Rapid Expansion of Immune-Related Gene Families in the House Fly, *Musca domestica*

Timothy B. Sackton,*,[1] Brian P. Lazzaro,[2] and Andrew G. Clark[3]

[1]Informatics Group, Faculty of Arts and Sciences, Harvard University, Cambridge, MA
[2]Department of Entomology, Cornell University, Ithaca, NY
[3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY

*Corresponding author: E-mail: tsackton@oeb.harvard.edu.
Associate editor: Patricia Wittkopp

Abstract

The house fly, *Musca domestica*, occupies an unusual diversity of potentially septic niches compared with other sequenced Dipteran insects and is a vector of numerous diseases of humans and livestock. In the present study, we apply whole-transcriptome sequencing to identify genes whose expression is regulated in adult flies upon bacterial infection. We then combine the transcriptomic data with analysis of rates of gene duplication and loss to provide insight into the evolutionary dynamics of immune-related genes. Genes up-regulated after bacterial infection are biased toward being evolutionarily recent innovations, suggesting the recruitment of novel immune components in the *M. domestica* or ancestral Dipteran lineages. In addition, using new models of gene family evolution, we show that several different classes of immune-related genes, particularly those involved in either pathogen recognition or pathogen killing, are duplicating at a significantly accelerated rate on the *M. domestica* lineage relative to other Dipterans. Taken together, these results suggest that the *M. domestica* immune response includes an elevated diversity of genes, perhaps as a consequence of its lifestyle in septic environments.

Key words: gene duplication, immunity, transcriptomics, Dipterans.

Introduction

The rapid increase in the number of sequenced genomes over the past decade has dramatically reshaped our understanding of the evolutionary dynamics of the insect innate immune system. It has long been recognized that genes involved in the immune response are among the most rapidly evolving in many organisms, including mammals (Hughes and Nei 1988; Nielsen et al. 2005; Kosiol et al. 2008), plants (Tiffin and Moeller 2006), and insects (Sackton et al. 2007; Lazzaro 2008; Obbard et al. 2009), with adaptation presumably driven by host–pathogen conflict. In the era of comparative genomics, it has become clear that this pattern of rapid evolution occurs against a backdrop of deeply conserved orthology in core signaling transduction pathways (Toll, imd, JAK/STAT, and JNK) across most insects studied to date (Evans et al. 2006; Sackton et al. 2007; Waterhouse et al. 2007; Werren et al. 2010), with only rare examples of secondary loss (Gerardo et al. 2010).

In addition to signaling cascades that are activated in response to infection, insect immune systems contain classes of proteins involved in pathogen recognition as well as classes of effector proteins whose role is to clear infections (e.g., antimicrobial peptides). Both recognition proteins and effector proteins are encoded by a diverse array of gene families with a variety of functions and specificities. In contrast, immune signaling tends to occur through only four primary signal transduction pathways—Toll, imd, JAK/STAT, and JNK (Buchon et al. 2014). While components of these primary signal transduction pathways are typically conserved in 1:1 orthology across all insects, gene families encoding recognition or effector proteins often vary considerably in copy number between species and exhibit substantial rates of duplication and deletion within evolutionary lineages (Ghosh et al. 2011). Several gene families, especially those encoding antimicrobial peptides, are restricted to particular insect clades (Bulet et al. 1999; Vizioli et al. 2001; Sackton et al. 2007), and the transcriptional response to infection in at least some insects results in the upregulation of numerous taxonomically restricted genes (Sackton et al. 2013).

The house fly, *Musca domestica*, is a particularly relevant insect to study in the context of the evolution of immune systems. Compared with other sequenced insects, they inhabit an unusually wide range of septic matter, including excreta, garbage, and diverse animal carcasses. House flies are also versatile mechanical vectors of numerous diseases of human and livestock, including bacterial, protozoan, viral, and helminthic infections ranging from cholera to tapeworms (Scott et al. 2009; Joyner et al. 2013; Nayduch et al. 2013). This lifestyle suggests that house flies contact and must successfully avoid a wide range of potentially damaging bacteria (Gupta et al. 2012), and implies that house flies may have an unusually effective immune system to cope with these challenges.

Open Access

Article

The recently sequenced house fly genome provides an ideal opportunity to test whether the highly septic lifestyle of this organism is correlated with increased diversity of immune genes. To do this, we first generated new RNA-seq data from experimentally infected and control (sterile-wounded) house flies to characterize the transcriptional response to infection in *M. domestica*. When combined with existing genomic resources in house flies and other Dipterans, these data reveal a striking expansion in the recognition and effector repertoires in *M. domestica*. We also develop a new statistical model for inference of gene family evolution, and show that these expanded repertoires in house flies are most likely associated with extremely elevated rates of gene duplication specifically in immune gene families along the house fly lineage, suggesting that the unusual lifestyle of house flies may be driving increased diversification of immunological molecules.

Results

Identifying Genes Regulated by Infection in *M. domestica*

To characterize the infection-regulated transcriptome in *M. domestica*, we used RNA-seq to quantify expression of genes and transcripts in infected and control (sterile-wounded) flies. We infected 4-day-old adult female flies by piercing the cuticle with a dissecting pin dipped in a mixed bacterial culture of *Serratia marsecens* and *Enterococcus faecalis*. Control flies were treated identically, except they were poked with a pin dipped in sterile LB broth. Six hours after treatment, we collected three replicate pools each of infected and control flies, and sequenced each pool using standard Illumina protocols. Combined, we sequenced 45.5 million reads from infected flies and 51.0 million reads from control flies, of which roughly 70% map to *M. domestica* gene models (NCBI annotation version 100) using RSEM (Li and Dewey 2011).

We identified genes differentially regulated between control and infected samples using the negative binomial approach implemented by DESeq2 (Love et al. 2014). We are able to detect expression for 13,621 genes, out of 14,466 annotated genes in the genome. Overall, we find 1,675 genes differentially regulated at a 5% FDR, with 784 upregulated and 891 downregulated (fig. 1), representing 5.4% and 6.2% of genes in the genome, respectively.

We used two approaches to identify genes in *M. domestica* with homology-based evidence for an immune function. First, we screened for homology to a curated list of genes with immune function in *Drosophila melanogaster* (supplementary table S1, Supplementary Material online). Second, we used an HMM-based approach (Waterhouse et al. 2007) to identify house fly proteins with homology to previously characterized Dipteran immune-related gene families. The gene families we analyzed are listed in (table 1), and alignments and HMMs are available online (https://github.com/tsackton/musca-immunity/tree/master/supplemental_methods/hmm). As expected, there is a highly significant overlap between these homology-annotated immune genes and the set of genes induced by infection. If we define *M. domestica* immune-related genes based on homology to *D.*
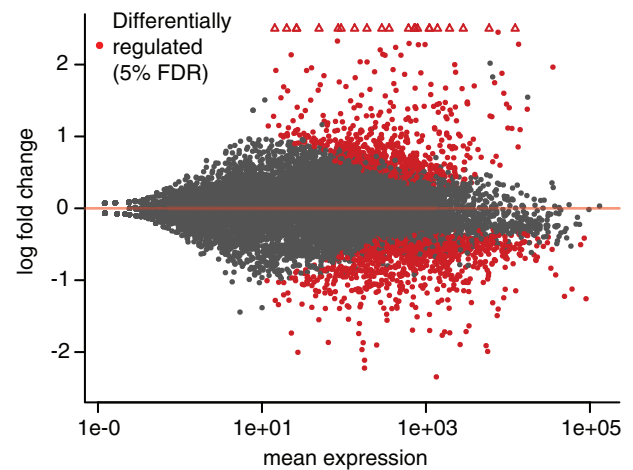
Fig. 1. MA plot of RNA-seq data comparing uninfected control (sterile wounded) samples to infected samples. The *x*-axis shows mean expression for each *Musca domestica* gene (as estimated in DESeq2), and the *y*-axis shows log2 fold change (infected vs. uninfected), also estimated in DESeq2. Points in red are differentially regulated between treatments at a 5% FDR. Open triangles represent points with log2 fold change >2.5.

melanogaster immune genes, we find 80 genes that are both induced and annotated as immune-related (25.8% of all induced genes and 10.3% of all immune-related genes; OR = 6.57, $P < 2.2 \times 10^{-16}$, Fisher's Exact test). If we define *M. domestica* immune-related genes based on HMM annotations, we find 92 genes that are both induced and annotated as immune-related (15.8% of all induced genes and 11.7% of all immune-related genes; OR = 3.55, $P < 2.2 \times 10^{-16}$, Fisher's Exact test). For comparison, 5.4% of genes overall are annotated as induced, 4.1% of genes overall are annotated as immune-related based on HMMs, and 2.2% of genes overall are annotated as immune related based on homology to *D. melanogaster*.

Looking at individual *M. domestica* genes induced by infection reveals a clear enrichment for genes with well-characterized immune annotations (fig. 2A). These include many homologs of consistently and strongly induced effector genes in *D. melanogaster*, such as those encoding cecropins (7 gene family members induced >2-fold in *M. domestica*), attacins (5 family members induced >2-fold in *M. domestica*), diptericins (2 family members induced >2-fold in *M. domestica*), and defensins (2 family members induced >2-fold in *M. domestica*). These also include homologs of genes that have immune roles in some animals including mosquitoes (Adema et al. 1997; Dong and Dimopoulos 2009; Vasta 2009; Romero et al. 2011) but that have not been experimentally characterized in *Drosophila*, including genes encoding FREPs (8 induced in *M. domestica*) and galectins (5 induced in *M. domestica*). A full list of genes with expression information is available at https://github.com/tsackton/musca-immunity/blob/master/results/mdom.difexp.tsv.

Combining all sources of evidence (HMMs, *D. melanogaster* homology, gene ontology, and regulation after infection), we identify and annotate a total of 1,392 putative immune-related genes in *M. domestica*. A full list of these

Table 1. Immune-Related Gene Families Annotated by Hidden Markov Models.

