nature genetics

Dynamic evolution of the innate immune system in *Drosophila*

Timothy B Sackton[1], Brian P Lazzaro[2], Todd A Schlenke[3], Jay D Evans[4], Dan Hultmark[5] & Andrew G Clark[1,6]

The availability of complete genome sequence from 12 *Drosophila* species presents the opportunity to examine how natural selection has affected patterns of gene family evolution and sequence divergence among different components of the innate immune system. We have identified orthologs and paralogs of 245 *Drosophila melanogaster* immune-related genes in these recently sequenced genomes. Genes encoding effector proteins, and to a lesser extent genes encoding recognition proteins, are much more likely to vary in copy number across species than genes encoding signaling proteins. Furthermore, we can trace the apparent recent origination of several evolutionarily novel immune-related genes and gene families. Using codon-based likelihood methods, we show that immune-system genes, and especially those encoding recognition proteins, evolve under positive darwinian selection. Positively selected sites within recognition proteins cluster in domains involved in recognition of microorganisms, suggesting that molecular interactions between hosts and pathogens may drive adaptive evolution in the *Drosophila* immune system.

Immune systems must constantly evolve in order to remain effective in the face of both changes in the suite of pathogens to which they are exposed and the evolution of virulence mechanisms. These dynamics can result in a strong signature of adaptive evolution in genes involved in the immune response[1,2]. However, general patterns have been difficult to discern, as most studies have focused on a small number of genes in a few particular species. The recent complete genome sequencing of ten new *Drosophila* species[3], coupled with extensive molecular knowledge of the mechanisms of *Drosophila* immunity, provides an opportunity to dissect the evolutionary history of the annotated *D. melanogaster* immune system across the *Drosophila* genus.

Drosophila mount both cellular and cell-free, or humoral, immune responses to pathogens[4]. The cellular immune response consists of phagocytosis of microbes and of cellular encapsulation and melanization of larger parasites such as parasitoid wasp eggs, by differentiated populations of hemocytes[5]. The humoral immune response is initiated by the recognition of conserved microbe-specific molecules such as peptidoglycan, leading to the activation of signaling cascades and the nuclear translocation of the NF-κB transcription factors Relish, dorsal and DIF, which induce the transcription of antimicrobial peptides (AMPs) and other effectors[6,7]. Although this response depends largely on the Toll and imd pathways[8], other signaling cascades, such as the JAK-STAT and JNK pathways, appear to have supplementary roles[9,10]. Many of these diverse immune responses are analogous to the innate immune responses of mammals, using many of the same components and regulatory pathways, although unlike mammals, insects such as *Drosophila* lack an antibody-mediated adaptive immune response[11,12].

Comparisons among the previously sequenced genomes of the dipterans *D. melanogaster*, *Anopheles gambiae* and *Aedes aegypti* and the hymenopteran *Apis mellifera* have revealed considerable variation in the size and diversity of immune-related gene families[13–15]. Complete genome sequences are now available for 12 species in the genus *Drosophila*: *D. melanogaster*, *Drosophila simulans*, *Drosophila sechellia*, *Drosophila yakuba*, *Drosophila erecta*, *Drosophila ananassae*, *Drosophila persimilis*, *Drosophila pseudoobscura*, *Drosophila willistoni*, *Drosophila virilis*, *Drosophila mojavensis* and *Drosophila grimshawi*[3]. The moderate divergence among these species (40 million years to the most recent common ancestor) provides considerable additional power for studies of molecular evolution, allowing tests for positive selection that are not possible with the much more divergent genomes previously available. Furthermore, the sequenced *Drosophila* species span a wide range of diverse habitats and ecologies, including tropical rain forest species (*D. erecta*, *D. yakuba*), island endemics (*D. sechellia*, *D. grimshawi*), cosmopolitan human commensals (*D. melanogaster*, *D. simulans*) and cactophilic desert species (*D. mojavensis*)[16]. *Drosophila* breed and lay eggs in rotting plant and fungal material, exposing them to a wide range of pathogens in these septic environments, including viruses, bacteria, fungi, protozoans, nematodes and parasitic wasps.

In this study, we annotate orthologs and paralogs of characterized and candidate immune-system genes across the genus *Drosophila*. We analyze patterns of gene family expansion and contraction in all 12 sequenced species and identify the origin of evolutionary novel immune-system genes. Using likelihood-based models of molecular evolution, we test for positive selection across immune-related genes in the *melanogaster*

[1]Field of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA. [2]Department of Entomology, Cornell University, Ithaca, New York 14853, USA. [3]Department of Biology, Emory University, Atlanta, Georgia 30322, USA. [4]US Department of Agriculture–Agricultural Research Service Bee Research Laboratory, Beltsville, Maryland 20705, USA. [5]Umeå Centre for Molecular Pathogenesis, Umeå University, S-901 87 Umeå, Sweden. [6]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA. Correspondence should be addressed to T.B.S. (tbs7@cornell.edu).

Received 25 July; accepted 1 November; published online 7 November 2007; doi:10.1038/ng.2007.60

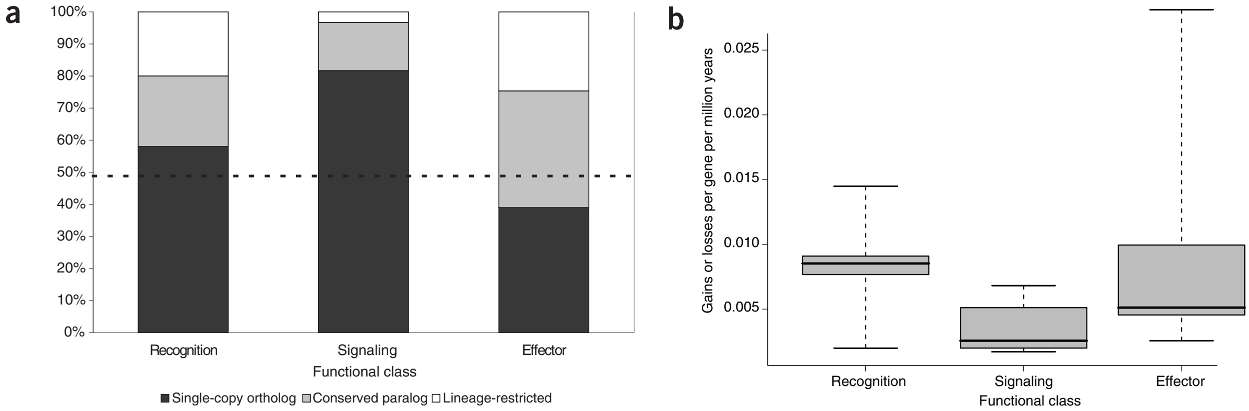

Figure 1 Variation in patterns of homology among immune-system genes. (**a**) Proportion of each functional class assigned to each homology class. The dashed line is the fraction of the entire genome estimated to be in the single-copy ortholog class. (**b**) Box plot of the estimated rate of gene turnover among genes in gene families for each functional class.

group (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*) and identify the protein domains that are the most likely targets of adaptive evolution.

