Sociality Is Linked to Rates of Protein Evolution in a Highly Social Insect

Brendan G. Hunt,[1] Stefan Wyder,[2,3] Navin Elango,[1] John H. Werren,[4] Evgeny M. Zdobnov,[2,3] Soojin V. Yi,*†,[1] and Michael A.D. Goodisman*†,[1]

[1]School of Biology, Georgia Institute of Technology

[2]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland

[3]Swiss Institute of Bioinformatics, Rue Michel-Servet, Geneva, Switzerland

[4]Department of Biology, University of Rochester

†These authors contributed equally to this work.

***Corresponding author:** E-mail: soojin.yi@gatech.edu; michael.goodisman@biology.gatech.edu.
Associate editor: David Irwin

Abstract

Eusocial insects exhibit unparalleled levels of cooperation and dominate terrestrial ecosystems. The success of eusocial insects stems from the presence of specialized castes that undertake distinct tasks. We investigated whether the evolutionary transition to societies with discrete castes was associated with changes in protein evolution. We predicted that proteins with caste-biased gene expression would evolve rapidly due to reduced antagonistic pleiotropy. We found that queen-biased proteins of the honeybee *Apis mellifera* did indeed evolve rapidly, as predicted. However, worker-biased proteins exhibited slower evolutionary rates than queen-biased or nonbiased proteins. We suggest that distinct selective pressures operating on caste-biased genes, rather than a general reduction in pleiotropy, explain the observed differences in evolutionary rates. Our study highlights, for the first time, the interaction between highly social behavior and dynamics of protein evolution.

Key words: *Apis mellifera*, caste, comparative genomics, eusocial, evolutionary rate, protein evolution.

Eusocial insects, which include ants, termites, some bees, and some wasps, exhibit unparalleled cooperation: Individuals act in distinct roles to increase colony-level success (Hamilton 1964). At the heart of this cooperation lies a division of labor among castes. Specifically, queens and males reproduce, whereas workers and soldiers engage in tasks related to brood rearing and colony defense (Wilson 1971). Thus, individual fitness is deeply intertwined within a colony, marking a major evolutionary transition in biological organization (Maynard Smith and Szathmary 1995). Eusociality has also proven to be immensely successful in ecological terms. Eusocial insects represent only 2% of insect taxa but may account for more than half of the total insect biomass (Fittkau and Klinge 1973; Wilson 1990).

Eusocial insect taxa with environmental caste determination serve as key examples of polyphenism (Wheeler 1986) and are well suited to the study of phenotypic plasticity and evolution. Advances in molecular biology have facilitated a wealth of insight into the molecular basis of caste polyphenisms (Goodisman et al. 2008; Kucharski et al. 2008; Smith et al. 2008). However, many questions remain concerning the link between eusocial evolution and the molecular processes driving caste polyphenisms. One major unanswered question is how the evolution of specialized castes has shaped protein evolution.

In general, a gene that functions in multiple phenotypes (e.g., sexes or tissues) can exhibit antagonistic fitness effects (Chippindale et al. 2001; Bonduriansky and Chenoweth 2009). When the expression of such a gene becomes limited to a single phenotype, its function may become more specialized, breaking antagonistic links. Thus, genes with phenotype-specific expression may undergo a reduction in pleiotropy, facilitating increased molecular evolutionary change through selection and drift (Winter et al. 2004; Ellegren and Parsch 2007). In accord with this idea, Gadagkar (1997) hypothesized that the divergence of caste phenotypes in eusocial organisms would reduce pleiotropic constraint and lead to "genetic release," which in turn would facilitate diversifying evolution and drive caste specialization (Gadagkar 1997). Here, we empirically address this hypothesis by investigating whether genes with caste-biased expression show elevated rates of protein evolution in a eusocial insect.

To determine evolutionary rates for caste-biased genes, we used gene expression data from brains of adult queens and workers of the eusocial bee, *Apis mellifera*, obtained by Grozinger et al. (2007). We identified orthologs of 1,511 genes from this data set in the noneusocial parasitoid wasp *Nasonia vitripennis* and one to three additional insect species (supplementary fig. S1, Supplementary Material online). Of these genes, 958 were not significantly biased in expression according to caste, 231 were worker biased (i.e., significantly more highly expressed in workers than in queens), and 322 were queen biased. Protein evolutionary rates for all *A. mellifera* genes and corresponding *N. vitripennis* orthologs were determined using protein phylogenies (fig. 1, supplementary fig. S2, supplementary table S1, Supplementary Material online).

