Published in final edited form as:

J Mol Evol. 2009 July ; 69(1): 94–114. doi:10.1007/s00239-009-9255-0.

Comparative Genomics of *Drosophila* mtDNA: Novel Features of Conservation and Change Across Functional Domains and Lineages

Kristi L. Montooth,

Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA

Dawn N. Abt,

Department of Ecology & Evolutionary Biology, Brown University, Providence, RI 02912, USA

Jeffrey W. Hofmann, and

Department of Ecology & Evolutionary Biology, Brown University, Providence, RI 02912, USA

David M. Rand

Department of Ecology & Evolutionary Biology, Brown University, Providence, RI 02912, USA

Kristi L. Montooth: montooth@indiana.edu; David M. Rand: David_Rand@brown.edu

Abstract

To gain insight on mitochondrial DNA (mtDNA) evolution, we assembled and analyzed the mitochondrial genomes of *Drosophila erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis* and *D. grimshawi* together with the sequenced mtDNAs of the *melanogaster* subgroup. Genomic comparisons across the well-defined *Drosophila* phylogeny impart power for detecting conserved mtDNA regions that maintain metabolic function and regions that evolve uniquely on lineages. Evolutionary rate varies across intergenic regions of the mtDNA. Rapidly evolving intergenic regions harbor the majority of mitochondrial indel divergence. In contrast, patterns of nearly perfect conservation within intergenic regions reveal a refined set of nucleotides underlying the binding of transcription termination factors. Sequencing of 5' cDNA ends indicates that *cytochrome C oxidase I (CoI)* has a novel (T/C)CG start codon and that perfectly conserved regions upstream of two NADH dehydrogenase (ND) genes are transcribed and likely extend these protein sequences. Substitutions at synonymous sites in the *Drosophila* mitochondrial proteomes reflect a mutation process that is biased toward A and T nucleotides and differs between mtDNA strands. Differences in codon usage bias across genes reveal that weak selection at silent sites may offset the mutation bias. The mutation-selection balance at synonymous sites has also diverged between the *Drosophila* and *Sophophora* lineages. Rates of evolution are highly heterogeneous across the mitochondrial proteome, with ND accumulating many more amino acid substitutions than CO. These oxidative phosphorylation complex-specific rates of evolution vary across lineages and may reflect physiological and ecological change across the *Drosophila* phylogeny.

Correspondence to: Kristi L. Montooth, montooth@indiana.edu.

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. EU189427-189434 and TPA: BK006335-006341.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-009-9255-0) contains supplementary material, which is available to authorized users.

Keywords

Comparative genomics; Codon bias; *Drosophila*; mtDNA; Oxidative phosphorylation; tRNA evolution

Introduction

Animal mitochondrial DNA (mtDNA) has maintained its critical role in aerobic metabolism across long evolutionary timescales in the context of dramatic change in the nuclear genome and in organismal physiology and ecology (Boore 1999; Saccone et al. 1999). The preservation of mitochondrial function, gene content and regulatory elements occurs despite a high mutation rate (Brown et al. 1982; Moriyama and Powell 1997), reduced effective population size (N_e), and little or no recombination that makes the mtDNA particularly prone to mutation accumulation (Muller 1964; Felsenstein 1974; Lynch 1996; Neiman and Taylor 2009). While molecular evolution and population genetic data suggest that the mtDNA does not evolve neutrally (reviewed in Meiklejohn et al. 2007), the unique genetics of the mitochondrion and its intimate interaction with the nuclear genome and the cytoplasm provide a challenge for determining the balance of selective pressures shaping the evolution of this small, but complex, genome.

Mutation accumulation experiments reveal that greater than 95% of *Drosophila melanogaster* mutations in mitochondrial protein-coding genes are non-synonymous (Haag-Liautard et al. 2008). However, the ratio of non-synonymous to synonymous differences between the mtDNA of two strains of *D. melanogaster* is only 10:36 and across the *D. melanogaster* subgroup is approximately 1:5 (Ballard 2000). This indicates a strong role for purifying selection in removing the majority of mutations that arise in the *Drosophila* mtDNA and is consistent with patterns of excess non-synonymous polymorphism relative to divergence in *Drosophila* and humans (Ballard and Kreitman 1994; Nachman 1998; Rand and Kann 1998). Yet, some of these slightly deleterious mutations do fix between species (e.g. Popadin et al. 2007) and more so in the mtDNA than in the nuclear genome (Lynch 1996). Theoretical models suggest that the bottleneck experienced by the mitochondrial genome during oogenesis increases the probability of fixation of these deleterious mutations but may also save the mitochondria from a “mutational meltdown” (Gabriel et al. 1993) by increasing the variance in mitochondrial fitness among the progeny of an individual and enhancing the efficacy of selection (Bergstrom and Pritchard 1998).

Other patterns suggest that positive selection shapes mtDNA evolution. Nucleotide diversity in mtDNA does not scale with population size across animal taxa as predicted under a mutation-drift equilibrium (Bazin et al. 2006). In species with large N_e , recurrent selective sweeps can erode the expected increase in polymorphism in larger populations via an increased rate of appearance and fixation of beneficial mutations (Gillespie 2000). However, this process of “genetic draft” is also predicted to increase the fixation rate of slightly deleterious mutations (Gillespie 2000). This may be particularly relevant for *Drosophila* mitochondrial genomes, if the mtDNA both harbors large amounts of slightly deleterious mutations and experiences repeated selective sweeps due to a large N_e .

The scope for positive and negative selection to shape mtDNA evolution will depend upon the distribution of functional effects of mitochondrial mutations. Understanding the main effects of mitochondrial mutations is complicated by the fact that oxidative phosphorylation (OXPHOS) complexes are jointly encoded by the mitochondrial and nuclear genomes, establishing a potential co-evolutionary dynamic between the genes encoded in each genome that interact to enable aerobic metabolism. Evidence is accumulating that mtDNA haplotypes

and nuclear-mtDNA interactions within species affect a variety of metabolic performance and life history traits (Rand et al. 2001; James and Ballard 2003; Ballard et al. 2007; Ellison and Burton 2008a, b). Mutations in mitochondrial-encoded OXPHOS genes are also linked to tRNA and rRNA genes encoded on the mtDNA that interact with nuclear-encoded mitochondrial proteins involved in transcription and translation. Inter-molecular coadaptation of these gene products may be important to maintain proper mitochondrial gene expression (Ellison and Burton 2008b), and intra-molecular coadaptation is important to maintain secondary structure and function of RNAs (Innan and Stephan 2001; Kern and Kondrashov 2004), providing a broad target for selection. Linkage to the cytoplasm also influences mitochondrial genome evolution. Cytoplasmic sweeps like those caused by cytoplasmic incompatibility-causing intracellular bacteria (e.g. *Wolbachia*) can increase the substitution rate and decrease diversity in mitochondrial genomes independent of the functional effects of mitochondrial mutations (Hoffmann and Turelli 1997; Shoemaker et al. 2004).

Here, we report the assembly of the *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis* and *D. grimshawi* mitochondrial genomes from shotgun sequences generated by the *Drosophila* 12 Genomes Consortium (2007) and the comparative analyses of these genomes with the previously sequenced mtDNAs of the *melanogaster* subgroup (Clary and Wolstenholme 1985; Garesse 1988; Ballard 2000). Mitochondrial genomic comparisons across the *Drosophila* phylogeny highlight novel aspects of positive, negative and weak selection, and allow annotation of conserved regions that maintain metabolic function. Combining this comparative approach with empirical 5' RACE data resolves previous uncertainty in the functional annotation of the *Drosophila* mtDNA. Evolutionary analysis of mitochondrial genomes across a phylogeny for which we also have complete nuclear genome sequence provides a foundation for understanding the relationships between molecular, biochemical and physiological divergence in a premier model system.

Materials and Methods

Mitochondrial Genome Assembly, Alignment and Annotation

mtDNA sequences were obtained using mega BLAST (Altschul et al. 1990) to query the *D. erecta*, *D. ananassae*, *D. persimilis*, *D. pseudoobscura*, *D. willistoni*, *D. mojavensis*, *D. virilis* and *D. grimshawi* trace archive databases generated by the *Drosophila* 12 Genomes Consortium (2007) with a complete *D. melanogaster* mtDNA sequence. Sequences were trimmed to remove vector and low-quality sequence and assembled in Sequencher v4.5. Assemblies do not include the mtDNA control region, which is A+T-rich, highly repetitive and divergent in sequence and length in *Drosophila* (Lewis et al. 1994). All assembled genomes contain polymorphic sites. Strains used for whole genome shotgun sequencing were inbred, but polymorphisms may persist despite inbreeding as mitochondrial haplotypes segregating either among or within sequenced individuals (heteroplasmy).

The overabundance of mitochondrial relative to nuclear DNA in a cell generated a high degree of coverage for the mtDNA (Table 1) and enabled straightforward assembly of a single large contig for *D. erecta*, *D. ananassae*, *D. mojavensis* and *D. grimshawi*. We manually sequenced a portion of the srRNA that was missing from the *D. ananassae* and *D. grimshawi* assemblies. The assemblies of *D. persimilis*, *D. willistoni* and *D. virilis* consisted of several large, non-overlapping contigs. We aligned these contigs to the complete mtDNAs and manually sequenced the remaining gaps. All assemblies were checked by hand. The high coverage suggests that the assemblies are not influenced by fragments of mtDNA that have transposed into the nuclear genome (NUMTs), which make up less than 1 kb of the *D. melanogaster* nuclear genome (Richly and Leister 2004). Manual sequencing of the srRNA from *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis* and *D. virilis* revealed no discrepancies with the assembled genomes.

We were not able to assemble contigs of significant length for *D. pseudoobscura*. The DNA used to sequence the *D. pseudoobscura* genome was prepared from purified embryonic nuclei (Richards et al. 2005), preventing the overrepresentation of mtDNA that we observed for the other species. *D. persimilis* and *D. pseudoobscura* share mtDNA haplotypes due to repeated introgressions where these species are sympatric (Powell 1983; Machado and Hey 2003), and we use the *D. persimilis* mtDNA as representative of this lineage.

For comparative analyses we included the following mtDNA lineages within the *melanogaster* subgroup (Ballard 2000): *D. melanogaster*, *D. simulans* sIII, *D. sechellia*, *D. mauritiana* malI and *D. yakuba* (Table 1). We aligned these genomes with the seven newly assembled mitochondrial genomes using ClustalX (Thompson et al. 1997). Less than 1.5% of the mitochondrial sequences contain non-coding DNA and the frequency of indels is low, making the alignment of these genomes straightforward. Alignments were checked by hand and the few, highly variable intergenic regions were left unaligned. We annotated the genomes based on the *D. melanogaster* and *D. yakuba* mtDNA annotations along with the data reported below from rapid amplification of 5' cDNA ends (5' RACE). The assembled and annotated genomes for *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis* and *D. grimshawi* are available in the Third Party Annotation Section of the DDBJ/EMBL/GenBank databases under the accession numbers TPA: BK006335-BK006341. The multi-species alignment and annotation is available online as Supplemental Material.

