligase activity; P:metabolic process	-	PF13193	Domain of unknown function (DUF4009)
CPIJ003921-RA	nacht and ankyrin domain protein	2	P:nucleoside metabolic process; F:catalytic activity	-	PF00023	Ankyrin repeat
CPIJ005969-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ002509-RA	meckel type 1	0	-	-	-	-
CPIJ014995-RA	initiator caspase	2	P:proteolysis; F:cysteine-type endopeptidase activity	EC:3.4.22.0	-	-
CPIJ012017-RA	serine protease	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ008062-RA	conserved hypothetical protein	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	-	-
CPIJ008578-RA	conserved hypothetical protein	4	C:cytoplasm; F:arginine-RNA ligase activity; P:arginyl-RNA aminoacylation; F:ATP binding	EC:6.1.1.19	-	-
CPIJ017797-RA	neurohypophysial hormones	2	P:proteolysis; F:serine-type endopeptidase activity	EC:3.4.21.0	PF00089	Trypsin
CPIJ017797-RA	neurohypophysial hormones	2	F:ATPase activity; F:ATP binding; C:integral to membrane; F:nucleotide binding; C:membrane;	-	-	-
CPIJ01801609-RA	conserved hypothetical protein	7	F:nucleoside-triphosphatase activity; P:ATP catabolic process	-	-	-
CPIJ006010-RA	FRA10AC1	0	-	-	PF09725	Folate-sensitive fragile site protein
CPIJ018955-RA	conserved hypothetical protein	0	-	-	-	-
CPIJ01814-RA	conserved hypothetical protein	5	F:structural constituent of ribosome; F:5S rRNA binding; P:translation; C:ribosome; C:intracellular	-	-	-
CPIJ010365-RA	fast myosin heavy chain hcii	0	-	-	-	-
CPIJ003556-RA	coiled-coil domain-containing protein 134	0	-	-	-	-
CPIJ002910-RA	retinoid-inducible serine carboxypeptidase	2	F:serine-type carboxypeptidase activity; P:proteolysis	EC:3.4.16.0	PF00450	Serine carboxypeptidase
CPIJ002022-RA	sperm flagellar protein 1	1	C:flagellum	-	PF06294	Domain of Unknown Function (DUF1042)
CPIJ007966-RA	conserved hypothetical protein	1	F:actin binding	-	PF02205	WH2 motif
CPIJ008746-RA	zinc finger protein	3	F:nucleic acid binding; C:intracellular; F:zinc ion binding	-	PF13894	C2H2-type zinc finger
CPIJ016621-RA	conserved hypothetical protein	1	F:nucleic acid binding	-	PF05485	THAP domain
CPIJ016278-RA	low quality protein: radial spoke head protein 4 homolog a-like	0	-	-	-	-
CPIJ011086-RA	conserved hypothetical protein	2	F:zinc ion binding; C:nucleus	-	PF04712	Radial spokehead-like protein
CPIJ015295-RA	juvenile hormone-inducible protein	1	F:transferase activity, transferring phosphorus-containing groups	-	PF13894	C2H2-type zinc finger
CPIJ007827-RA	carboxylesterase 2	1	F:hydrolase activity	-	PF02958	Ecdysteroid kinase
CPIJ014523-RA	coagulation factor ix: partial	2	F:serine-type endopeptidase activity; P:proteolysis	-	PF00135	Carboxylesterase family
CPIJ014523-RA	coagulation factor ix: partial	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ019592-RA	solute carrier family 2	3	F:substrate-specific transmembrane transporter activity; C:integral to membrane;	-	-	-
CPIJ017513-RA	conserved hypothetical protein	0	-	-	PF00083	Sugar (and other) transporter
CPIJ009618-RA	conserved hypothetical protein	0	-	-	PF07679	Immunoglobulin I-set domain
CPIJ009618-RA	conserved hypothetical protein	0	-	-	PF11779	Protein of unknown function (DUF3317)

**Table 2.5** GO terms enriched in the set of 83 genes common to diverged (top 95<sup>th</sup> percentile) sets of all three pairwise *Culex* species

Ka calculations.

GO Term	Name	Type	FDR	single test p-Value	# in test group	# in reference group	# non annot test	# non annot reference group	Over/Under
GO:0004252	serine-type endopeptidase activity	F	3.90E-02	6.10E-05	8	166	40	5773	over
GO:0006508	proteolysis	P	4.70E-02	8.50E-05	12	409	36	5530	over
GO:0005549	odorant binding	F	1.00E+00	2.90E-03	3	33	45	5906	over
GO:0004077	biotin-[acetyl-CoA-carboxylase] ligase activity	F	1.00E+00	8.00E-03	1	0	47	5939	over
GO:0004197	cysteine-type endopeptidase activity	F	1.00E+00	9.90E-03	2	17	46	5922	over

**Table 2.6** GO terms enriched in the set of 225 genes unique to the to diverged (top 95<sup>th</sup> percentile) set from the *Cx. pipiens pallens* –

*Cx. quinquefasciatus* pairwise Ka calculation.

GO Term	Name	Type	FDR	single test p-Value	# in test group	# in reference group	# non annot test	# non annot reference group	Over/Under
GO:0042246	tissue regeneration	P	1.20E-01	1.50E-04	3	2	148	5834	over
GO:0004861	cyclin-dependent protein serine/threonine kinase inhibitor activity	F	6.20E-01	1.90E-03	2	1	149	5835	over
GO:0005549	odorant binding	F	6.20E-01	1.90E-03	5	31	146	5805	over
GO:0050661	NADP binding	F	7.50E-01	2.90E-03	3	9	148	5827	over
GO:0004252	serine-type endopeptidase activity	F	9.80E-01	4.30E-03	11	163	140	5673	over
GO:0008061	chitin binding	F	1.00E+00	5.70E-03	6	59	145	5777	over
GO:0004499	N,N-dimethylaniline monooxygenase activity	F	1.00E+00	6.00E-03	2	3	149	5833	over

**Table 2.7** GO terms enriched in the set of 222 genes shared uniquely between the to diverged (top 95<sup>th</sup> percentile) sets of the *Cx. pipiens form molestus* – *Cx. quinquefasciatus* and *Cx. pipiens pallens* pairwise Ka calculation.

GO Term	Name	Type	FDR	single test p-Value	# in test group	# in reference group	# non annot test	# non annot reference group	Over/Under
GO:0000213	tRNA-intron endonuclease activity	F	5.50E-01	4.40E-04	2	0	124	5861	over
GO:0004252	serine-type endopeptidase activity	F	5.50E-01	1.00E-03	11	163	115	5698	over
GO:0007156	homophilic cell adhesion	P	8.90E-01	3.40E-03	3	12	123	5849	over
GO:0004568	chitinase activity	F	1.00E+00	5.90E-03	3	15	123	5846	over
GO:0006508	proteolysis	P	1.00E+00	6.70E-03	17	404	109	5457	over
GO:0090305	nucleic acid phosphodiester bond hydrolysis	P	1.00E+00	6.90E-03	3	16	123	5845	over
GO:0008061	chitin binding	F	1.00E+00	1.00E-02	5	60	121	5801	over



**Table 2.S3.** Annotations for genes enriched in the set of 225 genes unique to the to diverged (top 95<sup>th</sup> percentile) set from the Cx.

*piptiens pallens* – Cx. *quinquefasciatus* pairwise Ka calculation.

	# GO terms	GO terms	Enzyme Code	Top-scoring Pfam ID	Pfam description
<b>GO:0042246 Tissue regeneration</b>	3	Cintegral to membrane; P:cell adhesion; P:tissue regeneration	-	PF04923	Ninjurin
	3	Cintegral to membrane; P:cell adhesion; P:tissue regeneration	-	PF04923	Ninjurin
	3	Cintegral to membrane; P:cell adhesion; P:tissue regeneration	-	PF04923	Ninjurin
<b>GO:0004861 Cyclin-dependent protein serine/threonine kinase inhibitor</b>					
		F:cyclin-dependent protein kinase inhibitor activity; P:negative regulation of transcription from RNA polymerase II promoter; Cytoplasm; Nucleus;			
	5	F:snRNA binding	-	N/A	N/A
CPJ002766-RA		conserved hypothetical protein			
CPJ005739-RA		conserved hypothetical protein	-	PF02234	Cyclin-dependent kinase inhibitor
<b>GO:0005549 Odorant binding</b>					
	1	F:odorant binding	-	PF01395	PBP/GOBP family
	1	F:odorant binding	-	PF01395	PBP/GOBP family
	1	F:odorant binding	-	PF01395	PBP/GOBP family
	1	F:odorant binding	-	PF01395	PBP/GOBP family
	1	F:odorant binding	-	PF01395	PBP/GOBP family
<b>GO:0050661 NADP binding</b>					
	4	F:flavin adenine dinucleotide binding; F:N,N-dimethylamine monooxygenase activity; P:oxidation-reduction process; F:NADP binding	EC:1.14.13.8	PF00743	Flavin-binding monooxygenase-like
	4	F:flavin adenine dinucleotide binding; F:N,N-dimethylamine monooxygenase activity; P:oxidation-reduction process; F:NADP binding	EC:1.14.13.8	PF00743	Flavin-binding monooxygenase-like
	7	P:tetrahydrofolate biosynthetic process; P:oxidation-reduction process; P:nucleotide biosynthetic process; P:one-carbon metabolic process; P:glycine biosynthetic process; F:hydrofolate reductase activity; F:NADP binding	EC:1.5.1.3	PF00186	Dihydrofolate reductase
<b>GO:0004252 Serine-type endopeptidase activity</b>	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF12032	Regulatory CLIP domain of proteinases
	2	P:proteolysis; F:serine-type endopeptidase activity	EC:3.4.21.0	PF00089	Trypsin
	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF12032	Regulatory CLIP domain of proteinases
	4	C:membrane; F:scavenger receptor activity; P:proteolysis; F:serine-type endopeptidase activity	EC:3.4.21.0	PF00089	Trypsin
	4	F:kinase activity; P:phosphorylation; P:proteolysis; F:serine-type endopeptidase activity	EC:3.4.21.0	PF00089	Trypsin
	4	try1, anoga ame: full=trypsin-1			
	2	P:proteolysis; F:serine-type endopeptidase activity	EC:3.4.21.0	PF00089	Trypsin
	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF12032	Regulatory CLIP domain of proteinases
<b>GO:0008061 Chitin binding</b>					
	3	C:extracellular region; F:chitin binding; P:chitin metabolic process	-	PF01607	Chitin binding Peritrophin-A domain
	3	C:extracellular region; F:chitin binding; P:chitin metabolic process	-	PF01607	Chitin binding Peritrophin-A domain
	3	C:extracellular region; F:chitin binding; P:chitin metabolic process	-	PF01607	Chitin binding Peritrophin-A domain
	3	C:extracellular region; F:chitin binding; P:chitin metabolic process	-	PF01607	Chitin binding Peritrophin-A domain
	3	C:extracellular region; F:chitin binding; P:chitin metabolic process	-	PF01607	Chitin binding Peritrophin-A domain



**Table 2.S4.** Annotations for genes enriched in the set of 222 genes shared uniquely between the to diverged (top 95<sup>th</sup> percentile) sets of the *Cx. pipiens form molestus* – *Cx. quinquefasciatus* and *Cx. pipiens form molestus* – *Cx. pipiens pallens* pairwise Ka calculation.

GO:0004814 tRNA-intron endonuclease		# GO terms	GO terms	Enzyme Code	Top-scoring Pfam ID	Pfam description
CPIJ0801853-RA	trna-splicing endonuclease subunit sen2	4	P:nucleic acid phosphodiester bond hydrolysis; F:nucleic acid binding; P:trRNA splicing, via endonucleolytic cleavage and ligation; F:trRNA-intron endonuclease activity	EC:3.1.27.9	PF01974	tRNA intron endonuclease, catalytic C-terminal domain
CPIJ013589-RA	conserved hypothetical protein	5	P:nucleic acid phosphodiester bond hydrolysis; F:nucleic acid binding; P:trRNA-type intron splice site recognition and cleavage; C:trRNA-intron endonuclease complex; F:trRNA-intron endonuclease activity	EC:3.1.27.9	PF01974	tRNA intron endonuclease, catalytic C-terminal domain
GO:0004252 Serine-type endopeptidase / GO:0006508 Proteolysis						
CPIJ010072-RA	transmembrane protease	3	C:integral to membrane; P:proteolysis; F:serine-type endopeptidase activity			
CPIJ011382-RA	chymotrypsin 1	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ016220-RA	serine protease	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ006080-RA	trypsin theta	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Regulatory CLIP domain of proteinases
CPIJ013319-RA	zinc metalloproteinase nas-4	4	F:metal ion binding; F:metalloendopeptidase activity; P:proteolysis; F:zinc ion binding	EC:3.4.24.0	PF01400	Astatin (Peptidase family M12A)
CPIJ003717-RA	vitamin k-dependent protein c	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF12032	Regulatory CLIP domain of proteinases
CPIJ002489-RA	metalloproteinase	4	F:metal ion binding; F:metalloendopeptidase activity; P:proteolysis; F:zinc ion binding	EC:3.4.24.0	PF13855	Leucine rich repeat
CPIJ014110-RA	aael012157- partial	2	F:peptide activity; P:proteolysis	-	PF04389	Peptidase family M28
CPIJ012036-RA	aminopeptidase n	4	F:metalloendopeptidase activity; P:proteolysis; F:zinc ion binding; F:aminopeptidase activity	EC:3.4.24.0; EC:3.4.11.0	PF11838	Domain of unknown function (DUF3358)
CPIJ006946-RA	proteasome subunit alpha type 2	5	C:cytoplasm; C:proteasome core complex, alpha-subunit complex; P:ubiquitin-dependent protein catabolic process; C:nucleus; F:threonine-type endopeptidase activity	EC:3.4.25.0	PF00227	Proteasome subunit
CPIJ007871-RA	vitamin k-dependent protein c	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF13895	Immunoglobulin domain
CPIJ013396-RA	urokinase-type plasminogen activator	4	F:kinase activity; P:phosphorylation; P:proteolysis; F:serine-type endopeptidase activity	EC:3.4.21.0	PF00089	Trypsin
CPIJ004089-RA	conserved hypothetical protein	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ003622-RA	serine protease	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ013430-RA	serine protease	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ012713-RA	serine protease	2	F:serine-type endopeptidase activity; P:proteolysis	EC:3.4.21.0	PF00089	Trypsin
CPIJ006815-RA	serine protease inhibitor	3	F:serine-type endopeptidase inhibitor activity; P:proteolysis; F:peptidase activity	-	PF00079	Serpin (serine protease inhibitor)
GO:0007156 homophilic cell adhesion						
CPIJ005630-RA	conserved hypothetical protein	4	C:integral to membrane; P:homophilic cell adhesion; F:calcium ion binding; C:plasma membrane	-	N/A	N/A
CPIJ014101-RA	aael007478- partial	4	C:integral to membrane; P:homophilic cell adhesion; F:calcium ion binding; C:plasma membrane	-	N/A	N/A
CPIJ005629-RA	conserved hypothetical protein	4	C:integral to membrane; P:homophilic cell adhesion; F:calcium ion binding; C:plasma membrane	-	N/A	N/A

Table 2.S4 Continued

GO:0004568 chitinase activity						
CPIJ000009-RA	chitinase a1		6	F:chitinase activity; P:carbohydrate metabolic process; P:chitin catabolic process; F:cation binding; F:chitin binding; C:extracellular region	PF00704	Glycosyl hydrolases family 18
CPIJ013085-RA	sarcalumenin		7	F:cation binding; P:chitin catabolic process; P:carbohydrate metabolic process; P:GTP catabolic process; F:GTPase activity; F:chitinase activity; F:GTP binding	PF00350	Dynamin family
CPIJ800112-RA	endochitinase a		3	F:chitinase activity; P:chitin catabolic process; P:cell wall macromolecule catabolic process	PF00182	Chitinase class I
GO:0090305 nucleic acid phosphodiester bond hydrolysis						
CPIJ013589-RA	conserved hypothetical protein		5	P:nucleic acid phosphodiester bond hydrolysis; F:nucleic acid binding; P:tRNA-type intron splice site recognition and cleavage; C:tRNA-intron endonuclease complex; F:tRNA-intron endonuclease activity	PF01974	tRNA intron endonuclease, catalytic C-terminal domain
CPIJ011181-RA	anopheles gambiae pest agap012804-pa		4	P:nucleic acid phosphodiester bond hydrolysis; F:nucleic acid binding; C:cytoplasm; F:exonuclease activity	PF00929	Exonuclease
CPIJ801853-RA	trna-splicing endonuclease subunit sen2		4	P:nucleic acid phosphodiester bond hydrolysis; F:nucleic acid binding; P:tRNA splicing, via endonucleolytic cleavage and ligation; F:tRNA-intron endonuclease activity	PF01974	tRNA intron endonuclease, catalytic C-terminal domain
GO:0008061 chitin binding						
CPIJ000009-RA	chitinase a1		6	F:chitinase activity; P:carbohydrate metabolic process; P:chitin catabolic process	PF00704	Glycosyl hydrolases family 18
CPIJ009078-RA	conserved hypothetical protein		3	C:extracellular region; F:chitin binding; P:chitin metabolic process	PF01607	Chitin binding Peritrophin-A domain
CPIJ008500-RA	conserved hypothetical protein		3	C:extracellular region; F:chitin binding; P:chitin metabolic process	-	
CPIJ003202-RA	af373883_1mucin-like protein 2		3	C:extracellular region; F:chitin binding; P:chitin metabolic process	PF01607	Chitin binding Peritrophin-A domain
CPIJ004734-RA	conserved hypothetical protein		3	C:extracellular region; F:chitin binding; P:chitin metabolic process	PF01607	Chitin binding Peritrophin-A domain

**Table 2.8** Mutiple comparison test after Kruskal-Wallis one-way analysis of variance

Multiple comparison test after Kruskal-Wallis			
<i>p</i> -value: 0.05			
<b>Comparison</b>	<b>obs.dif</b>	<b>critical.dif</b>	<b>difference</b>
MOL-PAL/MOL-QUINQ	59.27419	61.10513	FALSE
MOL-PAL/PAL-QUINQ	480.10023	61.10513	TRUE
MOL-QUINQ/PAL-QUINQ	539.37442	61.10513	TRUE

### Chapter 3:

#### Characterization of the *doublesex* gene within the *Culex pipiens* complex suggests regulatory plasticity at the base of the mosquito sex determination cascade

##### Abstract

The *doublesex* gene controls somatic sexual differentiation in *Drosophila melanogaster*, and may function thusly in other metazoans including the malaria mosquito *Anopheles gambiae* and the dengue and yellow fever vector *Aedes aegypti* (Diptera: Culicidae). As in other studied dipteran *dsx* homologs, the gene maintains functionality via evolutionarily conserved protein domains and sex-specific alternative splicing. The upstream factors that regulate splicing of *dsx* and the manner in which they do so however remain variable even among closely related organisms. As the induction of sex ratio biases is a central mode of action in many emerging molecular insecticides, it is imperative to elucidate as much of the sex determination pathway as possible in the mosquito disease vectors. Here we report the full-length gene sequence of the *doublesex* gene in *Culex quinquefasciatus* (*Cxqdsx*) and its male and female-specific isoforms. *Cxqdsx* maintains characteristics possibly derived in the Culicinae and present in the *Aedes aegypti dsx* gene (*Aeadsx*) such as gain of exon 3b and the presence of Rbp1 *cis*-regulatory binding sites, and also retains presumably ancestral attributes present in *Anopheles gambiae* such as maintenance of a singular female-specific exon 5. Unlike in *Aedes aegypti*, we find no evidence for intron gain in the female transcript(s), yet recover a second female isoform generated via selection of an alternate splice donor. Utilizing next-gen

sequence (NGS) data, we complete the *Aeadsx* gene model and identify a putative core promoter region in both *Aeadsx* and *Cxqdsx*. Also utilizing NGS data, we construct a full-length gene sequence for the *dsx* homolog of the northern house mosquito *Culex pipiens* form *pipiens* (*Cxpipdsx*). Analysis of peptide evolutionary rates between *Cxqdsx* and *Cxpipdsx* (both members of the *Culex pipiens* complex) shows the male-specific portion of the transcript to have evolved rapidly with respect to female-specific and common regions. As in other studied insects, *doublesex* maintains sex-specific splicing and conserved *doublesex*/mab-3 domains in the mosquito *Culex quinquefasciatus* and *Cx. pipiens*. The *cis*-regulated splicing of *Cxqdsx* does not appear to follow either currently described mosquito model (for *An. gambiae* and *Ae. aegypti*); each of the three mosquito genera exhibit evidence of unique *cis*-regulatory mechanisms. The male-specific *dsx* terminus exhibits rapid peptide evolutionary rates, even among closely related sibling species.

## Introduction

The manifestation of distinct sexes is fundamentally conserved among most metazoans. However, the development of sex-specific somatic and gonadal tissues and neuronal processes (e.g. behaviors) is governed by a variety of factors both environmental and genetic, and often varying widely between and within taxa (Bull and Vogt 1979, Marín and Baker 1998, Zarkower 2001). Most animals direct sex-specific cell fate by function of the *Doublesex*/Mab-3 Related Transcription factor (DMRT) family of zinc-finger proteins (Raymond et al. 1999, Kopp 2012) and the genes they regulate. Within the insects, this process involves a genetic cascade first

elucidated in the model fly *Drosophila melanogaster* (Baker and Wolfner 1988) whereby a primary signal triggers sex-specific splicing of one or more regulatory factors which subsequently bind pre-mRNA of the conserved DMRT “major switch” gene, *doublesex*, and direct its sex-specific splicing, thus initiating development of male or female forms (Sánchez 2008).

Although there are many diverse primary signals that initiate the cascade (e.g. X:A ratio, M-factors, W/Y chromosomes; see (Marín and Baker 1998)), *dsx* appears to be conserved as the major switch at the base of the cascade (Geuverink and Beukeboom 2014, Wexler et al. 2014). In many insects the male and female-specific splicing of *dsx* is directed by the upstream regulator *transformer*, a serine/arginine rich (SR) protein which itself is transcribed in a sex-specific manner, as well as the constitutively expressed *transformer-2* (Inoue et al. 1990, Verhulst et al. 2010). The resultant TRA/TRA2 peptide complex binds the *dsx* mRNA at the *dsx* repeat element (*dsxRE*), facilitated by the purine-rich enhancer (PRE) element (Tian and Maniatis 1993, Lynch and Maniatis 1995), and directs sex-specific splicing of *dsx* mRNA for translation into male (DSX<sup>M</sup>) or female (DSX<sup>F</sup>) peptides. In *Drosophila*, an additional SR splicing enhancer component, RBP1, binds to target sites in the splice acceptor preceding the female-specific exon and is essential for efficient splicing of female *dsx* pre-mRNA (Heinrichs and Baker 1995). The downstream targets of insect *dsx* are not well elucidated, however 58 optimal binding sites and associated nearest genes have been identified for *D. melanogaster Dmdsx* (Luo et al. 2011). The red flour beetle *Tribolium castaneum Tcdsx* has been implicated in oocyte development including Vitellogenins and their associated

receptors (Shukla and Palli 2012), while Lepidopteran *dsx* has been shown to influence expression of pheromone-binding proteins and *hexamerin* storage proteins (Suzuki et al. 2003).

Orthologs of the *dsx* gene have currently been identified in seven orders of insects ranging from the primitive *Pediculus humanus* (human body louse) to several genera of Hymenoptera, however a functional *transformer* homolog has not always been recovered in these genomes (see Geuverink and Beukeboom [2014] for summary) leading to speculation that some lineages have recruited alternate or additional upstream regulators for *dsx* (Salvemini et al. 2011). For example, TRA/TRA-2 mediated splicing of *dsx* has been shown in the Brachyceran flies *Ceratitis capitata* (Salvemini et al. 2009), *Musca domestica* (Burghardt et al. 2005) and *Lucilia cuprina* (Concha and Scott 2009) yet transformer appears lost in the Nematoceran flies including mosquitoes (Geuverink and Beukeboom 2014). Despite varying primary signals and upstream regulatory mechanisms, male and female-specific DSX peptides of various Diptera including *Anastrepha* (Alvarez et al. 2009), *Ceratitis* (Saccone et al. 2008) and *Musca* (Hediger et al. 2004) effected partial masculinization and feminization of genetically female and male *D. melanogaster*, respectively, when expressed ectopically. This evolutionary conservation is due in part to the retention of two functional protein domains essential for peptide oligomerization: an atypical zinc-finger DNA-binding domain found in multiple members of the DMRT superfamily (DBD/OD1) and an oligomerization domain (OD2) unique to *dsx* (An et al. 1996). The DBD/OD1 domain functions to form a dimeric DNA-binding unit that maintains 92% sequence

similarity between Dipteran (*D. melanogaster*) and Lepidopteran (*Bombyx mori*) taxa while completely conserving the critical cysteine and histidine residues (Ohbayashi et al. 2001). The OD2 domain is likely responsible for sex-specific splicing activation or repression of downstream factors (An et al. 1996), and is modified by sex-specific splicing to maintain both common and male/female-specific portions; the common portion exhibits a greater degree of conservation within and among insect taxa than the C-terminal sex-specific portion (Ohbayashi et al. 2001, Salvemini et al. 2011).

Orthologs of *dsx* have been recovered from the mosquitoes *Anopheles gambiae* (*Angdsx* [Scali et al. 200]) and *Aedes aegypti* (*Aeadsx* [Salvemini et al. 2011]). Both genes show sex-specific splicing and contain multiple copies of TRA/TRA2 *cis*-regulatory elements including *dsxREs* and purine-rich enhancers, however they differ in several evolutionary aspects. The *Angdsx* gene (Fig. 1) spans an 85kb region of chromosome 2R and is composed of seven exons, of which the first four code for 5' UTR and a common non-sex specific region of the protein. Exon 5 is female-specific (i.e. is spliced out of the male mRNA) and contains an in-frame stop codon terminating the female peptide. Exon 6 contains male-specific coding sequence with termination codon and 3' UTR, and exon 7 contains only 3'UTR. Exons 6 and 7 are present in transcripts of both sexes but are transcribed entirely as UTR in the female isoform. Female-specific splicing of *Angdsx* and the retention of exon 5 relies on activation of a 5' splice donor (Fig. 1; see Black [2003] for alternative splicing mechanism review) of the downstream intron 5 following binding of a TRA/TRA2 complex to *dsxREs* which facilitates recruitment of the



spliceosomal machinery. This is in contrast to *doublesex* genes of *D. melanogaster*, *Bactrocera tryoni*, *M. domestica*, *M. scalaris* and *C. capitata*, which splice the female-specific isoform after activation of a weak (due to the presence of purines in the polypyrimidine tract) 3' splice acceptor upstream of the female-specific exon to facilitate its inclusion. This is evidenced by the location of the dsxRE in *Angdsx* being at the 3' end of the female-specific exon as opposed to the 5' end as in *Dmdsx*.

The current *Aeadsx* gene model (Salvemini et al. 2011) (Fig. 1) spans 450kb of genomic DNA of supercontig 1.370 and is composed of eight known exons, although nine are likely. Unlike other sequenced Dipteran *dsx* genes, *Aeadsx* was found to produce two female-specific isoforms by exon skipping, encoding peptides with alternative C-termini via inclusion of both exons 5a and 5b, or 5b alone. Additionally, analysis of *cis*-acting elements in *Aeadsx* revealed a cluster of TRA-2-ISS and RBP1 elements upstream of exon 5a, and Dipteran dsxRE binding sites and PRE elements present only in exon 5b (Fig. 1). Several instances of a motif strongly resembling a potential dsxRE element previously only recovered in the Hymenoptera (*NvdsxRE*, [Verhulst et al. 2010]) were found within exon and intron 5a. Unlike *An. gambiae* (and similar to *Drosophila*) *Aeadsx* possesses a weak splice acceptor upstream of exon 5b that is activated to splice both female isoforms. Salvemini et al. (2011) hypothesize that regulatory mechanisms governing the sex-specific splicing of the gene in *Ae. aegypti* are different than in other Diptera including *An. gambiae*, and that the two female-specific exons were each under the control of a different splicing regulator: A female-specific TRA-like protein acts in females as a splicing activator of exon 5b via dsxRE and PRE elements, while a splice

repressor acts on 5a (included by default splicing) in some transcripts. In the males, a male-specific factor may act to repress inclusion of exon 5a via TRA-2-ISS and *NvdsxRE* elements, while exon 5b is excluded due to lack of female-specific TRA.