Letter

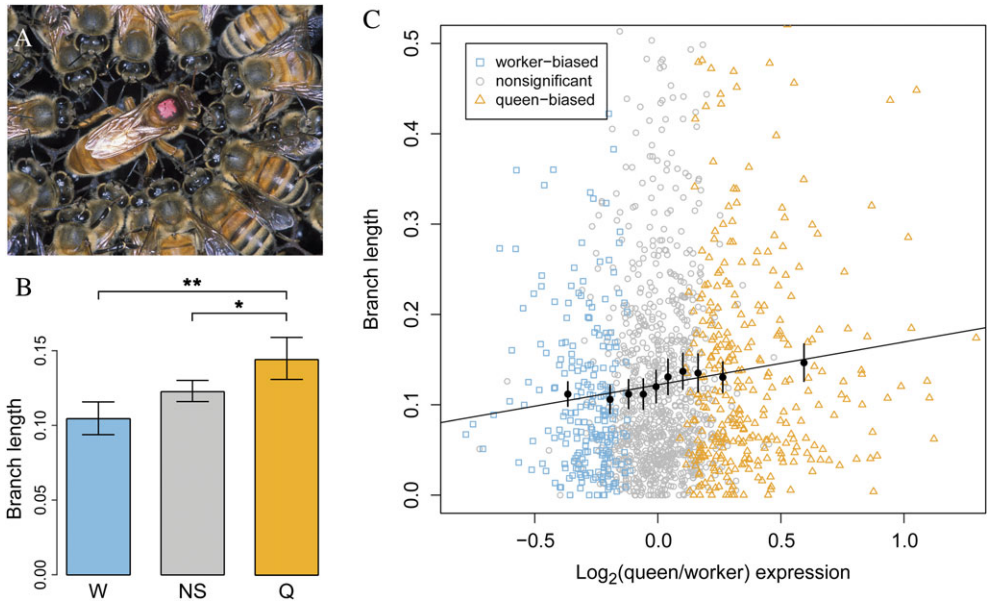

Fig. 1. Caste-biased gene expression is linked to protein evolutionary rate in *Apis mellifera*. (*A*) *Apis mellifera* workers surround a queen. (*B*) *Apis mellifera* evolutionary rates (branch lengths in amino acid substitutions per site) differ significantly among genes with worker-biased expression (W), nonsignificant bias (NS), and queen-biased expression (Q; Kruskal–Wallis $P = 0.0019$). Means with 95% confidence intervals are plotted, and significant differences are indicated (*$P < 0.05$ and **$P < 0.01$ pairwise Mann–Whitney *U* test with Bonferroni correction). (*C*) Log_2-transformed ratios of queen-to-worker gene expression are correlated with *A. mellifera* evolutionary rates (Spearman's rank correlation $r_s = 0.096$, $P = 0.0002$). A linear model best-fit line is plotted, and mean values for 10 equally sized bins of genes are shown as black dots with 95% confidence intervals. Outliers beyond the scaled axes contribute to plotted means and confidence intervals. Part *A* photo by Scott Bauer, the United States Department of Agriculture-Agricultural Research Service.

Proteins with queen-biased expression in *A. mellifera* exhibited significantly higher evolutionary rates than proteins with nonsignificant bias or worker bias (fig. 1*B*). In contrast, worker-biased proteins did not evolve at elevated rates (fig. 1*B*). In fact, worker-biased proteins had the lowest rates of amino acid substitution of the three gene expression classes, suggesting there is no overall positive relationship between caste specificity and evolutionary rate.

Protein evolutionary rate was not only significantly correlated with the ratio of queen-to-worker gene expression but was also strongly associated with several other factors (table 1; supplementary discussion; Pal et al. 2006). For example, there was a strong positive correlation between evolutionary rates in *A. mellifera* and noneusocial *N. vitripennis*. This suggests that protein evolution is heavily influenced by selective pressures shared with a noneusocial common ancestor (tables 1 and 2) and reveals that proteins with different evolutionary rates may have been co-opted for queen specialization and worker specialization during the origin and elaboration of eusociality. To further test this hypothesis, we analyzed the propensity of gene loss (PGL) in highly divergent eukaryotic orthologs of *A. mellifera* (Wolf et al. 2006). We found that orthologs of queen-biased genes were more likely to be lost during the course of evolution than worker-biased or nonbiased genes (table 3). This suggests that intrinsic properties of queen-biased orthologs strongly contribute to their evolutionary rates in *A. mellifera*.

In order to test whether queen-biased proteins were subject to additional rate increases specific to the *A. mellifera* lineage, we used multivariate analyses to control for shared effects on *A. mellifera* and *N. vitripennis*. Partial correlations (Kim and Yi 2006, 2007) showed that *A. mellifera* protein evolutionary rates were significantly correlated with the ratio of queen-to-worker gene expression when controlling for *N. vitripennis* branch length and several gene characteristics

Table 1. Spearman's Rank Correlation Coefficients (r_s) and Partial Correlations between *Apis mellifera* Evolutionary Rates[a] and Selected Gene Attributes.

Variable (*X*)	Correlation r_s *X*, *A. mellifera* branch length	Partial Correlation r_s *X*, *A. mellifera* branch length|all other variables
***Nasonia vitripennis* branch length**[a]	**0.706*******	**0.702*******
Synonymous third codon position GC content	**−0.177*******	**−0.118*****
Log_2(queen/worker) gene expression[b]	**0.096*****	**0.083****
Effective number of codons	**−0.142******	**0.059***
Coding sequence length	**−0.025**	**−0.015**
Brain expression level[c]	**−0.021**	**−0.003**

*$P < 0.05$; **$P < 0.01$; ***$P < 10^{-3}$; ****$P < 10^{-6}$; and *****$P < 10^{-9}$.

[a]Evolutionary rates are measured as branch lengths in units of amino acid substitutions per site.

[b]The log_2 ratio of queen-to-worker gene expression is significantly correlated with *A. mellifera* branch lengths when controlling for the combined effect of *N. vitripennis* branch lengths (a proxy for shared ancestral evolutionary determinants) and several *A. mellifera* gene characteristics associated with evolutionary rates.

[c]Mean of normalized gene expression levels in brains of queens and workers (see Methods).

Table 2. Principal Component Regression Analysis of *Apis mellifera* Evolutionary Rates and Selected Gene Attributes.

	Principal components[a]				
	1	2	3	4	All (6)
Percent variance explained in *A. mellifera* branch length[b]	**2.48*******	**0.90*****	**22.05*******	**4.17*******	**29.82*******
Percent contributions to principal components[c]					
Log_2(queen/worker) gene expression	**0.2**	**26.8**	**19.6**	**51.5**	
***Nasonia vitripennis* branch length**	**1.0**	**1.0**	**66.1**	**31.8**	
Effective number of codons	**44.2**	**3.0**	**0.3**	**3.3**	
Synonymous third codon position GC content	**45.4**	**2.5**	**0**	**1.2**	
Coding-sequence length	**6.7**	**26.3**	**8.1**	**9.3**	
Brain expression level[d]	**2.6**	**40.3**	**5.8**	**3.0**	

***$P < 10^{-3}$ and *****$P < 10^{-9}$.

[a]Principal components are numbered in the order of highest to lowest contribution to the variance of the independent variable in the partial correlation regression. Principal components 5 and 6 were not included because they do not contribute to the dependent variable, *A. mellifera* branch length, at a threshold of $P < 0.05$.