Sequencing Mitochondrial DNA and cDNA

To sequence gaps in the mtDNA assemblies, we PCR-amplified and sequenced genomic DNA extracted from pools of flies using a standard phenol:chloroform extraction. Additionally, we amplified and sequenced a fragment of the mitochondrial srRNA from *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis* and *D. virilis* that we compared to the assembled mitochondrial genomes to check the quality of our assemblies. Annotation of the aligned genomes revealed ambiguities in the start codon positions for *CoI*, *ND1* and *ND5*. To determine the 5' sequence of the mature transcripts for these genes we amplified the nucleic acid sequence at the 5'-end of *CoI*, *ND1* and *ND5* mature mRNA using 5' RACE (Invitrogen 18374-058.) All primer sequences are available online as Supplemental Material.

Comparative Genomic and Molecular Evolutionary Analyses

Sequence statistics and pairwise distances were calculated in MEGA 3.1 (Kumar et al. 2004), and sliding window analyses were performed using DnaSP 4.10 (Rozas and Rozas 1999). Phylogenetic relationships were inferred from a General Time Reversible model (Tavare 1986) with gamma distributed rate variation among sites and a proportion of invariant sites in Mr. Bayes 3.1 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). For each data set we ran four Markov chains, sampling for 19,000 iterations after an initial burn-in of 1000 iterations, which was sufficient for convergence. Polarized substitutions at synonymous sites were mapped onto the phylogeny using MacClade (Maddison and Maddison 2000) as in Rand and Kann (1998) and Ballard (2000). Codon-based models of divergence across the phylogeny (Goldman and Yang 1994; Yang and Nielsen 1998) were fit in PAML 3.14 (Yang 1997). Tajima's relative rate tests (RRT) (Tajima 1993) were performed in MEGA 3.1 and used *D. grimshawi* and *D. willistoni* as outgroup species for *D. mojavensis*-*D. virilis* and *D. persimilis*-*D. ananassae* comparisons, respectively. RRTs for ND were calculated using 2nd codon positions, but the small number of changes in CO genes required combining 1st and 2nd codon positions for CO RRTs. Cytochrome B and the ATPase genes contain only 377 and 274 codons, respectively, while ND and CO are encoded by 2054 and 999 codons, respectively. Because of the low number of changes observed in the small number of codons in Cytochrome B and the ATPases, we limited the analysis of codon bias and lineage-specific rate variation to the larger ND and CO gene complexes. We determined the secondary structures for the

majority of the 22 mitochondrial tRNAs using MFOLD (Zuker 2003) and tRNAscan (Lowe and Eddy 1997). The remaining structures were resolved by anchoring the tRNA at the anticodon and then identifying sequence regions that could form stems. To describe the process of compensatory mutation in tRNAs, we inferred the ancestral state of each tRNA at each node in the phylogeny using the baseml program in PAML 3.14. The *Drosophila* mtDNA encodes genes on both the major and minor strands. Analyses of protein-encoding and RNA genes were performed using the coding strand sequence.

Results

Genome Organization and Re-Annotation

Our assemblies reveal a high conservation of overall sequence length (Table 1) and no rearrangements in gene order (Fig. 1) across the 12 *Drosophila* mitochondrial lineages. Despite thorough comparative analyses in earlier studies of the *melanogaster* subgroup mtDNAs (Garesse 1988;Ballard 2000), our analyses across the 12 species phylogeny have uncovered several novel features of the *Drosophila* mtDNA. *Drosophila* *Col* does not contain a canonical in-frame start codon. It has been suggested that an out of frame ATA followed by the splicing of the subsequent A initiates translation in *D. melanogaster* and *D. yakuba* (de Bruijn 1983;Clary and Wolstenholme 1985). However, we find that the sequence ATAA is not conserved within the *melanogaster* subgroup nor is it present outside of the subgroup (Fig. 2). Sequencing the 5' end of the mature *Col* mRNA from *D. melanogaster*, *D. simulans*, *D. erecta*, *D. mojavensis* and *D. virilis* reveals that the start codon is downstream of the ATAA sequence in the *melanogaster* subgroup and is likely a novel (T/C)CG start codon that is in frame with a highly conserved nucleotide and amino acid sequence (Fig. 2). This confirms suspicions based on comparative sequence analysis in *Anopheles gambiae* that TCG was the “only remaining possibility” for the COI initiation codon (Beard et al. 1993). We find that the 12 *Drosophila* mtDNAs employ five different start codons and that the majority of genes have experienced substitutions in their start and/or stop codons across the phylogeny, with many genes also using incomplete stop codons (Fig. 1,Supplemental Table 1). Many invertebrate mtDNA employ a variety of start codons and incomplete stops (e.g. Boore and Brown 1994;Yamauchi et al. 2004;Breton et al. 2006). The tolerance of divergence in start and stop codons may reflect the nature of translation in a transcriptome that contains no characterized regulatory UTRs.

Two regions upstream of *ND1* and *ND5* that were previously annotated as noncoding DNA are nearly perfectly conserved across the *Drosophila* mitochondrial genomes (Fig. 3b). Both regions contain potential in-frame start codons that would extend the *ND1* and *ND5* proteins by three and five amino acids, respectively. Sequencing of the 5' end of the mature *ND1* and *ND5* mRNAs in *D. melanogaster*, *D. erecta* and *D. mojavensis* confirmed that both regions are included in the mature transcripts (Fig. 3b). The putative start codon for *ND1* is TTG and for *ND5* is GTG or ATG. Both TTG and GTG function as start codons in nematodes, mytilids and the black chiton *Katharina tunicata* (Boore and Brown 1994;Breton et al. 2006). The presence of in-frame start codons and the high degree of sequence conservation (only two putative synonymous substitutions) in these transcribed regions suggest that the amino acid sequences of these two proteins are longer than previously annotated. This result confirms previous suggestions that *ND5* could start with GTG in *D. yakuba* (Clary and Wolstenholme 1985) and *A. gambiae* (Beard et al. 1993), given the relatively large distance between *ND5* and the nearest tRNA.

mtDNA Nucleotide Composition and Pairwise Divergence

The *Drosophila* mtDNAs are A+T-rich, with third codon positions containing a remarkable 90.6 to 94.6% A+T (Table 1). The average pairwise Tamura–Nei distance (Tamura and Nei

1993) for the entire mtDNA is 0.123 changes per site (Table 2). Pairwise nucleotide distances range from 0.032 changes per site between *D. simulans* siII and *D. mauritiana* maII to 0.169 between *D. melanogaster* and *D. mojavensis* (Supplemental Material). The evolution of the mtDNA is dominated by changes in the protein-coding genes, with the RNA genes evolving relatively more slowly (Fig. 4a). Second codon positions are the slowest evolving element in the mtDNA, indicative of the maintenance of amino acid sequence (Fig. 4a), while third codon positions become rapidly saturated, with an average Tamura–Nei pairwise distance of 1.29 changes per site within the *melanogaster* subgroup.

Phylogenetic Analysis

The *Drosophila* species in this study represent the *melanogaster* (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. mauritiana*, *D. yakuba*, *D. erecta*, *D. ananassae*), *obscura* (*D. persimilis*) and *willistoni* (*D. willistoni*) groups of the subgenus *Sophophora* and the *virilis* (*D. virilis*), *repleta* (*D. mojavensis*) and *Hawaiian* (*D. grimshawi*) groups of the subgenus *Drosophila*. The phylogenetic relationships among these species groups are well resolved (Remsen and O'Grady 2002). However, the times separating speciation events in the *melanogaster* subgroup are short enough to allow for lineage sorting (Pamilo and Nei 1988), and different nuclear gene trees support different branching order between *D. yakuba*, *D. erecta* and the *D. melanogaster*–*D. simulans*–*D. sechellia*–*D. mauritiana* species clade (Ko et al. 2003; Pollard et al. 2006; Wong et al. 2007).

Bayesian estimation of the *Drosophila* mtDNA phylogeny strongly supports *D. yakuba* and *D. erecta* as sister taxa relative to *D. melanogaster* when either the entire mtDNA ($n = 15236$ base pairs) or the complete set of protein-encoding genes ($n = 11112$ bp) was used to estimate the tree topology. In each case a single topology (Fig. 5a) was estimated with each node having a posterior probability of one. This topology is consistent with the phylogenetic relationships inferred from smaller mtDNA sequence sets (Nigro et al. 1991; Kopp and True 2002; Lewis et al. 2005) and with the consensus from many nuclear loci (Ko et al. 2003; Pollard et al. 2006; *Drosophila* 12 Genomes Consortium 2007; Wong et al. 2007).

Intergenic and Regulatory DNA Evolution

Junctions between genes in the *Drosophila* mtDNAs contain less than 1.5% of the mitochondrial nucleotides but harbor the majority of the indel divergence (excluding the A+T-rich control region). Functional constraint varies greatly across intergenic regions, with some regions nearly perfectly conserved in both length and sequence and other regions unalignable due to expansion of repetitive tracks of As and Ts (Fig. 3a, d, e). The junction between *ATPase6* and *CoIII* contains no nucleotides in the *melanogaster* species group but ranges in length from 6 to 52 nucleotides in the remaining genomes (Fig. 3e). Clary and Wolstenholme (1985) suggested that a stem-loop configuration at this junction provides the splicing signal in *D. yakuba*. The expansion of this intergenic region in *Drosophila* species outside of the *melanogaster* subgroup suggests alternative mechanisms exist to ensure proper processing of the *CoIII* transcript. Splicing at the putative position at the base of the stem (Fig. 3e) would disrupt the reading frame or incorporate extra amino acids without additional post-transcriptional processing due to the presence of potential start codons in the intergenic region upstream of *CoIII*.

Two noncoding regions that bound gene clusters encoded on opposite strands of the mtDNA contain binding sites for the nuclear-encoded *D. melanogaster* transcription termination factor dmTTF that arrests progression of the mitochondrial polymerase (Roberti et al. 2003). While these two regions are more conserved than other highly variable non-coding regions, they still harbor a considerable amount of nucleotide and indel divergence (Fig. 3d). The entire intergenic region was shown to functionally bind dmTTF in *D. melanogaster* (Roberti et al.

2003). We find three motifs, AAMTT, ACTAA and AATTCA, that are shared by both intergenic regions and are nearly invariant across the 12 species (Fig. 3d), suggesting that these nucleotides are functionally conserved for binding dmTTF.