RESULTS

Annotation of immunity proteins in the genus *Drosophila*

We used an initial set of 245 *D. melanogaster* immune-related proteins (**Supplementary Table 1** online) to identify and manually curate 2,501 candidate homologs in the remaining *Drosophila* species studied. For many of our analyses, we grouped genes on the basis of molecular functions: 'recognition genes' that encode pathogen surveillance proteins (for example, peptidoglycan recognition proteins (PGRPs) and phagocytic receptors such as eater); 'signaling genes' that encode proteins in immune-related signaling pathways (for example, Toll and imd); and 'effector genes' that encode proteins that directly inhibit pathogen growth and survival (for example, AMPs).

Any broad functional classification is necessarily subjective to some degree, and some proteins could plausibly be assigned to multiple categories (for instance, some recognition proteins also initiate signal transduction). Furthermore, the molecular functions of many candidate immune-system genes are currently inferred only from sequence similarity, resulting in multiple equally plausible classifications of *D. melanogaster* immune-related proteins. Therefore, we have conducted our analyses using several classification schemes that were modified from the one presented here, either by including only the subset of genes with high-confidence functional annotations or by using alternative functional categories. These modified classification schemes did not substantially change our conclusions (**Supplementary Note**, **Supplementary Table 2** and **Supplementary Figs. 1** and **2** online).

Patterns of gene conservation across 12 *Drosophila* species

To initially assess gene conservation across the 12 sequenced species of *Drosophila*, we assigned homology patterns to one of three classes: 'single-copy ortholog' for genes conserved as single-copy orthologs in all 12 species; 'conserved paralog' for genes that vary in copy number across the phylogeny but that are inferred to have been present in the common ancestor of drosophilids; and 'lineage-restricted' for genes that have arisen since that common ancestor (**Supplementary Table 3** and **Supplementary Methods** online). The proportion of genes in each homology class varied significantly among recognition, signaling and effector classes, with the highest fraction of single-copy orthologs in the signaling class and the lowest in the effector class (χ^2 = 41.13, d.f. = 4, P = 2.53 × 10^{-8}; **Fig. 1a**; **Supplementary Fig. 1**). Furthermore, only the effector class had a deficit of single-copy orthologs relative to the genomic average (**Fig. 1a**). This is not an artifact of the general pattern that short proteins are less likely to be single-copy orthologs, as AMPs (N = 20) had significantly fewer single-copy orthologs than a control set of peptides (N = 2,878) of similar length (χ^2 = 14.92, d.f. = 2, P = 5.76 × 10^{-3}).

The variation in the proportions of genes in each homology class implies variation in the rates of gene duplication and loss among functional classes of immune-system genes. We used a recently developed maximum-likelihood model of birth-death evolution in gene families[17] to estimate λ, the rate of gene turnover (duplications and losses) per million years. The distribution of λ varied among functional classes (Kruskal-Wallis test; P = 0.012), even when only gene families with at least one duplication or loss ($\lambda > 0$) were considered (**Fig. 1b**; Kruskal-Wallis test; P = 0.038).

A prototypical example of the rapid changes in copy number among effector proteins is found in the cecropin gene family, a family of cationic peptides with antimicrobial activity against Gram-positive bacteria, Gram-negative bacteria and fungi. Cecropin homologs have been identified in all major endopterygote insect orders except *Hymenoptera*, and in many cases they seem to be organized in a single genomic cluster[18]. As expected, we find a syntenically conserved cecropin cluster in all 12 *Drosophila* species we studied here (**Supplementary Fig. 3** online). There seem to have been at least four independent expansions of this cluster within the *Sophophora* subgenus of drosophilids, three within the *Drosophila* subgenus, and at least two independent losses within the *melanogaster* group (**Supplementary Fig. 3**). In principle, paralogous gene conversion can create a phylogenetic pattern similar to that expected from gene duplication and deletion. However, previous studies have found no evidence of gene conversion among cecropin genes in *D. melanogaster*[19,20], and changes in gene order and orientation among species suggest rapid turnover, not gene conversion (**Supplementary Fig. 3**). This pattern of rapid gene turnover with many independent expansions is common, if in less extreme form, in other effector and recognition gene families, in sharp contrast to signaling genes, whose rates of gene duplication are markedly lower.

Evolutionary novelties in the *Drosophila* immune system

Comparisons among mosquitoes, fruit flies and honeybees have identified lineage-specific genes encoding both recognition and effector proteins, suggesting the emergence of evolutionary novelties in the insect immune system[13–15]. Based on the phylogenetic pattern of gene presence and absence in lineage-restricted gene families within *Drosophila*, we found evidence for the emergence of evolutionary novelties among recognition and effector gene families over roughly an order of magnitude shorter time scales. In contrast, the complement of signaling proteins in the immune system appears to be quite stable over the 40 million years of evolution since the root of the *Drosophila* genus, consistent with observations from more distant comparisons within insects[13–15].

Although the complement of proteins (such as PGRPs) that recognize microbe-specific molecules was essentially constant throughout the genus *Drosophila* (**Supplementary Table 3**), this was not the case for gene families thought to encode phagocytosis receptors. Of particular interest is the family that includes the genes encoding eater and nimrod, putative phagocytosis receptors characterized by a unique type of EGF-like repeat, the NIM repeat[21,22]. Members of this family, particularly *eater* and *nimrod C1* (*nimC1*), have independently expanded in several species (**Fig. 2a**).

The *Hemese (He)* gene is located within the *nimrod* cluster and is also expressed in the hemocyte plasma membrane, but lacks NIM repeats and instead has a short serine-and threonine-rich *O*-glycosylated extracellular domain[23]. *He* homologs were not detectable outside the *melanogaster* group (**Fig. 2a**), although one *nimC1* paralog in *D. willistoni*, *nimC1a*, has a similarly serine- and threonine-rich region and a reduced number of NIM repeats (**Fig. 2b**). A likely model is that the *He* gene originated from a truncated *nimC1* paralog that has lost all NIM repeats; *nimC1a* in *D. willistoni* may therefore represent a potential *He* analog. Within the *nimrod* family, *He* is not the only apparent evolutionary novelty: the *nimrod D* subfamily appears to be restricted to the *Sophophora* subgenus and the *nimrod E* subfamily to the *D. virilis–D. mojavensis* clade (**Fig. 2a**). Notably, the class C scavenger receptors (SR-Cs) in the *melanogaster* subgroup (a family of proteins related to SR-CI, a scavenger receptor expressed in the hemocyte plasma membrane and implicated in the phagocytosis of bacteria[24]) also seem to have diversified through partial or truncated duplications (see **Supplementary Note**). Novel genes often arise from rearrangements, truncations and fusions of existing genes[25], and the fixation of these novel genes may be a common mechanism to generate diversity in recognition proteins.