Cho et al. (2007) proposed that default female-specific *dsx* splicing by selective repression of the male isoform (i.e by the *feminizer* gene in *A. mellifera* [Gempe et al. 2009]) and the recently discovered piRNA precursor *Fem* in *B. mori* [Kiuchi et al. 2014]) is ancestral to holometabolous insects based on its conservation in taxa as phylogenetically distant as *A. mellifera* and *B. mori*, and that Diptera possess a derived splicing system where the male form is default and the female form must be ‘splice-activated’ by a TRA/TRA2-like factor. While this appears to be the case in *Anastrepha*, *Drosophila*, and *An. gambiae doublesex*, the data from Salvemini et al. (Salvemini et al. 2011) strongly suggest that the female spliceforms are default in *Ae. aegypti*; the “strong” exon 5a does not require TRA/TRA2 enhancement, and must be repressed by a male factor. Culicine mosquitoes (inclusive of the genera *Aedes* and *Culex*) determine sex at an autosomal locus (Newton et al. 1974), while Anopheline mosquitoes possess heteromorphic (XY) sex chromosomes (Gilchrist and Haldane 1947). The latter authors propose that this locus (the M-locus) may either act on intermediary factors or on the *dsx* gene itself (*transformer* appears to be either lost or extremely diverged in the mosquitoes [Geuverink and Beukeboom 2014], however *transformer2* is present) to suppress female-specific *dsx* splicing and generate the male form. Further, Salvemini et al (2011) posit that retention of the Hymenopteran-like *NvdsxRE* elements coupled with *Apis*-like splicing regulation (and a likely female-specific default

splicing) could represent a stably maintained ancestral state in *Ae. aegypti* exclusive of the rest of known Dipteran *doublesex*. Recently, analysis of the red flour beetle *Tribolium castaneum* (Shukla and Palli 2014) revealed three female-specific and one male-specific *dsx* isoform, with male default splicing occurring via suppression of maternally transferred zygotic TRA protein (required to activate female-specific splicing) by a dominant male factor. This variation in the top-level regulation of *dsx* among Hymenoptera, Diptera, Lepidoptera, and Coleoptera via upstream factors is in agreement with the theory of Wilkins (1995) stating that the cascade has evolved in reverse order, with the final double-switch gene (*doublesex*) remaining relatively conserved as additional elements are added and/or neofunctionalization occurs at the upper regulatory levels. As sex determination is critical to insect reproduction, deleterious mutations in *dsx* could therefore have strong effects on fitness and be selected against. Previous studies have shown the female-specific exon to be evolutionarily conserved (Lagos et al. 2005, Ruiz et al. 2007, Hughes 2011), yet disagree on evolutionary rate comparisons of the common and male-specific portions of the transcript over longer evolutionary time frames. Hughes (2011) found a much greater rate of non-synonymous substitutions within the male-specific region as compared to the common region, while Sobrinho Jr. and de Brito (2012) found nearly equivalent levels of positive selection between the two.

As the production of genetic sexing mosquito strains and molecular methods that create male bias and/or elimination of the female sex are ideal strategies for sterile insect technique (Gilles et al. 2014), it follows that a conserved sex regulator like *doublesex* (and *transformer*) would be optimal molecular targets for such

control programs (Dafa'alla et al. 2010). Elucidating the variable mechanisms by which *dsx* determines sexual fate in sequenced mosquito lineages is mandatory if progress is to be made towards a control strategy for the world's deadliest animals. Here we provide full-length gene sequence, sex-specific splicing analyses, and regulatory analysis of the *doublesex* gene from the southern house mosquito *Culex quinquefasciatus* (herein *Cxqdsx*) via RT-PCR and Illumina transcriptome data. Additionally, to discern the strength and location of early evolutionary drivers on *doublesex* within the *Culex pipiens* complex, we conduct an evolutionary analysis using *Cxqdsx* and a newly constructed *dsx* transcript from *Culex pipiens* form *pipiens* (*Cxpipdsx*). These results provide a comparative platform with which to study sex determination in those mosquitoes with currently sequenced genomes (*An. gambiae* [Holt et al. 2002], *Ae. aegypti* [Nene et al. 2007] and *Cx. quinquefasciatus* [Arensburger et al. 2010]).

## Methods

We used the conserved OD1 and OD2 peptide sequences of the *Aedes aegypti* *doublesex* gene (Salvemini et al. 2011) as a TBLASTN query to the *Cx. quinquefasciatus* genome assembly (Arensburger et al. 2010) and identified strong hits to both on genome supercontig 3.59. Further BLAST searches using the full peptide sequence of *Aeadsx* identified very weak local alignments to the supercontig representing putative female-specific (exon 5) and male specific/UTR (exon 6) coding sequence. A putative start codon in exon 2 was identified via homology with *Aeadsx*, and primers quinqOD12F, quinqOD12Rcom and quinqOD12Rfem (See Fig.

3.2 and Table 3.S1) were designed to amplify the putative 5' end of the common and female specific transcripts, respectively, and primers quinqDSX8F, quinqDSX7F, quinqDSX6R and quinqDSX7R were designed to amplify the 3' end of male and female specific transcripts.

*Culex quinquefasciatus* mosquitoes were obtained from a colony initiated in 2008 with egg rafts collected from Oahu, Hawaii, USA. Male and female total RNA was extracted separately from twenty adult mosquitoes of each sex using the Qiagen RNeasy Plus Universal Kit (Qiagen, Valencia CA) per manufacturer's protocol. Prior to extraction, samples were placed in a 2ml eppendorf tube containing a sterile steel bead + 800µl Qiazol solution and homogenized for 1 minute @ 20Hz on a Qiagen TissueLyser. Contaminant DNA was removed with the TURBO DNA-free DNA Removal Kit (Invitrogen, Carlsbad CA) and first-strand cDNA was generated using the Superscript First-Strand Synthesis System (Invitrogen) per manufacturer's protocol and diluted to 50µl in H<sub>2</sub>O. Four microliters of the cDNA was used in each 25µl PCR reaction containing 12µl H<sub>2</sub>O, 2.5µl Qiagen Q-solution, 2.5µl 10x PCR buffer, 0.5µl dNTPs, 2.5 units AmpliTaq DNA Polymerase (Invitrogen) and 0.5µl (200 µM final concentration) of each primer. Thermal cycling conditions were as follows: 1 minute @ 95°C, followed by 30 cycles x (30 seconds @ 94°C , 30 seconds @ 50-54°C primer-specific annealing, 60 seconds @ 68°C [120 seconds for products > 1kb]), 5 minutes @ 68°C final extension.

To recover the complete 5' end of the transcript, we performed 5' RACE PCR using the FirstChoice RLM-RACE Kit (Invitrogen) per manufacturer's protocol using internal gene-specific primers quinqDSX5RACE-GSP1 and quinqDSX5RACE-GSP2

placed adjacent to the OD1 domain. All RT-PCR and RACE-PCR amplicon products were visualized on a 1.5% agarose gel in TAE buffer and gel-purified using the QIAquick Gel Extraction Kit (Qiagen) prior to cloning via the TOPO TA Cloning Kit (Invitrogen) and PCR-enrichment using the M13 forward/reverse primer pair per manufacturer's protocol. PCR products were cleaned with ExoSap (Invitrogen) per manufacturer's protocol, and cycle sequencing was performed by GENEWIZ (South Plainfield, NJ) using the M13 primer pair.

The 3' end of *Cxqdsx* was predicted, and the entire gene sequence qualified by mapping the paired-end RNAseq data from NCBI SRA accession SRR991016 generated by Leal et al. (Leal et al. 2013) to *Cx. quinquefasciatus* supercontig 3.59 using the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark) large-gap read mapper (nucleotide similarity score of 95% over a 95% read length fraction) and manually examining the output. This process was repeated using the *Cx. pipiens* f. *pipiens* paired-end RNAseq library generated by Price and Fonseca (2014) and the *Cx. quinquefasciatus* reference generated above to create the full-length gene structure for *Cx. pipiens* f. *pipiens doublesex* (*Cxpipdsx*). To extend the gene model for *Aeadsx*, we repeated this protocol yet again with the *Ae. aegypti* NCBI short-read paired-end libraries SRR924024 and SRR789758 and AaegL1.4 supercontig 1.370.

To assess the distribution of the consensus dsxRE (TRA/TRA2) and RBP1 type-b motifs (derived from those of *D. melanogaster*, *An. gambiae* and *Ae. aegypti*), we screened all transcript coding (CDS) sequences corresponding with the *Cx. quinquefasciatus* Cpip1.3 dataset from VectorBase for their presence. The degenerate motif was broken down into all possible constituents, and each was

queried against the CDS dataset with BLASTn (e-val=999, word\_size=13 [dsxRE] or 7 [RBP1b]). The output was parsed via custom Perl scripts, and transcripts containing six copies of the motif in a 224bp (for the dsxRE; 546bp for RBP1b) window were retained. The AhoPro software utility (Boeva et al. 2007) was used to calculate the probability of observing the motif against a reference dataset of nucleotides randomly generated under a Bernoulli/0-order Markov model.

The synonymous substitutions per synonymous site and nonsynonymous substitutions per nonsynonymous site (Ks and Ka, respectively) and the Ka/Ks ratio were calculated in a pairwise comparison between *Cxqdsx* and *Cxpipdsx* using the KaKs Calculator v2.0 (Wang et al. 2010) under model averaging (MA). We recalculated these values for each sliding 30bp window while moving 3bp (1 amino acid) downstream at a time. To examine base composition of splice acceptor sites, we retrieved 52,278 internal (i.e. exclusive of exon 1) exons with 16nt of upstream sequence from the CpipJ 1.3 assembly (Vectorbase, (Megy et al. 2012)) and calculated the mean number of pyrimidines in the 12nt preceding the 4nt splice acceptor.

## Results and Discussion

### Structure and splicing of *Cxqdsx*

TBLASTN identified strong alignments to both *Aeadsx* OD1 and OD2 domains on *Cx. quinquefasciatus* supercontig 3.59. Further homology searches via TBLASTN (not shown) identified putative local alignments to both the common (exons 2 and 4 of *Aeadsx*), female specific (exon 5) and male-specific (exon 6) CDS sequence on that

same contig. The primer pair quinqOD12F/quinqOD12R (Fig. 2), designed to amplify the common regions of the OD1 and OD2 domains, produced a double-band in both male and female *Cx. quinquefasciatus* cDNA. Sequencing and genome alignment revealed this was due to the presence of a 75bp (25 amino acid) alternatively spliced in-frame intronic sequence within exon 2 that was present in some transcripts but spliced out of others (Fig. 3). An equivalent 63bp (21 amino acid) tract was reported from *Aeadsx* and a 72bp (24 amino acid) tract reported in *Angdsx* (Salvemini et al. 2011), however this appears to be specific to the Culicidae and has not been reported from sequenced *dsx* transcripts in other taxa. The conservation and evolution of this splicing event within the mosquitoes is evidence of an as yet undetermined functional role. Both male and female N-termini of the *Cxqdsx* gene contained two small 45bp exons homologous to exons 3a and 3b of *Aeadsx* (Fig. 1). The primer pair quinqOD12F/quinqOD12Rfem, designed to amplify the putative female-specific transcript by binding the 3' end of the OD2 domain in exon 5, generated product only in female cDNA (Fig. 4A) thus confirming the sex-specific splicing of the mRNA and the location of the female-specific exon. By using a forward primer located downstream of the in-frame intron in exon 4 (DSX8F) and reverse primer within the putative male-specific/common exon 6 (DSX6R), we generated an amplicon spanning a ca. 1,079bp exon (exon 5, Fig. 4B) specific to the female that was spliced to exon 6 after removal of 2.9kb of intronic sequence (Fig. 1). We find no evidence for an alternative female spliceform involving a second female-specific exon as is present in *Aeadsx* (Salvemini et al. 2011), indicating that the phenomenon is likely an intron gain in *Aedes* rather than a loss in *Anopheles*.



The male RT-PCR product lacked this exon, and consisted of a smaller amplicon splicing exons 4 and 6 (Fig. 4B). Both males and females shared the C-terminal male-specific/common exon 6 (as UTR in the female) as in *Angdsx* and *Aeadsx*.

5' RACE-PCR, after final amplification with primer DSX-5RACEGSP2, produced an identical ca. 1400bp amplicon from both male and female cDNA (Fig. 4C) that extends exon 2 of the transcript 451bp upstream of the start codon, meets 9,005bp of intronic sequence, and is spliced to an 856bp exon 1/UTR (Fig. 1). The transcription start site (TSS, position 671,074 of supercontig 3.59) falls on the adenine nucleotide of an initiator (Inr) sequence YYANWY (Fig. 5, (Smale and Baltimore 1989)) with a putative downstream promoter element (DPE) motif RGWY(T) at position +28. No TATA box was found. Pending functional validation, this region may thus represent a *Cxqdsx* promoter.

To qualify our *Cxqdsx* gene model, we mapped the short-read Illumina RNAseq data in NCBI SRA accession SRR991016 generated by Leal et al. (Leal et al. 2013) to supercontig 3.59 and manually annotated *Cxqdsx*. The transcript was well represented in these data, and the structure congrued with our RT-PCR and 5'RACE results in the placement and splicing of all previously described exons including the lack of additional spliceforms in female-specific exon 5 as well as the sequenced 5' common end of exon 6. Additionally, these data allowed us to define the C-terminus of *Cxqdsx*, including the full 1,016bp male-specific/common exon 6 and its splicing over 13,814bp of intron to a terminal 2,201bp 7<sup>th</sup> exon/UTR (Fig. 3.1, Fig. 3.S1). The final *Cxqdsx* protein product (Fig. 3.3) initiates translation in both females and males from the start codon in the common exon 2, and terminates in female mosquitoes at

the opal-ochre double stop codons (conserved in Diptera, see (Kuhn et al. 2000)) within exon 5 and in male mosquitoes at a stop codon within exon 6. Exons 6 and 7 are thus transcribed entirely as UTR in the female isoform, as has been shown in other Dipterans including *Megaselia scalaris* (Kuhn et al. 2000), *Anopheles gambiae* (Scali et al. 2005) and *Aedes aegypti* (Salvemini et al. 2011).

The RNAseq mapping revealed an additional alternative splicing event that was not reflected in our RT-PCR experiments; a 160bp extension of exon 4 (exon4ex) resulting in use of an alternate downstream splice donor (Fig. 3.S2) to the female-specific exon 5 acceptor. The putative peptide from this mRNA terminates within the extension at a double stop (TAATAA) codon 89bp from the previously recognized splice donor site and encodes 30 amino acids. This is the same number of amino acids encoded by the ORF within exon 5, thus both splice forms produce peptides of equivalent length (Fig. 3.3). Of the 166 reads in the library splicing exons 4 and 5, 47 (28.3%) splice exon 4 from the extension and 119 (71.7%) from the canonical position. Our RT-PCR using male cDNA and primers quinqOD12F/DSX6R produced only the expected double-band (with and without the 75bp in-frame intron) at 950bp, while the female reaction using primers DSX8F (downstream of the in-frame intron) and DSX6R produced the single band mentioned previously (Fig. 3.4B). To address the possibility that we failed to detect a second amplicon in the latter reaction, we performed a follow-up RT-PCR on female cDNA with primers quinqOD12F/quinqOD12RF; this generated four bands (Fig. 3.6) at sizes commensurate with those generated by removal of the exon 4 extension and/or the exon 2 in-frame intron (ca. 905, 830, 745 and 670bp). To

confirm the occurrence of the transcript variant and its restriction to female cDNA, we next searched Illumina RNAseq libraries prepared from *Cx. pipiens* f. *pipiens* and f. *molestus*, and *Cx. pipiens pallens* mosquitoes of mixed sex and life stages (Price and Fonseca 2014, in review; see additional file 2 for details and accession numbers) for presence of the exon4ex donor and for male-specific splicing of the exon 4 extension to exon 6. We found 58 amplicons (28.6%) that spliced the extension to the female-specific exon 5 (n=36, 13 and 9 in *Cx. p. f. pipiens*, f. *molestus* and *Cx. pip. pallens*, respectively) and 145 (71.4%) from the canonical position (n=100, 21 and 24 in *Cx. pip. f. pipiens*, f. *molestus* and *Cx. pip. pallens*, respectively). These numbers are nearly identical to those from *Cx. quinquefasciatus*, yet none were spliced to exon 6. This is evidence that the alternate isoform is likely specific to the female and comprises roughly 28% of female *dsx* isoforms in the mosquitoes studied. Additionally, the alternate splice donor appears to be conserved within the *Cx. pipiens* complex. The final *Cxqdsx* gene (Fig. 3.1) is composed of eight exons and spans 247,017bp of supercontig 3.59.

#### Completing the *Aeadsx* gene

To compare the size, structure, intron characteristics and putative promoter regions of our full-length gene model with that of the other sequenced Culicine mosquito, *Ae. aegypti*, we used publicly available Illumina short-read RNAseq data to discern in-silico the 5'UTR, transcription start site, exon 1 and full 3'UTR of *Aeadsx* (Salvemini et al. 2011). To predict the 5' end of *Aeadsx*, we mapped Illumina short-read RNAseq libraries from NCBI SRA accession SRR789758 to *Aedes aegypti*

strain Liverpool supercontig 1.370 as performed previously and located exon 2 defined by Salvemini et al. (Salvemini et al. 2011). By visual inspection of the mapping, we were able to extend the 2<sup>nd</sup> exon 472bp upstream of the start codon, define a splice junction spanning 14,481bp of intronic sequence, and locate a 1,388bp 1<sup>st</sup> exon/5'UTR (Fig. 3.1, Fig. 3.S3). As RNAseq mapping provides only approximate definition of transcript ends, we searched for a promoter motif within an area +/- 250bp from the point at which 5' short-read coverage for exon 1 ceased. We located an initiator element (Inr) of the form YYANWYY at position 109460 of the reverse-complemented supercontig 1.370 and a downstream promoter element (DPE) of the form RGWYV at canonical position +28 from the Inr adenine (Fig. 3.5), thus providing strong evidence for the *Aeadsx* transcription start site. As in *Cxqdsx*, no TATA box was found.

These transcriptome data disagree slightly with the C-terminus of the currently described *Aeadsx* transcript (see Table 3.1 of Salvemini et al. [2011]) in that we find 16,895bp of intronic sequence between exon 6 and the terminal/UTR exon 7 as opposed to the 22,437bp reported, and our data support a very large 6,382bp 7<sup>th</sup> exon (position 654172 – 660554 of reverse-complemented supercontig 3.59) as opposed to the reported 449bp (Fig. 3.S4). Additionally, we find the upstream splice acceptor to female-specific exon 5b to use canonical gt/ag splicing (Fig. 3.S5) as opposed to the suboptimal gt/gt splicing reported. This does not change the comparatively high number of purines in the polypyrimidine tract or the status of exon 5b as weak (and requiring splice activation). The final *Aeadsx* gene model (Fig. 3.1) spanned 471,155bp of supercontig 1.370.

## Repetitive elements

The genera *Aedes* and *Culex* are estimated to have diverged ca. 52 Mya (Arensburger et al. 2010). The genome size for *Cx. quinquefasciatus* currently stands at 540Mbp (Arensburger et al. 2010), while that of *Ae. aegypti* is estimated to be over twice that size at 1.3Gbp, largely due to the accumulation of transposable elements (TEs)(Nene et al. 2007). As TEs are not distributed randomly within chromosomes (Duret et al. 2000, Bartolomé et al. 2002), we assessed the frequency of repetitive elements within the *doublesex* gene in order to determine whether different classes have invaded the respective *dsx* genes of *Cx. quinquefasciatus* and *Ae. aegypti*. We used CENSOR (<http://www.girinst.org/censor/index.php>) to scan *Cxqdsx* introns 2-7 and compared the results to those for *Aeadsx* intron 2-8 (Salvemini et al. 2011) (Table 3.S2, Table 3.S3). The two genes contain nearly identical numbers of DNA transposons and similar numbers of LTR retrotransposons, however *Aeadsx* was found to contain nearly twice as many Non-LTR retrotransposons (or LINEs). These elements persist with great success in eukaryote genomes (Han 2010) and comprise 4% and 14% of the transposable elements in the *Cx. quinquefasciatus* and *Ae. aegypti* genomes, respectively (Nene et al. 2007, Arensburger et al. 2010), thus their abundance in *doublesex* likely reflects the genome-wide pattern.

## Regulatory mechanisms of *Cxqdsx*

All splice junctions of *Cxqdsx* use conserved GT-AG splice donor/acceptor motifs (Table 3.1). Interestingly, we find that the number of purines in the polypyrimidine tract of the 3' splice acceptor preceding the common/male-specific exon 6 (n=5) deviates significantly from the calculated mean (8.58, +/- 1.39 SE, see Methods) and constitutes a suboptimal splice acceptor. This is contrary to *Aeadsx*, which is hypothesized to activate a weak splice acceptor upstream of the female-specific exon 5b (Salvemini et al. 2011), and *Angdsx* which likely relies on activation of the 5' weak splice donor downstream of exon 5 (Scali et al. 2005).

To define putative regulatory mechanisms which may govern the sex-specific splicing of the female-specific exon 5 and/or the enhancement of the weak 3' splice acceptor preceding the male-specific exon 6, we searched intron 4, exon 5, intron 5 and exon 6 (8,045bp of sequence) for putative cis-acting elements derived from consensus alignments of *D. melanogaster*, *An. gambiae* (when available) and *Ae. aegypti* TRA/TRA2 binding sites (NMDNCRWNCWAYM), the *Nasonia vitripennis* TRA/TRA2 binding site (NGAAGAWN), the RBP1 type A and B motifs (DCADCTTTA and ATCYNNA) and the TRA-2-ISS motif (CAAGR, see Fig. 3.7 and Table 3.S4, Fig. 3.S6 for all *cis*-elements discussed below). Six copies of the TRA/TRA2 motif (two of which were overlapping) were found within a 224bp stretch at the 3' end of the female-specific exon 5. Three copies exhibit strong similarity ( $\geq 69\%$ ) to the *D. melanogaster* TRA/TRA2 sequence at the nucleotide level, while the remaining three deviated from *D. melanogaster* (46-61%) yet adhered to the consensus motif. Other dipterans including *Drosophila* maintain six copies of the dsxRE to facilitate

recruitment of splice factors to the female-specific splice site (Tian and Maniatis 1993, Lynch and Maniatis 1995). Their presence may thus be evidence for a functional significance in *Cxqdsx* splicing, and the action of a TRA-like factor in splicing *Cxqdsx* pre-mRNA. To assess the significance of this cluster, we searched the *Cx. quinquefasciatus* transcriptome for additional windows of 224 bp containing six copies of the consensus motif. Two genes (.01% of 19,019 total CDS sequences), CPIJ009301 (9 copies) and CPIJ007662 (8 copies) met this criterion. Both genes are currently annotated as 'hypothetical proteins' in VectorBase and maintain little homology to other peptides in the NCBI nr database (not shown). A single gene (CPIJ002327) contained 4 copies in 224 bp, while none remaining contained more than three. Additionally, we used the AhoPro software of Boeva et al. (Boeva et al. 2007) to determine the probability of observing six copies of the motif in 8,045bp (regardless of clustered distribution) to be  $4.8 \times 10^{-3}$ . The probability of observing six copies in 224bp is  $1.91 \times 10^{-7}$ . Six putative purine-rich elements (PREs) were identified, three of which were in the canonical position within exon 5 near the TRA/TRA2 binding sites, however two copies were found in intron 4 and one in intron 5. The function, if any, of the latter three elements currently remains unclear. Movement of the TRA/TRA2 enhancer sites (proximal to the splice acceptor of the female-specific exon in *Drosophila dsx*) downstream to the distal splice donor of the female-specific exon (exon 5b of *Aeaddsx*, see Fig. 3.1) appears to be conserved in the mosquitoes, however the exact effect of this placement on splicing to create the female isoform remains unknown. In *Drosophila*, they activate the splice acceptor of the female-specific exon (Lynch and Maniatis 1995), and are hypothesized to do the

same to exon 5b of *Aeadsx* (Salvemini et al. 2011); in *Anopheles*, they appear to activate the splice donor immediately downstream of the female-specific exon (Scali et al. 2005) (as they do in the *fruitless* gene of *D. melanogaster* [Lam et al. 2003]).

Twenty-two copies of an RBP1 type B motif were present; fourteen copies were located outside of exon 6, however these were represented by eleven different permutations of the consensus sequence. Each 7nt permutation had a BLASTn e-value of 1.3 when queried against the full *Cxqdsx* gene sequence, and (in the absence of a clustered distribution) can be expected to occur at least once by chance. Eight copies, however, were clustered in a 546bp stretch at the 5' end of the male-specific exon 6. Repeating the protocol used in the TRA/TRA2-like enrichment test above, we find 86 of 19,019 transcripts (0.45%) contain 8 or more copies of the RBP1b consensus in a 546bp window. Many of these contigs generated positive results due to tandem repeats however (Table 3.S5). Using AhoPro (Boeva et al. 2007), we determined the probability of observing this motif in 546bp of randomly generated sequence data to be  $2.89 \times 10^{-5}$ . A cluster of Rbp1 binding sites and TRA-2-ISS elements upstream of the “strong” female-specific exon 5a of *Aeadsx* are hypothesized to manage the differential splicing of this exon while other TRA/TRA2-like elements enhance the “weak” exon 5b (Salvemini et al. 2011) (see Fig. 3.1). The localization of this RBP1-binding cluster near the “weak” or suboptimal splice acceptor in *Cxqdsx* exon 6 indicates a SR-like factor may be involved in its splicing. This presents a curious model, as exon 6 is included in both male and female spliceforms. It is thus likely that if exon 6 requires activation by a SR-like factor, it would occur in the male-specific spliceform and facilitate excision



of the female-specific exon 5. This would require use of the exon 6 splice acceptor at the expense of exon 5, and could be facilitated by the Rbp1 elements. The functional TRA/TRA2-like factor present in the female would then suffice to maintain incorporation of exon 6 as UTR. Five copies of the TRA-2-ISS motif were found but were not in significant representation. Three copies of the NvTRA element were found, however unlike in *Aeadox* that maintains four copies within a cluster in exon 5, two copies were found in intron 4 and one in exon 5. The BLASTn e-value of each 8bp hit within the search area was 0.37, thus we cannot exclude this result as having occurred simply by chance.

#### Sequence evolution of *Cxqdsx*

Assembling the complete *doublesex* transcript from two members of the *Culex pipiens* complex (*Cx. quinquefasciatus* and *Cx. pipiens form pipiens*) allowed us to examine the rate of peptide evolution within this integral gene between closely related mosquito species. Using a sliding window approach along a pairwise codon alignment of the male and female *doublesex* isoforms (Fig. 3.S7, Fig. 3.S8) we graphed the Ka/Ks values along the gene length. The female isoform alignment, inclusive of the common OD1 and OD2 domains, was devoid of non-synonymous substitutions and thus both Ka and Ka/Ks indicated only purifying selection. The male isoform however exhibited elevated Ka and Ks values along the majority of the male-specific C-terminus of the peptide, with  $\omega$  reaching maximal values in several locations (Table 3.S6). These results indicate that particular regions of the isoform may be under positive selection. The 5' end of the male-specific region has been

shown to exhibit signs of positive selection in the *Anastrepha fraterculus* species group (Sobrinho and de Brito 2012), however unlike *Anastrepha*, we find significantly higher levels of peptide evolution (Ka) and potential positively selected sites (Ka/Ks) in the male-specific *doublesex* transcript as compared to the female-specific and common regions in these closely related mosquitoes. Hughes (2011) proposed a mechanism for this observation based on the fact that 1) *doublesex* influences not only development of insect genitalia but also of morphological and behavioral secondary sex characteristics (Siwicki and Kravitz 2009, Kijimoto et al. 2012, Devi and Shyamala 2013) and 2) these secondary traits are commonly exaggerated and diverge rapidly during sexual selection in response to female choice (Emlen 2008). If female choice itself were a product of neutral mutation (Nei 2007), the pleiotropic repercussions of evolving linked male characters in response could create “runaway” evolutionary pressures on the male-specific DSX protein and result in the Ka and Ka/Ks patterns witnessed in our data.