Animal mitochondrial rRNA genes are clustered and flanked by a tridecamer transcription terminator signal (TGGCAGA- - - -G) that is often embedded within the tRNA^{Leu} gene (Valverde et al. 1994). We find that the *Drosophila* tridecamer rRNA transcription termination signal is embedded in the tRNA^{Leu(CUN)} gene flanking the rRNA gene cluster and is perfectly conserved across the 12 *Drosophila* mtDNAs (Fig. 3c). We also identified an identical and similarly conserved 13 bp rRNA termination signal within the tRNA^{Leu(UUR)} gene (Fig. 6a, b).

tRNA Evolution

We quantified the evolution of each tRNA using both the average pairwise divergence across the 12 mtDNAs and the number of changes on the phylogeny estimated by inferring ancestral tRNA sequences at each node. Both measures revealed rate heterogeneity across the 22 tRNAs (Table 3), with tRNA^{Asn} having the largest and tRNA^{Ser(AGN)} the smallest number of inferred changes on the phylogeny (Fig. 6c, d). The fact that tRNA^{Ser(AGN)} is the most conserved tRNA in the *Drosophila* mtDNA is interesting, given its anomalous tRNA structure that is conserved in other insect mitochondrial genomes (e.g. Beard et al. 1993; Yamauchi et al. 2004) and its relatively low usage compared to tRNA^{Ser(UCN)} which is used over twice as often to encode serine in the *Drosophila* mitochondrial genomes. Neither measure of tRNA evolutionary rate was correlated with the usage of the amino acid associated with each tRNA (pairwise divergence $r = -0.013$, $P = 0.95$; inferred changes $r = -0.184$, $P = 0.41$), nor was evolutionary rate correlated with position along the coding strand (pairwise divergence $r = -0.111$, $P = 0.62$; inferred changes $r = -0.002$, $P = 0.99$).

Over 90% of inferred changes in tRNAs maintain Watson–Crick pairing (Table 3). We inferred 36 occurrences where both nucleotides of a Watson–Crick pair changed on a single branch in the phylogeny. In each case pairing was maintained. Due to G–U pairing in RNA, it is possible for the transition between nucleotide states at Watson–Crick pairs to avoid a non-pairing intermediate. For example, the transition from G–C to A–U can occur in the order G–C to G–U to A–U or through the non-pairing AC intermediate (Fig. 6c). When both nucleotides at a Watson–Crick pair have changed on the same lineage, we cannot infer which intermediate path was taken. If we assume, conservatively, that any pair of changes that could have transitioned through a G–U intermediate did so, and if we consider only compensation at the level of Watson–Crick pairing, we infer that 7 of the 36 pairs of changes must have transitioned through a non-pairing state (Fig. 6c, Table 3). Thus, 19% of the inferred changes at Watson–Crick pairs in the *Drosophila* mitochondrial tRNAs potentially traversed a fitness trough with the subsequent substitution of a compensatory mutation.

Evolution at Synonymous Sites in the mtDNA

The mitochondrial genome encodes genes on the major and minor strands in a non-overlapping fashion, with all of the protein-coding genes on the minor strand encoding components of the ND complex. Using both a six species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. mauritiana*, *D. yakuba*, *D. erecta*) and the complete 12 species mitochondrial phylogeny we inferred the frequency of all unambiguous polarized substitutions at synonymous sites for genes encoded on the major and minor strands (Table 4). There is a highly biased pattern of transitions at synonymous sites that differs between genes encoded on the two strands. Across the 12 species phylogeny the ratio of A ↔ G to C ↔ T transitions is 0.40 on the major strand but 2.76 on the minor strand. This difference cannot be explained simply by differences in the abundance of A+G or C+T at third positions on the two strands, as weighting the substitution frequencies

by the nucleotide frequencies does not remove the strand-bias (Table 4). There are many more A to G than G to A and T to C than C to T substitutions on both strands of the mtDNA. However, when nucleotide frequencies are accounted for, the reciprocal substitution rates between nucleotides are similar (Table 4).

We calculated codon usage bias as the relative synonymous codon usage (RSCU) across all 12 *Drosophila* mitochondrial genomes, for ND genes encoded on the two strands and for concatenated CO and ND gene sequences (Supplemental Table 2). A+T content at third positions exceeds 90% and codon usage is highly biased toward an A- or T-ending (A/T-ending) codon for every codon family across all 12 species (Supplemental Table 2). Because each codon family employs a single tRNA, the first position of the anticodon is always a wobble position that will only be a perfect match to one of the two possible codons for a twofold redundant codon family and to one of the four possible codons for a four-fold redundant codon family. The prevalence of G/C-beginning anticodons among the *Drosophila* tRNA genes combined with a mutation pressure towards A/T-ending codons results in only 7 of the 22 tRNA codon families having usage bias for the codon providing a perfect match to the tRNA anticodon (Supplemental Table 2).

Codon usage bias is greater in the *Sophophora* species than in the *Drosophila* species when measured either as a lower effective number of codons (ENC, Table 1; $t_{10} = 7.47$, $P < 0.0001$) or as greater RSCU for A/T-ending codons (Supplemental Table 2; paired $t_{21} = 4.84$, $P < 0.0001$). Consistent with an overall greater codon usage bias toward A/T-ending codons, the *Sophophora* species proteomes have significantly greater A+T content at third positions than do the *Drosophila* species proteomes (Table 1; Soph-Dros = 2.4%, $t_{10} = 6.52$, $P < 0.0001$).

There are 7 two-fold redundant codon families that end in a C/T and in each case the C-ending codon is the perfect match to the anticodon. When RSCU is averaged across all 12 genomes, we find that the ND genes encoded on the minor strand more often use the T-ending mismatch codon than do ND genes encoded on the major strand (diff. RSCU(- -T) major – minor = -0.22, paired $t_6 = 3.1$, $P = 0.021$) (Supplemental Table 2). Similarly, for A/G-ending two-fold redundant codon families, ND genes on the minor strand more often than ND genes on the major strand use the G-ending codon, which is the perfect match to the anticodon for only 1/3 of these codon families (diff. RSCU(- -A) major – minor = 0.123, paired $t_5 = 3.61$, $P = 0.015$) (Supplemental Table 2). With the exception of tRNA^{Ser(AGN)}, all fourfold redundant codon families in the *Drosophila* mtDNA are decoded by a tRNA that is a perfect match to the A-ending codon. Again, ND genes on the minor strand more often use the mismatch T-ending codon relative to ND genes on the major strand (diff. RSCU(- -T) major – minor = -0.834, paired $t_7 = 9.17$, $P < 0.0001$) (Supplemental Table 2). The strand-biased use of the mismatch codon results in the ND genes on the major strand using the perfect match anticodon more often than ND genes on the minor strand (diff. RSCU(match) major – minor = 0.37, paired $t_{21} = 4.08$, $P = 0.0005$) (Supplemental Table 2).

Despite the strong strand biases in both substitution pattern and codon usage bias, there are differences in codon usage between CO and ND genes that are independent of coding strand. There are ten synonymous codon families where a G/C-ending codon is the perfect match to the tRNA anticodon, at odds with the mutation and codon usage bias toward A/T-ending codons. RSCU for the A/T-ending mismatch codon is similar for ND genes on the major and minor strands (diff. RSCU(- -A/T) major – minor = -0.133, paired $t_9 = 1.83$, $P = 0.1$). However, when compared to CO genes, ND genes more often use the A/T-ending mismatch codon (diff. RSCU(- -A/T) CO – ND = -0.160, paired $t_9 = 4.87$, $P = 0.0009$). The increased use of the G/C-ending match codon by CO genes is also reflected in the overall lower third codon position A+T content in CO relative to ND genes across the 12 mtDNA (diff. A+T% CO – ND = -1.6%, paired $t_{11} = 5.81$, $P < 0.0001$).

Divergence Across the Mitochondrial Proteome

Average pairwise divergence rates (Li 1993; Pamilo and Bianchi 1993) across the *Drosophila* mitochondrial proteomes are 0.045 at non-synonymous sites (d_N) and 0.653 at synonymous sites (d_S), with an average d_N/d_S of 0.069 (Fig. 4b, Table 2). Average pairwise d_S is relatively similar across the four OXPHOS complexes when compared to d_N , which varies three-fold across OXPHOS complexes and results in OXPHOS complex-specific rates of d_N/d_S (Fig. 4b, Table 2). A sliding window of divergence rates across the concatenated mitochondrial proteome reveals that variation in d_N and d_N/d_S across the *Drosophila* proteome is highly OXPHOS complex-specific (Fig. 7a). Three distinct peaks of d_N and d_N/d_S correspond exclusively to ND genes in comparisons between *D. melanogaster* and the increasingly divergent mtDNAs of *D. yakuba*, *D. persimilis*, and *D. mojavensis* (Fig. 7a).

Rapid evolution at silent sites in the mtDNA is problematic for estimating d_N/d_S , and codon-based models of molecular evolution may better estimate this rate (Goldman and Yang 1994; Yang and Nielsen 1998). We fit codon-based models in PAML (Yang 1997) with either a single value of $\omega = d_N/d_S$ (model 0) or with three site classes each with a different ω value (model 3) across the *melanogaster* species group phylogeny (Table 5) using the phylogenetic relationships in Fig. 5. Codon-based estimates of d_N/d_S across the phylogeny were up to an order of magnitude lower than the Pamilo–Bianchi–Li pairwise divergence rates, suggesting that the codon-based models better account for the rapid evolution at silent sites. To test for rate heterogeneity among OXPHOS complexes, we compared the likelihood of the proteome data under model 0 to the sum of the likelihoods from model 0 fit to each of the four concatenated OXPHOS complex data sets (Parsch et al. 2001). Allowing a distinct d_N/d_S for each complex provides a significantly better fit to the mitochondrial proteome data than fitting a single rate ($\chi^2_3=818$, $P<0.0001$; Table 5). The rank ordering of d_N/d_S values across OXPHOS complexes was the same as that from estimates of pairwise divergence: ND, $\omega = 0.0126 >$ ATPase, 0.0076 > CytB, 0.0059 > CO, 0.0043 (Fig. 7). The difference in evolutionary rate between ND and CO is striking and persists across all 12 mitochondrial proteomes, with ND having four times the non-synonymous tree length as CO (Fig. 5, Table 5).

Lineage-Specific Rates of Proteome Evolution

While the ND complex evolves faster than the CO complex across the entire *Drosophila* phylogeny, there is also lineage-specific rate variation within both complexes (Fig. 5a). Models estimating a distinct ω on each lineage were a better fit to the ND and CO data across the *melanogaster* species group (Table 5). Across the 12 species phylogeny, the *D. mojavensis* ND genes have accumulated approximately twice as many non-synonymous substitutions as *D. virilis* (Tajima's RRT, $\chi^2_1=15.5$, $P<0.0001$), while these two species have a similar substitution rate in CO genes (RRT, $\chi^2_1=0.53$, $P = 0.466$) (Fig. 5). The *D. persimilis* mtDNA has accumulated the greatest amount of divergence in CO genes and twice as much as *D. ananassae* (RRT, $\chi^2_1=7.41$, $P = 0.006$), while these two species have accumulated a similar amount of divergence in ND genes (RRT, $\chi^2_1=0.02$, $P = 0.893$) (Fig. 5).

Discussion

The overarching evolutionary process acting on the animal mitochondrial genome is selection to maintain critical function that, in insects, enables a highly efficient aerobic metabolism that fuels energetically demanding behaviors such as flight and a high reproductive effort. Across the 12 *Drosophila* mitochondrial genomes we observe a strong signature of this process in highly conserved regulatory motifs, in the maintenance of tRNA secondary structure and in the low d_N/d_S values for genes encoding OXPHOS proteins. Yet we find that, in addition to negative selection against non-synonymous change, there is scope for weak selection and

positive selection in the mitochondrial genome that varies across functional domains and lineages.

