Molecular Population Genetics of Inducible Antibacterial Peptide Genes in *Drosophila melanogaster*

Brian P. Lazzaro and Andrew G. Clark

Molecular Biology and Genetics, Cornell University

Insects respond to septic infection in part by producing a suite of antimicrobial peptides that may be subject to host-pathogen coevolutionary dynamics. In order to infer population genetic forces acting on Drosophila antibacterial peptide genes, we examine global properties of polymorphism and divergence in the *Drosophila melanogaster defensin*, *drosocin*, *metchnikowin*, *attacin C*, *diptericin A*, and *cecropin A*, *B*, and *C* genes. As a functional class, antibacterial peptides exhibit low levels of interspecific amino acid divergence. There are multiple amino acid polymorphisms segregating within *D. melanogaster*, however, a high proportion of which change the charge or polarity of the variable residue. These polymorphisms are particularly prevalent in processed signal and propeptide domains. We find that models of coevolutionary ''arms races'' and selectively maintained hypervariability do not adequately describe the population dynamics of mature antibacterial peptides in *D. melanogaster*, but that a highly significant excess of high-frequency derived polymorphisms coupled with substantial intralocus linkage disequilibrium suggests that positive selection may act on antibacterial peptide genes. Some attributes of the data may be consistent with a simple demographic model of population founding followed by expansion, but departures from the equilibrium null tend to be more pronounced in the peptide genes than at other loci around the genome.

Introduction

Insects produce a battery of small, extracellularly secreted antimicrobial peptides as an important component of their innate immune defense. Six classes of antibacterial peptide have been characterized in *Drosophila melanogaster*, although recent whole-genome expression arrays suggest that there may be more (De Gregorio et al. 2001; Irving et al. 2001). Broadly speaking, drosocin and the attacins have activity directed against gram-negative bacteria (Bulet et al. 1993; Åsling, Dushay, and Hultmark 1995), whereas defensin is anti-gram-positive (Dimarcq et al. 1994). Diptericin A shows activity against both gram-positive and gram-negative bacteria (Wicker et al. 1990), and metchnikowin is active against gram-positive bacteria and filamentous fungi (Levashina et al. 1995). Cecropins have activity against gram-positive and gram-negative bacteria (Samakovlis et al. 1990) and fungi (Ekengren and Hultmark 1999). Production of all of these peptides is induced by septic infection. Drosophila antimicrobial peptides are typically synthesized in the fat body and circulating hemocytes of larvae and adults, although *cecropins B* and *C* are produced primarily during metamorphosis (Samakovlis et al. 1990; Tryselius et al. 1992). The different peptide classes vary in their mechanisms of microbial recognition and killing, but all of the peptides interact directly with the microbes they kill, creating the potential for host peptide genes to evolve coordinately with pathogens.

One major model of host-pathogen evolution driven by natural selection is the coevolutionary ''arms race'' (Dawkins and Krebs 1979). The premise of this model is that pathogens continually evolve to defeat host defenses, while the host continually evolves novel means of pathogen suppression. Therefore, new virulence and resistance alleles sequentially sweep through pathogen and host populations. The model posits that the host population should be in a continual state of recovery from selective sweep, so a generally low level of standing genetic variation is predicted. The expected degree of depression of variation depends on the intensity of selection and the frequency of favorable mutations (Wiehe and Stephan 1993). However, if the process is truly a ''race,'' the sweep events must be fairly common, even overlapping, with the selected amino acids frequently fixing in the population. If insect antimicrobial peptides evolve according to the arms race model, their genes should show elevated amino acid differentiation and low levels of standing variation, with indications of rapid and frequent allelic turnover via strong directional selection.

Under a second model, natural selection may favor genetic variability in a host population either if rare alleles are favored by virtue of their rarity or if variability in the host locus confers resistance to multiple distinct pathogens. A classic example of hypervariability generated by Darwinian selection is provided by the antigen recognition site of the vertebrate major histocompatibility complex (MHC) locus (Hughes and Nei 1988). Were insect antimicrobial peptide genes to conform to the hypervariability model, they would be expected to harbor substantial levels of amino acid polymorphism, perhaps exceeding even the level of silent variation in coding regions.

A third hypothesis is that insect antibacterial peptides may conform to the neutral model of molecular evolution (Kimura 1983). Under this model, the vast majority of mutations are sufficiently deleterious that they are rapidly removed from the population. The empirically observed mutations are thus neither favored nor disfavored by natural selection. Extensive theoretical work on this model makes it valuable as a null hypothesis, and there is some a priori evidence that *D. melanogaster* antibacterial peptides may evolve more or less neutrally. Prior surveys of natural variation in *D. melanogaster cecropin* and *diptericin A* genes have not detected marked departures from the neutral expectation (Clark and Wang 1997; Date et al. 1998; Ramos-Onsins and Aguadé 1998). The consistent and widespread observation of highly conserved antibacterial

Key words: *Drosophila melanogaster*, antibacterial peptide genes, polymorphisms, propeptide domains, innate immunity.

E-mail: brian.lazzaro@cornell.edu.

Mol. Biol. Evol. 20(6):914–923. 2003
DOI: 10.1093/molbev/msg109

peptide sequences across vast evolutionary distances (Boman 1995; Bulet et al. 1999) further argues against rapid, adaptive amino acid substitution as a general model of antibacterial peptide evolution.

This study revisits previously published surveys of natural variation in the *attacin C* (Lazzaro and Clark 2001), *cecropin A1*, *A2*, *B*, and *C*, and *diptericin A* (Clark and Wang 1997) genes and adds polymorphism and divergence data from the single-copy loci *defensin*, *drosocin*, and *metchnikowin*. The data from these genes are assembled and examined for systematic departures from a neutral evolutionary process. In particular, high rates of amino acid substitution and skews in the distribution of allele frequencies at polymorphic sites may be signatures of natural selection. It is likely that the pathogens *D. melanogaster* faces in North America are distinct from those found in sub-Saharan Africa, and because previous work has found significant population differentiation among alleles of the *diptericin A* and *cecropin* genes (Clark and Wang 1997) we focus here exclusively on North American alleles.

Materials and Methods

Sequence Collection

The *cecropin A1*, *A2*, *B*, and *C*, and *diptericin A* alleles sampled from Maryland, USA, were obtained from Clark and Wang (1997). For these loci only, *D. mauritiana* was used instead of *D. simulans* as an outgroup species. Polymorphism data for the *attacin C* locus were presented in Lazzaro and Clark (2001). These data were obtained from 12 lines of *D. melanogaster* derived from a population in Pennsylvania, USA, and one line of *D. simulans* derived from a population in California, USA. *Attacins A* and *B* are excluded from the analyses in this study because recurrent paralogous gene conversion between those genes results in strong departure from the models of independent mutation and infinite mutable sites (Lazzaro and Clark 2001).

New sequences were obtained for the *defensin*, *drosocin*, and *metchnikowin* loci. Sequence data were collected from the same 12 *D. melanogaster* lines and the same *D. simulans* line surveyed in Lazzaro and Clark (2001). Oligonucleotide primers for the *defensin*, *drosocin*, and *metchnikowin* genes were designed based on GenBank accession numbers Z27247, X98416, and AF030959. Primer sequences are available upon request. The survey region for *defensin* begins 1,123 bp upstream of the translational start codon, includes the entire 279 bp of coding sequence, and terminates 3 bp downstream of the stop codon. The *drosocin* region begins 933 bp upstream of translational start and continues to the end of the 195-bp coding region. The antibacterial peptide gene *attacin A* begins 1.2 kb downstream of the *drosocin* gene. Polymorphism and divergence in the sequence between *drosocin* and *attacin A* was described by Lazzaro and Clark (2001) and is qualitatively and quantitatively similar to the *drosocin* survey region described here. The *metchnikowin* survey region begins 1,499 bp upstream of the start codon, reads through the 159-bp coding sequence, and terminates 106 bp 3′ of the stop codon. *defensin*, *drosocin* and *metchnikowin* are all intronless. PCR-amplified templates were directly sequenced on either an Applied Biosystems 373 or a Beckman Coulter CEQ2000 automated sequencer, using modifications of the manufacturers' suggested protocols. All sequences were verified on both strands. The *defensin*, *drosocin*, and *metchnikowin* sequences have been deposited in GenBank under accession numbers AY224604 to AY224642.