Short Name	Description	Class
ATT	Attacin antimicrobial peptides	Effector
CEC	Cecropin antimicrobial peptides	Effector
DEF	Defensin antimicrobial peptides	Effector
DIPT	Diptericin antimicrobial peptides	Effector
GPX	Glutathione peroxidases	Effector
HPX	Heme peroxidases	Effector
LYS	Lysozymes	Effector
PPO	Prophenoloxidases	Effector
TPX	Thioredoxin peroxidases	Effector
TSF	Transferrins	Effector
BGBP	β-glucan binding proteins (GNBPs in Drosophila)	Recognition
CTL	C-type lectins	Recognition
FREP	Fibrinogen-related proteins	Recognition
GALE	Galectins	Recognition
IGSF	Ig superfamily proteins	Recognition
MD2L	MD2-like proteins	Recognition
NIM	Nimrods	Recognition
PGRP	Peptidoglycan recognition proteins	Recognition
SRCA	Scavenger receptor class A	Recognition
SRCB	Scavenger receptor class B	Recognition
SRCC	Scavenger receptor class C	Recognition
TEP	Thioester-containing proteins	Recognition
CLIPA	CLIP-domain serine protease class A	Signaling
CLIPB	CLIP-domain serine protease class B	Signaling
CLIPC	CLIP-domain serine protease class C	Signaling
CLIPD	CLIP-domain serine protease class D	Signaling
CLIPE	CLIP-domain serine protease class E	Signaling
NFKB	Nf-κB proteins	Signaling
SPRN	Serine protease inhibitors	Signaling
TLL	Toll family proteins	Signaling

genes, with annotations where possible, is available as supplementary table S2, Supplementary Material online.

Gene Ontology Analysis Suggests a Coordinated Shift from Metabolism to Protein Production after Infection

In addition to genes encoding proteins with specific immune functions, bacterial infection leads to broad changes in patterns of gene expression that may be reflective of physiological processes altered by infection. To better understand the overall biology of the transcriptional response to infection in house flies, we focused on the 613 induced genes (FDR < 0.05) and 568 repressed genes (FDR < 0.05) which were able to be annotated to GO terms based on homology (Scott et al. 2014).

As expected, genes induced by infection are enriched for GO classes related to immunity, including "response to biotic stimulus" (Holm's adjusted P value = 4.23×10^{-13}, Odds Ratio = 2.55), "response to stress" (adjP = 1.82×10^{-09}, odds ratio = 1.80), and "response to external stimulus" (adjP = 5.22×10^{-03}, odds ratio = 1.52). Additionally, genes induced by infection are enriched for a number of biological process GO categories that are suggestive of a coordinated upregulation of protein synthesis and export machinery. These include "translation" (adjP = 4.98×10^{-04}, odds ratio = 1.98), "transport" (adjP = 2.17×10^{-02}, odds ratio = 1.35), "cellular protein modification process" (adjP = 4.16×10^{-02}, odds ratio = 1.39), and "protein

metabolic process" (adjP = 7.70×10^{-07}, odds ratio = 1.63). In contrast, genes repressed by infection are enriched for GO terms suggestive of a role in metabolism. GO terms overrepresented in the downregulated gene set are primarily related to metabolism: "generation of precursor metabolites and energy" (adjP = 1.32×10^{-16}, odds ratio = 4.11), "lipid metabolic process" (adjP = 5.79×10^{-09}, odds ratio = 2.13), "catabolic process" (adjP = 3.24×10^{-08}, odds ratio = 1.83), and "secondary metabolic process" (adjP = 6.06×10^{-03}, odds ratio = 2.16). Molecular function GO terms paint a similar picture (supplementary table S3, Supplementary Material online). Taken together, these patterns point toward a pronounced physiological shift in house flies after infection, away from basal metabolism and toward protein production, transport, and secretion. This is consistent with recent work in *Drosophila* and other insects suggesting a close connection between metabolic control and immune system regulation (De Gregorio et al. 2001; Buchon et al. 2014; Unckless et al. 2015). A full list of GO terms enriched (at a Holms-adjusted P value < 0.05) for genes either upregulated or downregulated by infection is in supplementary table S3, Supplementary Material online.

Comparison to *D. melanogaster* RNA-seq Data Suggests *M. domestica* Induces a Larger Suite of Genes after Infection

To contextualize our observations about the genes induced by infection in *M. domestica*, we generated in parallel a new,
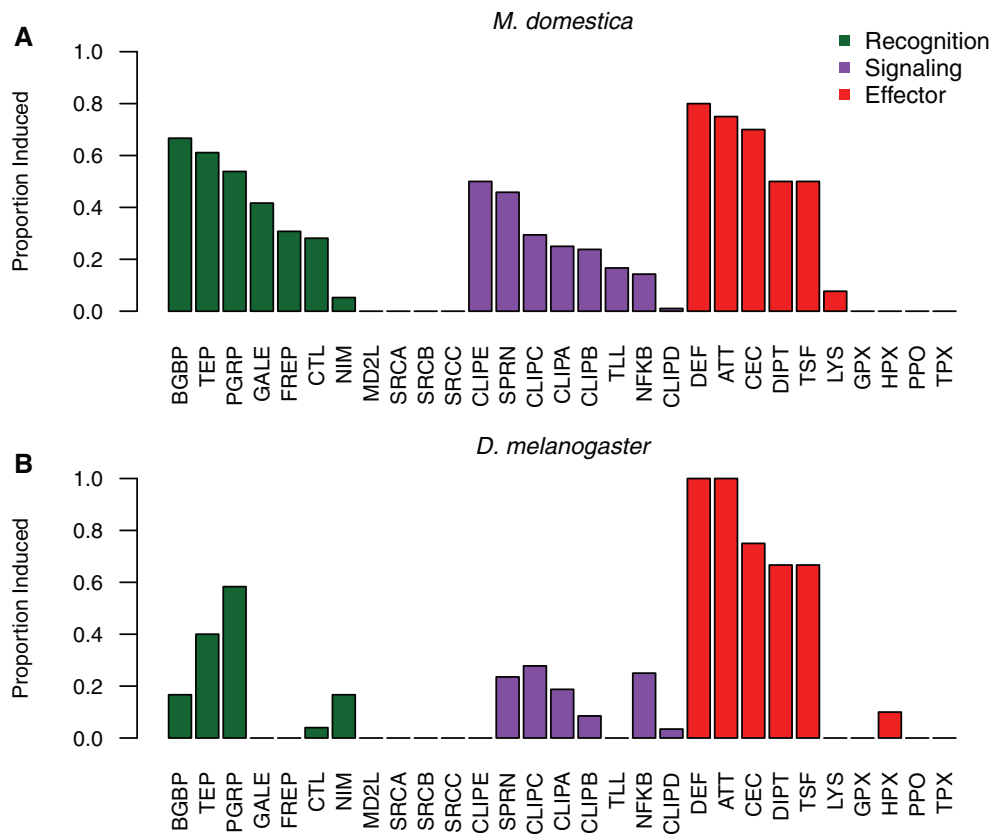
FIG. 2. (*A*) The proportion of each family of immune-related genes that are upregulated (at a 5% FDR, based on DESeq2 analysis) after infection relative to uninfected control (sterile wounded) at 6-h post-treatment in *Musca domestica*. Families are defined based on HMM profiles and ordered by category (recognition, signaling, effector) and proportion induced within each category. (*B*) The proportion of the each family of immune-related genes that are upregulated (at a 5% FDR, based on DESeq2 analysis) after infection relative to an uninfected control (naive, untreated) at 12-h post-treatment in *D. melanogaster*. Families are defined and ordered as in part A.

roughly comparable *D. melanogaster* RNA-seq data set. While previous studies have been conducted of the transcriptional response to infection in *D. melanogaster* (De Gregorio et al. 2001; Irving et al. 2001), a direct comparison has the benefit of using data generated with the technology, the same infection protocol, at a similar time point, and in the same laboratory as the *M. domestica* data (see "Materials and Methods" for details) to minimize technical artifacts. We also used the exact same analysis pipeline to analyze the *D. melanogaster* RNA-seq data. The RNA-seq data from *D. melanogaster* is of similar depth and quality (67.8 million reads for the infected replicates pooled, 75.6 million reads for the uninfected replicates pooled, 95% mapped to *D. melanogaster* gene models).

Despite these technical similarities, we caution that there are a number of important differences confounded with species in this design. In particular, differences in the nature of the control (sterile wound vs. naive), timing of sampling after infection (6 vs. 12 h), tissue allometry, and body size of the insects (which means that the bacterial inoculum proportional to body size is lower in the house fly) could all result in transcriptional differences between species. The nature of the control (untreated in *D. melanogaster*) is very likely to increase the number of genes detected as regulated by infection in *D. melanogaster*, but in the absence of detailed time course or dose response data for these particular bacteria it is

difficult to know conclusively the effect of timing or dosage differences on transcription. While these caveats limit our ability to infer quantitative differences in orthologous gene pair expression between the species, we believe that the general trends we report below are likely to be robust.

Of the 11,135 genes in *D. melanogaster* with detectable expression in our data, 156 are upregulated by infection and 150 are downregulated by infection, representing 1.4% and 1.35%, respectively, of expressed genes, and 0.9% and 0.87%, respectively, of all genes. This is notably fewer than in *M. domestica*, especially when taking into account the likely lower quality of the house fly annotations. Of induced genes, 27.6% are annotated as having an immune function. Unsurprisingly, the induced genes include many encoding known antimicrobial peptides (four attacins, three cecropins, defensin, two diptericins, drosomycin, and drosocin), recognition factors (two TEPs, seven PGRPs, and two Nimrods), and signaling components (cactus, Relish). A full list of genes with expression information is at https://github.com/tsackton/musca-immunity/blob/master/results/dmel.difexp.tsv.

