J Mol Evol (1996) 43:116–124

JOURNAL OF MOLECULAR EVOLUTION

Phylogenetic Analysis of DNA Length Mutations in a Repetitive Region of the Hawaiian *Drosophila* Yolk Protein Gene *Yp2*

Kin-Fan Ho,[1] Elysse M. Craddock,[2] Fabio Piano,[1] Michael P. Kambysellis[1]

[1] Department of Biology, New York University, 1009 Brown Building, Washington Square, New York, NY 10003, USA
[2] Division of Natural Sciences, Purchase College, State University of New York, 735 Anderson Hill Road, Purchase, NY 10577-1400, USA

Received: 7 December 1995 / Accepted: 4 March 1996

Abstract. Nucleotide sequence analysis has demonstrated that interspecific size variation in the YP2 yolk protein among Hawaiian *Drosophila* is due to in-frame insertions and deletions in two repetitive segments of the coding region of the *Yp2* gene. Sequence comparisons of the complex repetitive region close to the 5′ end of this gene across 34 endemic Hawaiian taxa revealed five length morphs, spanning a length difference of 21 nucleotides (nt). A phylogenetic character reconstruction of the length mutations on an independently derived molecular phylogeny showed clade-specific length variants arising from six ancient events: two identical insertions of 6 nt, and four deletions, one of 6 nt, one of 12 nt, and two identical but independent deletions of 15 nt. These mutations can be attributed to replication slippage with nontandem trinucleotide repeats playing a major role in the slipped-strand mispairing. Geographic analysis suggests that the 15 nt deletion which distinguishes the *planitibia* subgroup from the *cyrtoloma* subgroup occurred on Oahu about 3 million years ago. The homoplasies observed caution against relying too heavily on nucleotide insertions/deletions for phylogenetic inference. In contrast to the extensive repeat polymorphisms within other *Drosophila* and the human species, the more complex 5′ *Yp2* repetitive region analyzed here appears to lack polymorphism among Hawaiian *Drosophila,* perhaps due to founder effects, low population sizes, and hitchhiking effects of selection on the immediately adjacent 5′ region.

Correspondence to: M.P. Kambysellis

Key words: Hawaiian *Drosophila* — Yolk protein genes — Trinucleotide repeats — Nucleotide length mutations — Nucleotide insertions/deletions — Replication slippage — Founder effects — Genetic hitchhiking

Introduction

Repetitive DNA is ubiquitous in eukaryotic genomes, with a substantial fraction composed of simple sequence repeats (SSR) organized in tandem and distributed in both coding and noncoding regions of the genome. The length, location, and copy number of tandem repeats vary, distinguishing minisatellites (Jeffreys et al. 1985) and microsatellites (Litt and Luty 1989) from the blocks of highly repetitive or satellite DNA often found in the heterochromatin. In microsatellite regions, mono-, di-, tri-, and tetranucleotide motifs are the most common (Litt and Luty 1989). By virtue of their tandem nature, all classes of satellite DNAs are prone to expansion or contraction in the number of repeats as a result of mispairing, generating length mutations and oftentimes length polymorphisms among individuals of a population or species (Nakamura et al. 1987; Tautz 1989; Weber and May 1989; Caskey et al. 1992). Repeat number evolves when unequal crossing over (Smith 1976) follows misalignment between DNA strands, or when slipped-strand mispairing occurs during DNA replication (Streisinger et al. 1966; Levinson and Gutman 1987). Replication slippage seems to be a relatively frequent phenomenon in certain repetitive-sequence regions; these mutable regions are evolutionarily important components of the ge-

nome, and a major source of genetic variation (Jones and Kafatos 1982; Tautz et al. 1986).

To explore patterns in the evolution of repetitive regions in the expressed portion of the genome, we have undertaken a phylogenetic analysis of the repetitive segment in the 5′ coding sequences of the *Yp2* yolk protein gene among Hawaiian *Drosophila.* The YP2 protein size variation among Hawaiian species (Craddock and Kambysellis 1990) was the initial clue that the corresponding gene may have undergone a number of insertion/deletion mutations, a hypothesis that has now been confirmed by nucleotide sequence comparisons. Sequence analysis has revealed that a stretch of 75 nt just downstream of the 5′ end of the *Yp2* gene is rich in repeats with ~90% of the nucleotides forming tandem or interspersed di-, tri-, and hexanucleotide repeats. These repeats have generated an evolutionary hot spot for DNA length mutations which translates into variability in the number of residues at the amino end of the mature YP2 protein, contributing to the interspecific protein size variation.

The wealth of information available on the evolution and biogeography of the endemic Hawaiian *Drosophila* (Carson and Kaneshiro 1976) in conjunction with our phylogenetic analyses of this repetitive region has allowed us to identify the lineages in which particular insertion/deletion mutations have occurred, and in some cases to estimate the geological time frame and the geographic location of specific molecular events. Models of the replication slippage patterns responsible for the DNA length mutations are also proposed. Although this case from Hawaiian *Drosophila* exemplifies the phenomenon of replication slippage and validates its role in genome evolution, our observations differ markedly from previous observations of DNA length variation in *Drosophila* (Costa et al. 1991; Hey and Kliman 1993) and other organisms (Jones and Kafatos 1982; Nakamura et al. 1987; Tautz 1989) in the apparent lack of length polymorphism within the Hawaiian species. We present some possible explanations for this atypical pattern.