Statistical Analysis

Sites with alignment gaps were excluded from all statistical analyses of nucleotide polymorphism and divergence data. Four sites (three in *drosocin* and one in *cecropin C*) where three nucleotides are segregating within *D. melanogaster* were also excluded. At all other polymorphic sites, the parsimonious assumptions were made that the state of the *D. simulans* allele reflects the ancestral state of the polymorphism and that the probability of backmutation within *D. melanogaster* is negligible. Eleven sites where *D. simulans* has a third nucleotide, different from either state of a *D. melanogaster* polymorphism, were excluded from analyses that make use of outgroup information. Polymorphic sites tables for *diptericin A*, the *cecropins*, and *attacin C* can be found in Clark and Wang (1997) and Lazzaro and Clark (2001). Supplemental figures 1–3 for this manuscript show polymorphic sites and fixed differences observed in *defensin*, *drosocin*, and *metchnikowin* (see online Supplementary Material).

With the exception of $\hat{C}$ (Hudson 1987), the MK *G*-test (McDonald and Kreitman 1991), and a modified HKA test (Hudson, Kreitman, and Aguadé 1987), which were calculated in DnaSP 3.51 (Rozas and Rozas 1999), population genetic estimators and statistics were calculated using a program written in ANSI C. Probabilities of obtaining equivalent or more extreme test statistics were determined by simulation of neutral genealogies as in Hudson (1990) using his "ms" coalescence simulator (Hudson 2002; http://home.uchicago.edu/~rhudson1/source/mksamples.html). All simulations were conditioned on the empirical sample size, the empirically observed number of segregating sites, and the length in base pairs of the empirical sample. Each null distribution is based on 10,000 neutral genealogies simulated with the recombination parameter set to each of three values: (1) 0, that is, no recombination between sites; (2) $\hat{C}$, the recombination rate inferred from the empirical sample (Hudson 1987); and (3) a recombination rate, $4\hat{N}r$, where r is the meiotic recombination rate determined by Carvalho and Clark (1999) at the cytological position of each locus, and N is the effective population size, assumed to be 10^6 (Kreitman 1983; Andolfatto and Przeworski 2000).

Because extant North American *D. melanogaster* are believed to be derived from an ancestral African population (David and Capy 1988), we tested the empirically observed data against simple null models of population founding followed by expansion. These were approximated by simulating equilibrium neutral populations maintaining effective size N_0 for $4N_0$ generations, then introducing a single bottleneck of varying severity at various times before present. In all cases, the bottleneck was maintained for

Table 1
Per-Base Measures of Polymorphism, Divergence, and Recombination in *D. melanogaster* Antibacterial Peptide Genes

Locus	Cytological Position	N[a]	Length[b]	S[c]	$4\hat{N}r$[d]	$\hat{C}$[e]	θ[f]	π[f]	k[g]
Defensin	48C	12	1383	21	0.0468	0.0153	0.0050	0.0045	0.0525
Attacin C	50A	12	3097	100	0.0222	0.0107	0.0107	0.0120	0.0908
Drosocin	51A	12	1078	44	0.0688	0.0114	0.0135	0.0128	0.0734
Metchnikowin	55A	12	1738	48	0.0457	$<10^{-6}$	0.0091	0.0064	0.0529
Diptericin A	56A	12	345	12	0.0792	$<10^{-5}$	0.0115	0.0058	0.0527
Cecropin A1	99E	6	442	9	0.0400	0.071	0.0089	0.0094	0.0264
Cecropin A2	99E	5	457	19	0.0400	0.057	0.0200	0.0197	—
Cecropin B	99E	8	411	16	0.0400	0.143	0.0150	0.0129	0.0179
Cecropin C	99E	7	394	19	0.0400	0.651	0.0197	0.0181	0.0461

[a] Sample size.

[b] Length of the survey region measured in base pairs, not including alignment gaps.

[c] Number of polymorphic positions, excluding those inside alignment gaps and with three nucleotides segregating.

[d] Measure of population recombination rate measured from laboratory recombination experiments (see *Materials and Methods*).

[e] Measure of population recombination rate inferred from the empirical sequence data.

[f] Estimates of population-level heterozygosity.

[g] Average percent silent nucleotide divergence between *D. melanogaster* and *D. simulans* (*D. mauritiana* in *diptericin A* and the *cecropins*; *cecropin A2* is deleted in *D. melanogaster* sibling species preventing calculation of *k*).

0.0001 × N_0 (approximately 100 in *D. melanogaster*) generations, after which the population was allowed to grow to a size of 0.1 × N_0. Five parameter combinations were tested, with 10,000 genealogies simulated under each parameter set. In three cases, the bottleneck was set to have occurred 0.002 × N_0 (approximately 2,000) generations before present, and the population size was reduced to either 0.001 × N_0 (approximately 1,000), 0.0001 × N_0 (approximately 100), or 0.00001 × N_0 (approximately 10) individuals during the bottleneck phase. In two additional cases, the severity of the bottleneck was set to 0.0001 × N_0, the value under which the empirical data had the highest probability in the first sets of simulations, but the age of bottleneck was set to either 0.0005 × N_0 (approximately 500) or 0.05 × N_0 (approximately 5 × 10^4) generations before present. The conservative assumption of no recombination is made in all simulations incorporating a demographic component. Varying the recombination rate had substantial effect only after a very ancient bottleneck (2 × N_0 generations before present), and in this scenario simulations were similar to those assuming no demographic structure (data not shown). In no case was migration between the founded and ancestral population simulated.

Results

Previous studies have documented high levels of silent genetic variation in and around the *D. melanogaster cecropin* and *attacin* gene families (Date et al. 1998; Ramos-Onsins and Aguadé 1998; Lazzaro and Clark 2001). Polymorphism data obtained from three single-copy genes in this study, *defensin*, *drosocin*, and *metchnikowin*, reaffirm that silent variation is not depressed around antibacterial peptide genes, contrary to the expectation under recurrent strong directional selection. Estimates of θ range from 0.005 in *defensin* to 0.014 in *drosocin* (table 1). Interspecific silent divergence estimates for the antibacterial peptide genes range from 0.0179 at *cecropin B* to 0.1191 at *diptericin A* (table 1). Divergence in the peptide genes is consistent with, but generally smaller than, *D. melanogaster–D. simulans* divergence estimates reported for a variety of autosomal and X-linked loci, where the average silent divergence of *D. simulans* from *D. melanogaster* was 0.108 (see table 1 in Begun and Whitley 2000*a* for comparison). Further evidence against the recurrent directional selection hypothesis is provided by the very low nonsynonymous divergences in the antibacterial peptide genes (table 2). We observe no fixed amino acid replacements between *D. simulans* and *D. melanogaster* in *drosocin*, *metchnikowin*, *cecropin A1*, or *cecropin B*, one fixed amino acid replacement each in *cecropin A2* and *cecropin B*, and two fixed amino acid replacements in *Defensin*. This is despite the fact that there 16 amino acid polymorphisms segregating within these seven genes (table 2). Although the McDonald-Kreitman *G*-test (McDonald and Kreitman 1991) was not significant when calculated at any single locus, the short sequence lengths and small numbers of sites hamper the statistical power of the test for individual peptide genes. When data from all loci were pooled, the test approached significance ($P = 0.076$) in the direction of excess polymorphism or lack of divergence (table 2).

