Evidence for Recurrent Paralogous Gene Conversion and Exceptional Allelic Divergence in the *Attacin* Genes of *Drosophila melanogaster*

Brian P. Lazzaro and Andrew G. Clark

Institute of Molecular Evolutionary Genetics, Department of Biology, Penn State University, University Park, Pennsylvania 16802

Manuscript received February 8, 2001
Accepted for publication July 23, 2001

ABSTRACT

Insects produce a limited variety of antibacterial peptides to combat a wide diversity of pathogens. These peptides are often conserved across evolutionarily distant taxa, but little is known about the level and structure of polymorphism within species. We have surveyed naturally occurring genetic variation in the promoter and coding regions of three *Attacin* antibacterial peptide genes from 12 lines of *Drosophila melanogaster*. These genes exhibit high levels of silent nucleotide variations (1–3% per nucleotide heterozygosity), but are not excessively polymorphic at the amino acid level. There is extensive variation in the *Attacin* promoters, some of which may affect transcriptional efficiency, and one line carries a deletion in the *Attacin A* coding region that renders this gene nonfunctional. Two of the genes, *Attacins A* and *B*, are arranged in tandem and show evidence of repeated interlocus gene conversion. *Attacin C*, more divergent and located 1.3 Mbp upstream of *Attacins A* and *B*, does not appear to have been involved in such exchanges. All three genes are characterized by divergent haplotypes, and one *Attacin AB* allele appears to have recently increased rapidly in frequency in the population.

INSECTS fight bacterial infection, in part, through the generation and extracellular circulation of a variety of short, general antibacterial peptides. Although over 400 different innate immune peptides have been identified in eukaryotes (Hoffmann *et al.* 1999), most insects produce fewer than 10 peptide classes. This relatively small number of peptides must effectively combat a wide range of potential and actual pathogens. The conservation of antibacterial peptides, in amino acid sequence and in three-dimensional structure, has been well documented across evolutionarily distant taxa (Bulet *et al.* 1999). However, relatively little work has examined genetic variation within taxa. What studies have been done focus almost exclusively on the *Cecropin* cluster of *Drosophila melanogaster* (Clark and Wang 1997; Date *et al.* 1998; Ramos-Onsins and Aguadé 1998). Here, we present data on the quantity and origin of polymorphism in the *D. melanogaster Attacin* genes.

Attacins represent one of the most taxonomically widespread classes of antibacterial peptides. Families of Attacin-like peptides (usually two to four functional genes per haploid genome) have been identified in the lepidopteran species *Hyalophora cecropia* (Hultmark *et al.* 1983), *Bombyx mori* (Sugiyama *et al.* 1995), *Hyphantria cunea* (Shin *et al.* 1998), *Trichoplusia ni* (Kang *et al.* 1996), and *Heliothis virescens* (Ourth *et al.* 1994), as well as in the dipteran species *Sarcophaga peregrina* (Ando *et al.* 1987) and *D. melanogaster* (Åsling *et al.* 1995; Dushay *et al.* 2000; Hedengren *et al.* 2000). Mature Attacin peptides are typically ~190 amino acids in length (Sarcophaga peptides are longer) and adopt a "random coil" structure in solution (Gunne *et al.* 1990). This loose, flexible structure is devoid of disulfide bonds and does not take a rigid conformational shape. This lack of strict structural constraint may allow relatively free amino acid substitution, explaining the low level of amino acid identity between *Attacin* homologs in distant taxa. There is, however, conservation of general structure and functional activity. Attacins are lethal to Gram-negative bacteria, and *H. cecropia* Attacins have been shown to affect the growth rates of some Gram-positive bacteria (Engström *et al.* 1984). *H. cecropia* Attacins directly bind lipopolysaccharides in the outer membrane of Gram-negative bacteria, disrupting membrane integrity and leading to increased permeability (Engström *et al.* 1984). This binding also leads to the specific inhibition of several bacterial outer membrane proteins, probably through feedback inhibition, which limits the bacterium's ability to restore membrane function (Carlsson *et al.* 1991, 1998). The permeabilization of the bacterial outer membrane by Attacins allows large molecules such as Cecropins and lysozymes to more easily access the inner membrane, setting up a synergistic interaction between these antibacterial peptides (Engström *et al.* 1984). At sufficiently high concentrations, Attacins can cause bacterial cell lysis (Engström *et al.* 1984).

We identified two novel *D. melanogaster Attacins* by performing a BLAST query of *Attacin A* against the genome sequences released by the Berkeley Drosophila

Corresponding author: Brian Lazzaro, Department of Biology, Penn State University, University Park, PA 16802.
E-mail: bplazzaro@psu.edu

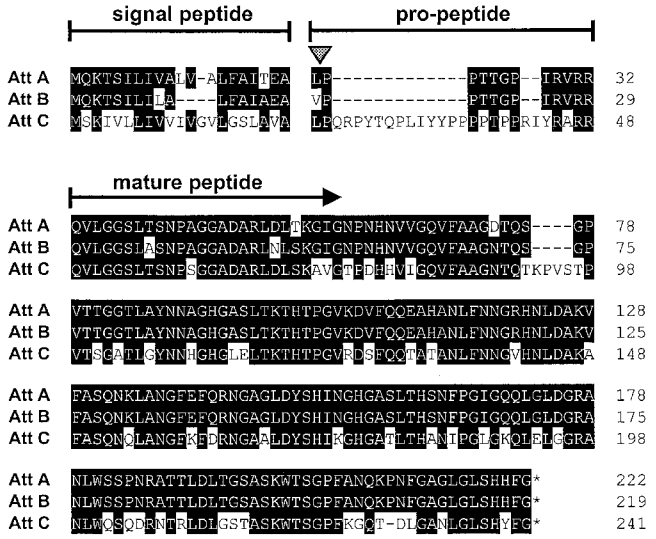


FIGURE 1.—Amino acid alignment of the consensus *Attacin A*, *Attacin B*, and *Attacin C* gene products. Strict amino acid identities are highlighted. The shaded triangle indicates the site of a polymorphic insertion of Pro-Ser-Leu in *D. melanogaster Attacin A*. This polymorphism may predate the species diversification of the *melanogaster* subgroup.

Genome Project (BDGP) and Celera Genomics (Rockville, MD; B. P. LAZZARO and A. G. CLARK, unpublished data; Flybase Report FBrf0126873, http://flybase.bio.inidiana.edu). One of these, the most similar to *Attacin A*, is the same sequence isolated by DUSHAY *et al.* (2000; GenBank accession no. AF220547) and probably represents the cross-hybridizing sequence inferred by ÅSLING *et al.* (1995) in the identification of *Attacin A*. This gene has been named *Attacin B*. The second match that we pursued is the same sequence identified by HEDENGREN *et al.* (2000) as *Attacin C*. A fourth match, termed *Attacin D* by HEDENGREN *et al.* (2000), was deemed too divergent (33% amino acid identity to *Attacin A*) to be considered here.

Attacins A and B are 96 and 97% identical at the nucleotide and amino acid levels, respectively. They are arranged head-to-tail, separated by just under 1.1 kb, at cytological position 51AB on chromosome 2R. The antibacterial peptide gene *Drosocin* lies 1.2 kb upstream of *Attacin A*. *Attacin C* is located at cytological position 50A and shows only 67% nucleotide and 70% amino acid identity to *Attacin A*. Although the mature peptides are similar in length, the propeptide region of *Attacin C* is 27 amino acids long compared to 11 amino acids in *Attacin A*, and the signal peptides are identical at only 6 of 20 positions, including the N-terminal methionine (Figure 1). Therefore, we suggest that *Attacin C* may be differently targeted or processed than *Attacins A* and *B*, although the high degree of similarity among the mature peptides implies a commonality of function.

We have surveyed natural genetic variation in alleles of *Attacins A*, *B*, and *C*, which were recovered from a wild population of North American *D. melanogaster*. The overall level of nucleotide diversity is quite high in each of these three genes, but there is not an excess of amino acid polymorphism. We find that *Attacins A* and *B* have experienced multiple paralogous gene conversion events and that a recent conversion has created a novel haplotype that subsequently increased rapidly in frequency. We also note a number of polymorphisms, including a null allele of *Attacin A*, that may affect the functional capacity of the immune response.

