MOLECULAR EVOLUTION IN THE SOCIAL INSECTS

A Dissertation
Presented to
The Academic Faculty

by

Brendan G. Hunt

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Biology in the
School of Biology

Georgia Institute of Technology
May 2011

MOLECULAR EVOLUTION IN THE SOCIAL INSECTS

Approved by:

Dr. Michael A. D. Goodisman, Co-advisor
School of Biology
Georgia Institute of Technology

Dr. Soojin V. Yi, Co-advisor
School of Biology
Georgia Institute of Technology

Dr. J. Todd Streelman
School of Biology
Georgia Institute of Technology

Dr. I. King Jordan
School of Biology
Georgia Institute of Technology

Dr. Todd A. Schlenke
Department of Biology
Emory University

Date Approved: February 8, 2011

For Colin.

ACKNOWLEDGEMENTS

I wish to thank, first and foremost, my wife Jennifer for making many sacrifices, paying bills, and offering steadfast emotional support in my pursuit of a PhD. My co-advisors, Mike Goodisman and Soojin Yi, have provided a unique balance of distinct personalities and expertise that has been a driving force behind this research. Their discussions have been invaluable. Mike has devoted an enormous amount of time to advising on a day-to-day basis, for which I am indebted. Todd Streelman, King Jordan, and Todd Schlenke contributed ideas and a sense of excitement and encouragement at committee meetings that made these meetings remarkably productive. Students in the Goodisman and Yi labs have provided moral support and scientific insight throughout this process. In particular, I would like to thank Jen Kovacs for helping me establish myself in the early years of my PhD, for taking part in countless scientific discussions, and for providing a social buffer at many conferences. More recently, Karl Glastad has been integral to the development of hypotheses, the discussion of papers, and the collection and feeding of ants. His title for this dissertation, *Genes with means: the power of love*, came in a close second place. Finally, I would like to thank Kevin Jansen for offering encouragement and guidance when needed most.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
BLAST	Basic Local Alignment Search Tool
cDNA	Complementary DNA
CDS	Coding Sequence
CI	Confidence Interval
CpG	Cytosine Immediately Followed by Guanine in 5' to 3' Direction
CpG o/e	Normalized CpG Dinucleotide Content
dN	Nonsynonymous Substitution Rate
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyl Transferase
dS	Synonymous Substitution Rate
EST	Expressed Sequence Tag
Fop	Frequency of Optimal Codons
GC Content	Guanine and Cytosine Content
GC3s	Third Codon Position Synonymous GC Content
GO	Gene Ontology
GpC	Guanine Immediately Followed by Cytosine in 5' to 3' Direction
KOG	Eukaryotic Cluster of Orthologous Groups
NCBI	National Center for Biotechnology Information
PC	Principal Component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PGL	Propensity of Gene Loss

QRT-PCR	Quantitative Real-Time Reverse Transcription PCR
RefSeq	Reference Sequence database
RNA	Ribonucleic Acid
r_s	Spearman's Rank Correlation Coefficient
SEM	Standard Error of the Mean

SUMMARY

Social insects are ecologically dominant because of their specialized, cooperative castes. Reproductive queens lay eggs, while workers take part in brood rearing, nest defense, and foraging. These cooperative castes are a prime example of phenotypic plasticity, whereby a single genetic code gives rise to variation in form and function based on environmental differences. Thus, social insects are well suited for studying mechanisms which give rise to and maintain phenotypic plasticity.

At the molecular level, phenotypic plasticity coincides with the differential expression of genes. This dissertation examines the molecular evolution of genes with differential expression between discrete phenotypic or environmental contexts, represented chiefly by female queen and worker castes in social insects. The studies included herein examine evolution at three important levels of biological information: (i) gene expression, (ii) modifications to DNA in the form of methylation, and (iii) protein-coding sequence.

From these analyses, a common theme has emerged: genes with differential expression among castes frequently exhibit signatures of relaxed selective constraint relative to ubiquitously expressed genes. Thus, genes associated with phenotypic plasticity paradoxically exhibit modest importance to overall fitness but exceptional evolutionary potential, as illustrated by the success of the social insects.

CHAPTER 1

INTRODUCTION

Social insects are ecologically dominant because of their specialized, cooperative castes (Wilson 1990). Reproductive queens lay eggs, while workers take part in brood rearing, nest defense, and foraging (Wilson 1971). These cooperative castes represent an important advance in biological organization (Maynard Smith and Szathmary 1995). Castes are also a prime example of phenotypic plasticity, whereby a single genetic code gives rise to variation in form and function based on environmental differences (West-Eberhard 2003). Indeed, social insect castes are most often environmentally determined (Wheeler 1986) and frequently exhibit discrete adult phenotypes among genetically indistinguishable individuals [i.e., polyphenism; (Evans and Wheeler 2001; Nijhout 2003)].

At the molecular level, phenotypic plasticity coincides with the differential expression of genes (Smith et al. 2008). Due to the evolutionary importance of phenotypic plasticity (West-Eberhard 2003; Pfennig et al. 2010), there has been a recent surge of interest in the molecular evolutionary properties of genes that exhibit conditional expression between phenotypic forms (Barker et al. 2005; Ellegren and Parsch 2007; Brisson and Nuzhdin 2008; Bonduriansky and Chenoweth 2009; Mank and Ellegren 2009; Snell-Rood et al. 2010; Van Dyken and Wade 2010). Important outstanding questions ask whether certain classes of genes are predisposed for involvement in phenotypic plasticity [e.g., (Mank and Ellegren 2009)] and whether phenotypic plasticity itself results in fundamental changes in molecular evolutionary processes [e.g., (Van Dyken and Wade 2010)]. The research presented in this dissertation sheds new light on the evolution of phenotypic plasticity and sociality, helping to address the above questions and to suggest new avenues of study.