Antibacterial peptides are typically composed of three domains: a signal peptide, a propeptide, and the mature peptide. The signal and propeptide domains are proteolytically cleaved to release the mature peptide in its active form, although additional posttranslational modification is sometimes required for complete activation (e.g., Bulet et al. 1993). The observed rate of amino acid substitution was highest in the processed domains, where there are a total of 18 amino acid polymorphisms in 286 sites across all nine genes. There are five fixed replacements in the 262 sites in processed domains of the eight genes from which divergence estimates could be obtained (*cecropin A2* is deleted in *D. melanogaster* sibling species; the last codon in *diptericin A* is absent from *D. mauritiana*). In contrast, in 520 residues of mature peptide across nine genes, there are only 11 amino acid polymorphisms observed, and six fixed differences in the 497 residues with outgroup information (fig. 1). Interestingly, 17 of the total 29 amino acid polymorphisms change polarity or charge at the variable residue (Lehninger, Nelson, and Cox 1993), and several of these are at intermediate frequency (fig. 1). Twelve of the 17 radical amino

Table 2
Polymorphism and Divergence in Antibacterial Gene Coding Regions

Locus	P_S[a]	P_R[a]	F_S[a]	F_R[a]	θ_S[b]	θ_R[b]	k_S[c]	k_R[c]	MK[d] (*P*-Value)
Defensin	4	5	7	2	0.0213	0.0076	0.1445	0.1500	2.157 (0.142)
Attacin C	18	4	25	6	0.0175	0.0074	0.0852	0.0417	0.012 (0.914)
Drosocin	4	2	1	0	0.0248	0.0047	0.0671	0.0048	undefined
Metchnikowin	2	2	1	0	0.0080	0.0058	0.0283	0.0015	undefined
Diptericin A	7	7	6	4	0.0286	0.0088	0.1102	0.0237	0.236 (0.627)
Cecropin A1	3	1	0	0	0.0068	0.0010	0.0052	0.0004	undefined
Cecropin A2	8	1	—	—	0.0200	0.0021	—	—	—
Cecropin B	5	3	0	0	0.0100	0.0060	0.0111	0.0020	undefined
Cecropin C	4	2	7	1	0.0085	0.0043	0.0030	0.0015	0.882 (0.348)
All pooled[e]	47	26	47	13					3.145 (0.076)

[a] Counts of silent and replacement polymorphisms and fixed differences found in each locus.

[b] Population estimates of polymorphism at silent and replacement sites.

[c] Average divergence of *D. melanogaster* from *D. simulans* (*D. mauritiana* in *Diptericin A* and *Cecropins*) at silent and replacement sites.

[d] McDonald-Kreitman *G*-test (McDonald and Kreitman 1991).

[e] *Cecropin A2* is deleted in *D. mauritiana* preventing estimation of divergence, and therefore is not included in pooled totals.

acid polymorphisms are located in processed peptide domains. Of the 11 amino acid replacements fixed between species, five are nonconservative. Two of these are found in the *diptericin A* signal peptide, with the other three in the *attacin C* mature peptide.

There is one position each in *diptericin A* and *cecropin C* with three amino acid residues segregating. At both sites, the two mutations in *D. melanogaster* are nonconservative with respect to the ancestral state inferred from *D. mauritiana.* The three-state position in the *cecropin C* signal peptide has additionally mutated a fourth time in the history of *Drosophila cecropin* genes, as this residue is fixed for a nonconservative replacement between *cecropins A1* and *A2* and *cecropins B* and *C* (fig. 1). Another *cecropin* residue shows evidence of convergent multiple mutations in the C-terminal portion of *cecropins B* and *C*, where an Ala→Gly mutation appears to have occurred independently in homologous positions of *D. melanogaster cecropin B* and *D. mauritiana cecropin C* (fig. 1). The ancestral state at this position is Ala in both loci, as determined from sequences obtained from GenBank of *D. simulans* (Y16860 [Ramos-Onsins and Aguadé 1998] and AB010790 [Date et al. 1998]), *D. yakuba*, *D. teissieri*, *D. orena*, *D. erecta*, *D. takahashii* (AB047059 to AB047063 [Date-Ito et al. 2001]) and *D. virilis* (U71249 [Zhou, Nguyen, and Kimbrell 1997]).

The nearly significant excess of amino acid polymorphisms relative to replacement fixations observed in the peptide data could potentially reflect a low level of purifying selection if the polymorphisms are nearly neutral or slightly deleterious, particularly since most of the polymorphic sites are in processed domains that may experience little functional constraint. If this is the case, then the allele frequency spectrum of polymorphic sites should resemble that predicted under selective neutrality. Derived mutations at high frequency are rare under a neutral process but may be more common if neutral polymorphisms ''hitchhike'' to high frequency when nearby sites are selectively favored (Fay and Wu 2000), particularly during and soon after the selective event (Przeworski 2002). Antibacterial peptide genes tend to contain an excess of high-frequency derived sites, measured by Fay and Wu's *H* (table 3). This tendency is especially pronounced when critical values are determined assuming a population recombination rate equivalent to meiotic recombination rates observed in the laboratory (Fisher's combined probability, $P = 6.04 \times 10^{-4}$) but is apparent even under the conservative assumption of no recombination (Fisher's combined probability, $P = 0.014$).

It has been suggested that demographic structure may cause empirically observed values of *H* to depart from the neutral panmictic null expectation (Przeworski 2002). Because North American *D. melanogaster* are founded from an ancestral African population (David and Capy 1988), we evaluated the empirically observed values of *H* under several simple null models of founding events followed by rapid population growth. These were approximated by simulation of population bottlenecks of varying severity and age (see *Materials and Methods*). The observed data are inconsistent with very ancient, very strong, and weak population bottlenecks before expansion (table 3). However, a moderate bottleneck that briefly constricted the population to 1/10,000 of its original size before allowing it to grow to 1/10 of its original size adequately fit the data. This model plausibly fits the natural history of *D. melanogaster*. The fit of the data to a moderate bottleneck model was little affected by setting the time of the bottleneck to 500, 2,000 or 50,000 Drosophila generations before present (table 3) or by varying the recombination rate (data not shown).