Apparent evolutionary novelties also exist in the effector class. The most notable example was the seven-member drosomycin antifungal peptide family, although we saw a similar pattern for less-well-characterized effector protein families, such as the Turandots (**Supplementary Note** and **Supplementary Fig. 4** online). Homologs of drosomycin have previously been identified within the *melanogaster* and *ananassae* subgroups[26] and in *Drosophila triauraria*, a member of the closely-related *montium* subgroup[27]. The genomic arrangement of the drosomycin family is conserved, and each drosomycin ortholog is monophyletic in the *melanogaster* subgroup; several rearrangements disrupt the drosomycin cluster in *D. ananassae*, suggesting independent expansion within the *ananassae* and *melanogaster* subgroups (**Supplementary Fig. 5** online). Despite strong conservation of amino acid sequence among the *D. melanogaster* drosomycins, BLAST searches against the newly sequenced *Drosophila* genomes (and other previously sequenced insect genomes) failed to identify putative homologs more distant than *D. ananassae*.

Unexpectedly, we found drosomycin-like sequences in EST databases from three different coleopteran species (for example, CB377292, DV767586 and CV160723 from dbEST). Given the lack of drosomycins in any completely sequenced non-*Drosophila* insect genome, it is possible that these beetle ESTs represent contaminants or microbial products. However, if these are actually genuine beetle drosomycins, we suggest at least three possible explanations: drosomycins have been independently introduced in the *Drosophila* and/or coleopteran lineages by horizontal gene transfer (perhaps by *Wolbachia*[28]); drosomycins have been lost independently in most flies and several other insect orders; or

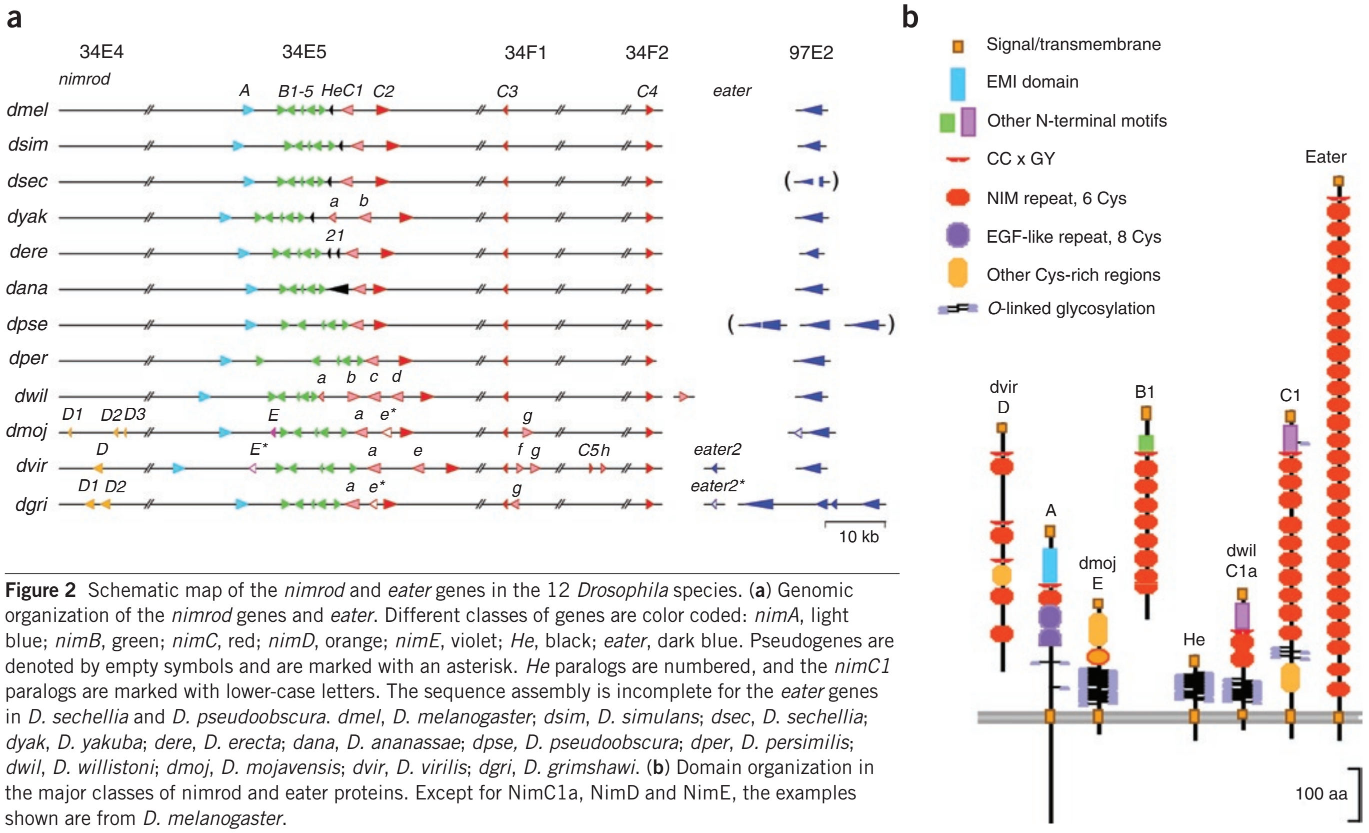

Figure 2 Schematic map of the *nimrod* and *eater* genes in the 12 *Drosophila* species. (**a**) Genomic organization of the *nimrod* genes and *eater*. Different classes of genes are color coded: *nimA*, light blue; *nimB*, green; *nimC*, red; *nimD*, orange; *nimE*, violet; *He*, black; *eater*, dark blue. Pseudogenes are denoted by empty symbols and are marked with an asterisk. *He* paralogs are numbered, and the *nimC1* paralogs are marked with lower-case letters. The sequence assembly is incomplete for the *eater* genes in *D. sechellia* and *D. pseudoobscura*. *dmel*, *D. melanogaster*; *dsim*, *D. simulans*; *dsec*, *D. sechellia*; *dyak*, *D. yakuba*; *dere*, *D. erecta*; *dana*, *D. ananassae*; *dpse*, *D. pseudoobscura*; *dper*, *D. persimilis*; *dwil*, *D. willistoni*; *dmoj*, *D. mojavensis*; *dvir*, *D. virilis*; *dgri*, *D. grimshawi*. (**b**) Domain organization in the major classes of nimrod and eater proteins. Except for NimC1a, NimD and NimE, the examples shown are from *D. melanogaster*.

