Molecular population genetics of *Sex-lethal* (*Sxl*) in the *D. melanogaster* species group - a locus that genetically interacts with *Wolbachia pipientis* in *Drosophila melanogaster*.

Vanessa L. Bauer DuMont[1], Simone L. White, Daniel Zinshteyn and Charles F. Aquadro[2]

Department of Molecular Biology and Genetics

Cornell University

Ithaca, NY 14853

[1] Present address: Department of Molecular, Cellular, and Developmental Biology and the BioFrontiers Institute, University of Colorado Boulder, Boulder, Colorado, USA

[2] Corresponding author: Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, E-mail: CFA1@CORNELL.EDU

Genbank accession numbers: KT935592-KT935663 (will be made public upon publication)

Running title: Molecular Evolution of *Sex-Lethal*

Key words: germline stem cells, *Wolbachia pipientis*, population genetics, natural selection, genetic conflict

Corresponding author: Charles F. Aquadro, Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, Phone: 607-254-4838, email: cfa1@cornell.edu

Abstract

Sex-lethal (*Sxl)* is the sex determination switch in *Drosophila*, and also plays a critical role in germ-line stem cell (GSC) daughter differentiation in *Drosophila melanogaster*. Three female-sterile alleles at *Sxl* in *Drosophila melanogaster* were previously shown to genetically interact to varying degrees with the maternally inherited endosymbiont *Wolbachia pipientis.* Given this genetic interaction and *W. pipientis'* ability to manipulate reproduction in *Drosophila*, we carried out a careful study of both the population genetics (within four *Drosophila* species) and molecular evolutionary analysis (across 20 *Drosophila* species) of *Sxl*. Consistent with earlier studies, we find that selective constraint has played a prominent role in *Sxl's* molecular evolution within *Drosophila*, but we also observe patterns that suggest both episodic bursts of protein evolution and recent positive selection at *Sxl*. The episodic nature of *Sxl's* protein evolution is discussed in light of its genetic interaction with *W. pipientis*.

Introduction

Reproductive success is a key fitness trait governed by a plethora of gene regulatory networks. Despite the presumed functional constraint for such critical genes, many of them have been shown to be evolving rapidly due to positive selection at the amino acid level (Clark et al. 2006; Chapman 2008; Wong and Rundle 2013; Popovic et al. 2014). For some sets of reproductive loci this observation makes intuitive sense, including loci involved in species-specific gamete recognition and those involved in coevolutionary conflict between the sexes. Interestingly, non-neutral patterns of amino acid evolution have also been detected at loci involved in the

differentiation of germ-line stem cells (GSCs) (Civetta et al. 2006; Bauer DuMont et al. 2007; Langley et al. 2012; Pool et al. 2012; Choi and Aquadro 2015; Flores et al. 2015a). The temporal and spatial expression of these GSC regulating loci does not coincide with that expected for genes influenced by sperm competition, sexual selection, inbreeding avoidance and gamete recognition (reviewed in Clark et al. 2006).

Several GSC regulating loci have functions outside the germline (e.g., Bell et al. 1988; Yi et al. 2008; Saito et al. 2010; Le Thomas et al. 2013) so their signatures of positive selection could be due to non-gametogenic functions. For many others it is possible that the positive selection is acting directly on gametogenic functions. We have previously hypothesized that interactions with maternally transmitted endosymbionts could be a gametogenesis-specific driver of positive selection (Bauer DuMont et al. 2007; Flores et al. 2015a and 2015b). One such endosymbiont, *Wolbachia pipientis,* infects an estimated 66% of arthropod species (Hilgenboecker et al. 2008) and is an obligate maternally transmitted endosymbiont that has been shown to manipulate host reproduction systems (Werren et al. 2008). *Wolbachia pipientis* infection has also been shown to have beneficial consequences. For example, infection in *Drosophila melanogaster* conveys greater resistance to viruses (Hedges et al. 2008; Teixeira et al. 2008; Chrostek et al. 2013) and has been shown to increase female fecundity on low and high iron diets relative to uninfected flies (Brownlie et al. 2009).

Here we explore the phylogenetic patterns of positive selection for amino acid diversification at *Sxl*, the sex determination master switch in *Drosophila* development (Bell et al. 1988) that also plays a critical role, along with *bag of marbles* (*bam*), in the maturation of cystoblasts during oogenesis (Chau et al. 2009). Proper *bam* function is essential for the start of cystoblast differentiation (McKearin and Spradling 1990). Epistasis experiments revealed that *bam* requires

73 *Sxl* activity for proper germline stem cell daughter differentiation and the presence of both
74 proteins is proposed to be responsible for regulating sex-specific gametogenesis (Chau et al.
75 2009; Chau et al. 2012; Shapiro-Kulnane et al. 2015).

76 In addition to genetically interacting with one another, hypomorphic alleles of both proteins have
77 been found to genetically interact with *W. pipientis* infection. Starr and Cline (2002) reported
78 that *W. pipientis* infection partially rescues the female-sterile phenotype caused by three *Sxl*
79 alleles, though to differing degrees for each allele. Similarly, *W. pipientis* infection can mitigate
80 the mutant phenotype of a hypomorphic allelic combination at the *bam* locus (Flores et al.
81 2015b). Starr and Cline (2002) also reported that *W. pipientis* infection did not rescue mutants in
82 three other genes with ovarian tumor phenotypes (snf^{1621}, otu^{11}, and $mei\text{-}P26^{fs1}$). These results
83 suggest a specific, possibly physical, interaction of *W. pipientis* or its gene products with the *Sxl*
84 and *bam* genes or gene products.

85 Using methods that consider both polymorphism and divergence we and others have previously
86 shown that in *D. melanogaster* and *D. simulans,* the *bam* locus is evolving rapidly at the amino
87 acid level due to positive selection (Civetta et al. 2006; Bauer DuMont et al. 2007). Phylogenetic
88 methods for detecting selection across a number of *Drosophila* species did not detect evidence of
89 positive selection, suggesting the selection pressures acting on *bam* are episodic. On the other
90 hand, recent studies of the molecular evolution of *Sxl* have focused on the evolution of *Sxl's* role
91 as the sex determination master switch in *Drosophila* development (Mullon et al. 2012; Zhang et
92 al. 2013). Mullon et al. (2012) used the maximum likelihood based phylogenetic methods
93 implemented in PAML (Yang 2007) within and between 3 families of Diptera (Drosophilidae,
94 Tephritidae and Muscidae), and detected both a relaxation of functional constraint and positive
95 selection acting across amino acid sites along the lineage leading to *Drosophila*. However, they

96 observe no evidence of positive selection among the 12 *Drosophila* species suggesting that the
97 positive selection they observed was associated with the acquisition of the sex determination
98 function in the common ancestor of *Drosophila* species.

99 Here we test for evidence departures from selective neutrality at *Sxl* within the genus *Drosophila*
100 specifically by incorporating sequence polymorphism and divergence data, and test for
101 departures consistent with lineage specific positive selection. First, we obtained high-quality
102 Sanger sequencing polymorphism data at *Sxl* in the following species: *Drosophila melanogaster,*
103 *D. simulans, D. ananassae* and *D. pseudoobscura*. Except for *D. pseudoobscura,* these species
104 are currently infected with *W. pipientis* (Mateos et al. 2006). Second, we searched 8 additional
105 sequenced *Drosophila* genomes to annotate *Sxl* orthologs, bringing the total number of
106 *Drosophila Sxl* orthologous sequences to 20 for phylogenetic analysis. The inclusion of
107 polymorphism data and additional *Drosophila* species allowed us greater power to detect
108 potential signatures of positive selection at *Sxl* within the genus *Drosophila*.