Because demography should have genome-wide effects, comparisons between the peptide loci and functionally unrelated *D. melanogaster* genes can reveal whether population structure causes the observed departures of *H* values from the null expectation. Andolfatto and Przeworski (2001) have assembled polymorphism data from *D. melanogaster* loci previously surveyed by other authors. We subsampled their data, retaining only North American alleles from each locus when five or more North American alleles were sampled and were segregating for five or more polymorphic sites. *H* was calculated for each of these 12 loci (*Acp26A*, *est6*, *G6PD*, *hsp83*, *mlc1*, *per*, *pgd*, *ref(2)p*, *SOD*, *tpi*, *v*, and *w*), and the probability of observing an *H* as or more negative than that observed was determined by simulation assuming panmixia and no recombination. The combined probability for the Andol-

```
Attacin C
                  signal peptide          propeptide                   mature peptide →
2CPA 1        MSKIVLLIVVIVGVLGSLAVA LPQRPYTQPLIYYPPPPTPPRIYRARR QVLGGSLTSNPSGGADARLDLSKAVGTPDHHVIGQVFAAGNTQTKPVSTPVTSGATLGYNNHGHGLELTKTH
2CPA 7        ..................... ........................... ........................................................................
2CPA 12       ..................... ............................................................................................................
2CPA 14       ..................... ........................... ........................................................................
2CPA 43       ...T................. ........................... ........................................................................
2CPA 46       ...T................. ........................... ........................................................................
2CPA 51       ..................... ........................... ...................................L....................................
2CPA 103      ..................... ........................... ........................................................................
2CPA 105      ..................... ........................... ........................................................................
2CPA 118      ..................... ........................... ...................................L....................................
2CPA 122      ..................... ........................... ...................................L....................................
2CPA 129      .......F............. ........................... ........................................................................
D.simulans    ...........I......... ........................... ........................I...............................................
                 *   *

                 mature peptide continued →
2CPA 1        TPGVRDSFQQTATANLFNNGVHNLDAKAFASQNQLANGFKFDRNGAALDYSHIKGHGATLTHANIPGLGKQLELGGRANLWQSQDRNTRLDLGSTASKWTSGPFKGQTDLGANLGLSHYFG
2CPA 7        .........................................................................................................................
2CPA 12       .........................................................................................................................
2CPA 14       .........................................................................................................................
2CPA 43       .........................................................................................................................
2CPA 46       ......................................................................................................................S..
2CPA 51C      .........................................................................................................................
2CPA 103      .........................................................................................................................
2CPA 105      .........................................................................................................................
2CPA 118      .........................................................................................................................
2CPA 122      ......................................................................................................................S..
2CPA 129      .........................................................................................................................
D.simulans    .......R......S.........................G.....N..........................................................................
                     *      *                               *                                                                       *
```

```
Diptericin A
                 signal peptide    pro                     mature peptide
B314          MQFTIAVALLCCAIASTLA YPMP DDMTMKPTPPPQYPLNLQGGGGGQSGDGFGFAVQGHQKVWTSDNGRHEIGLNGGYGQHLGGPYGNSEPSWKVGSTYTYRFPNF
B306          ................... .... ........................R......G...................................................
B226          ................... .... ...................................................................................
B225          ................... .... ...................................................................................
B222          ................... .... ...................................................................................
B208          ................... .... ...................................................................................
B202          ................... .... ...................................................................................
B141          ................... .... .................................................R.................................
B137          ................... .... ...................................................................................
B115          ................... .... ...................................................................................
B101          ................... .... ........................T.................................................N........
B009          ................... .... ...................................................................................
D.mauritiana  ......F........A... .... ........................R......G..................................................-
                    *        *                                 *                        *
```

```
Defensin
                 signal peptide           propeptide                     mature peptide
2CPA 1        MKFFVLVAIAFALLACMAQA QPVSDVDPIPEDHVLVHEDAHQEVLQHSRQKR ATCDLLSKWNWNHTACAGHCIAKGFKGGYCNDKAVCVCRN
2CPA 7        .....P..........V... .................D.............. ........................................
2CPA 12        ..............T..... ................................ ........................................
2CPA 14       .....P..........V... .................D.............. ........................................
2CPA 43       ..............T..... ................................ ........................................
2CPA 46       ..............T..... ................................ ........................................
2CPA 51       ..............T..... ................................ ........................................
2CPA 103      ..............T..... ................................ ........................................
2CPA 105      ..............T..... ................................ ........................................
2CPA 118      ................V... ................................ ........................................
2CPA 122      .................... ....................N........... ........................................
2CPA 129      ................V... ................................ ........................................
D.simulans    .................... .............A..........V....... ........................................
                            * *                        *
```

```
Metchnikowin
                 signal peptide           pro    mature peptide
2CPA 1        MQLNLGAIFLALLGVMATATSVLA EP HRHQGPIFDTRPSPFNPNQPRPGPIY
2CPA 7        ........................ .. ..........................
2CPA 12       ........................ .. ..........................
2CPA 14       ........................ .. ..........................
2CPA 43       ........................ .. ..........................
2CPA 46       ........................ .. ..........................
2CPA 51       ..................T..... .. ..R.......................
2CPA 103      ........................ .. ..........................
2CPA 105      ........................ .. ..........................
2CPA 118      ........................ .. ..........................
2CPA 122      ........................ .. ..........................
2CPA 129      ........................ .. ..........................
D.simulans    ........................ .. ..........................
                                *
```

```
Drosocin
                signal peptide     pro  mature peptide          propeptide
2CPA 1        MKFTIVFLLLACVFAMAVA TP GKPRPYSPRPTSHPRPIRV RREALAIEDHLAQAAIRPPPILPA
2CPA 7        ................... .. ................... ...........T............
2CPA 12       ................... .. ................... ........................
2CPA 14       ................... .. ................... ...........T............
2CPA 43       ................... .. ................... ...........T............
2CPA 46       ................... .. ................... ...........T............
2CPA 51       ................... .. ................... ...........T...........V
2CPA 103      ................... .. ................... ...........T............
2CPA 105      ................... .. ................... ........................
2CPA 118      ................... .. ................... ........................
2CPA 122      ................... .. ................... ........................
2CPA 129      ................... .. ................... ...........T............
D.simulans    ................... .. ................... ........................
                                                                    *
```

```
                   signal peptide               mature peptide
Cecropin A1
B115          MNFYNIFVFVALILAITIGQSEA GWLKKIGKKIERVGQHTRDATIQGLGIAQQAANVAATARG
B137          ....................... ........................................
B205          ....................... ........................................
B222          ....................... ........................................
B208          ..........P............ ........................................
B316          ..........P........E... ........................................
D. mauritiana ....................... ........................................
                                 *

Cecropin A2
B226          ......L................ ........................................
B225          ....................... ........................................
B208          ....................... ........................................
B141          ....................... ........................................
B115          ..........T............ ........................................
                    *   *

Cecropin B
B115          ...NK...........SL.N... ...R.L......I.......S..V................
B137          ...NK...........SL.N... ...R.L......I.......S..V................
B141          ...NK....L......SL.N... ...R.L......I.......S..V................
B202          ...NK...........SL.N... ...R.L......I.......S..V................
B108          ...NK......I....SL.N... ...R.L......I.......S..V................
B222          ...NK...........SL.N... ...R.L......I.......S..V...........G....
B225          ...NK...........SL.N... ...R.L......I.......S..V................
B226          ...NK...........SL.N... ...R.L......I.......S..V...........G....
D. mauritiana ...NK...........SL.N... ...R.L......I.......S..V................
                (**)

Cecropin C
B009          ....K...........S...... .....L..R...I...........................
B101          ....K...........S...... .....L..R...I...........................
B115          ....K...........S...... .....L..R...I...........................
B202          ....K...........S...... .....L..R...I.................T.........
B205          ....Q...........S...... .....L..R...I...........................
B208          ....E...........S...... .....L..R...I...........................
B316          ....K...........S...... .....L..R...I................R..........
D. mauritiana ....K...........S...... .....L..R...I......................G....
                  *                                            **
```

Table 3
Empirically Observed Values of *H* and Their Probabilities Under Various Recombinational and Demographic Null Models[a]

Locus	H[b]	nr[c]	$\hat{C}$[c]	$4\hat{N}r$[c]	WB[d]	B[e]	SB[f]	RB[g]	AB[h]
Defensin	−2.212	0.174	0.184	0.170	0.219	0.451	**0.010**	0.211	0.413
Attacin C	−9.636	0.168	0.093	0.068	0.228	0.474	**0.011**	0.473	0.440
Drosocin	−8.485	0.094	**0.049**	**0.018**	0.126	0.338	**0.009**	0.340	0.315
Metchnikowin	−8.606	0.105	0.072	**0.016**	0.143	0.343	**0.010**	0.352	0.336
Diptericin A	−7.273	**0.012**	**0.012**	**0.006**	**0.012**	0.096	**0.003**	0.098	0.086
Cecropin A1	−0.533	0.287	0.346	0.338	0.316	0.505	**0.003**	0.519	0.452
Cecropin B	−1.643	0.205	0.212	0.210	0.229	0.354	**0.010**	0.448	0.406
Cecropin C	4.6190	0.975	0.999	0.997	0.954	0.993	0.953	0.995	0.996
Combined probability[i]		0.014	0.006	6.04×10^{-4}	0.030	0.486	1.27×10^{-8}	0.424	0.439

[a] See *Materials and Methods* for details. Bold numbers indicate $P < 0.05$.

