

High divergence among *Drosophila simulans* mitochondrial haplogroups arose in midst of long term purifying selection

Matthew D. Dean*, J. William O. Ballard

University of Iowa, 208 Biology Building, Iowa City, IA 52242, USA

Received 18 September 2004; revised 9 February 2005

Available online 2 April 2005

Abstract

We characterize the type of selection acting within and among mitochondrial lineages in five closely related *Drosophila* species. We focus on *D. simulans*, where three genetically distinct mitochondrial haplogroups show high interhaplogroup divergence and low intrahaplogroup polymorphism. Using maximum likelihood models we find that the branches leading to these three distinct mitochondrial groups show a significantly reduced rate of nonsynonymous relative to synonymous substitution. This interhaplogroup rate is significantly reduced compared to the intrahaplogroup rate, and closely resembles the rate observed between distinct species. The data suggest that slightly deleterious mutations segregating within *D. simulans* haplogroups are removed by selection prior to their fixation among haplogroups. We explore several hypotheses to explain how lineages within a single species can be compatible with this model of slightly deleterious mutation. The most likely hypothesis is that *D. simulans* haplogroups have persisted in isolation, perhaps due to association with the bacterial symbiont *Wolbachia* and/or demographic history, introducing a bias against the fixation of slightly deleterious mutations.

© 2005 Elsevier Inc. All rights reserved.

Keywords: *Drosophila*; Simulans; Mitochondria; Slightly deleterious mutations

1. Introduction

Detecting shifts in the rate of nonsynonymous substitution has provided important insights into the evolutionary process (Gillespie, 1991; Kimura, 1983; Li, 1997). If the majority of nonsynonymous substitutions are slightly deleterious, they may segregate within a population but are removed by selection prior to fixation among species (Hasegawa et al., 1998; Nachman, 1998; Ohta, 1992). In contrast, if most nonsynonymous substitutions are selectively advantageous, then their fixation among species may be accelerated by selection (Aguadé, 1999; Begun and Whitley, 2000; Clark and

Wang, 1997; Coulthart and Singh, 1988; Schlenke and Begun, 2003; Swanson et al., 2001; Swanson and Vacquier, 1998).

In this study, we characterized the type of selection acting within and among mitochondrial lineages in five closely related *Drosophila* species, an important step towards understanding the discrepancies between polymorphism and divergence among these haplogroups. We focused on the three haplogroups harbored by *D. simulans* (siI, -II, and -III) which show average pairwise nucleotide and amino acid divergence on the order of 10^{-2} (Ballard, 2000a; Solignac et al., 1986). This amount of divergence approaches interspecific comparisons to the closely related *D. melanogaster* (Fig. 1). In contrast, there is a significant reduction of polymorphism within each *D. simulans* haplogroup compared to nuclear loci (Ballard, 2000a; Ballard et al., 2002). *D. melanogaster* also shows a significant reduction in mitochondrial

* Corresponding author. Present address: University of Arizona, 1041 East Lowell Avenue, Tucson, AZ 85741, USA. Fax: +1 520 621 9190.

E-mail address: mattdean@email.arizona.edu (M.D. Dean).

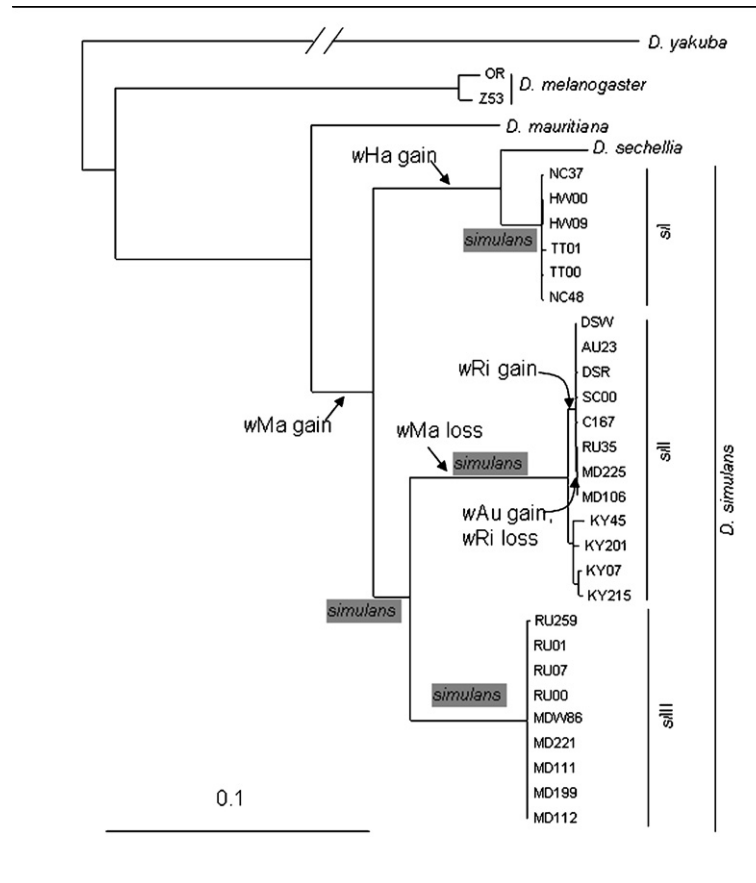


Fig. 1. Genealogical relationships inferred by Ballard (2004). Branches labeled “simulans” indicate those for which an additional ω was estimated under the simulans-specific model (see Section 2). Branch lengths drawn according to the number of nucleotide substitutions per codon, relative to scale bar shown. Branch leading to *D. yakuba* was truncated as there were ~ 0.60 substitutions per codon along this branch. Arrows indicate where the four strains of *Wolbachia* that infect *D. simulans* were either gained or lost. Although this phylogeny shows a gain and a loss of wRi, it should be noted that a single gain of wRi (no loss) cannot be rejected (Ballard, 2004).

polymorphism relative to divergence, while *D. yakuba* does not (Ballard et al., 1996).