Orchestrating Transcription of a Highly Streamlined Genome

Drosophila mtDNA are transcribed as polycistronic messages that are cleaved to produce the mature mRNA products (Berthier et al. 1986). As animal mtDNAs generally contain no introns and little room for regulatory information, the signals orchestrating this complex suite of transcription and post-transcriptional processing likely lie in the junctions between genes. These signals arise from transcription termination factor binding sites (Valverde et al. 1994; Roberti et al. 2003) or from the secondary structure of coding sequences between genes (Ojala et al. 1981; Clary and Wolstenholme 1985). Patterns of conservation and divergence in noncoding mtDNA across the 12 *Drosophila* species provide information on the regulatory capacity of intergenic nucleotides.

Many of the intergenic regions of the *Drosophila* mtDNA are rapidly evolving and reflect the high mutation pressure exerted on *Drosophila* mtDNA. However, other intergenic regions contain islands of conservation that imply regulatory function. Two intergenic regions of the *D. melanogaster* mtDNA bind the transcription termination factor dmTTF (Valverde et al. 1994; Roberti et al. 2003) and comparisons across the 12 *Drosophila* mtDNAs revealed that these two intergenic regions share three sequence domains that are conserved across all 12 species. Outside of these three domains, 20 of the 21 remaining intergenic nucleotide sites have experienced at least one change on the phylogeny. This pattern of high conservation in an otherwise rapidly evolving intergenic region suggests that these sites are necessary for binding dmTTF and ensuring proper transcriptional regulation of the polycistronic messages encoded on the two strands of the *Drosophila* mtDNA.

The junction between *ATPase6* and *CoIII* is highly conserved within the *melanogaster* subgroup and contains no intergenic nucleotides. A putative stem-loop structure at this junction is hypothesized to underlie splicing of these two transcripts (Clary and Wolstenholme 1985). Outside of the *melanogaster* subgroup, this intergenic region varies greatly in length and contains out-of-frame start codons. This provides an example of the mutational vulnerability of noncoding regions that is posited to underlie the streamlining of animal mtDNA (Lynch et al. 2006) and suggests either that this stem-loop structure does not underlie splicing of these gene transcripts or that other *Drosophila* species have acquired additional mechanisms for proper splicing of the downstream *CoIII* transcript.

Despite the polycistronic nature of transcription in the mtDNA, *Drosophila* mitochondrial genes within the same polycistronic message are differentially expressed (Berthier et al. 1986). Across animal mtDNA, a highly conserved tridecamer transcription terminator signal sequence located immediately downstream of the rRNA cluster is thought to ensure the proper accumulation of rRNA in excess of other mitochondrial transcripts (Valverde et al. 1994). The signal sequence resides in the *Drosophila* tRNA^{Leu(CUN)} gene immediately adjacent to the rRNA genes and is perfectly conserved across the 12 *Drosophila* genomes. Surprisingly, we found an identical and equally conserved signal sequence in the *Drosophila* tRNA^{Leu(UUR)} gene. In contrast, the human and mouse rRNA termination signal is found only in the tRNA^{Leu(UUR)} gene that abuts the rRNA cluster.

Comparing the mitochondrial tRNA genes of the cockroach *Periplaneta fuliginosa*, the dragonfly *Orthetrum triangulata melania*, the mosquito *Anopheles gambiae* and the flour beetle *Tribolium castaneum* (Beard et al. 1993; Friedrich and Muqim 2003; Yamauchi et al. 2004) revealed that both tRNA^{Leu} genes within each of these insect mtDNAs also contain the conserved transcription termination consensus sequence, TGGCAGA--AGTG. Strikingly, the tridecamer sequence is more conserved between paralogous tRNA^{Leu} genes within each of

these genomes than between orthologous tRNA^{Leu} genes. The perfectly conserved sequence spans the D-loop of both tRNA^{Leu} genes, suggesting conserved function of this signal sequence, as the D-loops of 18 of the remaining 20 mitochondrial tRNAs have accumulated multiple substitutions across the *Drosophila* phylogeny.

The tRNA^{Leu(UUR)} gene lies between *Col* and *ColI* in the *Drosophila* mitochondrial genome (Fig. 1). Relative transcript abundance is higher for CO genes than for other OXPHOS complexes in *D. melanogaster*, but *Col* mRNA does not accumulate to greater levels than CO genes downstream of the tRNA^{Leu(UUR)} transcriptional termination signal (Berthier et al. 1986). In contrast, the rRNA gene transcripts are seven times as abundant as transcripts of genes located downstream of the termination signal in the tRNA^{Leu(CUN)} gene (Berthier et al. 1986). This suggests that the signal sequence in the tRNA^{Leu(UUR)} gene is not serving as a transcriptional termination signal downstream of *Col*. However, the maintenance of this sequence in multiple tRNA^{Leu} genes, independent of location relative to the rRNA cluster, suggests an unknown functional capacity of this signal sequence in insect mtDNA.

Strand-Specific Mutation-Selection Balance at Synonymous Sites

Genes on the two strands of the mtDNA experience different patterns of substitution at silent sites. We find both a biased pattern of substitution in and a difference in synonymous codon usage by the genes encoded on the two strands. Consistent with previous observations in smaller *Drosophila* datasets (Rand and Kann 1998; Ballard 2000), across the 12 species phylogeny there are more than twice as many C ↔ T as A ↔ G transitions on the major strand, while there are nearly three times as many A ↔ G as C ↔ T transitions on the minor strand. Strand-specific substitution patterns could reflect both a bias in the mutation process (e.g. Reyes et al. 1998) and/or a difference in the process of weak selection acting on mutations at silent sites on the two strands. These processes are difficult to distinguish, as a true mutation bias on the two strands could generate strand-specific patterns of codon usage bias in the absence of weak selection on silent sites. Additionally, all genes on the minor strand encode components of the same OXPHOS-complex, ND, making it challenging to know whether ND-specific functional constraint is shaping strand-specific substitution patterns or whether strand-specific mutation bias is driving codon usage bias in ND genes.

Having the depth of information from all 12 genomes imparts power to detect differential patterns of codon usage bias between genes encoded on the two strands, between genes encoding different OXPHOS complexes and between species subgroups. Codon usage bias in nuclear genomes is thought to be maintained by a balance of selection increasing codon bias and mutation-drift eroding codon bias (Li 1987; Bulmer 1991; Akashi 1997, 2001). The nuclear genome employs multiple tRNAs to translate each codon family, and selection for translational efficiency and/or accuracy is hypothesized to explain the observation that the most highly used codon matches the most abundant tRNA (e.g. Ikemura 1985; Duret 2000). Mitochondrial genomes encode a single tRNA gene to decode each codon family. Assuming that there is no import of compatible nuclear tRNAs into the mitochondrion, this eliminates the selective pressure for codon preference to match tRNA abundance. Nevertheless, codon usage bias in animal mtDNA is widespread (e.g. Herbeck and Novembre 2003).

A strong mutation bias toward A+T (Martin 1995; Haag-Liautard et al. 2008) dominates codon usage bias in the *Drosophila* mtDNA. Yet, the codon that matches the anticodon of the majority of *Drosophila* mitochondrial tRNAs contains a G/C nucleotide at the third position. If there is selection for codon preference to match the tRNA anticodon (Bulmer 1991; Xia 2005), then the mutation process leading to high A+T content is moving the *Drosophila* mitochondrial proteome away from its ideal genetic code. This suggests a different mutation-selection balance than that of the nuclear genome. In the *Drosophila* mtDNA, mutation pressure generates a

codon usage bias toward A/T-ending codons that would oppose weak selection for codonanticodon matching.

Codon usage bias differs in genes encoded on the two strands of the *Drosophila* mtDNA. When compared to the ND genes on the major strand, ND genes encoded on the minor strand more often use T-ending codons that mismatch the anticodon for all two-fold redundant codon families ending in C/T and more often use the G-ending codon that mismatches the anticodon for 4 of the 6 twofold redundant codon families ending in A/G. This is consistent with the overabundance of T to C substitutions on the major strand and A to G substitutions observed on the minor strand. Because this difference is manifest between genes of similar function but encoded on different strands, this suggests that a biased mutation process rather than weak selection on silent sites underlies this pattern of codon usage and substitution bias on the two strands. In this scenario, mutation pressure causes higher usage of codons that mismatch the anticodon in ND genes encoded on the minor strand.

However, other patterns of codon usage are not consistent with a simple strand-biased mutation scenario. The majority of four-fold redundant codon families are decoded by a tRNA that matches the A-ending codon. ND genes on the minor strand more often use the mismatch T-ending codon, while ND genes on the major strand more often employ the matching A-ending codon. However, the frequency of A ↔ T substitutions is similar on both strands across the 12 mitochondrial genomes (major strand 0.482; minor strand 0.459). There are also patterns of codon usage between CO and ND genes that differ independent of strand. For codon families that are decoded by a tRNA that matches G/C-ending codons, CO genes more often use a matching codon than do ND genes, while ND genes on the two strands do not differ significantly in their codon usage. As a result of using more G/C-ending codons, CO genes have a decreased codon usage bias that is generally toward A/T-ending codons.

In nuclear genomes, codon bias is greater in more highly expressed genes (reviewed by Akashi 2001). CO gene transcripts are four times more abundant than ND gene transcripts in *D. melanogaster* mitochondria (Berthier et al. 1986). If there is weak selection for the perfect match codon to increase translational accuracy and/or efficiency, we predict that this pressure would be stronger on CO than ND genes. In contrast to the nuclear genome where weak selection for codon preference leads to increased codon usage bias, weak selection for codonanticodon matching in the A+T-rich *Drosophila* mtDNA is manifested as a lower codon usage bias. A similar mutation-selection balance may also exist in the A+T-rich *Plasmodium falciparum* nuclear genome, where more highly expressed genes more often use G/C-ending codons, decreasing a codon usage bias that is generally toward A/T-ending codons (Musto et al. 1999).

Evolution at synonymous sites in mtDNA is a complex balance of mutation bias and weak selection on codon usage that is modulated by the coding strand and the functional gene in which the codon resides. We also find that this mutation-selection balance at silent sites in the mitochondrial proteome has diverged across the *Drosophila* phylogeny, with higher codon usage bias in the *Sophophora* than in the *Drosophila* subgenus species (Supplemental Table 2).

Negative Selection and Variation in Functional Constraint Across OXPHOS Complexes

Comparing divergence rates estimated from codon-based models across nine *Drosophila* mtDNAs (*D. melanogaster*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. mojavensis*, *D. virilis*, *D. grimshawi*) to those estimated from 5067 nuclear genes (D. A. Pollard, V. N. Iyer and M. B. Eisen, personal communication) reveals that the mtDNA has 2.75 times the rate of synonymous change as the nuclear genome, yet only 0.37 times the rate of non-synonymous change, consistent with patterns in mammalian mtDNA (Saccone et al.