At the level of HMM-defined gene families, *D. melanogaster* induces many of the expected classes, with substantial overlap with the classes induced in *M. domestica* (fig. 2B). Notably, however, we find no evidence for induction of any gene encoding a FREP or galectin in *D. melanogaster*, nor are we
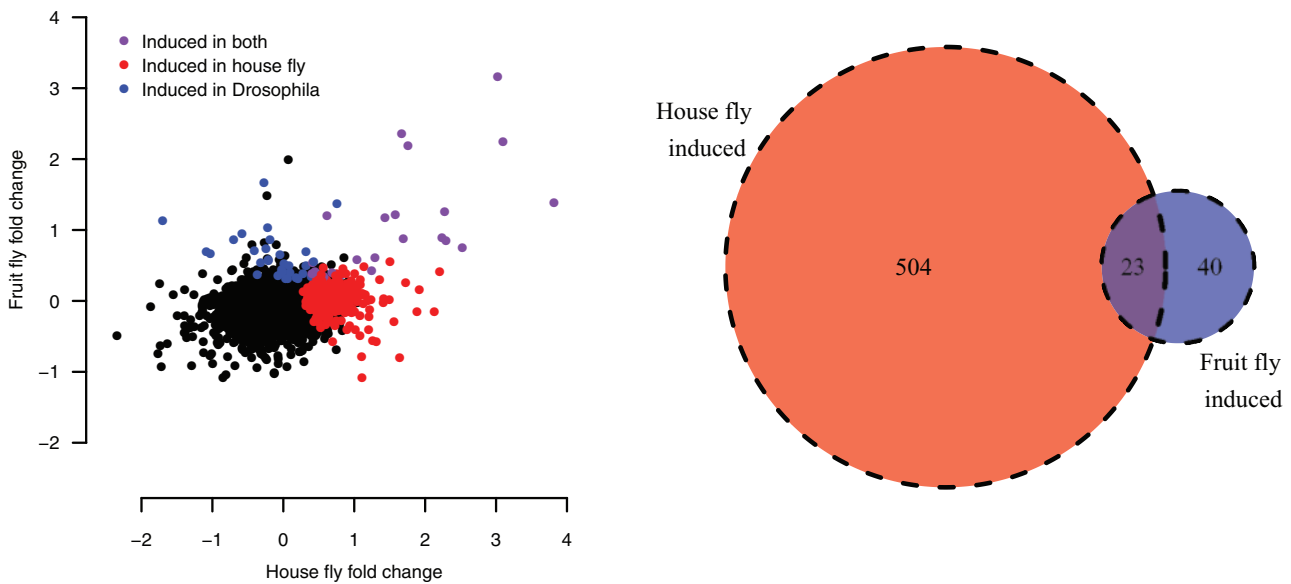
FIG. 3. The correlation between fold change after infection in house fly and fruit fly. Each point represents a single 1:1 ortholog with expression data in both species. Genes with significant upregulation after infection in either fruit fly, house fly, or both are colored; significant upregulation is defined based on a 5% FDR estimated with DESeq2. The Venn diagram shows the number of 1:1 orthologs in each induction class.

aware of any previously published reports of induction of members of either of these gene families in the *D. melanogaster* systemic immune response after bacterial infection. In contrast, 31% and 42%, respectively, of genes in these classes are induced by infection in *M. domestica*.

In our data set, there are 7,934 single-copy orthologs between *D. melanogaster* and *M. domestica* with detectable expression in both species. For these genes, we directly compared patterns of regulation after infection. While we find, as expected, highly significant overlaps in both induced genes ($P = 4.61 \times 10^{-12}$, Fisher's Exact test) and repressed genes ($P = 4.1 \times 10^{-07}$, Fisher's Exact test), there are many more genes induced in *M. domestica* alone than in *D. melanogaster* alone (fig. 3). This suggests that at least a portion of the greater number of genes regulated by infection in *M. domestica* may be attributable to regulatory evolutionary change in shared orthologs.

We also compared the set of gene ontology terms overrepresented among both upregulated and downregulated genes in *D. melanogaster* to those described for *M. domestica* above. In *D. melanogaster*, GO terms associated with immune functions dominate the list of terms overrepresented in the upregulated class (supplementary table S3, Supplementary Material online). However, we see no evidence for upregulation of GO terms associated with protein transport or translation. As discussed earlier, it is possible that differences in timing (6 vs. 12 h), or other technical or biological factors, could be associated with this difference, but it is also possible that this represents a reduced investment in immune protein production in *D. melanogaster* compared with *M. domestica*. For the downregulated genes, we see a similar set of GO categories associated with the *D. melanogaster* response as the *M. domestica* response (supplementary table S3, Supplementary Material online), supporting the idea that

the downregulation of basal metabolism is a broadly consistent response to infection in many Dipterans.

Taken as a whole, bacterial infection in *M. domestica* appears to result in differential expression of more genes (at 6-h post-infection) than in *D. melanogaster* (at 12-h post-infection). These additional regulated genes appear to include additional categories of immune-related genes (e.g., genes encoding FREPs and galectins), a broader range of biological processes (including protein translation and export machinery), and induction of more members of shared immune-related families that may have expanded in *M. domestica* (including genes encoding attacins, cecropins, TEPs, transferrins, and defensins).

The Infection-Induced Transcriptome of *M. domestica* Is Enriched for Taxonomically Young Genes

In several insects studied to date, the transcriptional response to infection includes a large number of young, taxonomically restricted genes (Sackton and Clark 2009; Sackton et al. 2013; Gupta et al. 2015). To test whether the data from *M. domestica* also show this pattern, we identified the phylogenetic age of each protein in the house fly genome using BLASTP and then inferring a date for gene origination based on the age of the deepest homolog identified (ages of species divergences from timetree.org). As has been seen in other insects, young genes in *M. domestica* are more likely to be induced by infection than old genes (Logistic regression: $\beta = -5.53 \times 10^{-4}$, $P = 7.88 \times 10^{-13}$, fig. 4A).

Recently, it has been suggested that phylostratigraphic methods such as this are prone to bias, since factors such as protein length and evolutionary rate can influence the probability of detecting ancient homologs (Moyers and Zhang 2015). To attempt to control for this effect, we normalized our age estimates based on the estimated effects of
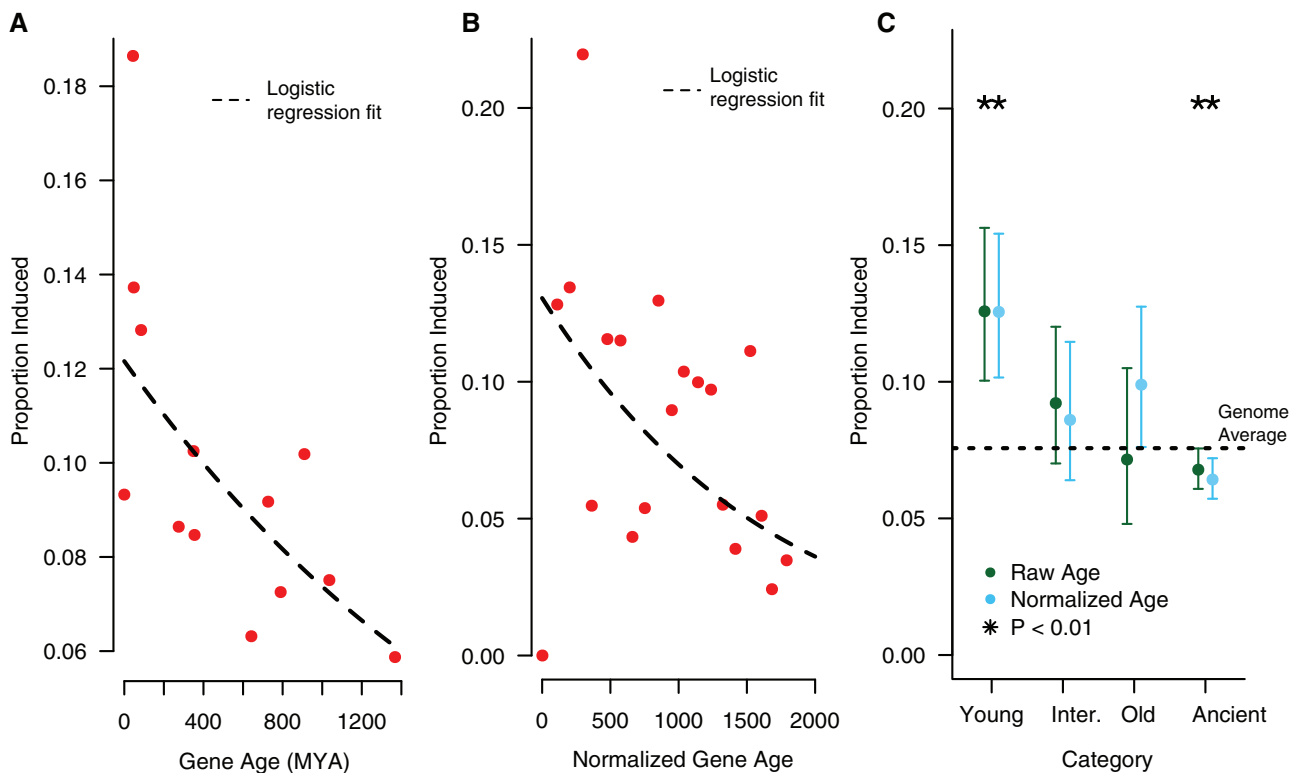
Fig. 4. (A) The proportion of genes induced by infection for each inferred gene age. The dashed line shows the logistic regression fit, which is highly significant (age $\beta = -5.53 \times 10^{-04}$, $P = 7.88 \times 10^{-13}$). (B) The proportion of genes induced by infection for each normalized gene age. The dashed line shows the logistic regression fit, which is highly significant (age $\beta = -6.94 \times 10^{-04}$, $P = 2 \times 10^{-16}$). Note that the normalization procedure generates a continuous distribution of ages, but for plotting purposes we converted this back to discrete age classes. (C) Proportion of genes induced by infection by age category. After classifying genes into one of four categories based on either raw (uncorrected) age (green points) or normalized (corrected) age (blue points), we estimated the proportion of each age class induced by infection. The dotted line shows the genome-wide average proportion genes induced by infection (0.081). To estimate significance, each category was compared with the remaining categories in turn using a chi-squared test. We get similar results using a logistic regression to estimate the effect of each category relative to the "ancient" group.

protein length and expression level as a proxy for evolutionary rate (Pál et al. 2001; Larracuente et al. 2008) in our data, and repeated our analysis (see "Materials and Methods" for details). After this correction, we still find strong evidence that younger genes are more likely to be induced by infection than older genes (Logistic regression: $\beta = -6.93 \times 10^{-04}$, $P < 2 \times 10^{-16}$, fig. 4B).

As an alternative approach, we also assigned genes to a small number of age categories (young = Schizophora-specific genes, intermediate = Insecta-specific genes, old = Protostomia-specific genes, ancient = Opisthokont-specific genes) and consider the patterns of expression in genes in each category. Using both uncorrected and corrected phylostratigraphic age categories, we find that genes in the "young" category are more likely to be induced after infection than genes in the other categories, and genes in the "ancient" category are less likely to be induced after infection than genes in the other categories (fig. 4C).