[b]Evolutionary rates are measured as branch lengths in units of amino acid substitutions per site.

[c]Bold indicates variables that contribute at least 10% to the principal component. *Nasonia vitripennis* branch length and the log_2 ratio of queen-to-worker gene expression make the greatest contributions to *A. mellifera* branch length.

[d]Mean of normalized gene expression levels in brains of queens and workers (see Methods).

associated with evolutionary rates in other taxa (table 1). The ratio of queen-to-worker gene expression was also a large contributor to principal components that significantly explained variance in *A. mellifera* branch lengths in our principal component regression analysis (table 2; Drummond et al. 2006). Together, these analyses suggest that queen-biased proteins incurred an additional rate increase in the *A. mellifera* lineage, during which queen and worker castes diverged.

Next, we investigated whether evolutionary rates of proteins associated with caste differences were associated with protein function. Although genes expressed in the brain are tightly linked to behavior (Robinson et al. 2008), they also represent diverse biological processes with far-reaching phenotypic consequences (Whitfield et al. 2002). Our gene ontology analysis revealed that many rapidly evolving queen-biased genes were involved in metabolic function (supplementary table S2, Supplementary Material online). This finding is bolstered by evidence that metabolic regulation is related to nutritional caste differences (Cristino et al. 2006; Grozinger et al. 2007; Hoffman and Goodisman 2007). Accordingly, the evolution of metabolic functions may help to explain queen-biased evolutionary rate increases.

Accelerated rates of evolution previously observed in sex-biased proteins and now observed in queen-biased proteins may be driven by similar processes (Ellegren and Parsch 2007). As in sexual dimorphism, caste dimorphism may give rise to ontogenetic conflict between phenotypic optima, which can in turn be resolved through caste-biased gene expression (Chippindale et al. 2001; Proschel et al. 2006; Haerty et al. 2007; Bonduriansky and Chenoweth 2009). This is the case because *A. mellifera* queen and worker castes are affected by selection in fundamentally different ways. Workers rely predominantly (but not exclusively) on indirect fitness by helping to rear queen-produced offspring, whereas queens rely on direct fitness components. Queens, like solitary females, are subject to sexual selection and evolutionary pressure for high fecundity, whereas workers are selected for their distinct roles, such as foraging (Wilson 1971). Our results are consistent with a scenario in which queen-biased proteins undergo adaptive evolution related to reproductive physiology and associated evolutionary arms races, whereas worker-biased proteins do not.

Alternatively, queen-biased proteins, like sex-biased proteins, may have had historically high levels of dispensability or few pleiotropic constraints relative to worker-biased and nonbiased proteins, causing an increase in evolutionary rates (Mank et al. 2008; Mank and Ellegren 2009). Subsequent to caste divergence, queen-biased proteins may have also undergone further reductions in pleiotropy, causing an additional increase in evolutionary rates. For example, queens of swarm founding species, which include *A. mellifera*, may have retained fewer ancestral behaviors related to foraging and maternal care than workers because they are assisted by workers throughout their life cycle (Peeters and Ito 2001; Linksvayer and Wade 2005).

If caste divergence is universally associated with a release of pleiotropic constraints or historical dispensability, these effects are overshadowed by the influence of caste-specific selective pressures. Our results suggest that queen-biased genes, in particular, were historically fast evolving and may

Table 3. Propensity of Gene Loss (PGL) for Eukaryotic Clusters of Orthologous Groups (KOGs) That Include *Drosophila melanogaster* Orthologs of *Apis mellifera* Proteins.

Mean KOG PGL[a] ± Standard Error of Mean			Rank Sum Test *P* Value			
Worker-Biased	**Nonsignificant**	**Queen-Biased**	**Overall[b]**	**P_{WN}[c]**	**P_{NQ}[c]**	**P_{WQ}[c]**
0.0862 ± 0.0155	**0.0964 ± 0.0080**	**0.1310 ± 0.0130**	**0.0328***	**1.0000**	**0.0637**	**0.0829**

*P value < 0.05.

[a]See Wolf et al. (2006).

[b]P value from Kruskal–Wallis rank sum test of worker-biased, nonsignificant, and queen-biased genes.

[c]P value from Bonferroni-corrected Mann–Whitney U test of worker-biased versus nonsignificant (P_{WN}), nonsignificant versus queen-biased (P_{NQ}), and worker-biased versus queen-biased (P_{WQ}). P_{NQ} and P_{WQ} were < 0.05 prior to conservative Bonferroni correction.

have attained an additional increase in evolutionary rate following caste divergence in *A. mellifera*. We have highlighted adaptive and nonadaptive explanations for these relationships that demand further investigation as comparative genomic resources expand.

Supplementary Material

Supplementary tables S1–S5, supplementary figures S1 and S2, methods, discussion, and references are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank C.M. Grozinger for providing assistance with microarray data analysis, E.V. Kriventseva for sharing orthology data, Y.I. Wolf for sharing PGL data, and T.W. Cunningham, I.K. Jordan, J.L. Kovacs, and three anonymous reviewers for comments that improved the manuscript. This research was supported by the US National Science Foundation (grant DEB 0640690 to M.G. and S.Y.) and the Swiss National Science Foundation (grant SNF 3100A0-112588 to E.Z.).

References

Bonduriansky R, Chenoweth SF. 2009. Intralocus sexual conflict. *Trends Ecol Evol*. 24:280–288.

Chippindale AK, Gibson JR, Rice WR. 2001. Negative genetic correlation for adult fitness between sexes reveals ontogenetic conflict in *Drosophila*. *Proc Natl Acad Sci USA*. 98:1671–1675.