Estimating the number of nonsynonymous substitutions per nonsynonymous site (dN) relative to the number of synonymous substitutions per synonymous site (dS) has proven an effective means to study selection (Yang and Bielawski, 2000). Under a neutral model, the null hypothesis is that the ratio of dN/dS (ω) is equal to one. An ω significantly less than one suggests purifying selection suppressed the fixation of nonsynonymous relative to synonymous substitutions over time. In contrast, an ω significantly greater than one suggests recurrent positive selection accelerated fixation of nonsynonymous substitutions over time.

These simple predictions assume that ω has remained constant among lineages, which in the case of these five *Drosophila* species spans some 5–10 million years (Li et al., 1999). By estimating separate ω 's for interspecific and intraspecific lineages, we partially relaxed this assumption and tested more specific hypotheses of selection. A hypothesis of slightly deleterious mutations predicts a higher intraspecific compared to interspecific ω

because selection removes nonsynonymous mutations segregating within a species prior to their fixation among species (Hasegawa et al., 1998; Ohta, 1992). Alternatively, a hypothesis of recurrent positive selection predicts accelerated fixation of nonsynonymous substitutions along interspecific, compared to intraspecific, lineages. Under either hypothesis, neutral mutations will continue to fix according to the neutral mutation rate, independently of linkage to strongly selected sites (Birky and Walsh, 1988; Gillespie, 2000). However, nearly neutral substitutions may experience relaxed selective constraint and fix more frequently than without linkage (Comeron and Kreitman, 1998).

The mechanism leading to large divergence among *D. simulans* haplogroups has remained a mystery ever since Solignac et al. (1986) first described the three distinct mitochondrial types. Until the current study, the hypothesis of repeated positive selection acting to accelerate interhaplogroup divergence has never been evaluated. Here, we rejected this hypothesis, showing instead strong evidence for a long history of purifying selection acting on lineages leading to these three haplo-

groups. High divergence among mitochondrial haplogroups is not predicted under a regime of long term purifying selection. However, the intimate association between distinct *D. simulans* mitochondria and strains of the bacterial symbiont *Wolbachia* may offer reconciliation between long term purifying selection and high interhaplogroup divergence.

The most common phenotype induced by *Wolbachia* is cytoplasmic incompatibility. In the simplest case, incompatibility occurs when an uninfected female mates with an infected male, causing a reduction in the egg hatch rate (reviewed by Hoffmann and Turelli, 1997). All other crossing types are compatible, giving a reproductive advantage to infected females which successfully reproduce with either infected or uninfected males. In a population of intermediate infection frequency, infected females quickly replace uninfected females in a few generations. Such *Wolbachia*-induced population sweeps have been observed theoretically (Caspari and Watson, 1959), in population cages (Kambhampati et al., 1992; Nigro and Prout, 1990), and in nature (Turelli and Hoffmann, 1991). During a *Wolbachia*-induced population sweep, all maternally inherited factors initially associated with infected females will increase in frequency. For this reason, mitochondrial polymorphism decreases in a way analogous to genetic hitchhiking (sensu Maynard Smith and Haigh, 1974). Interestingly, population dynamics following a *Wolbachia*-induced population sweep have not been explored in detail. Here we hypothesized that following a sweep, mitochondrial haplogroups remained isolated. Slightly deleterious mutations segregating within mitochondrial haplogroups were effectively removed by purifying selection prior to their fixation among haplogroups. This model of slightly deleterious mutation was originally applied to within and between species comparisons (Ohta, 1992). The low polymorphism observed within each mitochondrial haplogroup is consistent with a *Wolbachia*-induced sweep, and may be decreased even more through multiple sweep-recovery cycles (Ballard et al., 1996).

2. Materials and methods

2.1. DNA sequences used

We employed the mitochondrial genealogy proposed by Ballard (2004) to characterize the type of selection acting on the mitochondrial genome. That genealogy utilized complete mitochondrial genomes from one *D. yakuba* (Clary et al., 1982), two *D. melanogaster*, one *D. mauritiana* (haplogroup *maII*), 27 *D. simulans* (six *siI*, 12 *siII*, and nine *siIII*), and one *D. sechellia*. The *maI* haplogroup from *D. mauritiana* probably resulted from very recent introgression of the *D. simulans* *siIII* haplogroup into *D. mauritiana* (Ballard, 2000c) and

was not included. Of the 27 *D. simulans* genomes, 22 were from Ballard (2000a) and five from Ballard (2004). Four of these latter five *siII* came from a population with significantly higher mtDNA polymorphism than all previous investigations of *D. simulans* (Dean et al., 2003). From all genomes, the 13 protein coding regions in *Drosophila* were analyzed. Nonprotein-coding regions, including intervening spacer regions, tRNAs, and rRNAs (but not the unsequenced A + T-rich region), were used to assess background base composition.

2.2. Codon bias

Codon-based maximum likelihood models were used to investigate the role that selection has played throughout the mtDNA genealogy. These methods correct for codon usage bias, transition:transversion bias (κ), and the occurrence of multiple substitutions by taking a probabilistic approach to estimating ω (Yang and Bielawski, 2000). Changes between codons *i* and *j* occur at the instantaneous rate

$$q_{ij} = \begin{cases} 0 & \text{if the two codons differ at more than} \\ & \text{one position,} \\ \pi_j & \text{for synonymous transversion,} \\ \kappa\pi_j & \text{for synonymous transition,} \\ \omega\pi_j & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j & \text{for nonsynonymous transition,} \end{cases}$$

where π_j is the equilibrium frequency of codon *j*, κ is the transition:transversion rate ratio, and $\omega = dN/dS$. The equilibrium frequency of each codon was estimated using the observed nucleotide frequencies at the three codon positions (Yang and Nielsen, 2000).

Because there is little or no effective recombination in *Drosophila* mtDNA, one might consider calculating ω from the 13 protein coding regions as a pooled group. However, pooling genes assumes that the instantaneous rate matrix is homogeneous among genes. One way this assumption could be violated is if codon bias were heterogeneous among genes.