Gene Duplication and Loss in *M. domestica*
In addition to apparently inducing a broader suite of genes encoding immune-related proteins than many other insects, the *M. domestica* genome encodes a greater diversity of

immune-related genes than many other insects studied to date. For example, the *Musca* genome contains the highest number of genes encoding TEPs in a sequenced Dipteran genome (Scott et al. 2014), and in general has a high number of many immune-related gene families (table 2). To test whether this is a general pattern across the *Musca* immune system, and to determine whether the diversity of genes encoding immune proteins in *Musca* is driven by increased rates of gene duplication, decreased rates of gene loss, or both, we developed a phylogenetic framework to assess rates of copy number change using a Poisson regression approach (Koerich et al. 2008).

In this framework, we fit a Poisson regression to counts of gene gains and gene losses on each branch of the Dipteran phylogeny (fig. 5). We first verified the behavior of our model by simulation and by comparison to previous methods. Then, we focused on three different model parameterizations. In all three approaches, we allowed a different rate of duplication and loss on the *Musca* lineage compared with the rest of the tree. However, each approach differs in how we treat variation among genes. In the first approach, we estimated a single birth rate and death rate for all genes with a similar functional annotation (e.g., recognition, signaling, effector, nonimmune).

Table 2. Number of Genes Identified by HMM for Each Gene Family from Table 1.

		Mdom	Gmor	Aaeg	Adar	Agam	Aste	Cqui	Dana	Dmel	Dmoj	Dpse	Dvir	Dwil	Dyak
Canonical effectors	ATT	*10*	4	1	0	0	0	0	3	4	3	4	4	3	4
	DEF	*5*	0	4	1	2	1	1	1	1	2	1	2	1	1
	DIPT	4	0	1	1	1	1	1	3	3	*5*	4	4	4	3
	CEC	*12*	2	9	3	4	3	4	3	5	1	5	5	4	5
	LYS	*32*	4	7	3	8	4	5	13	13	13	18	14	12	12
Noncanonical effectors	TPX	6	6	5	5	5	5	5	8	8	6	6	6	*9*	7
	PPO	23	4	*25*	11	20	9	22	9	10	8	8	8	8	11
	GPX	1	0	3	3	3	3	*4*	3	2	1	2	1	2	2
	HPX	12	8	*19*	17	*19*	14	13	10	10	10	11	10	10	10
	TSF	6	3	*5*	3	4	4	4	3	3	3	4	3	3	3
Canonical recognition	NIM	*23*	10	8	10	9	10	15	18	17	19	19	19	15	18
	PGRP	*17*	4	10	8	7	7	10	13	13	14	15	13	13	15
	BGBP	3	3	7	6	7	5	*13*	6	7	6	8	6	10	8
	TEP	*22*	4	8	12	15	7	11	5	6	7	5	6	5	6
Other recognition	CTL	41	11	43	14	29	25	*54*	36	38	24	32	29	27	42
	FREP	38	7	34	22	50	27	*93*	42	14	22	31	34	35	15
	GALE	*13*	8	12	7	10	8	9	6	6	6	7	7	6	7
	IGSF	1	1	0	0	1	1	0	1	1	1	1	1	0	*2*
	MD2L	12	5	*26*	12	16	10	21	8	8	8	7	8	9	8
	SRCA	3	2	2	3	2	2	2	3	3	3	3	3	3	3
	SRCB	*18*	11	13	14	14	15	16	14	14	15	14	15	15	16
	SRCC	8	4	5	7	6	5	8	6	*9*	6	6	8	6	*9*
Signaling and modulation	NFKB	7	*8*	4	4	3	3	*8*	3	4	4	4	4	4	3
	SPRN	26	14	26	19	18	16	*51*	31	30	21	28	19	24	33
	TLL	7	8	*15*	9	14	8	13	11	10	8	*15*	9	*15*	13
	CLIPA	11	11	*28*	20	22	11	19	17	21	9	16	13	11	24
	CLIPB	27	19	99	36	82	37	*119*	41	80	22	57	30	30	74
	CLIPC	25	22	57	30	72	39	*81*	28	22	23	24	24	30	20
	CLIPD	166	66	182	103	150	79	*187*	139	132	115	116	130	141	145
	CLIPE	2	3	*6*	2	4	3	5	3	2	2	1	2	2	5

NOTE.—In each row, the highest value is italicized.

Species abbreviations: Mdom, *M. domestica*; Gmor, *G. morsitans*; Aaeg, *A. aegypti*; Adar, *A. darlingi*; Agam, *A. gambiae*; Aste, *A. stephensi*; Cqui, *C. quinquefasciastus*; Dana, *D. ananassae*; Dmel, *D. melanogaster*; Dmoj, *D. mojavensis*; Dpse, *D. pseudoobscura*; Dvir, *D. virilis*; Dwil, *D. willistoni*; Dyak, *D. yakuba*.

In the second approach, we focus on specific related gene classes (e.g., cecropins, TEPs). Finally, we fitted a separate birth and death rate for each individual gene family. In all analyses, we focused on gene families basally present in Diptera, and with at least one gain or loss on the tree.

Poisson Regression Is an Accurate Method for Estimating Rates of Gene Gain and Loss

To verify the behavior of our method, we simulated 1,000 gene trees, conditioned on a fixed species tree, for each of 12 different duplication/loss rates ranging from 0.00057 to 0.341 (supplementary table S4, Supplementary Material online), using the GuestTreeGen tool in jprime (https://github.com/arvestad/jprime). In our simulations, we fixed the duplication rate to equal the loss rate (so the total rate in events/MY is twice the input simulation rate), and after simulation restricted our analysis to the subset of simulations where the gene family was not lost entirely on one of the two branches leading from the root of the tree (to be consistent with our filtering of our analysis of the real data). This drastically reduces the number of simulation results we used for the highest turnover rates (supplementary table S4, Supplementary Material online), but up to a turnover rate of 0.023 events/ million years we retained at least 100 simulated trees. While we report results for all simulation values, those >0.023

events/MY should be treated with caution due to the low numbers of gene families passing our filters.

For each set of trees simulated under the same rate parameters, we estimated a fixed turnover parameter (birth rate + death rate), and also separate birth and death parameters, using our Poisson regression model. Even for very high turnover rates, we recovered overall turnover rates and duplication rates very similar to the simulated values (fig. 6A). For low to moderate turnover rates, our estimates of loss rates were also very accurate, but for very high turnover rates we began to underestimate loss rates (fig. 6A), probably because losses that extinguish the gene family are dropped from the analysis and thus not counted. We note that existing methods such as CAFE also perform poorly at very high turnover rates (fig. 6B), and this is not unexpected (De Bie et al. 2006). At low to moderate turnover rates, our method performs as well as CAFE and allows for more complex modeling of branch dependencies, although it also requires computationally intensive gene tree estimation and inference of duplications and losses on specific branches using phylogenetic methods.

In real data, considerable rate variation among individual gene families creates over-dispersion in the Poisson model and leads to serious underestimates of the SE of the coefficients of the model, and thus incorrect *P* values. To correct for this, we use a mixed model approach, specifying a random

intercept for each individual gene family; this is conceptually similar to using observation-level random effects (Harrison 2014). To verify the performance of our mixed model, we simulated 1,000 data sets with randomly selected
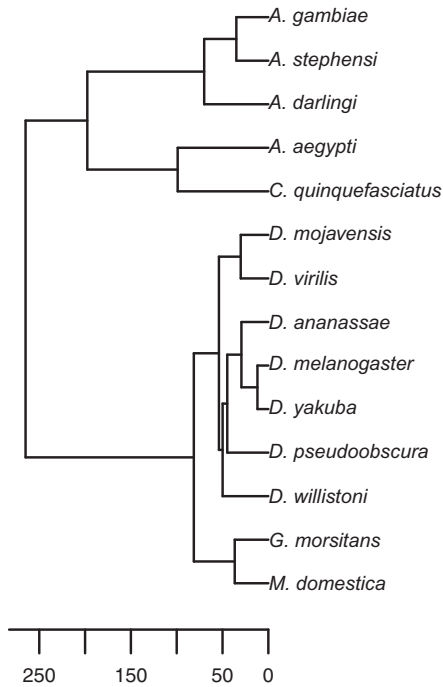

FIG. 5. Ultrametric tree of Dipteran species included in gene family analysis, estimated using the "chronos" function in the ape package for R. Scale bar is in millions of years ago, based on calibrations taken from timetree.org.

"immune" genes, as described in the "Materials and Methods". On average the effect of this "immune" classification on duplication or loss rates should be zero in these random permutations, so we expect to observe no more than 5% of simulations that reject the null hypothesis of no effect at a nominal alpha of 0.05. With the naive Poisson approach, we see a dramatic mis-calibration of the significance level (88.9% of all simulations have a P value < 0.05), which is completely eliminated by accounting for family-level rate variation using random effects (5.1% of all simulations have a P value < 0.05).

Genes Encoding Effector and Recognition Proteins Duplicate Rapidly on the *Musca* Lineage