## Conclusions

Our results show that the *Cx. quinquefasciatus doublesex* gene exhibits sex-specific splicing, as it does in the mosquitoes *Ae. aegypti* and *An. gambiae*, as well as in other Diptera. *Cxqdsx* shares characteristics of both *Aeadsx* (gain of exon 3b, Rbp1 cis-regulatory binding sites) and *Angdsx* (singular female-specific exon, shared 3' UTR), as well as a novel spliceform generated from an alternate exon 4 splice donor that appears to occur only in the female. Additionally, we complete the full-length *Aeadsx* model and identify a putative TATA-less Inr/DPE core promoter region in

both *Cx. quinquefasciatus* and *Ae. aegypti* mosquito genomes, allowing for future *in situ* validation and studies of *dsx* gene transcription.

We find that *cis*-regulatory splicing regulation of *Cxqdsx* does not appear to follow either currently described mosquito model, and instead involves activation of a weak splice acceptor of the male-specific/common exon 6, possibly involving a cluster of local Rbp1 binding sites as enhancers. This finding further exemplifies the diversity present in upstream splicing regulation of *dsx* within mosquitoes, as each of the three genera studied (*Anopheles*, *Aedes* and *Culex*) possess unique regulatory mechanisms despite maintaining TRA/TRAX-like binding sites in the 3' end of their respective female-specific exons (exon 5b in *Aeadsx*).

An analysis of peptide evolutionary rates between *Cxqdsx* and the *dsx* gene of the closely related *Cx. pipiens* form *pipiens* (*Cxpipdsx*, also generated in this study) shows that the male-specific component of the transcript has evolved at accelerated evolutionary rates relative to the female isoform, and contains sites exhibiting signs of positive selection. This result accentuates the rapid evolution of *doublesex* within the *Culex* species complex. Future research defining the degree to which *doublesex* influences the sexual selection cycle may shed light on the role (if any) that this integral gene plays in incipient speciation within insects.

#### *Availability of supporting data*

The nucleotide sequences for the male and female-specific *Cxqdsx* transcripts have been submitted to GenBank under accession numbers KP033512 and KP033513, respectively.

Sequences for male and female-specific *Cxpipdsx* transcripts have been submitted under accession numbers KP033514 and KP033515.

### **Acknowledgements**

We are grateful to Linda McCuiston for her unsurpassed expertise in rearing and colonizing the mosquitoes used in our study and to Nicole Wagner at the Rutgers University School of Environmental and Biological Sciences Genome Cooperative for performing our Illumina Sequencing. This work was funded by a New Jersey Mosquito Control Association Daniel M. Jobbins scholarship to DCP and by NE-1043 Multistate funds to DMF.

**Table 3.S1** Primer sequences used for 5' RACE-PCR and to amplify RT-PCR products of *Cxqdsx*.

<b>Primer</b>	<b>Sequence 5'-3'</b>
quinqOD12F	ATACCTGGATGGAGACGA
quinqOD12Rcomm	ACCCTTCAGTATCACGTACA
quinqOD12Rfem	AGATTGTGTAACCGTGAG
quinqDSX6R	TTGGCTGCTTTGGCTTGA
quinqDSX7F	TGTGAGTGAGTGAAAGTG
quinqDSX7R	GGAGTGCGTTTGATAGGG
quinqDSX8F	CCCCTGATGTACGTGATACT
quinqDSX5RACE-GSP1	GTGATCTTCGATGTAGTG
quinqDSX5RACE-GSP2	AGCTTGGGTATGTGAATGT

**Table 3.S2** CENSOR tabular output with heat-map diagram for *Cx. quinquefasciatus* *doublesex* introns 1 through 7.

Table 3.S2 is located in the online supplementary material.

**Table 3.S3** Comparison of mobile elements within *Ae. aegypti* dsx introns 2-8 and *Cx. quinquefasciatus* introns 2-7

<i>Aedes</i> introns 2-8					<i>Culex</i> introns 2-7				
All Elements	DNA Transposon	LTR Retrotransposon	Non-LTR Retrotransposon		All Elements	DNA Transposon	LTR Retrotransposon	Non-LTR Retrotransposon	
DNA	27	38	25	11	DNA	131	131	8	4
DNA Academ	1	36	LTR BEL	5	DNA Academ	1	LTR BEL	LTR Copta	5
DNA Chapae	20	27	LTR Copta	12	DNA Chapae	2	LTR Copta	LTR Gypsy	48
DNA Crypton	1	21	LTR DIRS	1	DNA Dada	1	Total	61	1
DNA EnSpm	21	DNA EnSpm	36	Non-LTR Daphne	DNA EnSpm	3	EnSpm CACTA	Non-LTR Jockey	12
DNA Gingerl	3	DNA Zator	79	Non-LTR I	DNA Gingerl	1	DNA EnSpm	Non-LTR Kiri	1
DNA Harbinger	3	DNA Sofia	13	Non-LTR Jockey	DNA Harbinger	3	DNA Harbinger	Non-LTR L1	11
DNA hAT	38	DNA Sola	12	Non-LTR L1	DNA hAT	10	DNA hAT	Non-LTR L2B	1
DNA Heltron	7	DNA Sola	9	Non-LTR Penelope	DNA Heltron	5	DNA Heltron	Non-LTR R2	2
DNA Heltron	36	DNA Heltron	7	Non-LTR R4	DNA IS3EU	1	DNA IS3EU	Non-LTR RTE	1
DNA ISL2EU	7	DNA ISL2EU	7	Non-LTR RTE	DNA Kolobok	1	DNA Kolobok	Non-LTR RTE	3
DNA Kolobok	2	DNA Mariner	7	Non-LTR SINE	DNA Mariner	10	DNA Mariner	Non-LTR SINE SINE2	24
DNA Kolobok	1	DNA P	5	Non-LTR SINE SINE2	DNA MuDR	3	DNA MuDR	Non-LTR Tadi	1
DNA Mariner	7	DNA P	5	Non-LTR Tx1	DNA P	2	DNA P	Non-LTR Tx1	2
DNA Merlin	4	DNA Polinton	1	Total	DNA Polinton	6	DNA Polinton	Total	69
DNA MuDR	4	DNA Merlin	4	132	DNA Sola	39	DNA Sola		
DNA P	5	DNA Harbinger	3		DNA Transib	6	DNA Transib		
DNA piggyBac	1	DNA Relavkus	3		DNA Zator	21	DNA Zator		
DNA Polinton	3	DNA Transib	3		ERV ERV1	2	Total		
DNA Relavkus	3	DNA Kolobok	2		ERV ERV2	1			
DNA Sofia	13	DNA Academ	1		Interspersed Repeat	43			
DNA Sola	12	DNA Crypton	1		LTR BEL	8			
DNA Transib	3	DNA Kolobok	1		LTR Copta	5			
DNA Zator	19	DNA piggyBac	1		LTR Gypsy	48			
ERV	2	Total	248		Non-LTR CR1	4			
ERV ERV2	23				Non-LTR Crack	3			
ERV ERV3	1				Non-LTR I	1			
Interspersed Repeat	7				Non-LTR Jockey	12			
LTR	25				Non-LTR Kiri	1			
LTR BEL	5				Non-LTR L1	11			
LTR Copta	12				Non-LTR L2B	1			
LTR DIRS	1				Non-LTR R2	2			
LTR Gypsy	36				Non-LTR RTE	1			
Non-LTR	11				Non-LTR RTE	3			
Non-LTR CR1	11				Non-LTR SINE	3			
Non-LTR Crack	2				Non-LTR SINE SINE2	24			
Non-LTR CRE	1				Non-LTR Tadi	1			
Non-LTR Daphne	1				Non-LTR Tx1	2			
Non-LTR I	1				Pseudogene tRNA	1			
Non-LTR Jockey	36				Simple Sat	1			
Non-LTR L1	9				Simple Sat MSAT	13			
Non-LTR Penelope	1				Total	437			
Non-LTR R4	1								
Non-LTR RTE	21								
Non-LTR SINE	34								
Non-LTR SINE SINE2	2								
Non-LTR Tx1	1								
Pseudogene tRNA	6								
Simple Sat	7								
Simple Sat SAT	3								
Total	508								

**Table 3.1** Splice donors and acceptors of the Cxqdsx gene. Coding (exon) sequences are in uppercase text, while the splice donor/acceptor and succeeding/preceding 12 nucleotides, respectively, are in lowercase. “Exon 4ex” denotes the alternate downstream splice donor of exon 4. Asterisk indicates the splice acceptor site deviates significantly from the genomic mean of 8.58 +/- 1.39 SE

	<b>exon end</b>	<b>Splice donor</b>	<b>intron</b>	<b>Splice acceptor</b>	<b>next exon begin</b>	<b>No. purines</b>
exon 1	AAAAAG	gtgggcttctttatct	intron 1	ctttttcccgtttcag	ATCCTTGCTT	11
exon 2	AAGGAG	gtaagttcgcaacctc	intron 2	cctcctctcttttgacg	CCAATCATGC	12
exon 3a	TACCAG	gtacgtgtcttccgct	intron 3a	cattatatcatcttcag	TCCCTCCAAA	8
exon 3b	GATCAG	gtgagtgtctagaagtc	intron 3b	tattatcccccttttcag	ACGATGAACT	10
exon 4	ACGAAG	gtatggccgagtggtc	intron 4	ttccggttcctacgcag	GTCAAAGCCGT	10
exon 4ex	TAAAT	gtacgcaagagagattcg	intron 4ex	ttccggttcctacgcag	GTCAAAGCCGT	10
exon 5	TGACAG	gtacttgaactaatta	intron 5	ccaaccaacaaaaacag	CTCAGGCTGT	5*
exon 6	GCGAAG	gtgagttgagcattgt	intron 6	cttatcatcattacag	ATGCCGCTAG	9
consensus		gtrnk-----		-----ncag		



**Table 3.S4** Putative *cis*-element motifs (TRA/TRA2, *N. vitripennis* TRA/TRA2, TRA-2-ISS and RBP1 type B) of *D. melanogaster*, *An. gambiae*, *Ae. aegypti* and *Cx. quinquefasciatus*. Genomic location of *Cx. quinquefasciatus* elements are listed.

TRA/TRA2	Sequence	Location	RBP1 type B	Sequence	Location
<i>D. melanogaster</i>	TCTTCAATCAACA		<i>D. melanogaster</i>	ATCCNNA	
	TCAACAATCAACA			ATCTNNA	
<i>An. gambiae</i>	TCGCCGATCAACC		<i>Ae. aegypti</i>	ATCCACA	
	CCATCGTTCAACC			ATCTCTA	
	TCTCCAATCAATC			ATCCGAA	
	ACATCAATCAATA			ATCTGAA	
	ACATCAATCAATC			ATCTAGA	
<i>Ae. aegypti</i>	AATACAAACAACA			ATCCACA	
	TCAACAAGCAACA			ATCTACA	
	TCTTCAACCAACC		Consensus	<u>ATCYNNA</u>	
	CCTACAATCTACA		<i>Cx. quinquefasciatus</i>	ATCTGAA	intron 4
	GCTGCAATCAACA			ATCTACA	intron 4
Consensus	<u>NMDNCRWNCWAYM</u>			ATCCTGA	intron 4
<i>Cx. quinquefasciatus</i>	AAAACAACCAACC	exon 5		ATCCAAA	intron 4
	AAACCGTGCAACC	exon 5		ATCCGTA	intron 4
	CAACCATGCAACC	exon 5		ATCTGAA	exon 5
	CATGCAACCAATA	exon 5		ATCTGAA	exon 5
	CAATCAAACAACC	exon 5		ATCCTTA	exon 5
	GCTTCAATCAACC	exon 5		ATCTTTA	intron 5
	AAATCAATCAACC	intron 5		ATCCGCA	intron 5
				ATCTCGA	intron 5
<b><i>N. vitripennis</i> TRA/TRA2</b>	<b>Sequence</b>	<b>Location</b>		ATCCTCA	intron 5
<i>N. vitripennis</i>	TGAAGATT			ATCTTTA	intron 5
	GGAAGATA			ATCTAAA	intron 5
	CGAAGATC			ATCCCGA	exon 6
<i>Ae. aegypti</i>	CGAAGATC			ATCTGGA	exon 6
	GGAAGAAG			ATCCTTA	exon 6
	AGAAGAAT			ATCTGCA	exon 6
	CGAAGAAA			ATCTAGA	exon 6
	AGAAGAAT			ATCCGCA	exon 6
Consensus	<u>NGAAGAWN</u>			ATCCCTA	exon 6
<i>Cx. quinquefasciatus</i>	TGAAGATT			ATCTCCA	exon 6
	AGAAGAAC				
	TGAAGAAA				
<b>TRA-2-ISS</b>	<b>Sequence</b>	<b>Location</b>			
<i>D. melanogaster</i>	CAAGG				
	GAAGA				
<i>Ae. aegypti</i>	CAAGG				
	GAAGA				
Consensus	<u>CAAGR</u>				
<i>Cx. quinquefasciatus</i>	CAAGA	intron 4			
	CAAGG	intron 4			
	CAAGG	intron 4			
	CAAGG	intron 4			
	CAAGA	intron 4			

**Table 3.S5** RBP1 type-b motif enrichment scan results. *Culex quinquefasciatus*

transcript id, with maximum number of RBP1b motifs per 547bp window, unique RBP1b motif sequence permutations present in the window, and nucleotide gene sequence are shown.

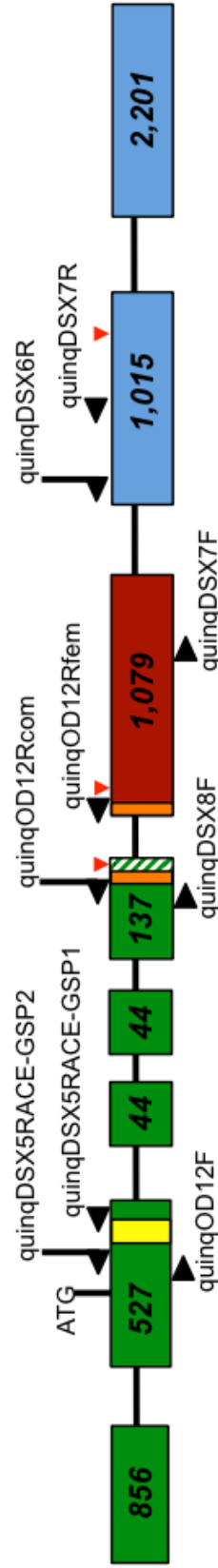
Table 3.S6 is located in the online supplementary material.

**Table 3.S6** Sliding window coordinates, Ka, Ks and Ka/Ks values calculated for each 30bp window of *Cx. quinquefasciatus* and *Cx. pipiens* form pipiens *dsx* CDS nucleotide alignment of male isoform.

Table 3.S6 is located in the online supplementary material.



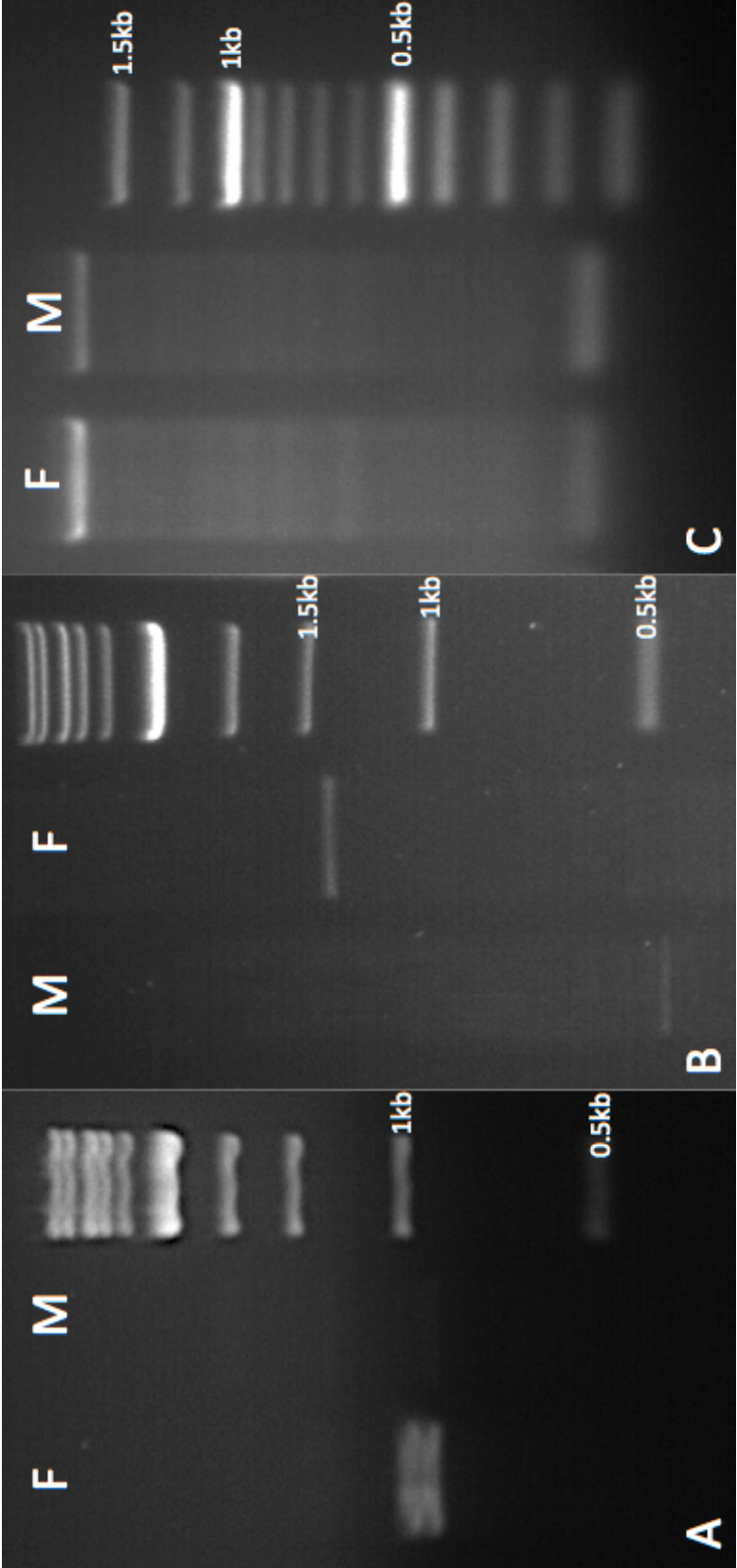
**Figure 3.2** Location of *Cxqdsx* RT-PCR primers (not to scale). Common exons are shown in green, the female-specific exon 5 in dark red, and male-specific (UTR in female) exons in blue. The exon4ex extension is represented with a green/white hatched box. The DBD/OD1 domain is indicated with a yellow box and OD2 with an orange box. Red triangles denote stop codons.







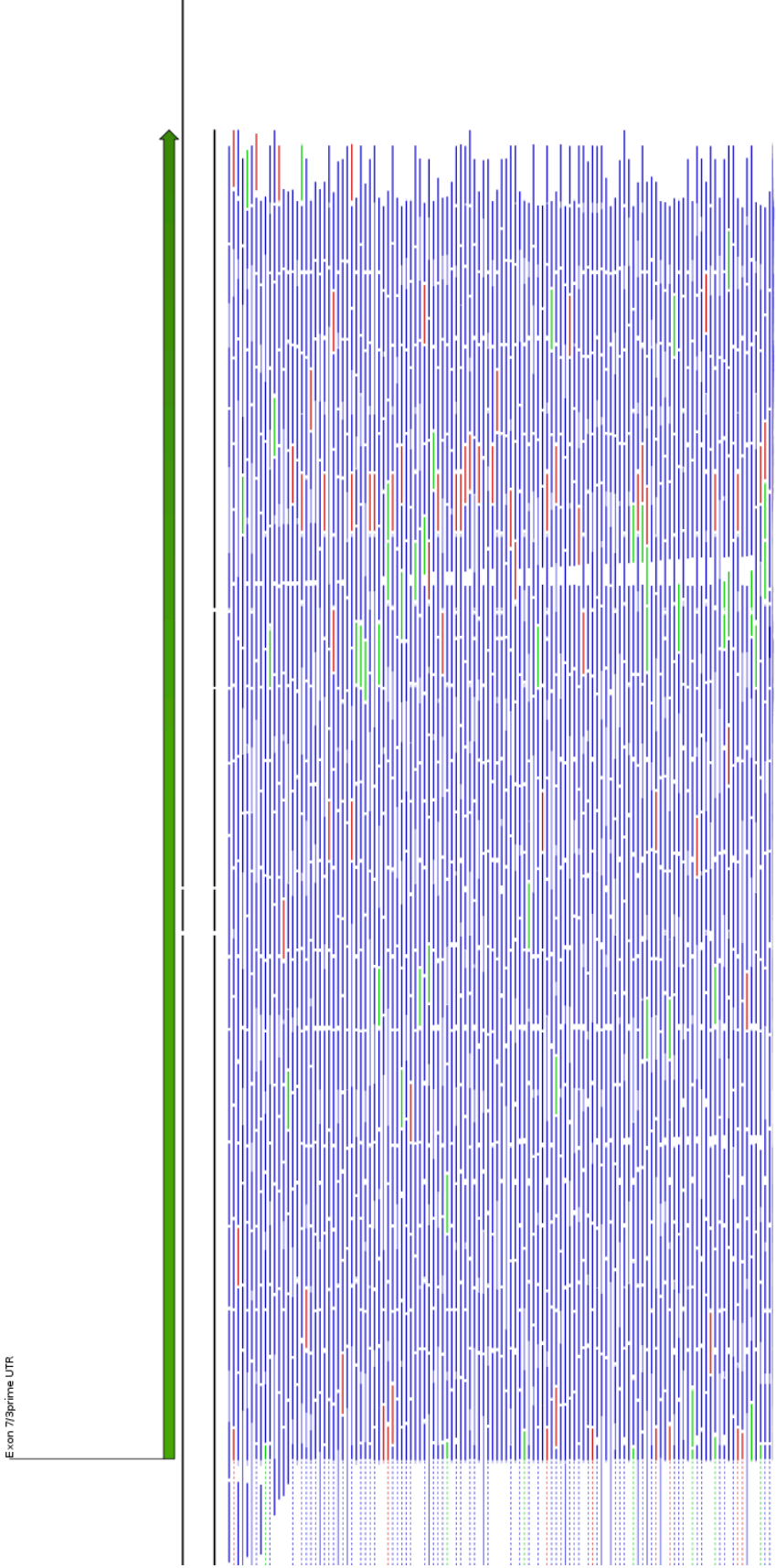
**Figure 3.4** RT-PCR gel visualizations. (A). RT-PCR products derived from primers quinqOD12F/quinqOD12Rfem used to amplify female (left) and male (right) cDNA. (B). RT-PCR products derived from primers dsx8F/dsx6R used to amplify male (left) and female (right) cDNA. (C.) RT-PCR products derived from 5'RACE-PCR reaction after final amplification with primer DSX-5RACEGSP2 on female (left) and male (right) cDNA.



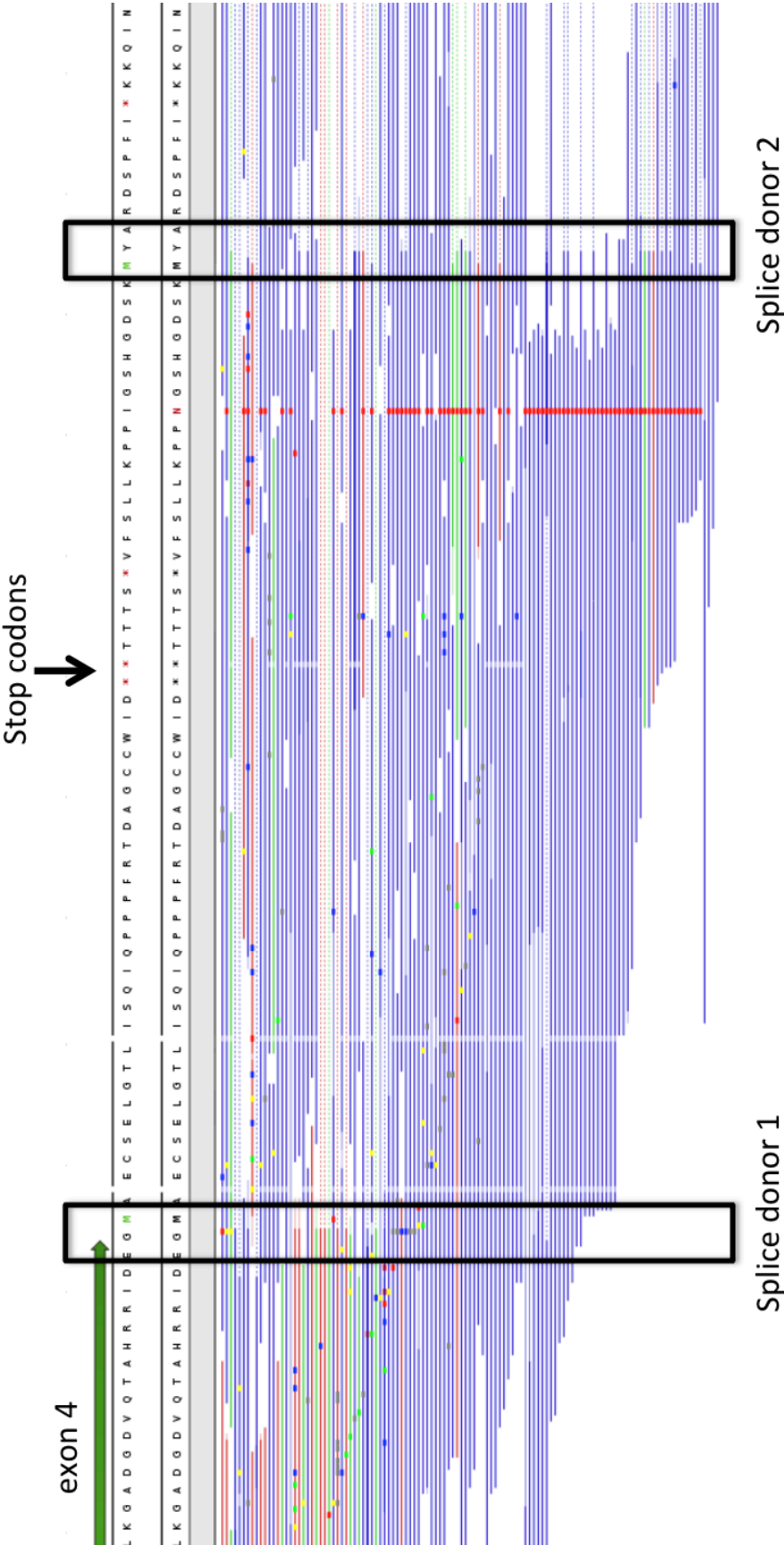




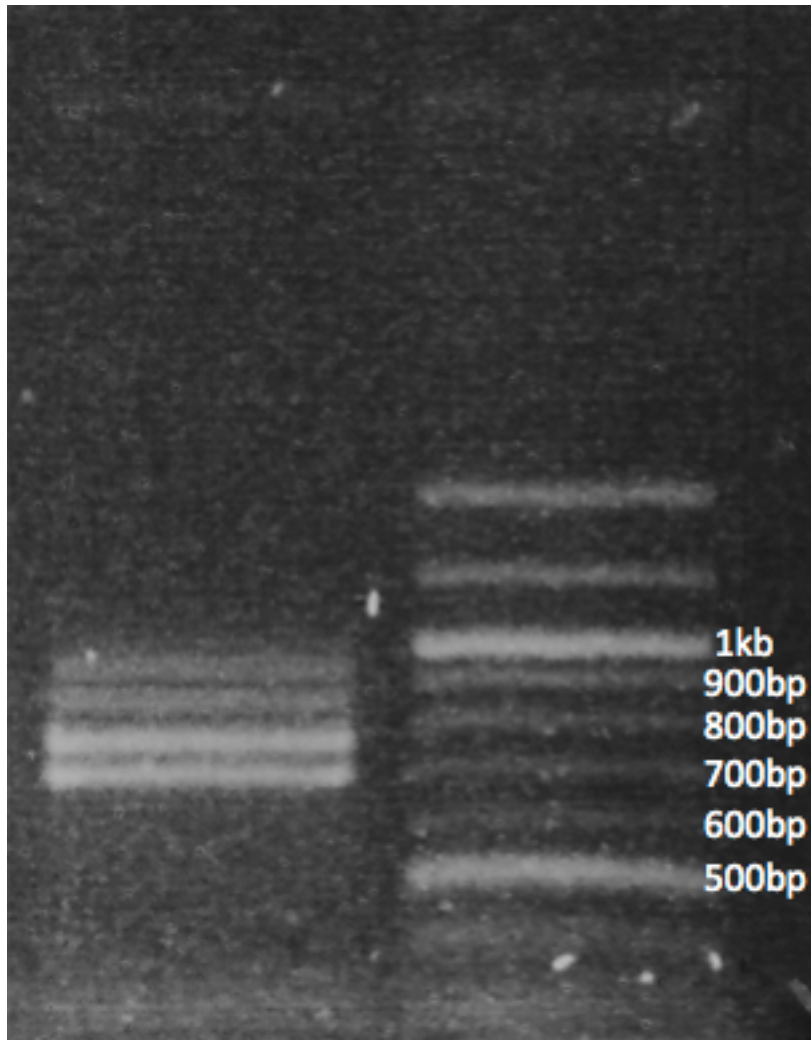
**Figure 3.S1** Short-read mapping of *Cx. quinquefasciatus* RNAseq data (below; paired-end reads in blue, single-end reads in red/green) generated by Leal et al. [42] to the derived location of *Cxqdsx* exon 7 (green arrow). Reads spanning the splice junction to exon 6 are indicated with dashes at left. Data are as visualized in the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark).