Chapter two of this dissertation (Hunt and Goodisman 2010) addresses the degree of conservation in gene expression patterns associated with caste differences in two species of social wasp from the genus *Vespula*. Contrasts of gene expression patterns in social insect taxa that diverged following the advent of queen and worker polyphenism can provide valuable insight into genotype-environment interactions and the evolutionary maintenance of phenotypic plasticity (Stearns 1989; Nijhout 1999; Gilbert 2001; Sultan 2007; Pfennig et al. 2010). In our study of gene expression in social wasps (Hunt and Goodisman 2010), we found that caste-biased gene expression patterns appear to be subject to particularly high turnover between species (also see Ometto et al. in press).

Chapters three and four (Elango, Hunt et al. 2009; Hunt et al. 2010a) address the relationship between DNA methylation and conditional gene expression. DNA methylation is an important epigenetic modification that regulates gene expression (Wolffe and Matzke 1999; Bird 2002; Weber et al. 2007) and is sensitive to environmental input. Thus, DNA methylation is an important mechanism by which phenotype can be influenced by external factors such as nutrition (Jaenisch and Bird 2003; Weaver et al. 2004; Jirtle and Skinner 2007; Kucharski et al. 2008; Maleszka 2008; Moczek and Snell-Rood 2008). One of the most striking examples of the phenotypic consequences of DNA methylation is the observed link between caste determination and DNA methylation in honeybees (Kucharski et al. 2008).

Surprisingly, our research revealed that caste-biased genes in the honeybee *Apis mellifera* preferentially lack evolutionary signatures of high levels of DNA methylation (Elango, Hunt et al. 2009). Our comparative analysis of *A. mellifera* and the pea aphid *Acyrthosiphon pisum* revealed that DNA methylation is widely targeted to ubiquitously expressed genes in insects [among tissues and alternate phenotypes; (Hunt et al. 2010a)]. A key investigation of intragenic methylation in mammals suggested that methylation prevents transcription from being initiated at alternate start sites (Maunakea et al. 2010). The buffering against spurious transcription is a process which may be especially critical for ubiquitously expressed genes (Hunt et al. 2010a; Maunakea et al. 2010). Although tissue-specific or developmentally- specific methylation may play a role in differential splicing (Lyko et al. 2010) or differential expression of some key genes associated with caste determination, the primary targets of DNA methylation in insects are not conditionally-expressed genes (Hunt et al. 2010a).

Finally, chapters five and six (Hunt et al. 2010b; Hunt et al. in preparation) directly address the relationship between conditional gene expression and molecular evolutionary changes in protein-coding sequences in *A. mellifera* and the fire ant *Solenopsis invicta*. Theory predicts that conditionally-expressed genes will be subject to diminished selective constraint because these genes exhibit limited phenotypic consequences in a subset of the population (Barker et al. 2005; Brisson and Nuzhdin 2008; Mank and Ellegren 2009; Snell-Rood et al. 2010; Van Dyken and Wade 2010). Furthermore, conditional gene expression can alleviate intralocus antagonism arising from differences in phenotypic optima for distinct developmental or environmental contexts (Chippindale et al. 2001; Ellegren and Parsch 2007; Bonduriansky and Chenoweth 2009). By alleviating antagonism between conditions, conditional gene expression could actually facilitate phenotypic specialization by allowing directional selection, rather than stabilizing selection, to shape protein sequences in such cases (Gadagkar 1997; Bonduriansky and Chenoweth 2009). Thus, either through population genetic processes that influence the efficiency of selection, or by a reduction in pleiotropy between conditions, conditionally-expressed genes are predicted to undergo diminished purifying selection relative to ubiquitously expressed genes.

Relaxed selection may also act as an evolutionary precursor to conditional expression (Mank and Ellegren 2009). Indeed, the costs of modifications to gene expression in specific developmental or environmental contexts may be lower for genes with relatively low importance to overall fitness (Castillo-Davis et al. 2004; Zhang et al. 2007; Mank and Ellegren 2009). Thus, whether relaxed selection on proteins is primarily a cause or consequence of conditional gene expression remains unknown. Our results from analysis of the fire ant *S. invicta* (Hunt et al. in preparation) suggest that conditional gene expression is one of the key factors associated with rates of protein evolution (Wall et al. 2005; Pal et al. 2006; Drummond and Wilke 2008). Even more importantly, we found that orthologs of genes with caste-biased expression exhibit high rates of protein evolution, even in taxa lacking castes. Thus, relaxed selective constraint on proteins is an important and underappreciated element in the evolutionary origin and elaboration of phenotypic plasticity.

From the analyses presented in this dissertation, a common theme has emerged: genes with differential expression among castes frequently exhibit signatures of relaxed selective constraint relative to ubiquitously expressed genes. This is evident from (i) evolutionary lability in gene expression patterns, (ii) relative scarcity of methylation targets, which preferentially target ubiquitously expressed genes in insects, and (iii) relative elevation in rates of protein evolution. Thus, genes associated with phenotypic plasticity paradoxically exhibit modest importance to overall fitness but exceptional evolutionary potential, as illustrated by the success of the social insects.