Using the 27 *D. simulans* taxa, we tested the assumption of homogeneity in codon bias using ENCprime (Novembre, 2002). This program corrects for background base composition, an important consideration given the A + T richness of the mitochondrial genome. Background base composition was calculated using 3926 bp of noncoding mtDNA in the mitochondrial genomes, which included intervening spacer regions, tRNAs, and rRNAs. Theoretically, the effective number of codons ranges from 20 (codon bias so extreme that only one codon is used per amino acid) to 62 (no codon bias where all redundant codons are used at equal frequency, minus the two stop codons in *Drosophila* mtDNA), assuming all amino acids are present at least as many times as their

redundant codons. The effective number of codons was calculated for each gene after excluding initiation and termination sites, then regressed onto gene length to test for heterogeneity among genes. Gene length has been shown to predict codon bias (Comeron et al., 1999; Duret and Mouchiroud, 1999; Eyre-Walker, 1996; Marin et al., 1998; Moriyama and Powell, 1998). Additionally, we tested whether codon bias differed among groups of mitochondrial genes sharing a particular expression level. Information on expression level was gathered indirectly, by considering that mitochondrial gene products differ in stoichiometry (Schägger and Pfeiffer, 2001).

2.3. Lineage-specific hypotheses

After testing for homogeneity in codon bias among mitochondrial genes, four codon-based maximum likelihood models were analyzed using PAML version 3.15 (Yang, 1997). The major differences among the four models are summarized in Table 1. The simplest model tested was the one-ratio-fixed model. In this model, an $\omega = 1$ was enforced along all branches in the mtDNA genealogy (Table 1). This is the expected ratio under a neutral model of evolution (Yang and Bielawski, 2000), so one-ratio-fixed served as the baseline model from which to begin explorations of evolutionary rates.

In the one-ratio-estimated model, we relaxed the constraint that $\omega = 1$. Instead, a single ω that best explained the data was estimated (Table 1). This model differed from the one-ratio-fixed model by one additional parameter, and comparing these two models was akin to asking if the overall ω deviated significantly from unity.

We compared two models to the one-ratio-estimated model to ask whether the lineages leading to *D. simulans* haplogroups (those labeled “simulans” in Fig. 1) were significantly more similar to intraspecific or interspecific rates of evolution. First, we compared the one-ratio-estimated model to an interspecific–intraspecific model, which estimated a single ω for all intraspecific lineages and another ω for all interspecific lineages (Table 1). In this model, we considered the lineages leading to *D. simulans* haplogroups as intraspecific. Because we had only a single representative of *D. mauritiana* and *D. sechellia* (Fig. 1) we included the branches leading to these terminal taxa in the interspecific category. This assignment was reasonable since no variation has been found within *D. sechellia* (Ballard, 2004) or within *D. mauritiana mal*

(unpublished data). Comparing the one-ratio-estimated to the interspecific–intraspecific model was akin to asking whether interspecific and intraspecific lineages showed significantly different rates of evolution. These two models differed by one additional parameter.

Additionally, we compared the one-ratio-estimated model to a simulans-specific model, in which an independent ω was estimated for each lineage leading to *D. simulans* haplogroups, as well as the branch leading from the common ancestor of *siII* + *siIII* to the immediately preceding ancestor (the branches labeled “simulans” in Fig. 1). The simulans-specific model estimated five additional parameters compared to the one-ratio-estimated model (Table 1).

All models were analyzed three times to assure convergence of the maximum likelihood estimation of parameters. The few cases where differences among runs occurred were negligible in magnitude, and interpretation of statistical significance was not affected. Statistical significance was judged using the likelihood ratio test statistic (LR) which is twice the difference in log-likelihoods of two models. This statistic approximately follows a χ^2 distribution, with degrees of freedom equal to the number of additional parameters estimated in the more complex model.

2.4. Site specific hypotheses

The lineage-specific models analyzed above assumed evolutionary forces acting along the length of a sequence were homogeneous. However, selection may act heterogeneously among codons (Nielsen and Yang, 1998; Yang et al., 2000). We employed the M7 and M8 likelihood models in PAML version 3.15 to detect heterogeneity of selection among codons (Yang et al., 2000). The M7 model assigns codons to one of 10 different rate categories, following a β distribution, with the constraint that no rate classes can be greater than $\omega = 1$. This null model can be directly compared to the alternative M8 model (Yang et al., 2000). There are two additional parameters estimated by the M8 model (the additional rate class and the probability that codons belong to that class). In cases where an $\omega > 1$ is detected, PAML uses Bayes theorem to assign probabilities that particular codons belong to particular ω classes (Yang, 2002). Both models were run three times to assure stability. Differences among runs were negligible and did not affect interpretation of statistical significance.

Table 1
The major differences among four lineage-specific maximum likelihood models tested

Model	ω 's estimated: description
One-ratio-fixed	0: (ω fixed at 1)
One-ratio-estimated	1: ω
Interspecific–intraspecific	2: ω_{inter} , ω_{intra}
Simulans-specific	6: ω_{inter} , ω_{intra} , ω_{siI} , ω_{siII} , ω_{siIII} , $\omega_{\text{siII+siIII}}$

3. Results

3.1. Codon bias

Interhaplogroup divergence and intrahaplogroup polymorphism is shown in Table 2. The regression of

Table 2
Interhaplogroup divergence and intrahaplogroup polymorphism

	Nucleotides		Amino acids	
	Total	%	Total	%
Intrahaplogroup				
<i>siI</i>	5.84	0.00039	2.00	0.00054
<i>siII</i>	4.78	0.00032	2.29	0.00062
<i>siIII</i>	0.60	0.00004	0.25	0.00007
Interhaplogroup				
<i>siI</i> vs. <i>siII</i>	392.05	0.02624	47.57	0.01285
<i>siI</i> vs. <i>siIII</i>	345.01	0.02309	26.68	0.00721
<i>siII</i> vs. <i>siIII</i>	305.27	0.02043	45.11	0.01219
Intraspecific				
<i>D. simulans</i>	236.07	0.01580	26.84	0.00725
<i>D. melanogaster</i>	53.06	0.00356	10.00	0.00270
Interspecific				
<i>D. simulans</i> vs. <i>D. melanogaster</i>	605.33	0.04064	89.89	0.02429

the effective number of codons on mitochondrial gene length was significant ($F = 9.95$, residual = 66.39, $df = 1$, $P < 0.01$), rejecting the null hypothesis of homogeneity in codon bias among genes (Fig. 2). This result argued against pooling mitochondrial genes. Therefore, we analyzed the three shortest (less than 150 codons) and 10 longest (greater than 150 codons) genes independently. These two categories formed natural groups with similar codon usage bias (Fig. 2), as there was not a significant regression within either group ($F = 1.26$, 0.53, $df = 1$, $P \approx 0.46$, 0.49, respectively). After pooling, the short genes combined for 485 codons while the long genes combined for 3398 codons.