2000). The overall pattern of rapid change at silent sites and near stasis at second codon positions indicates that negative selection on the *Drosophila* mtDNA maintains critical protein function in the face of a high mutation pressure. Yet, evolutionary rate heterogeneity across the *Drosophila* mitochondrial proteome indicates that the selective effects of mutations arising in the mitochondrial genome are clearly OXPHOS complex-dependent, with ND accumulating many more substitutions than CO. Genes encoding subunits of ND also evolve more rapidly than those of CO in mammals (Pesole et al. 1999), invertebrates (Ballard 2000; Breton et al. 2006) and fish (Doiron et al. 2002).

The scope for neutral mutations to become substitutions depends upon the degree of functional constraint for each complex. If functional constraint varies between OXPHOS complexes, then we might expect to see differences between ND and CO genes in the proportion of neutral nonsynonymous substitutions or in the number of slightly deleterious polymorphisms that can be tolerated. Our analysis of polymorphism data compiled by Bazin et al. (2006) reveals that across insect taxa ND genes have greater levels of segregating non-synonymous variation than do CO genes (median ND $\pi_N = 0.0037$, CO $\pi_N = 0.0014$, Mann–Whitney $P < 0.0001$; median ND $\pi_N/\pi_S = 0.0598$, CO $\pi_N/\pi_S = 0.035$, Mann–Whitney $P = 0.002$), supporting the hypothesis that replacement mutations generally have milder negative fitness consequences in ND than in CO genes.

While linkage does not affect the substitution rate of selectively neutral mutations, linkage does increase the rate of fixation of slightly deleterious mutations via Hill–Robertson effects (Hill and Robertson 1966; Felsenstein 1974; Birky and Walsh 1988). Either positive or negative selection operating on sites in the mtDNA will increase the variance in offspring produced by individuals in the population, decreasing the effective population size and increasing the effects of drift at linked sites segregating deleterious mutations (e.g. Charlesworth et al. 1993; Gillespie 2000). The increased non-synonymous substitution rate of ND relative to CO genes may then be due to the effects of strong linkage between selected sites across the molecule and an increased number of segregating slightly deleterious variants at ND relative to CO genes. The low absolute values of d_N/d_S indicate that negative selection is rampant in the mtDNA, but the observations of Bazin et al. (2006) suggest that taxa with large populations sizes, such as *Drosophila*, may also experience repeated bouts of positive selection. In this evolutionary scenario, the fixation rates for slightly deleterious variants in the mtDNA are likely dictated by selection at linked sites, increasing the number of slightly deleterious substitutions over what would be expected in a recombining genome.

Untangling the web of evolutionary forces acting on mtDNA requires understanding the distribution of functional and selective effects of mitochondrial mutations and how this differs across OXPHOS complexes. Amino acid substitutions in CO genes do have functional effects, particularly when combined with divergent nuclear backgrounds. CO is highly divergent between populations of the intertidal copepod *Tigriopus californicus* and this molecular divergence is manifested as a disruption in CO activity in mitochondrial introgressions between populations (Edmands and Burton 1999). Only a handful of amino acid substitutions have accumulated in CO genes between *D. simulans* and *D. mauritiana* mitochondrial haplotypes (Ballard 2000). Nevertheless, CO activity differs between mitochondrial haplotypes in *D. simulans* (Ballard et al. 2007) and is disrupted in mitochondrial introgressions between *D. simulans* and *D. mauritiana* (Sackton et al. 2003).

Less is known about the functional consequences of amino acid divergence in ND. The mtDNAs of the Brook and Arctic charr differ by 45 amino acids in ND and only a single amino acid in CO (Doiron et al. 2002). Yet, ND activity has not diverged between species, nor is it disrupted in hybrids (Blier et al. 2006). Clary and Wolstenholme (1985) observed that *D. yakuba* and mouse ND6 have identical hydrophobicity profiles despite the accumulation of

many amino acid differences. The effect of substitutions in ND6 appears neutral with respect to the maintenance of an overall hydrophobicity profile for ND6 across the *Drosophila* mitochondrial phylogeny (Fig. 8a). However, correlations between the hydrophobicity profiles of *D. melanogaster* and each of the 11 other species were lower for ND6 than for COI (Fig. 8; paired $t_{10} = 5.01$, $P = 0.0005$), indicating that the greater amino acid divergence in ND6 relative to COI does result in greater divergence in hydrophobicity for ND6 than COI. It remains to be determined whether this change in hydrophobicity causes functional divergence in ND activity with corresponding effects on fitness.

Scope for Positive Selection in the Mitochondrial Genome

Ballard (2000) identified three blocks in the mtDNA, all containing at least one ND gene, that were evolving inconsistently with the remainder of the genome along particular lineages of the *melanogaster* subgroup. This suggests that lineage-specific selection on a subset of genes contributes to the uncoupling of evolutionary rates across the mtDNA. Across the *Drosophila* mitochondrial genomes we find evidence for lineage-specific rates of evolution for both CO and ND genes. The lineage-specific rates do not indicate that any one species is evolving more quickly across the entire mitochondrial proteome, as would result from demographic effects or changes in mitochondrial mutation rate along a lineage. Rather, the selective effects of mitochondrial mutations differ not simply between complexes, but they do so in a lineage-specific manner. Evolutionary rates of ND and CO genes are accelerated along the *D. mojavensis* and *D. persimilis* mitochondrial lineages, respectively, and may reflect positive selection to maintain metabolic function in a dramatically different environment or in the face of repeated introgressions among sympatric species. Aerobic metabolism is a phenotype under strong stabilizing selection, yet the physiologies underlying metabolism must still evolve to maintain function in light of shifts in ecology, behavior and life history.

D. mojavensis is a xeric, cactophilic *Drosophila*. The sequenced strain inhabits the prickly pear *Opuntia* on Santa Catalina Island off the Southern California coast, and mitochondrial population genetic data suggest that this population has been small and isolated from its mainland counterparts for some time (Reed et al. 2007). *D. mojavensis* has accumulated the greatest amount of amino acid divergence in ND, resulting in decreased hydrophobicity in a hydrophilic region of ND6 (Figs. 5, 7, and 8). Whether these changes in ND may be advantageous for desert *Drosophila* like *D. mojavensis* or whether they have accumulated as a result of its small population size warrants further investigation. CO is evolving rapidly along the *D. persimilis* lineage (Fig. 5). Disruption of CO activity in mitochondrial introgressions between closely related *Drosophila* species suggests that mitonuclear epistases accumulate in CO genes (Sackton et al. 2003), and population studies comparing *D. persimilis* and *D. pseudoobscura* mitochondrial haplotypes indicate that there are fitness consequences of these intergenomic epistases (Hutter and Rand 1995). *D. pseudoobscura* and *D. persimilis* hybridize in nature (Powell 1983; Machado and Hey 2003), and one hypothesis is that repeated introgressions of mtDNA where these species are sympatric drives more rapid evolution of CO genes along the *pseudoobscura-persimilis* lineage.

There is also scope for positive selection to maintain tRNA structure in the *Drosophila* mtDNA. Nearly 20% of paired changes inferred at Watson–Crick pairs in tRNA stems likely traversed a state of decreased fitness, providing a selective pressure to fix compensatory mutations. It has been argued both that there is extensive positive selection for compensatory mutations in animal mitochondrial tRNAs (Kern and Kondrashov 2004) and that this process is overwhelmed by the accumulation of deleterious mutations via Muller's ratchet (Lynch 1996). Our estimate of 19% is consistent with Kern and Kondrashov's (2004) estimate that 10–50% of the substitutions in tRNAs along mammalian lineages are compensatory. The majority of paired changes inferred on single branches of the *Drosophila* phylogeny could have

proceeded via a G-U transition state. If the deleterious effects of this intermediate on fitness are small, this may increase the chance that a beneficial mutation arises on the same mitochondrial haplotype (Innan and Stephan 2001) making these pairs more likely to be observed than those changes that must traverse a trough of low fitness. However, 10% of the inferred changes on the *Drosophila* phylogeny do disrupt Watson–Crick pairing and are presumably deleterious variants that have not been removed by selection or restored by compensatory substitutions.

Conclusion

Analysis of 12 mitochondrial genomes across the well-defined *Drosophila* phylogeny provides evidence for both positive and negative selection shaping this genome. Positive and negative selection in the absence of recombination should strongly influence mitochondrial evolution via Hill–Robertson effects (Hill and Robertson 1966), and a subset of changes on the *Drosophila* phylogeny are likely to be those deleterious mutations that have fixed via linkage to selected sites in the genome (Birky and Walsh 1988; Charlesworth et al. 1993; Gillespie 2000). Other non-neutral substitutions may be accumulating along *Drosophila* mitochondrial lineages due to linkage with a cytoplasm that may harbor CI-inducing strains of *Wolbachia* (Hoffmann and Turelli 1997; Shoemaker et al. 2004). Understanding this balance of forces will require empirical study of the population dynamics of mtDNA haplotypes (e.g. levels of heteroplasmy, paternal mtDNA leakage and mitochondrial N_e), the distribution of functional and selective effects of mitochondrial mutations and the interaction of these mutations with the nuclear genetic variants that they encounter in nature. The 12 *Drosophila* mitochondrial genomes are embedded in a well-resolved phylogeny with complete nuclear genome sequence, and disruption of the mitochondrial–nuclear relationship in *Drosophila* is possible in the lab and occurs in natural populations (Machado and Hey 2003; Sackton et al. 2003; Bachtrog et al. 2006; Ballard et al. 2007). *Drosophila* are thus uniquely poised to investigate the population genetic parameters that are critical for models describing the fixation of beneficial and deleterious mutations via linkage, the maintenance of mitochondrial variation through mitochondrial–nuclear interactions (Rand et al. 2001) and the dynamics that lead both to the ratcheting of the mitochondria toward and the rescuing of the mitochondria from evolutionary decay (Bergstrom and Pritchard 1998).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge constructive discussion with the fly genomics community at large, as well as specific input from Brian Bettencourt, Rob Haney, Rob Kulathinal and Colin Meiklejohn and helpful comments from anonymous reviewers. The shotgun sequence data reported in this manuscript were generated by the *Drosophila* 12 Genomes Consortium. This work was supported by National Institutes of Health grants GM067862 to DMR and GM076812 to KLM and a Summer Research Fellowship to DMR in the Bay Paul Center for Comparative Molecular Biology and Evolution at the Marine Biological Laboratories.