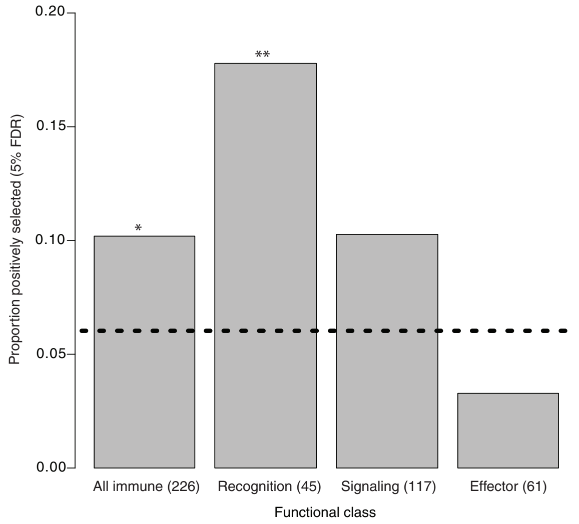

Figure 3 Variation in positive selection among immune-system genes. Proportion of positively selected genes (at a 5% false-discovery rate) within the immune system as a whole and in each functional class. Numbers in parentheses indicate the number of genes in each class. The dotted line is the estimate of the fraction of positively selected genes (at a 5% false-discovery rate) among all single-copy orthologs, from ref. 3. Asterisks indicate a significant difference from the genomic fraction (FET: *, $0.01 < P < 0.05$; **, $P < 0.01$). Among functional classes, the proportion of positively selected genes varies significantly ($P = 0.046$, χ^2-test).

drosomycins have arisen at least twice by convergent evolution, perhaps from a defensin-like precursor.

These instances of apparent lineage-specific gains of known and putative immune-related proteins in *Drosophila*, combined with the known diversity of AMPs across insects[14,29], suggest that there remain further immune components to discover in *Drosophila* species outside the *melanogaster* group. Although the core immune signaling pathways are deeply conserved as single-copy orthologs, there seems to be considerable flexibility in the inputs and outputs of the system that allows new components to be integrated into the immune response over evolutionarily short time scales.

Patterns of positive selection in innate immune-system genes

We used codon substitution models of molecular evolution, implemented in the software package PAML[30], to estimate ω (d_N/d_S, the relative rate of nonsynonymous to synonymous substitution) and infer patterns of positive selection (**Supplementary Table 4** online). These models require accurate nucleotide alignments and become less reliable at high synonymous divergence, limiting our analysis to the six species in the *melanogaster* group. To test for positive selection, we compared the likelihood of the data under a model that requires a subset of codons to have $\omega > 1$ (a pattern predicted only when adaptive fixations have occurred) to the likelihood of the data under a model that does not allow such codons[31]. Any gene for which the null model is rejected has some number of codons that have experienced significantly more nonsynonymous substitutions across the tree than expected. We used a false-discovery rate (FDR) of 5%, unless otherwise noted, to correct for multiple testing[32].

Immune-system genes evolve more rapidly than other genes

We compared ω estimated under the simplest model (a single ω per gene) between the immune-system genes in this study and all single-copy orthologs in the *Drosophila* protein-coding genome[3]. Immune-system genes were significantly less conserved than the set of all single-copy orthologs in the *melanogaster* group (immune-system genes: median $\omega = 0.080$, $N = 226$; all single-copy orthologs: median $\omega = 0.064$, $N = 8{,}510$; $P = 1.43 \times 10^{-5}$, Mann-Whitney *U*-test). This pattern did not appear to be the result of biases introduced by the manual curation of immunity genes compared to the computational curation of the whole-genome dataset, as the results were qualitatively identical when only computationally curated immunity gene models were included (see **Supplementary Note**). This elevated ω in immune-system genes is likely to be driven by adaptive evolution, as 514 of 8,510 single-copy orthologs in the *melanogaster* group (6.0%) showed evidence for positive selection after multiple-test correction[3], compared to 23 of 226 immunity genes (10.2%), a difference that is significant by Fisher's exact test (FET) ($P = 0.016$; **Fig. 3**). The strength of this effect depended slightly on what genes were classified as 'immunity' genes for this analysis (**Supplementary Note** and **Supplementary Fig. 2**).

Among immune-system genes, the proportion that are positively selected differed between the recognition, effector and signaling classes (**Fig. 3** and **Supplementary Fig. 2**). Compared to the genomic single-copy