While gene length has been shown to predict codon bias (Comeron et al., 1999; Duret and Mouchiroud, 1999; Eyre-Walker, 1996; Marin et al., 1998; Moriyama and Powell, 1998), gene expression may also influence codon bias (Gouy and Gautier, 1982; Ikemura, 1985). Mitochondrial protein products participate in four of five complexes in the electron transport chain and oxidative phosphorylation (reviewed in Scheffler, 1999). The products of these four complexes differ in stoichiometric levels (Schägger and Pfeiffer, 2001), suggesting differential gene expression among complexes. We therefore tested for differences in effective number of codons among the four mitochondrial gene complexes. Complex I includes the genes ND1, ND2, ND3, ND4, ND4L, ND5, and ND6. The genes ND1, ND4, ND4L, and ND5 were analyzed separately because they occur on the negative strand and show very different base composition compared to genes on the positive strand (Ballard, 2000b). Complex II is coded entirely by nuclear genes in *Drosophila* and is therefore not applicable to this analysis. Complex III includes only cytochrome *b*. Complex IV includes COI, COII, and COIII. Complex V includes ATPase 6 and ATPase 8. The effective number of codons did not differ among complexes (Kruskal–Wallis = 2.78, $df = 4$, $P \approx 0.60$), suggesting heterogeneous codon usage bias was not the result of differential gene expression. Pooling complex I genes into a single category, instead of separating them into positive strand and negative strand groups, did not change this result (Kruskal–Wallis = 2.77, $df = 3$, $P \approx 0.43$). The median effective number of codons was 51.33, 52.46, 48.59, 53.09, and 53.61 for Complexes I (positive strand only), I (negative strand only), III, IV, and V, respectively.

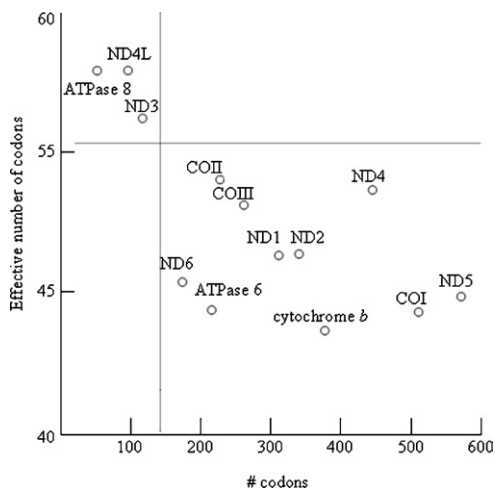


Fig. 2. Effective number of codons versus number of codons in mitochondrial genes. There is a significant correlation when all genes are considered, but not within groups of the three shortest and 10 longest genes (see Section 3).

chondrial protein products participate in four of five complexes in the electron transport chain and oxidative phosphorylation (reviewed in Scheffler, 1999). The products of these four complexes differ in stoichiometric levels (Schägger and Pfeiffer, 2001), suggesting differential gene expression among complexes. We therefore tested for differences in effective number of codons among the four mitochondrial gene complexes. Complex I includes the genes ND1, ND2, ND3, ND4, ND4L, ND5, and ND6. The genes ND1, ND4, ND4L, and ND5 were analyzed separately because they occur on the negative strand and show very different base composition compared to genes on the positive strand (Ballard, 2000b). Complex II is coded entirely by nuclear genes in *Drosophila* and is therefore not applicable to this analysis. Complex III includes only cytochrome *b*. Complex IV includes COI, COII, and COIII. Complex V includes ATPase 6 and ATPase 8. The effective number of codons did not differ among complexes (Kruskal–Wallis = 2.78, $df = 4$, $P \approx 0.60$), suggesting heterogeneous codon usage bias was not the result of differential gene expression. Pooling complex I genes into a single category, instead of separating them into positive strand and negative strand groups, did not change this result (Kruskal–Wallis = 2.77, $df = 3$, $P \approx 0.43$). The median effective number of codons was 51.33, 52.46, 48.59, 53.09, and 53.61 for Complexes I (positive strand only), I (negative strand only), III, IV, and V, respectively.

3.2. Lineage-specific hypotheses

The one-ratio-estimated model fit the data significantly better than the one-ratio-fixed model for both the short and long gene groups (Tables 3 and 4), revealing an ω significantly less than 1 (LR = 385.10 and 5767.90, $df = 1$, $P < 0.001$ for the short and long genes, respectively). The one-ratio-estimated model estimated $\omega = 0.0115$ and 0.0124 for the short and long genes, respectively.

Table 3
Likelihood scores and estimated parameters from four lineage-specific likelihood models

Gene group ^a	<i>L</i>	κ	ω	ω_{intra}	ω_{inter}	ω_{siI}	ω_{siII}	ω_{siIII}	$\omega_{\text{siII+siIII}}$
<i>One-ratio-fixed</i>									
Short genes	−1648.33	8.85	1.00						
Long genes	−23383.80	5.75	1.00						
<i>One-ratio-estimated</i>									
Short genes	−1455.78	5.65	0.0115						
Long genes	−20499.85	5.01	0.0124						
<i>Interspecific–intraspecific</i>									
Short genes	−1455.18	6.06		0.0183	0.0094				
Long genes	−20498.19	5.10		0.0160	0.0115				
<i>Simulans-specific</i>									
Short genes	−1448.98	6.26		0.0368	0.0115	0.0056	0.0076	0.0127	0.0328
Long genes	−20486.73	5.12		0.0893	0.0093	0.1218	0.0069	0.0001	0.0001

^a Three short genes and 10 long genes (see text and Fig. 2).