Materials and Methods

Experimental Material. Adult flies of 34 taxa of endemic Hawaiian *Drosophila* were obtained from field collections or from laboratory stocks maintained at the University of Hawaii or at Purchase College. Species included all 17 species of the *planitibia* species group, with multiple samples of two of these species, namely, *D. neopicta* (individuals from three populations from Maui and two populations from Molokai) and *D. silvestris* (laboratory isofemale lines U28T2, U26B9, and U34B4 from the Kilauea Forest, Kahuku, and Kohala populations, respectively). Other representatives of the picture-winged group were from the *primaeva* (*D. primaeva,* Kauai), *adiastola* (*D. adiastola,* Maui), and *grimshawi* species groups; the latter included the species *D. mulli* and *D. pullipes,* both from Hawaii, and *D. grimshawi* from the Kauai, East Maui, West Maui, and Lanai populations. Non–picture-winged representatives were from the *antopocerus* (*D. adunca,* Maui) and modified-mouthparts species groups (*D. mimica* and *D. infuscata,* both from Hawaii). Genomic DNA was isolated from individual males using the GNOME Kit (Bio 101) according to the supplier's recommendations.

PCR Amplification and Nucleotide Sequence Determination. The 5′ portion of the *Yp2* gene (~900 nt) was PCR-amplified (Saiki et al. 1985) in a Coy thermocycler from genomic DNA of single males using *Taq* polymerase (Perkin-Elmer Cetus) according to the supplier's specifications. The primers were V_{21} (5′-GCAGTACGGTTTGGTAC-3′), corresponding to the segment extending from +3 to +19 of the *D. planitibia Yp2* gene, and V_{24} (5′-GTCCGGCGACACCGGCAACGT-3′), located at +903 to +883. The PCR profile was: initial denaturation at 95°C for 2 min, followed by 30 cycles with denaturation at 95°C for 1.5 min, annealing at 55°C for 1.5 min, and extension at 72°C for 2 min. The amplified *Yp2* product included most of the leader, the first exon, the intron, and half of the second exon. An ~300 nt 5′ segment of this amplified product was sequenced in both directions using the CircumVent system (New England BioLabs) with the forward primer V_{21} and, for the opposite strand, the internal primer V_{28} (5′-CGAGAAGCGTGGC-3′), located at +274 to +262 of the *D. planitibia Yp2* gene, 9 nt upstream of the 5′ splice site. The nucleotide positions of the *D. planitibia* gene were determined from the alignment with the *D. grimshawi Yp2* gene for which the CAP site and intron splice sites have been experimentally determined (Parisi 1994). The signal sequence and the 3′ end sequence were identified by computer analysis, the latter being identified by the first in-frame stop codon. DNA sequences were aligned using CLUSTAL V (Higgins et al. 1992).

Phylogenetic Character Analysis. For analysis of the pattern of nucleotide insertions/deletions, we first applied maximum parsimony analysis (Swofford 1993) to ~1 kb of sequence data (1,028 nt) from the *Yp1* gene (Kambysellis et al. 1995) combined with 288 nt of the *Yp2* gene (not including the variable-length region) in order to derive a phylogenetic tree of the Hawaiian species. The resulting tree was the strict consensus of two equally parsimonious trees (length = 601; Consistency Index = 0.759; Retention Index = 0.786) that differ in the placement of the *Picticornis* subgroup relative to the *Grimshawi* group and the remaining members of the *Planitibia* group. The five distinct *Yp2* length morphs observed among the Hawaiian species analyzed were coded with reference to the length morph of *D. primaeva,* the most primitive picture-winged species (Carson and Kaneshiro 1976), which was arbitrarily assigned character state 0. Length differences of the corresponding sequence in each of the other species were interpreted as insertions (+) or deletions (−) of a specific number of nucleotides. The character states were then optimized on the independent molecular sequence tree using MacClade (Maddison and Maddison 1992).

Results

Role of Nucleotide Insertions/Deletions in Yolk Protein Size Differences

The three *Drosophila* vitellogenin or yolk proteins (YP1, YP2, and YP3) show extensive interspecific variability in their electrophoretic mobility when analyzed by SDS/PAGE (Craddock and Kambysellis 1990). Most commonly, these proteins are resolved in a three-band pattern with each major electrophoretic band representing one of the three yolk proteins. In a few species, electrophoresis of egg extracts results in a two-band pattern, due to co-electrophoresis of two of the three proteins. Such electrophoretic shifts could be due to variation in the lengths

a) 5' repetitive region of the *Yp2* gene

```
                        T   N   A   G   N   G   N   G   N   G   H   R
D. grimshawi    116  ACA AAT GCC GGC AAT GGC AAT GGC AAT GGT CAC AGA
D. planitibia   115  ... ... --- --- --- --- --- --- --- ... ... .A.
```

b) 3' repetitive region of the *Yp2* gene

```
                        R   Q   Q   Q   R   Q   Q   Q   Q   Q   K
D. grimshawi   1407  AGG CAG CAG CAG AGG CAG CAG CAG CAG CAG AAG
D. planitibia  1395  ... ... ... ... --- --- --- ... ... ... ...
```

Fig. 1. Nucleotide sequence comparison of the repetitive regions near **(A)** the 5′ end and **(B)** the 3′ end of the *Yp2* gene between the Hawaiian species *D. grimshawi* and *D. planitibia,* demonstrating the length difference between the two genes. *Periods* indicate identical nucleotides in the two species; gaps in the alignment are indicated by *dashes.* Repeated sequences in each region are *bracketed.* The encoded amino acids are identified by the *single-letter code* above the first nucleotide in each codon. The *D. grimshawi* sequence is from Parisi (1994). The numbers preceding each sequence identify the position of the first nucleotide shown in the *Yp2* gene of that species. In **A** the nucleotide positions differ by one because of the insertion of one nucleotide in the leader sequence of *D. grimshawi.* The 5′ segment shown here starts with the first codon immediately following the sequence encoding the signal peptide (i.e., the beginning of region II in Fig. 2); the first amino acid shown thus corresponds to the N-terminal of the mature processed protein.

of the genes encoding these proteins and/or to post-translational modifications of the yolk proteins.