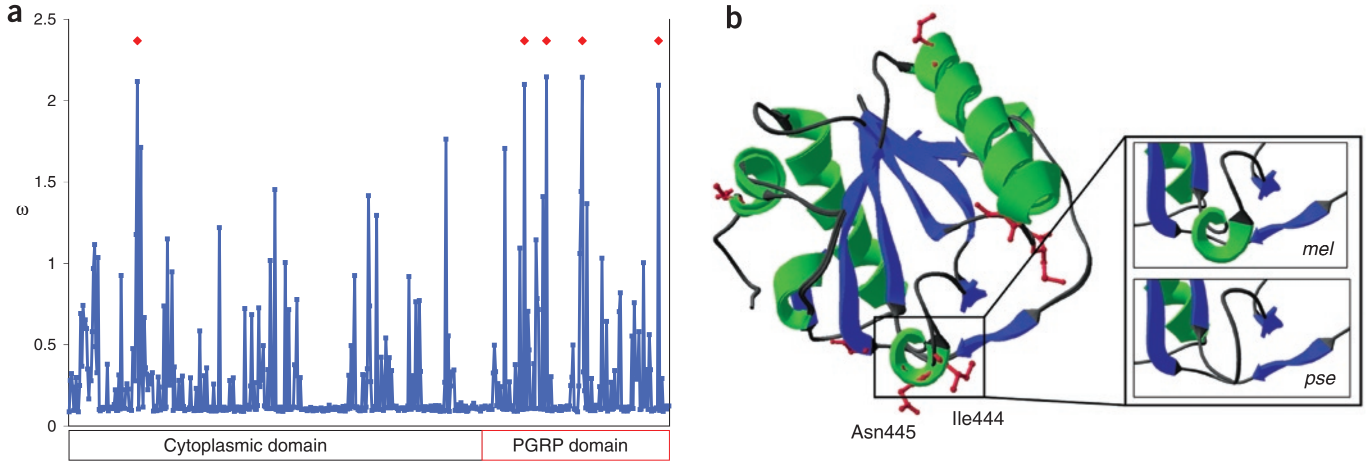

Figure 4 Positive selection in PGRP-LCa. (**a**) Variation in ω among codons of PGRP-LCa. The domain structure of the protein is represented on the *x* axis. Red diamonds mark codons with a Bayesian posterior probability of positive selection >0.75. No such positively selected sites are found in the alternative exons encoding PGRP-LCx and –LCy domains. (**b**) Structural model of the PGRP-LCa domain from *D. melanogaster* (*mel*) and *D. pseudoobscura* (*pse*). The *D. pseudoobscura* sequence is threaded onto the *D. melanogaster* structure using Swiss-PDB Viewer with the default settings. Both structures are color-coded by secondary structure. Side chains of positively selected sites (posterior probability >0.50) are shown in magenta, with the structure-altering residues Asn444 and Ile445 labeled. The insert shows the primary divergent region between the two structures.

Table 1 Distribution of positively selected sites among genes encoding recognition proteins

Gene	Domain encoding	Positively selected sites[a]	Total sites	*P* value[b]
PGRP-LCa	PGRP	4	166	
	Rest of gene	1	354	0.0379
Eater	N-terminal	2	199	
	Rest of gene	2	650	0.2354
NimC1	N-terminal	7	199	
	Rest of gene	2	429	0.0057
Receptor genes, pooled	'Pathogen interaction domain'	13	564	
	Rest of gene	5	1,433	0.0001
TepI	Hypervariable	8	63	
	Rest of gene	78	1,298	0.0552
TepII	Hypervariable[c]	6	200	
	Rest of gene	21	1,350	0.1473
TepIV	Hypervariable	4	47	
	Rest of gene	9	1,445	0.0005
Tep genes, pooled	Hypervariable	18	310	
	Rest of gene	108	4,093	0.0038
All pooled	'Pathogen interaction domain'	31	874	
	Rest of gene	113	5,526	0.0093

[a] Any site with a Bayesian posterior probability of positive selection >0.75 is considered a 'positively selected site'. [b] Calculated by Fisher's exact test. [c] Includes all splice forms.

ortholog dataset, genes that encode recognition proteins were significantly more likely to show evidence of positive selection (17.8% versus 6.0%, $P = 0.005$, FET), a result that was robust with respect to a number of different alternative classifications of immune-system genes (see **Supplementary Note**). Genes that encode signaling proteins trended toward excess positive selection relative to the genomic set (10.3% versus 6.0%, $P = 0.076$, FET), and genes that encode effector proteins were less likely, although not significantly so, to be in the positive selection class (3.3% versus 6.0%, $P = 0.585$, FET). Although signaling proteins likely have more pleiotropic nonimmune functions than recognition or effector proteins, differences in the degree of immune specificity among functional classes did not seem to account for the variation in positive selection that we observe (see **Supplementary Note**).

Positive selection drives the evolution of recognition proteins

Notably, of the ten recognition genes ascertained to be evolving under positive selection (with a 10% FDR), two encode proteins that have been directly shown to participate in phagocytosis of foreign microorganisms (*NimC1* (ref. 22), *TepII*[33]), and seven of the remaining eight are homologous to genes encoding proteins involved in phagocytosis in *Drosophila* or mammals (TEPs: *TepI*, *TepIV*; nimrods: *NimB1*, *NimB4*; CD36 homologs: *crq*, *CG31217*, *emp*). Furthermore, two additional experimentally identified genes encoding phagocytic receptors (*eater*[21] and *peste* (*pes*)[34]) showed some evidence for positive selection (*eater* nominal $P = 0.019$; *pes* nominal $P = 0.019$; **Supplementary Table 4**). In contrast, among the genes encoding PGRPs or Gram-negative binding proteins (GNBPs), only *PGRP-LC* and *PGRP-LB* showed any evidence for positive selection. This excess of positive selection in putative phagocytosis genes was significant ($P = 0.034$; FET). One possible hypothesis to explain this difference is that the molecules recognized by PGRPs and GNBPs (peptidoglycan and β-glucan) are evolutionarily static and thus unlikely to trigger coevolutionary arms races. In contrast, the targets of phagocytosis receptors may be more variable in structure and thus be more likely to lead to bouts of host-pathogen coevolution.

Of the 14 recognition genes showing evidence for positive selection at a nominal α of 0.05, we have reasonable hypotheses regarding what protein domain might interact directly with pathogens for six of them (*TepI*, *TepII*, *TepIV*, *NimC1*, *eater* and *PGRP-LC*; **Supplementary Fig. 6** online). The TEP proteins are members of the α2-macroglobin superfamily, and they contain a hypervariable region that is likely important for interactions with pathogens[35]. Nimrod C1 and eater are both type I membrane proteins characterized by a large number of NIM repeats and a more divergent N-terminal region. The 200 N-terminal amino acids have been experimentally determined to be sufficient for bacterial binding in eater[21] and are likely to have a similar function in NimC1. Both molecular and structural data show that the PGRP domain is required for binding to peptidoglycan in PGRP-LC[36,37]. Using Bayesian estimates of the probability of positive selection for each codon in these six proteins[38], we found that codons encoding residues in these 'pathogen interaction domains' were significantly more likely to evolve by positive selection than codons encoding residues outside these domains (**Table 1**), suggesting that adaptive evolution of these *Drosophila* recognition proteins is driven by interactions with pathogen-associated molecules.

One of these recognition proteins, PGRP-LC, is alternatively spliced in *D. melanogaster* to produce three isoforms with different PGRP domains attached to the same cytoplasmic domain[39]. All three splice forms (PGRP-LCa, PGRP-LCx and PGRP-LCy) were conserved in all 12 species, although we found evidence for positive selection only in the PGRP-LCa isoform, particularly in the PGRP domain (**Table 1; Fig. 4a**). Two of the putative positively selected sites in PGRP-LCa (Ile444 and Asn445) are part of an insertion (relative to PGRP-LCx and PGRP-LCy) that induces a structural change resulting in altered binding properties[36,37]. This two-amino-acid insertion is present in the five species of the *melanogaster* subgroup but not in any more distant species (**Fig. 4b**). It thus appears that the structural conformation induced by this insertion is evolutionary recent and that selection may have acted to fine-tune the modified structure for improved stability, binding affinity or some similar property.