MATERIALS AND METHODS

Line construction: All polymorphism data were determined in 12 lines derived from a natural population of *D. melanogaster* in State College, Pennsylvania. Individual females were collected in the wild in 1998 and allowed to oviposit in vials containing standard cornmeal medium. F_1 or F_2 male progeny were mated to females carrying the second chromosome balancer CyO. Individual male progeny from this cross were backcrossed to the CyO stock. CyO/+ backcross progeny were then sib-mated and the CyO chromosome was eliminated in the following generation. The remaining wild-type progeny were recurrently sib-mated, establishing stocks that are homozygous for a single, naturally occurring second chromosome that has experienced a minimum of recombination and selection in the laboratory. These stocks have been named "2CPA" for "2nd chromosome, Central Pennsylvania."

D. simulans was used as an outgroup for the sequence analyses. Sequence was obtained from a *D. simulans* isofemale line established in 1992 from a natural population in Winters, California.

Sequence analysis: We surveyed nucleotide sequence variation among all 12 2CPA lines in the coding and promoter regions of *Attacins A*, *B*, and *C*. The survey region for the *Attacin A* upstream begins 1215 bp upstream of the *Attacin A* start codon (this is the first nucleotide following the *Drosocin* stop codon) and ends 151 bp upstream of the *Attacin A* start (1064 bp total). The surveyed *Attacin A* coding region begins at the 33rd nucleotide following the start codon and ends 16 bp before the stop codon (682 bp total, including the 64 bp intron). *Attacins A* and *B* are separated by ~1.1 kb, which must include the *Attacin B* promoter. A total of 1091 bp of this region was sequenced, beginning 62 bp after the *Attacin A* stop codon and continuing to the *Attacin B* start codon. All 721 bp of the *Attacin B* gene, including the 64-bp intron, plus 37 bp downstream of the *Attacin B* stop codon, was surveyed as the *Attacin B* coding region (758 bp total). A total of 3104 bp of the *Attacin C* region also was sequenced. Of this 3104 bp, 2024 bp is 5′ of the start codon, 726 bp is coding, 63 bp is intronic, and 291 bp is 3′ of the stop codon. A small window, beginning 84 bp and ending 54 bp upstream of the *Attacin C* start codon, was not surveyed.

DNA was isolated from pools of 100–200 flies using a standard phenol-chloroform extraction followed by ethanol precipitation. Gene-specific PCR was carried out using oligonucleotide primers that were designed from sequences deposited in GenBank by D. Hultmark, the Berkeley Drosophila Genome Project, and Celera Genomics (accession nos. Z46893, AE003813.2, and AE003818.2). Primer sequences corresponding to the sequenced regions are available as supplemental material at http://www.genetics.org/supplemental/ or can be obtained from the authors upon request. All amplifications were carried out with at least one primer annealing to noncoding DNA to prevent accidental nonspecific amplification of gene paralogs. Ampli-

fication products were purified by ethanol precipitation and directly sequenced on either an ABI 373 or a Beckman CEQ 2000 automated sequencer, using slight modifications of the manufacturers' protocols. The entire survey region was sequenced on both strands in each line, except in rare instances where sequencing reactions failed to give reliable sequence reads long enough to reach the next primer on the strand. In these instances, sequence was inferred from the complementary strand. All sequences have been deposited in GenBank under accession nos. AY056843–AY056902.

RESULTS

Polymorphism in and gene conversion between the *Attacin A* and *B* genes: For the purpose of analyzing the *Attacin A* and *Attacin B* polymorphism data, we have broken the *Attacin AB* array into four segments. These will be referred to as the *Attacin A* upstream region (1064 bp), the *Attacin A* coding sequence (682 bp), the *Attacin AB* intergenic spacer (1091 bp), and the *Attacin B* coding sequence (758 bp). Although these units will be referred to as discrete segments, it is important to keep in mind that they are adjacent in the genome.

The level of nucleotide variability is strikingly high in the *Attacin A* coding region (Figure 2B). Without distinguishing between silent and replacement substitutions, we obtained per-base estimates (± 1 SD, assuming no recombination) of $\hat{\theta} = 0.0315$ ($\pm 1.29 \times 10^{-2}$) from the number of segregating sites in the sample and $\hat{\pi} = 0.0243$ ($\pm 7.10 \times 10^{-3}$) from the average pairwise distance between alleles in the sample (Table 1). We uncovered 13 amino acid replacement polymorphisms within the *D. melanogaster Attacin A* coding region, 7 of which are concentrated in exon 2 of line 2CPA 129 (Figure 2B).

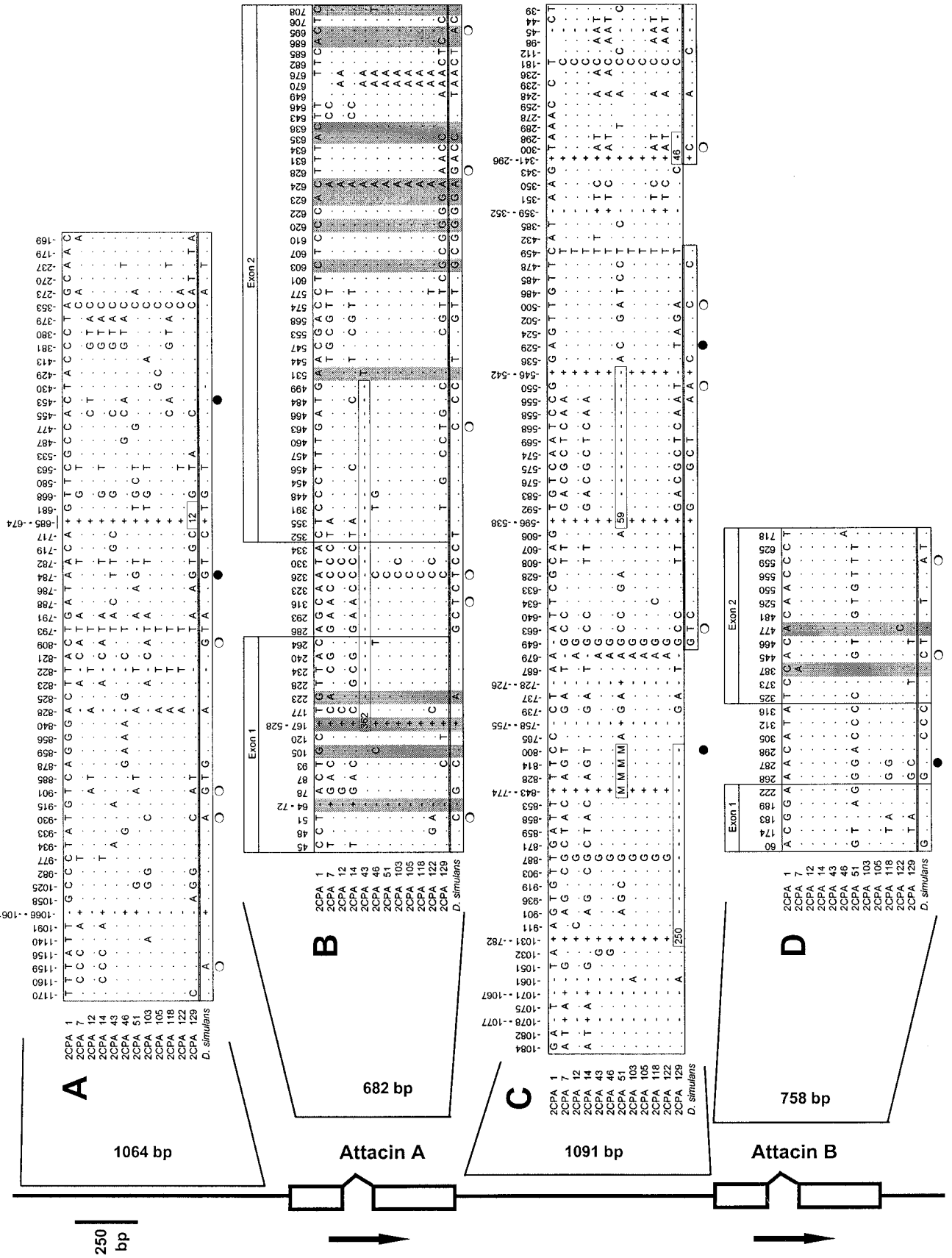
We also detected two insertion/deletion polymorphisms in the *Attacin A* coding sequence. The first of these is a 9-bp insertion (relative to *D. simulans*) in the *Attacin A* pro-peptide sequence. This polymorphism results in the insertion of three amino acids (Pro-Ser-Leu) after position 21 (Figure 1) in lines 2CPA 7 and 2CPA 14. *Attacin A* sequences obtained from the *D. melanogaster* sibling species *D. simulans*, *D. sechellia*, and *D. mauritiana* suggest that this polymorphism may be as old as the origin of those species. We found that *D. simulans* and *D. sechellia* carry the deletion state for this *D. melanogaster* polymorphism, while *D. mauritiana* has the insertion (B. P. Lazzaro and A. G. Clark, unpublished results). We did not, however, survey a sufficient number of alleles from the sibling species to determine whether or not this indel has been maintained as a polymorphism within each species.