References

- Akashi H. Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene* 1997;205:269–278. [PubMed: 9461401]
- Akashi H. Gene expression and molecular evolution. *Curr Opin Genet Dev* 2001;11:660–666. [PubMed: 11682310]
- Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410. [PubMed: 2231712]

- Bachtrog D, Thornton K, Clark A, Andolfatto P. Extensive introgression of mitochondrial DNA relative to nuclear genes in the *Drosophila yakuba* species group. *Evolution* 2006;60:292–302. [PubMed: 16610321]
- Ballard JW. Comparative genomics of mitochondrial DNA in members of the *Drosophila melanogaster* subgroup. *J Mol Evol* 2000;51:48–63. [PubMed: 10903372]
- Ballard JW, Kreitman M. Unraveling selection in the mitochondrial genome of *Drosophila*. *Genetics* 1994;138:757–772. [PubMed: 7851772]
- Ballard JW, Melvins RG, Katewa SD, Maas K. Mitochondrial DNA variation is associated with measurable differences in life-history traits and mitochondrial metabolism in *Drosophila simulans*. *Evolution* 2007;61:1735–1747. [PubMed: 17598752]
- Bazin E, Glemin S, Galtier N. Population size does not influence mitochondrial genetic diversity in animals. *Science* 2006;312:570–572. [PubMed: 16645093]
- Beard CB, Hamm DM, Collins FH. The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Mol Biol* 1993;2:103–124. [PubMed: 9087549]
- Bergstrom CT, Pritchard J. Germline bottlenecks and the evolutionary maintenance of mitochondrial genomes. *Genetics* 1998;149:2135–2146. [PubMed: 9691064]
- Berthier F, Renaud M, Alziari S, Durand R. RNA mapping on *Drosophila* mitochondrial DNA: precursors and template strands. *Nucleic Acids Res* 1986;14:4519–4533. [PubMed: 3086843]
- Birky CW, Walsh JB. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci USA* 1988;85:6414–6418. [PubMed: 3413105]
- Blier PU, Breton S, Desrosiers V, Lemieux H. Functional conservatism in mitochondrial evolution: insight from hybridization of arctic and brook charrs. *J Exp Zool B Mol Dev Evol* 2006;306:425–432.
- Boore JL. Animal mitochondrial genomes. *Nucleic Acids Res* 1999;27:1767–1780. [PubMed: 10101183]
- Boore JL, Brown WM. Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. *Genetics* 1994;138:423–443. [PubMed: 7828825]
- Breton S, Burger G, Stewart DT, Blier PU. Comparative analysis of gender-associated complete mitochondrial genomes in marine mussels (*Mytilus* spp.). *Genetics* 2006;172:1107–1119. [PubMed: 16322521]
- Brown WM, Prager EM, Wang A, Wilson AC. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J Mol Evol* 1982;18:225–239. [PubMed: 6284948]
- Bulmer M. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 1991;129:897–907. [PubMed: 1752426]
- Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. *Genetics* 1993;134:1289–1303. [PubMed: 8375663]
- Clary DO, Wolstenholme DR. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J Mol Evol* 1985;22:252–271. [PubMed: 3001325]
- Crooks G, Hon G, Chandonia J, Brenner S. WebLogo: a sequence logo generator. *Genome Res* 2004;14:1188–1190. [PubMed: 15173120]
- de Bruijn MH. *Drosophila melanogaster* mitochondrial DNA, a novel organization and genetic code. *Nature* 1983;304:234–241. [PubMed: 6408489]
- Doiron S, Bernatchez L, Blier PU. A comparative mitogenomic analysis of the potential adaptive value of Arctic charr mtDNA introgression in brook charr populations (*Salvelinus fontinalis* Mitchell). *Mol Biol Evol* 2002;19:1902–1909. [PubMed: 12411599]
- Drosophila* 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 2007;450:203–218. [PubMed: 17994087]
- Duret L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* 2000;16:287–289. [PubMed: 10858656]
- Edmands S, Burton RS. Cytochrome C oxidase activity in interpopulation hybrids of a marine copepod: a test for nuclear-nuclear or nuclear-cytoplasmic coadaptation. *Evolution* 1999;53:1972–1978.
- Ellison CK, Burton RS. Interpopulation hybrid breakdown maps to the mitochondrial genome. *Evolution* 2008a;62:631–638. [PubMed: 18081717]

- Ellison CK, Burton RS. Genotype-dependent variation of mitochondrial transcriptional profiles in interpopulation hybrids. *Proc Natl Acad Sci USA* 2008b;105:15831–15836. [PubMed: 18843106]
- Felsenstein J. The evolutionary advantage of recombination. *Genetics* 1974;78:737–756. [PubMed: 4448362]
- Friedrich M, Muqim N. Sequence and phylogenetic analysis of the complete mitochondrial genome of the flour beetle *Tribolium castaneum*. *Mol Phylogenet Evol* 2003;26:502–512. [PubMed: 12644407]
- Gabriel W, Lynch M, Burger R. Muller's ratchet and mutational meltdowns. *Evolution* 1993;47:1744–1757.
- Garesse R. *Drosophila melanogaster* mitochondrial DNA: gene organization and evolutionary considerations. *Genetics* 1988;118:649–663. [PubMed: 3130291]
- Gillespie JH. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 2000;155:909–919. [PubMed: 10835409]
- Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 1994;11:725–736. [PubMed: 7968486]
- Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B, Keightley PD. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. *PLoS Biol* 2008;6:e204. [PubMed: 18715119]
- Herbeck JT, Novembre J. Codon usage patterns in *cytochrome oxidase I* across multiple insect orders. *J Mol Evol* 2003;56:691–701. [PubMed: 12911032]
- Hill WG, Robertson A. The effect of linkage on limits to artificial selection. *Genet Res* 1966;8:269–294. [PubMed: 5980116]
- Hoffmann, AA.; Turelli, M. Cytoplasmic incompatibility in insects. In: O'Neill, S.; Hoffmann, A.; Werren, J., editors. *Influential passengers: inherited microorganisms and arthropod reproduction*. Oxford University Press; New York: 1997. p. 42-80.
- Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 2001;17:754–755. [PubMed: 11524383]
- Hutter CM, Rand DM. Competition between mitochondrial haplotypes in distinct nuclear genetic environments: *Drosophila pseudoobscura* vs. *D. persimilis*. *Genetics* 1995;140:537–548. [PubMed: 7498735]
- Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* 1985;2:13–34. [PubMed: 3916708]
- Innan H, Stephan W. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. *Genetics* 2001;159:389–399. [PubMed: 11560913]
- James AC, Ballard JW. Mitochondrial genotype affects fitness in *Drosophila simulans*. *Genetics* 2003;164:187–194. [PubMed: 12750331]
- Kern AD, Kondrashov FA. Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. *Nat Genet* 2004;36:1207–1212. [PubMed: 15502829]
- Ko WY, David RM, Akashi H. Molecular phylogeny of the *Drosophila melanogaster* species subgroup. *J Mol Evol* 2003;57:562–573. [PubMed: 14738315]
- Kopp A, True JR. Phylogeny of the Oriental *Drosophila melanogaster* species group: a multilocus reconstruction. *Syst Biol* 2002;51:786–805. [PubMed: 12396591]
- Kumar S, Tamura K, Nei M. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 2004;5:150–163. [PubMed: 15260895]
- Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;157:105–132. [PubMed: 7108955]
- Lewis DL, Farr CL, Farquhar AL, Kaguni LS. Sequence, organization, and evolution of the A+T region of *Drosophila melanogaster* mitochondrial DNA. *Mol Biol Evol* 1994;11:523–538. [PubMed: 8015445]
- Lewis RL, Beckenbach AT, Mooers AØ. The phylogeny of the subgroups within the *melanogaster* species group: likelihood tests on *COI* and *COII* sequences and a Bayesian estimate of phylogeny. *Mol Phylogenet Evol* 2005;37:15–24. [PubMed: 16182148]

- Li WH. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J Mol Evol* 1987;24:337–345. [PubMed: 3110426]
- Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 1993;36:96–99. [PubMed: 8433381]
- Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997;25:955–964. [PubMed: 9023104]
- Lynch M. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol Biol Evol* 1996;13:209–220. [PubMed: 8583893]
- Lynch M, Koskella B, Schaack S. Mutation pressure and the evolution of organelle genomic architecture. *Science* 2006;311:1727–1730. [PubMed: 16556832]
- Machado CA, Hey J. The causes of phylogenetic conflict in a classic *Drosophila* species group. *Proc Biol Sci* 2003;270:1193–1202. [PubMed: 12816659]
- Maddison, DR.; Maddison, WP. MacClade 4: analysis of phylogeny and character evolution Version 4.0. Sinauer Associates; Sunderland, MA: 2000.
- Martin AP. Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. *Mol Biol Evol* 1995;12:1124–1131. [PubMed: 8524045]
- Meiklejohn CD, Montooth KL, Rand DM. Positive and negative selection on the mitochondrial genome. *Trends Genet* 2007;23:259–263. [PubMed: 17418445]
- Moriyama EN, Powell JR. Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *J Mol Evol* 1997;45:378–391. [PubMed: 9321417]
- Muller H. The relation of recombination to mutational advance. *Mutat Res* 1964;1:2–9. [PubMed: 14195748]
- Musto H, Romero H, Zavala A, Jabbari K, Bernardi G. Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. *J Mol Evol* 1999;49:27–35. [PubMed: 10368431]
- Nachman MW. Deleterious mutations in animal mitochondrial DNA. *Genetica* 1998;102–103:61–69.
- Neiman M, Taylor DR. The causes of mutation accumulation in mitochondrial genomes. *Proc Biol Sci* 2009;276:1201–1209. [PubMed: 19203921]
- Nigro L, Solignac M, Sharp PM. Mitochondrial DNA sequence divergence in the *Melanogaster* and oriental species subgroups of *Drosophila*. *J Mol Evol* 1991;33:156–162. [PubMed: 1920452]
- Ojala D, Montoya J, Attardi G. tRNA punctuation model of RNA processing in human mitochondria. *Nature* 1981;290:470–474. [PubMed: 7219536]
- Pamilo P, Bianchi NO. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol Biol Evol* 1993;10:271–281. [PubMed: 8487630]
- Pamilo P, Nei M. Relationships between gene trees and species trees. *Mol Biol Evol* 1988;5:568–583. [PubMed: 3193878]
- Parsch J, Meiklejohn CD, Hauschteck-Jungen E, Hunziker P, Hartl DL. Molecular evolution of the *ocnus* and *janus* genes in the *Drosophila melanogaster* species subgroup. *Mol Biol Evol* 2001;18:801–811. [PubMed: 11319264]
- Pesole G, Gissi C, De Chirico A, Saccone C. Nucleotide substitution rate of mammalian mitochondrial genomes. *J Mol Evol* 1999;48:427–434. [PubMed: 10079281]
- Pollard DA, Iyer VN, Moses AM, Eisen MB. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genet* 2006;2:e173. [PubMed: 17132051]
- Popadin K, Polishchuk LV, Mamirova L, Knorre D, Gunbin K. Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *Proc Natl Acad Sci USA* 2007;104:13390–13395. [PubMed: 17679693]
- Powell JR. Interspecific cytoplasmic gene flow in the absence of nuclear gene flow: evidence from *Drosophila*. *Proc Natl Acad Sci USA* 1983;80:492–495. [PubMed: 6300849]
- Rand DM, Kann LM. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol Biol Evol* 1996;13:735–748. [PubMed: 8754210]
- Rand DM, Kann LM. Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. *Genetica* 1998;102–103:393–407.