**Figure 3.S2** Short-read mapping of *Cx. quinquefasciatus* RNAseq data (below; paired-end reads in blue, single-end reads in red/green) generated by Leal et al. [42] illustrating alternate exon 4 splice donor (boxed). Reads spanning the splice junction to exon 5 are indicated with dashes. Data are as visualized in the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark).



**Figure 3.6** Female RT-PCR products. RT-PCR products derived from amplification of female cDNA with primers quinqOD12F/quinqOD12Rfem illustrating the four female-specific amplicons obtained by splicing of the exon 4 extension and/or the exon 2 in-frame intron (ca. 905, 830, 745 and 670bp).



**Figure 3.S3** Short-read mapping of *Ae. aegypti* RNAseq data (below; paired-end reads in blue, single-end reads in red/green) from NCBI SRA accession SRR789758 illustrating the derived location of *Aeadsx* exon 1 (green arrow). Reads spanning the splice junction to exon 2 are indicated with dashes at right. The exon 1 annotation begins at the transcription start site (TSS), or the first adenine nucleotide of the initiator (Inr) sequence. Data are as visualized in the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark).





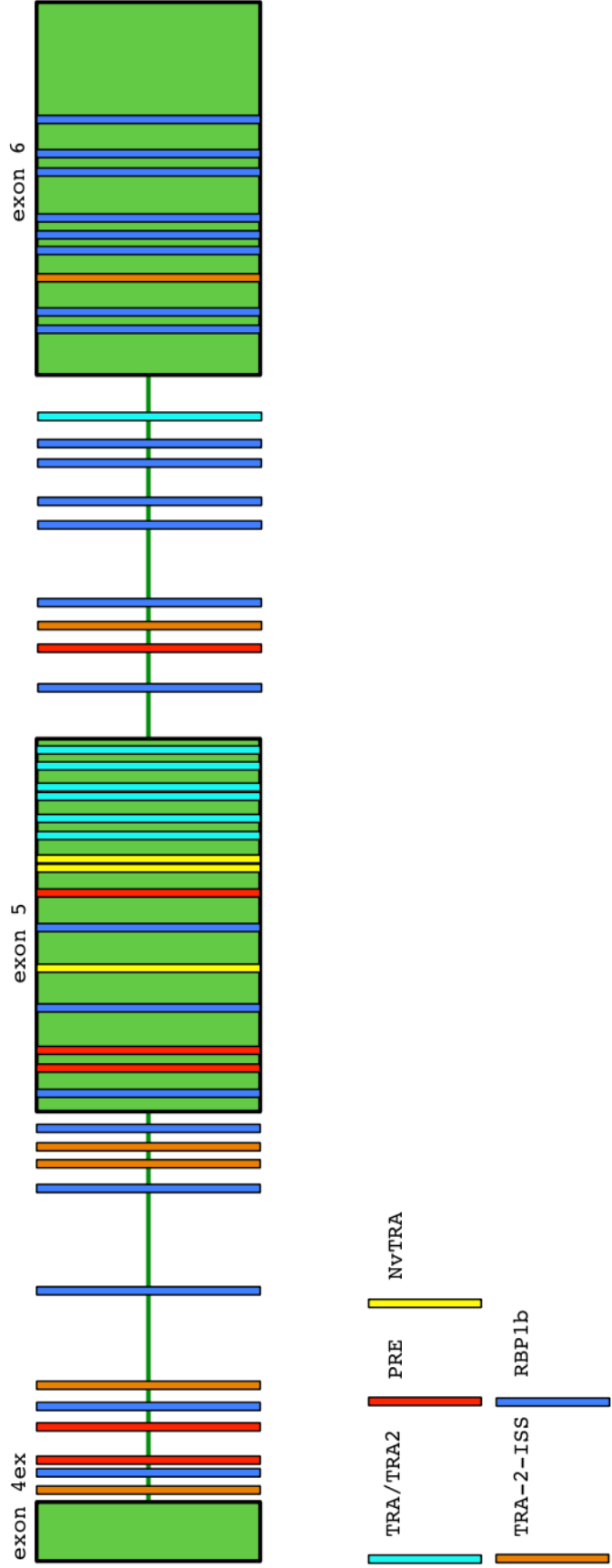
**Figure 3.S4** Short-read mapping of *Ae. aegypti* RNAseq data (below; paired-end reads in blue, single-end reads in red/green) from NCBI SRA accession SRR789758 illustrating the derived location of *Aeadsx* exon 7 (green arrow). Reads spanning the splice junction to exon 6 are indicated with dashes at left. Data are as visualized in the CLC Genomics Workbench (CLC Bio, Aarhus, Denmark).



**Figure 3.S5** Splicing and alignment of RNAseq reads (numbered 1 through 14) from NCBI SRA accession SRR789758 to the *Aeadsx* exon5a/5b junction (exon 5a in yellow, 5b in green) illustrating canonical gt/ag splice donor/acceptor.

	Exon 5a	intron 5	Exon 5b
CONSENSUS	TGTGAAACAGAAATAGAGCCAACTGTGCGCGGAGAAATGTTGAG	gtcat....cacag	TGCAAAATGCTGTTTAACGATAATAGCGACATGCAGC
1	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaac
2	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacg
3	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacg
4	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcg
5	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgaca
6	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc
7	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc
8	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc
9	tgtgaaacagagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc
10	-----acagaatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc
11	-----aatagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc
12	-----tagagccaaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc
13	-----ccaacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc
14	-----aacctgtgcgcggagagaatgttgag		tgcaaatgctgttttaacgataatagcgacatgcagc

**Figure 3.7** *Cis*-element distribution. Graphical representation of *cis*-element distribution within the exon 4 extension, intron 4, exon 5 (female-specific), intron 5 and exon 6 (male-specific/common UTR). Transformer/transformer 2 complex (TRA/TRA2) binding sites are colored light blue, purine-rich elements (PRE) are colored red, Nasonia-like TRA/TRA2 sites in yellow, TRA-2-ISS elements in orange and RBP2b elements in dark blue. Exons are represented as green boxes.

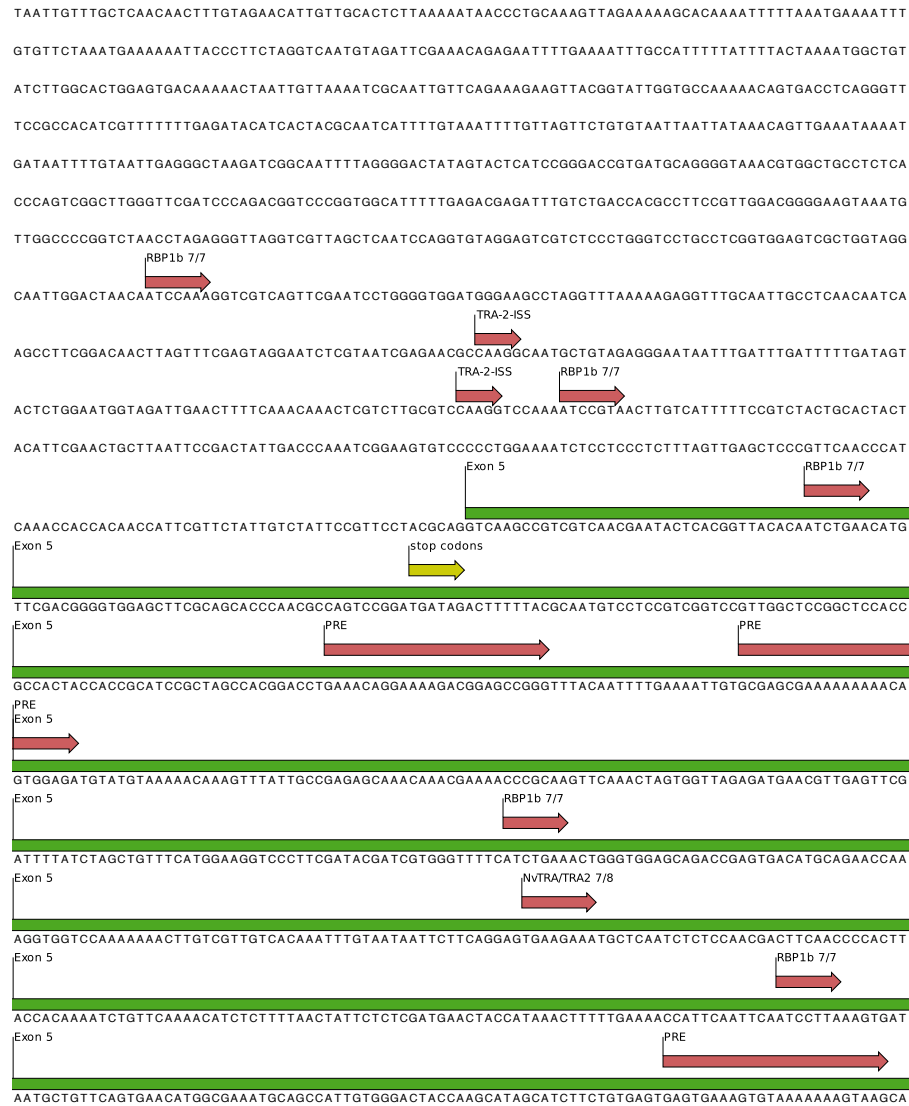




**Figure 3.S6** Nucleotide sequence of exon 4 extension, intron 4, exon 5, intron 5 and exon 6 with putative *cis*-element binding sites annotated. See Fig. 3.7 for graphical representation.



Figure 3.S6 Continued from previous page

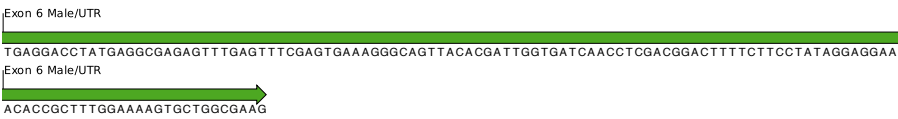




**Figure 3.S6** Continued from previous page



**Figure 3.S6** Continued from previous page



**Figure 3.S7** Amino acid (A) and nucleotide (B) alignment of *Cx. quinquefasciatus* and *Cx. pipiens* form pipiens female doublesex isoforms. Female-specific portion of protein highlighted in yellow.

#### A. Protein

*Cxqdsx* MYSQDTWMTSESGYEGRPDCGASGSSNSLNPRTPPNCARCNRNHLKIGLKGHRKYCYRSCNCEKCCLTAEQRVMALOTALRRAQTOEQRALNDGEVAPEPVHNIHIPKLSLKE  
*Cxpipdsx* MYSQDTWMTSESGYEGRPDCGASGSSNSLNPRTPPNCARCNRNHLKIGLKGHRKYCYRSCNCEKCCLTAEQRVMALOTALRRAQTOEQRALNDGEVAPEPVHNIHIPKLSLKE  
 \*\*\*\*\*

*Cxqdsx* MKHNLHNSQQORSLIDCDSSSTGSMNSTPCTSSMALPLHRRSPTGPVHPGEAQHLGANHASVSEPEANLLPVPPNIRVHHGPDSSRSDDELVKRSQYLLLEKLNYPWEMMPLMYVILKGADG  
*Cxpipdsx* MKHNLHNSQQORSLIDCDSSSTGSMNSTPCTSSMALPLHRRSPTGPVHPGEAQHLGANHASVSEPEANLLPVPPNIRVHHGPDSSRSDDELVKRSQYLLLEKLNYPWEMMPLMYVILKGADG  
 \*\*\*\*\*

*Cxqdsx* DVQTAHRRIDE**GQAVVNEYSRLHNLNMF**DGVELRSTQRQSG  
*Cxpipdsx* DVQTAHRRIDE**GQAVVNEYSRLHNLNMF**DGVELRSTQRQSG  
 \*\*\*\*\*

#### B. Nucleotide

*Cxqdsx* atgggtttcgcaagataacctggatggagacgatgtcagaatcgggatacgaagccgcccgcgacggggccagcgggtgcattccagcagtaactcgtgaaccgggagcggcccccgaactgt  
*Cxpipdsx* atgggtttcgcaagataacctggatggagacgatgtcagaatcgggatacgaagccgcccgcgacggggccagcgggtgcattccagcagtaactcgtgaaccgggagcggcccccgaactgt  
 \*\*\*\*\*

*Cxqdsx* gcccgctgcgaaaccacgggctcaagattggcctgaaggacacaaaggcttactgcaagtatcgacgtgcaactgcagaaatgctgcctgaaggccggaacggcgagcggtcatggcc  
*Cxpipdsx* gcccgctgcgaaaccacgggctcaagattggcctgaaggacacaaaggcttactgcaagtatcgacgtgcaactgcgaaatgctgcctgaaggccggaacggcgagcggtcatggcc  
 \*\*\*\*\*

*Cxqdsx* ctgcagacggccctgcggcgggcccaaaactcagagacgacgaacgacctcaacgatggcggaagtggccccgaaccgggtacataaacattcacatacccaagctatccgaactgaaaagag  
*Cxpipdsx* ctgcagacggccctgcggcgggctcaaaactcagagacgacgaacgacctcaacgatggcggaagtggccccgaaccgggtacataaacattcacatacccaagctatccgaactgaaaagag  
 \*\*\*\*\*

*Cxqdsx* atgaaacataatttgatgcataattctcagcagcaacgcgtcgttgatcgactgcgattcgtcgaccggatcgatgaactccacaccgggcaacctcgtccatggcactaccactacatcga  
*Cxpipdsx* atgaaacataatttgatgcataattctcagcagcaacgcgtcgttgatcgactgcgattcgtcgaccggatcgatgaactccacaccgggcaacctcgtccatggcactaccactacatcga  
 \*\*\*\*\*

*Cxqdsx* agatcacccgaggggtccgggtacatcccgcgagcgagcgcaacatcttggaccgaatcgtccagcgtatctcccggaaccggcaacctgttacccagtcctccaaacatcagagtacatcac  
*Cxpipdsx* agatcacccgaggggtccgggtacatcccgcgagcgagcgcaacatcttggaccgaatcgtccagcgtatctcccggaaccggcaacctgttacccagtcctccaaacatcagagtacatcac  
 \*\*\*\*\*

*Cxqdsx* ggtccagattctcgatcagacgatgaactgggtgaaacgatctcagtatctgctggagaagctcaactaccctgggagatgatgcccttgatgtacgtgatactgaagggtgccgacggg  
*Cxpipdsx* ggtccagattctcgatcagacgatgaactgggtgaaacgatctcagtatctgctggagaagctcaactaccctgggagatgatgcccttgatgtacgtgatactgaagggtgccgacggg  
 \*\*\*\*\*

*Cxqdsx* gacgtccaaacggcgaccggcggtatcgacgaaggtcaagcgtcgtcaacgaatactcagcgttacacaactctgaacatgttcgacggggtggaggttcgcagagcccaacgacgacgtcc  
*Cxpipdsx* gacgtccaaacggcgaccggcggtatcgacgaaggtcaagcgtcgtcaacgaatactcagcgttacacaactctgaacatgttcgacggggtggaggttcgcagagcccaacgacgacgtcc  
 \*\*\*\*\*

*Cxqdsx* gga  
*Cxpipdsx* gga  
 \*\*\*

**Figure 3.S8** Amino acid (A) and nucleotide (B) alignment of *Cx. quinquefasciatus* and *Cx. pipiens* form *pipiens* male *doublesex* isoforms. Bolded text denotes male-specific portion of protein.

#### A. Peptide

```

CxqdsxM  MVSQDTWMTMESGYEGRPDGASGSSNSLNPRTPPNCARCNHGLKIGLGHKRYCKYRSCNCEKCLTAERQRMALQTAALRRQQTQDEORALNDGEVAPEPVHNIHIPKLSELKE
CxipdsxM  MVSQDTWMTMESGYEGRPDGASGSSNSLNPRTPPNCARCNHGLKIGLGHKRYCKYRSCNCEKCLTAERQRMALQTAALRRQQTQDEORALNDGEVAPEPVHNIHIPKLSELKE
*****

CxqdsxM  MKHNLHNSQQORSLLIDCDSSTGSMNSTPCTSSMALPLHRRSPTGPVHPGEAQLGANHASVSEPPANLLFPVPPNIRVHHGPDSSRSDDELVKRSQYLLEKLNYPWEMMPLMYVILKGADG
CxipdsxM  MKHNLHNSQQORSLLIDCDSSTGSMNSTPCTSSMALPLHRRSPTGPVHPGEAQLGANHASVSEPPANLLFPVPPNIRVHHGPDSSRSDDELVKRSQYLLEKLNYPWEMMPLMYVILKGADG
*****

CxqdsxM  DVQTAHRRIDEAQAVLLHSRIGRDDIDDENISVTGRTNSTLSRCSSTYRSRSRSPHPDEEGVLNLDTKSAKNAASDDSSAFNDVKPKQPSSEHQSRLEEAYQSSVEQHHSAKSXSKKH
CxipdsxM  DVQTAHRRIDEAQAV--LHRRIGRNDIDDENISVTGRTNSTLSRCSSTYRSRSRSPHPDEEGVLNLDTKSAKNAASDDSSAFNDVKPKQPSSEHQSRLEEAYQSSVEQHHSAKSXSKKH
*****

CxqdsxM  SVADDAEPVSQVAPHETNGFEKGLKLFNNTKASRRSTHKDDSVNESFAARDKPLSLFPRHLHLAENLELLKTPLSLPSAANFLPFTIPLTNMEAIRSYPOFFYPYQTHSGSDPPHPLIS
CxipdsxM  SVADDAEPVSQVAPHENMGFEKGLGLFKNNKANRRSTHKDDSAIESFAARDKPLSLFPRHLHLAENLELLKTPLSLPSAANFLPFSISPLTNMEAIRSYPOFFYPYQTHSGSDAPHPPLIS
*****

CxqdsxM  SPFMNTPHHLLFPDGYRKELPKFPCPTSPSRSTSPPKVGAQPFSGSRVEPVLPQNQPSVAPIH
CxipdsxM  SPFMNTPHHLLFPDGYRKELPKFPCPTSPSRSTSPPKVGAQPFAGSRGEPLVLPQNQPSVAPIH
*****

```

#### B. Nucleotide

```

CxqdsxM  atggtttcgcaagataacctggatggagacgatgtcagaatcgggatacgaagccggcgacggggccagcgggtgcataccagcagtaactcgtgaacccgcgagcgcgcccaaaatgt
CxipdsxM  atggtttcgcaagataacctggatggagacgatgtcagaatcgggatacgaagccggcgacggggccagcgggtgcctccagcagtaactcgtgaacccgcgagcgcgcccaaaatgt
*****

CxqdsxM  gcccgctgccgaataaccacggggtcaagattggcctgaaggacacacacagcgttactgcaagtatcgagctgcaactgcgagaaaatgctgcctgacggccgaacgagcgggtcatggcc
CxipdsxM  gcccgctgccgaataaccacggggtcaagattggcctgaaggacacacacagcgttactgcaagtatcgagctgcaactgcgagaaaatgctgcctgacggccgaacgagcgggtcatggcc
*****

CxqdsxM  ctgcagacggccctgcggcgggcccaactcaggacgagacacagcagcctcaacgatggcgaagtggccccgaacgggtacataaacattcacatacccaagctatccgaactgaaagag
CxipdsxM  ctgcagacggccctgcggcgggctcaactcaggacgagacacagcagcctcaacgatggcgaagtggccccgaacgggtacataaacattcacatacccaagctatccgaactgaaagag
*****

CxqdsxM  atgaacacataatttgatgcataattctcagcagcaacgcctgttgatgcactgcgattcgtgcaccggatcgatgaactccacacccggcacctcgtccatggcactaccactacatcga
CxipdsxM  atgaacacataatttgatgcataattctcagcagcaacgcctgttgatgcactgcgattcgtgcaccggatcgatgaactccacacccggcacctcgtccatggcactaccactacatcga
*****

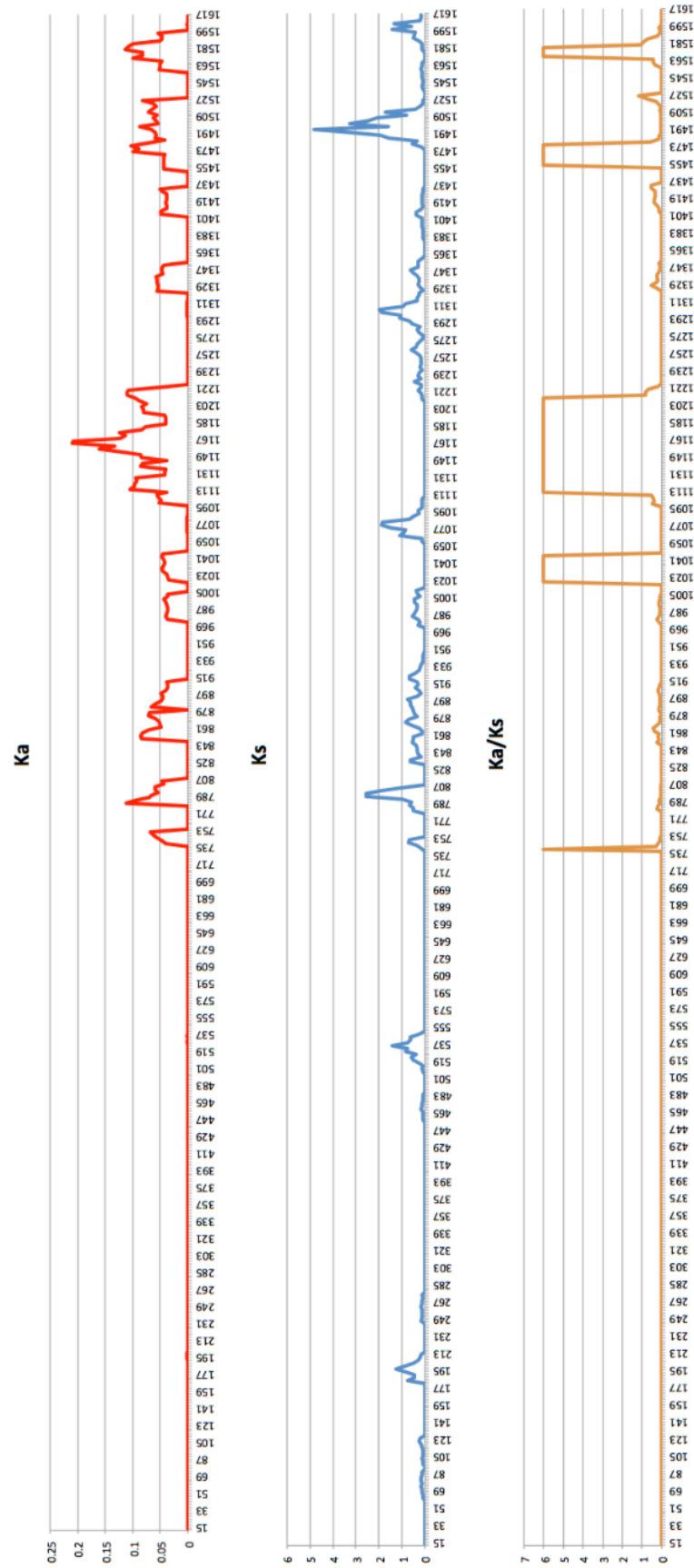
```







**Figure 3.8** Ka/Ks graph. Graphical representation of Ka (top), Ks (middle) and Ka/Ks (bottom) values calculated for male-specific *Cx. quinquefasciatus* / *Cx. pipiens* form *pipiens doublesex* pairwise codon alignment. Values are recalculated for each 30 nucleotide (10 amino acid) window, which slides 3nt (1 AA) at a time. Numbers on the x-axis denote the coordinate of the central nucleotide in the window. Ka/Ks values were truncated at a maximal value of six for display purposes. The common portion of the transcript ends and male-specific sequence begins with the window centered at nucleotide position 735.



## Chapter 4:

### ***Doublesex* is ubiquitous and ancestral in the insects**

#### **Abstract**

The *doublesex* (*dsx*) gene functions as a master switch at the base of the insect sex determination cascade. It is acted upon by diverse upstream genetic signals that result in its sex-specific splicing, and ultimately serves to trigger male or female somatic sexual differentiation. To the best of our knowledge, this gene has only been reported from seven current insect orders, and thus the phylogenetic distribution within the largest Arthropod sub-phylum, the Hexapoda, is unknown. In contrast to what has been shown in insects, other arthropods such as chelicerates and crustaceans express at least two copies of *dsx*, both of which lack alternative splicing. To understand the evolution of this integral gene relative to other arthropods we qualified the presence of *dsx* from all 32 hexapod orders using public EST and genome sequencing projects. We find the *dsx* gene to be ubiquitous, with likely *dsx*-encoding EST contigs recovered from 29 orders. Additionally, we recovered both alternatively spliced and putative multi-copy *dsx* transcripts from several orders of hexapods, including basal lineages, indicating the likely presence of these characteristics in the hexapod common ancestor. Furthermore, we discovered a large degree of length heterogeneity in *dsx* coding sequences, both within and among orders, which we argue likely result from lineage-specific sexual selective pressures inherent to each taxon. Our work serves as a valuable resource for understanding the evolution of sex determination in insects, and in future research aimed at developing genetic sexing strains of insect disease vectors and agricultural pests.

## Introduction

Somatic sexual differentiation is fundamentally conserved among metazoans. The development of distinct male and female phenotypes integral to sexual reproduction is directed by the *Doublesex*/Mab-3 Related Transcription factor (DMRT) family of zinc-finger proteins (Raymond et al. 1999, Kopp 2012). The DMRT gene family has been described from most animal genomes studied thus far, and likely has pre-eumetazoan origins (Wexler et al. 2014). Despite conservation of DMRT presence, the specific manner in which members of the family act upon the multigene cascade governing sexual development varies widely among animal phyla (Matsuda et al. 2007, Marshall Graves 2008, Matson et al. 2011). One of the most functionally characterized DMRT genes is insect *doublesex*, the gene that acts as the terminal “double-switch” in the insect sex determination cascade.

Perhaps best elucidated in the model fly *Drosophila melanogaster* (Baker and Wolfner 1988, Baker 1989), the insect sex-determination molecular cascade consists of a primary signal (e.g X:A ratio, W/Y chromosomes, male-determining loci; see Marin et al. (1998) that facilitates sex-specific splicing of intermediary factors such as *sexlethal*, *transformer* and unknown others (*sexlethal* is limited to *Drosophila* while *transformer* appears lost in several insect lineages [(Verhulst et al. 2010, Saccone et al. 2011)]). These factors then facilitate splicing of *doublesex* to male- and female-specific mRNAs that in turn encode sex-specific DSX<sup>M</sup> and DSX<sup>F</sup> peptides, respectively. The genes targeted by DSX transcription factors remain elusive, however recent work has shown that *D. melanogaster* male and female DSX binds thousands of targets from multiple tissues (Clough et al. 2014). Many of these targets overlap with those identified for

mouse DMRT1, attesting to the conserved functional nature of eukaryote sex determination. In addition to determining sexual fate, *dsx* is also known to influence the development of both behavioral and morphological secondary sex characteristics (Siwicki and Kravitz 2009, Kijimoto et al. 2012, Devi and Shyamala 2013) and therefore plays several key roles in determining the specific sexual characteristics of a species.