The second *D. melanogaster Attacin A* indel is a 362-bp deletion in line 2CPA 43. This deletion eliminates about half of the coding region and the entire intron. It is out of frame, causing a premature stop codon 25 amino acids into the mature peptide (the wild-type peptide is 190 amino acids long), and presumably results in a nonfunctional protein product. This deletion is unique among the sequenced lines and was not detected among an additional 85 central Pennsylvania chromosomes screened by PCR.

Although the level of polymorphism in the *Attacin A* gene is quite high, the substitutions are not distributed evenly among the sampled alleles. Twenty-six polymorphisms are unique to one line, 2CPA 129 (Figure 2B). Fu and Li (1993) showed that, under neutrality, the expected number of singletons in a sample is equal to the per-base estimate of θ multiplied by the length in base pairs of the region surveyed. Our estimate of θ for the entire *Attacin A* coding region is 21.5. Thus, the 26 unique polymorphisms contributed to the sample by line 2CPA 129 alone exceeds the total number of singletons expected in the sample under neutrality, and all but one of these 2CPA 129-specific mutations are found in the second exon. Lines 2CPA 7 and 2CPA 14 contribute nearly all of the remaining segregating sites. If these three lines are excluded, $\hat{\theta}$ drops from 0.0315 ($\pm 1.29 \times 10^{-2}$) for the entire data set to 0.0091 ($\pm 4.29 \times 10^{-3}$) for the more homogenous set of alleles. We applied the haplotype test of Hudson *et al.* (1994) to estimate the probability that the extreme haplotype divergence observed in the *Attacin A* region was generated by a neutral process. This test specifically measures the probability, under neutrality, that a subset of alleles in a sample is monomorphic for a high proportion of the sites that segregate in the complete data set. We simulated sets of 10,000 neutral genealogies conditioning on the empirical sample size of 11 (the null allele was excluded from this analysis) and the observed number of polymorphisms (64). Three out of 10,000 genealogies simulated under the conservative assumption of no recombination had subsamples of seven alleles with 15 or fewer sites segregating, as is observed in the empirical data ($P = 0.0003$).

Although we found the *Attacin B* coding region to be less variable than the *Attacin A* coding region, the same pattern of haplotype structure exists there (Figure 2D). In the *Attacin B* gene, line 2CPA 51 contributes most of the polymorphisms, carrying 15 unique nucleotide substitutions, although only 24 sites segregate in the entire sample. Only 4 out of 10,000 simulated samples of 12 alleles with a total of 24 segregating sites had a clade of 9 alleles segregating for 9 or fewer sites ($P = 0.0004$).

The observation of such divergent haplotypes in the *Attacin A* and *B* genes begs the question as to whether the polymorphisms involved are ancestral or derived. That is, do the outlier haplotypes represent lineages that have experienced multiple new mutations or are these comparatively older alleles that have been maintained in the population? Comparison to the *D. simulans* sequence at each locus reveals that many of the polymorphisms are in their ancestral states in the outlier haplotypes, so that the more common state at these positions is derived. For instance, 15 of the 26 polymorphisms


A
B
C
D
Exon 1
Exon 2
1064 bp
682 bp
1091 bp
758 bp
Attacin A
Attacin B
250 bp
D. simulans

unique to line 2CPA 129 in the *D. melanogaster Attacin A* sample are identical in state to the *D. simulans* sequence, including 5 of the 7 positions that cause amino acid replacements in the 2CPA 129 allele (Figure 2B). Likewise in *Attacin B*, line 2CPA 51 is identical in state to *D. simulans* at 8 of the 19 positions at which 2CPA 51 carries a rare or unique substitution (Figure 2D). Such a large number of high frequency, derived polymorphisms are unexpected under a neutral process (Fay and Wu 2000), and the excess in *Attacins A* and *B* is statistically significant ($H = -31.254$, $P = 0.0054$ in *Attacin A*; $H = -9.848$, $P = 0.0188$ in *Attacin B*; critical values determined by simulation) (Fu 1997; Fay and Wu 2000).

Alignment of all of the *Attacin A* sequences with all of the *Attacin B* sequences shows that many of the high frequency, derived sites in *Attacin A* are identical in state to the common *Attacin B* sequence (Figure 3). We also note that the *Attacin A* outlier haplotypes 2CPA 7 and 2CPA 14 share state at several positions with the *Attacin B* outlier haplotype 2CPA 51 (Figure 3). The sharing of polymorphisms between *Attacins A* and *B* suggests that paralogous gene conversion may be acting to exchange sequence between the two genes. Gene conversion can be identified from aligned sequences by the clustering of sites that share different phylogenetic partitions (Stephens 1985). In applying this method to the *Attacin AB* data, we find that a set of 12 segregating sites involved in the partition of lines 2CPA 7, 2CPA 14 (in *Attacin A*), and 2CPA 51 (in *Attacin B*) are significantly clustered. The probability of a span of sites sharing this partition by chance is extremely unlikely [$P(d < d_0) = 0.00096$; Stephens 1985]. Similarly, the method of Sawyer (1989) yields a probability of 0.006 that a larger sum of squared lengths of shared sequence runs could be obtained by chance without recombination or gene conversion.

These observations lead us to a model where recurrent paralogous gene conversion shapes the pattern of polymorphism in the *Attacin A* and *B* genes. Specifically, in a recent event, the 3′ end of the *Attacin B* gene has converted the 3′ end of the *Attacin A* gene and the conversion product has reached high frequency in the sample. In an independent conversion event, sequence was exchanged between the *Attacin A* and *B* genes, creating the partial haplotype that lines 2CPA 7 and 2CPA 14 in *Attacin A* share with line 2CPA 51 in *Attacin B*. It is not clear whether this sequence originated in *Attacin A* or *Attacin B*. It is also noteworthy that the *D. simulans Attacin A* and *Attacin B* genes share several substitutions relative to the *D. melanogaster* sequences at both genes (Figure 3), which could result from intergenic exchange in either or both species. Such recurrent paralogous gene conversions tend to homogenize the genes within a species and obscure the phylogenetic relationships of homologous genes across species (Figure 4).

The degree of nucleotide homology between the *Attacin A* and *B* genes drops off sharply outside the peptide coding region, making it unlikely that paralogous gene conversion events extend beyond the boundaries of coding sequence. Nevertheless, the four alleles that have apparently participated in paralogous exchange events also display divergent haplotypes in the *Attacin AB* intergenic spacer (which must include the *Attacin B* promoter; Figure 2C). Lines 2CPA 7, 2CPA 14, and 2CPA 129 are distinguished from the remaining alleles by 27 nucleotide substitutions. The intergenic spacer in line 2CPA 129 additionally carries a series of unique deletions, the largest of which is 250 bp and which sums to 331 bp. Line 2CPA 51 contributes the vast majority of the remaining polymorphisms segregating in the *Attacin AB* intergenic spacer, including 15 unique nucleotide substitutions and a series of unique deletions that range in size from 9 to 58 bp and total 95 bp. Line 2CPA 51 also carries a large, complex mutation. This mutation reduces a 69-bp window (843–774 bp upstream of the *Attacin B* start codon) to 31 bp, although those 31 bp are completely unalignable with the remaining sequences. The mechanism for generating such a mutation is not clear, but we consider this as a single evolutionary event in our analyses. Notably, there are several deletions and polymorphic sites among lines 2CPA 7, 2CPA 14, 2CPA 129, and 2CPA 51 that overlap each other, creating