[b] Empirically observed value of H (Fay and Wu 2000).

[c] Neutral process simulated with recombination rate equal to 0, $\hat{C}$, or $4\hat{N}r$ (see *Materials and Methods*).

[d] Weak bottleneck, 1,000 individuals 2,000 generations ago.

[e] Bottleneck, 100 individuals 2,000 generations ago.

[f] Severe bottleneck, 10 individuals 2,000 generations ago.

[g] Recent bottleneck, 100 individuals 500 generations ago.

[h] Ancient bottleneck, 100 individuals 50,000 generations ago.

[i] Fisher's combined probability (Sokal and Rohlf 1995, pp. 794–797) across all loci. *Cecropin A2* is absent in close relatives of *D. melanogaster*, preventing the calculation of *H* at that locus.

fatto and Przeworski loci is marginally significant ($\chi^2_{(24)} = 36.72, P = 0.046$), although less so than the combined probability of the observed peptide data ($\chi^2_{(18)} = 35.55$, $P = 0.014$). When two nonpeptide loci that have individually significant negative H values (*vermilion*, $P = 0.040$; *white*, $P = 0.046$) are excluded, the combined probability across the genome sample is nonsignificant ($\chi^2_{(20)} = 24.09$, $P = 0.237$). When the single peptide locus with an individually significant H is excluded (*Diptericin A*, $P = 0.012$), the combined probability of the remaining peptide loci is still nearly significant ($\chi^2_{(14)} = 22.02, P = 0.078$).

Excepting the *cecropins*, the antibacterial peptide genes tend to show an excess of linkage disequilibrium, with $\hat{C}$ estimated from the data (Hudson 1987) ranging from two to four orders of magnitude smaller than $4\hat{N}r$ calculated from the laboratory meiotic recombination rate (table 1). The amount of linkage disequilibrium in a data set can be measured using Z_{nS}, a statistic based on the sum of coefficients of linkage disequilibrium across all pairs of sites in the sample (Kelly 1997). As expected, there is no evidence of excess linkage disequilibrium in any of the antibacterial peptide genes when critical values of Z_{nS} are determined by simulations assuming either no recombination or a recombination rate equal to the empirically estimated C. More extreme P-values are observed, however, when the null distribution of Z_{nS} is determined using the recombination parameter estimated from laboratory recombination rates (Fisher's combined probability $P = 2.77{\times}10^{-5}$ [table 4]), indicating that the antibacterial peptide genes have an overall excess of linkage disequilibrium, given our best estimates of their actual meiotic recombinational environments. Excess linkage disequilibrium can be generated by natural selection or under numerous demographic scenarios, although the effect of positive selection on disequilibrium is expected to be short-lived (Przeworski 2002). The *cecropin* genes differ from the remainder of the peptide genes in that $\hat{C}$ is approximately equal to $4\hat{N}r$ in *cecropins A1* and *A2*, and $\hat{C}$ is greater than $4\hat{N}r$ in *cecropins B* and *C*. Gene function, genome arrangement, and other possible explanations for this discrepancy are considered in the *Discussion*.

A genome-wide reduction in $\hat{C}$ relative to $4\hat{N}r$ has previously been noted in *D. melanogaster* (Andolfatto and Przeworski 2000). The same tendency is observed in the non-*cecropin* peptide genes, although the extremely small ratios of $\hat{C}/4\hat{N}r$ observed at *metchnikowin* and *diptericin A* far exceed the smallest ratios observed in cosmopolitan or exclusively North American samples of nonpeptide genes. Excluding peptide genes *metchnikowin* and *diptericin A* and nonpeptide *Ref(2)P*, the distribution of $\hat{C}/4\hat{N}r$ ratos is similar between the antibacterial peptide genes and North American alleles of nonpeptide loci distributed around the genome. *Ref(2)P*, involved in Drosophila immunity to rhabdovirus sigma, departs in the opposite direction from most genes with $\hat{C}$ much larger than $4\hat{N}r$ (Wayne, Contamine, and Kreitman 1996). The distributions of Z_{nS} values calculated from the peptide data and from North American alleles of the Andolfatto and Przeworski data are qualitatively and quantitatively similar (data not shown).

The *drosocin* locus has one of the most extreme values of Z_{nS} and of H among the peptide genes. There are only two amino acid polymorphisms and no fixed differences in *drosocin*, one of the polymorphisms being a polarity-changing Ala/Thr polymorphism at intermediate frequency in the propeptide domain (fig. 1). A preliminary

←

FIG. 1.—Amino acid alignments of North American alleles of *attacin C*, *diptericin A*, *defensin*, *metchnikowin*, *drosocin*, and the *cecropins*. The bottom sequence in each gene is from the outgroup species, except for *cecropin A2*, which is absent in *D. melanogaster* sibling species. Residues identical to the uppermost allele in an alignment are indicated with periods. Stars indicate substitutions that change charge or polarity at the variable residue. Two residues in *cecropin B* that contain nonconservative fixed differences relative to *cecropins A1* and *A2* are starred with parentheses. Two positions, in *diptericin A* and *cecropin C*, are segregating for three amino acids in *D. melanogaster*. Both *D. melanogaster* mutations are nonconservative with respect to the ancestral state inferred from *D. mauritiana* at both positions.

Table 4
Empirically Observed Values of Z_{nS} and Their Probabilities Under Three Recombinational Scenarios

Locus	Z_{nS}[a]	nr[b]	$\hat{C}$[b]	$4\hat{N}r$[b]
Defensin	0.176	0.891	0.530	0.138
Attacin C	0.175	0.937	0.359	0.071
Drosocin	0.223	0.765	0.339	**0.006**
Metchnikowin	0.318	0.430	0.436	**0**
Diptericin A	0.564	0.097	0.091	**0**
Cecropin A1	0.329	0.729	0.299	0.275
Cecropin A2	0.370	0.784	0.377	0.305
Cecropin B	0.196	0.948	0.503	0.682
Cecropin C	0.160	0.999	0.888	0.993
Combined probability[c]		0.971	0.461	**2.77×10^{-5}**

[a] Empirically observed value of Z_{nS} (Kelly 1997).

[b] Neutral, panmictic process simulated with recombination rate equal to 0, $\hat{C}$, or $4\hat{N}r$ (see *Materials and Methods* for details). Bold numbers indicate $P < 0.05$.

[c] Fisher's combined probability (Sokal and Rohlf 1995, pp. 794–797) across all loci. Simulated *P*-values of 0 were considered to be 0.001 for the calculation of Fisher's combined probability.

analysis suggested that the Ala alleles were deficient in polymorphism compared with the Thr alleles, although Ala is inferred to be the ancestral state. We pursued this observation by sequencing approximately 205 bp upstream and 357 bp downstream of the Ala/Thr site in an additional 20 chromosomes collected in Pennsylvania, USA, in 2001 (these sequences have been deposited into GenBank under accession numbers AY224643 to AY224662). Surprisingly, only three of these additional lines had Thr at the variable position, although seven of the original 12 alleles encoded Thr alleles. It is significantly unlikely that these two samples (7 Thr:5 Ala and 3 Thr:17 Ala) were taken from populations with the same allele frequency ($\chi^2_{(1)} = 6.56$, $P = 0.011$), raising the possibility that the Ala allele might have substantially increased in frequency in only two years.