109

110 **Materials and Methods**

111 All DNA was extracted using the Qiagen puregene kit A DNA isolation kits (Qiagen). For *D.*
112 *melanogaster,* we sequenced *Sxl* in 20 extracted X-chromosome lines from Zambia Africa,
113 which was made through a series of crosses using the balancer X chromosome line Fm7a and
114 isofemale Zambia lines (Pool et al. 2012). For *D. simulans*, we used 10 lines from a Madagascar
115 population sample collected in 1998 by J.W.O. Ballard. For *D. pseudoobscura* we included 29
116 lines collected from Mesa Verde National Park, Kaibab National Park and the Bosque Del
117 Apache National Wildlife Refuge provided by Stephen Schaeffer. Our line of *D. miranda* was

provided by Doris Bachtrog. Finally, we also surveyed *Sxl* variation across 12 lines of *D. ananassae* collected from Bangkok, Thailand provided by Wolfgang Stephan, and a single line of *D. bipectinata* from the UCSD Species Stock Center (stock 0000-1029.01). We also sequenced a single *Sxl* allele from *D. guanche* (obtained from the *Drosophila* Species Stock Center at University of California San Diego; stock number 14011-0095.01), and *D. atripex* (provided by Artyom Kopp), which were used for analyzes that require divergence for the *D. pseudoobscura* and *D. ananassae* datasets respectively. Our sequences are available via Genbank accession numbers KT935592 - KT935663. All other sequences used in our analyses were those from *Drosophila* 12 Genomes Consortium et al. (2007) or Chen et al. (2014) and downloaded from Flybase.

While there are multiple *Sxl* transcripts, due to alternative splicing, most of them include exons 5 through 8 in *D. melanogaster* according to Flybase genome annotations (St Pierre et al. 2014). We PCR amplified and sequenced exons 5 through 8. We used Promega GoTaq for amplification following their standard protocol. Sanger sequencing was performed using the PCR primers and internal sequencing primers (primer sequences available upon request) through the Cornell Institute of Biotechnology Genomics Facility (https://www.biotech.cornell.edu/core-facilities-brc/facilities/genomics-facility).

The following triplets of species were independently aligned using MegAlign (DNASTAR Inc., Madison WI): *D. melanogaster, D. simulans, D. yakuba; D. ananassae, D. atripex*, *D. bipectinata*; and *D. pseudoobscura*, *D. miranda, D. guanche.* Gaps within coding regions were manually adjusted to ensure the sequences remained in-frame. Population genetic analyses were performed using DNAsp 5.10.1 (Librado and Rozas 2009). When using coalescent simulations to obtain the *P*-values of site frequency based tests we incorporated recombination rate estimates

(following Przeworski et al. 2001). For the *D. melanogaster* and *D. simulans* datasets, we used the *Drosophila melanogaster* Recombination Rate Calculator (Comeron et al. 2012) estimate of r = 3.34×10^{-8} recombinants per base-pair per generation for the *Sxl* region of the X chromosome. This translates to an estimated R = 152 for the 1500 base pair region sequenced ($R = 3N_e r$ since *Sxl* is X-linked and assuming a population size of 1.0×10^6). For *D. ananassae* and *D. pseudoobscura* the values of R were estimated from the polymorphism data directly using DNAsp 5.10.1 (Librado and Rozas 2009) as R=194 and R=6, respectively.

For *D. melanogaster* we also incorporated demography into our neutral simulations when obtaining our significance cut-offs. These simulations were done using the program msABC (Pavlidis et al. 2010). There is growing evidence that African populations of *D. melanogaster* have experienced changes in effective population size over time (Glinka et al 2003; Li and Stephan 2006; Hutter et al 2007; Haddrill et al 2008; Duchen et al 2013; Singh et al 2013). However, given the large effective population size of these species and signatures of a high rate of adaptation (e.g., Begun et al 2007; Langley et al 2012), inferring demographic parameters is challenging. Because of this we simulated three different scenarios: standard neutral equilibrium model, standard neutral with exponential growth as estimated by Hutter et al (2007), and standard neutral with a 3 phase bottleneck as estimate by Duchen et al. (2013). We supplied msABC with uniform prior distributions for theta and all demographic parameters. The prior distribution for theta for *D. melanogaster* was obtained from Pool et al. (2012) and ranged between 0.006 and 0.009 per site. The resulting *P*-values are the proportion of simulated datasets that were less than (for negative statistics) or greater than (for positive statistics) our observed test statistic for *Sxl*. The *P*-values were adjusted for multiple testing following the Bonferroni method.

The McDonald-Kreitman test (MKT) was done manually following the method's original implementation (McDonald and Kreitman 1991) by combining polymorphism from multiple species if it was available. If a position in the alignment had more than one nucleotide segregating within a species' population sample, it was labeled as polymorphic. Divergent sites were those for which all alleles from one species differed from all the alleles of the other two species.

To test for evidence of departures from neutrality for synonymous sites at *Sxl*, we used the method of Bauer DuMont et al. (2005). This method looks for differences in the rates of preferred and unpreferred codon substitutions per site in a manner similar to a dN/dS comparison (Nei and Gojobori 1986). Statistical significance is assessed by a 2x2 contingency table comparison.

In order to test for evidence of departures from selective neutrality in rates and patterns of sequence evolution at *Sxl* across a broader group of *Drosophila* species, we first retrieved the *Sxl* gene region sequences from FlyBase for the following 20 *Drosophila* species: *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D. eugracilis, D. ficusphila, D. rhopaloa, D. elegans, D. takahashii, D. biarmipes, D. kikkawae, D. bipectinata, D. ananassae, D. miranda, D. pseudoobsura, D. persimilis, D. willistoni, D. mojavensis, D. virilis, and D. grimshawi*. *Sxl* is an alternatively spliced locus. To ensure we are analyzing orthologous exons, we first made alignments of the entire gene region (introns and exons) using the web-based versions of the alignment programs Muscle (Edgar 2004) (http://www.ebi.ac.uk/Tools/msa/muscle/). We then used the annotated exons of *D. melanogaster* as a guide to identify orthologous coding sequences (CDS) for *Sxl's* Isoform L (6 exons - the female specific splice variant; Bell et al. 1988) from

each aligned sequence. This isoform contains the poly-proline region of the Sxl protein where the female-sterile *Sxl* variants are located (Starr and Cline 2002).

We estimated a *Sxl* gene tree across these 20 *Drosophila* species using Mega 5.1 (Tamura et al. 2011). A maximum likelihood tree was estimated using all nucleotide sites, default parameters, and the GTR substitution model with gamma distributed site variation. To test for selection across the estimated *Sxl* tree we used Hyphy (Kosakovsky Pond et al. 2005) run online using the DataMonkey website (http://www.datamonkey.org/). A model selection procedure was conducted to determine that the best nucleotide substitution model for the data was TrN93 (Tamura and Nei 1993) which was used in all subsequent analyzes. We ran GARD (Kosakovsky Pond et al. 2006a; Kosakovsky Pond et al. 2006b) to look for evidence of recombination across these species at *Sxl* (that would reflect incomplete ancestral polymorphism sorting) using a general discrete site-to-site rate variation and 3 rate classes. To detect evidence of purifying and/or diversifying selection across sites and lineages we used the following Hyphy programs: BranchRel, GAbranch, FUBAR and MEME.