CHAPTER 2

EVOLUTIONARY VARIATION IN GENE EXPRESSION IS ASSOCIATED WITH DIMORPHISM IN EUSOCIAL VESPID WASPS[1]

Abstract

Phenotypic diversity is frequently generated by differences in gene expression. In this study, we addressed the relationship between homology in gene expression and phenotype among four species of eusocial wasps. Specifically, we investigated the evolution of caste-specific and sex-specific gene expression patterns associated with caste polyphenisms and sexual dimorphisms. We also identified several genes with functions relevant to their phenotype-specific roles. Our results suggest that gene expression profiles associated with caste polyphenisms may evolve rapidly relative to those associated with sexes. Thus, caste-biased genes may undergo less regulatory constraint or be subject to greater neutral variation in expression than sex-biased genes.

Introduction

Understanding the molecular basis of phenotypic diversity is one of the fundamental challenges of evolutionary biology. In recent years, a surge of attention and debate has focused on the relative evolutionary roles played by protein-coding sequence, cis-regulatory elements, epigenetic inheritance, and developmental processes (West-Eberhard 2003; Pigliucci et al. 2006; Hoekstra and Coyne 2007; Moczek 2007; Carroll

[1] Hunt BG, Goodisman MAD. 2010. Evolutionary variation in gene expression is associated with dimorphism in eusocial vespid wasps. Insect Mol Biol 19:641-652.

2008; Jablonka and Raz 2009). This attention has not been in vain. Among other findings, a large body of research now demonstrates that phenotypic diversification frequently arises through changes in the spatial and temporal expression of functionally conserved proteins (Carroll 2008).

Polyphenisms, which arise when environmental differences give rise to discrete phenotypic classes from a single genotype, provide an opportunity to study phenotypic divergence in the context of a shared genetic background. Insects, in particular, provide numerous examples of polyphenisms marked by the canalization of highly elaborate, plastic phenotypes (Evans and Wheeler 2001; Nijhout 2003). Examples include beetle horns (Moczek et al. 2007), female aphid dispersal and parthenogenetic morphs (Brisson and Stern 2006), and eusocial insect castes (Wilson 1971; Wheeler 1986). In each of these cases differences in gene expression, rather than differences in genotype, are associated with the presence of alternate phenotypes (Sumner 2006; Brisson et al. 2007; Goodisman et al. 2008; Smith et al. 2008; Kijimoto et al. 2009). In addition, sex-specific differences in phenotype, although often triggered by genetic differences, are also associated with widespread differences in gene expression (Jin et al. 2001; Ellegren and Parsch 2007; Ayroles et al. 2009). Thus gene regulation has been widely linked to the developmental fate and adult function of distinct phenotypic classes.

Many comparative evolutionary studies have investigated the degree of similarity in the molecular basis of independently arising, convergent phenotypes. For example, discrete queen and worker castes have arisen multiple times independently in eusocial insects and exhibit limited, but notable, convergence in their underlying gene expression patterns (Toth et al. 2007; Smith et al. 2008). Interestingly, many studies have also observed surprising flexibility in the molecular basis of similar phenotypes with a shared evolutionary origin in different species, such as gene expression patterns associated with sexual dimorphisms (Zhang et al. 2007)and caste polyphenisms (Abouheif and Wray 2002; Weil et al. 2009). An important set of evolutionary questions thus concerns the levels of functional drift and neutral variation influencing gene expression profiles associated with homologous phenotypes (Wray and Abouheif 1998; True and Haag 2001; Khaitovich et al. 2004).

In this study, we sought to understand how gene expression patterns evolve in homologous phenotypic classes of eusocial insects. Hymenopteran eusocial insects, which include ants, some bees, and some wasps, are attractive candidates for studying the association between variation in phenotype and gene expression for several reasons. First, the caste system that defines eusociality has arisen at least nine independent times within the Hymenoptera (Wilson 1971; Hughes et al. 2008). Specifically, in all hymenopteran eusocial insects, labor is divided within a colony such that queens and males reproduce and female workers engage in tasks related to brood rearing and colony defense (Wilson 1971). Thus, queen and worker castes, which often exhibit extreme polyphenism, are prime targets for studying the evolutionary maintenance of alternate phenotypes arising in a single sex. Second, hymenopterans do not exhibit chromosomal sex determination. Instead, males are generally produced by unfertilized, haploid eggs (arrhenotokous parthenogenesis) and females are produced by fertilized, diploid eggs (Heimpel and de Boer 2008). Thus, males and females possess the same genes, and sex differences, like caste differences, develop in large part through variation in gene expression associated with differential gene activation and differences in ploidy (Aron et al. 2005; Hoffman and Goodisman 2007). Finally, although castes and sexes are both examples of dramatic dimorphisms, castes have arisen far more recently than sexes during the course of evolution. Our study takes advantage of the unique characteristics of eusocial hymenopterans to reveal the potential evolutionary consequences of differences between caste polyphenisms and sexual dimorphisms on the evolution of caste-specific and sex-specific gene expression patterns.

We focused our investigation on wasps of the family Vespidae because they exhibit wide variation in social complexity. For example, wasps in the subfamily

Table 2.1. Gene annotation and relative expression patterns among males (M), queens (Q), and workers (W) in *V. squamosa* and *V. maculifrons*, between workers and males in *D. maculata*, and between late-season females (F) and males in *P. exclamans*.