Twenty-one of the 22 Ala alleles cluster in a single clade distinct from the Thr alleles. The remaining Ala allele creates the most basal *D. melanogaster* branch in the entire genealogy when the tree is rooted with the *D. simulans* sequence (fig. 2). Only seven sites are segregating within the 21-allele internal Ala clade. Six of these are unique to that clade, and none have a frequency higher than 0.095 within the clade. In contrast, there are 21 sites segregating among the 10 Thr alleles (fig. 3). Despite the small number of sites, Tajima's *D* statistic (Tajima 1989) is significantly negative within the internal Ala clade, indicating a significant excess of rare polymorphisms ($D = -1.70$, $P = 0.030$; all simulations in this short sequence window assume no recombination). Among the Thr alleles only, $D = -0.16$ ($P = 0.477$). Inclusion of the basal Ala allele with those in the internal Ala clade adds several more rare polymorphisms ($D = -2.15$, $P = 0.005$) and results in a highly significant excess of high-frequency derived mutations ($H = -11.67$, $P = 0.001$). With the exclusion of the basal Ala allele, however, the common derived sites are no longer polymorphic but become fixed differences relative to *D. simulans*, resulting in nonnegative value of *H* ($H = 0.848$, $P = 0.661$) and illustrating a peculiarity of the *H* test. *H* is significantly negative over the entire 32 allele data set ($H = -11.94$, $P = 0.016$) but not among Thr alleles alone ($H = -0.089$, $P = 0.270$). Overall, these data are consistent with a model of positive selection driving the expansion in frequency of the internal Ala clade. *D. simulans* and *D. yakuba* (not shown) sequences both indicate that Ala is the ancestral state at this residue, suggesting that this position may have mutated from Ala to Thr early in the history of the *D. melanogaster* lineage and then mutated back to Ala in the expanding clade (fig. 2). If the Ala/Thr polymorphism is itself the target of selection, this history implies that selection pressure has changed over time. The presence of the basal Ala allele also complicates the interpretation that the Ala/Thr polymorphism is the target of selection, although this allele may be a recombinant. Alternatively, selection may be acting not on the Thr/Ala polymorphism, but rather on a site linked to the high-frequency Ala allele.

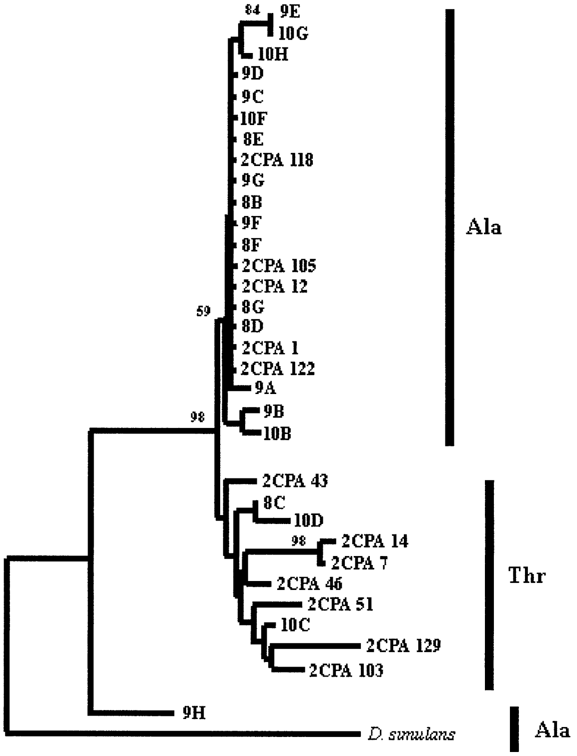

FIG. 2.—Gene genealogy of an expanded sample of *drosocin* alleles. Genealogy was constructed with neighbor-joining method using p-distance and pairwise deletion and significance tested with 1,000 bootstrap replicates using MEGA software version 2.1 (Kumar et al. 2001). Alleles are labeled with their state at the Ala/Thr polymorphism at *drosocin* position 52. Bootstrap support is listed for nodes with greater than 55% support.

Discussion

Despite their different protein structures, bacterial targets, and bactericidal mechanisms, there are some evolutionary commonalities observed across antibacterial peptide genes. The sequence polymorphism data presented here and in previous studies (Clark and Wang 1997; Date et al. 1998; Ramos-Onsins and Aguadé 1998; Lazzaro and Clark 2001) allow the rejection of a simple coevolutionary

```
               5'  Drc cds                            3' of Drc stop
               -   ------
               1   11111                             1111111111111111111112222222223333333
               8   874224  3   4555555566666668899999000012223467788999999033557890112233
               8   352419087   5145678901245677823456012440182195608456789258288367476706

8C             C   CTCTTCATC   GC--ACCGCGGGAAGTTAAA--TTAAATTTACTTCCA------GCATTCCGGAGTATG
10D            .   ....C....   ..--................--...............------A..............
2CPA 129       .   ..TGCT...   T.--.............C..--...............------AG...G.......C.
2CPA 51        .   .AT.....T   T.--................--...............AAAACT.G.............
2CPA 43        .   .A...T...   T.--................--...............------...........A...
2CPA 46        .   .A.......   T.--................--...............AAAACT............G..
2CPA 14        .   .A.......   T.--................--CC.C......A....AAAACT......T........
2CPA 7         .   .........   T.--................--CC.C......A....AAAACT......T........
10C            .   .........   T.--................--...............------.G...G.........
2CPA 103       .   .........   T.--................--.....A.........------.G...G.......C.
9B             .   .....TG..   T.--................--.....A.........------.A.............
10B            .   A....TG..   T.--................--...............------.A.............
8B             .   .....TG..   T.--................--...............------...............
9A             .   .....TG..   T.--................--...............------..............T
2CPA 1         .   .....TG..   T.--................--...............------...............
2CPA 12        .   .....TG..   T.--................--...............------...............
2CPA 105       .   .....TG..   T.--................--...............------...............
2CPA 118       .   .....TG..   T.--................--...............------...............
2CPA 122       .   .....TG..   T.--................--...............------...............
8D             .   .....TG..   T.--................--...............------...............
9C             .   .....TG..   T.--................--...............------...............
10H            G   .....TG..   T.--................--...............------...............
9D             .   .....TG..   T.--................--...............------...............
8E             .   .....TG..   T.--................--...............------...............
9E             ?   ??...TG..   T.--................--...............------......TT.......
9F             .   .....TG..   T.--................--...............------...............
8F             .   .....TG..   T.--................--...............------...............
8G             .   .....TG..   T.--................--...............------...............
9G             .   .....TG..   T.--................--...............------...............
10F            .   .....TG..   T.--................--...............------...............
10G            ?   ??...TG..   T.--................--...............------......TT.......
9H             .   ....CTG..   TG--C.........A..TTTTTCC...A..C....T.AAAACT...........A...
D.simulans     .   ...GC.GC.   TGCACATCTTAATGAG.TTTTT.CT.C.CG.T.AA.GAAAACT..CGA....TA....
                         *
```

FIG. 3.—Polymorphic sites segregating within *D. melanogaster* and fixed differences relative to *D. simulans* in a 563-bp window surrounding the Ala/Thr polymorphism in *drosocin*. Residues identical to the uppermost allele are indicated with periods. Positions are numbered relative to the Ala/Thr polymorphism, which is indicated with a star.