Results

We surveyed DNA variability at the population level at *Sxl* for four species of *Drosophila* (*D. melanogaster, D. simulans, D. ananassae* and *D. pseudoobscura*), the first three of which have evidence of current *W. pipientis* infection (Mateos et al. 2006). Even though 37 lines of *D. pseudoobscura* were surveyed for infection, this species appears to be currently uninfected with *W. pipientis* (Mateos et al. 2006).

We find that Sxl is a very conserved protein with little to no nonsynonymous polymorphism or divergence within or between the 20 *Drosophila* species surveyed (Table 1). In addition, levels

208 of synonymous polymorphism and divergence between *D. melanogaster* and *D. simulans* are
209 below the average reported by Andolfatto (2005). The same is true for synonymous variation
210 observed within and between *D. pseudoobscura* and *D. miranda* as compared to that reported by
211 Haddrill et al. (2010). While synonymous polymorphism is slightly lower at *Sxl* within *D.*
212 *ananassae* the level of divergence between *D. ananassae* and *D. atripex* is similar to values
213 previously reported (Grath et al. 2009; Choi and Aquadro 2014).

214 Previous studies have reported a skew in the Site Frequency Spectrum (SFS) toward rare alleles,
215 as illustrated by a general negative Tajima *D* (Tajima 1989) test statistic, in all of the species
216 included in our study (Kliman et al. 2000; Machado et al. 2002; Das et al. 2004; Andolfatto
217 2007; Grath et al. 2009; Haddrill et al. 2010; Jensen and Bachtrog 2011;). At *Sxl,* we observe a
218 negative Tajima *D* for all species, except *D. ananassae*. Tajima *D* in *D. simulans* rejects the
219 hypothesis of a SFS at equilibrium with 67% (33/49) of the polymorphisms being singletons.
220 These singletons are evenly distributed across synonymous (sevem singleton/10 total) and intron
221 (26 singleton/39 total) sites. At *Sxl,* Fay and Wu's *H* is negative in all species but *D. simulans*.
222 Fay and Wu's *H* statistic is significantly negative in *D. melanogaster* even when considering two
223 different demographic scenarios estimated for African populations of this species (Hutter et al.
224 2007; Duchen et al. 2013).

225 The McDonald-Kreitman Test (MKT; McDonald and Kreitman 1991) is used to detect
226 departures from the neutral expectation that synonymous and nonsynonymous variants will have
227 similar ratios of within to between species variation. A rejection in the direction of an excess of
228 nonsynonymous divergence is typically interpreted as evidence of repeated amino acid
229 substitutions due to positive selection. As seen in Table 1 there are few nonsynonymous changes
230 at *Sxl*. The MKT does not reject neutral expectations when *D. ananassae* and *D pseudoobscura*

polymorphism is compared to divergence to *D. atripex* and *D. miranda*, respectively (Table 2). We observe no amino acid differences within or between *D. melanogaster* and *D. simulans*. However, when considering the *Sxl* sequence from two additional and closely related Drosophila species, *D. yakuba* and *D. erecta,* we observe seven amino acid substitutions along the lineage leading to *D. melanogaster* and *D. simulans*. Given this observation, we chose to apply the MKT to these species in a manner similar to its first implementation (McDonald and Kreitman 1991) by combining the polymorphism from *D. melanogaster* and *D. simulans*. Our divergent changes in this MKT comparison are all differences since the most recent common ancestor of the *D. melanogaster* and *D. simulans* lineage rooted by the *D. yakuba* and *D. erecta* lineages. This combined polymorphism MKT rejects neutral expectations, in the direction suggestive of an excess of amino acid substitutions.

The significant MKT for *D. melanogaster/D. simulans* could be due to selective fixation of amino acid differences or to selection acting on synonymous changes. While synonymous sites have traditionally been assumed as the neutral yardstick of molecular evolution, there is evidence that this assumption may be invalid, for at least some genes in *Drosophila* (e.g., (Bauer DuMont et al. 2004; Bauer DuMont et al. 2009; Poh et al. 2012; Lawrie et al. 2013). We looked for evidence of selection acting on synonymous changes at *Sxl* using a per site counting method (CF-test) similar to a dN/dS comparison (Bauer DuMont et al. 2004), except in this test we are comparing the number of changes toward unpreferred or preferred codons per the number of unpreferred and preferred "sites". Along the *D. simulans*, *D. ananassae* and *D. pseudoobscura* lineages we observe a significant departure from neutrality in the direction suggesting a selective advantage of mutations toward preferred codons at *Sxl* (Table 3). The test did not reject in *D. melanogaster*.

Selection acting on synonymous sites can lead to false positive MKT results by elevating the synonymous polymorphism cell of the 2x2 table, due to segregating slightly deleterious synonymous variants. One method proposed to mitigate the effects of selective constraint on the MKT is to remove low frequency polymorphisms from the analysis (Fay et al. 2002). Given the CF-test results for *D. simulans*, we also carried out the MKT removing two derived unpreferred polymorphic sites at low frequency (singletons – with a frequency of 10% in sample) in *D. simulans*. The *D. melanogaster/D. simulans* MKT remains significant (*P*-value = 0.031).

To further assess the molecular evolution at *Sxl* we made a coding sequence (CDS) alignment using the program Muscle for *Sxl's* Isoform L (6 exons - the female specific splice variant; Bell et al 1988) for the following 20 Drosophila species: *D. melanogaster, D. sechellia, D. yakuba, D. erecta, D. eugracilis, D. ficusphila, D. rhopaloa, D. elegans, D. takahashii, D. biarmipes, D. kikkawae, D. bipectinata, D. ananassae, D. miranda, D. pseudoobsura, D. persimilis, D. willistoni, D. mojavensis, D. virilis, and D. grimshawi*. We observe 86 amino acid substitutions at 53 codon positions among these species at *Sxl*. Roughly 38% of the codons that have experienced an amino acid substitution have been hit multiple times (20/53; note that some multiply hit amino acid positions had more than three different amino acids segregating among the species). The conservation of the RNA binding domain of the Sxl protein has been previously noted (Zhang, Klein, Nei 2013). In agreement, 84 out of the 86 amino acid changes occurred outside the RNA binding domain region, between codons 1 - 136 and 304 – 373 in our alignment (Figures 1 and 2). We will call these non-RNA binding regions of the Sxl protein the N-terminal and C-terminal regions, respectively. The amino acid substitutions at Sxl have not occurred equally between the N-terminal and C-terminal regions after taking into account their differences in total codon length. We observe significantly more amino acid substitutions in the C-terminal

region (24 codons with an amino acid substitution out of 137 total codons in N-terminal region versus 27 substituted codons out of 70 total codons in C-terminal region; 2x2 table chi-square = 11.1, *P*-value = 0.001). The two regions have experienced a similar proportion of multiple hit codons (10 multiple hit codons out of 24 total codons with an amino acid substitution in N-terminal region versus 10 out of 27 such codons in C-terminal region).