Table 4
Likelihood ratio test statistics comparing two models

Gene group ^a	One-ratio-fixed vs. one-ratio-estimated (df ^b = 1)	One-ratio-estimated vs. interspecific–intraspecific (df = 1)	One-ratio-estimated vs. simulans-specific(df = 5)
Short genes	385.10*	1.20	13.60*
Long genes	5767.90*	3.32	26.24*

^a Three short genes and 10 long genes (see text and Fig. 2).

^b df = degrees of freedom.

* $P < 0.05$.

We compared two models, interspecific–intraspecific and simulans-specific, to the one-ratio-estimated model. The interspecific–intraspecific model did not fit the data significantly better than the one-ratio-estimated model (LR = 1.20 and 3.32, df = 1, $P > 0.05$ for the two gene groups, Tables 3 and 4). The intraspecific/interspecific ω 's for the short and long genes, respectively, were estimated as 0.0183/0.0094 and 0.0160/0.0115 (Table 3).

In contrast to the interspecific–intraspecific model, the simulans-specific model explained the data significantly better than the one-ratio-estimated model (LR = 13.60 and 26.24, df = 5, $P < 0.05$ for the two gene groups, Tables 3 and 4). The simulans-specific model showed that in general, the intraspecific rate was greater than the background interspecific rates and the simulans-specific rates (Table 3). The lone

exception to this trend was the high $\omega_{\text{siI}} = 0.1218$ estimated from the long genes, which was greater than the intraspecific rate $\omega = 0.0893$. Therefore, after moving the “simulans” lineages out of the intraspecific category, the intraspecific rate of evolution increased substantially (from 0.0183 and 0.0160 to 0.0368 and 0.0893 for the short and long genes, respectively). This shift suggested that the lineages leading to *D. simulans* haplogroups have a reduced rate of evolution compared to background intraspecific rates.

3.3. Site-specific hypotheses

No evidence for positive selection acting on any codons was found for either gene group (Table 5). Analyses of site-specific models showed selection has acted

Table 5
Likelihood scores and estimated parameters from two site-specific likelihood models

Gene group ^a	<i>L</i>	ω	Estimates of parameters
M7			
Short genes	−1446.96	0.0124	$p = 0.048, q = 2.562$
Long genes	−20341.35	0.0140	$p = 0.039, q = 1.731$
M8			
Short genes	−1446.89	0.0126	$p_0 = 0.960, p_1 = 0.040, p = 0.081, q = 11.229, \omega_2 = 0.190$
Long genes	−20335.91	0.0144	$p_0 = 0.398, p_1 = 0.602, p = 0.030, q = 0.789, \omega_2 = 0.006$

^a Three short genes and 10 long genes (see text and Fig. 2).

heterogeneously among codons in the long genes, but in neither group was there evidence for positive selection acting on any codons. Model M8 did not fit the data significantly better than M7 for short genes (LR = 0.14, df = 2, $P \approx 0.93$), but did for the long genes (LR = 10.88, df = 2, $P < 0.005$), which resulted from estimation of the additional rate class $\omega = 0.006$ (Table 5). Neither gene group showed evidence of rate categories where $\omega > 1$, and zero out of 485 or 3398 codons for the short or long gene groups, respectively, were under positive selection with a posterior probability greater than 0.50.

4. Discussion

Selection has apparently acted upon distinct mitochondrial lineages within *D. simulans* in a way that is characteristic of interspecific lineages. Previously, Ballard (2000a) found a significant excess of nonsynonymous changes and a significant deficiency of synonymous changes within *siI* and *siII* haplogroups. Ballard and Kreitman (1994) argued the *siI* lineage evolved in a manner usually observed among species. This discovery made intuitive sense given *siI* is more closely related to *D. sechellia* than either is to *siII* or *siIII* (Fig. 1). Furthermore, *siI* is the only haplogroup that has never been found in sympatry with another haplogroup (Solignac, 2004, also see <http://www.myweb.uiowa.edu/bballard/map.htm>). With maximum likelihood models, we extended this finding, estimating ω 's for the "simulans" lineages that were significantly reduced and more similar to interspecific lineages (Fig. 1). Our findings suggest slightly deleterious mutations segregating within *D. simulans* haplogroups were removed by selection prior to their fixation among haplogroups (Hasegawa et al., 1998; Ohta, 1992). No evidence was found to suggest any codons experienced recurrent positive selection on this background of purifying selection.

One plausible hypothesis is that mitochondrial haplogroups represent isolated lineages, so that the model of slightly deleterious mutations applies to the three "simulans" lineages just as if they were distinct species. Such isolation may have arisen by association with distinct strains of *Wolbachia* and/or historical demographic events. Association with distinct strains of *Wolbachia* provides a mechanism which makes this hypothesis viable. In *D. simulans*, each mitochondrial haplogroup is associated with a genetically distinct strain of *Wolbachia* (Ballard, 2004). The *siI* haplogroup may be infected with *wHa*, *wMa*, or both simultaneously. The *siII* haplogroup may be infected with *wRi* or *wAu*, but never both. The *siIII* haplogroup may be infected with *wMa*. These strains differ by up to 12% at the highly variable *wsp* locus (Zhou et al., 1998), but some are invariant at the more conserved 16S (O'Neill et al., 1992). A likely recon-

struction of the gains and losses of these strains is presented in Fig. 1 (Ballard, 2004).

As mentioned previously, *Wolbachia*-induced population sweeps result in a reduction of mitochondrial polymorphism, in a way analogous to genetic hitchhiking (Hoffmann and Turelli, 1997; Turelli et al., 1992). Following a *Wolbachia*-induced population sweep, the associated mitochondrial type will persist in a population indefinitely. Incoming migrants will be selected against unless they are infected with the same strain of *Wolbachia*, in which case they are likely to carry the same mitochondrial variant. This may lead to long term persistence of the mitochondrial haplotype associated with the infection since it is largely removed from competition with individuals carrying other mitochondrial haplotypes. For example, it has been shown that the *siI* haplogroup is competitively inferior in population cage experiments (Ballard and James, 2004) yet it remains fixed in Hawaii and Tahiti, perhaps due to association with *wHa* which induces strong incompatibility. Long term persistence of a mitochondrial haplotype may in turn bias against the fixation of slightly deleterious mutations among haplogroups since such mutations must also persist an irregularly long time in a population. Our results contrast with patterns found in *D. recens*, where an elevated dN/dS was associated with *Wolbachia* infection (Shoemaker et al., 2004).