DSX peptides retain conserved domains essential for transcription factor activity and protein oligomerization: an atypical zinc-finger DNA-binding domain (herein OD1) found throughout the DMRT gene family that exhibits a characteristic C2H2C4 Cys-His configuration (Erdman and Burtis 1993, Zhu et al. 2000), and a dimerization domain (herein OD2) specific to *doublesex* (An et al. 1996). The OD2 domain is primarily responsible for enhancing dimerization strength and thus OD1-DNA recognition (Cho and Wensink 1998), a function achieved via ubiquitin-associated domain folds (Bayrer et al. 2005). Mutations blocking dimer formation have produced intersex individuals (Erdman et al. 1996), thus demonstrating the conformation is integral to peptide function.

*Dsx* transcripts have been recovered in the crustacean *Daphnia magna* (Kato et al. 2011), as well as the arachnid *Metaseiulus occidentalis* (Pomerantz et al. 2014), both of which are hexapod outgroups within the Arthropoda. In these organisms, *dsx* pre-mRNA shows no sign of alternative splicing and directs sexual fate via sex-biased expression in males. The extent of conservation of this critical gene throughout hexapods however remains unknown, as thus far *dsx* homologs have been recovered from genome and/or transcriptome data of only seven insect orders (see Geuverink and Beukeboom 2014). As the development of genetic sexing strains of insect vectors is a major focus of sterile insect technique (SIT, a method of population control for disease vectors and other pest

species [Gilles et al. 2014]), conserved genes that influence sexual development such as *dsx* are optimal targets for emerging molecular insecticides (Dafa'alla et al. 2010).

To understand the breadth and functional conservation of these genes throughout insects (as well as other Metazoa) we used public transcriptome datasets, including recent EST data generated for all 32 current hexapod orders (Misof et al. 2014) to qualify the presence of *dsx* in this diverse group. We report transcript fragments encoding domain motifs indicative of *dsx* from 30 orders. Additionally, we identified EST contigs encoding both an OD1 and downstream OD2 domain from 22 orders, both within the primitive entognathous members (i.e. Protura) and the derived pterygote insects, suggesting *doublesex* was present in the common ancestor of the Hexapoda.

## Materials and Methods

We queried the NCBI Transcriptome Shotgun Assembly (TSA), Whole-Genome Shotgun (WGS), and non-redundant (nr) databases via tBLASTn/BLASTp (e-val = 1.0) with the Pfam seed alignments for pfam00751 (DM DNA binding domain [OD1]) and pfam08828 (*Doublesex* dimerization domain [OD2]). Contigs with both domains present were extracted and translated in the proper reading frame as reported by BLAST. We then extracted and manually aligned the two protein domain motifs with their respective orthologs from other orders. These data compose the ‘high-confidence’ dataset, as the presence of both domains on a single EST is strong evidence for *doublesex* homology. We entered the remaining EST contigs encoding a singular OD1 or OD2 domain into the alignment in the same manner. To search for remote (or diverged) homologs, an alignment of OD2 domain sequences (OD1 maintained strong sequence similarity) from

the high-confidence dataset was used to construct a profile hidden Markov model (HMM) using HMMER v3.1b1 (Eddy 2011). This profile HMM was then used to query local copies of the above NCBI databases. We examined manually the resultant hits scoring below the inclusion threshold e-value of 0.01, and added candidate ESTs representing orders not recovered in the BLAST analysis to the dataset. Contigs encoding both an OD1 and OD2 domain were added to the high-confidence set.

We used several data sources outside of the NCBI TSA and nr databases as secondary sources of evidence for orders in which we failed to recover a high-confidence *dsx* EST. The *Lepismachilis y-signata* (Archaeognatha) Illumina-generated RNAseq data deposited in NCBI SRA accessions ERR424579, ERR392013, ERR392014, and ERR392008 were downloaded and assembled using the CLC Genomics Workbench de-novo assembler (CLC Bio, Aarhus, DK). A BLAST database was created from the resultant contigs and queried with the OD1 and OD2 consensus sequences. Both domains were recovered on assembled contigs 383881 and 336923, respectively (Fig. 4.S6). To ensure our inability to recover an OD2 domain from the Embioptera was not an assembly artifact (i.e that ‘singleton’ unassembled transcriptome reads may contain a domain hit) we retrieved and queried the *Haploembia palaui* and *Aposthonia japonica* short-read RNAseq libraries (NCBI accessions SRR921605 and SRR921566, respectively) of Misof et al. (2014) via TBLASTN and profile HMM as described previously.

The predicted proteome of the dampwood termite *Zootermopsis nevadensis* (Terrapon et al. 2014; <http://termitegenome.org>) was retrieved from <http://termitegenome.org> and queried via BLASTP with the OD1/2 peptide sequences

reported here for the roaches (Blattodea, the sister taxon to the termites (Misof et al. 2014)), and with profile HMM as described previously. As a final check, we searched the *Nasutitermes takasagoensis* raw sequencing reads generated by Hayashi et al. (2013) in NCBI SRA accession DRR013047 again using TBLASTN and profile HMM. This process was repeated using the genome contigs of the German cockroach *Blattella germanica* (Blattodea; Baylor College of Medicine Human Genome Sequencing Center [<https://www.hgsc.bcm.edu/arthropods/german-cockroach-genome-project>]) in an attempt to isolate a high-confidence genomic contig encoding both OD1 and OD2 domains. We queried the assembled genome scaffolds via TBLASTN with roach OD1/2 sequences identified within the NCBI TSA, and with the OD2 profile HMM.

To search for alternative splicing of *dsx* transcripts within the basal hexapods, we retrieved the sequencing reads from the NCBI SRA corresponding to species for which we had recovered an OD2 domain: Protura (SRR921562 [*Acerentomon sp. AD-2013*]), Collembola (SRR921564 [*Anurida maritime*], SRR921647 [*Tetrodontophora bielanensis*], SRR921641 [*Sminthurus viridis*], SRR921635 [*Pogonognathellus sp. AD-2013*]), Diplura (SRR921624 [*Occasjapyx japonicas*]), Archaeognatha (ERR424579, ERR392013, ERR392014, ERR392008 [*Lepismachilis y-signata*], SRR921617 [*Meinertellus cundinamarcensis*]) and Zygentoma (SRR921568 [*Atelura formicaria*], SRR921648 [*Thermobia domestica*], SRR921654 [*Tricholepidion gertschi*]). These reads were six-frame translated and pattern matched to the 3' end of the OD2 domain to discern divergent splice forms.

## Results and Discussion

### *Sequence and phylogenetic conservation of hexapod doublesex*

*Doublesex* EST contigs containing both an OD1 and OD2 domain were designated ‘high-confidence’ contigs and recovered from 22 hexapod orders (Figs. 4.1, 4.S1). We also recovered evidence for both domains existing as singleton (or separate) contigs in an additional 7 orders (Figs. 4.S2, 4.S3, 4.S4). The remaining three orders lacked one of the two domains: we failed to find evidence for the OD2 domain from the Embioptera and Isoptera, while OD1 appears absent from the Mantophasmatodea. In addition to the EST contigs deposited in NCBI, we denovo assembled contigs from short read RNAseq libraries (see Methods) for *Sipyloidea sipyilus* (Phasmatodea) and *Lepismachilis y-signata* (Archaeognatha) that contained high-confidence ESTs.

With the exception of four orders, all high-confidence *doublesex* ESTs recovered maintained a conserved Lysine amino acid residue at position 14 of the DM/OD1 domain alignment, most frequently with a Thr-Pro-Pro-Asn-like motif at positions 1-4 (Figs. 4.S1, 4.S2; herein referred to as a type-A OD1 motif). There were multiple instances, however, where we identified ESTs encoding an OD1 domain (often from species for which we had already identified a type-A OD1-encoding EST) with an Ile-Ser-Cys beginning at position 14 and an Arg-Thr-Pro-Lys-like motif at positions 1-4 (Fig. 4.S3; herein a type-B OD1 motif). As the type-B OD1 motif was not present in any high-confidence *dsx* transcript (members of the Zoraptera and Hemiptera contained proximal matches), we find type-B singletons to be poor evidence for the presence of insect *dsx* and may represent a convergent zinc-finger domain from an alternative DMRT-superfamily gene. Of note, the conserved lysine at position 21 responsible for



maintaining a salt bridge with the target DNA molecule (Zhang et al. 2006) was maintained in all putative OD1 domain sequences of A and B types.

We failed to recover conclusive evidence for the presence of *doublesex* from the termites (Isoptera) as only a single type-A OD1 motif was recovered, with no evidence for OD2. Two BLAST hits to a type-B-like OD1 motif (contigs Znev\_05388 and Znev\_16235) were recorded from the *Zootermopsis nevadensis* predicted proteome (Terrapon et al. 2014), however these peptides encode a CUE-like (coupling of ubiquitin to ER degradation) DMA domain in the C-terminus (cd14370, Pfam family PF03474) characteristic of the DMRT family but not of *doublesex*. No BLAST hits to OD2 were reported from *Z. nevadensis*. A profile HMM search using the model we created (see Methods) failed to identify an OD2 motif. As a final check, we searched the *Nasutitermes takasagoensis* raw sequencing reads generated by Hayashi et al. (2013) in NCBI SRA accession DRR013047 again using TBLASTN and profile HMM. This returned a single read (G5ZWOJF02FLJ2Z; Fig. 4.S3) encoding a type B-like OD1 motif, with no evidence for OD2.

Similarly, we did not recover an OD2 domain from webspinners (Embioptera) thus we scanned the *Haploembia palaui* and *Aposthonia japonica* short-read RNAseq libraries of Misof et al. (2014) via TBLASTN and profile HMM as described previously. No OD2 hits were recovered in either test from either library. The *H. palaui* EST contig that was found to encode a type-A OD1 (Fig. 4.S2) encodes a stop codon 423bp downstream of the domain, thus terminating the putative *dsx* peptide without evidence of a recognizable OD2 motif. Future work (and likely genomic sequence) will be required to determine the exact genome organization and splicing of *dsx* within the Isoptera and

Embioptera, and to confirm whether the dimerization domain has indeed been lost or instead has diverged significantly (beyond recognition) from the other insects.

Since we recovered only singleton OD1 and OD2 domains from the roaches (Blattodea), we chose to query the assembled genome of the German cockroach *Blattella germanica* in an attempt to isolate a genomic contig encoding both OD1 and OD2 domains that could be considered high-confidence. A single type-A OD1 motif was recovered on contig JPZV01136765, while a single OD2 domain was found on contig JPZV01136787 (Figs. 4.S2, 4.S3). Three additional type-B OD1 motifs were recovered, however no hits shared a common genomic contig and thus we were unable to link them.

Several taxa were found to possess modified zinc finger structures responsible for binding target DNA and modulating transcription. The canonical C2H2C4 Cys-His motif (Erdman and Burtis 1993) was notably modified in the high-confidence structures of the Orthoptera (C2H2CHC2) and Zoraptera (C2H3C3) (Fig. 4.S1). Additional cysteine and histidine residues were observed in members of multiple orders that may affect zinc ion coordination (e.g *Metallyticus splendidus*, *Mantis religiosa*, *Corydalus cornutus*, *Xenophysella greensladeae*), however further protein biochemistry will be required to qualify the exact effects on DNA-binding affinity.

#### *Gene copy number*

The nearest hexapod outgroup taxa with reported *doublesex* homologs are the water flea *Daphnia magna* (Arthropoda: Crustacea: Branchiopoda: Cladocera: Daphniidae) (Kato et al. 2011) and the predatory mite *Metaseiulus occidentalis* (Arthropoda: Chelicerata: Acari: Phytoseiidae) (Pomerantz et al. 2014). Both organisms

express two copies of *dsx* and lack sex-specific alternative splicing in favor of sex-biased expression in males. The possibility thus exists that the hexapod common ancestor may have encoded two or more *dsx* genes.

Our analyses recovered at least one high-confidence *dsx* transcript and an additional divergent OD2-encoding singleton (i.e. two putative divergent *dsx* transcripts) from eight hexapod orders (Protura, Zygentoma, Ephemeroptera, Zoraptera, Phasmatodea, Mantodea, Hymenoptera and Diptera; see Fig. 4.S1). In all three species of Zygentoma reported here, the OD2 singleton transcript held a greater degree of similarity to the “canonical” OD2 amino acid sequence observed, i.e. maintained a double Leucine in the 5’ end with a double Valine and conserved Glutamic acid in the 3’ direction. The second transcript, also in all three species, contained both OD1/2 domains (i.e. was a high-confidence *dsx* EST) yet diverged at several conserved amino acid positions and produced significantly higher domain e-values in homology searches against the profile HMM. Expression of the diverged transcript also appears to be much higher than that of the canonical form; the zygentoman libraries SRR921568, SRR921654 and SRR921648 contained 148, 114 and 164 respective reads that mapped to the OD2 domain of the diverged singleton EST, while 2, 2 and 5 reads from the same libraries mapped to the domain on the high-confidence EST (not shown). Interestingly, the diverged high-confidence transcript sequences of the Zygentoma exhibit a greater degree of amino acid similarity to the high-confidence ESTs identified for the Ephemeroptera than to their conspecific singleton reads encoding an OD2 domain (Fig. 4.S5). In turn, the singleton reads encoding an OD2 domain are more similar between the two orders than within,

indicating these transcripts represent an alternate *dsx* gene motif shared between the two orders.

The Zoraptera appear to maintain two copies of *dsx*, both of which were recovered as high-confidence ESTs from *Zorotypus caudelli* and *Zorotypus gurneyi* (Fig. 4.S1). One copy exhibits a C2H2C4 zinc finger motif (a hallmark of the DMRT family, see Introduction) while the other has lost a conserved cysteine to form a modified C2H2C3 finger and encodes a diverged OD2 domain. A third high-confidence EST was recovered from *Zorotypus caudelli* that contained a modification to the zinc finger as above, yet differed in amino acid sequence.

Although previous studies have reported *dsx* in single-copy from the derived orders of holometabolous insects (Scali et al. 2005, Shukla and Palli 2012), we find evidence for multiple copies in single members of the Hymenoptera (*Sphaerophthalma orestes*), Coleoptera (*Meloe violaceus*) and Diptera (*Belgica antarctica*). Further taxon sampling will be required to determine whether these represent individual cases of gene duplication, or retention of an ancestral state. Additionally, multiple members of the Coleoptera (*Gyrinus marinus*, *Meligethes aeneus* and *Onthophagus nigriventris*) were found to encode a second OD2 domain downstream of the first (Fig. 4.S1, 4.S4). The duplicate domain recovered from *Meligethes aeneus* overlaps a portion of the upstream OD2 and begins at the first amino acid located 3' of the splice donor involved in alternative splicing of other coleopteran *dsx* (Shukla and Palli 2012). The remaining examples from *G. marinus* and *O. nigriventris* encode the second domain 425nt and 30nt respectively downstream of the first, however both are in a different reading frame. In the case of *Meloe violaceus*, we recovered a high-confidence *dsx* EST and an additional

singleton OD2 contig that encodes the domain downstream of a stop codon. In the absence of sequencing or assembly artifacts, these data suggest that the duplicated OD2 domains are not translated, yet have undergone domain duplication and/or gene elongation at some point in the evolutionary history of beetles.

### *Alternative splicing*

Insect male and female-specific sexual development has been shown in multiple species to proceed via sex-specific splicing of *doublesex* pre-mRNA (Baker and Wolfner 1988, Ohbayashi et al. 2001, Scali et al. 2005, Salvemini et al. 2011, Shukla and Palli 2012). The closest insect outgroups for which a *dsx* homolog has been recovered are the branchiopod crustacean genus *Daphnia* and the chelicerate *Metaseiulus occidentalis*, both of which regulate sexual dimorphism via transcript abundance with no evidence of sex-specific alternative splicing (Kato et al. 2011, Pomerantz and Hoy 2015). To assess the extent of *dsx* sexually dimorphic splicing within the hexapods (currently reported only from within the holometabolous insects), we mapped sequencing reads from basal orders (Protura, Collembola, Diplura, Archaeognatha and Zygentoma) to the 3' end of the respective OD2 domain identified for each. We chose this location as it has been shown to harbor the sex-specific splice donor in the Diptera (Nagoshi and Baker 1990, Scali et al. 2005, Ruiz et al. 2007, Salvemini et al. 2011) and Coleoptera (Shukla and Palli 2012). We identified multiple reads from both the Archaeognatha (*Lepismachilis y-signata*) and Collembola (*Anurida maritima*) that support a divergent transcript isoform immediately adjacent to the Glutamic acid residue in the 3' OD2 mRNA. This site is identical to that from which alternative splicing occurs in higher insects (Fig. 4.2; see above references).

These results suggest that not only *dsx* but also its alternative splicing were present in the hexapod common ancestor. Further work (i.e. RT-PCR) will be required to confirm these alternate isoforms, and to qualify their sex-specific expression.

### *Evolution of doublesex*

We observed significant length heterogeneity between *doublesex* homologs. The length (in nucleotides) between the OD1 and OD2 domains varied widely both within and between orders (Fig. 4.3). The average distance among all taxa, excluding putative alternate isoforms, was  $356 \pm 243$ bp. The parasitic hymenopteran *Orussus abietinus* exhibited the largest *dsx* fragment recovered at 6,488bp with an OD1-to-OD2 distance of 1,724bp (more than 6-fold greater than the other hymenopteran *dsx* ESTs recovered). Whether this pattern is maintained throughout the Orussidae is of future interest, as a second member of the “Symphyta” (*Tenthredo koehleri*) had a much shorter inter-domain distance. The phasmid *Aretaon asperimus* had the smallest distance bridging the two domains at only 84bp. This extreme length variation attests to the rapid evolution of the gene throughout the insect tree of life. Although the function of *dsx* remains conserved, it has been shown to be under positive selection in multiple orders (Ruiz et al. 2007, Hughes 2011, Sobrinho and de Brito 2012) and implicated in a system of “runaway evolution” due to its developmental influence on secondary sex characteristics and the response of female preference to genetic drift (Hughes 2011). As the upstream regulators of *dsx* are diverse and differ even among closely related taxa (Scali et al. 2005, Salvemini et al. 2011), the variation in gene length witnessed here is likely a product of evolution under these lineage-specific constraints.

## Conclusions

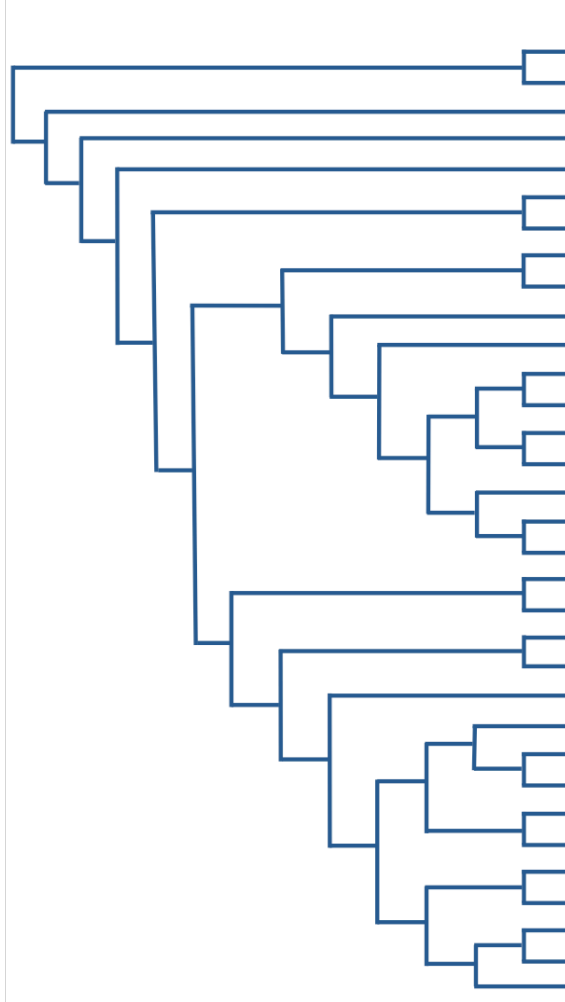
Our results show *doublesex* to be present throughout the Hexapoda, and likely present in the last common ancestor of the subphylum. The recovery of EST contigs that encode both the DNA-binding (DBD/OD1) and dimerization domains (OD2) from members of the primitive Entognatha to the derived orders of holometabolous insects attests to the conservation of this integral double-switch gene in the sex-determination cascade. Of note, we were unable to recover evidence for the OD2 domain in the embiopteran and isopteran datasets. Despite the use of alternative short-read libraries in addition to transcriptome sequence archives, the domain (if present) has diverged beyond our ability to isolate it via profile hidden Markov model and/or BLAST in these orders.

We present evidence that *dsx* likely existed as a multi-copy gene in the hexapod common ancestor, as at least two copies are transcribed in the basal Entognatha (among other orders), and also present in sequenced crustacean and chelicerate outgroups. Additionally, the phenomenon of sex-specific alternative splicing, currently undescribed outside of the insects, appears to occur in the most basal orders and is thus likely to have been present in the proto-hexapod. The advent of sex-specific alternative splicing coupled with mutation and eventual loss of *dsx* gene copy in the higher insect orders are perhaps concomitant, with neo-functionalization of the alternate isoform replacing the lost copy. Tandem duplications of the OD2 domain in the Coleoptera reported here, perhaps once functional, appear to have been inactivated via accrual of mutations. The extreme variation in coding sequence length observed between and within orders is evidence of the punctuated molecular response to sexual selection pressures influenced

by the phenotypes that *dsx* modulates. The transcripts identified here serve as initial candidates for studying pan-insect evolution, and for development of transgenic sexing techniques for insect pests.



**Figure 4.1** Summary of evidence for hexapod *doublesex* recovered in this study. Orders for which we identified a high-confidence *dsx* EST encoding both OD1 and OD2 domains are marked 'LINKED'; those for which only a Type-B OD1 domain was recovered (with Type-A absent) are annotated as such. Phylogeny as per Misof et al. (2014).



A phylogenetic tree of hexapod orders is shown on the left, with branches extending to the right towards a table. The tree is rooted at the top left and branches downwards and to the right. The orders are listed in the table to the right of the tree, with their corresponding OD1 and OD2 domain evidence.

Order	OD1	OD2
Protura	LINKED	LINKED
Collembola	YES	YES
Diplura	Type B only	YES
Archaeognatha	YES	YES
Zygentoma	LINKED	LINKED
Odonata	LINKED	LINKED
Ephemeroptera	LINKED	LINKED
Zoraptera	LINKED	LINKED
Dermaptera	LINKED	LINKED
Plecoptera	YES	YES
Orthoptera	LINKED	LINKED
Mantophasmatodea	NO	YES
Grylloblattodea	YES	YES
Embioptera	YES	NO
Phasmatodea	LINKED	LINKED
Mantodea	LINKED	LINKED
Blattodea	YES	YES
Isoptera	YES	NO
Thysanoptera	Type B only	YES
Hemiptera	LINKED	LINKED
Psocodea (Phthiraptera)	LINKED	LINKED
Psocodea (Psocoptera)	LINKED	LINKED
Hymenoptera	LINKED	LINKED
Raphidioptera	LINKED	LINKED
Megaloptera	LINKED	LINKED
Neuroptera	LINKED	LINKED
Strepsiptera	LINKED	LINKED
Coleoptera	LINKED	LINKED
Trichoptera	LINKED	LINKED
Lepidoptera	LINKED	LINKED
Siphonaptera	LINKED	LINKED
Mecoptera	LINKED	LINKED
Diptera	LINKED	LINKED







**Figure 4.S2** Type-A OD1-encoding singletons recovered in our analyses. Highly conserved amino acid positions are in red text. The conserved Lysine residue characteristic of the Type-A *dsx* DM domain (OD1, see Results and Discussion) is highlighted in yellow. E-values were calculated against our profile HMM using HMMer v3.1; asterisks indicate e-value was calculated via NCBI Conserved Domain Database search.

Order	Species	[Accession].[frame]/[range]	Sequence	HMM e-val
Protura	<i>Acerentomon</i> sp. AD-2013	GAXE01105806.1_2/58-86	RLPKCARCINHGKIPILKGHKRYCTNKYCDP-----	8.5e-13
Collembola	<i>Sminthurus viridis</i>	FCDOGKACXX:7:1301:17137:108047:3	TPPNCARCNHRLKIPILKGHKRYCTFRCTCKQKILTAERQVRVMAQ	2.0e-14*
Archaeognatha	<i>Lepismachilis y-signata</i>	Contig_383881 <sup>2</sup>	APPNCARCNHRLKVELKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	7.2e-28*
Zygentoma	<i>Thermobia domestica</i>	GASN01365969.1_4/81-127	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.3e-29
Odonata	<i>Calopteryx splendens</i>	GAXA01080693.1_4/36-82	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.4e-29
Ephemeroptera	<i>Isonychia bicolor</i>	GAXA01091320.1_6/79-119	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.9e-24
Ephemeroptera	<i>Baetis</i> sp. AD-2013	GATU01010243.1_2/88-133	SNRCAQCNHGKIPVGRHKRFCKFRICNQCCLLVKQRIVLH	1.2e-20
Ephemeroptera	<i>Eurylophella</i> sp. AD-2013	GAG01080309.1_5/8-52	ASRLCAFRCNHSIKIPVGRHKRFCKFRICNQCCLLVKQRIVLH	1.9e-20
Dermaptera	<i>Apachyus charceus</i>	GAUW01001274.1_3/14-59	KTPNCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	4.8e-28
Dermaptera	<i>Forficula auricularia</i>	GAYO01017359.1_5/107-152	KAPNCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	2.3e-27
Plecoptera	<i>Leuctra</i> sp. AD-2013	GAUF01079838.1_5/141-187	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	3.0e-27
Plecoptera	<i>Cosmioperla kuna</i>	GAYL01095698.1_6/28-74	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	3.4e-29
Grylloblattodea	<i>Grylloblatta bifratrilecta</i>	GAUF01142412.1_3/59-104	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	5.4e-29
Grylloblattodea	<i>Galloisiana yuasai</i>	GAWN01171825.1_4/47-93	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	2.0e-30
Embioptera	<i>Haploembia palaui</i>	GAZA01236824.1_5/84-130	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	3.3e-28
Mantodea	<i>Empusa pennata</i>	GAUT01309511.1_4/111-157	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	2.8e-30
Blattodea	<i>Periplaneta americana</i>	GANS01015027.1_4/1-44	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.2e-28
Blattodea	<i>Blattella germanica</i> genome	JPZV01136765.1_2/621-668	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.7e-15*
Isopoda	<i>Prohrioterme simplex</i>	GASE01075660.1_4/15-55	KIKYCSKSAHTKVOIKNKPKFDCCLCKDCIPILVRRDSTALA	7.6e-08
Hemiptera	<i>Xenophysella greensladeae</i>	GAYT01136691.1_3/24-70	TAPNCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.7e-27
Hemiptera	<i>Planococcus citri</i>	GAXF01116661.1_1/146-191	TAPNCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.2e-22
Hemiptera	<i>Planococcus citri</i>	GAXF01103858.1_1/43-88	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	4.9e-20
Hemiptera	<i>Acanthosoma haemorrhoidale</i>	GAUV01009588.1_3/70-108	KVKYCAKCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	7.1e-12
Psocodea (Psocoptera)	<i>Ectopsocus briggsi</i>	GAPT01011620.1_2/92-137	KIPYCGRCNHSIKIPVGRHKRFCKFRICNQCCLLVKQRIVLH	8.3e-20
Psocodea (Psocoptera)	<i>Ectopsocus briggsi</i>	GAPT01002163.1_6/155-201	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	4.3e-15
Neuroptera	<i>Pseudomallada prasinus</i>	GAUV01019538.1_2/83-128	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	2.2e-28
Neuroptera	<i>Conwentzia psociformis</i>	GAYH01007786.1_3/137-182	APPNCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	3.7e-22
Neuroptera	<i>Euroleon nostras</i>	GAXW01087396.1_4/25-71	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	7.1e-29
Coleoptera	<i>Lepicerus</i> sp. AD-2013	GABZ01138597.1_2/45-91	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	6.4e-28
Trichoptera	<i>Rhyacophila fasciata</i>	GAXX01046516.1_4/10-56	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	4.8e-31
Trichoptera	<i>Platycentropus radiatus</i>	GASO01092876.1_3/184-204	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	4.0e-08
Trichoptera	<i>Hydroptila</i> sp. AD-2013	GAUV01073298.1_1/112-158	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	3.2e-30
Lepidoptera	<i>Yponomeuta evonymellus</i>	GASG01093141.1_4/23-68	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.1e-19
Lepidoptera	<i>Polyommatus icarus</i>	GAST01056365.1_4/55-86	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	2.5e-29
Lepidoptera	<i>Dysierocrania subpurpurella</i>	GAUV01015364.1_1/256-302	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.9e-30
Lepidoptera	<i>Trodia sylvina</i>	GAYB01150705.1_2/60-106	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	6.6e-29
Siphonaptera	<i>Ctenocephalides felis</i>	GAYP01006643.1_1/27-73	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	1.5e-28
Mecoptera	<i>Panorpa vulgaris</i>	GAUH01062093.1_3/57-103	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	2.7e-25
Diptera	<i>Bibio marci</i>	GATJ01056020.1_6/139-184	TPPKCARCNHRLKIPILKGHKRYCKYKRYCNCCKRLTADRQVRVMAQ	



**Figure 4.S3** Type-B OD1-encoding singletons recovered in our analyses. Highly conserved amino acid positions are in red text. The conserved Isoleucine residue characteristic of the Type-B *dsx*-like OD1 domain is highlighted in yellow. E-values were calculated against our profile HMM using HMMer v3.1; asterisks indicate e-value was calculated via NCBI CDD search.