To determine if the apparent heterogeneity in amino acid substitution at *Sxl* is the result of positive selection, we analyzed the data using the phylogeny-based Hyphy method (Kosakovsky Pond et al. 2005; Delport et al. 2010). Phylogenetic incongruence between species along a gene sequence, due to sorting of ancestral polymorphisms, can have adverse effects on phylogenetic based inferences of positive selection (Wong et al. 2007). We first performed the Hyphy GARD method (Kosakovsky Pond et al. 2006b), which is designed to detect such incongruences. None were detected at a *P*-value cutoff of 0.10. Therefore, we used a maximum-likelihood gene tree, made from 3rd codon positions, in subsequent analyzes.

We applied the following methods of the Hyphy package to our data: GA-branch (Kosakovsky Pond and Frost 2005), BranchRel (Kosakovsky Pond et al. 2011), MEME (Murrell et al. 2012) and FUBAR (Murrell et al. 2013). GAbranch detects significant heterogeneity across the *Sxl* phylogeny in the rate of nonsynonymous compared to synonymous evolution (the dN/dS ratio). The best fitting model includes 3 rate classes, yet the dN/dS for the highest class is only 0.115 across the *Sxl* locus. The posterior probabilities suggest that the following branches are within the highest dN/dS rate class and that they are evolving at a significantly different rate than other branches in the tree: the branch leading to *D. melanogaster* and *D. sechellia*, the branch leading to *D. elegans* and *D. rhopaloa*, the branch leading to the melanogaster species group, and the *D. kikkawei* lineage (Figure 2).

The BranchRel method pools information across sites to estimate selection parameters along branches. The method reports the proportion of codons along each lineage that have evolved under three selection regimes: negative selection, neutral/nearly neutral or episodic positive selection. BranchRel confirmed ubiquitous evidence of amino acid constraint (negative selection) across the *Sxl* phylogeny. After multiple testing corrections, no lineage has significant evidence of positive diversifying selection. Similar results were obtained using MEME.

FUBAR is used to detect selection pressure acting on individual codons. The strength of this method is that it does not restrict the parameter space for which nonsynonymous and synonymous rates are drawn from during the maximum likelihood process. FUBAR does not detect any sites under diversifying selection with a posterior probability greater than 0.90. However, it does detect 258 codons under negative selection with a posterior probability greater than 0.90, which is roughly 70% of the protein. Just over half of these negatively selected sites (140) are located within the RNA binding domain. We observe no significant difference in the number of negatively selected sites between the N-terminal and C-terminal regions of *Sxl* relative to their respective lengths (N-terminal: 77 negative selective sites in 137 codons versus C-terminal: 41 negative selected sites in 70 codons; 2x2 chi-square = 0.023, *P*-value = 0.879).

Interestingly, the seven amino acid differences observed on the lineage leading to the ancestor of the *D. melanogaster* and *D. simulans* species group (with which we observe the significant MKT), cluster with the location of the mutations previously shown to genetically interact with *W. pipientis* (Figure 2).

Discussion

In this study we use population and phylogenetic based methods to examine the molecular population genetics and evolution of the *Sxl* locus within the genus *Drosophila* for which *Sxl* is the master switch in sex-determination. We were motivated by the observation of a genetic interaction between *W. pipientis* infection and some mutant alleles at *Sxl* in *D. melanogaster* (Starr and Cline 2002; Sun and Cline 2009). It is not known if this interaction is due to a ubiquitous effect of *W. pipientis* on overall egg production, or if it is due to a direct interaction between the endosymbiont and the *Sxl* locus or protein product.

Considering polymorphism data alone, we detect patterns consistent with a recent selective sweep in both *D. simulans* and *D. melanogaster* with the Tajima *D* and Fay and Wu *H* tests, respectively. These significant skews in the frequency spectrum in these species could be due to a variety of evolutionary forces including demography, a selective sweep associated with the fixation of a linked positively selected mutation, or segregating weakly deleterious mutations. *D. simulans* is thought to have experienced a recent population expansion resulting in a general tendency for loci in this species to have negative Tajima *D* test statistics (Kliman et al. 2000). The significant Fay and Wu's *H* test in *D. melanogaster* remains significant even when demography is incorporated into the null distribution of the test, suggesting that in this species we are detecting a recent selective sweep. These signatures of positive selection would not be due to an amino acid fixation, given that there are no amino acid differences between *D. melanogaster* and *D. simulans* at *Sxl.*

Sxl's long-term evolution within *Drosophila* largely reflects strong conservation of protein sequence, as noted previously by Mullon et al. (2012). We detect the action of negative selection both along lineages and at specific codons. The FUBAR method estimates that 70% of the *Sxl* codons are selectively constrained, suggesting that negative selection has had a pervasive effect

on *Sxl's* molecular evolution. However, we do observe amino acid differences across these *Drosophila* species and the pattern of these substitutions is heterogeneous. For example, GA-branch method detects significant variation in the dN/dS ratio across the *Drosophila* species included in this analysis. This heterogeneity at *Sxl* could be due to sporadic relaxations of the negative selection that dominates *Sxl's* molecular evolution or due to sporadic bursts of positive selection.

The Hyphy methods used to detect recurring or episodic positive selection fail to do so after multiple testing corrections. However, we note that the pervasive negative selection observed at *Sxl* could confound these methods, especially if the positive selection is weak or if only a few sites are affected (Kosakovsky Pond et al. 2011). Our current data shows no evidence of long-term or recent positive selection at *Sxl* along the *D. ananassae* and *D pseudoobscura* lineages. In contrast, there are weak signatures of both types of selection in *D. melanogaster* and *D. simulans*, so we focus on the molecular evolution of these species and the lineage leading to their common ancestor.

The lineage leading to *D. melanogaster and D. simulans* shows a decoupling of synonymous and nonsynonymous evolution with the MKT in the direction of an excess of nonsynonymous divergence. This result is due to seven amino acid substitutions on the lineage leading to the *D. melanogaster/D. simulans* clade. These nonsynonymous substitutions cluster with the locations of the *Sxl* alleles that genetically interact with *W. pipientis.* In addition, this region of the Sxl protein (the C-terminal non-RNA binding region) has experienced significantly more amino acid substitutions than the N-terminal region. This elevation in amino acid substitutions does not appear to be due to a simple relaxation of constraint as we observe no difference between the C- and N-terminal regions in the number of codons predicted to be experiencing negative selection.

These results are suggestive of positive selection being at least partially responsible for the fixations of these seven *D. melanogaster/D. simulans* amino acid substitutions. However, there are other possible explanations for these results such as synonymous site evolution and/or changes in effective population. The seven amino acid fixations could be due to the fixation of slightly deleterious mutations if the effective population size was smaller on the branch leading to the *D. melanogaster/D. simulans*. However, the relaxation of constraint is expected to affect both synonymous and nonsynonymous substitutions. We assume the ancestral state in codon preference is toward preferred synonymous codons at *Sxl*, given the results of the CF-test. If relaxation of constraint were responsible for the burst of amino acid fixations on the *D. melanogaster/D. simulans* lineage we may also expect a burst of derived unpreferred substitutions, but we do not observe this. We observe an equal number of preferred and unpreferrred changes on this lineage (data not shown).