Interestingly, the association between distinct mitochondrial haplogroups and *Wolbachia* strains may also explain the high divergence among *D. simulans* haplogroups. Independent *Wolbachia*-induced population sweeps may have eliminated once abundant mitochondrial polymorphism, resulting in the preservation of three mitochondrial haplogroups. Based on studies of nuclear genes (Andolfatto, 2001; Dean and Ballard, 2004), it is feasible that an ancestral population of *D. simulans* held mitochondrial diversity on the order of 2%. Incoming *Wolbachia* strains may have randomly associated with mitochondrial variants averaging 2% genetic difference. After independent *Wolbachia*-induced population sweeps removed intermediate variation, *D. simulans* would then hold three distinct mitochondrial haplogroups.

An obvious question is whether the original association between distinct mitochondrial haplogroups and *Wolbachia* strains arose by chance alone or if there was some functional reason for the particular associations. The latter hypothesis predicts a mutually increased rate of evolution (Erlich and Raven, 1964; Frank, 2000), especially if *Wolbachia* are parasitic. Our analyses of maximum likelihood models above showed no evidence for an increase in evolutionary rate, and no evidence for any codons experiencing repeated positive selection. These patterns argue that the original association between distinct mitochondrial haplogroups and distinct *Wolbachia* strains arose by chance alone.

The hypothesis that *Wolbachia*-induced sweeps isolate mitochondrial lineages and lead to a decreased rate of interhaplogroup evolution must be reconciled with the pattern of gains and losses of particular strains (Fig. 1). As shown, the branch leading from the common ancestor of all *siII* to the common ancestor of all *siII* + *siIII* was probably uninfected since no *siII* are infected with *wMa*, while *siI* and *siIII* are regularly infected with *wMa*. Nevertheless, our analyses above identified a significantly reduced rate along this branch. At first impression, this may seem to contradict the hypothesis that *Wolbachia* infections caused long term isolation, but we cannot be sure when the *wMa* infection was lost along this branch. If *wMa* was lost nearer the common ancestor of all *siII*, then these haplogroups may have been historically isolated from other mitochondrial lineages. Even though the *wMa* strain of *Wolbachia* does not induce strong incompatibility in the laboratory (James and Ballard, 2000), it may have persisted through other means such as fitness benefits to the host (Hoffmann and Turelli, 1997) and lead to long term persistence of the associated mitochondrial type. In addition, it may have induced incompatibility in the past.

Another pattern that requires reconciliation is the high genetic diversity in the basal *siII* group of KY45, KY201, KY07, and KY215 (Fig. 1). Dean et al. (2003) showed these came from a population with unusually high mitochondrial diversity, and that *Wolbachia*-infected individuals were effectively absent. From that study, we might expect the branch leading to these four taxa to show evidence of relaxed purifying selection, since their mitochondrial haplogroups have apparently not been genetically isolated. However, a model that estimated an additional ω along the branch leading to this basal *siII* group did not fit the data significantly better compared to the *simulans*-specific model for either gene group.

Population dynamics following a *Wolbachia*-induced sweep have not been well studied. Theoretical work suggests that a *Wolbachia* variant may re-sweep through an already infected population if acquired mutations allow for stronger incompatibility, lower fitness costs to their hosts, and/or higher maternal transmission fidelity (Turelli, 1994). Empirical evidence suggests sweeps of mitochondrial variation have occurred in both infected and uninfected populations. Ballard et al. (1996) showed mitochondrial variability was significantly reduced (relative to divergence among haplogroups) compared to a noncoding region of a nuclear gene, in both uninfected and infected flies. They suggested the dramatic reduction in mitochondrial diversity may be related to *Wolbachia* or the fixation of an advantageous mitochondrial mutation. In this study we reject the latter hypothesis as a general explanation.

Another way multiple *Wolbachia*-induced sweeps may occur is through natural fluctuation of infection

frequency. Founder effects may reduce infection frequency, making that population susceptible to another *Wolbachia*-induced population sweep in the future. Similarly, selection may favor a host that can clear *Wolbachia* infections or can minimize the strength of incompatibility, which would make the population susceptible to another *Wolbachia*-induced population sweep (Hurst and McVean, 1996; Turelli, 1994). Weakened incompatibility has been observed in several species (Bourtzis et al., 1998; Charlat et al., 2002; Giordano et al., 1995; Merçot and Poinso, 1998; Rousset and Solignac, 1995; Turelli and Hoffmann, 1995).

Several alternative hypotheses are now discussed that seem less likely to explain why the lineages leading to *D. simulans* haplogroups showed a significantly reduced rate of evolution compared to the background intraspecific rate. One hypothesis is that the three *D. simulans* mitochondrial haplogroups represent cryptic species, thus showing a rate of evolution more similar to the interspecific rate. This hypothesis was rejected by Ballard et al. (2002), who showed that neither nuclear DNA data, mate choice, nor genitalic shape showed any evidence of organismal subdivision similar to mitochondrial subdivision.

A second alternative hypothesis is that the mitochondrial haplogroups diverged in other species and were horizontally transferred into *D. simulans*, carrying with them the signature of interspecific selection. This hypothesis seems unlikely. While *D. mauritiana* shares a haplotype identical to the canonical *siIII*, Ballard (2000c) argued that this haplotype was transferred from *D. simulans* to *D. mauritiana*. The only viable hybrids between these two species result when a male *D. mauritiana* mates with a female *D. simulans*. It is unlikely that unsampled taxa would change this interpretation since this group is well studied.