Order	Species	[Accession]_lframe/[range]	HMM aligned sequence
Archaeognatha	<i>Meinertellus cundinamarcensis</i>	GAUG01133643.1_3/44-81	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRLCRW <sup>R</sup> EC <sup>R</sup> CPN <sup>C</sup> CLLVVE-----
Blattodea	<i>Periplaneta americana</i>	GAWS01189376.1_5/2-47	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRLCRW <sup>R</sup> EC <sup>R</sup> CPN <sup>C</sup> QLVVERQ <sup>R</sup> VMAAQ
Blattodea	<i>Cryptocercus wright</i>	GAZN01162439.1_6/11-56	-RPKCARCRN <sup>H</sup> GMI <sup>H</sup> SLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Calanoida	<i>Calanus finmarchicus</i>	GAXR01154012.1_1/25-70	-RPKCARCRN <sup>H</sup> GMI <sup>H</sup> SLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Calanoida	<i>Calanus finmarchicus</i>	GAXR01154011.1_1/25-70	-RPKCARCRN <sup>H</sup> GMI <sup>H</sup> SLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Coleoptera	<i>Aleochara curtula</i>	GATW01017006.1_5/27-72	-VPKCARCRN <sup>H</sup> GMI <sup>H</sup> SLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Coleoptera	<i>Ips typographus</i>	GACR01003199.1_2/53-98	-VPKCARCRN <sup>H</sup> GMI <sup>H</sup> SLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Collembola	<i>Tetradontophora bielaniensis</i>	GAXI01024987.1_5/8-53	-VPKCARCRN <sup>H</sup> GMI <sup>H</sup> SLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Collembola	<i>Pogonognathellus sp. AD-2013</i>	GATD01092185.1_3/9-54	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Collembola	<i>Anurida maritima</i>	GAUR01052744.1_3/128-173	-APKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Diptera	<i>Culicoides sonorensis</i>	GAWN01017604.1_1/28-73	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Diptera	<i>Trichocera saltator</i>	GAXZ01130618.1_6/9-54	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Ephemeroptera	<i>Isonychia bicolor</i>	GAXA01106317.1_1/122-167	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Ephemeroptera	<i>Eurylophella sp. AD-2013</i>	GAZG01012200.1_6/32-77	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Grylloblattodea	<i>Galloisiana yuasai</i>	GAWN01063629.1_6/2-47	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hemiptera	<i>Graminella nigrifrons</i>	GAQX01025711.1_6/87-133	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hemiptera	<i>Cercopis vulnerata</i>	GAUN01102337.1_4/98-144	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hemiptera	<i>Macropsiphum euphorbiae</i>	GAOM01101235.1_1/7-52	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hemiptera	<i>Cercopis vulnerata</i>	GAUN01003127.1_1/48-53	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hemiptera	<i>Nilaparvata lugens</i>	GAYF01062266.1_2/4-49	-HPKCARCRN <sup>H</sup> GMI <sup>H</sup> SLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hemiptera	<i>Acanthosuarina muellerianae</i>	GAYT01137565.1_1/301-346	-TLPNCARCRN <sup>H</sup> GMI <sup>H</sup> SLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hemiptera	<i>Cercopis vulnerata</i>	GAUN01092972.1_3/100-120	-PPNCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hemiptera	<i>Ganaspis sp. G1</i>	GAIW01023601.1_3/89-135	-TPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Ganaspis sp. G1</i>	GAIW01012036.1_5/215-260	-STNCGYCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Leptopilina clavipes</i>	GAXY01053312.1_1/30-75	-RPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Stigmatomma oregonense</i>	GAXR01000833.1_2/51-96	-RPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Ganaspis sp. G1</i>	GAIW01012035.1_5/481-526	-STNCGYCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Cotesia vestalis</i>	GAUP01064990.1_1/26-71	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Cotesia vestalis</i>	GAUP01004774.1_1/35-80	-RPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Cotesia vestalis</i>	GAGK01004932.1_5/23-68	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Osmia cornuta</i>	GAGH01047348.1_5/369-414	-RPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Mischocyttarus flavitarsis</i>	GAXW01016837.1_6/60-105	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Sphaerophthalma orestes</i>	GAXP01040437.1_2/70-113	-RPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Stigmatomma oregonense</i>	GAXR01010442.1_4/19-64	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Orussus abietinus</i>	GAUW01091006.1_5/57-102	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Tenthredo koehleri</i>	GAWW01088411.1_2/215-260	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Chyphotes mellipes</i>	GAXL01043535.1_1/29-74	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Pseudomasaris vespoidea</i>	GAXQ01011632.1_5/152-197	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Pseudomasaris vespoidea</i>	GAXQ01011634.1_5/152-197	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Pseudomasaris vespoidea</i>	GAXQ01011633.1_6/152-197	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Pseudomasaris vespoidea</i>	GAXQ01011635.1_6/152-197	-SPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ
Hymenoptera	<i>Cotesia vestalis</i>	GAGK01010867.1_5/96-130	-RPKCARCRN <sup>H</sup> GVISCLGK <sup>H</sup> KRSQY <sup>R</sup> DFKSCVCAKCN <sup>L</sup> IAERQ <sup>R</sup> VMAAQ

Figure 4.S3 Continued.

Hymenoptera	<i>Cotesia vestalis</i>	GAKG01010867.1_5/96-130	-RPKCARCRNEGLISWLRGHKRCRYKECLCPKCSL-----
Isoptera	<i>Mastoterme darwiniensis</i>	GAZE01022571.1_3/6-51	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Isoptera	<i>Prorehinotermes simplex</i>	GASE01241946.1_5/28-73	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Isoptera	<i>Nasutitermes takasagoensis</i>	G5ZW0JF02FLJ72.4/133-171	TPPKCARCRIBKELTFPVKHKRDCPHRECHCRCPQYVAKGMYRALQ
Lepidoptera	<i>Spodoptera exigua</i>	GARL01010958.1_6/32-77	-VPKCARCRNEGLISLKGHKCAAYRLCQCPKCGLIKERQVRVMAAQ
Lepidoptera	<i>Polyommatus icarus</i>	GAST01022246.1_6/28-73	-TPKCARCRNEGVSCLKGHKRLCRWDCRCPSCLLVLERQVRVMAAQ
Lepidoptera	<i>Aethis lepigone</i>	GARB01040060.1_3/51-96	-VPKCARCRNEGLISLKGHKCAAYRLCQCPKCGLIKERQVRVMAAQ
Lepidoptera	<i>Yponomeuta evonymellus</i>	GASG01107059.1_5/89-129	-TPKCARCRNEGVSCLKGHKRLCRWDCRCPGCLLVLERQVRVMAAQ
Mantodea	<i>Metalliticus splendendus</i>	GATB01281539.1_1/31-76	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Mecoptera	<i>Nannochorista philpotti</i>	GADB01005720.1_1/40-85	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Mecoptera	<i>Nannochorista philpotti</i>	GADB01001122.1_6/166-211	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Mecoptera	<i>Panorpa vulgaris</i>	GAUH01056862.1_5/41-86	-VPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Neuroptera	<i>Conwentzia psociformis</i>	GAYH01081098.1_6/60-105	-VPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Neuroptera	<i>Osmylus fulvicephalus</i>	GAYC01008034.1_6/33-78	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Orthoptera	<i>Tetrix subulata</i>	GASQ01096846.1_1/5-50	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Orthoptera	<i>Ceuthophilus sp. AD-2013</i>	GAUX01271954.1_3/33-78	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Orthoptera	<i>Prosarthria teretirostris</i>	GAZT01143051.1_3/24-69	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Orthoptera	<i>Gryllopalpa sp. AD-2013</i>	GAWZ01138841.1_4/2-47	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Phasmatodea	<i>Timema cristinae</i>	GAUX01115626.1_6/41-86	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Phthiraptera	<i>Menopon gallinae</i>	GAWR01095172.1_5/146-191	-KPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Phthiraptera	<i>Menopon gallinae</i>	GAUF01093074.1_6/87-132	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Plecoptera	<i>Leuctra sp. AD-2013</i>	GAUF01064618.1_6/8-53	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Psocodea (Psocoptera)	<i>Liposcelis bostrychophila</i>	GAYV01083169.1_5/23-68	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Raphidioptera	<i>Inocellia crassicornis</i>	GAZH01032725.1_5/40-78	-IPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Siphonaptera	<i>Ceratophyllus gallinae</i>	GAWK01044190.1_6/53-98	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Siphonostomatoida	<i>Caligus rogercresseyi</i>	GAZX01023775.1_4/32-77	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Thysanoptera	<i>Trips palmi</i>	GAXC01041671.1_5/35-80	-EPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Trichoptera	<i>Rhyacophila fasciata</i>	GAXX01069925.1_6/105-138	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Zygentoma	<i>Occasjapex japonicas</i>	GAXJ01053929.1_3/25-69	--PKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Zygentoma	<i>Tricholepidion gertschi</i>	GASO01237223.1_5/12-57	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Zygentoma	<i>Occasjapex japonicas</i>	GAXJ01108043.1_3/66-111	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ
Zoraptera	<i>Zorotypus caudelli</i>	GAYA0137377.1_2/34-79	-TPKCARCRNEGVSCLKGHKRLCRWRECQPCNCQLVVERQVRVMAAQ



**Figure 4.S4** Dimerization (OD2) domain-encoding singletons recovered in our analyses. Highly conserved amino acid positions are in red text. E-values were calculated against our profile HMM using HMMer v3.1; asterisks indicate e-value was calculated via NCBI Conserved Domain Database search.

Order	Species	NCBI accession	Sequence	HMM e-val
Protura	<i>Acerentomon</i> sp. AD2013	GAXE01016537_1	AATLTFQRLHLLNKKFLPVESLPAIYAVLKDAKGV <b>EASER</b> MAAYEEELHEISVRLGLL	9.8e-14
Collembola	<i>Anurida maritima</i>	GAUE01011497_6	MSVDGKMKSHALLEFRFLESPLVYRILKDYRADPFL <b>AS</b> SKMEAQTELSALAEAR	2.7e-17
Collembola	<i>Pogonognathellus</i> sp. AD-2013	GATD01095902_2	GAELPDGKMKESLOTLMDPEPPLVLYLILKDSRADPFL <b>ASTR</b> ILEAQELRSALAEAR	7.8e-11
Collembola	<i>Ochesella cincta</i>	GAMM01004046_1	TAIDEMKESVKALMEFRLPSETDPLIYRLKDYRGDF <b>QA</b> SAKIFEAQCELSWELLREAA	8.6e-12
Diplura	<i>Oceasjapyx japonicus</i>	GAXJ01108506_1	DNVEYVQAKVILKIFLHPDESCHLLWVILKADGDL <b>AS</b> SRIMEAQSELSMAFDVTR	1.7e-15
Archaeognatha	<i>Meinertellus cundinamarcensis</i>	GAUG01186505_6	ENVENKESVYGLMDRFRLPSECLPLVYLKADAFD <b>KEA</b> NRIVDAQDLHAYTLREAQY	1.6e-20
Archaeognatha	<i>Lepismachilis y-signata</i>	Contig_336923_1	VSLETIVENCNKLEEFHYSWEMPLVILVYAGSD <b>LEAS</b> RKIDEGQWRSQTCCAA*	5.6e-27*
zygentoma	<i>Tricholepidion gertschi</i>	GASO01035827_5	EEGDVWMSLHALLDFQLPLETLPLIYVVLKADRS <b>DVKEA</b> NRIMEAQSELSIF-----	3.3e-23
zygentoma	<i>Thermobia domestica</i>	GASN01345779_5	ENVELLKDSIHALLDFRLPLETLPLIYVVLKADRS <b>DVKEA</b> NRIMEGYQCMNGIIPAG	2.4e-22
Odonata	<i>Calopteryx splendens</i>	GAYM01004193_6	DNMFKDAILALLQWRLPVETPLPLIYAILOGAR <b>DVKEA</b> NRIONAQEHLRAMALRMYP	1.1e-19
Ephemeroptera	<i>Isonychia bicolor</i>	GAXA01058460_2	DNIDSRLDSIQTLDDFLRPMETPLVYVVLKVSRS <b>DVQEA</b> NRISEGCIYHMTSDVPVAAA	1.2e-20
Plecoptera	<i>Leuctra</i> sp. AD2013	GAUF01067748_5	DGHEVPFESIRKLLERFLPSEAPQLLAILKNSDY <b>DS</b> EAASKQIEIGYNSHPQGAISV	2.0e-14
Plecoptera	<i>Perla marginata</i>	GATV01133007_4	FYDPVHLGSIATKIDMCHLSPETQPLILAILKLS <b>FDVKEA</b> YKQIMAGGYFNKNTSN---	3.6e-10
Manoptasmatodea	<i>Tanzaniophasma</i> sp. AD2013	GAXB01144312_2	ENMDGLREAIHTLLKIFQLPLETLPLIYVVLKADRS <b>DVFEA</b> YNRIRNEARNELETARREAC	6.6e-20
Grylloblattodea	<i>Galloisiana yuasai</i>	GAWN01150275_2	DHGKLWSESIOALRENFHLSDQSVPLIFLVAF <b>SHFVNEA</b> SIRIKEGFMVIEIYLVGTIV	2.3e-15
Phasmatodea	<i>Areton asperimus</i>	GAWC01030481_5	VSVEAMMDGVYTLHMFHYHVEMLPLLVLLSDAH <b>CDVSEA</b> YNRILQVLGHLSVLPVTSS	4.3e-18
Phasmatodea	<i>Ramulus artemis</i>	GAWC01101272_2	VSVEAMMDGVYTLHMFHYHVEMLPLLVLLSDAH <b>CDVSEA</b> YNRILQVECAAA-----	1.5e-18
Blattodea	<i>Blaberus atropos</i>	GAYD01026473_5	ESLEVTQSHVYLMNTFRPLEALPLVYVVLQ <b>LSHS</b> DVSEAFARIKGSYST-----	6.0e-16
Blattodea	<i>Periplaneta americana</i>	GAWS01246226_6	ENQEVPLQSQMLMDFRMPMEALPLIYVVLQ <b>LSHS</b> DVSNVAYNRILQAEQSMALREAR	1.6e-15
Blattodea	<i>Blattella germanica</i> genome	JPZV01136787_1	ENQEVPLQSQMLMDFRMPMEALPLIYVVLQ <b>LSHS</b> DVSNVAYNRILQAEQSMALREAR	6.3e-10*
Thysanoptera	<i>Frankliniella occidentalis</i>	GAXD01013559_1	ENGRVHRESIORLVEKNFPFVALPLIYVVL <b>KDHGS</b> DVEVKN-----	1.1e-12
Thysanoptera	<i>Frankliniella cephalica</i>	GAYE01005232_5	ENGRVHRESIORLVEKNFPFVALPLIYVVL <b>KDHGS</b> DVEVKNLLEASSELDETSYNOERR	6.3e-16
Thysanoptera	<i>Thrips palmi</i>	GAXC01002564_6	ENGRVHRESIORLVEKNFPFVALPLIYVVL <b>KDHGS</b> DVEVKNLLEASSELDETSYNOERR	1.4e-14
Thysanoptera	<i>Gynaikothrips ficorum</i>	GAXG01131335_3	-NERVPWQSTRLLOQNFENALPLIYVVLAD <b>HKAD</b> EVKVALLEDGEF*-----	3.1e-07
Megaloptera	<i>Sialis lutaria</i>	GABR01005713_2	HSLDVLEKTRLLDDFNYPWEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	8.4e-25
Megaloptera	<i>Corydalinae</i> sp. KMRSPM-2012	GADH01041729_2	-NIELLEDCSKLLDFTLYPWEMLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	2.1e-24
Strepsiptera	<i>Mengenilla moldrzyki</i>	JP085896_4	NETERILDCSKLLDFTLYPWEMLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	5.7e-22*
Strepsiptera	<i>Stylops mellittae</i>	GAZM01003845_3	-----FXOKLMEYKYPIYEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	3.0e-16
Coleoptera	<i>Meloe violaceus</i>	CATA01002860_6	RGAEINLEFCORLKDFQLSWKMSI <b>SLVD</b> VLKYAK <b>DQEA</b> WRQIDEAFLEIRALAAVEAR	2.3e-08
Coleoptera	<i>Dendroctonus ponderosae</i>	GATF01002861_6	RGAEINLEFCORLKDFQLSWKMSI <b>SLVD</b> VLKYAK <b>DQEA</b> WRQIDEAFLEIRALAAVEAR	2.4e-08
Coleoptera	<i>Onthophagus nigriiventris</i>	GAFX01005311_3	DRSMDLLEDCSKLLDFTLYPWEMLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	1.9e-25
Coleoptera	<i>Onthophagus nigriiventris</i>	GAQW01004722_2	IGOKDLIOESLQLEKERYSWEMPLIYVY <b>AD</b> TPD- <b>LEEAS</b> RKIDEGKDAEQVDFLIK*	3.1e-14
Trichoptera	<i>Hydroptilia</i> sp. AD2013	GAQM01004722_1	*RR*TTSGFNFNKTDRFHLUSKMSI <b>SLIH</b> VLKAK <b>DQ</b> QKAFQIDEAFLEVQALAKYPTP	5.9e-05
Trichoptera	<i>Platycentropus radiatus</i>	GAVM01009688_4	ALMSEILENCYTLLEFNFFEMPLIYAILK <b>ATH</b> ID <b>EA</b> SRIDEQOTFYLYVDVMTAA	1.3e-19
Lepidoptera	<i>Dysericocrania subpurpurella</i>	GASS01085500_2	ASLETILENCCKLLEKHYSWEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	9.9e-29
Lepidoptera	<i>Dysericocrania subpurpurella</i>	GASY01001893_6	TSLETILENCCKLLEKHYSWEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	2.2e-16
Lepidoptera	<i>Polyommatus icarus</i>	GAST0101894_6	TSLETILENCCKLLEKHYSWEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	7.1e-18
Lepidoptera	<i>Yponomeuta evonymellus</i>	GAST01019581_4	VSLETILENCCKLLEKHYSWEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	3.4e-19
Lepidoptera	<i>Empusa pennata</i>	GASG01061525_3	-----LVENCNCKLLEKHYSWEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	1.4e-22
Siphonaptera	<i>Ctenocephalides felis</i>	GAWT01314774_6	ENTAVGRESILELLDRLFPWETPLPLIYVVL <b>ODAR</b> CDV <b>NEA</b> SRILKCYAKANDKVIYNAI	2.1e-17
Mecoptera	<i>Bittacus pilicornis</i>	GATH01008256_1	EYKVALLESQQIQLEKYPFFEMPLIYAIL <b>GLSG</b> - <b>ISEA</b> KNINQOLVYSEYSRMNLN	1.6e-15
Mecoptera	<i>Boreus hyemalis</i>	GAYK01002938_3	SEVNLLEDCCKLLEKHYSWEMPLIYAIL <b>GLSG</b> - <b>ISEA</b> SRIDEGLVYVERLEIRIM	1.1e-20
Mecoptera	<i>Boreus hyemalis</i>	GAYK01015512_3	YSKRVVFDLSQKLEKHYSWEMPLIYAIL <b>GLSG</b> - <b>ISEA</b> SRIDEGLVYVERLEIRIM	3.6e-10
Diptera	<i>Bombus major</i>	GATY01017535_5	YFDKRVFDLSQKLEKHYSWEMPLIYAIL <b>GLSG</b> - <b>ISEA</b> SRIDEGLVYVERLEIRIM	7.4e-10
Diptera	<i>Bombus major</i>	GATY01017535_5	YFDKRVFDLSQKLEKHYSWEMPLIYAIL <b>GLSG</b> - <b>ISEA</b> SRIDEGLVYVERLEIRIM	2.0e-20
Diptera	<i>Bibio marci</i>	GATJ01028801_1	-----VDVLEKSTKLELFOYPWEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	1.8e-19
Diptera	<i>Bibio marci</i>	GATJ01028801_1	-----VDVLEKSTKLELFOYPWEMPLMVLVYAG <b>GN</b> IDEAARSIREGKRVVNEYSRIHNL	3.7e-11

**Figure 4.S5** Alignment of both putative *doublesex* transcripts from each of the Zygentoma and Ephemeroptera species reported in this study. The high-confidence ESTs (black text) share similarity with each other, while the alternate motif (red text) segregates.

<i>Atelura formicaria</i>	GAYJ01004010*	ARERMLYESLVVMRQAFPVGEEAMPILLICILKHRS--VQEASLKIYQGHYDLNAKGFVGLDG	Zygentoma
<i>Thermobia domestica</i>	GASN01031735*	ARERMLYESLVVMRQAFPVGEEAMALLICILKHRS--VQEASWKIYQGTYDLTSRFGVGLDN	Zygentoma
<i>Tricholepidion gertschi</i>	GASO01256612*	EREQMLYESLVVMRQAFPVGEEAMPILLICILKHRS--VQEASWKICQGTDRDLARGIIGIEN	Zygentoma
<i>Ephemera danica</i>	GAUK01006643*	TRERMLYESLMQLRQLYPVPEASLPILLICILKSSRS--IDEASAKIVQGTTHLSTRGFPGIEA	Ephemeroptera
<i>Isonychia bicolor</i>	GAXA01111429*	TRERMLYESLLELRQSPVPEAGLPILLICILKSSRS--VHEASAKIVQGTTHLSTGSLGLDG	Ephemeroptera
<i>Baetis sp. AD-2013</i>	GATU01002143*	QRKNLMSLLELRSHHPVPPDAALPILLICILNISNS--VKEASEKIQGLQELSLAASNEL	Ephemeroptera
<i>Atelura formicaria</i>	FCDOKPIACXX_7_2202_17946_4233_6 <sup>1</sup>	-----KESLQTLDDMERLPMETPLPIYVVVLKDARSDVKEASNRIIEGGYWIMTQMAA-----	Zygentoma
<i>Thermobia domestica</i>	GASN01345779	ENVELLKDSLHALLDMEFRLPLETLPPIYVVVLKDARSDVKEASNRIIEGGYQCMNQGIIPIG	Zygentoma
<i>Tricholepidion gertschi</i>	GASO01035827	EEGDMMESLHALLDMEFQLPLETLPPIYVVVLKDARSDVKEASNRIIEAQSELSIF-----	Zygentoma
<i>Isonychia bicolor</i>	GAXA01058460	DNIDSLRDSIQTLDDTFLRPMETPLPIYVVVLKVSRSVDVQEA FNRIIEGGYHMTSSDVPVAAA	Ephemeroptera

\* - high-confidence EST contig

<sup>1</sup> - transcriptome sequencing read obtained from SRA accession SRR921568



**Figure 4.2** Alternative splicing of *doublesex* transcripts from *A. Anurida maritime* (Collembola) and *B. Lepismachilis y-signata* (Archaeognatha). The consensus EST contig from each species (top, bold text), with short-read data mapped below illustrates the diverging transcript isoforms. The common 5' OD2 sequence in each species is in black text, while the diverging 3' ends are in red and blue text.

**A. Collembola**

***Anurida maritime* GAUE01011497**  
FCD0KMHACXX\_8\_2204\_18193\_73442\_5  
FCD0KMHACXX\_8\_1303\_19208\_5498\_1  
FCD0KMHACXX\_8\_1305\_9211\_60447\_1  
FCD0KMHACXX\_8\_1104\_6890\_161026\_3  
FCD0KMHACXX\_8\_2104\_12102\_172846\_1  
FCD0KMHACXX\_8\_2104\_6210\_40876\_2

MSVDGMKESLHALLMFPPLESILPVYRIILKDYRADFRFLASSKIME**AQTELRSLALLEAAAR**  
ILKDYRADFRFLASSKIME**AQTELRSLALLEAAARVASYPTNYISYPQG**  
LLEMFRPPLESILPVYRIILKDYRADFRFLASSKIME**AQTELRSLALLEAAAR**  
ESLHALLMFPPLESILPVYRIILKDYRADFRFLASSKIME**AQTELRSLAL**  
VDGMKESLHALLMFPPLESILPVYRIILKDYRADFRFLASSKIME**AQTE**  
PLVYRIILKDYRADFRFLASSKIMEGKSFKKKGKTCKM\*YHKF\*IGFLHFD  
SKIMEGKSFKKKGKTCKM\*YHKF\*IGFLHFDKAQTELRSLALLEGARVA

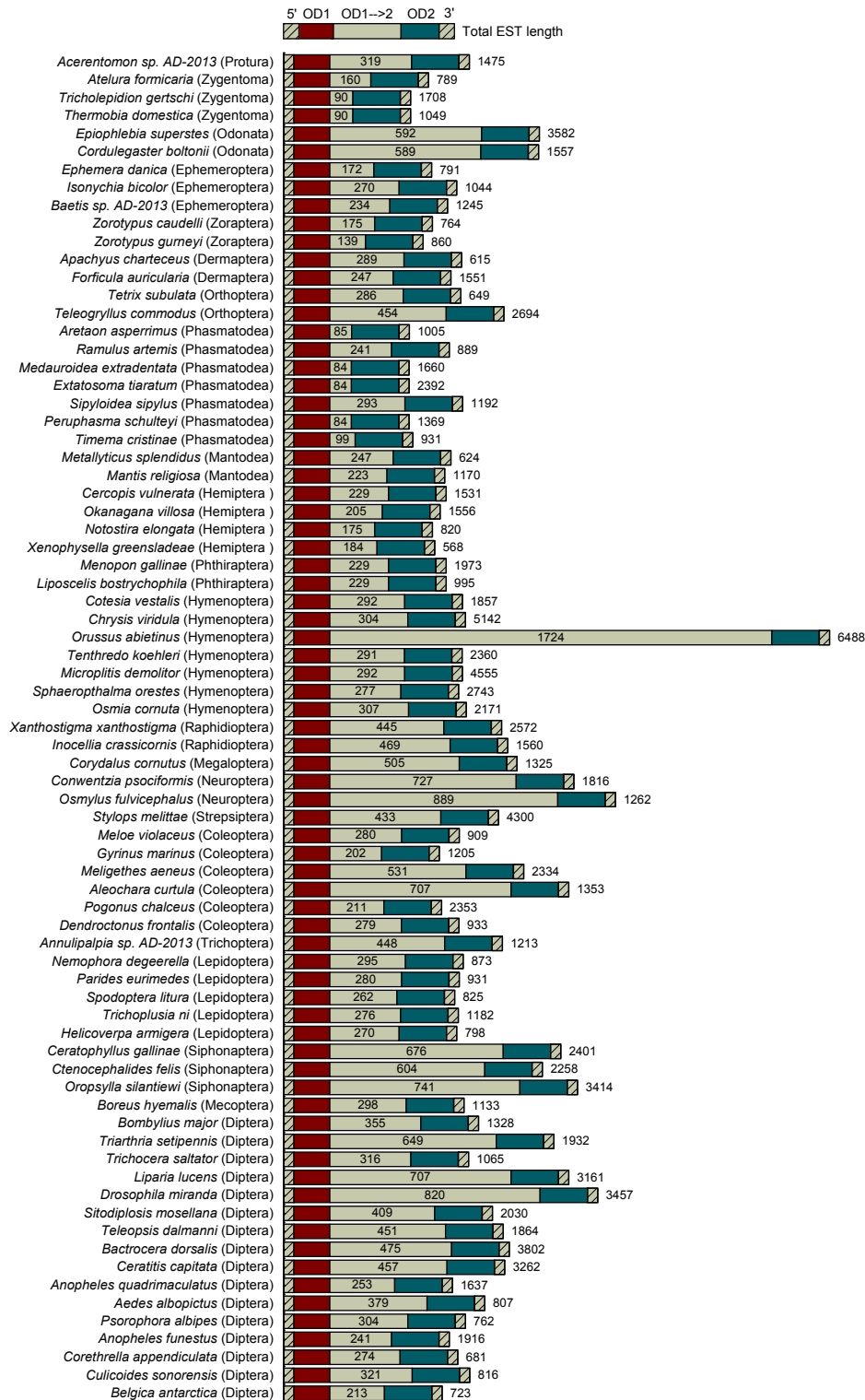
**B. Archaeognatha**

***Lepismachilis y-signata* Contig\_336923**  
HWI-ST863\_254\_D235LACXX\_7\_1310\_6553\_41213\_5  
HWI-ST863\_254\_D235LACXX\_7\_2204\_2322\_41551\_5  
HWI-ST863\_254\_D235LACXX\_7\_2213\_16217\_84596\_4  
HWI-ST863\_254\_D235LACXX\_7\_1216\_13158\_72514\_4  
HWI-ST863\_254\_D235LACXX\_7\_1304\_2921\_53027\_4

VSLETLVENCNKLEKHFYSWEMMPLVLVILNYAGSDLDASRKIDEGQWRSQTFCCAAN\*-  
ILNYAGSDLDASRKIDEGQWRSQTFCCAAN\*E  
ILNYAGSDLDASRKIDEGQWRSQTFCCAAN\*E  
LVIILNYAGSDLDASRKIDEGKMMVDEYARKHN  
LVIILNYAGSDLDASRKIDEGKMMVDEYARKHN  
SDLDEASRKIDEGKMMVDEYARKHNLNIFAGLE

Collembola reads from SRR921564  
Archaeognatha reads from ERR392013/ERR392014

**Figure 4.3** Distance in nucleotides (grey) between the OD1 (red) and OD2 domains (blue), for each high-confidence *doublesex* transcript reported in this study. Total EST length (may include untranslated regions) is to the right of each bar.