- Rand DM, Clark AG, Kann LM. Sexually antagonistic cytonuclear fitness interactions in *Drosophila melanogaster*. *Genetics* 2001;159:173–187. [PubMed: 11560895]
- Reed LK, Nyboer M, Markow TA. Evolutionary relationships of *Drosophila mojavensis* geographic host races and their sister species *Drosophila arizonae*. *Mol Ecol* 2007;16:1007–1022. [PubMed: 17305857]
- Remsen J, O'Grady PO. Phylogeny of Drosophilinae (Diptera: Drosophilidae), with comments on combined analysis and character support. *Mol Phylogenet Evol* 2002;24:249–264. [PubMed: 12144760]
- Reyes A, Gissi C, Pesole G, Saccone C. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* 1998;15:957–966. [PubMed: 9718723]
- Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, et al. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res* 2005;15:1–18. [PubMed: 15632085]
- Richly E, Meister D. NUMTs in sequenced eukaryotic genomes. *Mol Biol Evol* 2004;21:1081–1084. [PubMed: 15014143]
- Roberti M, Polosa PL, Bruni F, Musicco C, Gadaleta MN, Cantatore P. DmTTF, a novel mitochondrial transcription termination factor that recognises two sequences of *Drosophila melanogaster* mitochondrial DNA. *Nucleic Acids Res* 2003;31:1597–1604. [PubMed: 12626700]
- Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003;19:1572–1574. [PubMed: 12912839]
- Rozas J, Rozas R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* 1999;15:174–175. [PubMed: 10089204]
- Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 1999;238:195–209. [PubMed: 10570997]
- Saccone C, Gissi C, Lanave C, Larizza A, Pesole G, Reyes A. Evolution of the mitochondrial genetic system: an overview. *Gene* 2000;261:153–159. [PubMed: 11164046]
- Sackton TB, Haney RA, Rand DM. Cytonuclear coadaptation in *Drosophila*: disruption of cytochrome c oxidase activity in backcross genotypes. *Evolution* 2003;57:2315–2325. [PubMed: 14628919]
- Schneider T, Stephens R. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res* 1990;18:6097–6100. [PubMed: 2172928]
- Shoemaker DD, Dyer KA, Ahrens M, McAbee K, Jaenike J. Decreased diversity but increased substitution rate in host mtDNA as a consequence of *Wolbachia* endosymbiont infection. *Genetics* 2004;168:2049–2058. [PubMed: 15611174]
- Tajima F. Simple methods for testing molecular clock hypothesis. *Genetics* 1993;135:599–607. [PubMed: 8244016]
- Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 1993;10:512–526. [PubMed: 8336541]
- Tavare S. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect Math Life Sci* 1986;17:57–86.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997;25:4876–4882. [PubMed: 9396791]
- Valverde JR, Marco R, Garesse R. A conserved heptamer motif for ribosomal RNA transcription termination in animal mitochondria. *Proc Natl Acad Sci USA* 1994;91:5368–5371. [PubMed: 7515499]
- Wong A, Jensen JD, Pool JE, Aquadro CF. Phylogenetic incongruence in the *Drosophila melanogaster* species group. *Mol Phylogenet Evol* 2007;43:1138–1150. [PubMed: 17071113]
- Xia X. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. *Gene* 2005;345:13–20. [PubMed: 15716092]
- Yamauchi MM, Miya MU, Nishida M. Use of a PCR-based approach for sequencing whole mitochondrial genomes of insects: two examples (cockroach and dragonfly) based on the method developed for decapod crustaceans. *Insect Mol Biol* 2004;13:435–442. [PubMed: 15271216]

- Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 1997;13:555–556. [PubMed: 9367129]
- Yang Z, Nielsen R. Synonymous and non-synonymous rate variation in nuclear genes of mammals. *J Mol Evol* 1998;46:409–418. [PubMed: 9541535]
- Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003;31:3406–3415. [PubMed: 12824337]

Fig. 1.

Genome content and organization are conserved across the 12 *Drosophila* mtDNAs. Arrows indicate the coding strand for each element on the major (outer) and minor (inner) strands. Where intergenic regions occur, the range of intergenic lengths (bp) observed across species is given. tRNAs are labeled with the one-letter amino acid code. ND1-6, 4L encode components of NADH dehydrogenase (Complex I); Cyt-B, cytochrome B (Complex III); COI-III, cytochrome C oxidase (Complex IV); ATPase6, 8, ATP synthase (Complex V)

	tRNA ^{Tyr}		Col									
<i>D. mel</i>	acttaatc	-----aataa	tcg	cga	caa	tga	tta	ttt	tct	aca	aat	cat
<i>D. sim</i>	acttaatc	-----agtga	cgc	cga	caa	tga	tta	ttt	tct	aca	aat	cat
<i>D. sec</i>	acttaatc	-----aataa	cgc	cga	caa	tgg	tta	ttt	tct	aca	aat	cat
<i>D. mau</i>	acttaatc	-----attaa	cgc	cga	caa	tga	tta	ttt	tct	aca	aat	cat
<i>D. yak</i>	acttaatc	-----cataa	tcg	cga	caa	tgg	tta	ttt	tct	aca	aat	cat
<i>D. ere</i>	acttaatc	aaaagcggttaa	tcg	cga	caa	tga	tta	ttt	tct	aca	aat	cat
<i>D. ana</i>	acttaa--	-----	tcg	cga	caa	tga	tta	ttt	tca	aca	aat	cat
<i>D. per</i>	acttaa--	-----	tcg	cga	caa	tga	tta	ttt	tct	aca	aat	cat
<i>D. wil</i>	acttaa--	-----	tcg	cga	caa	tgg	tta	ttc	tca	aca	aat	cat
<i>D. moj</i>	acttaa--	-----	tcg	cga	caa	tgg	tta	ttc	tca	aca	aat	cat
<i>D. vir</i>	acttaa--	-----	tcg	cga	caa	tga	tta	ttt	tca	aca	aat	cat
<i>D. gri</i>	acttaa--	-----	tcg	cga	caa	tga	tta	ttt	tca	aca	aat	cat
<i>A. gam</i>	atttaa--	-----	tcg	cga	caa	tga	tta	ttt	tca	aca	aat	cat
			S/P	R	Q	W	L	F	S	T	N	H
	conserved, <i>Tribolium</i>					*	*	*	*	*	*	*
	conserved, <i>Apis</i>					*			*	*	*	*

Fig. 2.

5' RACE data suggest a novel start codon for *Drosophila CoI*. The previously annotated start codon for *D. melanogaster* and *D. yakuba* is underlined. 5' RACE sequence data are shaded and begin with a TCG or CCG (boxed). Outside of the *melanogaster* subgroup, the coding regions for *CoI* and tRNA^{Tyr} overlap by two nucleotides. Also shown are the conserved nucleotide sequence of *Anopheles gambiae* and the conserved amino acids with *Tribolium* and *Apis*

Fig. 3.

Functional constraint varies across intergenic and regulatory regions of the *Drosophila* mtDNA. **a** A representative non-coding region that is divergent in both sequence and length. **b** Two regions previously annotated as non-coding that are nearly perfectly conserved and are included in the mature mRNA transcript. Shaded sequences indicate 5' RACE data. The putative new start codon inferred from 5' RACE data is boxed. **c** The rRNA transcription termination signal embedded in the tRNA^{Leu(CUN)} gene is perfectly conserved. **d** Two binding regions for the *D. melanogaster* terminator transcription factor (DmTTF) share three nearly invariant motifs that are shaded and can be seen in the sequence logo (Schneider and Stephens 1990) generated by WebLogo (Crooks et al. 2004). **e** The *ATPase6* and *ColIII* gene junction contains a putative secondary structure that may signal splicing of the mRNA (indicated by an

arrow) but is not conserved across the phylogeny. All sequences are 5' to 3' on their coding strand. Minor strand, *top*; major strand, *bottom*

Fig. 4.

Evolutionary rates are heterogeneous across elements of the mtDNA. **a** Pairwise Tamura–Nei nucleotide distances for different categories of sites. **b** Pamilo–Bianchi–Li estimates of synonymous and non-synonymous distances for each OXPHOS complex reveal heterogeneous rates of non-synonymous divergence across complexes, but relative homogeneity at silent sites. Plotted are the estimates between *D. melanogaster* and each of the 11 other species, as well as the average pairwise distance across all species comparisons (*right*)

Fig. 5.

Evolutionary rates are OXPHOS complex- and lineage-specific. **a** Distance trees (substitutions/codon) for the concatenated genes of ND and CO estimated from a codon-based substitution model that allows three discrete site classes with different ω . Topologies are from the Bayesian phylogenetic analysis reported in this study. Numbers represent ω and the number of amino acid changes, respectively, estimated from a model that allows ω to vary across branches (Table 5). Highlighted are the relatively long branches of *D. persimilis* CO and *D. mojavenensis* ND. **b** Amino acid alignments represent the most divergent window of 70 amino acids of *CoI* and *ND6* and exemplify the difference in evolutionary rate between these two complexes. An asterisk indicates a premature stop codon in *D. persimilis* *ND6*. The perfect conservation downstream suggests that this codon does not function as a translational stop

Fig. 6.

Inferred changes mapped onto ancestral tRNAs indicate strong maintenance of secondary structure. Changes in loops are in *orange* with the number of hits at each site indicated. **a, b** Both tRNA^{Leu} genes contain a perfectly conserved tridecamer rRNA transcription termination sequence indicated in *blue*. The lesser used tRNA^{Leu(CUN)} has nearly twice the number of inferred changes on the phylogeny than the more highly used tRNA^{Leu(UUR)}. **c, d** tRNA^{Asn} and tRNA^{Ser} are the most diverged and conserved tRNAs, respectively. Pairs of changes at a Watson–Crick stem pair in tRNA^{Asn} that occurred on a single branch of the phylogeny include those that potentially traversed an intermediate that maintains G–U pairing (*green*), as well as those for which the Watson–Crick pairing must have been disrupted (*red*). These categories

of change correspond to the counts in Table 3. *RSCU*, relative synonymous codon usage. (Color figure online)

Fig. 7.

Three peaks of d_N and d_N/d_S in the mitochondrial proteome overlap regions encoding components of NADH dehydrogenase (ND2,4,5 and 6). **a** Sliding window plots of d_N , d_S and d_N/d_S between *D. melanogaster* and either *D. yakuba*, *D. persimilis* or *D. mojavensis* across the concatenated protein-coding genes. Sliding windows contained 250 sites and a step size of 25 sites. **b** $\omega = d_N/d_S$ estimated from Model 0 in PAML for each protein-coding gene across the *melanogaster* species group (7 spp) or the entire phylogeny (12 spp)

Fig. 8.