Infection dynamics of *W. pipientis* appear to be sporadic and variable both between and within lineages, with uninfected species interspersed with infected species throughout the phylogeny (e.g., Mateos et al. 2006) consistent with multiple losses or gains of infection. The resulting uncertainty in the infection history of species unfortunately prevents us from reliably testing for correlations between *W. pipientis* infection status and burst of positive selection; there are many factors that could weaken our ability to detect an association.

In this study we present data revealing both similarities and difference between the molecular evolution at *bam* and *Sxl*, two loci that genetically interact with *W. pipientis* infection. For both loci the fixation of amino acid variants appears to be heterogeneous, potentially weakening phylogenetic methods to detect positive selection or associations with character states. The MKT does reject neutrality for both loci in a manner suggestive of an acceleration of nonsynonymous

391 fixations. However, the extent of amino acid differences is very different between these loci with
392 there being 59 fixed amino acid substitutions between *D. melanogaster* and *D. simulans* at *bam*
393 and none at *Sxl*.

394 Our results do not allow us to draw strong conclusions regarding the role of positive selection on
395 the molecular evolution of the *Sxl* locus within *Drosophila*, but it also does not allow us to
396 discount the influence of both long term (MKT) or recent (Tajima D and Fay and Wu H tests)
397 positive selection in *D. melanogaster* and *D. simulans*. Current data and methodology also do not
398 allow us to make a direct connection between *W. pipientis* infection and selective pressures
399 acting on *Sxl* or *bam*. So, it remains open whether the genetic interaction between mutant *Sxl* and
400 *bam* alleles and *W. pipientis* is due to a direct interaction between Sxl and Bam protein and this
401 endosymbiont. Our results do motivate screening for genetic interactions between *W. pipientis*
402 and other mutant alleles at *Sxl* and other GSC loci because the observation that *W. pipientis*
403 rescues some but not other mutations (e.g., Starr and Cline 2002) will help refine candidates for
404 the mechanism(s) by which *W. pipientis* is manipulating *Drosophila* reproduction.

405

406 **Acknowledgements**

407 We would like to thank Jae Young Choi for helpful input regarding data collection and
408 manuscript preparation, and Helen K. Salz for her feedback on the manuscript. Research was
409 supported by National Institute of Health grant number R01GM095793 to C.F.A.

410 **Literature Cited**

411 Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437: 1149-
412 1152.

413 Andolfatto, P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the
414 *Drosophila melanogaster* genome. Genome Res. 17: 1755-1762.
415 Bauer DuMont, V., J. C. Fay, P. P. Calabrese, and C. F. Aquadro, 2004. DNA variability and
416 divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans*: A case of
417 accelerated synonymous site divergence. Genetics 167: 171-85.
418 Bauer Dumont V. L., N. D. Singh, M. H. Wright and C. F. Aquadro, 2009 Locus-specific
419 decoupling of base composition evolution at synonymous sites and introns along the
420 *Drosophila melanogaster* and *Drosophila sechellia* lineages. Genome Biol Evol. 1: 67-74.
421 Bauer Dumont V.L., H. A. Flores, M. H. Wright, and C. F. Aquadro, 2007 Recurrent positive
422 selection at *bgcn*, a key determinant of germ line differentiation, does not appear to be
423 driven by simple coevolution with its partner protein *bam*. Mol. Biol. Evol. 24: 182-91.
424 Bell L. R., E. M. Maine, P. Schedl, and T. W. Cline, 1988 *Sex-lethal*, a Drosophila sex
425 determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to
426 RNA binding proteins. Cell 55:1037-46.
427 Bennett, G. M., N. A. Pantoja, and P. M. O'Grady, 2012 Diversity and phylogenetic
428 relationships of Wolbachia in Drosophila and other native Hawaiian insects. Fly 6: 273-
429 283.
430 Brownlie, J. C., B. N. Cass, M. Riegler, J. J. Witsenburg, I. Iturbe-Ormaetxe, E. A. McGraw, and
431 S. O'Neill, 2009 Evidence for metabolic provisioning by a common invertebrate
432 endosymbiont, *Wolbachia pipientis*, during periods of nutritional stress. PLoS Pathog 5:
433 e1000368.
434 Chapman, T., 2008 The soup in my fly: Evolution, form and function of seminal fluid proteins.
435 PLoS Biol. 6: e179.

436 Chau, J., L. S. Kulnane L. S., and H. K. Salz, 2009. *Sex-lethal* facilitates the transition from
437 germline stem cell to committed daughter cell in the *Drosophila* ovary. Genetics 182: 121-
438 32.
439 Chau, J., L. S. Kulnane, and H. K. Salz, 2012 *Sex-lethal* enables germline stem cell
440 differentiation by down-regulating Nanos protein levels during *Drosophila* oogenesis. Proc.
441 Natl. Acad. Sci. USA 109: 9465-9470.
442 Chen, Z. X., ZD. Sturgill, J. Qu, H. Jiang, S. Park, N. Boley, A. M. Suzuki, et al. 2014.
443 Comparative validation of D. melanogaster modENCODE transcriptome annotation.
444 Genome Res. 24: 1209-1223.
445 Choi, J. Y., and C. F. Aquadro, 2014 The coevolutionary period of *Wolbachia pipientis* infecting
446 *Drosophila ananassae* and its impact on the evolution of the host germline stem cell
447 regulating genes. Mol. Biol. Evol. 31: 2457-2471.
448 Chrostek E, Marialva MSP, Esteves SS, Weinert LA, Martinez J, Jiggins FM, Teixeira L. 2013.
449 *Wolbachia* variants induce differential protection to viruses in *Drosophila melanogaster*: A
450 phenotypic and phylogenomic analysis. PLoS Genet 9: e1003896.
451 Civetta, A., S. A. Rajakumar, B. Brouwers, and J. P. Bacik, 2006 Rapid evolution and gene-
452 specific patterns of selection for three genes of spermatogenesis in *Drosophila*. Mol. Biol.
453 Evol. 23: 655-662.
454 Clark, N. L., J. E. Aagaard, and W. J. Swanson, 2006 Evolution of reproductive proteins from
455 animals and plants. Reproduction 131: 11-22.
456 Comeron, J.M., R. Ratnappan, and S. Bailin, 2012 The many landscapes of recombination in
457 *Drosophila melanogaster*. PLoS Genet. 8: e1002905.