Cristino AS, Nunes FMF, Lobo CH, Bitondi MMG, Simoes ZLP, Costa LD, Lattorff HMG, Moritz RFA, Evans JD, Hartfelder K. 2006. Caste development and reproduction: a genome-wide analysis of hallmarks of insect eusociality. *Insect Mol Biol*. 15:703–714.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*. 23:327–337.

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet*. 8:689–698.

Fittkau EJ, Klinge H. 1973. On biomass and trophic structure of the central amazonian rain forest ecosystem. *Biotropica*. 5:2–14.

Gadagkar R. 1997. The evolution of caste polymorphism in social insects: genetic release followed by diversifying evolution. *J Genet*. 76:167–179.

Goodisman MAD, Kovacs JL, Hunt BG. 2008. Functional genetics and genomics in ants (Hymenoptera: Formicidae): the interplay of genes and social life. *Myrmecol News*. 11:107–117.

Grozinger CM, Fan YL, Hoover SER, Winston ML. 2007. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (*Apis mellifera*). *Mol Ecol*. 16:4837–4848.

Haerty W, Jagadeeshan S, Kulathinal RJ, et al. (11 co-authors). 2007. Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. *Genetics* 177:1321–1335.

Hamilton WD. 1964. The genetical evolution of social behaviour. II. *J Theor Biol*. 7:17–52.

Hoffman EA, Goodisman MAD. 2007. Gene expression and the evolution of phenotypic diversity in social wasps. *BMC Biol*. 5:23.

Kim SH, Yi SV. 2006. Correlated asymmetry of sequence and functional divergence between duplicate proteins of *Saccharomyces cerevisiae*. *Mol Biol Evol*. 23:1068–1075.

Kim SH, Yi SV. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. *Genetica*. 131: 151–156.

Kucharski R, Maleszka J, Foret S, Maleszka R. 2008. Nutritional control of reproductive status in honeybees via DNA methylation. *Science* 319:1827–1830.

Linksvayer TA, Wade MJ. 2005. The evolutionary origin and elaboration of sociality in the aculeate Hymenoptera: maternal effects, sib-social effects, and heterochrony. *Q Rev Biol*. 80:317–336.

Mank JE, Ellegren H. 2009. Are sex-biased genes more dispensable? *Biol Lett*. 5:409–412.

Mank JE, Hultin-Rosenberg L, Zwahlen M, Ellegren H. 2008. Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression. *Am Nat*. 171:35–43.

Maynard Smith J, Szathmary E. 1995. The major transitions in evolution. New York: Oxford University Press.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet*. 7:337–348.

Peeters C, Ito F. 2001. Colony dispersal and the evolution of queen morphology in social Hymenoptera. *Annu Rev Entomol*. 46:601–630.

Proschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. *Genetics*. 174: 893–900.

Robinson GE, Fernald RD, Clayton DF. 2008. Genes and social behavior. *Science* 322:896–900.

Smith CR, Toth AL, Suarez AV, Robinson GE. 2008. Genetic and genomic analyses of the division of labour in insect societies. *Nat Rev Genet*. 9:735–748.

Wheeler DE. 1986. Developmental and physiological determinants of caste in social Hymenoptera: evolutionary implications. *Am Nat*. 128:13–34.

Whitfield CW, Band MR, Bonaldo MF, Kumar CG, Liu L, Pardinas JR, Robertson HM, Soares MB, Robinson GE. 2002. Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Res*. 12:555–566.

Wilson EO. 1971. The insect societies. Cambridge: Harvard University Press.

Wilson EO. 1990. Success and dominance in ecosystems: the case of the social insects. Oldendorf/Luhe (Germany): Ecology Institute.

Winter EE, Goodstadt L, Ponting CP. 2004. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res*. 14:54–61.

Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. *Proc R Soc B*. 273:1507–1515.

Sociality is linked to rates of protein evolution in a highly social insect

Brendan G. Hunt, Stefan Wyder, Navin Elango, John H. Werren, Evgeny M. Zdobnov, Soojin V. Yi, Michael A. D. Goodisman

Supplementary Material

Methods

Identification of Apis mellifera caste-biased genes

Genes differentially expressed in *A. mellifera* brains of adult queens and adult sterile workers were identified previously by Grozinger et al. (2007) using cDNA microarrays. We obtained normalized relative expression levels and *P*-values for 2,684 genes (excluding information from untranslated regions) from CM Grozinger. Significance designations from Grozinger et al.'s supplementary material (2007) were used to assign genes to the following classes based on relative expression levels: (i) nonsignificant difference in expression between *A. mellifera* queen and worker castes (nonsignificant or non-biased), (ii) significantly higher expression in workers than queens (worker-biased), and (iii) significantly higher expression in queens than workers (queen-biased). The ratios of queen to worker gene expression levels were $\log_2$ transformed for further analysis.

Assignment of orthologous proteins

Proper assignment of orthology is a necessary precondition for comparative sequence analysis and measurement of relative evolutionary rates. We identified orthologous genes in five insect

species with sequenced genomes and robust global protein sequence data: the eusocial bee *A. mellifera*, the non-eusocial parasitoid wasp *Nasonia vitripennis,* the beetle *Tribolium castaneum*, the fly *Drosophila melanogaster*, and the louse *Pediculus humanus* (Fig. *S1*). Protein sets were retrieved from Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu) for *A. mellifera* (release 1) and *T. castaneum*, from FlyBase (http://www.flybase.org) for *D. melanogaster*, and from VectorBase (http://www.vectorbase.org) for *P. humanus*. Orthologs were determined using 27,403 NCBI predicted *N. vitripennis* genes (RefSeq together with Gnomon *ab initio* models) following the OrthoDB method described by Kriventseva et al. (2008). We identified 4,836 orthologous groups with a 1:1 relationship in *A. mellifera*, *N. vitripennis*, and 1-3 of the outgroup insect species listed above. 1,511 of these orthologous groups included *A. mellifera* genes with expression data from Grozinger et al. (2007).