To address this question, we first compared the DNA sequences of the *Yp2* gene of two Hawaiian *Drosophila* species, *D. grimshawi,* a three-band species, and *D. planitibia,* a two-band species in which the YP2 and YP3 proteins coelectrophorese. The aligned sequences of the coding region are interrupted in two positions, one near the 5′ end of the gene and the other near the 3′ end (30 nt from the TAA stop codon), as a result of two short deletions (21 nt and 9 nt, respectively) in the *D. planitibia* sequence (Fig. 1). These in-frame deletions correspond to 10 aa, generating an estimated 1.3-kDa molecular weight difference that is largely responsible for the electrophoretic shift observed between the YP2 proteins of these two Hawaiian species. Interestingly, both length mutations in the *Yp2* gene are in repetitive regions. The 3′ sequence is rich in d(CAG) trinucleotides (aa Gln), while at the 5′ end, a d(GGC-AAT) hexanucleotide (aa Gly-Asn) is involved.

Interspecific Nucleotide Length Variation in the 5′ Repetitive Region of the Yp2 Gene

In this paper we analyze the more complex repetitive region at the 5′ end of the *Yp2* gene. Figure 2 presents aligned sequences from an approximately 100-nt portion of the amplified and sequenced region (which begins at the first position of the sixth codon of the gene) for 34 Hawaiian taxa; 23 of these are members of the *planitibia* species group that exhibits both two- and three-band electrophoretic profiles. The sequence composition of this DNA fragment is remarkably complex, consisting almost exclusively of a variety of simple sequence repeats (SSR), including di-, tri-, and hexanucleotide motifs, some in tandem but many interspersed. For the purposes of discussion, we divide the sequence into three regions.

Region I (42 nt) extends to the end of the signal peptide and consists almost entirely of consecutive, interspersed, or overlapping SSRs, but lacks length variation in Hawaiian *Drosophila* despite the rich array of repeats. Likewise, region III lacks length variation, although it includes two copies of the d(AAG) trinucleotide and four copies of the d(AAC) trinucleotide (see shaded blocks, Fig. 2). Only region II shows nucleotide length variation: All of the insertions and deletions are in-frame, and the difference between the longest and shortest gene in this region is 21 nt, or seven codons. There are five discrete length morphs among the Hawaiian species analyzed, and in general, closely related species share the same length morph. Notably, there is no length polymorphism among individuals or populations of the three multiply sampled species, *D. grimshawi, D. neopicta,* and *D. silvestris.* Much of the interspecific nucleotide length variation in this gene region is due to variable numbers (zero, two, or three) of the consensus hexanucleotide d(GGC-AAT).

Phylogenetic Tracing of the Yp2 Nucleotide Length Variants

In order to trace the evolution of the nucleotide events responsible for the interspecific length differences, we have undertaken a phylogenetic analysis. The length variation is treated as a set of discrete characters coded as