	Gene ontology similarity	Relative expression pattern[a]				Gene ontology annotation[b]
Contig ID	**Putative ortholog**	***V. squamosa***	***V. maculifrons***	***D. maculata***	***P. exclamans***	**Biological process terms**
contig52	regucalcin	W ≈ M > Q	-	W > M	-	cellular calcium ion homeostasis, positive regulation of ATPase activity, regulation of calcium-mediated signaling
contig61	Ribosomal protein L13	Q > W ≈ M	NS	W > M	-	mitotic spindle elongation, mitotic spindle organization, translation
contig88	Vitellogenin-2	-	Q > W ≈ M	W > M	-	determination of adult life span
contig155	Ribosomal protein L32	NS	Q ≈ W > M	W > M	F > M	mitotic spindle elongation, mitotic spindle organization, translation
contig170	lipoma HMGIC fusion partner-like 3	NS	M > Q ≈ W	-	F > M	biological process
contig272	Ribosomal protein L8	NS	-	W > M	-	mitotic spindle elongation, mitotic spindle organization, translation
contig339	Tropomyosin 2	M > W > Q	M > Q ≈ W	M > W	-	heart development
contig350	Stubarista	Q > M	Q ≈ W > M	W > M	F > M	mitotic spindle elongation, mitotic spindle organization, translation
contig422	Ribosomal protein S18	-	Q > W ≈ M	W > M	F > M	mitotic spindle organization, translation, translational initiation

contig430	Twinstar	W > Q ≈ M	W ≈ M > Q	W > M	F > M	actin filament organization, axonogenesis, leg segmentation
contig446	Elongation factor 1alpha100E	W > M	W > Q > M	W > M	F > M	translation, translational elongation
contig462	Ribosomal protein L15	-	Q ≈ W > M	W > M	F > M	Translation
contig498	string of pearls	NS	-	W > M	-	Translation
contig504	CG4692	NS	Q > W > M	NS	NS	proton transport
contig506	Ribosomal protein L23	NS	Q > M	-	F > M	mitotic spindle elongation, mitotic spindle organization, translation
contig531	Ribosomal protein LP0	-	-	W > M	F > M	DNA repair, translation
contig583	Ribosomal protein L17	W > M	Q ≈ W > M	NS	F > M	mitotic spindle elongation, mitotic spindle organization, translation
contig626	wings up A	W ≈ M > Q	M > W ≈ Q	NS	F > M	muscle maintenance, muscle organ development, skeletal muscle tissue development
contig644	Superoxide dismutase	NS	NS	W > M	F > M	aging, determination of adult life span, removal of superoxide radicals
contig664	40S ribosomal protein S27	NS	NS	NS	F > M	cell proliferation, translational elongation
contig677	60S ribosomal protein L23a	W > Q	Q ≈ W > M	W > M	-	translational elongation

[a] A caste's level of expression was designated as less-than or greater-than another if $P < 0.05$. When a caste had intermediate but nonsignificant expression differences to two significantly different groups, it was not included in the Table. - = no data, NS = nonsignificant differences (see Fig. 2.1).

[b] Where more than three GO biological process terms existed for a gene, only three terms were chosen.

Vespinae exhibit dramatic polyphenisms and have large colony sizes, whereas wasps in the subfamily Polistinae tend to exhibit relatively small colony sizes and behaviorally-defined castes (Hunt 2007). In this study, we examined relative gene expression levels for 21 genes in castes and sexes of four vespid wasp species with a common eusocial ancestor (Hines et al. 2007; Hunt 2007) using quantitative real-time PCR. Two of the four species were informative of differences arising between female castes and sexes, whereas the other two species provided information on sex differences alone, due to partial sampling of female castes. We predicted that gene expression profiles specific to a given female caste would evolve as fast or faster than gene expression profiles specific to a given sex (Jin et al. 2001; Khaitovich et al. 2004; Cutter and Ward 2005; Zhang et al. 2007; Ayroles et al. 2009). Gene expression associated with castes may be particularly fast-evolving because castes are relatively recent evolutionary innovations (Hughes et al. 2008), in contrast with sexes, and recent experimental evidence suggests that castes exhibit flexible underlying patterns of gene regulation across taxa (Weil et al. 2009).

Results

We analyzed gene expression in three vespine wasps, *Vespula maculifrons*, *V. squamosa* (commonly known as yellowjackets; Macdonald and Matthews 1981, 1984) and *Dolichovespula maculata* (hornets; Archer 2006), as well as one polistine wasp, *Polistes exclamans* (paper wasps; Hunt 2007; Turillazzi and West-Eberhard 1996). Relative levels of gene expression were determined for adult queens, workers, and males in *V. maculifrons* and *V. squamosa*, in adult workers and males in *D. maculata*, and in adult females (caste undetermined) and males in *P. exclamans* (Table 2.1, Figs. 2.1 and 2.2). Data were collected for up to 21 genes in each species (Fig. 2.1, see experimental procedures).