458 Das, A., S. Mohanty, and W. Stephan, 2004 Inferring the population structure and demography
459 of *Drosophila ananassae* from multilocus data. Genetics 168: 1975-1785.
460 Delport, W., A. F. Y. Poon, S. D. W. Frost, and S. L. Kosakovsky Pond, 2010 Datamonkey
461 2010: A suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:
462 2455-2457.
463 Drosophila 12 Genomes Consortium, A. G. Clark, M. B. Eisen, D. R. Smith, C. M. Bergman, B.
464 Oliver, T. A. Markow, et al. 2007. Evolution of genes and genomes on the Drosophila
465 phylogeny. Nature 17: 1932-1942
466 Duchen, P., D. Zivkovic, S. Hutter, W. Stephan, and S. Laurent, 2013 Demographic inference
467 reveals African and European admixture in the North American *Drosophila melanogaster*
468 population Genetics 193: 291-301.
469 Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high
470 throughput. Nucleic Acids Res. 32: 1792-1797.
471 Fay, J.C., G. J. Wyckoff, and C. I. Wu, 2001 Positive and negative selection on the human
472 genome. Genetics 158: 1227-1234.
473 Flores, H. A. F., V. L. Bauer DuMont, A. Fatoo, D. Hubbard, M. Hijji, D. A. Barbash, and C. F.
474 Aquadro. 2015a. Adaptive evolution of genes involved in the regulation of germline stem
475 cells in *Drosophila melanogaster* and *D. simulans*. G3 (Bethesda) 5: 583-592.
476 Flores, H. A. F., J. E. Bubnell, C.F . Aquadro, and D. A. Barbash. 2015b. The *Drosophila bag*
477 *of marbles* gene interacts genetically with *Wolbachia* and shows female-specific effects of
478 divergence. PLoS Genet. 11: e1005453.
479 Grath, S., J. F. Baines, and J. Parsch, 2009 Molecular evolution of sex-biased genes in the
480 *Drosophila ananassae* subgroup. BMC Evol. Biol. 9: 291.

481 Haddrill, P. R. , L. Loewe, and B. Charlesworth, 2010. Estimating the parameters of selection on
482 nonsynonymous mutations in *Drosophila pseudoobscura* and *D. miranda*. Genetics 185:
483 1381-1396.
484 Hedges, L. M., J. C. Brownlie, S. J. O'Neill, and K. N. Johnson, 2008. *Wolbachia* and virus
485 protection in insects. Science 322: 702.
486 Hilgenboecker, K., P. Hammerstein, P. Schlattmann, A. Telschow, and J. H. Werren, 2008 How
487 many species are infected with *Wolbachia*? -- a statistical analysis of current data. FEMS
488 Microbiol. Lett. 281: 215-220.
489 Hutter, S., H. Li, S. Beisswanger, D. De Lorenzo, and W. Stephan, 2007 Distinctly different sex
490 ratios in African and European populations of *Drosophila melanogaster* inferred from
491 chromosomewide single nucleotide polymorphism data. Genetics 177: 469-480.
492 Jensen, J. D., and D. Bachtrog, 2011 Characterizing the influence of effective population size on
493 the rate of adaptation: Gillespie's Darwin domain. Genome Biol. Evol. 3: 687-701.
494 Jiggins, F. M., and K. W. Kim, 2005 The evolution of antifungal peptides in *Drosophila*.
495 Genetics 171: 1847-1859.
496 Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, J.
497 Wakeley, and J. Hey, 2000 The population genetics of the origin and divergence of the
498 *Drosophila simulans* complex species. Genetics 156: 1913-1931.
499 Kosakovsky Pond, S. L., and S. D. W. Frost, 2005 A genetic algorithm approach to detecting
500 lineage-specific variation in selection pressure. Mol. Biol. Evol. 32: 478-485.
501 Kosakovsky Pond, S. L. K., S. D. W. Frost, and S. V. Muse, 2005 HyPhy: Hypothesis testing
502 using phylogenies. Bioinformatics 21: 676-679.

503 Kosakovsky Pond, S, L., B. Murrell, M. Fourment, S. D. W. Frost, W. Delport, and K. Scheffler,
504 2011 A random effects branch-site model for detecting episodic diversifying selection. Mol.
505 Biol. Evol. 28: 3033-3043.
506 Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. Frost, 2006a.
507 Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol.
508 Evol. 23: 1891-1901.
509 Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk, and S. D. W. Frost, 2006b.
510 GARD: A genetic algorithm for recombination detection. Bioinformatics 22: 3096-3098.
511 Lawrie, D. S., P. W. Messer, R. Hershberg, and D. A. Petrov, 2013 Strong purifying selection at
512 synonymous sites in *D. melanogaster*. PLoS Genet. 9: e1003527.
513 Le Thomas, A., A. K. Rogers, A. Webster, G. K. Marinov, S. E. Liao et al., 2013 Piwi induces
514 piRNA-guided transcriptional silencing and establishment of a repressive chromatin state.
515 Genes Dev. 27: 390-399.
516 Librado, P., and J. Rozas, 2009 DnaSP v5: A software for comprehensive analysis of DNA
517 polymorphism data. Bioinformatics 25: 1451-1452.
518 Machado, C. A., R. M. Kliman, J. A. Markert, and J. Hey 2002 Inferring the history of speciation
519 from multilocus DNA sequence data: The case of *Drosophila pseudoobscura* and close
520 relatives. Mol. Biol. Evol. 19: 472-488.
521 Mateos, M., S. J. Castrezana, B. J. Nankivell, A. M. Estes, T. A. Markow *et al*, 2006 Heritable
522 endosymbionts of *Drosophila*. Genetics 174: 363-376.
523 McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in
524 *Drosophila*. Nature 351: 652-654.

McKearin, D. M., and A. C. Spradling,1990 *bag-of-marbles*: A Drosophila gene required to initiate both male and female gametogenesis. Genes Dev. 4: 2242-2251.

Muller, M.J., N. C. Drebes Dorr, M. Depra, H. J. Schmitz, V. H. Valiati and V. L. da Siva Valente, 2013 Reevaluating the infection status by the *Wolbachia* endosymbiont in *Drosophila* Neotropical species from the *willistoni* subgroup. Infect. Genet. Evol. 19: 232-239.

Mullon, C., A. Pomiankowski, and M. Reuter, 2012 Molecular evolution of *Drosophila Sex-lethal* and related sex determining genes. BMC Evol. Biol. 12: 5.

Murrell, B., J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, and S. L. Kosakovsky Pond, 2012 Detecting individual sites subject to episodic diversifying selection. PLoS Genetics 8: e1002764.

Poh, Y. P., C. T. Ting, H. W. Fu, C. H. Langley, and D. J. Begun, 2012 Population genomic analysis of base composition evolution in *Drosophila melanogaster*. Genome Biol. Evol. 4: 1245-1255.

Pool, J. E., A. Wong, and C. F. Aquadro, 2006 Finding of male-killing *Spiroplasma* infecting *Drosophila melanogaster* in Africa implies transatlantic migration of this endosymbiont. Heredity 97: 27-32.

Pool, J. E., R. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, et al., 2012 Population genomics of sub-saharan *Drosophila melanogaster*: African diversity and

Popovic, I., P. B. Marko, J. P. Wares, and M. W. Hart, 2014 Selection and demographic history shape the molecular evolution of the gamete compatibility protein bindin in *Pisaster* sea stars. Ecol. Evol. 4: 1567-1588.