At the broadest level, we find evidence that the *Musca* lineage has experienced a significantly higher rate of gene turnover (duplication + loss) than other Dipteran lineages for both immune genes (defined based on homology to *D. melanogaster*) and nonimmune genes (table 3). Notably, the increased turnover rate along the *Musca* lineage is significantly higher for immune genes than for nonimmune genes (interaction $\beta = 0.22$, $P = 0.0164$, table 3), suggesting that immune genes in particular experience rapid turnover along the *Musca* lineage. The *Musca*-specific increase in turnover of immune genes appears to be driven by an increased duplication rate (duplications only, interaction $\beta = 0.43$, $P = 3.45 \times 10^{-05}$, table 3) rather than a change in the rate of gene loss (losses only, interaction $\beta = -0.0054$, $P = 0.983$, table 3). When we define immune genes more broadly to include both genes with homology to *D. melanogaster* immune-related genes and members of HMM-defined immune gene classes, the same trends hold albeit somewhat
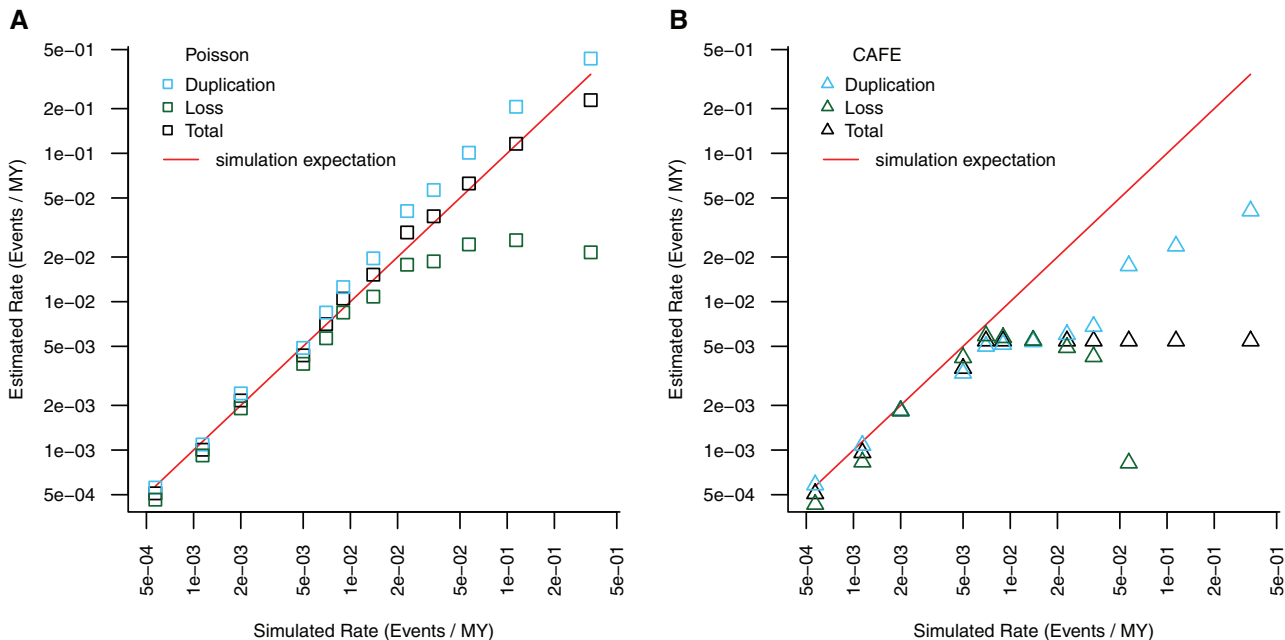
FIG. 6. Analysis of simulated gene duplication data. The same simulation inputs (supplementary table S5, Supplementary Material online) were analyzed with our Poisson regression framework (*A*) and with CAFE (*B*) for a range of simulated duplication/loss rates. Points represent estimated turnover (duplication + loss, black), loss (green) and duplication (blue) rates estimated with each method; the red line is the expectation based on the simulated input values.

Table 3. Poisson Regression Models for the Analysis of Duplication and Loss Data.

	Model Term	Effect Size	SE	P value
Response variable: turnover rate	(Intercept)	−5.912	0.017	<2e-16
	Musca branch=TRUE	0.253	0.025	<2e-16
	Immune gene=TRUE	0.517	0.103	4.62e-07
	Musca×immune interaction	0.222	0.092	0.0164
Response variable: duplication rate	(Intercept)	−6.388	0.019	<2e-16
	Musca branch=TRUE	0.637	0.028	<2e-16
	Immune gene=TRUE	0.517	0.112	3.68e-06
	Musca×immune interaction	0.428	0.103	3.46e-05
Response variable: loss rate	(Intercept)	−7.412	0.033	<2e-16
	Musca branch=TRUE	−0.930	0.070	<2e-16
	Immune gene=TRUE	0.600	0.171	0.000459
	Musca×immune interaction	−0.005	0.255	0.983

more weakly (supplementary table S5, Supplementary Material online: duplication and loss interaction $\beta = 0.13$, $P = 0.0868$; duplications only interaction $\beta = 0.31$, $P = 1.6 \times 10^{-04}$; losses only interaction $\beta = -0.26$, $P = 0.256$).

To rule out the possibility that our results are driven by unusual rates in non-*Musca* lineages, we repeated these analyses with a model that allows for separate rates for each family of Dipterans included in our analysis (Muscidae, Drosophilidae, Glossinidae, and Culicidae), excluding events that occurred in basal lineages that pre-date the divergence of these families. In this analysis, we treated the Muscidae as the reference level; while genes with immune annotation have higher duplication rates in general than other gene families in the genome, in all non-Muscidae lineages the increase in duplication rates associated with immune function is significantly lower than the increase in duplication rates associated with immune function in Muscidae (family×immune interaction $\beta = -0.58$ for Drosophilidae, -1.136 for Glossinidae, and -0.62 for Culicidae, all P values $< 1 \times 10^{-05}$).

To initially determine if particular components of the innate immune system are responsible for this pattern, we estimated separate rates for different functional classes of immune proteins (recognition, signaling, modulation, and effectors; based on homology to *D. melanogaster* proteins with annotated functions in these classes). Gene families encoding recognition and effector proteins have elevated duplication rates in the Muscidae lineage compared with other Dipterans, but gene families encoding signaling or modulation proteins do not (fig. 7, left-most panel). Using simultaneous tests of linear contrasts, we determined whether this increase persists in models that allow separate rates for each major Dipteran lineage included in our data set. We found the increase in duplication rates of genes encoding effector or recognition proteins (compared with the duplication rate of nonimmune genes) is consistently elevated in the Muscidae lineage in all comparisons (fig. 7, right three panels).

In order to understand the specific drivers of this pattern, we analyzed rates of gene duplication and loss in HMM-defined gene families that make up the broader homology-based classes, focusing on effector and recognition classes. Among genes encoding recognition or effector proteins, the TEPs, lysozymes, and cecropins show the most striking pattern, with significantly larger increases in duplication rates

(relative to the baseline nonimmune duplication rate) along the *M. domestica* lineage than in other Dipterans pooled (table 4). Furthermore, for all three of these gene classes the increase in duplication rates in the Muscidae lineage is significantly greater than the increase in duplication rates in either the mosquito or the Drosophila lineages, relative to the baseline rate of all genes not in the family in question (table 5).

We can also fit our birth/death model to individual gene families (orthogroups), although in these cases we have substantially reduced power to estimate rates accurately, and thus will likely only detect the most extreme effects. We used this approach to estimate for each gene family the relative turnover rate (birth + death) on the *Musca* lineage compared with the rest of the tree; this is positive for gene families with a higher turnover rate on the *Musca* lineage and negative for gene families with a lower turnover rate on the *Musca* lineage. Immune-related genes (combining HMM-based and homology-to-Drosophila based annotations) are overrepresented among gene families with individually significant accelerations in turnover rate along the *M. domestica* lineage (6/154 immune families, 53/4,565 nonimmune families, $P = 0.012$, Fisher's Exact test), including orthogroups containing TEPs, lysozymes, and cecropins (consistent with our HMM-class rate estimation; table 6 has the full set of immune-related gene families with elevated turnover rates in *Musca*). Thus, all our modeling approaches consistently demonstrate a specific acceleration of rates of gene duplication in certain key classes of genes encoding recognition and effector proteins along the *M. domestica* lineage.

As an additional line of evidence, we also examined the counts of each HMM-defined gene family detected in each species. Here, we do not focus on rates of duplication or the phylogenetic relationships among genes, but rather just the absolute count of the number of genes with evidence for protein homology to particular immune gene families. We did this in as unbiased a way as possible, by using the same input set of HMM profiles to screen the full set of annotated proteins for each target species with HMMER. The counts of each gene family for each species are listed in table 2. For all three gene families (cecropins, TEPs, lysozymes) where we infer a dramatically increased rate of gene duplication on the *Musca* lineage, we note that *M. domestica* has the most
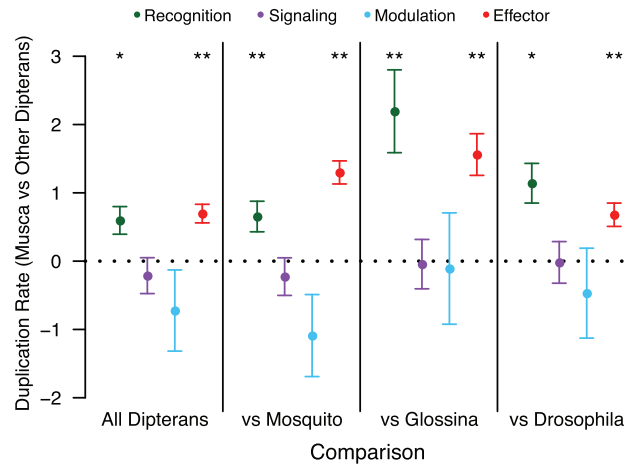
Fig. 7. Linear contrasts testing the relative duplication rate in the *Musca* lineage versus other Dipterans for specific immune classes. Each point represents the estimated linear contrast (±SE) for the duplication rate of genes in that category in *Musca*, compared with the duplication rate of genes in that category in all other Dipterans together or in individual non-*Musca* lineages. *P* values are listed above each point for the test of whether the contrast is equal to 0, which is the expectation if the duplication rate for that category is equal on the *Musca* branch and the rest of the tree (* 0.01 < *P* < 0.05, ** *P* < 0.01).

Table 4. Linear Contrasts Testing Duplication Rate in the *Musca* Lineage versus Other Dipterans for Specific Immune HMM Families.

Family	Estimate	SE	P value
BGBP	−4.034	8.03	1
CEC	1.794	0.41	0.000294
CLIPA	−3.185	8.95	1
CLIPB	0.066	0.24	1
CLIPC	−3.608	10.76	1
CLIPD	−0.745	0.52	0.98
CTL	0.216	0.21	1
FREP	0.182	0.21	1
GALE	−7.821	57.48	1
HPX	−0.631	1.05	1
IGSF	−0.101	1.05	1
LYS	1.427	0.25	<0.000001
MD2L	0.733	0.37	0.72
NFKB	1.795	0.93	0.75
NIM	−4.268	7.42	1.00
PGRP	1.201	0.44	0.16
PPO	1.036	0.36	0.11
SPRN	0.429	0.39	1.00
SRCA	−2.236	9.93	1.00
SRCB	0.120	0.77	1.00
SRCC	1.791	0.96	0.80
TEP	1.568	0.30	0.00000515
TLL	−4.964	8.64	1.00
TPX	−79.750	47450000.00	1.00
TSF	2.887	1.25	0.40

members of the full set of annotated Dipteran genomes we investigated. In general, the house fly immune system appears to encode a larger number of both effectors (including antimicrobial peptides, PPO pathway genes, and lysozymes) and recognition proteins (including TEPs, Nimrods, and PGRPs) than any other Dipteran included in our analysis.