We first investigated whether sexes or castes showed more conserved patterns of gene regulation across taxa. Comparisons of relative gene expression levels for the nine

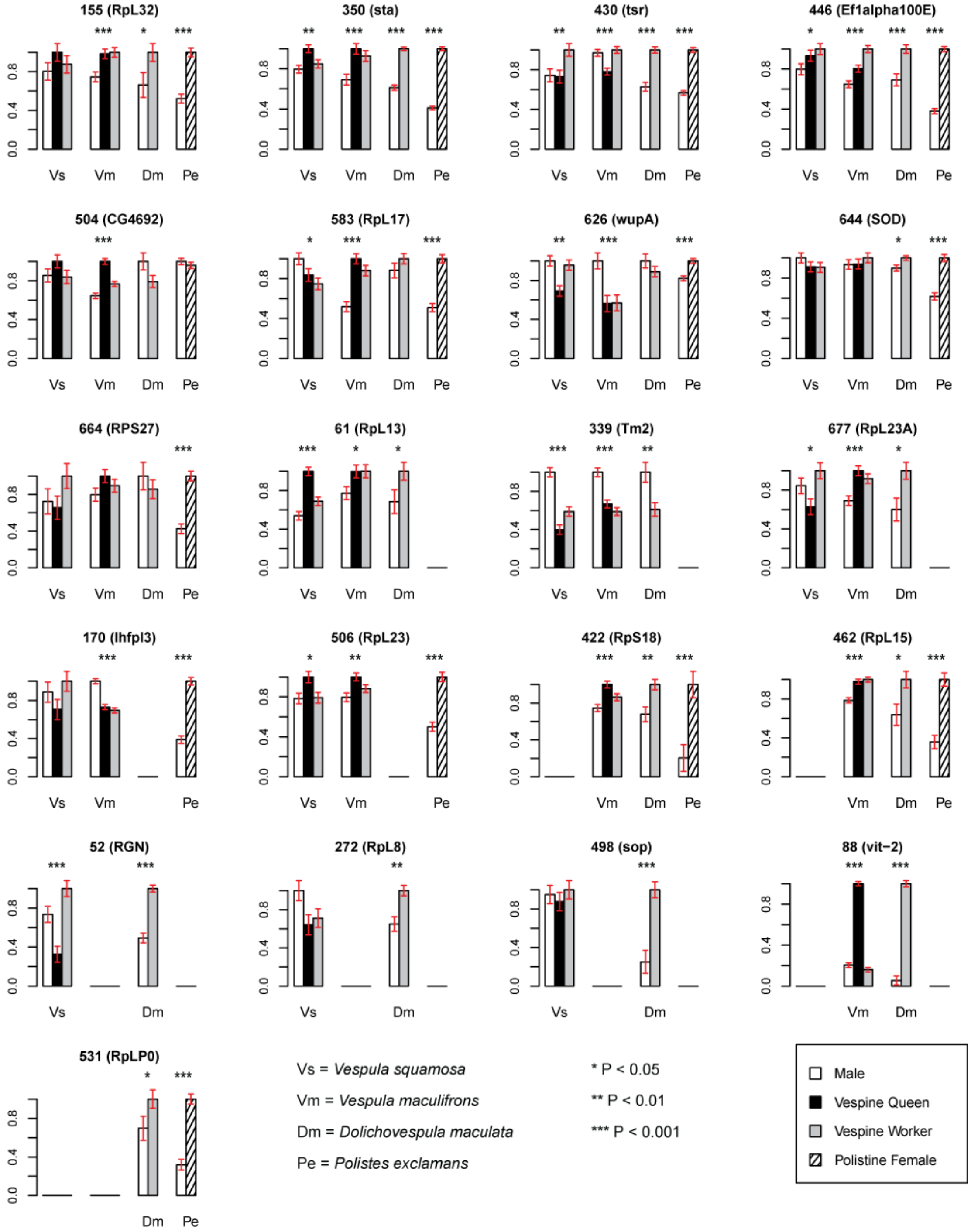

Figure 2.1. Relative expression differences between castes and sexes in vespid wasps. Normalized least squares means of relative expression levels are plotted with 95% confidence intervals for four vespid wasp species. P-values denote the outcome of an ANOVA test of differences between castes.