"arms race" model of peptide evolution. Under this model, antibacterial peptides would be expected to diverge rapidly at the amino acid level as new virulence and resistance mutations arise and fix in the host and pathogen populations. The genes encoding the peptides would also harbor low levels of standing variation. In fact, standing silent variation in peptide loci is not depressed within *D. melanogaster* and amino acid divergence is quite low between Drosophila species. In *metchnikowin* and *drosocin* there are no fixed amino acid replacements between *D. melanogaster* and *D. simulans*, and there are only 14 fixed replacements across all nine genes. This observation can be extended to evolutionarily very distant taxa. For example, substantial amino acid homology is observed between dipteran and lepidopteran attacins (Åsling, Dushay, and Hultmark 1995; Sugiyama et al. 1995) and between cecropins isolated from dipterans, basal chordates, and vertebrates (Lee et al. 1989; Lee, Cho, and Lehrer 1997; Zhao et al. 1997). We therefore find no support for the rapid allelic turnover characteristic of arms races on either a short evolutionary scale or a long one.

Neither do the data support the classical model of selectively maintained hypervariability in mature antibacterial peptides, as the rate of silent substitution is higher than the nonsynonymous rate. Nevertheless, these genes have a high level of amino acid polymorphism relative to interspecific divergence (table 2), with most amino acid variation located in domains that are proteolytically removed to activate the peptide (fig. 1). Several of these polymorphisms are radical, changing charge or polarity at the variable residue. This could be attributable to an absence of purifying selection, making the observed polymorphisms effectively neutral, although in that case, a higher proportion of the amino acid substitutions might be expected to drift to fixation. Alternatively, the segregating amino acid polymorphisms might be slightly deleterious, preventing their fixation, although deleterious mutations are rarely expected to achieve intermediate frequency as several of the peptide polymorphisms do (fig. 1). The nearly significant excess of amino acid polymorphism relative to fixation might be achieved if rare amino acid variants are selectively favored, but only when rare, and if selective advantage is lost as the variant becomes more common. If this is true, the peptide genes could be expected to show other indications of selection. Some evidence of that positive selection affects allele frequencies is provided by the general excess of high-frequency derived mutations and the high degree of linkage disequilibrium in the peptide genes. Potential selection is most clearly illustrated at the *drosocin* locus, where one allele seems to have recently and rapidly increased in frequency.

The genes showing the strongest departure from the equilibrium null model (*drosocin*, *metchnikowin*, and *diptericin A*) are completely unlinked, being distributed across 4.1 Mbp of chromosome 2. The only naturally occurring inversion polymorphism of appreciable frequency in this chromosomal region is In(2R)NS. Both *drosocin* and *metchnikowin* are outside the breakpoints of this inversion, and independent genetic evidence (unpublished data) suggests that the 12 alleles sequenced at these loci are all of the standard arrangement. The cytological arrangements of the 12 *diptericin A* alleles, which are inside In(2R)NS, are not known. It is possible that In(2R)NS polymorphism within the sample could affect estimates of $\hat{C}/4\hat{N}r$ and Z_{nS}, but inversion polymorphism alone is not expected to affect H. We therefore find it unlikely that inversion polymorphism underlies the observed data.

It is important to note that skewing of the site frequency spectrum and departure from linkage equilibrium can have demographic causes as well as selective ones. Przeworski (2002) has shown that an extreme degree of population subdivision with unequal sampling across subpopulations can give a significant departure of H from the neutral expectation, but, as acknowledged by Przeworski, such a model is not likely to represent real Drosophila populations. Our simulations demonstrate that a more plausible model of population bottleneck followed by expansion can also generate values of H reflecting an excess of high-frequency derived polymorphisms. One way of distinguishing demographic from selective effects is by comparison with other loci in the genome. The departure of the pooled peptide data from neutral panmictic expectations is qualitatively similar to, although more extreme than, that of the pooled genome-wide data in terms of linkage disequilibrium and skew in site frequency spectrum. Even this comparison, however, is problematic because of variability in the power to detect departure from the null among loci due to differences in number of alleles surveyed and polymorphic sites observed. Furthermore, far from being randomly chosen, the genome-wide "control" loci are subject to both experimenter and publication bias. This complication is illustrated by the fact that the two loci that drive the marginal combined probability significance of H in the genome-wide data, *vermilion* (Begun and Aquadro 1995) and *white* (Kirby and Stephan 1995), were surveyed in anticipation of detecting natural selection.

Additionally, Fay, Wyckoff, and Wu (2002) have used the Andolfatto and Przeworski (2000) data to argue that natural selection is pervasive in the *D. melanogaster* genome. A convincing distinction between selective and demographic effects would require comparison to polymorphism data from loci sampled throughout the genome without regard to expected selective history, and such a control data set does not currently exist for *D. melanogaster*.

The *cecropin* genes, in particular *cecropin C*, differ noticeably from the remainder of the genes in their comparative lack of both linkage disequilibrium (tables 1 and 4) and skew in the site frequency spectrum towards common-derived variants (table 3). The contrasting patterns of variability may reflect functional differences among the genes. While the rest of the genes are induced by larvae and adults in response to systemic infection, *cecropins B* and *C* are expressed in pupae during metamorphosis, where they may be exposed to less pathogenic or variable bacteria, for instance those residing in the larval gut. This interpretation is consistent with the observation that *cecropins B* and *C* are more similar to each other at the amino acid level than either is to *cecropin A1* and *A2* (fig. 1). The departure of the *cecropins* from the remainder of the antibacterial peptides may be partially attributable to genomic arrangement, as well. *Cecropins A1*, *A2*, and *B* are within a 4-kb segment of chromosome 3R, with *cecropin C* less than 4 kb away. With this in mind, it might be better to consider the *Cecropin* genes as a single superlocus with respect to recombination and *H*. However, inconsistencies in sample size and composition from gene to gene within the *cecropin* cluster (Clark and Wang 1997) preclude their concatenation into a single data set. The treatment of the tightly linked *cecropin* genes as separate loci may be justified by the fact that these genes show the lowest levels of intragenic linkage disequilibrium.

Individual data from *metchnikowin*, *diptericin A*, and *drosocin* and combined data from all of the peptide loci suggest the effects of natural selection in the recent past, although demographic history is also likely to have played a role in the evolution of these genes. We observed virtually no amino acid differentiation between species and little amino acid polymorphism in the mature peptide domains. If selection is acting on these loci it likely acts either on regulatory variants or on the substantial number of nonconservative amino acid polymorphisms in proteolytically processed domains. Selection might favor such polymorphisms if they provide protection against immunomodulatory molecules injected by pathogenic bacteria into the host cell. Bacterial injection of proteins that interfere with host cell signaling pathways and immune responses have been well documented in plants and animals (Hueck 1998; Cornelis and Van Gijsegem 2000; Ernst 2000). Data from an immunity-related Drosophila transcription factor, Relish, conceptually supports the bacterial interference model. Relish proteins have an autoinhibitory domain, which is proteolytically cleaved to activate the transcription factor (Dushay, Åsling, and Hultmark 1996). The amino acids surrounding the site of Relish cleavage have an extremely high rate of amino acid substitution (Begun and Whitley 2000*b*), suggesting that these amino acids may also be an intracellular site of host-pathogen coevolution. The Relish data, however, document rapid fixation of amino acid substitutions, whereas the peptide genes evolve very slowly. Amino acid variation could conceivably be maintained with no increase in the rate of fixation if selection is dependent on the allele frequency of the targeted site. More direct experiments on pathogen-Drosophila biology are obviously required to test this hypothesis.