Rapid divergence of signal modulation proteins

Although signaling genes overall showed only a nonsignificant trend toward excess positive selection relative to genomic averages, a different pattern emerged when the signaling class was divided into genes encoding proteins with a modulation function and genes encoding proteins with a signal transduction function. Six out of 26 modulation proteins showed evidence for positive selection (23.8%), a significantly greater proportion than either the genomic average (6.0%; $P = 0.004$, FET) or signal transduction proteins (6.4%; $P = 0.0217$, FET). Modulation proteins also had a very different pattern of copy number conservation: 51.7% were found as single-copy orthologs, as compared to 88.4% of signal transduction proteins. These differences may result from modulation proteins occupying less central, and therefore less constrained, positions in innate immune signaling networks.

Positive selection in the Relish cleavage complex

Previous work has suggested that signaling proteins, and particularly genes in the imd pathway, evolve by positive selection in the *D. simulans* lineage[2,40]. Although the median *P* value for the test of positive selection was marginally lower for genes in the imd pathway than for other signaling genes (imd median: 0.1376; other signaling median: 0.2954, $P = 0.050$, Mann-Whitney *U*-test, one-tailed), neither signaling genes as a group nor genes in the imd pathway alone were over-represented among positively selected genes in this study. This discrepancy would be expected if imd pathway genes experience positive selection in only a subset of the species examined. We tested this hypothesis by fitting codon models that allow for lineage-specific variation in ω to test for an acceleration of protein evolution along a particular lineages in the phylogeny[41], and by fitting codon models that test explicitly for positive selection that is restricted to particular branches in the phylogeny[42].

A number of genes in the imd pathway (*Relish* (*Rel*), *ird5*, *key* and *Dredd*) showed evidence for accelerated rates of evolution specifically in the *D. melanogaster* lineage at a nominal $P < 0.01$ (**Supplementary Fig. 7** online); *Rel* was also accelerated in the ancestral *simulans-sechellia* lineage. Of those four genes, *Rel* and *ird5* also had a subset of codons with $\omega > 1$ specifically in the *melanogaster* lineage; *Dredd*, *Dnr1* and *ird5* showed evidence for positive selection in the entire phylogeny (**Supplementary Table 4**). *BG4* showed evidence for positive selection in both the *melanogaster* and *simulans-sechellia* lineages ($P = 0.019$ and $P = 0.048$, respectively), although not in the whole phylogeny. Taken together, these results suggest that a substantial fraction of genes in the imd pathway have experienced positive selection in the *melanogaster* species group, selection that has occurred since the divergence of these species from *D. yakuba* and *D. erecta*; this is consistent with recent results obtained for *Rel* using a different methodology[43].

Many of the positively selected proteins in the imd pathway are thought to physically interact. Relish is cleaved at a caspase cleavage site located in the spacer region between the N-terminal REL homology domain (encoding the functional transcription factor) and the C-terminal ANK repeat region (encoding an autoinhibitory domain)[44]. Cleavage requires phosphorylation of Relish by the kinase ird5, and also requires Dredd, a caspase that forms a complex with Relish[44]. Thus, molecular coevolution may drive positive selection in the spacer region of Relish, the caspase domain of Dredd and the kinase domain of ird5. Pooling across all putatively interacting domains, positively selected sites significantly cluster inside the interacting domains (**Supplementary Table 5** and **Supplementary Fig. 6** online), suggesting that, at least in *D. melanogaster*, the entire complex is evolving by positive selection. The apparent restriction of positive selection in at least some of these genes to the *melanogaster* species group suggests that it may stem from a taxon-specific host-pathogen interaction.

DISCUSSION

A number of the genes we identified here as positively selected also evolve adaptively in other organisms, suggesting that widely disparate taxa may yet reveal similarities in the evolution of the immune system and raising the tantalizing possibility that certain kinds of immune proteins may generally be involved in host-pathogen 'arms races'. The most notable example of such commonalities is in the TEP proteins: analysis of fragments of thioester-containing proteins have suggested that these have undergone adaptive evolution in anopheline mosquitoes[45] and the cladoceran crustacean *Daphnia*[46], suggesting that the TEP superfamily is commonly the target of positive selection in arthropods and motivating further study in mammals of TEP, α2-macroglobin and complement superfamily proteins.

We and others[2,40,43] have also identified *Rel* and its protein interactors as targets of positive selection in *Drosophila*, apparently in only a subset of lineages. In an interesting parallel, positively selected codons have also been detected in the linker, PEST domain and caspase cleavage site of *Rel* in termites of the genus *Nasutitermes*, with a similar clade-restricted pattern[47], suggesting that *Rel* may commonly be involved in taxon-specific host-pathogen interactions.

Neither our study, nor any previous study[19,26,48], has found any evidence for adaptive evolution among AMPs in *Drosophila*. In contrast, AMPs in frogs, termites and mammals have all been shown to evolve both by rapid gene duplication and by positive selection[49]. Although we see extensive gene duplication and high rates of gene turnover in *Drosophila*, the lack of positive selection is puzzling. AMPs in *Drosophila* and other insects, in contrast to organisms with adaptive immune systems, serve as the primary microbial- and fungal-killing proteins and may be particularly important in preventing infection by non-coevolving saprophytic organisms, as opposed to more specific pathogens that would be expected to drive rapid coevolutionary arms races. Furthermore, in the *Drosophila* immune response, a large number of different AMPs are induced to high systemic levels after infection. These two factors may lead to stronger selection for speed and efficiency of transcription and translation of AMPs after infection, as opposed to modifications of the protein sequence by positive selection.

The intersection of comparative genomics and molecular evolution provides fertile ground to explore the evolution of innate immune pathways along a multispecies phylogeny. Although considerable attention has been focused on the evolutionary dynamics of components of the adaptive immune system in vertebrates, the results presented here suggest that the *Drosophila* innate immune system experiences similar selective pressures, driven in similar ways by host-pathogen coevolutionary dynamics. Specifically, we find that proteins involved in pathogen recognition, and in particular regions of these proteins that interact with pathogens, undergo significantly more positive selection than other components of the innate immune system, reminiscent of classic examples of adaptive evolution in vertebrate immunity. Further work will be needed to assess the generality of the patterns we observe in taxa that have both adaptive and innate immune systems. Nonetheless, the deep genetic resources of *Drosophila* provide a unique opportunity to further understand the functional divergence of innate immune pathways.