Discussion

The house fly is unique among Dipteran insects sequenced to date in that it lives primarily in highly septic environments

such as excreta, garbage, and carcasses. These environments have the potential to significantly impact the evolutionary dynamics of innate immune defense in this species. Organisms might deal with a potentially infectious environment by strengthening the barriers to initial infection, generating a more impermeable cuticle that is tougher or less prone to breaches that may allow bacterial invasion. They might also simply become more tolerant of bacterial presence, and not expend the energy entirely on strengthened resistance (Schneider and Ayres 2008; Medzhitov et al. 2012). In this study, we combine transcriptome sequencing before and after infectious challenge with homology based annotations to characterize the genes involved in the *M. domestica* immune response and elucidate their evolutionary history. Numerous studies have reported that genes encoding proteins in the insect immune response are exceptionally likely to evolve by repeated positive selection (Schlenke and Begun 2003; Sackton et al. 2007; Lazzaro 2008; Obbard et al. 2009; Keebaugh and Schlenke 2012; Roux et al. 2014); here, we focus particularly on rates of gene gain and loss.

Several lines of evidence suggest that the *M. domestica* immune response is unusual, at least when compared with the standard Dipteran model *D. melanogaster*. First, house flies appear to induce a broader range of putative immune genes than *D. melanogaster*. In addition to upregulating genes encoding a number of conserved antimicrobial peptides (e.g., defensins, cecropins) after infection, *M. domestica* also induces large numbers of genes encoding FREPs and galectins. While both FREPs and galectins have been associated with immune function in other insects, we find no evidence that are induced in *D. melanogaster*, at least under the conditions we assayed, and there is little published evidence to link any FREP or galectin to immune function in *Drosophila*. No member of either of these gene classes was induced in early microarray studies of whole flies (De Gregorio et al. 2001; Irving et al. 2001; Boutros et al. 2002), and to our knowledge there are no

Table 5. Linear Contrasts Testing Elevated Duplication Rate in the *Musca* Lineage versus Other Dipteran Families for Cecropins, TEPs, and Lysozymes.

Family	Estimate	*Musca* versus *Drosophila*			*Musca* versus Mosquitoes		P value
		SE		P value	Estimate	SE	
TEP	2.45	0.63		0.000192	1.21	0.33	0.000546
CEC	1.22	0.52		0.036	2.06	0.55	0.000348
LYS	0.8	0.3		0.0143	2.94	0.49	4.52e-09

Table 6. Orthologous Groups with Putative Immune Function and Evidence for Accelerated Duplication Rates in the *Musca domestica* Lineage.

Orthogroup ID	*Musca* Rate	Q value	HMM Class	Homology Class
14504	1.40	8.92E-08	LYS	Effector
17260.2	1.82	0.0053	PPO	Effector
24756.1.1	2.24	0.0078	MD2L	None
4938	1.61	2.50E-06	TEP	Recognition
7079	2.04	8.652E-06	CEC	Effector
9685	1.55	0.0278	SPRN	None

clear functional studies linking any member of either of these gene classes to an immune phenotype in *D. melanogaster*, although there is some evidence that galectin is expressed in circulating hemocytes (Pace et al. 2002) and one FREP gene appears to be induced after gut infection (Buchon et al. 2009).

Second, we find some suggestion that house flies may induce a stronger immune response than *D. melanogaster* based on the function of nonimmune genes that are regulated by infection. After challenge, we find that *M. domestica* upregulates a large number of genes with functions related to protein transport, protein synthesis, and protein export, and downregulates a large number of genes with functions related to oxidative phosphorylation and metabolism. The downregulation of genes with functions related to metabolism has been previously noted in *Drosophila* (e.g., De Gregorio et al. 2001), and is likely related to metabolic shifts associated with infection in *Drosophila* (Ayres and Schneider 2009; Chambers et al. 2012) and other animals (Exton 1997; Hart 1998). However, to our knowledge, the upregulation of protein transport machinery has not be previously shown in Dipterans and is not detectable in our *D. melanogaster* expression data. It is possible that this transcriptional pattern relates to the ability of house flies to manage resource allocation demands associated with immune activation (Moret and Schmid-Hempel 2000; Lee et al. 2006; Bashir-Tanoli and Tinsley 2014; Bajgar et al. 2015).

The results from the above-mentioned infection-related gene expression experiments are likely robust, but there are some caveats to their interpretation due to experimental design and other differences between the species (see "Results" for discussion). However, genomic analysis also supports the uniqueness of *Musca* immunity when compared with *Drosophila*. We find clear evidence that genes encoding both recognition and effector components of the insect immune response are duplicating more rapidly along the *M. domestica* lineage than in other Dipterans, and that among the gene families in its genome, those involved in recognition

and effector functions are among the fastest to expand. In particular, we find evidence for dramatic expansions of TEPs (thioester-containing proteins involved in recognition and phagocytosis), cecropins (antimicrobial peptides), and lysozymes along the house fly branch. Thioester-containing proteins in particular are notable as they are likely subject to rapid adaptive evolution in Drosophilids (Sackton et al. 2007) and mosquitoes (Little and Cobbe 2005; Obbard et al. 2008), suggesting that diversification of the TEP repertoire in house flies may be particularly beneficial in broadening the diversity of pathogens they can handle.

The rapid expansion of these gene families, and others involved in recognition and effector immune functions, could be due to either selective or mutational processes, which are difficult to disentangle. It is tempting to speculate that this is driven by selection for either increased diversity or increased dosage in house flies, perhaps in response to their septic habitats. In an intriguing parallel, a high diversity of novel putative effectors is induced by LPS stimulation in the rat-tailed maggot (Altincicek and Vilcinskas 2007), which also inhabits a highly septic environment. Ultimately, however, more studies will be needed to test whether immune gene duplication rates are indeed increased generally in insects that live in particularly septic habitats.

A number of studies have suggested that the life history traits of insects can have a large impact on the gene content and structure of the immune system. Early work in honeybees suggested a depauperate immune system (Evans et al. 2006), and pea aphids appear to lack several key immune system genes (Altincicek et al. 2008; Gerardo et al. 2010). In both cases, it is possible that behavioral or life history factors (eusociality in honeybees, association with microbial symbionts in aphids) are responsible for loss of immune genes. However, recent work (Barribeau et al. 2015) suggests the reduction in immune gene content in honeybees may not be unique to social Hymenoptera. Further study will be necessary to better understand the broader connections between changes in immune gene content and external factors across diverse insect lineages.

Finally, this study confirms the pattern observed in other insects that genes induced by infection have a general tendency to be taxonomically restricted. However, what drives this pattern is still an open question. At least two hypotheses seem viable. First, it could be the case that young genes are in general less tightly regulated at the transcriptional level. As a consequence, in conditions of strong transcriptional activation (such as during an immune response), these genes have a tendency to be upregulated even without a clear function. Alternatively, this pattern could be driven by selective

recruitment of novel genes to the immune system in response to the particular challenges that diverse insect lineages experience.

Ultimately, these conclusions solidify emerging evidence that rapid host–pathogen evolutionary dynamics are not limited to rapid sequence evolution. While it is difficult to know the ultimate cause of evolutionary change, this and other recent work makes clear that insect immune systems are labile, not just at the level of protein sequence, but also at the expression level and even at the level of gene content. It seems likely that much of these rapid changes are indeed driven by host–pathogen conflict, and that the evolutionary consequences of these arms races are broader than traditionally assumed.

Materials and Methods

Data Collection
In order to detect genes induced by infection in *M. domestica*, we infected adult female flies (the genome-sequenced strain, abyss; Scott et al. 2014) 4-d post-eclosion with a 50:50 mixture (by volume of O.D. 1.0 samples) of *Serratia marcescens* and *Enterococcus faecalis*. These are same bacterial strains used in previous similar studies (Sackton et al. 2013), and were chosen to capture responses to both Gram-positive and Gram-negative infections. Bacteria were delivered by pricking the thorax with a 0.1-mm dissecting pin to penetrate the cuticle of the flies. Control flies were pricked using the same protocol, but with sterile LB broth instead of bacterial cultures. Both control and infected flies were infected between 12:00–1:00 PM in a single day and frozen in liquid nitrogen 6 h after treatment. We collected three independent biological replicates (where each replicate was a pool of five flies) for both the control and infected samples.

To inform our analysis of the transcriptional response to infection in *M. domestica*, we also generated RNA-seq data for infected and control *D. melanogaster* (iso-1 strain). For the *D. melanogaster* study, we used the same bacterial strains and concentration as above and did the experiments with females 3- to 5-day post-eclosion, but control flies remained unpricked and flies (control and infected) were frozen 12 h after treatment. As for *M. domestica*, we collected three independent biological replicates of each condition, where each replicate represents a pool of three to six flies.

Subsequently, we extracted RNA from whole frozen flies in TRIZOL following standard protocols. RNA-seq libraries were made using the Illumina TruSeq RNA sample prep kit, and sequenced (single end 50 bp) on a HiSeq 2500 platform. Raw sequencing reads are available from NCBI under BioProjects PRJNA348189 (*M. domestica*) and PRJNA348190 (*D. melanogaster*).

Updating *M. domestica* Gene Annotations
We first sought to update the existing *M. domestica* gene annotations to detect gene models that might have been missing in the initial published annotation. To do this, we used a pipeline based on the Trinity-assisted PASA workflow (Haas et al. 2003, 2011, 2013) described at http://pasa.source

forge.net/ and in more detail at the Github page associated with this manuscript (https://github.com/tsackton/musca-immunity). In brief, this pipeline takes RNA-seq reads, maps them to the genome, and looks for regions of the genome with evidence for spliced transcripts. These regions are then tested for overlap with existing gene annotations, and used to extend gene models or predict new transcripts (either novel splice forms or novel genes). We started with the *M. domestica* GFF, protein, and transcript files produced by NCBI during the initial annotation of the *Musca* genome (NCBI release 100) (Scott et al. 2014), available at ftp://ftp.ncbi.nlm.nih.gov/genomes/Musca_domestica/ARCHIVE/ANNOTATION_RELEASE.100/.