A third alternative hypothesis is that each haplogroup has a very large effective population size relative to the mutation rate, and that selection is exceedingly efficient at removing slightly deleterious mutations prior to fixation among haplogroups. This hypothesis also seems unlikely given that *siI* and *siIII* have only been found in small geographic regions (Ballard, 2004; Montchamp-Moreau et al., 1991; Solignac, 2004). Over 95% of the range of *D. simulans* is inhabited by *siII*-carrying flies, so *siII* probably has the largest effective population size relative to *siI* and *siIII*. Yet all three lineages show a characteristic reduction in ω . In order for this hypothesis to remain viable, one must posit that historic effective population sizes were much larger than present and that they were relatively equal in size throughout most of their history. This is an unlikely ad hoc justification given that *D. simulans* is a human commensal that probably expanded from an ancestral origin in Madagascar (Dean and Ballard, 2004; Lachaise et al., 1988; Lachaise and Silvain, 2004).

In summary, the lineages leading to *D. simulans* haplogroups have experienced a significantly reduced rate of molecular evolution, as measured by the rate of nonsynonymous relative to synonymous substitutions. Association with the bacterial symbiont *Wolbachia* may have caused associated mitochondrial types to persist an irregularly long time in isolation, and demographic history may also have lead to isolation. Long term isolation may bias against the fixation of slightly deleterious mutations segregating within *D. simulans* haplogroups, since such mutations must also persist an irregularly long period of time. Future investigations should compare patterns found here to species groups that have not been affected by *Wolbachia*.

Acknowledgments

Funding was provided by National Science Foundation Grants DEB-9702824 and DEB-0296086 awarded to J.W.O.B., and a Lester Armour Fellowship awarded to M.D.D.

References

- Aguadé, M., 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. *Genetics* 152, 543–551.
- Andolfatto, P., 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. *Mol. Biol. Evol.* 18, 279–290.
- Ballard, J.W.O., James, A.C., 2004. Differential fitness of mitochondrial DNA in perturbation cage studies correlates with global abundance and population history in *Drosophila simulans*. *Proc. R. Soc. Lond. B* 271, 1197–1201.
- Ballard, J.W.O., 2000a. Comparative genomics of mitochondrial DNA in *Drosophila simulans*. *J. Mol. Evol.* 51, 64–75.
- Ballard, J.W.O., 2000b. Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. *J. Mol. Evol.* 51, 48–63.
- Ballard, J.W.O., 2000c. When one is not enough: introgression of mitochondrial DNA in *Drosophila*. *Mol. Biol. Evol.* 17, 1126–1130.
- Ballard, J.W.O., 2004. Sequential evolution of a symbiont inferred from the host: *Wolbachia* and *Drosophila simulans*. *Mol. Biol. Evol.*
- Ballard, J.W.O., Chernoff, B., James, A.C., 2002. Divergence of mitochondrial DNA is not corroborated by nuclear DNA, morphology, or behavior in *Drosophila simulans*. *Evolution* 56, 527–545.
- Ballard, J.W.O., Hatzidakis, J., Karr, T.L., Kreitman, M., 1996. Reduced variation in *Drosophila simulans* mitochondrial DNA. *Genetics* 144, 1519–1528.
- Ballard, J.W.O., Kreitman, M., 1994. Unraveling selection in the mitochondrial genome of *Drosophila*. *Genetics* 138, 757–772.
- Begun, D.J., Whitley, P., 2000. Adaptive evolution of relish, a *Drosophila* NF- κ B/I κ B protein. *Genetics* 154, 1231–1238.
- Birky Jr., C.W., Walsh, J.B., 1988. Effects of linkage on rates of molecular evolution. *Proc. Natl. Acad. Sci. USA* 85, 6414–6418.
- Bourtzis, K., Dobson, S.L., Braig, H.R., O'Neill, S.L., 1998. Rescuing *Wolbachia* have been overlooked. *Nature* 391, 852–853.
- Caspari, E., Watson, G.S., 1959. On the evolutionary importance of cytoplasmic sterility in mosquitoes. *Evolution* 13, 568–570.
- Charlat, S., Nirgianaki, A., Bourtzis, K., Mercot, H., 2002. Evolution of *Wolbachia*-induced cytoplasmic incompatibility in *Drosophila simulans* and *D. sechellia*. *Evolution* 56, 1735–1742.
- Clark, A.G., Wang, L., 1997. Molecular population genetics of *Drosophila* immune system genes. *Genetics* 147, 713–724.
- Clary, D.O., Goddard, J.M., Martin, S.C., Fauron, C.M., Wolstenholme, D.R., 1982. *Drosophila* mitochondrial DNA: a novel gene order. *Nucleic Acids Res.* 10, 6619–6637.
- Comeron, J.M., Kreitman, M., 1998. The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: mutation, selection or relaxed constraints?. *Genetics* 150, 767–775.
- Comeron, J.M., Kreitman, M., Aguadé, M., 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. *Genetics* 151, 239–249.
- Coulthart, M.B., Singh, R.S., 1988. High level of divergence of male-reproductive-tract proteins, between *Drosophila melanogaster* and its sibling species, *D. simulans*. *Mol. Biol. Evol.* 5, 182–191.
- Dean, M.D., Ballard, J.W.O., 2004. Linking phylogenetics with population genetics to reconstruct the geographic origin of a species. *Mol. Phylogenet. Evol.* 32, 998–1009.
- Dean, M.D., Ballard, K.J., Glass, A., Ballard, J.W.O., 2003. Influence of two *Wolbachia* strains on population structure of east African *Drosophila simulans*. *Genetics* 165, 1959–1969.
- Duret, L., Mouchiroud, D., 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* 96, 4482–4487.
- Erlich, P.R., Raven, P.H., 1964. Butterflies and plants: a study in coevolution. *Evolution* 18, 586–608.
- Eyre-Walker, A., 1996. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy?. *Mol. Biol. Evol.* 13, 864–872.
- Frank, S.A., 2000. Polymorphism of attack and defense. *Trends Ecol. Evol.* 16, 167–171.
- Gillespie, J.H., 1991. The Causes of Molecular Evolution. Oxford University Press, Oxford.
- Gillespie, J.H., 2000. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* 155, 909–919.
- Giordano, R., O'Neill, S.L., Robertson, H.M., 1995. *Wolbachia* infections and the expression of cytoplasmic incompatibility in *Drosophila sechellia* and *D. mauritiana*. *Genetics* 140, 1307–1317.
- Gouy, G., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10, 7055–7074.
- Hasegawa, M., Cao, Y., Yang, Z., 1998. Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. *Mol. Biol. Evol.* 15, 1499–1505.
- Hoffmann, A.A., Turelli, M., 1997. Cytoplasmic incompatibility in insects. In: O'Neill, S.L., Hoffmann, A.A., Werren, J.H. (Eds.), *Influential Passengers*. Oxford, New York, Tokyo, pp. 42–80.
- Hurst, L.D., McVean, G.T., 1996. Clade selection, reversible evolution and the persistence of selfish elements: the evolutionary dynamics of cytoplasmic incompatibility. *Proc. R. Soc. Lond. B* 263, 97–104.
- Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2, 13–24.
- James, A.C., Ballard, J.W.O., 2000. Expression of cytoplasmic incompatibility in *Drosophila simulans* and its impact on infection frequencies and distribution of *Wolbachia pipientis*. *Evolution* 54, 1661–1672.
- Kambhampati, S., Rai, K.S., Verleye, D.M., 1992. Frequencies of mitochondrial DNA haplotypes in laboratory cage populations of the mosquito, *Aedes albopictus*. *Genetics* 132, 205–209.
- Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.
- Lachaise, D., Cariou, M.L., David, J.R., Lemeunier, F., Tsacas, L., Ashburner, M., 1988. Historical biogeography of the *Drosophila melanogaster* species subgroup. *Evol. Biol.* 22, 159–225.