METHODS

Annotation of immune-related proteins in *D. melanogaster*. We generated an initial list of immune-system proteins in *D. melanogaster* from recent reviews, FlyBase annotations and the published literature, including any protein for which there is direct molecular evidence for an immune role in *D. melanogaster* as well as proteins homologous to known immune proteins of *D. melanogaster* or other organisms. The full list of immune-system genes included in this study is presented as **Supplementary Table 1**; further details of our functional classifications are described in the **Supplementary Methods**. All analyses, unless otherwise noted, used this manually curated gene set.

Initial dataset and alignments. Our annotation of homologs of immune-system genes in non-*melanogaster* species started with predicted GLEANR models and homology clusters derived from the computational analysis described in ref. 3 and the **Supplementary Methods**. We then improved these annotations manually. In many cases, we were able to extend partial computationally defined gene models (although this was not always possible, as occasionally assembly gaps prevented the extension of gene models), merge single models inappropriately split into multiple models, eliminate erroneous paralogy calls introduced by assembly duplications, and find homologs not identified by the computational pipeline. In some cases, we corrected erroneous frameshift or nonsense mutations by examining raw sequence traces in the NCBI trace archive.

We derived initial homology assignments from the fuzzy reciprocal BLAST homology clusters described in ref. 3, and we refined them by manual annotation as described in the **Supplementary Methods**. For cases in which we judged GLEANR models to be correct, and no paralogs were identified, we used the alignments produced by ref. 3 for all subsequent analyses. In all other cases, we used alignments produced by T-COFFEE and manually edited. We then masked these alignments as described in ref. 3 (and the **Supplementary Methods**) before molecular evolutionary analysis. Gene models and alignments used in this study are available upon request from T.B.S.

Gene family evolution. Many of the genes involved in innate immunity are organized in clusters of related genes having similar function, which can expand or contract in number across the *Drosophila* phylogeny. Phylogenetic hypotheses were used to assign paralogy and orthology within these genes families, as described in the **Supplementary Methods**.

We used birth-death models, which assume that gene families evolve by duplication to create new gene copies with some birth rate and by pseudogenization and loss of existing gene copies with some death rate, to test for variation in rates of gene turnover across families. Using the EM algorithm implemented by the CAFE software[50], we estimated λ, the rate of gene copy turnover per million years, for each immune-system gene family in our dataset, assuming the time to the most recent common ancestor of drosophilids is 40 million years. Gene families with no copy number variation are assumed to have $\lambda = 0$. CAFE assumes a model of gene family evolution with a single constant rate for both gene duplication and gene loss that is homogeneous across the phylogeny. We also used this analysis to test for nonhomogeneity of the birth-death process, although we did not reject nonhomogeneity for any gene family in our dataset after multiple-test correction.

PAML analysis. All PAML analyses were carried out with PAML version 3.15 on the *melanogaster* group alignments, as described below and in the **Supplementary Methods**. For all alignments, we ran PAML model M0, M7 and M8. Model M0 assumes a single ω for each gene, whereas M7 and M8 allow ω to vary among codons in a gene. In general, we used per-gene estimates of ω from M0 unless otherwise noted, and we used more complicated models primarily to test for evidence for positive selection. M7 assumes that ω follows a beta(0,1) distribution, with shape parameters estimated by maximum likelihood. M8 makes the same initial assumption, and adds a class of codons with $\omega \geq 1$. Our test for positive selection was a comparison of twice the difference in likelihoods between model M7, which does not allow for positive selection, and M8, which does. We estimated *P* values by simulation under the null model, as described in the **Supplementary Methods**. We corrected for multiple testing using two different false-discovery-rate approaches, as described in the **Supplementary Methods**.

For the genes in the imd pathway, we also analyzed a series of branch models that allow ω to vary among branches[41]: one in which the *melanogaster* terminal lineage has one ω and the rest of the tree has another, and one in which the *simulans* and *sechellia* lineages have one ω and the rest of the tree has another. These models test for changes in constraint along a particular branch, not for positive selection *per se*. To explicitly test for positive selection, we used a branch-site model[42], which allows four classes of codons: a strictly conserved class ($\omega < 1$), a class that is conserved in the 'background' lineages but under positive selection in the 'foreground' lineage of interest, a class that is strictly neutral ($\omega = 1$), and a class that is neutral on the 'background' lineages but under positive selection in the 'foreground' lineage of interest. When compared to the null model, which does not allow positive selection in the foreground lineage, this model provides a robust test for positive selection in a subset of codons on a particular lineage[42]. We applied this branch-site model to two sets of foreground lineages: the *melanogaster* terminal branch and the *simulans-sechellia* clade. For both the branch test and the branch-site test, significance was assessed using standard asymptotic assumptions, because both tests are well behaved[42].

We used the Bayesian empirical Bayes approach implemented in PAML model M8 and the PAML branch-site models to estimate the probabilities of positive selection for specific codons[38], as described in the **Supplementary Methods**.

Statistical analysis. All statistical analyses were carried out in R (version 2.4.1), with the exception of some permutation tests, which we implemented with custom Perl scripts.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (A.G.C. and B.P.L.); the National Sciences Foundation (B.P.L.); and the Swedish Foundation for Strategic Research, the Wallenberg Consortium North and the Swedish Research Council (D.H.). T.B.S. is a Howard Hughes Predoctoral Fellow.

AUTHOR CONTRIBUTIONS

T.B.S. and A.G.C. designed this study; T.B.S., B.P.L., T.A.S., J.D.E., D.H., and A.G.C. generated the data and analyzed the results; T.B.S. wrote this paper; B.P.L., T.A.S., J.D.E., D.H., and A.G.C. contributed to the writing and editing of this paper.