ND6 (a) and COI (b) hydropathy profiles (Kyte and Doolittle 1982) are conserved between *D. melanogaster*, *D. persimilis*, and *D. mojavensis*. However, *ND6* sequence divergence along the *D. mojavensis* lineage is manifested as a decrease in hydrophobicity between amino acids 100–125. A region of perfect conservation was removed from the COI profile in order to preserve the scale. Hydrophobic regions have hydropathy values >1. *Horizontal bars* refer to the region of amino acid alignment presented in Fig. 5. Window size is 11 amino acids

Table 1

Features of the 12 *Drosophila* mitochondrial genomes

Species	Genome data source	No. traces	Coverage (max)	Length (bp) ^a	Non- coding (bp) ^b	% A+T					Codon Usage (ENC)		
						Overall	Coding	Codon position			tRNA	rRNA	
								1st	2nd	3rd			
Subgenus <i>Sophophora</i>													
<i>D. melanogaster</i>	NC_001709 ^c	–	–	14916	163	78.0	77.1	70.2	66.7	94.6	77.1	82.0	30.21
<i>D. simulans</i> sIII	AF200840 ^c	–	–	14946	187	77.6	76.7	70.1	66.4	93.7	76.9	81.6	30.86
<i>D. sechellia</i>	NC_005780 ^c	–	–	14950	194	77.5	76.6	69.8	66.2	93.7	76.8	81.6	30.57
<i>D. mauritiana</i> mAll	AF200830 ^c	–	–	14964	207	77.7	76.8	69.8	66.3	94.3	76.1	81.9	30.73
<i>D. yakuba</i>	X03240 ^d	–	–	14942	181	77.6	76.6	69.5	66.3	94.0	76.5	81.8	30.74
<i>D. erecta</i>	Trace Archive ^e	517	16.6x (35x)	14952	200	77.2	76.4	69.7	66.4	92.8	75.7	81.4	31.29
<i>D. ananassae</i>	Trace Archive ^e	563	21.3x (46x)	14920	172	77.4	76.7	69.5	66.6	93.8	76.0	81.0	31.41
<i>D. persimilis</i>	Trace Archive ^e	252	17.3x (28x)	14930	179	77.3	76.4	69.3	66.4	93.5	76.1	81.3	31.89
<i>D. willistoni</i>	Trace Archive ^e	1094	53.4x (100x)	14915	167	77.2	76.3	69.0	66.7	93.5	76.0	81.2	31.33
Subgenus <i>Drosophila</i>													
<i>D. mojavensis</i>	Trace Archive ^e	890	33.4x (51x)	14904	150	76.5	75.3	68.9	66.4	90.6	76.0	81.2	33.76
<i>D. virilis</i>	Trace Archive ^e	614	16.7x (31x)	14949	200	76.7	75.6	68.8	66.4	91.7	76.5	81.4	33.11
<i>D. grimshawi</i>	Trace Archive ^e	430	15.7x (38x)	14874	123	76.7	75.7	68.6	66.4	92.0	76.8	81.4	33.36

^aExcluding the A+T-rich control region

^bUsing new annotation of coding regions reported in this article

^cBallard (2000)

^dClary and Wolstenholme (1985)

^e*Drosophila* 12 Genomes Consortium (2007)

Table 2Average pairwise distances across the 12 *Drosophila* mtDNAs

Sequence element	Nucleotide ^a	Non-synonymous (d_N) ^b	Synonymous (d_S) ^b	d_N/d_S ^c
Entire mtDNA ^d	0.123	–	–	
rRNA	0.064	–	–	
lrRNA	0.058	–	–	
srRNA	0.075	–	–	
tRNA	0.055	–	–	
Proteome	0.145	0.045	0.653	0.069
1st position	0.086	–	–	
2nd position	0.023	–	–	
3rd position	3.695	–	–	
Complex I (ND)	0.152	0.059	0.639	0.091
Complex III (CytB)	0.158	0.037	0.782	0.048
Complex IV (CO)	0.132	0.020	0.660	0.031
Complex V (ATPase)	0.147	0.042	0.642	0.072

^a Average Tamura–Nei distances ($\gamma = 0.5$) across all pairwise comparisons^b Average Pamilo–Bianchi–Li distances across all pairwise comparisons^c Average d_N/d_S from Pamilo–Bianchi–Li distances across all pairwise comparisons^d Excluding the A+T-rich control region and a small number of unalignable intergenic nucleotides

Table 3
Divergence of tRNA sequences and conservation of tRNA secondary structure across the 12 *Drosophila* mtDNAs

tRNA	Length (bp)	Fraction of basepairs in stems	Usage ^d	Avg. no. pairwise differences	Inferred number of changes on the phylogeny ^b			
					Total	Fraction that occur in stems	Fraction that maintain pairing	No. paired stem changes ^c No. compensatory changes ^d
Major strand								
I, Ile	65	0.58	9.61	5.3	26	0.46	1.00	4
M, Met	69	0.58	5.67	2.0	12	0.08	1.00	
W, Trp	68	0.62	2.72	6.6	34	0.38	0.92	2
L, Leu (UUR)	67	0.60	15.1	1.9	9	0.89	0.88	3
K, Lys	71	0.54	2.30	3.0	20	0.15	0.67	1 1
D, Asp	69	0.61	1.77	2.4	24	0.00		
G, Gly	65	0.65	5.96	1.7	12	0.50	1.00	2 1
A, Ala	66	0.61	4.67	1.9	14	0.21	1.00	
R, Arg	66	0.61	1.60	4.6	30	0.33	0.80	2 1
N, Asn	67	0.57	5.53	6.1	41	0.24	0.80	4 1
S, Ser (AGN)	68	0.56	2.83	1.3	8	0.50	1.00	2
E, Glu	69	0.58	2.13	4.9	36	0.17	1.00	2 1
T, Thr	65	0.62	4.95	1.8	12	0.17	1.00	1
S, Ser (UCN)	67	0.60	6.25	1.6	10	0.10	0.00	
Total major	942	0.59		40.1	288	0.27	0.90	23 5
Minor strand								
V, Val	72	0.58	5.20	1.5	8	0.63	1.00	
L, Leu (CUN)	65	0.62	1.55	2.6	18	0.61	0.82	1
P, Pro	68	0.62	3.51	6.9	13	0.31	1.00	1
H, His	67	0.63	2.08	2.9	26	0.19	1.00	1
F, Phe	67	0.60	8.94	6.8	36	0.53	0.95	8 2
Y, Tyr	66	0.64	4.51	2.9	20	0.15	1.00	
C, Cys	64	0.63	1.11	3.6	24	0.38	1.00	1
Q, Gln	69	0.64	1.93	2.5	15	0.33	0.80	1
Total minor	538	0.62		21.3	160	0.38	0.93	13 2
Overall	1480	0.60		61.3	448	0.31	0.91	36 7

^aFrequency (%) of amino acid in proteome. Usage of Leu and Ser was calculated as this frequency multiplied by the average relative synonymous codon usage for the two tRNAs

^bInferred using ancestral state reconstruction for each tRNA at each node in the phylogeny

^cInstances where both nucleotides of a stem pair change on a single lineage

^dPaired stem changes that must have gone through a state of disrupted pairing (see text and Fig. 6c)

Table 4
Strand-biased synonymous substitution patterns within non-overlapping coding regions on the major and minor strands

Substitution class	Six species						Twelve species					
	Major strand ^d			Minor strand ^d			Major strand			Minor strand		
	Subs ^b	Nuc ^c	Subs/nuc	Subs	Nuc	Subs/nuc	Subs	Nuc	Subs/nuc	Subs	Nuc	Subs/nuc
A → G	0.187	46.1	0.0041	0.416	43.5	0.0096	0.125	45.9	0.0027	0.309	43.5	0.0071
G → A	0.009	1.8	0.0050	0.050	4.4	0.0114	0.005	2	0.0025	0.021	5.1	0.0041
A ↔ G	0.196	47.9	0.0041	0.466	47.9	0.0097	0.130	47.9	0.0027	0.331	48.6	0.0068
C → T	0.042	5	0.0084	0	1.1	0.0000	0.028	5.5	0.0051	0	1.2	0.0000
T → C	0.392	47	0.0083	0.173	51	0.0034	0.299	46.6	0.0064	0.120	50.3	0.0024
C ↔ T	0.434	52	0.0083	0.173	52.1	0.0033	0.327	52.1	0.0063	0.120	51.5	0.0023
A ↔ G/C ↔ T	0.45		0.49	2.69		2.94	0.40		0.43	2.76		2.96

^aConcatenated sets of genes encoded on the major and minor strands of the mtDNA do not overlap
^bFrequency of all unambiguous polarized substitutions inferred using a 6 (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. mauritiana*, *D. yakuba*, *D. erecta*) or 12 species phylogeny
^cPercent nucleotide frequency of the initial nucleotide or, for the unpolarized class, the sum of the nucleotide frequencies

Table 5

Evolutionary rates across the *melanogaster* species group estimated by codon-based substitution models

Sequence subset	No. codons	Model ^a	λ	Estimated parameters	Best ω^b	Tree length (d_N)
Proteome	3704	M0	-24254	$\omega = 0.0096$		0.058
		M3	-23964***	$\omega_{0,1,2} = 0.0012, 0.056, 0.237$ $p_{0,1,2} = 0.870, 0.114, 0.015$	0.0111	0.070
		Branch	-24238**	$\omega_{\min} = 0.0069, \omega_{\max} = 0.0183$		0.058
		Complex	-23845***	$\omega_{I,III,IV,V} = 0.0105, 0.0059, 0.0046, 0.0090$		
Complex I (ND)	2054	M0	-13430	$\omega = 0.0106$		0.079
		M3	-13240***	$\omega_{0,1,2} = 0, 0.0294, 0.2068$ $p_{0,1,2} = 0.732, 0.241, 0.027$	0.0126	0.101
		Branch	-13418*	$\omega_{\min} = 0.0079, \omega_{\max} = 0.024$		0.079
		M0	-2395	$\omega = 0.0055$		0.044
Complex III (CytB)	377 ^c	M3	-2380***	$\omega_{0,1,2} = 0, 0, 0.038$ $p_{0,1,2} = 0.407, 0.437, 0.156$	0.0059	0.051
		M0	-6323	$\omega = 0.004$		0.023
		M3	-6275***	$\omega_{0,1,2} = 0, 0.0005, 0.079$ $p_{0,1,2} = 0.450, 0.499, 0.051$	0.0043	0.025
		Branch	-6308**	$\omega_{\min} = 0, \omega_{\max} = 0.016$		0.023
Complex IV (CO)	999	M0	-1683	$\omega = 0.0052$		0.054
		M3	-1649***	$\omega_{0,1,2} = 0, 0.0044, 0.179$ $p_{0,1,2} = 0.680, 0.284, 0.036$	0.0076	0.138

^aM0, single $\omega = d_N/d_S$ for all branches and sites; M3, ω estimated for three distinct site classes each containing proportion p of the total number of sites; branch, ω estimated for each branch; complex, ω estimated for each OXPHOS complex with the transition/transversion ratio fixed at the value estimated for the proteome by M0 (Parsch et al. 2001). Tests of model significance are against M0, using $\chi^2 = -2\Delta\lambda$.

^b ω estimated from the best fitting model. For model 3 this is the weighted average across the three site classes

^cBranch-specific rate model was not run due to the small number of codons

* $P < 0.01$;

** $P < 0.001$;

*** $P < 0.0001$