- Lachaise, D., Silvain, J.-F., 2004. How two afro-tropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster*-*D. simulans* paleogeographic riddle. *Genetica* 120, 17–39.
- Li, W.-H., 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, MA.
- Li, Y.-J., Satta, Y., Takahata, N., 1999. Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method. *Genes Genet. Syst.* 74, 117–127.
- Marin, A., Gonzalez, F., Gutierrez, G., Oliver, J.L., 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.* 26, 4540.
- Maynard Smith, J., Haigh, J., 1974. The hitchhiking effect of a favorable gene. *Genet. Res.* 23, 23–25.
- Merçot, H., Poinot, D., 1998. ...and discovered on Mount Kilimanjaro. *Nature* 391, 853.
- Montchamp-Moreau, C., Ferveur, J.-F., Jacques, M., 1991. Geographic distribution and inheritance of three cytoplasmic incompatibility types in *Drosophila simulans*. *Genetics* 129, 399–407.
- Moriyama, E.N., Powell, J.R., 1998. Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res.* 26, 3188–3193.
- Nachman, M.W., 1998. Deleterious mutations in animal mitochondrial DNA. *Genetica* 102–103, 61–69.
- Nielsen, R., Yang, Z., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936.
- Nigro, L., Prout, T., 1990. Is there selection on RFLP differences in mitochondrial DNA? *Genetics* 125, 551–555.
- Novembre, J.A., 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Mol. Biol. Evol.* 19, 1390–1394.
- Ohta, T., 1992. The nearly neutral theory of molecular evolution. *Ann. Rev. Ecol. Syst.* 23, 263–286.
- O'Neill, S.L., Giordano, R., Colbert, A.M., Karr, T.L., Robertson, H.M., 1992. 16S rRNA phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. *Proc. Natl. Acad. Sci. USA* 89, 2699–2702.
- Rousset, F., Solignac, M., 1995. Evolution of single and double *Wolbachia* symbioses during speciation in the *Drosophila simulans* complex. *Proc. Natl. Acad. Sci. USA* 92, 6389–6393.
- Schägger, H., Pfeiffer, K., 2001. The ratio of oxidative phosphorylation complexes I–V in bovine heart mitochondria and the composition of respiratory chain supercomplexes. *J. Biol. Chem.* 276, 37861–37867.
- Scheffler, I.E., 1999. *Mitochondria*. Wiley, New York.
- Schlenke, T.A., Begun, D.J., 2003. Natural selection drives the evolution of the *Drosophila* immune system. *Genetics* 164, 1471–1480.
- Shoemaker, D.D., Dyer, K.A., Ahrens, M., McAbee, K., Jaenike, J., 2004. Decreased diversity but increased substitution rate in host mtDNA as a consequence of *Wolbachia* endosymbiont infection. *Genetics* 168, 2049–2058.
- Solignac, M., 2004. Mitochondrial DNA in the *Drosophila melanogaster* complex. *Genetica* 120, 41–50.
- Solignac, M., Monnerot, M., Mounolou, J.C., 1986. Mitochondrial DNA evolution in the *melanogaster* species subgroup of *Drosophila*. *J. Mol. Evol.* 23, 31–40.
- Swanson, W.J., Aquadro, C.F., Vacquier, V.D., 2001. Polymorphism in abalone fertilization proteins is consistent with the neutral evolution of the egg's receptor for lysin (VERL) and positive Darwinian selection of sperm lysin. *Mol. Biol. Evol.* 18, 376–383.
- Swanson, W.J., Vacquier, V.D., 1998. Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. *Science* 281, 710–712.
- Turelli, M., 1994. Evolution of incompatibility-inducing microbes and their hosts. *Evolution* 48, 1500–1513.
- Turelli, M., Hoffmann, A.A., 1991. Rapid spread of an inherited incompatibility factor in California *Drosophila*. *Nature* 353, 440–442.
- Turelli, M., Hoffmann, A.A., 1995. Cytoplasmic incompatibility in *Drosophila simulans*: dynamics and parameter estimates from natural populations. *Genetics* 140, 1319–1338.
- Turelli, M., Hoffmann, A.A., McKechnie, S.W., 1992. Dynamics of cytoplasmic incompatibility and mtDNA variation in natural *Drosophila simulans* populations. *Genetics* 132, 713–723.
- Yang, Z., 1997. PAML, a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13, 555–556.
- Yang, Z., 2002. Inference of selection from multiple species alignments. *Curr. Opin. Genet. Dev.* 12, 688–694.
- Yang, Z., Bielawski, J.P., 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15, 496–503.
- Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43.
- Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.-M.K., 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–449.
- Zhou, W., Rousset, F., O'Neill, S.L., 1998. Phylogeny and PCR-based classification of *Wolbachia* strains using *wsp* gene sequences. *Proc. R. Soc. Lond. B* 265, 509–515.