After running the PASA pipeline (https://github.com/tsackton/musca-immunity/tree/master/supplemental_methods/pasa), our primary goal was to extract novel gene annotations: we only added new gene models that may have been excluded from prior annotation, and did not to update existing gene models. The rationale for this decision is that in the absence of paired-end data or higher coverage data, determining biologically real novel splice forms is a challenging problem subject to a high false-positive rate. Thus we focused exclusively on novel gene annotations, that is, gene models predicted by PASA from Trinity alignments to the *M. domestica* genome that do not overlap existing annotations. We identified 70 new protein-coding transcript models with this approach. By definition these tend to be predicted proteins with little homology evidence (as genes with strong homology to other Dipterans would likely be annotated by the NCBI pipeline), and they are significantly shorter than previously annotated proteins (median length 223 aa vs. 389 aa, $P = 1.58 \times 10^{-6}$, Mann–Whitney U test). The transcripts encoding these novel predicted proteins tend to be more highly expressed than those encoding previously annotated proteins (adjusted count 186 vs. 98, $P = 0.00014$, Mann–Whitney U test). An updated GFF file, isoform-to-gene key, protein fasta file, and transcript fasta file are available as supplemental data and online at https://github.com/tsackton/musca-immunity/tree/master/input_data/annotations

Differential Expression Analysis
We used RSEM to quantify differential expression after infection in *M. domestica* and in *D. melanogaster*. Briefly, we first trimmed reads using Trimmomatic, then computed expression for each transcript in our updated annotation described above using RSEM v1.2.16 (Li and Dewey 2011) using bowtie2 as the read mapper. The full code to run our RSEM pipeline is available at https://github.com/tsackton/musca-immunity/tree/master/supplemental_methods/difexp, and the raw RSEM output is available in the supplemental data and at https://github.com/tsackton/musca-immunity/tree/master/input_data/rsem. To infer differential expression, we used DESeq2 (Love et al. 2014) with standard options. The full scripts for differential expression inference and related statistical analysis are available at https://github.com/tsackton/musca-immunity/tree/master/R.

Bioinformatic Characterization of Predicted *M. domestica* Proteins

We focused on characterizing three properties of *M. domestica* proteins that can be determined from sequence and comparative information: the presence of a signal peptide, the phylogenetic age of the gene, and the presence of immune-related protein domains. All scripts are available at https://github.com/tsackton/musca-immunity/tree/master/supplemental_methods.

To identify signal peptides, we used signalp v4.1 with default options run on all predicted *M. domestica* proteins.

To define phylogenetic age (specifically, phylostratigraphic age, *sensu*; Domazet-Loso et al. 2007), we started with a series of blastp searches and defined age as the node of the tree of life at which the most distant blastp hit is detectable. This is conservative in the sense that we do not screen for any kind of parsimonious pattern, so a spurious deep hit will mean we consider a protein to be ancient even in the absence of any more closely related hits. When we say a gene is young, we simply mean that no homologs can be detected by BLAST to older lineages; other factors, such as length or overall rate of sequence evolution, can thus impact gene age estimation if they increase the probability that distant homologs will be missed (Moyers and Zhang 2015). In particular, proteins that are rapidly evolving will tend to appear younger than their true age, and proteins that are short may also appear younger than their true age, due to biases inherent in detecting distant homologies of short and/or rapidly diverging sequences (Moyers and Zhang 2015). While in most cases our results focus on relatively recent homologs (i.e., within Diptera or within insects), which are likely relatively unaffected by these biases (Moyers and Zhang 2015), we also corrected for these effects (at least partially) by modeling the impact of protein length and evolutionary rate (using expression level in *M. domestica* as a proxy) on our estimates of age. Formally, we first log-transformed and mean-recentered length and expression level, and then computed model coefficients for separate regressions with either scaled expression or scaled length as the predictor variable and age as the response. These coefficients are equivalent to the change in estimated age expected for a unit deviation from the mean (on a log scale) of either expression level or length. Length is essentially uncorrelated with estimated age in our data set (Kendall's tau $= 0.02$, $P = 0.0002$), but expression level is correlated with estimated age (Kendall's tau $= 0.267$, $P < 2.2 \times 10^{-16}$). To calculate scaled ages, we computed the normalized age as the real estimate age minus the predicted effect of expression; normalized age is no longer strongly correlated with expression, as expected (Kendall's tau $= -0.03$, $P = 2.67 \times 10^{-9}$).

To define phylogenetic age, we began with a curated set of complete proteomes (listed at https://github.com/tsackton/musca-immunity/blob/master/supplemental_methods/strata/strata_key.txt) and ran blastp against each complete proteome. For each set of BLAST results (representing the best hit of each *M. domestica* protein against a target database), we considered a hit as indicating the presence of a putative homolog if the alignment length is at least 40% of the *M. domestica* protein length and the alignment has at least 20% identity. We then extracted the deepest node for which we found evidence for a putative homolog, and defined that as the phylogenetic age of each *M. domestica* protein.

In order to quantify the presence of domains that have putative immune function, we first built a set of HMM profiles based on ImmunoDB curated alignments (http://cegg.unige.ch/Insecta/immunodb) (Waterhouse et al. 2007) and additional alignments for the Nimrod domain, IGSF proteins, and transferrins based on sequences downloaded from FlyBase. The non-ImmunoDB alignments, as well as the final alignment file of all immune-related proteins, are available in the Github repository associated with this paper. We then searched the complete set of predicted *M. domestica* proteins for matches to predicted immune-related HMMs using HMMER 3.0. We then processed the HMMER output to: (1) exclude cases where the *E* value of the best domain is >0.001, (2) exclude cases where the overall *E* value is $>1 \times 10^{-5}$, and (3) assign proteins that match multiple HMMs to the single HMM with the best *E* value. To provide comparative information for the analysis of *M. domestica*, we also searched the predicted proteomes of the other Dipterans listed in supplementary table S6, Supplementary Material online, against our immune-related HMM database, and inferred the presence of domains with putative immune function using the same protocol.

Determining Orthologs and Paralogs of *M. domestica* Proteins across Dipterans

To determine patterns of orthology and paralogy of *M. domestica* proteins among Dipterans, we built a gene-tree-based pipeline for identifying gene families and determining the relationships among genes. This pipeline is described in full at https://github.com/tsackton/musca-immunity/supplemental_methods/orthology and in brief below.

First, we used OMA version 0.99 (Altenhoff et al. 2013) with default options to generate an initial set of homologous groups (HOGs), using as input the longest protein translation of each annotated protein in the *M. domestica* genome along with 13 other Dipterans (5 mosquitoes, 7 Drosophilids, and *Glossina moristans*; supplementary table S6, Supplementary Material online).

After running OMA, we refined orthogroups as follows. First we generated an alignment of each initial orthogroup using MAFFT (Katoh and Standley 2013), and then created HMMs for each group using HMMER version 3. We then refined orthogroup assignment by searching each protein against each HMM, and merging orthogroups linked by a well-supported HMM hit. We also added genes to orthogroups when a gene was not initially assigned to any group, but has a significant HMM hit to a group (see part 1 of readme at Github site).

After orthogroup updating, we realigned each orthogroup with MAFFT (–auto option) and then computed a gene tree using RAxMLHPC-SSE3 version 7.75 (Stamatakis 2014), with the default options except -m PROTGAMMAAUTO and -N 10 (see part 2 of readme at Github site).

In some cases, our pipeline led to large gene families with one or more duplications at the base of Diptera. To both increase the computational efficiency of Treefix (see below), and improve the accuracy of our rate estimation, we used a custom Perl script (treesplit.pl on Github) to split trees where the deepest node was inferred to be a duplication rather than a speciation event. After the first round of tree splitting, we used the programs Treefix (v. 1.1.8; default options except -m PROTGAMMAWAG, -niter = 1,000, and -maxtime) and tree-annotate (part of the treefix package) to reconcile the species tree with each gene tree and compute the likely number of gains and losses on the tree (Wu et al. 2012). Treefix attempts to produce the most parsimonious tree with respect to duplications and losses while remaining consistent with the maximum likelihood gene tree. It does this by searching the neighborhood of the maximum likelihood tree for topologies that reduce the number of duplication and loss events without significantly reducing the likelihood of the tree under the evolutionary model specified. Because this process is inefficient on large trees, we set a maximum time for the program to run (~1 week), which means that for large trees we sample fewer iterations than for small trees. To partially account for this, we ran a second round of tree splitting with our treesplit script after our first round of TreeFix (which led to some large trees being split into smaller trees), and then repeated treefix on any altered trees. We then ran tree-annotate to produce duplication/loss inference on this final set of trees.

Analysis of Gene Family Dynamics

To determine rates of gene duplication and loss across the phylogeny, we used both previously published, count-based methods such as CAFE (De Bie et al. 2006) and we implemented a Poisson regression model using duplication and loss events inferred from gene tree/species tree reconciliation. To account for differences in branch lengths, we constructed an ultrametric tree as follows (https://github.com/tsackton/musca-immunity/tree/master/supplemental_methods/ultrametric). First, we identified orthogroups with no duplications or losses across the phylogeny. Second, we concatenated the trimmed alignments of these orthogroups to produce a single Dipteran alignment for tree estimation. Finally, we used RAXML (version 7.7.5) with the -f e option (to estimate branch lengths on a fixed phylogeny) to estimate branch lengths from the known Dipteran phylogeny. Finally, we used the "chronos" function from the ape package in R to convert the tree to an ultrametric tree with arbitrary edge units.

To test for variation in rates of duplication and loss among different classes of genes along different lineages, we use a mixed model Poisson regression. Specifically, we fit a model which includes both fixed effects (functional class, lineage of interest), branch length as an offset, and a separate random intercept for each gene family, to control for overdispersion caused by rate variation among gene families, using the "glmer" function in the R package "lme4". R code to implement this approach, and containing the full models used for each analysis, is available at https://github.com/tsackton/musca-immunity/tree/master/R. This approach allows us to use the full power of general linear models to test hypotheses concerning lineage-specific rates of duplication.