At present, the data allow the firm rejection of arms races and maintained hypervariablility as appropriate models to describe the evolution of antibiotically active domains of Drosophila antibacterial peptides. But there is suggestive evidence that natural selection may act on these genes, perhaps favoring radical amino acid variability in a frequency-dependent manner and in response to pressure from pathogens. Further research is necessary to test this model and to conclusively separate demographic from selective effects.

Acknowledgments

This work was supported by National Institute of Health Grant AI46402 to A.G.C., National Science Foundation Dissertation Improvement Award DEB0073598 to B.P.L. and A.G.C., and a National Science Foundation Graduate Research Fellowship to B.P.L. B.P.L. is a founding member of the Institute of Drosophila Immunomics in Ithaca, New York.

Literature Cited

Andolfatto, P., and M. Przeworski. 2000. A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156**:257–268.

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.

Begun, D. J., and C. F. Aquadro. 1995. Molecular variation at the vermilion locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. Genetics **140**: 1019–32.

Begun, D. J., and P. Whitley. 2000*a*. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **97**:5960–5965.

———. 2000*b*. Adaptive evolution of Relish, a Drosophila NF-κB/IκB protein. Genetics **154**:1231–1238.

Boman, H. G. 1995. Peptide antibiotics and their role in innate immunity. Annu. Rev. Immunol. **13**:61–92.

Bulet, P., J.-L. Dimarcq, C. Hetru, M. Lagueux, M. Charlet, G. Hegy, and A. Van Dorsselaer. 1993. A novel inducible antibacterial peptide of Drosophila carries an O-glycosylated substitution. J. Biol. Chem. **268**:14893–14897.

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

Carvalho, A. B., and A. G. Clark. 1999. Intron size and natural selection. Nature **401**:343–344.

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

Cornelis, G. R. and F. Van Gijsegem. 2000. Assembly and function of type III secretory systems. Annu. Rev. Microbiol. **54**:735–774.

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.

Date-Ito, A., K. Kasahara, H. Sawai, and S. I. Chigusa. 2002. Rapid evolution of the male-specific antibacterial protein *andropin* gene in Drosophila. J. Mol. Evol. **54**:665–670.

David, J. R., and P. Capy. 1988. Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4**:106–111.

Dawkins, R., and J. R. Krebs. 1979. Arms races between and within species. Proc. R. Soc. Lond. B Biol. Sci. **205**:489–511.

De Gregorio, E., P. T. Spellman, G. M. Rubin, and B. Lemaitre. 2001. Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. Proc. Natl. Acad. Sci. USA **98**:12590–12595.

Dimarcq, J.-L., D. Hoffmann, M. Meister, P. Bulet, R. Lanot, J.-M. Reichhart, and J. A. Hoffmann. 1994. Characterization and transcriptional profiles of a Drosophila gene encoding an insect defensin: a study in insect immunity. Eur. J. Biochem. **221**:201–209.

Dushay, M. S., B. Åsling, B., and D. Hultmark. 1996. Origins of immunity: *Relish*, a compound Rel-like gene in the antibacterial defense of Drosophila. Proc. Natl. Acad. Sci. USA **93**:10343–10347.

Ekengren, S., and D. Hultmark. 1999. *Drosophila cecropin* as an antifungal agent. Insect Biochem. Mol. Biol. **29**:965–972.

Ernst, J. D. 2000. Bacterial inhibition of phagocytosis. Cell. Microbiol. **2**:379–386.

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

Fay, J. C., G. J. Wyckoff, and C.-I Wu. 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature **415**:1024–1026.

Hudson, R. R. 1987. Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50**:245–250.

———. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* D. Futuyma and J. Antonovics, eds. Oxford surveys in evolutionary biology. Oxford University Press, Oxford.

———. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**:337–338.

Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

Hueck, C. J. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. Microbiol. Mol. Biol. Rev. **62**:379–433.

Hughes, A. L., and M. Nei. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. 1988. Nature **335**:167–70.

Irving, P., L. Troxler, T. S. Heuer, M. Belvin, C. Kopczynski, J.-M. Reichhart, J. A. Hoffmann, and C. Hetru. 2001. A genome-wide analysis of immune responses in Drosophila. Proc. Natl. Acad. Sci USA **98**:15119–15124.

Kelly, J. 1997. A test of neutrality based on interlocus associations. Genetics **146**:1197–1206.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kirby, D. A., and W. Stephan. 1995. Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **141**:1483–90.

Kreitman, M. 1983. Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster.* Nature **304**:412–417.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244–1245.

Lazzaro, B. P., and A. G. Clark 2001. Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the *attacin* genes of *Drosophila melanogaster*. Genetics **159**:659–671.

Lee, J.-Y., A. Boman, S. Chuanxin, M. Andersson, H. Jörnvall, V. Mutt, and H. G. Boman. 1989. Antibacterial peptides from pig intestine: isolation of a mammalian cecropin. Proc. Natl. Acad. Sci. USA **86**:9159–9162.

Lee, I. H., Y. Cho, and R. I. Lehrer. 1997. Styelins, broad-spectrum antimicrobial peptides from the solitary tunicate, *Styela clava*. Comp. Biochem. Physiol. **118B**:515–521.

Lehninger, A. L., D. L. Nelson, and M. M. Cox. 1993. Principles of biochemistry. Worth Publishers, New York.

Levashina, E. A., S. Ohresser, P. Bulet, J.-M. Reichhart, C. Hetru, and J. A. Hoffmann. 1995. Metchnikowin, a novel immune-inducible proline-rich peptide from Drosophila with antibacterial and antifungal properties. Eur. J. Biochem. **233**:694–700.

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

Przeworski, M. 2002. The signature of positive selection at randomly chosen loci. Genetics **160**:1179–1189.

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.

Samakovlis, C., D. A. Kimbrell, P. Kylsten, Å. Engström, and D. Hultmark. 1990. The immune response in Drosophila: pattern of *cecropin* expression and biological activity. EMBO J. **9**:2969–2976.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry, 3rd edition. W. H. Freeman and Company, New York.

Sugiyama, M., H. Kuniyoshi, E. Kotani et al. (14 co-authors). 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.

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

Tryselius, Y., C. Samakovlis, D. A. Kimbrell, and D. Hultmark. 1992. *CecC*, a cecropin gene expressed during metamorphosis in Drosophila pupae. Eur. J. Biochem. **204**:395–399.

Wayne, M. L., D. Contamine, and M. Kreitman. 1996. Molecular population genetics of *Ref(2)P*, a locus which confers viral resistance in Drosophila. Mol. Biol. Evol. **13**:191–199.

Wicker, C., J-M. Reichhart, D. Hoffmann, D. Hultmark, C. Samakovlis, and J. A. Hoffmann. 1990. Insect immunity: characterization of a Drosophila cDNA encoding a novel member of the diptericin family of immune peptides. J. Biol. Chem. **265**:22493–22498.

Wiehe T. H., and W. Stephan. 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster.* Mol. Biol. Evol. **10**: 842–854.

Zhao, C., L. Liaw, I. H. Lee, and R. I. Lehrer. 1997. cDNA cloning of three cecropin-like antimicrobial peptides (styelins) from the tunicate, *Styela clava.* FEBS Lett. **412**: 144–148.

Zhou, X., T. Nguyen, and D. A. Kimbrell. 1997. Identification and characterization of the *cecropin* antibacterial protein gene locus in *Drosophila virilis.* J. Mol. Evol. **44**:272–281.

David Rand, Associate Editor

Accepted January 31, 2003