Estimates of divergence times between insect taxa in our analyses (Fig. *S1*) were taken from the literature (Grimaldi and Engel 2005; Nasonia Genome Sequencing Consortium *in preparation*). A basal position of Hymenoptera relative to the Coleoptera in the holometabolous insect tree is supported by multiple independent genome characters (Savard et al. 2006; Krauss et al. 2008).

Determination of evolutionary rates

We used comparisons of *N. vitripennis* and *A. mellifera* (both from the order Hymenoptera) with outgroup taxa to determine evolutionary rates. We obtained evolutionary rate measures by first generating multiple protein alignments and then generating phylogenies for each orthologous group as follows. Multiple protein alignments were generated using MUSCLE with default settings (Edgar 2004). Confidently aligned gap-free columns were then extracted using Gblocks

(Talavera and Castresana 2007), and only long alignments (≥ 100 amino acids) were kept for analysis. Individual phylogenies were generated using the maximum likelihood method implemented in PHYML using the JTT model of amino acid substitution to correct for multiple substitutions and a gamma distribution over four rate categories to account for variable substitution rates among sites (Guindon and Gascuel 2003). Trees composed of 3-5 species were rooted with an outgroup (*P. humanus*, *D. melanogaster*, or *T. castaneum*) and branch lengths were extracted in units of amino acid substitutions per site. *N. vitripennis* and *A. mellifera* terminal branch lengths were used to compare evolutionary rates between genes and taxa (Table *S1*).

Propensity of gene loss

We examined the propensity of gene loss (PGL) in orthologous groups (KOGs; Tatusov et al. 2003), which include seven highly divergent eukaryotic taxa. PGL values were calculated previously by Wolf, Carmel, and Koonin (2006). To assign KOGs to *A. mellifera*, we first identified orthologous proteins between *A. mellifera* and *D. melanogaster*. *A. mellifera* official gene set identifiers were converted to protein GI accessions using the gene_info database from the NCBI FTP site (http://www.ncbi.nlm.nih.gov/ftp/). Next, *D. melanogaster* orthologs of *A. mellifera* genes were downloaded from the Roundup database of orthology (DeLuca et al. 2006), which uses the robust reciprocal smallest distance algorithm for ortholog determination. KOG PGL values were assigned to 225 nonsignificantly biased proteins, 94 queen-biased proteins, and 59 worker-biased proteins. A Kruskal-Wallis test was then used to determine if PGL differed significantly among protein categories.

Apis mellifera gene attributes

We downloaded protein-coding nucleotide sequences for *A. mellifera* from BeeBase (release 1; http://www.beebase.org). We determined several characteristics of each gene using the software package codonW (http://codonw.sourceforge.net) to assess whether these characteristics influenced evolutionary rate. For each gene, we determined synonymous third codon position GC content (Jorgensen, Schierup, and Clark 2007), effective number of codons (Wright 1990; Duret and Mouchiroud 1999), and coding sequence length (Lemos et al. 2005).

The magnitude of gene expression levels in brains of females was estimated from microarray data obtained by Grozinger et al. (2007). Background intensities for each microarray spot were first subtracted from median spot intensities for seven two-dye microarrays representing hybridizations of cDNA from pooled sterile worker brains and pooled queen brains (see Grozinger et al. 2007 for detailed experimental design). We used the 'maanova' package (version 1.14.0) in R to $\log_2$ transform the data in order to stabilize variance in high intensity spots. Joint lowess smoothing was then applied to normalize differences in intensities arising from microarray spatial heterogeneity (Cui, Kerr, and Churchill 2003). Mean of duplicate spots within arrays were then taken as a measure of intensity for each gene. Next, data was normalized across microarrays by quantile normalization, as implemented in the 'preprocessCore' package in R (Bolstad et al. 2003). Finally, the expression level of each gene in queens and workers was calculated as the median of normalized values across all microarrays. The brain gene expression value used in subsequent statistical analyses was taken as the mean of queen and worker expression levels. We note that the peak expression level between queens and workers was highly correlated with the average of queens and workers (Spearman's rank

correlation > 0.99, $P < 10^{-15}$) and using peak expression levels, as opposed to mean expression levels, would thus not significantly alter our results.

Analysis of evolutionary rate correlates

All statistical analyses of evolutionary rates were performed using R (R Development Core Team 2008). We first used partial correlations to test whether caste-biased gene expression is associated with *A. mellifera* protein evolutionary rates while controlling for other gene characteristics (see above) and *N. vitripennis* evolutionary rates. We then used principal component regression for the same purpose. We used R code from Drummond et al. (2006) supplementary material and the 'pls' R package to perform principal component regression. Variables other than *A. mellifera* branch length, which was not used for determining principal components, were standardized to zero mean and unit variance prior to principal component regression.

Analysis of gene ontology functional terms

We investigated associations between gene function, evolutionary rates, and queen- or worker-bias in gene expression using biological process gene ontology terms (Gene Ontology Consortium 2000). Specifically, we tested whether there were gene ontology terms that were (i) overrepresented in queen-biased, nonsignificant, or worker-biased genes compared to a background population of all the genes, and (ii) overrepresented in one of five equally-sized 'bins' of genes with similar evolutionary rates compared to a background population of all the genes. The intersection of significant gene ontology terms in these two analyses revealed

functions that are enriched in a caste-biased class and a particular evolutionary rate class (e.g., rapidly evolving queen-biased genes; Table *S2*).

We used *D. melanogaster* orthologs of *A. mellifera* genes in this analysis because the *D. melanogaster* genome is better annotated. Analysis of overrepresentation was performed using the DAVID bioinformatics database functional annotation tool (Dennis et al. 2003). A modified Fisher Exact *P*-Value called the EASE score was used to determine statistical significance of overrepresentation for a given gene ontology term in a given group compared to the background population at a threshold of $P < 0.05$ (Hosack et al. 2003). For analysis of overrepresentation according to evolutionary rate, we assigned genes to five equal bins according to increasing evolutionary rate (302-303 genes each). The numbers of *D. melanogaster* orthologs in each bin, from lowest to highest evolutionary rate, were 220, 223, 222, 213, and 202, respectively. The numbers of *D. melanogaster* orthologs for each gene expression class were 156 for worker-biased genes, 698 for nonsignificant genes, and 226 for queen-biased genes.