Published online at http://www.nature.com/naturegenetics

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Hughes, A.L. & Nei, M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
2. Schlenke, T.A. & Begun, D.J. Natural selection drives *Drosophila* immune system evolution. *Genetics* **164**, 1471–1480 (2003).
3. *Drosophila* 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* doi:10.1038/06341 (published online 8 November 2007).
4. Lemaitre, B. & Hoffmann, J. The host defense of *Drosophila melanogaster. Annu. Rev. Immunol.* **25**, 697–743 (2007).
5. Meister, M. & Lagueux, M. *Drosophila* blood cells. *Cell. Microbiol.* **5**, 573–580 (2003).
6. Steiner, H. Peptidoglycan recognition proteins: on and off switches for innate immunity. *Immunol. Rev.* **198**, 83–96 (2004).
7. Hultmark, D. *Drosophila* immunity: paths and patterns. *Curr. Opin. Immunol.* **15**, 12–19 (2003).
8. De Gregorio, E., Spellman, P.T., Tzou, P., Rubin, G.M. & Lemaitre, B. The Toll and Imd pathways are the major regulators of the immune response in *Drosophila. EMBO J.* **21**, 2568–2579 (2002).
9. Boutros, M., Agaisse, H. & Perrimon, N. Sequential activation of signaling pathways during innate immune responses in *Drosophila. Dev. Cell* **3**, 711–722 (2002).
10. Agaisse, H. & Perrimon, N. The roles of JAK/STAT signaling in *Drosophila* immune responses. *Immunol. Rev.* **198**, 72–82 (2004).
11. Silverman, N. & Maniatis, T. NF-kappaB signaling pathways in mammalian and insect innate immunity. *Genes Dev.* **15**, 2321–2342 (2001).
12. Evans, C.J., Hartenstein, V. & Banerjee, U. Thicker than blood: conserved mechanisms in *Drosophila* and vertebrate hematopoiesis. *Dev. Cell* **5**, 673–690 (2003).
13. Christophides, G.K. *et al.* Immunity-related genes and gene families in Anopheles gambiae. *Science* **298**, 159–165 (2002).
14. Evans, J.D. *et al.* Immune pathways and defence mechanisms in honey bees *Apis mellifera. Insect Mol. Biol.* **15**, 645–656 (2006).
15. Waterhouse, R.M. *et al.* Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* **316**, 1738–1743 (2007).
16. Markow, T.A. & O'Grady, P.M. *Drosophila* biology in the genomic age. *Genetics* doi:10.1534/genetics.107.074112 (in the press).
17. Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15**, 1153–1160 (2005).
18. Hultmark, D. Immune reactions in *Drosophila* and other insects: a model for innate immunity. *Trends Genet.* **9**, 178–183 (1993).
19. Clark, A.G. & Wang, L. Molecular population genetics of *Drosophila* immune system genes. *Genetics* **147**, 713–724 (1997).
20. Ramos-Onsins, S. & Aguade, M. Molecular evolution of the Cecropin multigene family in *Drosophila.* functional genes vs. pseudogenes. *Genetics* **150**, 157–171 (1998).
21. Kocks, C. *et al.* Eater, a transmembrane protein mediating phagocytosis of bacterial pathogens in *Drosophila. Cell* **123**, 335–346 (2005).
22. Kurucz, E. *et al.* Nimrod, a putative phagocytosis receptor with EGF repeats in *Drosophila* plasmatocytes. *Curr. Biol.* **17**, 649–654 (2007)
23. Kurucz, E. *et al.* Hemese, a hemocyte-specific transmembrane protein, affects the cellular immune response in *Drosophila. Proc. Natl. Acad. Sci. USA* **100**, 2622–2627 (2003).
24. Ramet, M. *et al. Drosophila* scavenger receptor CI is a pattern recognition receptor for bacteria. *Immunity* **15**, 1027–1038 (2001).
25. Long, M., Betran, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**, 865–875 (2003).
26. Jiggins, F.M. & Kim, K.W. The evolution of antifungal peptides in *Drosophila. Genetics* **171**, 1847–1859 (2005).
27. Daibo, S., Kimura, M.T. & Goto, S.G. Upregulation of genes belonging to the drosomycin family in diapausing adults of *Drosophila* triauraria. *Gene* **278**, 177–184 (2001).
28. Hotopp, J.C. *et al.* Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753–1756 (2007).
29. Bulet, P., Hetru, C., Dimarcq, J.L. & Hoffmann, D. Antimicrobial peptides in insects; structure and function. *Dev. Comp. Immunol.* **23**, 329–344 (1999).
30. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
31. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A.M. Codon-substitution models for het-

erogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).
32. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
33. Stroschein-Stevenson, S.L., Foley, E., O'Farrell, P.H. & Johnson, A.D. Identification of *Drosophila* gene products required for phagocytosis of *Candida albicans*. *PLoS Biol.* **4**, e4 (2006).
34. Philips, J.A., Rubin, E.J. & Perrimon, N. *Drosophila* RNAi screen reveals CD36 family member required for mycobacterial infection. *Science* **309**, 1251–1253 (2005).
35. Blandin, S. & Levashina, E.A. Thioester-containing proteins and insect immunity. *Mol. Immunol.* **40**, 903–908 (2004).
36. Chang, C.I. *et al.* Structure of the ectodomain of *Drosophila* peptidoglycan-recognition protein LCa suggests a molecular mechanism for pattern recognition. *Proc. Natl. Acad. Sci. USA* **102**, 10279–10284 (2005).
37. Chang, C.I., Chelliah, Y., Borek, D., Mengin-Lecreulx, D. & Deisenhofer, J. Structure of tracheal cytotoxin in complex with a heterodimeric pattern-recognition receptor. *Science* **311**, 1761–1764 (2006).
38. Yang, Z., Wong, W.S. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118 (2005).
39. Werner, T., Borge-Renberg, K., Mellroth, P., Steiner, H. & Hultmark, D. Functional diversity of the *Drosophila* PGRP-LC gene cluster in the response to lipopolysaccharide and peptidoglycan. *J. Biol. Chem.* **278**, 26319–26322 (2003).
40. Begun, D.J. & Whitley, P. Adaptive evolution of relish, a *Drosophila* NF-kappaB/IkappaB protein. *Genetics* **154**, 1231–1238 (2000).
41. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998).
42. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479 (2005).
43. Levine, M.T. & Begun, D.J. Comparative population genetics of the immunity gene, Relish: is adaptive evolution idiosyncratic? *PLoS ONE* **2**, e442 (2007).
44. Stoven, S. *et al.* Caspase-mediated processing of the *Drosophila* NF-kappaB factor Relish. *Proc. Natl. Acad. Sci. USA* **100**, 5991–5996 (2003).
45. Little, T.J. & Cobbe, N. The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioester-recognizing protein. *Insect Mol. Biol.* **14**, 599–605 (2005).
46. Little, T.J., Colbourne, J.K. & Crease, T.J. Molecular evolution of daphnia immunity genes: polymorphism in a gram-negative binding protein gene and an alpha-2-macroglobulin gene. *J. Mol. Evol.* **59**, 498–506 (2004).
47. Bulmer, M.S. & Crozier, R.H. Variation in positive selection in termite GNBPs and Relish. *Mol. Biol. Evol.* **23**, 317–326 (2006).
48. Lazzaro, B.P. & Clark, A.G. Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster. Mol. Biol. Evol.* **20**, 914–923 (2003).
49. Tennessen, J.A. Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. *J. Evol. Biol.* **18**, 1387–1394 (2005).
50. De Bie, T., Cristianini, N., Demuth, J.P. & Hahn, M.W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).