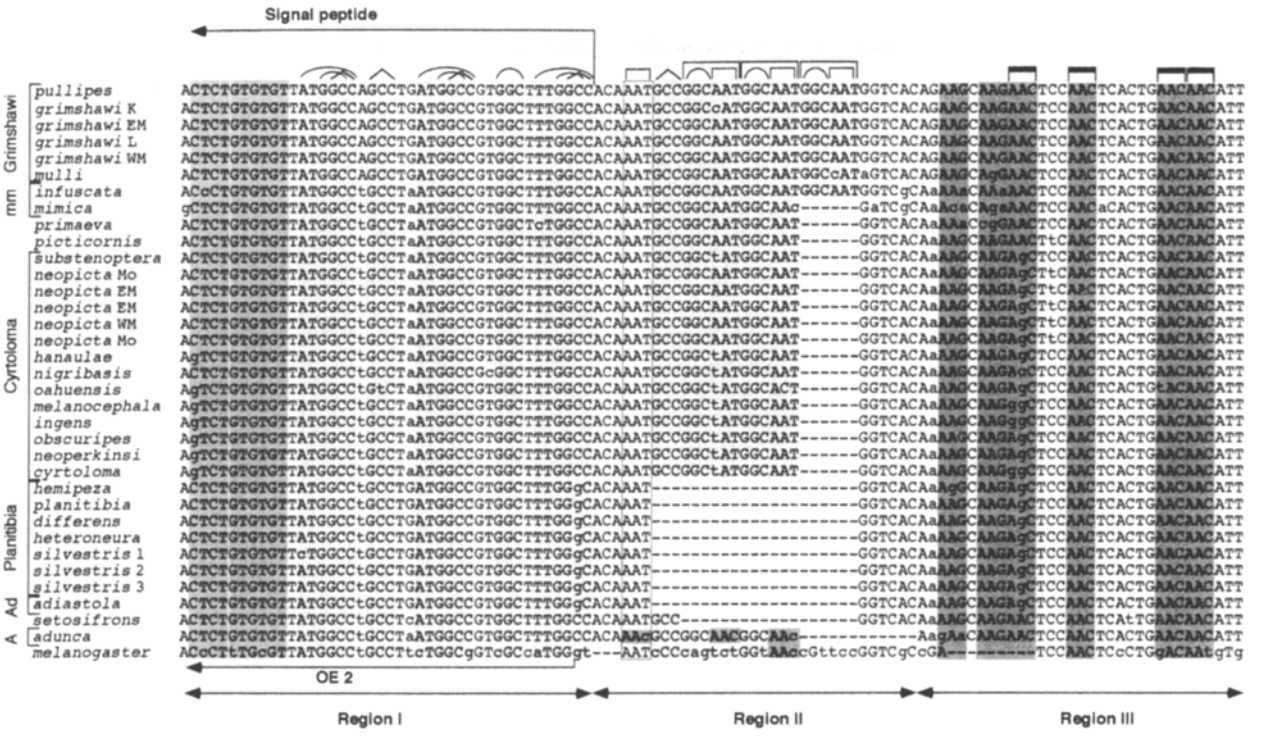

Fig. 2. Aligned nucleotide sequences near the 5′ end of the *Yp2* gene from 34 Hawaiian taxa and *D. melanogaster* (Hung and Wensink 1983), showing the numerous repeats indicated by the *shaded blocks.* Four specific trinucleotides and two different hexanucleotide repeats are identified by the *symbols* above the first sequence. For the most part species are grouped according to subgroup (see *far left*), but they are ordered primarily by length rather than strictly phylogenetically. The four *D. grimshawi* sequences are derived from populations on the islands of Kauai (*K*), East Maui (*EM*), Lanai (*L*), and West Maui (*WM*). The *D. neopicta* sequences are from Molokai (*Mo*), East Maui (*EM*), and West Maui (*WM*). The three *D. silvestris* sequences are from three geographic populations on the island of Hawaii (see Materials and Methods). For the purposes of discussion, the nucleotide segment is divided into regions I, II, and III (see text). The SSR-rich region I encodes the terminal portion of the signal peptide (Hung and Wensink 1983) and part of the ovarian enhancer OE2 (Logan et al. 1989). Although repeats (AAG/C) are also prevalent in region III, length variation is restricted to region II, which begins with the codon corresponding to the N-terminal of the mature processed protein. The partial sequences shown above comprise about a third of the sequences available from Genbank under Accession Nos. U61697–U61722.

the insertion (+) of deletion (−) of a specific number of nucleotides relative to the sequence of *D. primaeva,* and these characters are then traced (Maddison and Maddison 1992) on the independent *Yp1–Yp2* tree. The conclusions presented below are unaffected by the topological differences between the two equally parsimonious trees that are the basis for the consensus tree. Character analysis of the length variation in the repetitive region II (Fig. 3) demonstrates that each insertion/deletion is a discrete event, and the resulting length variant is shared by all members of a clade. For example, in the *planitibia* species group for which all 17 species have been analyzed, the ten species of the *cyrtoloma* subgroup are uniformly of the same length as the *D. primaeva* sequence (Fig. 3), whereas the five species of the *planitibia* subgroup share a sequence that is 15 nucleotides shorter, due to deletion of a d(GCC) trinucleotide and two d(GGC-AAT) hexanucleotide repeats (see Fig. 2) in the ancestral population that led to this subgroup. The two remaining species of the anomalous *picticornis* subgroup depart from this pattern: *D. setosifrons* differs from *D. picticornis* by a deletion of two hexanucleotide repeats (−12).

In addition to the nucleotide deletions of −12 and −15 identified in the *planitibia* species group, our phylogenetic analysis reveals two other deletions and two insertion events among this sampling of Hawaiian *Drosophila* species. Among the more primitive non–picture-wings, there is one insertion (+6) and one deletion (−6). *D. adunca* of the *antopocerus* species group displays a novel DNA length variant due to deletion (−6) of d(GGT-CAC) from the 3′ end of the variable region (Fig. 2). *D. infuscata* of the modified-mouthparts group shows a hexanucleotide insertion of d(GGC-AAT). Remarkably, the same insertion is observed in the picture-winged *grimshawi* group. Figure 3 and the sequences in Fig. 2 clearly show that these two insertions are phylogenetically independent events.

Within the picture-winged group, *D. adiastola* of the *adiastola* species group shows the same length morph as species of the *planitibia* subgroup (Fig. 2) due to a 15-nt deletion from the ancestral sequence of *D. primaeva.* However, phylogenetic analysis (Fig. 3) demonstrates that this deletion cannot be a synapomorphy since *D. adiastola* and the five *planitibia*-subgroup species are differentiated by 57 informative substitutions (33 transversions and 24 transitions in the *Yp1* sequence), and