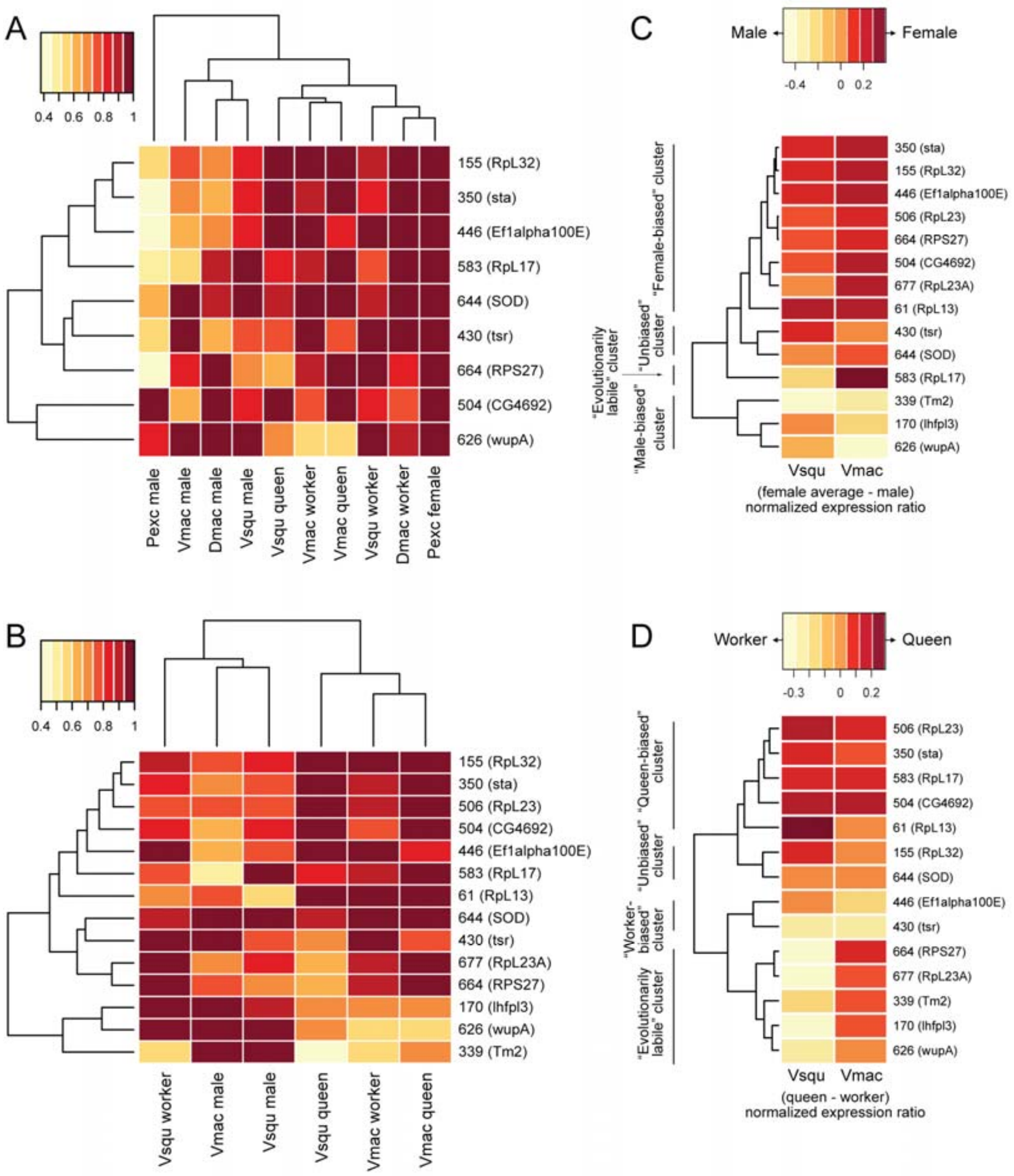

Figure 2.2. Gene expression profiles associated with castes and sexes in vespid wasps. Increasing or decreasing block saturation indicates higher or lower relative levels of expression for a particular gene in a given caste. (A) Normalized gene expression intensities in four species illustrate that male and female gene expression profiles cluster across taxa; all genes with data from every phenotypic class in all four species are shown. (B) Normalized gene expression intensities in *V. squamosa* (Vsqu) and *V. maculifrons* (Vmac) illustrate that female queen and worker gene expression profiles cluster in *V. maculifrons* but not *V. squamosa*. (C) Differences in normalized gene expression intensities between females and males indicate that male-biased and female-biased genes cluster among *Vespula* species. (D) Differences in normalized gene expression intensities between queens and workers indicate that a limited proportion of genes exhibit conserved differences with respect to caste. All genes with data for queens, workers, and males in both *Vespula* species are shown in panels B-D.

genes with data from sexes and castes of all four species revealed an overall clustering of male and female profiles across taxa (Fig. 2.2A). Comparisons of relative gene expression levels for the 14 genes with data from sexes and castes of both *Vespula* species revealed that seven genes (contigs 61, 155, 350, 446, 504, 506, and 677) were consistently expressed more highly in females than males (with statistically higher expression in female castes of at least one taxon; Figs. 2.2B and 2.2C). Meanwhile, two of 14 genes (contigs 339, and 626) were consistently more highly expressed in males than females in both species (with statistically higher expression in males of at least one taxon; Figs. 2.1, 2.2B, and 2.2C).

We next determined whether genes showed conserved expression relationships between queen and worker castes in both *Vespula* species. Comparisons of relative gene expression levels in queens and workers revealed that some genes were consistently differentially expressed between castes in both species. Specifically, four of 14 genes (contigs 350, 504, 506, and 583) were more highly expressed in queens than workers in *V. maculifrons* and *V. squamosa* (with statistically significant queen upregulation in one species in all four cases; Figs. 2.1 and 2.2D). In contrast, seven of 14 genes had divergent relationships with respect to queen and worker gene expression differences in *V. maculifrons* and *V. squamosa*, demonstrating the labile nature of caste-biased gene expression over evolutionary time (Figs. 2.1 and 2.2D). Five of these genes (contigs 61, 339, 446, 626, and 677) exhibited statistically different expression levels between queens and workers within at least one taxon. Furthermore, queen and worker gene expression profiles had qualitatively different relationships in *V. maculifrons* and *V. squamosa*; queen and worker gene expression profiles clustered together more closely in *V. maculifrons* than in *V. squamosa* (Fig. 2.2A and 2.2B). Together these results suggest that relative gene expression levels between castes vary widely among taxa and gene expression profiles associated with sexes are less labile than those associated with castes.

We further tested whether *Vespula* genes in our study were convergent, with respect to caste-specific regulation, with the independently eusocial honeybee *Apis mellifera* (Table 2.2; Barchuk et al. 2007; Grozinger et al. 2007). Two of the seven genes we compared had convergent relationships between expression pattern and caste, both of which were upregulated in queens. These were orthologs of *vitellogenin* and CG4692 (see discussion). These results are consistent with the presence of few genes showing convergent caste-specific expression patterns in distantly related, independently eusocial taxa.