Alternate phylogenetic analysis

For our primary analysis of evolutionary rates, we constructed protein phylogenies that included between three and five insect species (Fig. *S1*). In order to ensure that our conclusions were not influenced by differences in the taxonomic composition of phylogenies, we repeated our analysis using only orthologous groups composed of all five insect species. We used 370 orthologous groups with 1:1 relationships among all five species. The results of this conservative approach corroborated the findings of our primary analysis (Table *S3*, Table *S4*, Table *S5*, supplementary discussion).

Discussion

Synonymous third codon GC content (GC3s) and effective number of codons (Nc) were significantly negatively correlated with protein evolutionary rate in *A. mellifera* (Table 2). This result is somewhat surprising because codon usage bias (reflected by lower Nc values) is generally positively correlated with overall gene expression levels (Duret and Mouchiroud 1999) and negatively correlated with evolutionary rate (Drummond et al. 2005; Drummond, Raval, and Wilke 2006; Larracuente et al. 2008). However, we find the opposite pattern with respect to Nc and evolutionary rate. This correlation may arise because Nc is heavily affected by mutational bias (Wright 1990), which is evident in GC-poor regions of the *A. mellifera* genome (Jorgensen, Schierup, and Clark 2007). As a result, Nc may not be a reliable proxy of overall gene expression level in *A. mellifera*. Another possible contributor to the observed positive correlation between codon usage bias and evolutionary rate may be queen-biased gene expression. Codon usage bias is positively correlated with the degree of female-biased gene expression in *D. melanogaster* (Hambuch and Parsch 2005). Queen-biased genes, which evolve rapidly, may bear a similar relationship to codon usage bias because queens are reproductive females.

To more directly test for the effect of gene expression level, we calculated absolute gene expression levels in adult female brains using microarray data (Grozinger et al. 2007). Surprisingly, expression levels were not correlated with evolutionary rates in our analyses (Table 1). Queen-biased genes also exhibited significantly higher expression levels than worker-biased or non-biased genes ($P < 0.001$, pairwise Bonferroni corrected Mann-Whitney *U* tests). However, we caution that more data from additional tissues and developmental stages are

necessary to sufficiently test whether highly expressed genes evolve slowly in *A. mellifera*, as has been observed in other eukaryotes (Drummond, Raval, and Wilke 2006; Pal, Papp, and Lercher 2006; Larracuente et al. 2008).

An interesting result from our primary principal component regression analysis was the covariance of the ratio of queen to worker gene expression, brain gene expression level, and coding sequence length (Table 2). This finding highlights the possibility of a number of characteristics that may be specific to caste-biased genes, warranting further study (e.g., Elango et al 2009). In our principal component regression analysis using only genes with five-species orthologous groups, we found that the ratio of queen to worker gene expression had an effect independent of *N. vitripennis* branch length on evolutionary rate (Table *S5*). This finding strongly supports our interpretation that queen-biased proteins of the *A. mellifera* lineage have undergone an evolutionary rate increase subsequent to divergence from a non-eusocial common ancestor shared with *N. vitripennis*. This alternate principal component regression analysis also identified a link between gene expression level and queen to worker gene expression ratio (note that queen-biased genes are more highly expressed than other classes of genes; see main text).

References

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. **19**:185-193.

Cui, X., M. K. Kerr, and G. A. Churchill. 2003. Transformations for cDNA microarray data. Stat Appl Genet Mol Biol. **2**:Article 4.

DeLuca, T. F., I. H. Wu, J. Pu, T. Monaghan, L. Peshkin, S. Singh, and D. P. Wall. 2006. Roundup: a multi-genome repository of orthologs and evolutionary distances. Bioinformatics. **22**:2044-2046.

Dennis, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. 2003. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol. **4**:R60.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci USA. **102**:14338-14343.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. **23**:327-337.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, *Arabidopsis*. Proc Natl Acad Sci USA. **96**:4482-4487.

Edgar, R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792-1797.

Elango, N., B. G. Hunt, M. A. D. Goodisman, and S. V. Yi. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. Proc Natl Acad Sci U S A. **106**:11206-11211.

Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. **25**:25-29.

Grimaldi, D., and M. Engel. 2005. Evolution of the Insects. Cambridge University Press, Cambridge.

Grozinger, C. M., Y. L. Fan, S. E. R. Hoover, and M. L. Winston. 2007. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (*Apis mellifera*). Mol Ecol. **16**:4837-4848.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. **52**:696-704.

Hambuch, T. M., and J. Parsch. 2005. Patterns of synonymous codon usage in *Drosophila melanogaster* genes with sex-biased expression. Genetics. **170**:1691-1700.

Hosack, D. A., G. Dennis, B. T. Sherman, H. C. Lane, and R. A. Lempicki. 2003. Identifying biological themes within lists of genes with EASE. Genome Biol. **4**:R70.

Jorgensen, F. G., M. H. Schierup, and A. G. Clark. 2007. Heterogeneity in regional GC content and differential usage of codons and amino acids in GC-poor and GC-rich regions of the genome of *Apis mellifera*. Mol Biol Evol. **24**:611-619.

Kampstra, P. 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. J Stat Software, Code Snippets. **28**:1-9.

Kim, S. H., and S. V. Yi. 2007. Understanding relationship between sequence and functional evolution in yeast proteins. Genetica. **131**:151-156.

Krauss, V., C. Thummler, F. Georgi, J. Lehmann, P. F. Stadler, and C. Eisenhardt. 2008. Near intron positions are reliable phylogenetic markers: An application to Holometabolous insects. Mol Biol Evol. **25**:821-830.