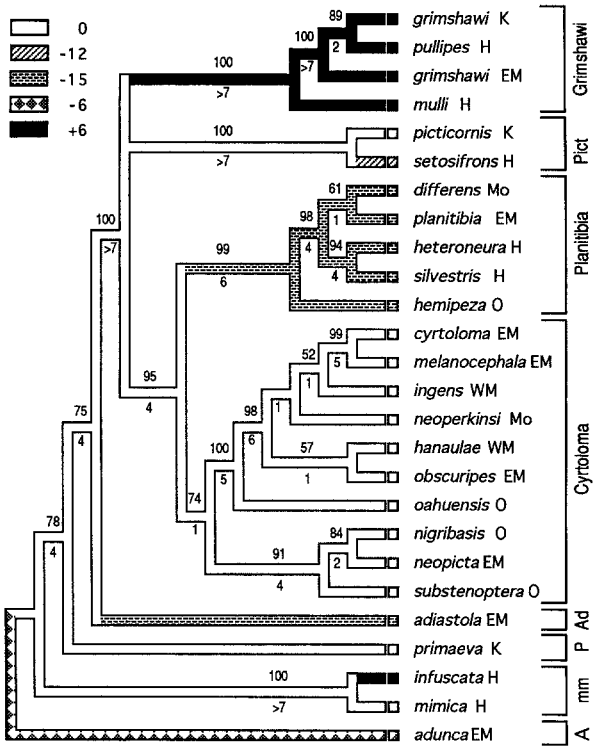

Fig. 3. Mapping the 5′ insertions/deletions in *Yp2* on the *Yp* nucleotide sequence phylogeny of Hawaiian *Drosophila* species via MacClade (Maddison and Maddison 1992). Decay index values (Donoghue et al. 1992) are shown below the branches, and bootstrap values (Felsenstein 1985) from 1,000 replications are shown above the branches, indicating the level of statistical support for the topology determined by that branch. Only bootstrap values above 50 are included. The non–picture-winged species *D. adunca, D. mimica,* and *D. infuscata,* were used as the outgroups. The length of the 5′ *Yp2* sequence in the most primitive picture-winged species *D. primaeva* was assigned as the zero-character state. The sizes in nucleotides of the insertion (+) or deletion (–) events relative to the *D. primaeva* sequence are as shown. The island to which each Hawaiian species is endemic is shown to the right of the species names, abbreviated as follows: *K,* Kauai; *O,* Oahu; *Mo,* Molokai; *WM,* West Maui; *EM,* East Maui; *H,* Hawaii. Species cluster into the formerly recognized species groups and subgroups as shown to the far right (*A* = *antopocerus; mm* = modified mouthparts; *P* = *primaeva; Ad* = *adiastola; Pict* = *picticornis*).

thus comprise phylogenetically distinct lineages. Nonetheless, there are two possible interpretations of the observed evolutionary pattern. Given the fact that other intermediate groups between *D. adiastola* and the *planitibia* subgroup lack the deletion, it is most likely that the *D. adiastola Yp2* sequence originated from an independent deletion event of the same 15-nt fragment that was deleted in the *planitibia* subgroup. Alternatively, it could be hypothesized that this particular 15-nt deletion was widely maintained as a polymorphism in ancestral populations, with the deleted morph becoming fixed in the *adiastola* and *planitibia* subgroup lineages, whereas the standard nondeleted morph was fixed in the remaining lineages. The lack of evidence for longstanding length polymorphisms in this region among the Hawaiian species (see Discussion) makes this scenario improbable, although it cannot be discounted entirely. If, however, the first alternative is correct, two instances of the same event are evident in two independent evolutionary lineages. Regardless, the frequency of length mutations in this region suggests that it is an evolutionary hot spot for nucleotide insertions/deletions.

Discussion

Constraints on Length Variation in Repetitive DNA

Although more common in noncoding regions of the genome (Litt and Luty 1989), microsatellites are also present in coding regions, where they are generally constrained to trinucleotide, hexanucleotide, or higher-multiple motifs (Costa et al. 1991; Caskey et al. 1992; Hey and Kliman 1993). Repetitive elements based on trinucleotide repeats characterize a number of human genetic disease genes (Caskey et al. 1992); these regions are subject to replication errors resulting in interindividual variation and intraindividual instability in the number of repeats. For example, expansion of tandem d(CAG) repeats encoding poly(Gln) can result in Huntington's disease (Huntington's Disease Collaborative Research Group 1993) or Kennedy's disease (La Spada et al. 1991). No abnormalities are, however, associated with the similar but fewer d(CAG) repeats near the 3′ ends of the *D. melanogaster Notch* locus (Wharton et al. 1985) or the *Drosophila Yp2* gene (Fig. 1).

As reported here, the region of the hexanucleotide repeat d(GGC-ATT) in the 5′ coding sequences of the *Yp2* gene represents an evolutionary hot spot for DNA length mutations among Hawaiian *Drosophila* species. This compound repetitive element is embedded in a more extensive SSR-rich region, but neither the upstream (region 1) or downstream (region III) segments show any length variation (Fig. 2) despite an apparent potential for replication slippage provided by the multiple repeats. The lack of DNA length variation in repetitive region I can be attributed to dual functional constraints. First, in *D. melanogaster* this sequence coincides with the terminal part of the 105-nt transcriptional enhancer OE2 located in the first exon, which has been demonstrated to regulate ovarian expression of the *Yp* genes (Logan et al. 1989). Presumably, the homologous sequences in Hawaiian species serve a similar function. Second, this repetitive sequence also encodes the terminal 14 aa of the 19 aa signal peptide (Hung and Wensink 1983; Parisi 1994; Fig. 2) needed to direct secretion of the yolk proteins from their sites of synthesis (female fat body and ovarian follicle cells) prior to uptake by the maturing oocytes (Bownes 1982). These two functions must impose stringent selection on this sequence, limiting nucleotide substitutions and precluding DNA length mutations. We are unaware of any particular function of region III.