**Figure 4.S6** De-novo assembled contigs referenced in this study.

>Sipyloidea\_sipylus\_contig22122

CTGTCCCGCGGAAGAAGGGGAGGGCCCCGGCTTGGGGCTTGCTTGCTGCCA  
CTTACTTGCACTCGCGGGCGGGCAGTGAGTGCCGCGCGAACAAGAGTGCGTCA  
CAAAACACCCGGTTAACCTCAAATCAACAAAACACAGCTTTAGCCGCGCGCG  
CATGTTCGGACACCGACACCAGCACGGACGCGACAGCCGCCGCCGCCGCC  
ACGTGCTCGTCATCGTCGGCGGGCGGTGGCCGCCAGCAGCTCGCAGAACCCGC  
GCACGCCGCCGAACCTGCGCTCGTTGCCGCAACCACCGTCTCAAGATCGCCTT  
GAAAGGCCACAAGAGATACTGCAAATACCGCTACTGCACTTGCAAGTGC  
CGGCTGACGGCGGAGCGACAACGCGTCATGGCGCTGCAGACGGCGTTGCGG  
CGGGCACAAGCACAGGACGAGAGGTACCTGGCGCAGCAGGTGGCAGCCGTC  
ACCGCCTCCGGCGTATCACCTCCACCGCCGCCGTTATCCTTGCACTGCTGCTGC  
AGGTACGTGCTCGCCGCCGCCGCTGCAGCCTACGTCGCTCGCCCCCGCTCGCT  
CGGCCGACGGCAGCTGTGACTCGTCGTCATCGTCTCCGTGCTCGGCTAGGGT  
AATCTCCGTGCCGGCGCCTCGTAAAGCGATGGCACCCGTGCACCCACCTGAC  
ATCGCCGCCTCCACGGGGATGAGCGTGGATGCCATGATGGACGGAGTTTACA  
CTTTATTGCACATGTTCCACTATCATGTGGAGATGCTGCCCTTACTATTGGTC  
GTACTCAGTGACGCACACTGTGACGTCACAGAAGCCTATAACCGCATCCTGC  
AGGAAACCCGGCCTCGGAAATTCGGGAATGCTAAGCTATGGTTGTCAGTATC  
ATTCGTACGTGTTTGCTGAAGTAGATATGGCGACTCCAAACTGCTGATGAAG  
ACTGTCATCGGCTGGGAACAACCTCAACTACGGACCAGATTGGCGCAGCTTCT  
AGAATGCCAAGGAAGAGGTGATAGCGATGGGGCGGCGGGAGGCGGCGCGAT  
TGATGCAGTACCATCCGGCGGGCGGCGGCGACCTACGGTGGCTGCTACGGATC  
CGCAGCGCCCGCATACCTACCGGGGATGACCGCTAACGGGGCAATCTACCCC  
ACGCCGCCTCCAAGTCTCATGTTCTCACCTCCTAGCGCCGCCGCC

>Lepismachilis\_y-signata\_contig\_383881

CGACCCCGGCGCCTCCAGCTCAAGCTCGGCCGTGCCGCGTGCGCCGCCAAC  
TGCGCGCGCTGCCGCAACCACCGGCTCAAGGTCGAGCTGAAGGGTCACAAGC  
GCTACTGCAAGTACCGCTACTGCAACTGCGAGAAGTGCCGGCTGACGGCGGA  
CCGGCAGCGTGTGATGGCGCTGCAGACGGCGCTGCGGCGGGCGCAGGCGCA  
GGACGAGGCGCGCGCGCAACAACGGCCATCCGCCGCCGGGCGTGGAGCT  
GGAATGCCCCGGAGCCGCCGGTGGTGAAGGCGCCGCGCAGCCCGGTGATCCC  
GCCGCGGTCCGTGGGCTCTACCAGCGGCGAGTCGGTGCCGGGGTCGCCGGTG  
GTGTCTCCGTACGCGGCGCCGCCGCCCTCTGCGCCGCCGCCAACCATGCCGC  
CGCTTCTCCCGCCACAACAGCCTGT

>Lepismachilis\_y-signata\_contig\_336923

CCTTCATCTATCTTCCGTGACGCCTCATCCAGGTCGCTCCCCGCATAGTTGAG  
GATCACCAGGACTAAGGGCATCATCTCCAGGAGTAGTGGAACCTTCTCCAGG  
AGTTTGTTCAGTTTTCAACCAGCGTCTCGAGGGATACAGGCTGTTGTGGCGG  
GGGAAGCGGGGGCACGGGTGGGGGGGGGGCAGAGGGCGGGGGGGCCGCGG  
ACGGGGGCACCAACCCGCG

### Literature Cited

- Abascal, F., R. Zardoya, and M. J. Telford. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* 38: W7-13.
- Abbott, R., and L. Rieseberg. 2012. *Hybrid Speciation*, eLS. John Wiley & Sons, Ltd., Chichester.
- Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287: 2185-2195.
- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* 215: 403-410.
- Alvarez, M., M. F. Ruiz, and L. Sánchez. 2009. Effect of the gene doublesex of *anastrepha* on the somatic sexual development of *Drosophila*. *PLoS One* 4: e5141.
- An, W., S. Cho, H. Ishii, and P. C. Wensink. 1996. Sex-specific and non-sex-specific oligomerization domains in both of the doublesex transcription factors from *Drosophila melanogaster*. *Mol Cell Biol* 16: 3106-3111.
- Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, E. M. Zdobnov, and I. Consortium. 2000. InterPro--an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 16: 1145-1150.
- Araki, T., and J. Milbrandt. 1996. Ninjurin, a novel adhesion molecule, is induced by nerve injury and promotes axonal growth. *Neuron* 17: 353-361.
- Arensburger, P., K. Megy, R. M. Waterhouse, J. Abrudan, P. Amedeo, B. Antelo, L. Bartholomay, S. Bidwell, E. Caler, F. Camara, C. L. Campbell, K. S. Campbell, C. Casola, M. T. Castro, I. Chandramouliswaran, S. B. Chapman, S. Christley, J. Costas, E. Eisenstadt, C. Feschotte, C. Fraser-Liggett, R. Guigo, B. Haas, M.

- Hammond, B. S. Hansson, J. Hemingway, S. R. Hill, C. Howarth, R. Ignell, R. C. Kennedy, C. D. Kodira, N. F. Lobo, C. Mao, G. Mayhew, K. Michel, A. Mori, N. Liu, H. Naveira, V. Nene, N. Nguyen, M. D. Pearson, E. J. Pritham, D. Puiu, Y. Qi, H. Ranson, J. M. Ribeiro, H. M. Roberston, D. W. Severson, M. Shumway, M. Stanke, R. L. Strausberg, C. Sun, G. Sutton, Z. J. Tu, J. M. Tubio, M. F. Unger, D. L. Vanlandingham, A. J. Vilella, O. White, J. R. White, C. S. Wondji, J. Wortman, E. M. Zdobnov, B. Birren, B. M. Christensen, F. H. Collins, A. Cornel, G. Dimopoulos, L. I. Hannick, S. Higgs, G. C. Lanzaro, D. Lawson, N. H. Lee, M. A. Muskavitch, A. S. Raikhel, and P. W. Atkinson. 2010. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science* 330: 86-88.
- Arnold, M. 1997. Natural Hybridization and evolution.
- Gilchrist, B.M. and J.B.S Haldane. 1947. Sex linkage and sex determination in a mosquito, *Culex molestus*. *Hereditas* 33: 175-190.
- Bahnck, C. M., and D. M. Fonseca. 2006. Rapid assay to identify the two genetic forms of *Culex* (*Culex*) *pipiens* L. (Diptera: Culicidae) and hybrid populations. *Am J Trop Med Hyg* 75: 251-255.
- Bailey, T. C., M. W. Merritt, and F. Tediosi. 2015. Investing in Justice: Ethics, Evidence, and the Eradication Investment Cases for Lymphatic Filariasis and Onchocerciasis. *Am J Public Health*: e1-e8.
- Baker, B. S. 1989. Sex in flies: the splice of life. *Nature* 340: 521-524.
- Baker, B. S., and M. F. Wolfner. 1988. A molecular analysis of doublesex, a bifunctional gene that controls both male and female sexual differentiation in *Drosophila melanogaster*. *Genes Dev* 2: 477-489.
- Barr, A. 1957. The distribution of *Culex p. pipiens* and *Culex p. quinquefasciatus* in North America. *American Journal of Tropical Medicine and Hygiene* 6: 153-165.
- Bayrer, J. R., W. Zhang, and M. A. Weiss. 2005. Dimerization of doublesex is mediated by a cryptic ubiquitin-associated domain fold: implications for sex-specific gene regulation. *J Biol Chem* 280: 32989-32996.
- Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. 2004. Ultraconserved elements in the human genome. *Science* 304: 1321-1325.
- Bekku, H. 1956. Studies on the *Culex pipiens* group of Japan. *Nagasaki Medical Journal* 31: 956-966.
- Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57: 289-300.
- Bhatnagar, R. K., N. Arora, S. Sachidanand, M. Shahabuddin, D. Keister, and V. S. Chauhan. 2003. Synthetic propeptide inhibits mosquito midgut chitinase and blocks sporogonic development of malaria parasite. *Biochem Biophys Res Commun* 304: 783-787.
- Black, D. L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* 72: 291-336.

- Blandin, S. A., E. Marois, and E. A. Levashina. 2008. Antimalarial responses in *Anopheles gambiae*: from a complement-like protein to a complement-like pathway. *Cell Host Microbe* 3: 364-374.
- Borovsky, D. 2003. Biosynthesis and control of mosquito gut proteases. *IUBMB Life* 55: 435-441.
- Borovsky, D., and Y. Schlein. 1987. Trypsin and chymotrypsin-like enzymes of the sandfly *Phlebotomus papatasi* infected with *Leishmania* and their possible role in vector competence. *Med Vet Entomol* 1: 235-242.
- Briley, P. A., M. T. Filbin, G. G. Lunt, and P. D. Turner. 1981. Glutamate receptor binding in insects and mammals. *Mol Cell Biochem* 39: 347-356.
- Brinen, L. S., and T. J. Stout. 2003. Can mosquitoes be bitten? A new hope for anti-malarial drug design. *Structure* 11: 1309-1310.
- Broderick, S., X. Wang, N. Simms, and A. Page-McCaw. 2012. *Drosophila* Ninjurin A induces nonapoptotic cell death. *PLoS One* 7: e44567.
- Bull, J., and R. Vogt. 1979. Temperature-Dependent Sex Determination in Turtles. *Science* 206: 1186-1188.
- Bulmer, M. S. 2010. Evolution of Immune Proteins in Insects. In: *Encyclopedia of Life Sciences*, John Wiley & Sons, Ltd., Chichester.
- Burghardt, G., M. Hediger, C. Siegenthaler, M. Moser, A. Dübendorfer, and D. Bopp. 2005. The transformer2 gene in *Musca domestica* is required for selecting and maintaining the female pathway of development. *Dev Genes Evol* 215: 165-176.
- Calvo, E., B. J. Mans, J. F. Andersen, and J. M. Ribeiro. 2006. Function and evolution of a mosquito salivary protein family. *J Biol Chem* 281: 1935-1942.
- Calvo, E., B. J. Mans, J. M. Ribeiro, and J. F. Andersen. 2009a. Multifunctionality and mechanism of ligand binding in a mosquito antiinflammatory protein. *Proc Natl Acad Sci U S A* 106: 3728-3733.
- Calvo, E., V. M. Pham, O. Marinotti, J. F. Andersen, and J. M. Ribeiro. 2009b. The salivary gland transcriptome of the neotropical malaria vector *Anopheles darlingi* reveals accelerated evolution of genes relevant to hematophagy. *BMC Genomics* 10: 57.
- Chapman-Smith, A., and J. E. Cronan. 1999. In vivo enzymatic protein biotinylation. *Biomol Eng* 16: 119-125.
- Charlab, R., J. G. Valenzuela, E. D. Rowton, and J. M. Ribeiro. 1999. Toward an understanding of the biochemical and pharmacological complexity of the saliva of a hematophagous sand fly *Lutzomyia longipalpis*. *Proc Natl Acad Sci U S A* 96: 15155-15160.
- Chaves, L. F., L. C. Harrington, C. L. Keogh, A. M. Nguyen, and U. D. Kitron. 2010. Blood feeding patterns of mosquitoes: random or structured? *Front Zool* 7: 3.
- Cho, S., and P. C. Wensink. 1998. Linkage between oligomerization and DNA binding in *Drosophila* doublesex proteins. *Biochemistry* 37: 11301-11308.
- Cho, S., Z. Y. Huang, and J. Zhang. 2007. Sex-specific splicing of the honeybee doublesex gene reveals 300 million years of evolution at the bottom of the insect sex-determination pathway. *Genetics* 177: 1733-1741.

- Choe, K. M., H. Lee, and K. V. Anderson. 2005. *Drosophila* peptidoglycan recognition protein LC (PGRP-LC) acts as a signal-transducing innate immune receptor. *Proc Natl Acad Sci U S A* 102: 1122-1126.
- Clough, E., E. Jimenez, Y. A. Kim, C. Whitworth, M. C. Neville, L. U. Hempel, H. J. Pavlou, Z. X. Chen, D. Sturgill, R. K. Dale, H. E. Smith, T. M. Przytycka, S. F. Goodwin, M. Van Doren, and B. Oliver. 2014. Sex- and tissue-specific functions of *Drosophila* doublesex transcription factor target genes. *Dev Cell* 31: 761-773.
- Concha, C., and M. J. Scott. 2009. Sexual development in *Lucilia cuprina* (Diptera, Calliphoridae) is controlled by the transformer gene. *Genetics* 182: 785-798.
- Conesa, A., S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674-3676.
- Cornel, A. J., R. D. McAbee, J. Rasgon, M. A. Stanich, T. W. Scott, and M. Coetzee. 2003. Differences in extent of genetic introgression between sympatric *Culex pipiens* and *Culex quinquefasciatus* (Diptera: Culicidae) in California and South Africa. *J Med Entomol* 40: 36-51.
- Dafa'alla, T., G. Fu, and L. Alphey. 2010. Use of a regulatory mechanism of sex determination in pest insect control. *J Genet* 89: 301-305.
- Darriba, D., G. L. Taboada, R. Doallo, and D. Posada. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9: 772.
- de Boer, J. P., A. A. Creasey, A. Chang, J. J. Abbink, D. Roem, A. J. Eerenberg, C. E. Hack, and F. B. Taylor. 1993. Alpha-2-macroglobulin functions as an inhibitor of fibrinolytic, clotting, and neutrophilic proteinases in sepsis: studies using a baboon model. *Infect Immun* 61: 5035-5043.
- Devenport, M., P. H. Alvarenga, L. Shao, H. Fujioka, M. L. Bianconi, P. L. Oliveira, and M. Jacobs-Lorena. 2006. Identification of the *Aedes aegypti* peritrophic matrix protein AeIMUCI as a heme-binding protein. *Biochemistry* 45: 9540-9549.
- Devi, T. R., and B. V. Shyamala. 2013. Male- and female-specific variants of doublesex gene products have different roles to play towards regulation of Sex combs reduced expression and sex comb morphogenesis in *Drosophila*. *J Biosci* 38: 455-460.
- Dieckmann, U., J. Metz, M. Sabelis, and K. Sigmund. 2002. *Adaptive Dynamics of Infectious Diseases*, Cambridge: Cambridge University Press.
- Dobrotworsky, N. 1967. The problem of the *Culex pipiens* complex in the South Pacific (including Australia). *Bulletin of the World Health Organization* 37: 251-255.
- Dong, Y., and G. Dimopoulos. 2009. Anopheles fibrinogen-related proteins provide expanded pattern recognition capacity against bacteria and malaria parasites. *J Biol Chem* 284: 9835-9844.
- Eddy, S. R. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* 7: e1002195.
- Emlen, D. 2008. The Evolution of Animal Weapons. *Annual Review of Ecology, Evolution and Systematics* 39: 387-413.
- Erdman, S. E., and K. C. Burtis. 1993. The *Drosophila* doublesex proteins share a novel zinc finger related DNA binding domain. *EMBO J* 12: 527-535.

- Erdman, S. E., H. J. Chen, and K. C. Burtis. 1996. Functional and genetic characterization of the oligomerization and DNA binding properties of the *Drosophila* doublesex proteins. *Genetics* 144: 1639-1652.
- Farajollahi, A., D. M. Fonseca, L. D. Kramer, and A. Marm Kilpatrick. 2011. "Bird biting" mosquitoes and human disease: a review of the role of *Culex pipiens* complex mosquitoes in epidemiology. *Infect Genet Evol* 11: 1577-1585.
- Felix, C. R., B. Betschart, P. F. Billingsley, and T. A. Freyvogel. 1991. Post-feeding induction of trypsin in the midgut of *Aedes aegypti* L. (Diptera: Culicidae) is separable into two cellular phases. *Insect Biochemistry* 21: 197-203.
- Felsenstein, J. 1989. Mathematics vs. Evolution: Mathematical Evolutionary Theory. *Science* 246: 941-942.
- Fisher, R. 1930. *The Genetical Theory of Natural Selection*, Oxford, Clarendon Press, Great Britain.
- Fonseca, D., N. Keyghobadi, C. Malcolm, M. Mogi, F. Schaffner, R. Fleischer, and R. Wilkerson. 2004a. Response to Outbreak of West Nile virus in North America. *Science* 306: 1473-1475.
- Fonseca, D. M., J. L. Smith, R. C. Wilkerson, and R. C. Fleischer. 2006. Pathways of expansion and multiple introductions illustrated by large genetic differentiation among worldwide populations of the southern house mosquito. *Am J Trop Med Hyg* 74: 284-289.
- Fonseca, D. M., J. L. Smith, H. C. Kim, and M. Mogi. 2009. Population genetics of the mosquito *Culex pipiens pallens* reveals sex-linked asymmetric introgression by *Culex quinquefasciatus*. *Infect Genet Evol* 9: 1197-1203.
- Fonseca, D. M., N. Keyghobadi, C. A. Malcolm, C. Mehmet, F. Schaffner, M. Mogi, R. C. Fleischer, and R. C. Wilkerson. 2004b. Emerging vectors in the *Culex pipiens* complex. *Science* 303: 1535-1538.
- Fritz, M. L., E. D. Walker, J. R. Miller, D. W. Severson, and I. Dworkin. 2015. Divergent host preferences of above- and below-ground *Culex pipiens* mosquitoes and their hybrid offspring. *Med Vet Entomol*.
- Garske, T., M. D. Van Kerkhove, S. Yactayo, O. Ronveaux, R. F. Lewis, J. E. Staples, W. Perea, N. M. Ferguson, and Y. F. E. Committee. 2014. Yellow Fever in Africa: estimating the burden of disease and impact of mass vaccination from outbreak and serological data. *PLoS Med* 11: e1001638.
- Gehring, W. J. 1992. The homeobox in perspective. *Trends Biochem Sci* 17: 277-280.
- Gempe, T., M. Hasselmann, M. Schjøtt, G. Hause, M. Otte, and M. Beye. 2009. Sex determination in honeybees: two separate mechanisms induce and maintain the female pathway. *PLoS Biol* 7: e1000222.
- Geuverink, E., and L. W. Beukeboom. 2014. Phylogenetic distribution and evolutionary dynamics of the sex determination genes doublesex and transformer in insects. *Sex Dev* 8: 38-49.
- Gilles, J. R., M. F. Schetelig, F. Scolari, F. Marec, M. L. Capurro, G. Franz, and K. Bourtzis. 2014. Towards mosquito sterile insect technique programmes: exploring genetic, molecular, mechanical and behavioural methods of sex separation in mosquitoes. *Acta Trop* 132 Suppl: S178-187.



- Gomes, B., R. Parreira, C. A. Sousa, M. T. Novo, A. P. Almeida, M. J. Donnelly, and J. Pinto. 2012. The *Culex pipiens* complex in continental Portugal: distribution and genetic structure. *J Am Mosq Control Assoc* 28: 75-80.
- Grant, P. R., and B. R. Grant. 2002. Unpredictable evolution in a 30-year study of Darwin's finches. *Science* 296: 707-711.
- Guha, M. 2012. Cyclin-dependent kinase inhibitors move into Phase III. *Nat Rev Drug Discov* 11: 892-894.
- Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59: 307-321.
- Han, J. S. 2010. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mob DNA* 1: 15.
- Hanington, P. C., and S. M. Zhang. 2011. The primary role of fibrinogen-related proteins in invertebrates is defense, not coagulation. *J Innate Immun* 3: 17-27.
- Harbach, R., B. Harrison, and A. Gad. 1984. *Culex* (*Culex*) *molestus* Forskal (Diptera: Culicidae): neotype designation, description, variation, and taxonomic status. *Proc Entomol Soc Wash* 86: 521-542.
- Hayashi, Y., S. Shigenobu, D. Watanabe, K. Toga, R. Saiki, K. Shimada, T. Bourguignon, N. Lo, M. Hojo, K. Maekawa, and T. Miura. 2013. Construction and characterization of normalized cDNA libraries by 454 pyrosequencing and estimation of DNA methylation levels in three distantly related termite species. *PLoS One* 8: e76678.
- Hediger, M., G. Burghardt, C. Siegenthaler, N. Buser, D. Hilfiker-Kleiner, A. Dübendorfer, and D. Bopp. 2004. Sex determination in *Drosophila melanogaster* and *Musca domestica* converges at the level of the terminal regulator doublesex. *Dev Genes Evol* 214: 29-42.
- Heinrichs, V., and B. S. Baker. 1995. The *Drosophila* SR protein RBP1 contributes to the regulation of doublesex alternative splicing by recognizing RBP1 RNA target sequences. *EMBO J* 14: 3987-4000.
- Hekmat-Safe, D. S., C. R. Safe, A. J. McKinney, and M. A. Tanouye. 2002. Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. *Genome Res* 12: 1357-1369.
- Hodges, T. K., L. V. Cosme, G. Athrey, S. Pathikonda, W. Takken, and M. A. Slotman. 2014. Species-specific chemosensory gene expression in the olfactory organs of the malaria vector *Anopheles gambiae*. *BMC Genomics* 15: 1089.
- Holt, R. A., G. M. Subramanian, A. Halpern, G. G. Sutton, R. Charlab, D. R. Nusskern, P. Wincker, A. G. Clark, J. M. Ribeiro, R. Wides, S. L. Salzberg, B. Loftus, M. Yandell, W. H. Majoros, D. B. Rusch, Z. Lai, C. L. Kraft, J. F. Abril, V. Anthonard, P. Arensburger, P. W. Atkinson, H. Baden, V. de Berardinis, D. Baldwin, V. Benes, J. Biedler, C. Blass, R. Bolanos, D. Boscus, M. Barnstead, S. Cai, A. Center, K. Chaturverdi, G. K. Christophides, M. A. Chrystal, M. Clamp, A. Cravchik, V. Curwen, A. Dana, A. Delcher, I. Dew, C. A. Evans, M. Flanigan, A. Grundschober-Freimoser, L. Friedli, Z. Gu, P. Guan, R. Guigo, M. E. Hillenmeyer, S. L. Hladun, J. R. Hogan, Y. S. Hong, J. Hoover, O. Jaillon, Z. Ke, C. Kodira, E. Kokoza, A. Koutsos, I. Letunic, A. Levitsky, Y. Liang, J. J. Lin, N. F.

- Lobo, J. R. Lopez, J. A. Malek, T. C. McIntosh, S. Meister, J. Miller, C. Mobarry, E. Mongin, S. D. Murphy, D. A. O'Brochta, C. Pfannkoch, R. Qi, M. A. Regier, K. Remington, H. Shao, M. V. Sharakhova, C. D. Sitter, J. Shetty, T. J. Smith, R. Strong, J. Sun, D. Thomasova, L. Q. Ton, P. Topalis, Z. Tu, M. F. Unger, B. Walenz, A. Wang, J. Wang, M. Wang, X. Wang, K. J. Woodford, J. R. Wortman, M. Wu, A. Yao, E. M. Zdobnov, H. Zhang, Q. Zhao, et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129-149.
- Huang, d. W., B. T. Sherman, and R. A. Lempicki. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1-13.
- Huang, S., G. Molaei, and T. G. Andreadis. 2008. Genetic insights into the population structure of *Culex pipiens* (Diptera: Culicidae) in the Northeastern United States by using microsatellite analysis. *Am J Trop Med Hyg* 79: 518-527.
- Hughes, A. L. 2011. Runaway evolution of the male-specific exon of the doublesex gene in Diptera. *Gene* 472: 1-6.
- Hulpiau, P., and F. van Roy. 2009. Molecular evolution of the cadherin superfamily. *Int J Biochem Cell Biol* 41: 349-369.
- Inoue, K., K. Hoshijima, H. Sakamoto, and Y. Shimura. 1990. Binding of the *Drosophila* sex-lethal gene product to the alternative splice site of transformer primary transcript. *Nature* 344: 461-463.
- James, A. A. 2002. Engineering mosquito resistance to malaria parasites: the avian malaria model. *Insect Biochem Mol Biol* 32: 1317-1323.
- Jasrapuria, S., Y. Arakane, G. Osman, K. J. Kramer, R. W. Beeman, and S. Muthukrishnan. 2010. Genes encoding proteins with peritrophin A-type chitin-binding domains in *Tribolium castaneum* are grouped into three distinct families based on phylogeny, expression and function. *Insect Biochem Mol Biol* 40: 214-227.
- Joyce, D. A., D. H. Lunt, M. J. Genner, G. F. Turner, R. Bills, and O. Seehausen. 2011. Repeated colonization and hybridization in Lake Malawi cichlids. *Curr Biol* 21: R108-109.
- Kalume, D. E., M. Okulate, J. Zhong, R. Reddy, S. Suresh, N. Deshpande, N. Kumar, and A. Pandey. 2005. A proteomic analysis of salivary glands of female *Anopheles gambiae* mosquito. *Proteomics* 5: 3765-3777.
- Kato, Y., K. Kobayashi, H. Watanabe, and T. Iguchi. 2011. Environmental sex determination in the branchiopod crustacean *Daphnia magna*: deep conservation of a Doublesex gene in the sex-determining pathway. *PLoS Genet* 7: e1001345.
- Katoh, K., and H. Toh. 2010. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 26: 1899-1900.
- Kawamata, N., R. Morosetti, C. W. Miller, D. Park, K. S. Spirin, T. Nakamaki, S. Takeuchi, Y. Hatta, J. Simpson, and S. Wilczynski. 1995. Molecular analysis of the cyclin-dependent kinase inhibitor gene p27/Kip1 in human malignancies. *Cancer Res* 55: 2266-2269.
- Kijimoto, T., A. P. Moczek, and J. Andrews. 2012. Diversification of doublesex function underlies morph-, sex-, and species-specific development of beetle horns. *Proc Natl Acad Sci U S A* 109: 20526-20531.