547 Przeworski, M., J. D. Wall, and P. Andolfatto, 2001 Recombination and the frequency spectrum
548 in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. 18: 291-298.
549 Puigbo, P., S. Garcia-Vallve, and J. O. McInerney, 2007. TOPD/FMTS: A new software to
550 compare phylogenetic trees. Bioinformatics 23: 1556-1558.
551 Ravikumar, H., B. M. Prakash, S. Sampathkumar, and H. P. Puttaraju, 2011 Molecular
552 subgrouping of *Wolbachia* and bacteriophage WO infection among some Indian *Drosophila*
553 species. J. Genet. 90: 507-510.
554 Richardson, M. F., L. A. Weinert, J. J. Welch, R. S. Linheiro, M. M. Magwire, et al., 2012
555 Population genomics of the *Wolbachia* endosymbiont in *Drosophila melanogaster*. PLoS
556 Genet. 8: e1003129.
557 Saito, K., H. Ishizu, M. Komai, H. Kotani, Y. Kawamura, et al., 2010 Roles for the Yb body
558 components Armitage and Yb in primary piRNA biogenesis in *Drosophila*. Genes Dev. 24:
559 2493-2498.
560 Shapiro-Kulnane, L., A. E. Smolko, and H. K. Salz, 2015 Maintenance of *Drosophila* germline
561 stem cell sexual identity in oogenesis and turmorigenesis. Development 143: 1073-1082,
562 Starr, D. J., and T. W. Cline, 2002 A host parasite interaction rescues *Drosophila* oogenesis
563 defects. Nature 418: 76-79.
564 Sun, S., and T. W. Cline, 2009 Effects of *Wolbachia* infection and ovarian tumor mutations on
565 *sex-lethal* germline functioning in *Drosophila*. Genetics 181: 1291-1301.
566 Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA
567 polymorphism. Genetics 123: 585.-595.
568 Tamura, K., and M. Nei, 1993 Estimation of the number of nucleotide substitutions in the control
569 region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10: 512-526.

570 Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, 2011 MEGA5:
571 Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance,
572 and maximum parsimony methods. Mol. Biol. Evol .28: 2731-2739.
573 Teixeira, L., A. Ferreira, and M. Ashburner, 2008 The bacterial symbiont *Wolbachia* induces
574 resistance to RNA viral infections in *Drosophila melanogaster.* PLoS Biol. 6: e1000002.
575 Werren, J. H., L. Baldo, and M. E. Clark, 2008 *Wolbachia*: Master manipulators of invertebrate
576 biology. Nat. Rev. Microbiol. 6: 741-751.
577 Wong, A. and H. Rundle, 2013 Selection on the *Drosophila* seminal fluid protein Acp62F. Ecol.
578 Evol. 3: 1942-1950.
579 Wong, A., J. D. Jensen, J. E. Pool, and C. F. Aquadro, 2007 Phylogenetic incongruence in the
580 *Drosophila melanogaster* species group. Mol. Phylogenet. Evol. 43: 1138-1150.
581 Yang, Z, 2007 PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:
582 1586-1591.
583 Yi, X., H. I. de Vries, K. Siudeja, A. Rana, W. Lemstra, J. F. Brunsting, et al. 2008 *Stwl* modifies
584 chromatin compaction and is required to maintain DNA integrity in the presence of
585 perturbed DNA replication. Mol. Biol. Cell 20: 983-994.
586 Zabalou, S., S. Charlat, A. Nirgianaki, D. Lachaise, H. Mercot, and K. Bourtzis, 2004 Natural
587 *Wolbachia* infections in the *Drosophila yakuba* species complex do not induce cytoplasmic
588 incompatibility but fully rescue the *w*RI modification. Genetics 167: 827-834.
589 Zhang, Z., J. Klein, and M. Nei, 2013 Evolution of the *sex-lethal* gene in insects and origin of the
590 sex-determination system in *Drosophila*. J. Mol. Evol. **78**: 50-65.
591

592

593 Figure legends

594 Figure 1 Schematic of Sxl protein. Green box denotes the location of the RNA binding domain
595 of the Sxl protein. Vertical lines show the locations of all amino acid substitutions at *Sxl* across
596 20 Drosophila species with the red lines being those that have occurred specifically on the
597 lineage leading to *D. melanogaster* and *D. sechellia*. Stars at C-terminal end of protein denote
598 the relative location of the mutations that generate the mutant alleles shown to interact
599 genetically with *Wolbachia pipientis*.

600

601 Figure 2 *Sxl* gene tree schematically showing the branches for which amino acid substitutions
602 have occurred. The rectangles denote the Sxl protein with the vertical black lines indicating the
603 location of the amino acid change(s) along that lineage. Hatched box denotes the location of the
604 RNA binding domains of the Sxl protein. Stars at C-terminal end of protein denote the relative
605 location of the three mutations that generate the mutant alleles shown to interact genetically with
606 *W. pipientis*. Species names in bold indicate *Wolbachia* has been detected with the numbers of
607 *Wolbachia* positive lines relative the total number of lines screened given in parenthesis.
608 *Wolbachia* data for all species are from Mateo et al. (2006), as well as for *D. erecta* (Zabalou
609 2004), *D. kikkawai* (Bennett et al. 2012), *D. bipectinata* (Ravikumar et al. 2011), and *D.*
610 *willistoni* (Muller et al. 2013).

Figure 1

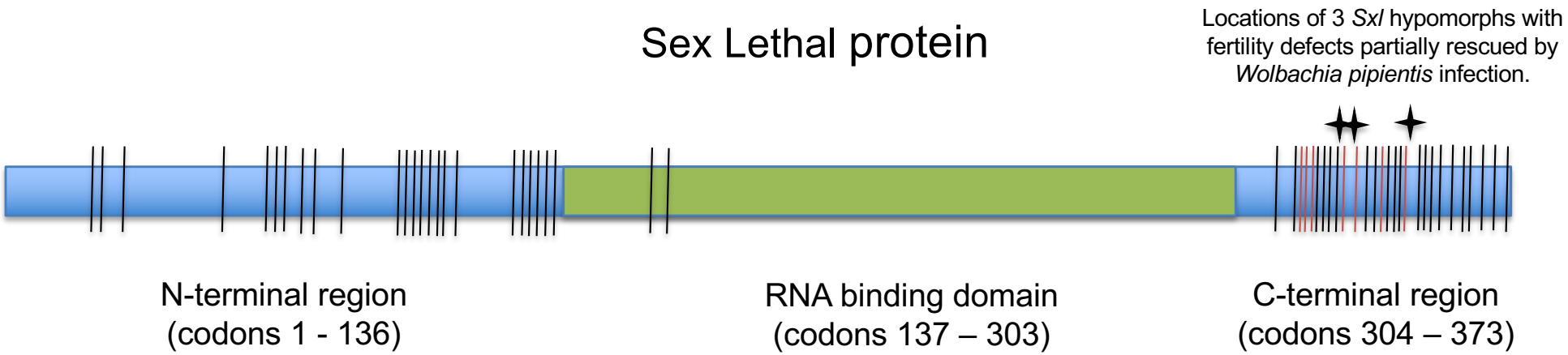
Sex Lethal protein
Locations of 3 *Sxl* hypomorphs with fertility defects partially rescued by *Wolbachia pipientis* infection.
N-terminal region
(codons 1 - 136)
RNA binding domain
(codons 137 – 303)
C-terminal region
(codons 304 – 373)