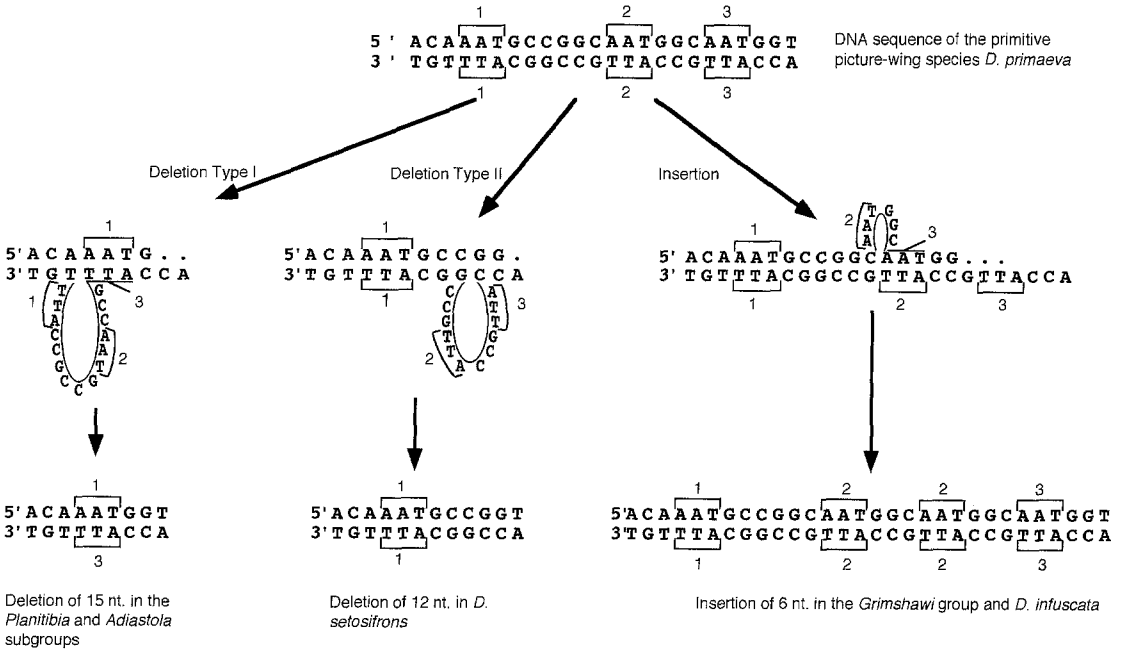

Fig. 4. Proposed slippage events responsible for generating DNA length variation at the 5′ end of the *Yp2* gene of Hawaiian *Drosophila* species. The diagrams show the strand mispairings that most parsimoniously account for the 15-nt (left) and 12-nt (center) deletions (3′ to 5′ template strand looped out), and to the right, a 6-nt insertion (5′ to 3′ nascent strand looped out) relative to the *D. primaeva* sequence of region II. In deletion type I (left) the first copy of d(AAT) in the nascent strand mispairs with the third copy of the complementary sequence in the template. In the insertion event (right), the third copy of d(AAT) in the daughter strand mispairs with the second complementary copy in the template strand, generating a duplication of the second copy.

Restriction of nucleotide length variation to the short intervening region II is intriguing in light of the protein secondary-structure predictions (Hung and Wensink 1983; Parisi 1994) which show that this segment encodes a random coil lying between the alpha-helix of the signal peptide and another alpha-helix toward the C-terminal. Interestingly, the repetitive (Thr-Gly) element in the *Drosophila per* gene product forms a series of turns that subdivide the protein into two main globular domains (Costa et al. 1991). It should be noted that, following cleavage of the signal peptide and secretion, the variable-length region of the *Yp2* gene product would be located right at the amino terminus of the mature YP2 protein. This segment varies in length by up to seven amino acids among the Hawaiian species we have examined, and, given that additional or fewer amino acids can be tolerated, it is tempting to suggest that the length variation and the sequence variation in this segment are effectively neutral. This is further supported by the degenerate nature of the sequence in *D. melanogaster* (Fig. 2) which suggests that the repeats *per se* have no obvious function.

Mechanism of DNA Length Mutation

Given the repetitive nature of the 5′ segment of *Yp2* in the Hawaiian *Drosophila* species, replication slippage (Levinson and Gutman 1987) is a likely mechanism for generating the length variants, as suggested for the chorion (Jones and Kafatos 1982) and other genes. Figure 4 presents a scheme that proposes how replication slippage events could have generated three of the observed DNA length variants most parsimoniously from the putative ancestral sequence in *D. primaeva* and *D. mimica.* The three nontandem d(AAT) repeats present in the primitive sequence appear to have been pivotal to the mispairings between the template strand and the replicating daughter strand that led to the two independent 15-nt deletions in the *planitibia* and *adiastola* species groups (Fig. 4, left) and the hexanucleotide insertions in the *grimshawi* group and in *D. infuscata* (Fig. 4, right). The GC-rich sequence d(CCGG) appears to have been involved in the mispairing that looped out two copies of the hexanucleotide from the template strand and generated the 12-nt deletion in *D. setosifrons* (Fig. 4, center). We need more data from the non–picture-winged species to propose a mechanism for evolution of the sequence observed in *D. adunca,* but conceivably the triplet d(AAC/T) was involved here.

Homoplasy in Nucleotide Length Mutations

It should be pointed out that of the six length mutations identified by our phylogenetic analysis (two insertions and four deletions), only two events are unique. These data demonstrate that insertion/deletion mutations display significant homoplasy, and that, in the absence of other phylogenetic information, shared microsatellite variants cannot be reliably used for inferring evolution-

ary relationships. However, if an independent phylogeny is available, the evolution of DNA length variants can be traced, as we have demonstrated here. What is surprising in our data is the degree to which particular length variants are shared by whole clades (Fig. 3), given the potential for replication slippage in tandemly repeated regions.

Timing and Geography of Nucleotide Length Mutations

Our phylogenetic analysis suggests that insertion/deletion events are comparatively rare in the 5′ repetitive region of *Yp2* of Hawaiian *Drosophila,* and moreover, rather ancient. A unique advantage of our knowledge of the biogeography of the endemic Hawaiian *Drosophila* (Carson and Kaneshiro 1976) and the geological ages of the Hawaiian volcanoes (Carson and Clague 1995) is that we can estimate the geographic location and the time frame for particular nucleotide events. From consideration of the island distributions of the sampled *Drosophila* species in the context of the molecular phylogeny (Fig. 3), it can be concluded that the branching that divided the *planitibia* species group into the two main evolutionary lineages, the *cyrtoloma* and *planitibia* subgroups, took place on Oahu. The interpretation from the chromosomal phylogeny (Carson 1992) that this occurred on Molokai and that *D. hemipeza* resulted from a back migration from Molokai to Oahu is not supported by our molecular phylogeny. Since all ten species of the *cyrtoloma* subgroup share the putative ancestral *Yp2* length variant (and display three YP electrophoretic bands), whereas all five species of the *planitibia* subgroup share a variant (Fig. 3) that is 15 nt shorter (and display two YP bands), the most parsimonious interpretation is that this 15-nt deletion occurred once, in an ancestral population on Oahu. This deletion cannot be older than 3.7 million years (Myr), the age of the oldest volcanic range on Oahu, the Waianae Mountains. Alternatively, if this deletion took place in *planitibia*-group populations in the more easterly Koolau Mountains, the upper limit on the age of this deletion would be 2.6 Myr.