Figure 2.—Aligned polymorphic sites segregating in the *Attacin AB* gene region of 12 lines of *D. melanogaster* and 1 line of *D. simulans*. Sites are numbered relative to the start codon of the nearest downstream gene. Minus (−) and plus (+) signs indicate deletions and multiple base insertions. Deletions that span sites that are polymorphic among other lines are boxed and are labeled with the length of the deletion. Amino acid replacement polymorphisms are shaded. Solid circles (●) indicate sites that are segregating for more than two nucleotides within *D. melanogaster*. Open circles (○) indicate polymorphisms in *D. melanogaster* at which a third nucleotide is found in *D. simulans*. (A) Polymorphisms segregating in 1064 bp surveyed in the *Attacin A* promoter region. Forty-nine fixed differences between *D. simulans* and *D. melanogaster* are not shown. (B) Polymorphisms segregating in 682 bp surveyed in the *Attacin A* coding region, including both exons (boxed) and the 64-base intron (open). Twenty-six fixed differences are not shown. (C) Polymorphisms segregating in 1091 bp surveyed between the *Attacin A* and *Attacin B* genes, including the *Attacin B* promoter. A complex mutation in line 2CPA 51 is indicated by M. This polymorphism reduces a 69-bp window to 31 bp that are unalignable with the original 69-bp sequence. Positions at which no *D. simulans* sequence is shown reflect regions in which the interspecific alignment was not reliable enough to infer the ancestral state of a *D. melanogaster* polymorphism. Twenty-nine fixed differences among the 616 alignable bases are not shown. (D) Polymorphisms segregating in 758 bp of the *Attacin B* coding region, including both exons (boxed) and the 64-base intron. There were no polymorphisms in 37 bp surveyed downstream of the *Attacin B* stop codon. Twenty-one fixed differences between *D. simulans* and *D. melanogaster* are not shown.

TABLE 1

Summary statistics describing polymorphism in *D. melanogaster Attacin* genes

Region	Length	n	S	$\hat{\theta}$ (±1 SD)	$\hat{\pi}$ (±1 SD)	Tajima's D	$4Nc$	$\hat{\pi}_s$	$\hat{\pi}_n$	MK G-test
Attacin A promoter	1043	12	51	0.0162 (±0.007)	0.0148 (±0.002)	−0.509	0.0486			
Attacin A cds[a]	682	11	63	0.0315 (±0.013)	0.0243 (±0.007)	−1.088	0.0012	0.0890	0.0080	0.026 ($P = 0.87$)
Attacin B promoter	700	12	42	0.0199 (±0.008)	0.0178 (±0.003)	−0.577	0.0076			
Attacin B promoter[b]	965	11	77	0.0272 (±0.011)	0.0218 (±0.007)	−1.047	0.0003			
Attacin B promoter[c]	1064	10	47	0.0156 (±0.006)	0.0161 (±0.004)	0.172	0.0007			
Attacin B cds	758	12	24	0.0105 (±0.004)	0.0066 (±0.003)	−1.748*	0	0.0197	0.0007	0.043 ($P = 0.84$)
Attacin C promoter	1963	12	64	0.0108 (±0.004)	0.0117 (±0.001)	0.404	0.0126			
Attacin C cds	1134	12	36	0.0105 (±0.004)	0.0123 (±0.002)	0.785	0.0139	0.0385	0.0022	0.012 ($P = 0.91$)

Length is measured in base pairs after elimination of sites with alignment gaps, n is the number of lines, and S is the number of polymorphic nucleotides. Standard deviations of $\hat{\theta}$ and $\hat{\pi}$ are calculated assuming no recombination. Tajima's D (Tajima 1989) is the normalized difference between $\hat{\theta}$ and $\hat{\pi}$, and $4Nc$ represents the per-base estimate of Hudson's recombination parameter (Hudson 1987). $\hat{\pi}_s$ and $\hat{\pi}_n$ are estimates of $\hat{\pi}$ at silent and nonsynonymous positions, respectively. The MK G-test is described in McDonald and Kreitman (1991). * $P < 0.05$.

[a] The null allele of *Attacin A* (2CPA 43) was not used in the calculation of statistics in the *Attacin A* coding region.

[b] Line 2CPA 129 was excluded from this calculation of statistics summarizing variation in the *Attacin B* promoter.

[c] Lines 2CPA 129 and 2CPA 51 were excluded from this calculation of statistics summarizing variation in the *Attacin B* promoter.

many positions in the sequence at which more than two nucleotide states exist. There are several additional positions in the *Attacin AB* data set where three nucleotide states are segregating at a position in *D. melanogaster* or where *D. simulans* displays a third state at a position that is polymorphic in *D. melanogaster* (Figure 2). Through gene conversion and multiple hits, the *Attacin AB* empirical data set shows departure from the theoretical model of independent mutations and infinite sites in which mutations may occur.

Interestingly, the strong haplotype structure observed in the *Attacin A* and *B* coding regions and in the intergenic spacer is apparently absent upstream of the *Attacin A* gene, although variability is not greatly diminished. We found 52 nucleotide polymorphisms in the *Attacin A* upstream region (Figure 2A), yielding estimates of $\hat{\theta} = 0.0162$ ($\pm 6.53 \times 10^{-3}$) and $\hat{\pi} = 0.0149$ ($\pm 1.78 \times 10^{-3}$). We also detected five small insertion/deletion polymorphisms, ranging in size from 2 to 12 bp.

Considering the high level of variability and apparent lack of haplotype structure in the *Attacin A* upstream region, the alleles that have the common haplotype in both the *Attacin A* and *Attacin B* coding regions show a curious deficiency of polymorphism in both coding regions and in the intergenic spacer. The paucity of variation could possibly be explained if the common *Attacin AB* allele has only recently expanded to high frequency, such that there has not been time for it to accumulate mutations or to recombine appreciably. Rapid expansion models predict that most polymorphisms within the expanding clade should be rare, a prediction that is borne out in the empirical data. Tajima's D (Tajima 1989), a measure of skew in the site frequency spectrum, is negative for the region beginning with the *Attacin A* coding region and continuing through the *Attacin B* coding region ($D = -0.613$) when alleles 2CPA 7, 2CPA 14, 2CPA 129, and 2CPA 51 (and also the *Attacin A* null allele, 2CPA 43) are excluded. The negative value of D indicates an excess of rare polymorphisms, although the skew is not statistically significant ($P = 0.312$; determined by simulation). If the *Attacin A* upstream region is included in the calculation, D for the putatively expanding clade is -0.746 ($P = 0.271$). Fay and Wu (2000) show that the rapid rise in frequency of an allele results in an excess of high frequency, derived polymorphisms within the expanding clade. Fay and Wu's statistic, H, assumes that mutations are independent (Fay and Wu 2000), an assumption that is violated by the gene conversions in the coding regions. However, the noncoding flanking regions are free from this violation. Unfortunately, in some sequence windows (including that containing the complex mutation in 2CPA 51), the *D. melanogaster Attacin AB* intergenic spacer is unalignable with the *D. simulans* sequence, making it impossible to infer the ancestral state of the *D. melanogaster* polymorphisms. Even so, in both flanking noncoding regions there is a nearly