Finally, we examined the variability of gene regulation among castes and sexes to see if relative gene expression levels differed among these phenotypic classes within each taxon. We found that all 21 genes exhibited significant differences in expression among some phenotypic classes for at least one species, suggesting that gene expression differences are widespread among dimorphic adult phenotypes (Fig. 2.1). We also observed widespread differences in gene expression between individuals of the same caste, even for *V. squamosa* samples, which were all collected at a single time point, providing additional support to the finding that gene expression is highly variable between individuals (Oleksiak et al. 2002).

Discussion

Evolution of gene expression in castes and sexes

One of the most striking patterns that emerged from our analysis is that caste-biased gene expression profiles appear to have undergone greater rates of change than sex-biased gene expression profiles in the *Vespula* genus (Figs. 2.2B, 2.2C, and 2.2D). Previous studies have demonstrated that sex-biased genes undergo elevated gene expression divergence between taxa relative to non-sex biased genes (Meiklejohn et al. 2003; Ranz et al. 2003). However, our data suggest that caste-biased gene expression may evolve at even faster rates than sex-biased gene expression. The observation that

Table 2.2. Comparison of gene expression between *A. mellifera* queens (Q) and workers (W) and two species of *Vespula* wasps

Contig ID	*A. mellifera* gene name (ID)	BLAST E-value (score)	*A. mellifera* expression	*V. squamosa* expression[c]	*V. maculifrons* expression[c]
contig88	vitellogenin (GB13999-PA)	6e-63 (236)	Q > W[a]	-	Q > W
contig350	similar to stubarista CG14792-PA, isoform A (GB19082-PA)	1e-149 (524)	W > Q[a]	Q ≈ W	W ≥ Q
contig430	similar to Cofilin/actin-depolymerizing factor homolog (Protein D61) (Protein twinstar) (GB18917-PA)	3e-82 (299)	Q ≈ W[a]	W > Q	W > Q
contig446	elongation factor 1-alpha (GB16844-PA)	6e-74 (274)	Q ≈ W[a], Q > W[b]	W ≥ Q	W > Q
contig498	ribosomal protein S2e (GB10861-PA)	1e-151 (530)	Q > W[a]	Q ≈ W	-
contig504	similar to CG4692-PB, isoform B (GB18417-PA)	4e-30 (126)	Q > W[a]	Q ≈ W	Q > W
contig644	CuZn superoxide dismutase (GB10133-PA)	6e-43 (170)	Q > W[a]	Q ≈ W	Q ≈ W

> indicates significantly higher expression, ≈ indicates nonsignificant expression difference, - indicates no data.

[a] data comparing expression levels in adult queen and worker brains (Grozinger et al. 2007)
[b] data comparing 4th instar queen-destined and worker-destined larvae; other genes in Table 2 are unreported (either due to nonsignificant expression differences or technical issues) in comparisons of queen- and worker-destined 3rd, 4th, and 5th instar larvae (Barchuk et al. 2007)
[c] according to a post-hoc Tukey's HSD analysis for the effect of caste in our whole-body expression analysis (see experimental procedures).

caste-biased gene expression patterns are largely species-specific agrees with recent work comparing caste-biased gene expression patterns among closely related termites (Fig. 2.2D; Weil et al. 2009). Considering the flexibility in caste-specific gene expression patterns we observed within a single genus (Figs. 2.2B and 2.2D), the fact that we observed little convergence between vespid wasps and *A. mellifera* is not surprising (Table 2.2). Together, these results suggest that only a limited subset of caste-biased genes may contribute to a common 'genetic toolkit' underlying the evolution of eusocial castes (Toth and Robinson 2007). However, we also caution that differences observed between *Vespula* and *A. mellifera* gene expression patterns likely arise, in part, because the honeybee data were generated from adult brains (Grozinger et al. 2007) and whole-body larvae (Barchuk et al. 2007) while our *Vespula* data were generated from whole-body adults (Table 2.2).

Because female caste dimorphisms are relatively recent innovations, whereas sexual dimorphisms are widespread and ancient among animals, it is possible that greater resolution has been achieved in intralocus conflict over phenotypic optima arising between sexes than between castes (Bonduriansky and Chenoweth 2009). This may help to explain the labile gene expression patterns observed among castes of different species. Alternatively, genes that exhibit caste-biased expression patterns may simply be subject to fewer pleiotropic constraints or exhibit higher dispensability, and a lower impact on fitness, than genes associated with sex-bias(Mank et al. 2008; Mank and Ellegren 2009). Accordingly, although prior transcriptomic studies have revealed that variation in gene expression is not inherently neutral (Denver et al. 2005; Rifkin et al. 2005), the flexibility in caste-biased gene regulation could represent a larger role for neutral variation in shaping caste-specific patterns of gene expression relative to sex-specific gene expression (Khaitovich et al. 2004).

We also observed that queen and worker gene expression profiles clustered together more closely in *V. maculifrons* than in *V. squamosa* (Fig. 2.2A and 2.2B). This may reflect species-level differences in the degree of dimorphism between queen and worker castes because *V. squamosa* castes are among the most dimorphic of vespid wasps (Greene 1991). Alternatively, the clustering of worker and male phenotypes in *V. squamosa* may be an effect of the limited number of loci used in our study (Hoffman and Goodisman 2007). Thus, the relationship between levels of caste-biased gene expression and the degree of caste dimorphism should be more fully addressed on a transcriptome-wide scale in the future.