To identify where and when the identical but probably independent 15-nt deletion in the *adiastola* group took place, additional species from this group as well as its immediate ancestors will need to be sequenced. If a shared ancestral polymorphism were responsible for the origin of the same deleted morph in *D. adiastola* and the *planitibia* subgroup, the time and place of origin of the putative 15-nt deletion could not be reliably determined. However, for reasons outlined more fully below, this alternative hypothesis is not favored.

In dating specific *Yp2* length mutations, the assumption is that they are unique events that arose at a discrete time and place and then rapidly became fixed in that local population, ultimately coming to characterize that species and all its descendant species in the ensuing clade. Although our data are limited, no intraspecific length variation has been found within *D. neopicta, D. silvestris,* or *D. grimshawi,* either within or between geographic populations. Further, the homogeneity in sequence length observed among species within the *cyrtoloma, planitibia,* and *grimshawi* subgroups argues against any widespread polymorphism. If any slippage events are occurring currently, the resulting sequences must not be maintained in current-day populations, either as segregating polymorphisms or as newly fixed variants. We conclude that this repetitive region is not a contemporary hot spot, although it might be considered a hot spot for DNA length mutations in an evolutionary time frame.

Why Does the Hawaiian Drosophila Yp2 Gene Lack Length Polymorphism?

Our observations on the 5′ repetitive region in the *Yp2* gene of Hawaiian *Drosophila* are in striking contrast to most previous observations, which indicate that SSRs are routinely associated with DNA length polymorphisms, with the number of repeats varying between chromosomes of heterozygous individuals and among individuals of a species. Polymorphism is the norm in most of the available data on tandem dinucleotide (noncoding) and trinucleotide (coding) repeats in humans (Tautz 1989; Weber and May 1989; La Spada et al. 1991; Caskey et al. 1992), *Mus domesticus* (Reue and Leete 1991), *Drosophila* (Costa et al. 1991; Hey and Kliman 1993), and other organisms. Although our data are limited, they pose a question as to why the Hawaiian *Drosophila Yp2* gene apparently lacks length polymorphism.

Founder effects (Carson and Templeton 1984) following rare interisland migrations have played a significant role in the biology and speciation of the Hawaiian *Drosophila.* Furthermore, because of volcanic effects, local Hawaiian populations suffer frequent extinctions and recolonizations (Carson et al. 1990): There is the potential for loss of polymorphisms at each genetic bottleneck and for the rapid fixation of novel DNA length variants due to random drift. The small size of the founding population (at an extreme, a single fertilized female) would greatly increase the chance of these nucleotide variants being established; they could then come to characterize the new species resulting from the founder event. *Yp2* length variants arising later in the history of the species when population sizes are large would seem to have little chance of reaching fixation, given that they do not appear to confer any particular selective advantage.

Rapid genetic drift predicts low levels of neutral or nearly neutral genetic variation, and nucleotide polymorphism levels can, in fact, be interpreted to be an indicator of effective population size (Hey and Kliman 1993). A low level or an apparent lack of (presumably neutral) *Yp2* nucleotide length polymorphism in Hawaiian *Dro-*

sophila species is consistent with their founder origins and historically small and patchy populations.

Although founder events have been common in the history of individual Hawaiian clades such as the *planitibia* group, the number of discrete fixations of nucleotide insertions/deletions in the *Yp2* gene is comparatively small. Such events must therefore be rare. For example, the Hawaiian species pair *D. silvestris–D. heteroneura* originated from one or perhaps two "*planitibia*-like" founders from the Maui complex (Carson and Kaneshiro 1976), but the length variant common to these two Hawaiian species is the same as that in the three older members of the *planitibia* subgroup. In the chorion genes of these same species (Martinez-Cruzado 1990), however, length mutations have been much more frequent; the two species from the island of Hawaii differ from *D. planitibia* by several length mutations in the coding and noncoding regions, the two Hawaiian species differ from each other, and the eastern Hilo populations of *D. silvestris* differ from the western Kona *silvestris* populations by two length variants. Clearly the chorion and the *Yp2* genes of Hawaiian *Drosophila* display quite different frequencies and patterns of DNA length variation. Thus other factors, besides founder events, must be responsible for the relative constraint on DNA length variation in 5′ coding regions of the *Yp2* gene. It should be noted that length polymorphism has been detected within and between populations in a repetitive segment in the adjacent *Yp1–Yp2* intergenic region (unpublished data).

The infrequent occurrence of length variants in the 5′ region of *Yp2* may be due in part to the fact that this repetitive region is comprised of more complex nontandem rather than simple tandem repeats, which may provide greater opportunity for slippage. Moreover, a strong argument for the lack of length polymorphism is the likelihood of genetic hitchhiking effects (Maynard Smith and Haigh 1974; Kaplan et al. 1989) coincident with selective sweeps involving the immediately upstream region. This tightly linked sequence is clearly under strong selection due to its dual functional contraints; any advantageous mutations occurring in this sequence would be subject to strong directional selection that would rapidly fix the favored allele plus any neutral variants nearby, either substitutions or length variants. Evidence supporting past hitchhiking events can be found by examination of the sequence variation in region 1 (Fig. 2). It can be observed that the occurrence and fixation of the 6-nt insertion in the *grimshawi* clade was associated with fixation of a T → A transversion in the 12th codon of the gene, mutating the d(TGC) encoding cysteine in the signal peptide to the d(AGC) encoding serine. In the *planitibia* subgroup, the deletion of 15 nt was associated with a second position C → G transversion, converting the 19th d(GCC) codon to d(GGC) and replacing alanine by glycine. Remarkably, the same C → G transversion seems to have occurred independently in the *adiastola* species group, and this transversion may have been associated with the identical but phylogenetically independent 15-nt deletion (Figs. 2, 3). These observations provide strong support for the role of genetic hitchhiking in limiting length polymorphism in this particular region of the *Yp2* gene.