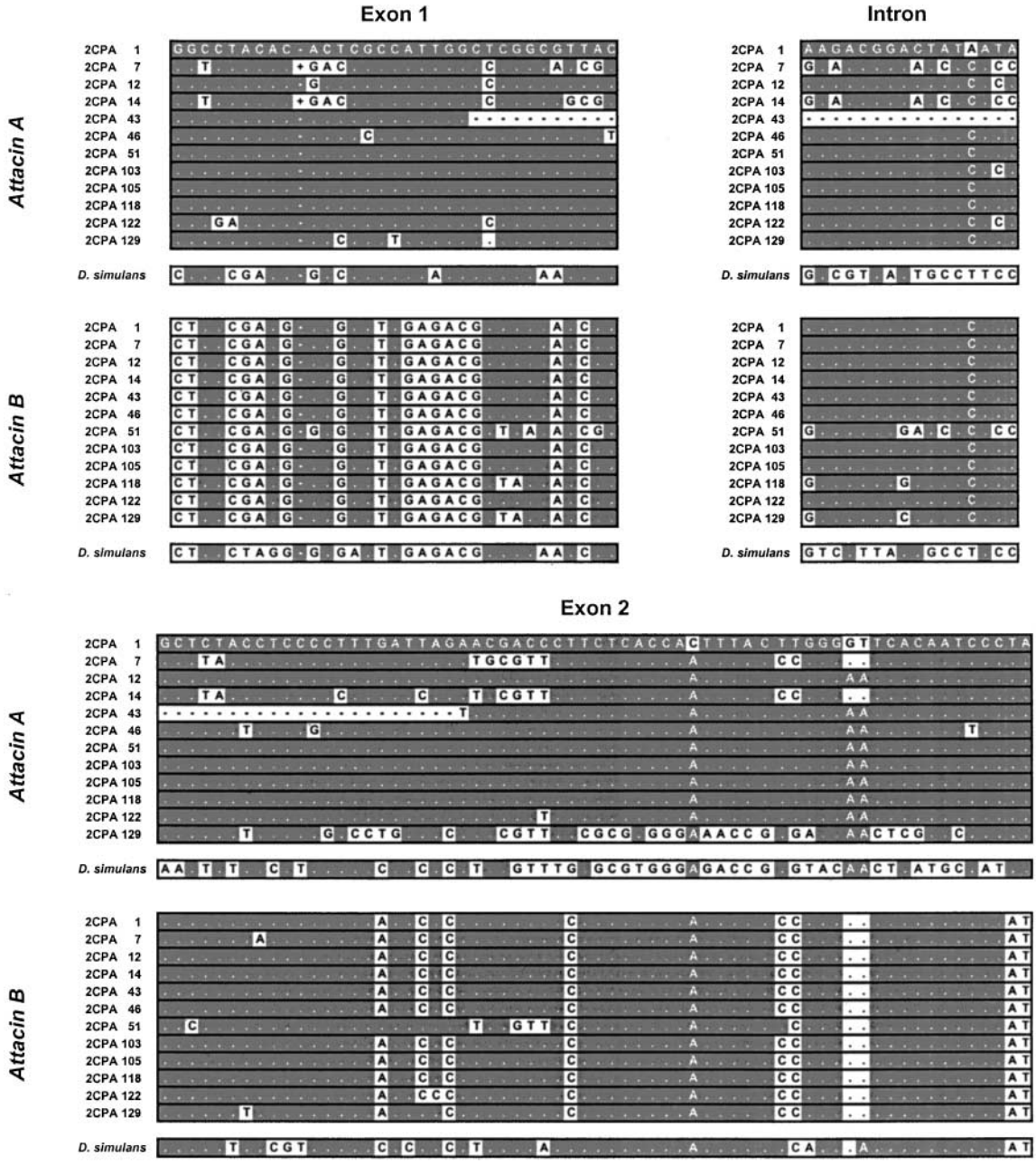


FIGURE 3.—Alignments of the variable sites within and between *D. melanogaster* and *D. simulans Attacins A* and *B*. The most common nucleotide state in the *Attacin A* gene at each position is shaded. A paralogous gene conversion event that subsequently rose in frequency has homogenized exon 2 of *Attacins A* and *B,* but has not disrupted sequence divergence in exon 1. The large number of sites line 2CPA 129 shares with *D. simulans* in exon 2 of *Attacin A* suggests that this allele was not affected by gene conversion and probably is ancestral to the remaining *D. melanogaster Attacin A* sequences. The number of positions at which lines 2CPA 7 and 2CPA 14 in *Attacin A* share state with line 2CPA 51 in *Attacin B* suggests that these alleles reflect an independent exchange event between *Attacins A* and *B*.

significant excess of derived polymorphisms at high frequency within the expanding clade ($H = -9.500$, $P = 0.0748$ in the *Attacin A* upstream; $H = -3.142$, $P = 0.080$ in the intergenic spacer; critical values determined by simulation) when only sites where confident inference of the ancestral state is possible are considered. When all alleles are considered, there is a much smaller excess of high frequency, derived polymorphisms ($H = -0.667$, $P = 0.298$ in the *Attacin A* upstream; $H = -4.364$, $P = 0.132$ in the intergenic spacer, complete deletion analysis). The power of this calculation is hampered in the intergenic spacer, the region that would be expected to show the strongest effect, by the small number of sites for which ancestry can be inferred. Still, the D and H values reflect a strong trend supporting the hypothesis that an *Attacin AB* allele created by para-

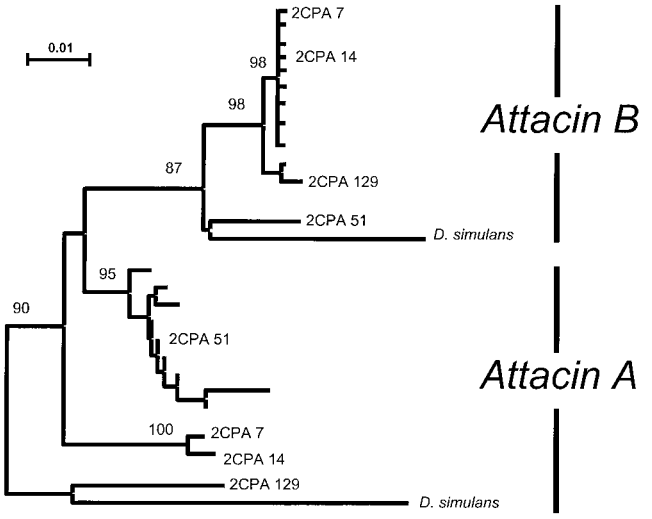


FIGURE 4.—Neighbor-joining tree (SAITOU and NEI 1987) of the aligned *Attacin A* and *Attacin B* coding sequences from *D. melanogaster* and *D. simulans*. Indicated nodes have >85% bootstrap support, as determined using 1000 bootstrap replicates in the MEGA2 software (http://www.megasoftware.net). Recurrent gene conversion events obscure the phylogenetic ancestry of alleles of *Attacins A* and *B*, causing most *D. melanogaster Attacin A* alleles to cluster significantly with *Attacin B* alleles, instead of with *D. simulans Attacin A*. Similarly, line 2CPA 51 in *Attacin B* falls outside the main *Attacin B* clade by virtue of shared polymorphisms with *Attacin A* alleles 2CPA 7 and 2CPA 14.

logous gene conversion has recently increased rapidly in frequency.

Polymorphism in the *Attacin C* gene: The *Attacin C* promoter and coding sequences are less variable at the nucleotide level than are *Attacins A* and *B*, with $\hat{\theta} = 0.011$ ($\pm 4.20 \times 10^{-3}$) and $\hat{\pi} = 0.012$ ($\pm 1.07 \times 10^{-3}$). This relative reduction in variability may partially be due to the fact that *Attacin C* is far enough from *Attacins A* and *B* on the chromosome (and sufficiently divergent at the nucleotide level) to escape any paralogous exchanges. There are five sites in the *Attacin C* survey region at which the *D. simulans* sequence displays a third nucleotide state in a position that is polymorphic in *D. melanogaster*, again showing multiple mutational hits at a position and a departure from the infinite sites model.

Interestingly, the *Attacin C* sequences we obtained were fixed for an 8-bp coding region insertion relative to the BDGP genome sequence. This insertion retains the open reading frame, whereas the BDGP allele should terminate in a premature stop, supporting the assertion of HEDENGREN *et al.* (2000) that the *y*; *cn bw*; *sp* stock (iso-1) used by the genome project probably carries a null allele of the *Attacin C* gene.

As in the *Attacin AB* region, we see marked haplotype dimorphism among the *Attacin C* alleles. In the *Attacin C* case, however, the two common allelic classes are at intermediate frequency and line 102G is an apparent recombinant between them (Figure 5). A sliding window of the level of nucleotide diversity along the *Attacin C* gene region reveals a region of ~200 bp beginning 400 bp upstream of the translational start codon where nucleotide heterozygosity is increased ~10-fold (Figure 6A). The polymorphic sites within this region are in strong linkage disequilibrium with one another and with several sites flanking the window of elevated diversity (Figure 6B). Comparison to the *D. simulans Attacin C* sequence reveals that neither of the two primary haplotypes generated by this linkage disequilibrium is obviously more ancestral than the other, but instead that both are composed of a combination of apparent ancestral and derived sequence states (Figure 5). A sliding window analysis of the divergence between the *D. simulans* and *D. melanogaster Attacin C* sequences shows an ~4-fold increase in divergence overlying the peak of *D. melanogaster* variability (Figure 6). This spike in divergence is preceded by a 31-bp insertion in the *D. simulans* sequence (the insertion does not compromise the ability to align the *D. melanogaster* and *D. simulans* sequences). The window of high polymorphism and divergence is relatively AT-rich (~20% GC compared to ~43% GC over the entire *Attacin C* region) and is mildly repetitive, but does not have any characterized function.