To test our Poisson regression approach, we simulated 1,000 trees each with 1 of 12 different rates of gene duplication (assuming equal birth and death rates), ranging from 0.00057 events/MYA to 0.3412 events/MYA. To do these simulations, we fixed the species tree and estimate a gene tree within the species tree using the GuestTreeGen tool (part of jprime) with options -minper 0 -min 4 -maxper 10,000 -max 10,000 (code: https://github.com/tsackton/musca-immunity/tree/master/supplemental_methods/sims). The simulation approach of GuestTreeGen is based on a duplication-loss model where duplications and losses each occur with a specified Poisson rate along branches of a phylogeny, and speciation events result the simulated lineage splitting into two child lineages that continue to evolve by duplication and loss independently (Sjöstrand et al. 2013). After simulating data, we estimated gain/loss rates using both CAFE and our Poisson regression in order to estimate the duplication, or duplication and loss rates independently, for each simulated data set. In order to calibrate the statistical properties of our regression approach, we also simulated 1,000 datasets in which a random sample of 100 trees with different rates were selected to represent "immune genes." We then test whether we find a significant difference between rates of duplication in "immune genes" compared with "nonimmune genes", using the Poisson regression approach described earlier.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

References

Adema CM, Hertel LA, Miller RD, Loker ES. 1997. A family of fibrinogen-related proteins that precipitates parasite-derived molecules is produced by an invertebrate after infection. *Proc Natl Acad Sci U S A.* 94:8691–8696.

Altenhoff AM, Gil M, Gonnet GH, Dessimoz C. 2013. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One* 8:e53786.

Altincicek B, Gross J, Vilcinskas A. 2008. Wounding-mediated gene expression and accelerated viviparous reproduction of the pea aphid *Acyrthosiphon pisum. Insect Mol Biol.* 17:711–716.

Altincicek B, Vilcinskas A. 2007. Analysis of the immune-inducible transcriptome from microbial stress resistant, rat-tailed maggots of the drone fly Eristalis tenax. *BMC Genomics* 8:326–312.

Ayres JS, Schneider DS. 2009. The role of anorexia in resistance and tolerance to infections in *Drosophila*. *PLoS Biol*. 7:e1000150.

Bajgar A, Kucerova K, Jonatova L, Tomcala A, Schneedorferova I, Okrouhlik J, Dolezal T. 2015. Extracellular adenosine mediates a systemic metabolic switch during immune response. *PLoS Biol*. 13:e1002135.

Barribeau SM, Sadd BM, du Plessis L, Brown MJF, Buechel SD, Cappelle K, Carolan JC, Christiaens O, Colgan TJ, Erler S, et al. 2015. A depauperate immune repertoire precedes evolution of sociality in bees. *Genome Biol*. 16:83.

Bashir-Tanoli S, Tinsley MC. 2014. Immune response costs are associated with changes in resource acquisition and not resource reallocation. *Funct Ecol*. 28:1011–1019.

Boutros M, Agaisse H, Perrimon N. 2002. Sequential activation of signaling pathways during innate immune responses in Drosophila. *Dev. Cell* 3:711–722.

Buchon N, Broderick NA, Poidevin M, Pradervand S, Lemaitre B. 2009. *Drosophila* intestinal response to bacterial infection: activation of host defense and stem cell proliferation. *Cell Host Microbe* 5:200–211.

Buchon N, Silverman N, Cherry S. 2014. Immunity in *Drosophila melanogaster*–from microbial recognition to whole-organism physiology. *Nat Rev Immunol*. 14:796–810.

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

Chambers MC, Song KH, Schneider DS. 2012. *Listeria monocytogenes* infection causes metabolic shifts in *Drosophila melanogaster*. *PLoS One* 7:e50679.

De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22:1269–1271.

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

Domazet-Loso T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet*. 23:533–539.

Dong Y, Dimopoulos G. 2009. *Anopheles* fibrinogen-related proteins provide expanded pattern recognition capacity against bacteria and malaria parasites. *J Biol Chem*. 284:9835–9844.

Evans JD, Aronstein K, Chen YP, Hetru C, Imler J-L, Jiang H, Kanost M, Thompson GJ, Zou Z, Hultmark D. 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol Biol*. 15:645–656.

Exton MS. 1997. Infection-induced anorexia: active host defence strategy. *Appetite* 29:369–383.

Gerardo NM, Altincicek B, Anselme C, Atamian H, Barribeau SM, de Vos M, Duncan EJ, Evans JD, Gabaldón T, Ghanim M, et al. 2010. Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. *Genome Biol*. 11:R21–R17.

Ghosh J, Lun CM, Majeske AJ, Sacchi S, Schrankel CS, Smith LC. 2011. Invertebrate immune diversity. *Dev Comp Immunol*. 35:959–974.

Gupta AK, Nayduch D, Verma P, Shah B, Ghate HV, Patole MS, Shouche YS. 2012. Phylogenetic characterization of bacteria in the gut of house flies (*Musca domestica* L.). *FEMS Microbiol Ecol*. 79:581–593.

Gupta SK, Kupper M, Ratzka C, Feldhaar H, Vilcinskas A, Gross R, Dandekar T, Förster F. 2015. Scrutinizing the immune defence inventory of *Camponotus floridanus* applying total transcriptome sequencing. *BMC Genomics* 16:302–321.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 31:5654–5666.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 8:1494–1512.

Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Wortman JR. 2011. Approaches to fungal genome annotation. *Mycology* 2:118–141.

Harrison XA. 2014. Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ* 2:e616.

Hart BL. 1998. Biological basis of the behavior of sick animals. *Neurosci Biobehav Rev*. 12:123–137.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–170.

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

Joyner C, Mills MK, Nayduch D. 2013. *Pseudomonas aeruginosa* in *Musca domestica* L.: temporospatial examination of bacteria population dynamics and house fly antimicrobial responses. *PLoS One* 8:e79224.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. *Mol Biol Evol*. 30:772–780.

Keebaugh ES, Schlenke TA. 2012. Adaptive evolution of a novel *Drosophila* lectin induced by parasitic wasp attack. *Mol Biol Evol*. 29:565–577.

Koerich LB, Wang X, Clark AG, Carvalho AB. 2008. Low conservation of gene content in the *Drosophila* Y chromosome. *Nature* 456:949–951.

Kosiol C, Vinař T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet*. 4:e1000144.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in *Drosophila*. *Trends Genet*. 24:114–123.

Lazzaro BP. 2008. Natural selection on the *Drosophila* antimicrobial immune system. *Curr Opin Microbiol*. 11:284–289.

Lee KP, Cory JS, Wilson K, Raubenheimer D, Simpson SJ. 2006. Flexible diet choice offsets protein costs of pathogen resistance in a caterpillar. *Proc B* 273:823–829.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323.

Little TJ, Cobbe N. 2005. 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.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15:550.

Medzhitov R, Schneider DS, Soares MP. 2012. Disease tolerance as a defense strategy. *Science* 335:936–941.

Moret Y, Schmid-Hempel P. 2000. Survival for immunity: the price of immune system activation for bumblebee workers. *Science* 290:1166–1168.

Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol*. 32:258–267.

Nayduch D, Cho H, Joyner C. 2013. Staphylococcus aureus in the house fly: temporospatial fate of bacteria and expression of the antimicrobial peptide defensin. *J Med Entomol*. 50:171–178.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*. 3:e170–e110.

Obbard DJ, Callister DM, Jiggins FM, Soares DC, Yan G, Little TJ. 2008. The evolution of TEP1, an exceptionally polymorphic immunity gene in *Anopheles gambiae*. *BMC Evol Biol*. 8:274.

Obbard DJ, Welch JJ, Kim K-W, Jiggins FM. 2009. Quantifying adaptive evolution in the *Drosophila* immune system. *PLoS Genet*. 5:e1000698–e1000613.

Pace KE, Lebestky T, Hummel T, Arnoux P, Kwan K, Baum LG. 2002. Characterization of a novel *Drosophila melanogaster* galectin. Expression in developing immune, neural, and muscle tissues. *J Biol Chem*. 277:13091–13098.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genetics* 158:927–931.

Romero A, Dios S, Poisa-Beiro L, Costa MM, Posada D, Figueras A, Novoa B. 2011. Individual sequence variability and functional activities of fibrinogen-related proteins (FREPs) in the Mediterranean mussel (*Mytilus galloprovincialis*) suggest ancient and complex immune recognition models in invertebrates. *Dev Comp Immunol.* 35:334–344.

Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. *Mol Biol Evol.* 31:1661–1685.

Sackton TB, Clark AG. 2009. Comparative profiling of the transcriptional response to infection in two species of *Drosophila* by short-read cDNA sequencing. *BMC Genomics* 10:259–217.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in *Drosophila*. *Nature Genetics* 39:1461–1468.

Sackton TB, Werren JH, Clark AG. 2013. Characterizing the infection-induced transcriptome of *Nasonia vitripennis* reveals a preponderance of taxonomically-restricted immune genes. *PLoS One* 8:e83984–e83911.

Schlenke TA, Begun DJ. 2003. Natural selection drives *Drosophila* immune system evolution. *Genetics* 164:1471–1480.

Schneider DS, Ayres JS. 2008. Two ways to survive infection: what resistance and tolerance can teach us about treating infectious diseases. *Nat Rev Immunol.* 8:889–895.

Scott JG, Liu N, Kristensen M, Clark AG. 2009. A case for sequencing the genome of *Musca domestica* (Diptera: Muscidae). *J Med Entomol.* 46:175–182.

Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, Hediger M, Jones AK, Kasai S, Leichter CA, et al. 2014. Genome of the house fly, *Musca domestica* L., a global vector of diseases with adaptations to a septic environment. *Genome Biol.* 15:466.

Sjöstrand J, Arvestad L, Lagergren J, Sennblad B. 2013. GenPhyloData: realistic simulation of gene family evolution. *BMC Bioinformatics* 14:209.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Tiffin P, Moeller DA. 2006. Molecular evolution of plant immune system genes. *Trends Genet.* 22:662–670.

Unckless RL, Rottschaefer SM, Lazzaro BP. 2015. The complex contributions of genetics and nutrition to immunity in *Drosophila melanogaster*. *PLoS Genet.* 11:e1005030.

Vasta GR. 2009. Roles of galectins in infection. *Nat Rev Microbiol.* 7:424–438.

Vizioli J, Bulet P, Hoffmann JA, Kafatos FC, Müller HM, Dimopoulos G. 2001. Gambicin: a novel immune responsive antimicrobial peptide from the malaria vector *Anopheles gambiae*. *Proc Natl Acad Sci U S A.* 98:12630–12635.

Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, et al. 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* 316:1738–1743.

Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, The Nasonia Genome Working Group, Beukeboom LW, Desplan C, Elsik CG, et al. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* 327:343–348.

Wu YC, Rasmussen MD, Bansal MS, Kellis M. 2012. TreeFix: statistically informed gene tree error correction using species trees. *Syst Biol.* 62:110–120.