Kriventseva, E. V., N. Rahman, O. Espinosa, and E. M. Zdobnov. 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res. **36**:D271-275.

Larracuente, A. M., T. B. Sackton, A. J. Greenberg, A. Wong, N. D. Singh, D. Sturgill, Y. Zhang, B. Oliver, and A. G. Clark. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. **24**:114-123.

Lemos, B., B. R. Bettencourt, C. D. Meiklejohn, and D. L. Hartl. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. **22**:1345-1354.

Nasonia Genome Sequencing Consortium. *in preparation*. Main genome paper.

Pal, C., B. Papp, and M. J. Lercher. 2006. An integrated view of protein evolution. Nat Rev Genet. **7**:337-348.

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Savard, J., D. Tautz, S. Richards, G. M. Weinstock, R. A. Gibbs, J. H. Werren, H. Tettelin, and M. J. Lercher. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Res. **16**:1334-1338.

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. **56**:564-577.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, et al. (17 co-authors). 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. **4**:41.

Wolf, Y. I., L. Carmel, and E. V. Koonin. 2006. Unifying measures of gene function and evolution. Proc R Soc B. **273**:1507-1515.

Wright, F. 1990. The effective number of codons used in a gene. Gene. **87**:23-29.

Table *S1*. Evolutionary rates (branch lengths in units of amino acid substitutions per site) for *Apis mellifera* and *Nasonia vitripennis* orthologs grouped according to caste-biased expression class in *A. mellifera*. *P*-values for rank sum tests demonstrate that mean branch lengths differed among *A. mellifera,* but not *N. vitripennis*, expression classes (significant *P*-values in bold).

	Mean branch length ± SEM			Rank sum test *P*-value			
	Worker-biased	Non-biased	Queen-biased	Overall[a]	P_{WN}[b]	P_{NQ}[b]	P_{WQ}[b]
A. mellifera	0.1042 ± 0.0056	0.1225 ± 0.0036	0.1438 ± 0.0072	0.0019**	0.2993	0.0206*	0.0020**
N. vitripennis	0.1314 ± 0.0077	0.1459 ± 0.0050	0.1542 ± 0.0078	0.0742	1.0000	0.2032	0.0853

[a] Kruskal-Wallis rank sum test *P*-value of worker-biased, nonsignificant, and queen-biased genes

[b] Bonferroni corrected Mann-Whitney *U* test *P*-value of worker-biased versus nonsignificant (P_{WN}), nonsignificant versus queen-biased (P_{NQ}), and worker-biased versus queen-biased (P_{WQ})

Table *S2*. Functional gene ontology (GO) biological process terms overrepresented by caste-biased gene expression class and evolutionary rate class in *Apis mellifera*, which reveal putative functional links between caste and evolutionary rate. Notably, we find that rapidly evolving queen-biased genes are enriched for functions related to metabolism. These functional terms correspond with the observed rate differences between queen-biased and worker-biased genes and may help to explain the discrepancy in their evolutionary rates.

Branch length		Gene expression class		
Bin	Mean ± SEM	Worker-biased	Nonsignificant	Queen-biased
		GO biological process terms significantly ($P < 0.05$) overrepresented in both gene expression class and branch length bin		
1	0.0202 ± 0.0008	gene expression (GO:0010467); ARF protein signal transduction (GO:0032011); regulation of ARF protein signal transduction (GO:0032012)	cell morphogenesis (GO:0000902); anatomical structure morphogenesis (GO:0009653); organelle organization and biogenesis (GO:0006996); cellular structure morphogenesis (GO:0032989); cellular component	

			organization and biogenesis (GO:0016043)	
2	0.0541 ± 0.0006	transcription from RNA polymerase II promoter (GO:0006366)		Rho protein signal transduction (GO:0007266)
3	0.0925 ± 0.0010	multicellular organismal development (GO:0007275)		
4	0.1523 ± 0.0014	nitrogen compound metabolic process (GO:0006807)		acetyl-CoA metabolic process (GO:0006084); cellular catabolic process (GO:0044248); cofactor catabolic process (GO:0051187); tricarboxylic acid cycle (GO:0006099); coenzyme catabolic process (GO:0009109); aerobic respiration (GO:0009060); acetyl-

			CoA catabolic process (GO:0046356); organic acid metabolic process (GO:0006082); carboxylic acid metabolic process (GO:0019752); cellular respiration (GO:0045333)
5	0.2918 ± 0.0068		carbohydrate metabolic process (GO:0005975); generation of precursor metabolites and energy (GO:0006091); electron transport (GO:0006118)

Table *S3*. Comparison of evolutionary rates (branch lengths in units of amino acid substitutions per site) for *Apis mellifera* and *Nasonia vitripennis* orthologs with branch lengths calculated using five-species phylogenies, grouped according to caste-biased expression class in *A. mellifera*.

	Mean branch length ± SEM			Rank sum test *P*-value			
	Worker-biased[a]	Non-biased[a]	Queen-biased[a]	Overall[b]	P_{WN}[c]	P_{NQ}[c]	P_{WQ}[c]
A. mellifera	0.0777 ± 0.0093	0.1066 ± 0.0064	0.1509 ± 0.0140	0.0003***	0.0524	0.0179*	0.0007***
N. vitripennis	0.0939 ± 0.0094	0.1266 ± 0.0088	0.1422 ± 0.0115	0.0178*	0.2384	0.2259	0.0148*

[a] Worker-biased $n = 58$; non-biased $n = 231$; queen-biased $n = 81$

[b] Kruskal-Wallis rank sum test *P*-value of worker-biased, nonsignificant, and queen-biased

[c] Bonferroni corrected Mann-Whitney *U* test *P*-value of worker-biased versus nonsignificant (P_{WN}), nonsignificant versus queen-biased (P_{NQ}), and worker-biased versus queen-biased (P_{WQ})

Table *S4*. Spearman's rank correlation coefficients (r_s) and partial correlations between *Apis mellifera* evolutionary rates (branch lengths in units of amino acid substitutions per site) and selected gene attributes for genes with branches calculated using five-species phylogenies. The $\log_2$ ratio of queen to worker gene expression remains significantly correlated with *A. mellifera* branch lengths when controlling for the combined effect of *Nasonia vitripennis* branch lengths (a proxy for shared ancestral evolutionary determinants) and several *A. mellifera* gene characteristics associated with evolutionary rates.