DISCUSSION

The predominant pattern that emerges from the Drosophila *Attacin* sequence data is the presence of highly divergent haplotypes in alleles sampled from a single population. The most divergent haplotypes in the *Attacin A* and *Attacin B* coding sequences have apparently been generated by paralogous gene conversion events that have introduced tracts of segregating sites into the converted locus. Most dramatically, the 3′ end of *Attacin B* has converted the 3′ end of *Attacin A*, introducing ~30 segregating sites and seven amino acid replacements into the second exon of *Attacin A*. The converted allele is found at a frequency of 0.92 in our sequence sample, effectively displacing the unconverted *D. melanogaster Attacin A* allele in the sampled population (and establishing itself as the *Attacin A* reference sequence).

The *Attacin* genes reside in a region of the Drosophila second chromosome that experiences high levels of meiotic recombination (KLIMAN and HEY 1993; CARVALHO and CLARK 1999). Given the relative sequence homogeneity and the absence of apparent historical recombination between conversion types, the rise in frequency of the converted allele must have been rapid. Precise estimates of the age of the conversion event are difficult to make, as tests of allele age and the rate of clade expansion rely on assumptions of independent mutation and infinite sites. Both of these assumptions are clearly violated in our *Attacin AB* data. Some resolution may be provided by examining associations between sites in the *Attacin A* and *B* coding regions and sites in the flanking noncoding regions, which are unlikely to

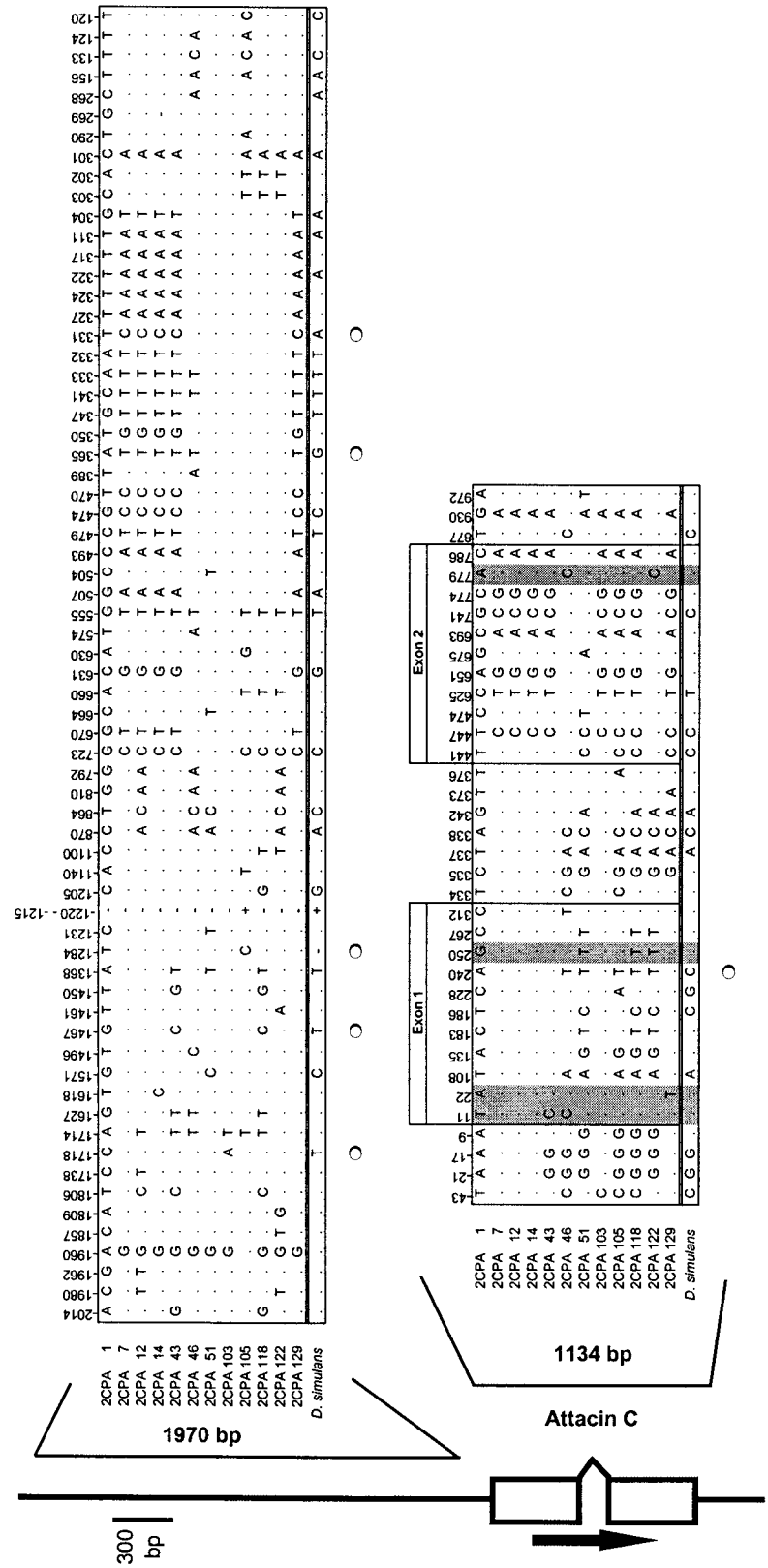


FIGURE 5.—Table of polymorphic sites in the *Attacin C* promoter and coding regions. A total of 3097 bp of the *Attacin C* region was surveyed in 12 lines of *D. melanogaster*. The bottommost sequence in each alignment reflects the *D. simulans* sequence state at positions that are polymorphic in *D. melanogaster*. Totals of 149 sites in the promoter region, 36 sites in the coding region, and 23 sites that contain fixed differences between the two species are not illustrated. Positions are numbered relative to the *Attacin C* start codon. Minus (−) and plus (+) signs indicate deletions and multiple base insertions. Amino acid replacement polymorphisms are shaded. Open circles (○) indicate sites that are polymorphic within *D. melanogaster* at which a third nucleotide is found in *D. simulans*.