- Kilpatrick, A. M., L. D. Kramer, M. J. Jones, P. P. Marra, and P. Daszak. 2006. West Nile virus epidemics in North America are driven by shifts in mosquito feeding behavior. *PLoS Biol* 4: e82.
- Kiuchi, T., H. Koga, M. Kawamoto, K. Shoji, H. Sakai, Y. Arai, G. Ishihara, S. Kawaoka, S. Sugano, T. Shimada, Y. Suzuki, M. G. Suzuki, and S. Katsuma. 2014. A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* 509: 633-636.
- Koo, B. H., Y. D. Sohn, K. C. Hwang, Y. Jang, D. S. Kim, and K. H. Chung. 2002. Characterization and cDNA cloning of halyxin, a heterogeneous three-chain anticoagulant protein from the venom of *Agkistrodon halys breviceaudus*. *Toxicon* 40: 947-957.
- Koonin, E. V., and R. L. Tatusov. 1994. Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J Mol Biol* 244: 125-132.
- Koonin, E. V., and I. B. Rogozin. 2003. Getting positive about selection. *Genome Biol* 4: 331.
- Kopp, A. 2012. Dmrt genes in the development and evolution of sexual dimorphism. *Trends Genet* 28: 175-184.
- Kothera, L., B. M. Nelms, W. K. Reisen, and H. M. Savage. 2013. Population genetic and admixture analyses of *Culex pipiens* complex (Diptera: Culicidae) populations in California, United States. *Am J Trop Med Hyg* 89: 1154-1167.
- Kramer, L. D., L. M. Styer, and G. D. Ebel. 2008. A global perspective on the epidemiology of West Nile virus. *Annu Rev Entomol* 53: 61-81.
- Krørup, A., S. Thiel, A. Hansen, T. Fujita, and J. C. Jensenius. 2004. L-ficolin is a pattern recognition molecule specific for acetyl groups. *J Biol Chem* 279: 47513-47519.
- Kryazhimskiy, S., and J. B. Plotkin. 2008. The population genetics of dN/dS. *PLoS Genet* 4: e1000304.
- Kuhn, S., V. Sievert, and W. Traut. 2000. The sex-determining gene doublesex in the fly *Megaselia scalaris*: conserved structure and sex-specific splicing. *Genome* 43: 1011-1020.
- Kumar, B. A., and K. P. Paily. 2008. Identification of immune-responsive genes in the mosquito *Culex quinquefasciatus* infected with the filarial parasite *Wuchereria bancrofti*. *Med Vet Entomol* 22: 394-398.
- Lagos, D., M. F. Ruiz, L. Sánchez, and K. Komitopoulou. 2005. Isolation and characterization of the *Bactrocera oleae* genes orthologous to the sex determining Sex-lethal and doublesex genes of *Drosophila melanogaster*. *Gene* 348: 111-121.
- Lam, B. J., A. Bakshi, F. Y. Ekinci, J. Webb, B. R. Graveley, and K. J. Hertel. 2003. Enhancer-dependent 5'-splice site control of fruitless pre-mRNA splicing. *J Biol Chem* 278: 22740-22747.
- Langer, R. C., and J. M. Vinetz. 2001. Plasmodium ookinete-secreted chitinase and parasite penetration of the mosquito peritrophic matrix. *Trends Parasitol* 17: 269-272.
- Leal, W. S. 2005. Pheromone Reception. *Topics in Current Chemistry* 240: 1-36.

- Leal, W. S., Y. M. Choo, P. Xu, C. S. da Silva, and C. Ueira-Vieira. 2013. Differential expression of olfactory genes in the southern house mosquito and insights into unique odorant receptor gene isoforms. *Proc Natl Acad Sci U S A* 110: 18704-18709.
- Li, M., J. Liu, and C. Zhang. 2011. Evolutionary history of the vertebrate mitogen activated protein kinases family. *PLoS One* 6: e26999.
- Low, M. G., and P. W. Kincade. 1985. Phosphatidylinositol is the membrane-anchoring domain of the Thy-1 glycoprotein. *Nature* 318: 62-64.
- Lu, B. 1997. *Fauna Sinica, Insecta. Vol. 9, Diptera. Culicidae II.*, Science Press, Beijing.
- Luo, S. D., G. W. Shi, and B. S. Baker. 2011. Direct targets of the *D. melanogaster* DSXF protein and the evolution of sexual development. *Development* 138: 2761-2771.
- Lynch, K. W., and T. Maniatis. 1995. Synergistic interactions between two distinct elements of a regulated splicing enhancer. *Genes Dev* 9: 284-293.
- Madala, P. K., J. D. Tyndall, T. Nall, and D. P. Fairlie. 2010. Update 1 of: Proteases universally recognize beta strands in their active sites. *Chem Rev* 110: PR1-31.
- Mallet, J. 2007. Hybrid speciation. *Nature* 446: 279-283.
- Manoharan, M., M. Ng Fuk Chong, A. Vaïtinadapoulé, E. Frumence, R. Sowdhamini, and B. Offmann. 2013. Comparative genomics of odorant binding proteins in *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus*. *Genome Biol Evol* 5: 163-180.
- Marchler-Bauer, A., J. B. Anderson, P. F. Cherukuri, C. DeWeese-Scott, L. Y. Geer, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, G. H. Marchler, M. Mullokandov, B. A. Shoemaker, V. Simonyan, J. S. Song, P. A. Thiessen, R. A. Yamashita, J. J. Yin, D. Zhang, and S. H. Bryant. 2005. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res* 33: D192-196.
- Marshall Graves, J. A. 2008. Weird animal genomes and the evolution of vertebrate sex and sex chromosomes. *Annu Rev Genet* 42: 565-586.
- Marín, I., and B. S. Baker. 1998. The evolutionary dynamics of sex determination. *Science* 281: 1990-1994.
- Matson, C. K., M. W. Murphy, A. L. Sarver, M. D. Griswold, V. J. Bardwell, and D. Zarkower. 2011. DMRT1 prevents female reprogramming in the postnatal mammalian testis. *Nature* 476: 101-104.
- Matsuda, M., A. Shinomiya, M. Kinoshita, A. Suzuki, T. Kobayashi, B. Paul-Prasanth, E. L. Lau, S. Hamaguchi, M. Sakaizumi, and Y. Nagahama. 2007. DMY gene induces male development in genetically female (XX) medaka fish. *Proc Natl Acad Sci U S A* 104: 3865-3870.
- Mavárez, J., and M. Linares. 2008. Homoploid hybrid speciation in animals. *Mol Ecol* 17: 4181-4185.
- McBride, C. S., F. Baier, A. B. Omondi, S. A. Spitzer, J. Lutomia, R. Sang, R. Ignell, and L. B. Vosshall. 2014. Evolution of mosquito preference for humans linked to an odorant receptor. *Nature* 515: 222-227.

- McKay, M. M., and D. K. Morrison. 2007. Integrating signals from RTKs to ERK/MAPK. *Oncogene* 26: 3113-3121.
- Megy, K., S. J. Emrich, D. Lawson, D. Campbell, E. Dialynas, D. S. Hughes, G. Koscielny, C. Louis, R. M. MacCallum, S. N. Redmond, A. Sheehan, P. Topalis, D. Wilson, and V. Consortium. 2012. VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res* 40: D729-734.
- Mensch, J., F. Serra, N. J. Lavagnino, H. Dopazo, and E. Hasson. 2013. Positive selection in nucleoporins challenges constraints on early expressed genes in *Drosophila* development. *Genome Biol Evol* 5: 2231-2241.
- Micieli, M. V., A. C. Matarachiero, E. Muttis, D. M. Fonseca, M. T. Aliota, and L. D. Kramer. 2013. Vector competence of Argentine mosquitoes (Diptera: Culicidae) for West Nile virus (Flaviviridae: Flavivirus). *J Med Entomol* 50: 853-862.
- Minh, B. Q., M. A. Nguyen, and A. von Haeseler. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* 30: 1188-1195.
- Minshull, J., J. Pines, R. Golsteyn, N. Standart, S. Mackie, A. Colman, J. Blow, J. V. Ruderman, M. Wu, and T. Hunt. 1989. The role of cyclin synthesis, modification and destruction in the control of cell division. *J Cell Sci Suppl* 12: 77-97.
- Misof, B., S. Liu, K. Meusemann, R. S. Peters, A. Donath, C. Mayer, P. B. Frandsen, J. Ware, T. Flouri, R. G. Beutel, O. Niehuis, M. Petersen, F. Izquierdo-Carrasco, T. Wappler, J. Rust, A. J. Aberer, U. Aspöck, H. Aspöck, D. Bartel, A. Blanke, S. Berger, A. Böhm, T. R. Buckley, B. Calcott, J. Chen, F. Friedrich, M. Fukui, M. Fujita, C. Greve, P. Grobe, S. Gu, Y. Huang, L. S. Jermiin, A. Y. Kawahara, L. Krogmann, M. Kubiak, R. Lanfear, H. Letsch, Y. Li, Z. Li, J. Li, H. Lu, R. Machida, Y. Mashimo, P. Kapli, D. D. McKenna, G. Meng, Y. Nakagaki, J. L. Navarrete-Heredia, M. Ott, Y. Ou, G. Pass, L. Podsiadlowski, H. Pohl, B. M. von Reumont, K. Schütte, K. Sekiya, S. Shimizu, A. Slipinski, A. Stamatakis, W. Song, X. Su, N. U. Szucsich, M. Tan, X. Tan, M. Tang, J. Tang, G. Timelthaler, S. Tomizuka, M. Trautwein, X. Tong, T. Uchifune, M. G. Walz, B. M. Wiegmann, J. Wilbrandt, B. Wipfler, T. K. Wong, Q. Wu, G. Wu, Y. Xie, S. Yang, Q. Yang, D. K. Yeates, K. Yoshizawa, Q. Zhang, R. Zhang, W. Zhang, Y. Zhang, J. Zhao, C. Zhou, L. Zhou, T. Ziesmann, S. Zou, X. Xu, H. Yang, J. Wang, K. M. Kjer, and X. Zhou. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346: 763-767.
- Mogi, M. 2012. The forms of the *Culex pipiens* complex in East Asia, with ecological thoughts on their origin and interrelation. *J Am Mosq Control Assoc* 28: 28-52.
- Moreno-Hagelsieb, G., and K. Latimer. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* 24: 319-324.
- Mugal, C. F., J. B. Wolf, and I. Kaj. 2014. Why time matters: codon evolution and the temporal dynamics of dN/dS. *Mol Biol Evol* 31: 212-231.
- Nagoshi, R. N., and B. S. Baker. 1990. Regulation of sex-specific RNA splicing at the *Drosophila* doublesex gene: cis-acting mutations in exon sequences alter sex-specific RNA splicing patterns. *Genes Dev* 4: 89-97.

- Nei, M. 2007. The new mutation theory of phenotypic evolution. *Proc Natl Acad Sci U S A* 104: 12235-12242.
- Nelson, N., N. Perzov, A. Cohen, K. Hagai, V. Padler, and H. Nelson. 2000. The cellular biology of proton-motive force generation by V-ATPases. *J Exp Biol* 203: 89-95.
- Nene, V., J. R. Wortman, D. Lawson, B. Haas, C. Kodira, Z. J. Tu, B. Loftus, Z. Xi, K. Megy, M. Grabherr, Q. Ren, E. M. Zdobnov, N. F. Lobo, K. S. Campbell, S. E. Brown, M. F. Bonaldo, J. Zhu, S. P. Sinkins, D. G. Hogenkamp, P. Amedeo, P. Arensburger, P. W. Atkinson, S. Bidwell, J. Biedler, E. Birney, R. V. Bruggner, J. Costas, M. R. Coy, J. Crabtree, M. Crawford, B. Debruyne, D. Decaprio, K. Eiglmeier, E. Eisenstadt, H. El-Dorri, W. M. Gelbart, S. L. Gomes, M. Hammond, L. I. Hannick, J. R. Hogan, M. H. Holmes, D. Jaffe, J. S. Johnston, R. C. Kennedy, H. Koo, S. Kravitz, E. V. Kriventseva, D. Kulp, K. Labutti, E. Lee, S. Li, D. D. Lovin, C. Mao, E. Mauceli, C. F. Menck, J. R. Miller, P. Montgomery, A. Mori, A. L. Nascimento, H. F. Naveira, C. Nusbaum, S. O'leary, J. Orvis, M. Pertea, H. Quesneville, K. R. Reidenbach, Y. H. Rogers, C. W. Roth, J. R. Schneider, M. Schatz, M. Shumway, M. Stanke, E. O. Stinson, J. M. Tubio, J. P. Vanzee, S. Verjovski-Almeida, D. Werner, O. White, S. Wyder, Q. Zeng, Q. Zhao, Y. Zhao, C. A. Hill, A. S. Raikhel, M. B. Soares, D. L. Knudson, N. H. Lee, J. Galagan, S. L. Salzberg, I. T. Paulsen, G. Dimopoulos, F. H. Collins, B. Birren, C. M. Fraser-Liggett, and D. W. Severson. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316: 1718-1723.
- Newton, M. E., D. I. Southern, and R. J. Wood. 1974. X and Y chromosomes of *Aedes aegypti* (L.) distinguished by Giemsa C-banding. *Chromosoma* 49: 41-49.
- Noriega, F. G., A. E. Colonna, and M. A. Wells. 1999. Increase in the size of the amino acid pool is sufficient to activate translation of early trypsin mRNA in *Aedes aegypti* midgut. *Insect Biochem Mol Biol* 29: 243-247.
- Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302: 205-217.
- Ohbayashi, F., M. G. Suzuki, K. Mita, K. Okano, and T. Shimada. 2001. A homologue of the *Drosophila* doublesex gene is transcribed into sex-specific mRNA isoforms in the silkworm, *Bombyx mori*. *Comp Biochem Physiol B Biochem Mol Biol* 128: 145-158.
- Osório, H. C., L. Zé-Zé, F. Amaro, A. Nunes, and M. J. Alves. 2014. Sympatric occurrence of *Culex pipiens* (Diptera, Culicidae) biotypes *pipiens*, *molestus* and their hybrids in Portugal, Western Europe: feeding patterns and habitat determinants. *Med Vet Entomol* 28: 103-109.
- Parmley, J. L., and L. D. Hurst. 2007. How common are intragene windows with  $KA > KS$  owing to purifying selection on synonymous mutations? *J Mol Evol* 64: 646-655.
- Pascoa, V., P. L. Oliveira, M. Dansa-Petretski, J. R. Silva, P. H. Alvarenga, M. Jacobs-Lorena, and F. J. Lemos. 2002. *Aedes aegypti* peritrophic matrix and its interaction with heme during blood digestion. *Insect Biochem Mol Biol* 32: 517-523.
- Peiser, L., M. P. De Winther, K. Makepeace, M. Hollinshead, P. Coull, J. Plested, T. Kodama, E. R. Moxon, and S. Gordon. 2002. The class A macrophage

- scavenger receptor is a major pattern recognition receptor for *Neisseria meningitidis* which is independent of lipopolysaccharide and not required for secretory responses. *Infect Immun* 70: 5346-5354.
- Pelosi, P., and R. Mařda. 1995. [Physiological functions of odorant-binding proteins]. *Biofizika* 40: 137-145.
- Peterson, G. I., and J. Masel. 2009. Quantitative prediction of molecular clock and  $k_a/k_s$  at short timescales. *Mol Biol Evol* 26: 2595-2603.
- Pomerantz, A. F., and M. A. Hoy. 2015. Expression analysis of *Drosophila* doublesex, transformer-2, intersex, fruitless-like, and vitellogenin homologs in the parahaploid predator *Metaseiulus occidentalis* (Chelicerata: Acari: Phytoseiidae). *Exp Appl Acarol* 65: 1-16.
- Pomerantz, A. F., M. A. Hoy, and A. Y. Kawahara. 2014. Molecular characterization and evolutionary insights into potential sex-determination genes in the western orchard predatory mite *Metaseiulus occidentalis* (Chelicerata: Arachnida: Acari: Phytoseiidae). *J Biomol Struct Dyn*: 1-15.
- Porat, A., Y. Sagiv, and Z. Elazar. 2000. A 56-kDa selenium-binding protein participates in intra-Golgi protein transport. *J Biol Chem* 275: 14457-14465.
- Price, D., and D. Fonseca. 2015. Genetic divergence between populations of feral and domestic forms of a mosquito disease vector assessed by transcriptomics. *PeerJ* 3: e807.
- Rawlings, N. D., and A. J. Barrett. 1994. Families of serine peptidases. *Methods Enzymol* 244: 19-61.
- Raymond, C. S., J. R. Kettlewell, B. Hirsch, V. J. Bardwell, and D. Zarkower. 1999. Expression of *Dmrt1* in the genital ridge of mouse and chicken embryos suggests a role in vertebrate sexual development. *Dev Biol* 215: 208-220.
- Reid, W. R., L. Zhang, F. Liu, and N. Liu. 2012. The transcriptome profile of the mosquito *Culex quinquefasciatus* following permethrin selection. *PLoS One* 7: e47163.
- Ribeiro, J. M., R. Charlab, V. M. Pham, M. Garfield, and J. G. Valenzuela. 2004. An insight into the salivary transcriptome and proteome of the adult female mosquito *Culex pipiens quinquefasciatus*. *Insect Biochem Mol Biol* 34: 543-563.
- Roux, J., E. Privman, S. Moretti, J. T. Daub, M. Robinson-Rechavi, and L. Keller. 2014. Patterns of positive selection in seven ant genomes. *Mol Biol Evol* 31: 1661-1685.
- Ruiz, M. F., J. M. Eirín-López, R. N. Stefani, A. L. Perondini, D. Selivon, and L. Sánchez. 2007. The gene doublesex of *Anastrepha* fruit flies (Diptera, Tephritidae) and its evolution in insects. *Dev Genes Evol* 217: 725-731.
- Saccone, G., M. Salvemini, and L. C. Polito. 2011. The transformer gene of *Ceratitis capitata*: a paradigm for a conserved epigenetic master regulator of sex determination in insects. *Genetica* 139: 99-111.
- Saccone, G., M. Salvemini, A. Pane, and L. C. Polito. 2008. Masculinization of XX *Drosophila* transgenic flies expressing the *Ceratitis capitata* DoublesexM isoform. *Int J Dev Biol* 52: 1051-1057.
- Salvemini, M., M. Robertson, B. Aronson, P. Atkinson, L. C. Polito, and G. Saccone. 2009. *Ceratitis capitata* transformer-2 gene is required to establish and

- maintain the autoregulation of *Cctra*, the master gene for female sex determination. *Int J Dev Biol* 53: 109-120.
- Salvemini, M., U. Mauro, F. Lombardo, A. Milano, V. Zazzaro, B. Arcà, L. C. Polito, and G. Saccone. 2011. Genomic organization and splicing evolution of the doublesex gene, a *Drosophila* regulator of sexual differentiation, in the dengue and yellow fever mosquito *Aedes aegypti*. *BMC Evol Biol* 11: 41.
- Scali, C., F. Catteruccia, Q. Li, and A. Crisanti. 2005. Identification of sex-specific transcripts of the *Anopheles gambiae* doublesex gene. *J Exp Biol* 208: 3701-3709.
- Schultze, A., P. Pregitzer, M. F. Walter, D. F. Woods, O. Marinotti, H. Breer, and J. Krieger. 2013. The co-expression pattern of odorant binding proteins and olfactory receptors identify distinct trichoid sensilla on the antenna of the malaria mosquito *Anopheles gambiae*. *PLoS One* 8: e69412.
- Serra, F., L. Arbiza, J. Dopazo, and H. Dopazo. 2011. Natural selection on functional modules, a genome-wide analysis. *PLoS Comput Biol* 7: e1001093.
- Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51: 492-508.
- Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17: 1246-1247.
- Shin, D., A. Civana, C. Acevedo, and C. T. Smartt. 2014. Transcriptomics of differential vector competence: West Nile virus infection in two populations of *Culex pipiens quinquefasciatus* linked to ovary development. *BMC Genomics* 15: 513.
- Shukla, J. N., and S. R. Palli. 2012. Doublesex target genes in the red flour beetle, *Tribolium castaneum*. *Sci Rep* 2: 948.
- Shukla, J. N., and S. R. Palli. 2014. Production of all female progeny: evidence for the presence of the male sex determination factor on the Y chromosome. *J Exp Biol* 217: 1653-1655.
- Silverbush, D., and R. Sharan. 2014. Network orientation via shortest paths. *Bioinformatics* 30: 1449-1455.
- Siwicki, K. K., and E. A. Kravitz. 2009. Fruitless, doublesex and the genetics of social behavior in *Drosophila melanogaster*. *Curr Opin Neurobiol* 19: 200-206.
- Smale, S. T., and D. Baltimore. 1989. The "initiator" as a transcription control element. *Cell* 57: 103-113.
- Smith, J. L., and D. M. Fonseca. 2004. Rapid assays for identification of members of the *Culex* (*Culex*) *pipiens* complex, their hybrids, and other sibling species (Diptera: culicidae). *Am J Trop Med Hyg* 70: 339-345.
- Sobrinho, I. S., and R. A. de Brito. 2012. Positive and purifying selection influence the evolution of doublesex in the *Anastrepha fraterculus* species group. *PLoS One* 7: e33446.
- Sottrup-Jensen, L., O. Sand, L. Kristensen, and G. H. Fey. 1989. The alpha-macroglobulin bait region. Sequence diversity and localization of cleavage sites for proteinases in five mammalian alpha-macroglobulins. *J Biol Chem* 264: 15781-15789.
- Spielman, A. 2001. Structure and seasonality of nearctic *Culex pipiens* populations. *Ann N Y Acad Sci* 951: 220-234.



- Spielman, A., T. G. Andreadis, C. S. Apperson, A. J. Cornel, J. F. Day, J. D. Edman, D. Fish, L. C. Harrington, A. E. Kiszewski, R. Lampman, G. C. Lanzaro, F. R. Matuschka, L. E. Munstermann, R. S. Nasci, D. E. Norris, R. J. Novak, R. J. Pollack, W. K. Reisen, P. Reiter, H. M. Savage, W. J. Tabachnick, and D. M. Wesson. 2004. Outbreak of West Nile virus in North America. *Science* 306: 1473-1475; author reply 1473-1475.
- Stavrou, E., and A. H. Schmaier. 2010. Factor XII: what does it contribute to our understanding of the physiology and pathophysiology of hemostasis & thrombosis. *Thromb Res* 125: 210-215.
- Suzuki, M. G., S. Funaguma, T. Kanda, T. Tamura, and T. Shimada. 2003. Analysis of the biological functions of a doublesex homologue in *Bombyx mori*. *Dev Genes Evol* 213: 345-354.
- Sánchez, L. 2008. Sex-determining mechanisms in insects. *Int J Dev Biol* 52: 837-856.
- Tanaka, K., K. Mizusaka, and E. Saugstad. 1979. A revision of the adult and larval mosquitoes of Japan (including the Ryukyu archipelago and the Ogasawara islands) and Korea (Diptera: Culicidae). *Contributions of the American Entomological Institute* 16: 1-987.
- Terra, W. R., and C. Ferreira. 1981. The physiological role of the peritrophic membrane and trehalase: Digestive enzymes in the midgut and excreta of starved larvae of *Rhynchosciara*. *Journal of Insect Physiology* 27: 325-331.
- Terrapon, N., C. Li, H. M. Robertson, L. Ji, X. Meng, W. Booth, Z. Chen, C. P. Childers, K. M. Glastad, K. Gokhale, J. Gowin, W. Gronenberg, R. A. Hermansen, H. Hu, B. G. Hunt, A. K. Huylmans, S. M. Khalil, R. D. Mitchell, M. C. Munoz-Torres, J. A. Mustard, H. Pan, J. T. Reese, M. E. Scharf, F. Sun, H. Vogel, J. Xiao, W. Yang, Z. Yang, J. Zhou, J. Zhu, C. S. Brent, C. G. Elsik, M. A. Goodisman, D. A. Liberles, R. M. Roe, E. L. Vargo, A. Vilcinskas, J. Wang, E. Bornberg-Bauer, J. Korb, G. Zhang, and J. Liebig. 2014. Molecular traces of alternative social organization in a termite genome. *Nat Commun* 5: 3636.
- Tian, M., and T. Maniatis. 1993. A splicing enhancer complex controls alternative splicing of doublesex pre-mRNA. *Cell* 74: 105-114.
- Turell, M., D. Dohm, and D. Fonseca. 2014. Comparison of the potential for different genetic forms in the *Culex pipiens* complex (Diptera: Culicidae) in North America to transmit Rift Valley fever virus. *Journal of the American Mosquito Control Association* In press.
- Urbanelli, S., F. Silvestrini, W. K. Reisen, E. De Vito, and L. Bullini. 1997. Californian hybrid zone between *Culex pipiens pipiens* and *Cx. p. quinquefasciatus* revisited (Diptera: Culicidae). *J Med Entomol* 34: 116-127.
- Valenzuela, J. G., V. M. Pham, M. K. Garfield, I. M. Francischetti, and J. M. Ribeiro. 2002. Toward a description of the sialome of the adult female mosquito *Aedes aegypti*. *Insect Biochem Mol Biol* 32: 1101-1122.
- Verhulst, E. C., L. van de Zande, and L. W. Beukeboom. 2010. Insect sex determination: it all evolves around transformer. *Curr Opin Genet Dev* 20: 376-383.
- Vieira, F. G., and J. Rozas. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and

- evolutionary history of the chemosensory system. *Genome Biol Evol* 3: 476-490.
- Wang, D., Y. Zhang, Z. Zhang, J. Zhu, and J. Yu. 2010. KaKs\_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* 8: 77-80.
- Wang, D., F. Liu, L. Wang, S. Huang, and J. Yu. 2011. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. *Biol Direct* 6: 13.
- Wang, D., S. Zhang, F. He, J. Zhu, S. Hu, and J. Yu. 2009. How do variable substitution rates influence Ka and Ks calculations? *Genomics Proteomics Bioinformatics* 7: 116-127.
- Wang, E., H. Ni, R. Xu, A. D. Barrett, S. J. Watowich, D. J. Gubler, and S. C. Weaver. 2000. Evolutionary relationships of endemic/epidemic and sylvatic dengue viruses. *J Virol* 74: 3227-3234.
- Wang, P., and R. Granados. 2001. Molecular structure of the peritrophic membrane (PM): Identification of potential PM target sites for insect control. *Archives of Insect Biochemistry and Physiology* 47: 110-118.
- Wang, X., Q. Zhao, and B. M. Christensen. 2005. Identification and characterization of the fibrinogen-like domain of fibrinogen-related proteins in the mosquito, *Anopheles gambiae*, and the fruitfly, *Drosophila melanogaster*, genomes. *BMC Genomics* 6: 114.
- Weeda, E., A. Koopmanschap, C. de Kort, and A. Beenakkers. 1980. Proline synthesis in fat body of *Leptinotarsa decemlineata*. *Insect Biochemistry* 10: 631-636.
- Wexler, J. R., D. C. Plachetzki, and A. Kopp. 2014. Pan-metazoan phylogeny of the DMRT gene family: a framework for functional studies. *Dev Genes Evol* 224: 175-181.
- Whyard, S., C. Erdelyan, A. Partridge, A. Singh, N. Beebe, and R. Capina. 2015. Silencing the buzz: a new approach to population suppression of mosquitoes by feeding larvae double-stranded RNAs. *Parasites & Vectors* 8.
- Wilkins, A. S. 1995. Moving up the hierarchy: a hypothesis on the evolution of a genetic sex determination pathway. *Bioessays* 17: 71-77.
- Zarkower, D. 2001. Establishing sexual dimorphism: conservation amidst diversity? *Nat Rev Genet* 2: 175-185.
- Zhang, W., B. Li, R. Singh, U. Narendra, L. Zhu, and M. A. Weiss. 2006. Regulation of sexual dimorphism: mutational and chemogenetic analysis of the doublesex DM domain. *Mol Cell Biol* 26: 535-547.
- Zhu, L., J. Wilken, N. B. Phillips, U. Narendra, G. Chan, S. M. Stratton, S. B. Kent, and M. A. Weiss. 2000. Sexual dimorphism in diverse metazoans is regulated by a novel class of intertwined zinc fingers. *Genes Dev* 14: 1750-1764.
- Zolnierowicz, S. 2000. Type 2A protein phosphatase, the complex regulator of numerous signaling pathways. *Biochem Pharmacol* 60: 1225-1235.