Notwithstanding the lack of polymorphism, the length variation documented in the *Yp2* gene of the 26 Hawaiian *Drosophila* species analyzed here substantiates the significant role of replication slippage in generating length mutations in regions of short tandem and nontandem repeats (Fig. 4), adding another example to the growing list of cases of variable SSRs in coding regions. Our phylogenetic analysis also demonstrates that although rare, such nucleotide events are not necessarily unique, and care must be exercised in using microsatellite length variants for phylogenetic inference.

Acknowledgments. We thank Hamp Carson for comments on an early draft of the manuscript and an anonymous reviewer for constructive comments that improved the final version. Thanks also to Jacob Cohen for assistance with the computer alignments. This research was supported by a grant from the National Science Foundation to M.P.K. and E.M.C.

References

Bownes M (1982) Hormonal and genetic regulation of vitellogenesis in *Drosophila.* Q Rev Biol 57:247–274

Carson HL (1992) Inversions in Hawaiian *Drosophila.* In: Krimbas CB, Powell JR (eds) Drosophila inversion polymorphism. CRC Press, Boca Raton, pp 407–439

Carson HL, Clague DA (1995) Geology and biogeography of the Hawaiian Islands. In: Wagner WL, Funk VA (eds) Hawaiian biogeography. Evolution on a hot spot Archipelago. Smithsonian Institution Press, Washington, DC, pp 14–29

Carson HL, Kaneshiro KY (1976) *Drosophila* of Hawaii: systematics and ecological genetics. Annu Rev Ecol Syst 7:311–345

Carson HL, Templeton AR (1984) Genetic revolutions in relation to speciation phenomena: the founding of new populations. Annu Rev Ecol Syst 15:97–131

Carson HL, Lockwood JP, Craddock EM (1990) Extinction and recolonization of local populations on a growing shield volcano. Proc Natl Acad Sci USA 87:7055–7057

Caskey CT, Pizzuti A, Fu Y-H, Fenwick RG, Nelson DL (1992) Triplet repeat mutations in human disease. Science 256:784–789

Costa R, Peixoto AA, Thackeray JR, Dalgleish R, Kyriacou CP (1991) Length polymorphism in the threonine-glycine encoding repeat region of the *period* gene in *Drosophila.* J Mol Evol 32:238–246

Craddock EM, Kambysellis MP (1990) Vitellogenin protein diversity in the Hawaiian *Drosophila.* Biochem Genet 28:415–432

Donoghue MJ, Olmstead RG, Smith JF, Palmer JD (1992) Phylogenetic relationships of Dipsacales based on rbcL sequences. Ann Mo Bot Gardens 79:333–345

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Hey J, Kliman RM (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol Biol Evol 10:804–822

Higgins DG, Bleasby AJ, Fuchs R (1992) CLUSTAL V: improved

software for multiple sequence alignment. Comput Appl Biosci 8:189–191

Hung M-C, Wensink PC (1983) Sequence and structure conservation in yolk proteins and their genes. J Mol Biol 164:481–492

Huntington's Disease Collaborative Research Group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72:971–983

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73

Jones CW, Kafatos FC (1982) Accepted mutations in a gene family: evolutionary diversification of duplicated DNA. J Mol Evol 19:87–103

Kambysellis MP, Ho K-F, Craddock EM, Piano F, Parisi M, Cohen J (1995) Pattern of ecological shifts in the diversification of Hawaiian *Drosophila* inferred from a molecular phylogeny. Curr Biol 5:1129–1139

Kaplan NL, Hudson RR, Langley CH (1989) The "hitchhiking effect" revisited. Genetics 123:887–899

La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352:77–79

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4(3):203–221

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401

Logan SK, Garabedian MJ, Wensink PC (1989) DNA regions that regulate the ovarian transcriptional specificity of *Drosophila* yolk protein genes. Genes Dev 3:1453–1461

Maddison WP, Maddison DR (1992) MacClade. Analysis of phylogeny and character evolution, version 3. Sinauer, Sunderland, MA

Martinez-Cruzado JC (1990) Evolution of the autosomal chorion cluster in *Drosophila*. IV. The Hawaiian *Drosophila*: rapid protein evolution and constancy in the rate of DNA divergence. J Mol Evol 31:402–423

Maynard Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23:23–35

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, White R (1987) Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235:1616–1622

Parisi M (1994) Nucleotide sequence of the *Drosophila grimshawi Yolk protein* genes and analysis of the regulation of the *Yp1–Yp2* gene cluster. PhD thesis, New York University, New York

Reue K, Leete TH (1991) Genetic variation in mouse apolipoprotein A-IV due to insertion and deletion in a region of tandem repeats. J Biol Chem 266:12715–12721

Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H, Arnheim N (1985) Enzymatic amplification of B-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. Science 230:1350–1354

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M (1966) Frameshift mutations and the genetic code. Cold Spring Harb Symp Quant Biol 31:77–84

Swofford DL (1993) Phylogenetic analysis using parsimony (PAUP), version 3.1.1. University of Illinois, Champaign, IL

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res 17:6463–6471

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Wharton KA, Yedvobnick B, Finnerty VG, Artavanis-Tsakonas S (1985) *opa*: a novel family of transcribed repeats shared by the *Notch* locus and other developmentally regulated loci in *D. melanogaster*. Cell 40:55–62