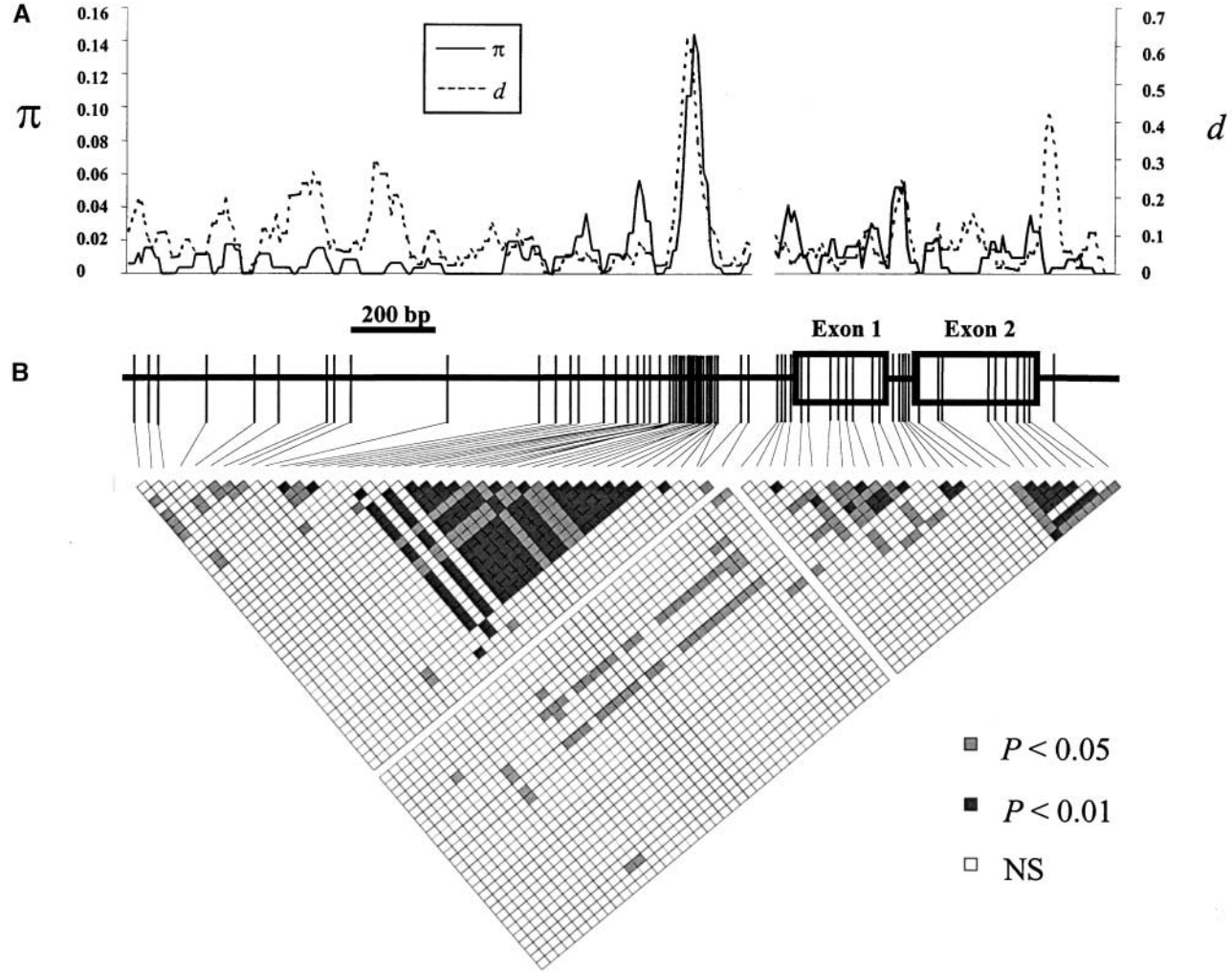


FIGURE 6.—(A) A sliding window analysis of *Attacin C* shows that nucleotide heterozygosity within *D. melanogaster* (π) and divergence of *D. melanogaster* from *D. simulans* (*d*) both peak in a short window ~200 bp upstream of the translational start codon. The *x*-axis is labeled with the midpoint of each window, relative to the *Attacin C* start codon. The short gap in the plot represents a 30-bp region that was not surveyed for sequence variation. The second, smaller peak in divergence occurs downstream of the *Attacin C* stop codon in the putative 3′ UTR. The sliding window analysis was performed with the DnaSP 3.14 software package (ROZAS and ROZAS 1999) using a 50-bp window taking 10-bp steps. (B) Linkage disequilibrium between sites present in two or more alleles was measured using Fisher's exact test, uncorrected for multiple comparisons. There is a general lack of significant disequilibrium across the *Attacin C* region, except among closely linked sites in the coding region and among the sites that underlie the spike in polymorphism and divergence, where linkage disequilibrium is strong.

be subject to the same intergenic conversion events that affect the coding regions. Forces, such as natural selection, that cause rapid changes in allele frequency in the coding sequences should cause perturbations in the site frequencies and associations in the noncoding regions. The four lines that have aberrant haplotypes in the *Attacin A* or *B* coding regions, lines 2CPA 7, 2CPA 14, 2CPA 129, and 2CPA 51, also have divergent haplotypes in the intergenic spacer. The continued association of the outlier coding region haplotypes with the outlier haplotypes in the intergenic spacer, despite a high rate of meiotic recombination, could indicate that the conversion events only recently (and fortuitously) occurred on chromosomes that were already divergent in the spacer. However, the presence of 10 fixed or nearly fixed differences between exon 2 of *Attacin A* and *B* (Figure 3) and the imperfect haplotype sharing of *Attacin A* alleles 2CPA 7 and 2CPA 14 with *Attacin B* allele 2CPA 51 argue against the young conversion hypothesis. Rather, the relative homogeneity of the intergenic spacer in the other eight lines, the lines that are also homogeneous in the *Attacin A* and *B* coding regions, probably reflects the fact that this *Attacin AB* allele has increased in frequency sufficiently recently that there has not been time for it to recombine or accumulate novel mutations. It is not clear why the *Attacin A* 5′ region does not retain the haplotype structure found in the remainder of the locus. It is possible that the

Drosocin gene, immediately upstream of the *Attacin A* 5′ survey region, affects the pattern of polymorphism in the *Attacin A* 5′ region.

One alternative mechanism for generating the deep genealogical branches we observe in our data could be provided if chromosomal inversions segregated among the lines in our sample. Inverted chromosomes are known to segregate in natural *D. melanogaster* populations (Ashburner 1989). These rearrangements disrupt the gene order along the chromosome and prevent homologous recombination during meiosis. If the *Attacin* genes were locked up in such an inversion, mutations could accumulate independently and without recombination between the two inversion types. For instance, if line 2CPA 129 was inversion type "A" and the remainder of the lines were inversion type "B," mutations could accumulate and fix in inversion type B without ever influencing the sequence of line 2CPA 129. In this way, line 2CPA 129 could continue to maintain the ancestral sequence (as inferred from *D. simulans*) while the other lines (inversion type B) could diverge markedly. However, the inversion hypothesis cannot explain how lines 2CPA 7 and 2CPA 14 at *Attacin A* share site state at so many positions with line 2CPA 51 at *Attacin B*. The only inversion on the right arm of the second chromosome that segregates in natural populations with any appreciable frequency is In(2R)NS, which spans cytological positions 52A to 56F (Lindsley and Grell 1967). This inversion is near, but does not include, the *Attacin* genes. Furthermore, the fact that the haplotype structure observed in the *Attacin AB* region is not maintained in the nearby *Attacin C* region argues for intergenic recombination, which is inconsistent with the inversion hypothesis. The lack of strong associations between polymorphic sites in the *Attacin A* promoter region also argues for free recombination between alleles in the sample and against the inversion hypothesis.

A second alternative explanation for the divergent haplotypes in the second exon of *Attacin A* involves the introgression of a segment of the *D. simulans Attacin A* gene into its *D. melanogaster* homolog. Under this scenario, all of the sampled alleles except 2CPA 129 represent the "true" *D. melanogaster Attacin A* sequence, which is highly divergent from the *D. simulans Attacin A* gene sequence. Interspecific hybridization between *D. melanogaster* and *D. simulans* would have introduced the *D. simulans Attacin A* sequence into the *D. melanogaster* genome, where it has since attained polymorphic frequency in the *D. melanogaster* population. Such an introgression event has been proposed in the history of the Drosophila *Cecropin* gene family (Date *et al.* 1998). Like the inversion hypothesis, however, the introgression hypothesis fails to explain the allele shared between *Attacins A* and *B*. The lack of a clear mechanism for hybridization between these two species (hybrid females are sterile, and hybrid males are inviable; Sturtevant 1920) makes the introgression hypothesis unattractive. Since intergenic exchange is known to occur among tandemly repeated genes (Leigh Brown and Ish-Horowicz 1981; Mellor *et al.* 1983), we favor paralogous gene conversion as the logical mechanism for creating the pattern we observe in the *Attacin A* and *Attacin B* genes.

Gene conversion, however, cannot explain the haplotype dimorphism we observe in *Attacin C*. For most of the 2 kb that we surveyed upstream of the *Attacin C* start codon, the level of nucleotide heterozygosity is low and there is little association among sites. Then, for a window beginning 400 bp upstream of the translational start and continuing for ~200 bp, the level of polymorphism increases 10-fold from ~0.015 to a peak of 0.15. Linkage disequilibrium is strong in this region, generating two primary haplotypes that are at frequencies of 0.58 and 0.42 in the sample (Figures 5 and 6). There is no clear explanation for this pattern. Nucleotide divergence from *D. simulans* is also substantially increased in this region from ~0.15 for the remainder of the region to a remarkable peak of 0.62, consistent with a sharply localized increase in the mutation rate. However, an increase in mutational pressure alone should not create the degree of linkage disequilibrium observed in the region. A balanced polymorphism might create such a pattern (Kreitman and Hudson 1991), but this would not explain the increase in interspecific divergence, and the spike in diversity is too sharply defined for the balancing selection explanation to be likely. There are no identified regulatory elements under the spike in either haplotype. There are three amino acid polymorphisms in the 5′ end of the *Attacin C* coding sequence (Figure 5), but none of these is at an appropriate frequency in the sample or in adequate linkage disequilibrium with the promoter haplotypes to be a good candidate for involvement in the maintenance of the observed haplotype dimorphism. The effects of natural selection acting on a linked locus outside the survey region could conceivably have generated the observed data, but there is no independent evidence suggesting such a selected locus. It is possible that such a pattern could be generated by an extremely local increase in mutation rate coupled with strong population subdivision, but such severe subdivision should be detectable at other, unrelated loci, and no such population structure has been documented. Further study will be required to explain the structure of haplotypes in *Attacin C*.