Figure 2

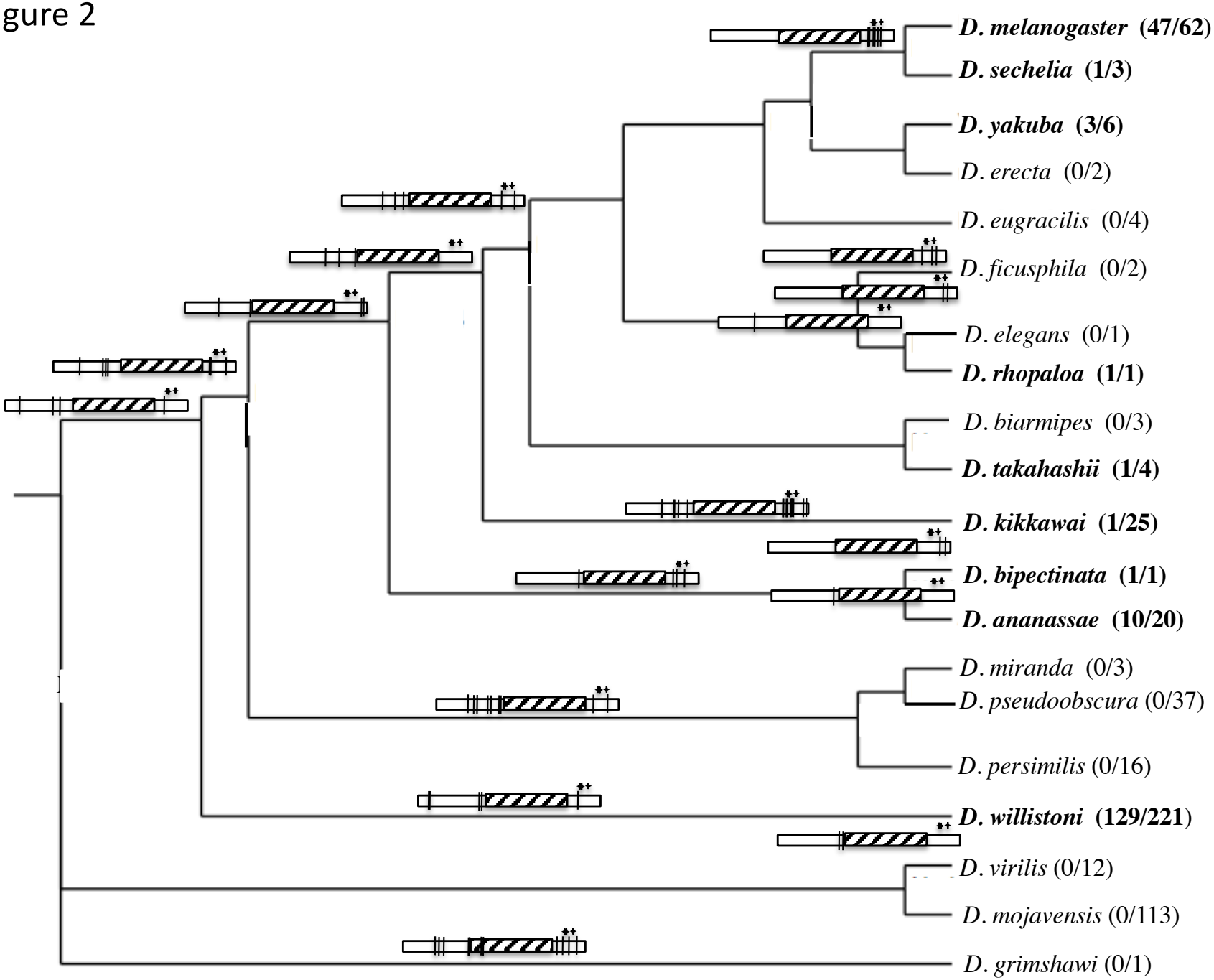
D. melanogaster (47/62)
D. sechelia (1/3)
D. yakuba (3/6)
D. erecta (0/2)
D. eugracilis (0/4)
D. ficusphila (0/2)
D. elegans (0/1)
D. rhopaloa (1/1)
D. biarmipes (0/3)
D. takahashii (1/4)
D. kikkawai (1/25)
D. bipectinata (1/1)
D. ananassae (10/20)
D. miranda (0/3)
D. pseudoobscura (0/37)
D. persimilis (0/16)
D. willistoni (129/221)
D. virilis (0/12)
D. mojavensis (0/113)
D. grimshawi (0/1)

Table 1 Levels of within species variation and divergence between *Drosophila* species and results of tests to detect departures from a neutral site frequency spectrum

	θ	π	Div[c]	Taj-D[a] (*P*-value[b])	FW-H[a] (*P*-value[b])
D. melanogaster (n = 20)					
Synonymous	0.012	0.011	0.069	-0.267	-2.26
Nonsynonymous	0	0	0	(0.214; 0.80; 0.70)	(**<0.0002; <0.0002; <0.0002**)
Intron	0.019	0.019	0.083		
D. simulans (n = 10)					
Synonymous	0.017	0.013	0.069	-1.045	3.47
Nonsynonymous	0	0	0	(**0.001**)	(0.87)
Intron	0.031	0.025	0.083		
D. ananassae (n = 13)					
Synonymous	0.010	0.010	0.209	0.349	-4.333
Nonsynonymous	0.000	0.000	0.003	(0.166)	(0.038)
Intron	0.021	0.025	0.299		
D. pseudoobscura (n = 29)					
Synonymous	0.011	0.007	0.073	-1.293	-0.539
Nonsynonymous	0.0004	0.0001	0.002	(0.038)	(0.311)
Intron	0.014	0.008	0.111		

[a] test statistic when using all sites in the analysis: Tajima's D (Taj-D) and Fay and Wu's H (FW-H)

[b] Proportion of simulated datasets that were equal to or less than (for negative statistics) or equal to or greater than (for positive statistics) our observed test statistic for *Sxl*. These are two-sided tests, for which we do two tests per species resulting in a significant cut-off level of 0.0125 (0.025/2). For *D. melanogaster* the *P*-value was calculated by simulating different demographic scenarios listed in this order in parenthesis: Standard neutral; Bottleneck with exponential growth and a 3 Epoch bottleneck (as described in Materials and Methods). *P*-values significant after multiple testing corrections in bold

[c] Uncorrected divergence between the following species pairs: *D. melanogaster* and *D. simulans*; *D. ananassae* and *D. atripex*; and *D. pseudoobscura* and *D. miranda*

Table 2 Results of the McDonald-Kreitman Test at *Sxl* between 3 different sets of *Drosophila* species.

	Synonymous	Nonsynonymous	*P*-value [a]
Single species polymorphism			
D. ananassae/D. atripex			
polymorphic	8	0	
Fixed divergent	44	2	1.00
D. pseudoobscura/D. miranda			
polymorphic	9	1	
Fixed divergent	14	0	0.417
Two species polymorphism combined			
(D. melanogaster+D. simulans)/D. yakuba			
polymorphic	19	0	
Fixed divergent	19	7	**0.015**

[a] Fisher exact test *P*-value

Table 3 Results of the CF Test at *Sxl* along four *Drosophila* lineages.

	Unpreferred	Preferred	*P*-value [a]	direction of departure
***D. melanogaster* lineage**				
Fixed substitutions	5	0		
Sites	150	31	0.303	
***D. simulans* lineage**				
Fixed substitutions	1	2		
Sites	150	31	0.022	preferred codons favored
***D. ananassae* lineage**				
Fixed substitutions	6	11		
Sites	174	29	<0.001	preferred codons favored
***D. pseudoobscura* lineage**				
Fixed substitutions	1	3		
Sites	154	31	0.002	preferred codons favored

[a] Fisher exact test *P*-value