Variable (*X*)	**Correlation** r_s *X,A. mellifera* branch length	**Partial correlation** r_s *X,A. mellifera* branch length|all other variables
N. vitripennis branch length	0.679^{*****}	0.670^{*****}
Log_2(queen/worker) gene expression	0.253^{****}	0.187^{***}
Synonymous 3rd codon position GC content	-0.221^{***}	-0.116^{*}
Effective number of codons	-0.198^{***}	0.038
Coding sequence length	-0.032	-0.072
Brain expression level	0.024	0.036

$^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 10^{-3}$; $^{****}P < 10^{-6}$; $^{*****}P < 10^{-9}$

Table *S5*. Principal component regression analysis of *Apis mellifera* evolutionary rates (branch lengths in units of amino acid substitutions per site) and selected characteristics for genes with branches calculated using the same 3 insect outgroups in each case.

	Principal components				
	1	2	3	4	All (6)
Percent variance explained in					
A. mellifera branch length	3.97***	5.66****	3.21***	15.43*****	28.48*****
Percent contributions					
Log_2(queen/worker) gene expression	0.1	**62.2**	**32.5**	5.1	
N. vitripennis branch length	0.7	3.2	1.8	**93.7**	
Effective number of codons	**47.8**	2.0	1.8	0	
Synonymous 3rd codon position GC content	**45.3**	1.5	1.0	1.0	
Coding sequence length	1.2	2.3	5.3	0.1	
Brain expression level	4.9	**28.8**	**57.7**	0.1	

$^{***}P < 10^{-3}$; $^{****}P < 10^{-6}$; $^{*****}P < 10^{-9}$. Bold indicates variables that contribute at least 10% to the principal component.

Fig. *S1*. Species tree with approximate divergence times. Dotted lines depict terminal branch lengths for *A. mellifera* and *N. vitripennis* used in this study. Insect orders are labeled on internal branches.

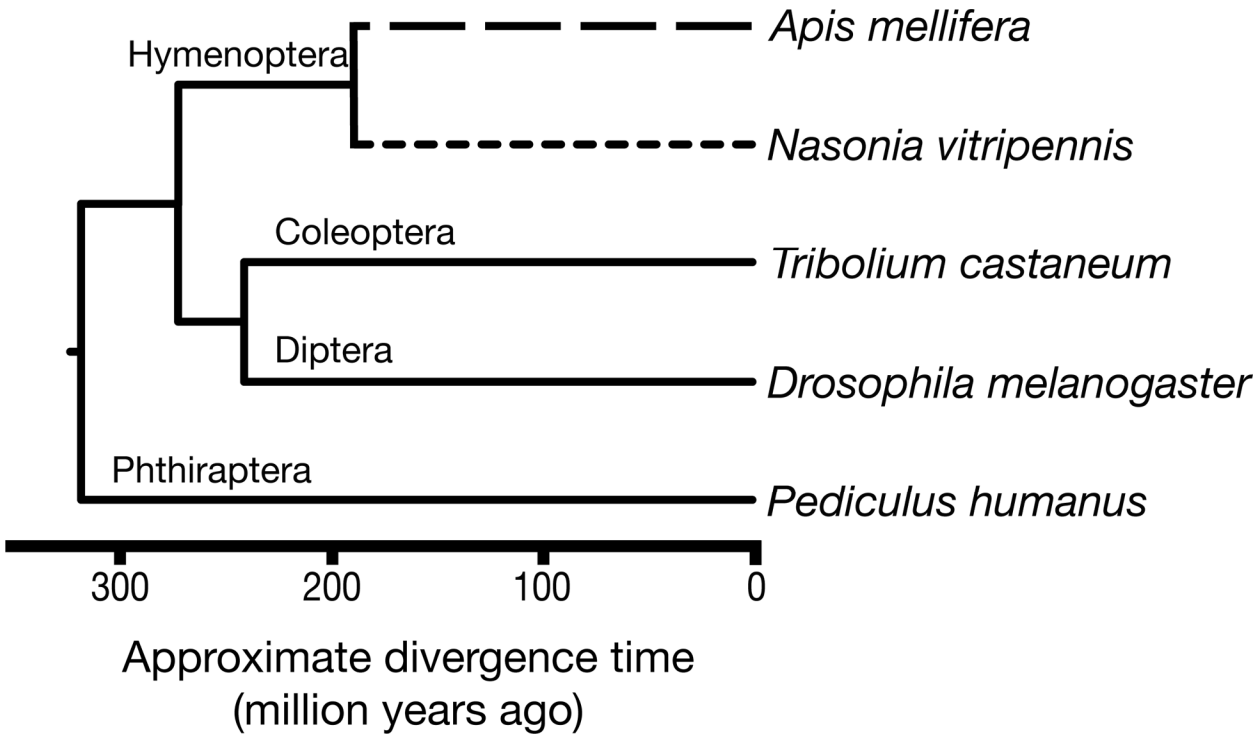

Fig. *S2*. 'Beanplots' illustrating the distributions of evolutionary rates (branch lengths in units of amino acid substitutions per site) for *A. mellifera* and *N. vitripennis* orthologs according to caste-biased expression class in *A. mellifera* (Kampstra 2008). Individual observations are shown as small horizontal lines in a one-dimensional scatter plot (increased line width indicates multiple observations for a given value). Black bars indicate group means while the dotted line indicates the overall mean of the plot.