Despite the fact that there is not an excess of amino acid substitutions in the *Attacin* genes, we have uncovered a number of polymorphisms that are likely to have some functional effect. In particular, the seven amino acid changes segregating between the converted and unconverted alleles in the second exon of *Attacin A* may be functionally important. The observation that the converted allele is rapidly increasing in frequency, perhaps through the action of natural selection, bolsters this assertion. Additionally, the analysis of the newly dis-

covered null allele of *Attacin A* should prove insightful with respect to the functional redundancy of antibacterial peptides. The fact that the null allele occurred on the background of the most common *Attacin AB* haplotype (which we infer to be young) suggests that this mutation is probably recent, and our failure to detect a second occurrence in a much larger sample indicates that it may be in mutation-selection balance. As pointed out by Hedengren *et al.* (2000), the *D. melanogaster* stock used for sequencing by the Berkeley Drosophila Genome Project carries a probable null allele of *Attacin C*. *H. cecropia* carries the nonfunctional remnants of two *Attacin* genes (Sun *et al.* 1991), and *D. melanogaster* harbors pseudogenes derived from two *Cecropin* antibacterial peptide genes (Kylsten *et al.* 1990). Additionally, *D. simulans*, *D. mauritiana*, and *D. sechellia* each carry a third young *Cecropin* pseudogene (Clark and Wang 1997; Date *et al.* 1998; Ramos-Onsins and Aguadé 1998). Therefore, it seems loss-of-function mutations may be relatively common in antibacterial peptide genes. More subtle phenotypic effects may result from mutations in regulatory regions, especially in light of the need for antibacterial peptide genes to be rapidly transcribed upon infection. A complex series of insertion/deletions and substitutions in the *Attacin B* promoter, which may not directly eliminate or create transcription factor binding sites, may alter transcription factor binding efficiency through changes in chromatin structure or physical spacing between bound factors.

The picture that emerges from this analysis is that *Attacin* antibacterial peptide genes retain high levels of polymorphism in *D. melanogaster* populations by exhibiting an unusual level of genomic instability, including paralogous gene conversion, insertion/deletion events, and gene duplication and loss. The ramifications of this instability on the flies' capacity to mount an effective antibacterial response is not easy to determine, but some of the observed variation is likely to have a functional effect. Several features of the data suggest departures from simple neutral evolution. Ultimately, the most convincing assessment of the functional consequences of polymorphism in insect antibacterial response genes will come from careful studies associating genotypic with phenotypic variation.

We thank Manolis Dermitzakis and Malia Fullerton for insightful discussion regarding the frequencies and allelic distributions of polymorphic sites. Comments from M. Dermitzakis, K. Montooth, M. Aguadé, and two anonymous reviewers improved the quality of the manuscript. This work was supported by grant AI46402 to A.G.C. from the National Institutes of Health and by a Dissertation Improvement Award DEB0073598 to B.P.L. and A.G.C. from the National Science Foundation. B.P.L. is supported by a National Science Foundation Graduate Research Fellowship.

LITERATURE CITED

Ando, K., M. Okada and S. Natori, 1987 Purification of Sarcotoxin II, antibacterial proteins of *Sarcophaga peregrina* (flesh fly) larvae. Biochemistry **26:** 226–230.

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Åsling, B., M. S. Dushay and D. Hultmark, 1995 Identification of early genes in the *Drosophila* immune response by PCR-based differential display: the *Attacin A* gene and the evolution of attacin-like proteins. Insect Biochem. Mol. Biol. **25:** 511–518.

Bulet, P., C. Hetru, J.-L. Dimarq and D. Hoffmann, 1999 Antimicrobial peptides in insects: structure and function. Dev. Comp. Immunol. **23:** 329–344.

Carlsson, A., P. Engström, E. T. Palva and H. Bennich, 1991 Attacin, an antibacterial protein from *Hyalophora cecropia*, inhibits synthesis of outer membrane proteins in *Escherichia coli* by interfering with *omp* gene transcription. Infect. Immun. **59:** 3040–3045.

Carlsson, A., T. Nyström, H. de Cock and H. Bennich, 1998 Attacin—an insect immune protein—binds LPS and triggers the specific inhibition of bacterial outer-membrane protein synthesis. Microbiology **144:** 2179–2188.

Carvalho, A. B., and A. G. Clark, 1999 Intron size and natural selection. Nature **401:** 344.

Clark, A. G., and L. Wang, 1997 Molecular population genetics of Drosophila immune system genes. Genetics **147:** 713–724.

Date, A., Y. Satta, N. Takahata and S. I. Chigusa, 1998 Evolutionary history and mechanism of the *Drosophila Cecropin* gene family. Immunogenetics **47:** 417–429.

Dushay, M. S., J. B. Roethele, J. M. Chaverri, D. E. Dulek, S. K. Syed *et al.*, 2000 Two *Attacin* antibacterial genes of *Drosophila melanogaster*. Gene **246:** 49–57.

Engström, P., A. Carlsson, A. Engström, Z. J. Tao and H. Bennich, 1984 The antibacterial effect of attacins from the silk moth *Hyalophora cecropia* is directed against the outer membrane of *Escherichia coli*. EMBO J. **3:** 3347–3351.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Fu, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915–925.

Fu, Y. X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gunne, H., M. Hellers and H. Steiner, 1990 Structure of preproattacin and its processing in insect cells infected with a recombinant baculovirus. Eur. J. Biochem. **187:** 699–703.

Hedengren, M., K. Borge and D. Hultmark, 2000 Expression and evolution of the *Drosophila Attacin/Diptericin* gene family. Biochem. Biophys. Res. Comm. **279:** 574–581.

Hoffmann, J. A., F. C. Kafatos, C. A. Janeway and R. A. B. Ezekowitz, 1999 Phylogenetic perspectives in innate immunity. Science **284:** 1313–1318.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 45–50.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the *superoxide ismutase* (*Sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

Hultmark, D., A. Engström, K. Andersson, H. Steiner, H. Bennich *et al.*, 1983 Insect immunity. Attacins, a family of antibacterial proteins from *Hyalophora cecropia*. EMBO J. **2:** 571–576.

Kang, D., G. Liu, H. Gunne and H. Steiner, 1996 PCR differential display of immune gene expression in *Trichoplusia ni*. Insect Biochem. Mol. Biol. **26:** 177–184.

Kliman, R. M., and J. Hey, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239–1258.

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

Kylsten, P., C. Samakovlis and D. Hultmark, 1990 The *Cecropin* locus in *Drosophila*: a compact gene cluster involved in the response to infection. EMBO J. **9:** 217–224.

Leigh Brown, A. J., and D. Ish-Horowicz, 1981 Evolution of the 87A and 87C heat-shock loci in *Drosophila*. Nature **290:** 677–681.

Lindsley, D. L., and E. H. Grell, 1967 *Genetic Variations of Drosophila melanogaster*. Carnegie Institution of Washington, Washington, DC.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:** 652–654.

Mellor, A. L., E. H. Weiss, K. Ramachandran and R. A. Flavell, 1983 A potential donor gene for the *bml* gene conversion event in the C57BL mouse. Nature **306:** 792–795.

Ourth, D. D., T. D. Lockey and H. E. Renis, 1994 Induction of cecropin-like and attacin-like antibacterial but not antiviral activity in *Heliothis virescens* larvae. Biochem. Biophys. Res. Commun. **200:** 35–44.

Ramos-Onsins, S., and M. Aguadé, 1998 Molecular evolution of the *Cecropin* multigene family in Drosophila: functional genes *vs.* pseudogenes. Genetics **150:** 157–171.

Rozas, J., and R. Rozas, 1999 DnaSP Version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for constructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Sawyer, S., 1989 Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6:** 526–538.

Shin, S. W., S. S. Park, D. S. Park, M. G. Kim, S. C. Kim *et al.*, 1998 Isolation and characterization of immune-related genes from the fall webworm, *Hyphantria cunea*, using PCR-based differential display and subtractive cloning. Insect Biochem. Mol. Biol. **28:** 827–837.

Stephens, J. C., 1985 Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:** 539–556.

Sturtevant, A. H., 1920 Genetic studies on *Drosophila simulans*. I. Introduction. Hybrids with *Drosophila melanogaster*. Genetics **5:** 488–500.

Sugiyama, M., H. Kuniyoshi, E. Kotani, K. Taniai, K. Kadono-Okuda *et al.*, 1995 Characterization of a *Bombyx mori* cDNA encoding a novel member of the Attacin family of insect antibacterial peptides. Insect Biochem. Mol. Biol. **25:** 385–392.

Sun, S. C., I. Lindstrom, J. Y. Lee and I. Faye, 1991 Structure and expression of the *Attacin* genes in *Hyalophora cecropia*. Eur. J. Biochem. **196:** 247–254.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Communicating editor: M. Aguadé