There were readily apparent differences in the gene expression profiles of males and females in our data (Fig. 2.2A), as has been observed previously in taxa such as *D. melanogaster* (Ayroles et al. 2009). In hymenopteran insects, where males are haploid and females are diploid, differences in gene expression between the sexes arise both as a function of differential gene activation and differential gene dosage related to ploidy levels (Birchler et al. 2005; Mileyko et al. 2008; Mank 2009). Based purely on ploidy levels, one would expect many genes to be more highly expressed in females than males. However, hymenopterans may undergo dosage compensation, sometimes only in only limited tissues (Aron et al. 2005) or genes (Mank 2009). Furthermore, recent theory suggests that changes in gene copy number may result in disproportionally large changes in gene expression levels in the context of biological gene interaction networks (Mileyko et al. 2008). Thus, it remains unclear precisely how gene dosage in haplodiploid organisms alters gene regulation in males and females.

It is notable that we observed widespread female-biased gene expression at the loci we assayed (Figs. 2.2A, 2.2C). This is particularly evident for *P. exclamans,* which exhibited increased female bias relative to the three vespine species in our study (Fig. 2.2A). This is consistent with prior research demonstrating that sex-specific gene expression patterns change in line with increasing evolutionary divergence (Zhang et al. 2007). However, we do not expect that female bias in our data could arise from differences in ploidy levels because the total numbers of gene transcripts were held constant for each of our samples using global cDNA normalization. Thus, this female bias potentially reflects differences in the variance in gene expression levels among genes within sexes, where relatively few male-biased loci contribute disproportionately to the total pool of male gene transcripts. Alternatively, this could reflect our limited sampling of loci.

As with ploidy, effects of experimental design must be considered to appropriately interpret our results. In particular, given the limited number of genes in our study our conclusions must be vetted against the possible effects of ascertainment bias. First, our methods preferentially favored genes with high levels of sequence conservation because we selected only those loci with enough conservation to facilitate the design of cross-species primers (see experimental procedures). For example, 12 of the 21 genes were chosen as candidates for primer development based on sequence similarity to genes in an ant with developmental gene expression data (Goodisman et al. 2005). Accordingly, we cannot say whether the observed differences in regulatory lability between caste-biased genes and sex-biased genes would hold true for rapidly evolving genes. Second, EST sequences used for primer development in this study were chosen based on preliminary analysis according to diverse biological interests, including differential expression between queen and worker castes. Specifically, nine genes in our dataset had preliminary gene expression data for *V. squamosa* (Hoffman and Goodisman 2007). Of these, four genes were previously found to exhibit caste-biased gene expression patterns and five had no *a priori* predictions with respect to caste. None of the genes in our study had *a priori* predictions with respect to sex. Regardless, we do not believe that the selection of genes previously shown to display caste-biased expression has spuriously given rise to more labile regulation for caste-biased genes than sex-biased genes because two of the three predicted caste-biased genes analyzed in both *Vespula* species exhibited conserved expression patterns among castes (contigs 504 and 583; Figs. 2.1 and 2.2D).

Functional significance of gene expression

Several genes in our study displayed particularly informative patterns of expression in the context of functional annotation and prior studies of the molecular differences underlying eusocial insect castes. For instance, orthologs of a few genes have previously been linked to the division of labor in eusocial insects. Namely, *vitellogenin* is a yolk protein believed to play an important role in egg production and antioxidant function related to queen longevity and is strikingly upregulated in queens of diverse eusocial insect taxa (Table 2.2; Corona et al. 2007; Graff et al. 2007; Scharf et al. 2005; Sumner et al. 2006). Here we find that *vitellogenin* is dramatically upregulated in queens relative to workers and males in *V. maculifrons* and upregulated in *D. maculata* workers relative to males (Fig. 2.2). Although *V. squamosa* data is not presented for *vitellogenin* because its dilution curve did not meet our standards (see experimental procedures), *V. squamosa vitellogenin* expression patterns were qualitatively similar to *V. maculifrons*, with dramatically higher expression in queens than males or workers. The upregulation of *vitellogenin* in *D. maculata* workers relative to males may be explained by high levels of worker reproduction in this species (Foster et al. 2001). In contrast, *V. maculifrons* and *V. squamosa* workers rarely reproduce (Kovacs and Goodisman 2007) and workers and males in these species show limited *vitellogenin* expression (Fig. 2.1).

Superoxide dismutase (SOD) is another well-studied gene that has been hypothesized to play a key role in aging because of its association with longevity in *Drosophila* (Orr and Sohal 1994). Although SOD is upregulated in brains of queens in the honeybee *A. mellifera* (Table 2.2; Grozinger et al. 2007), a SOD ortholog was found to be down-regulated in queens relative to workers and males in the ant *Lasius niger* (Parker et al. 2004). In *V. maculifrons* and *V. squamosa*, SOD is not significantly differently expressed between queens, workers, and males (Fig. 2.1). However, SOD expression likely varies among tissues and over time. We did not investigate variation in expression in different tissues or developmental stages. Thus, we can only state that our results generally support the view that antioxidant activity by SOD does not contribute to longevity in social insects (Parker et al. 2004), but further work is needed to fully address this question in vespid wasps.

Finally, CG4692, which has been found to be upregulated in *A. mellifera* queens, is also upregulated in queens of *V. maculifrons* (Table 2.2; Grozinger et al. 2007). CG4692 is associated with proton transport, which may bear functional significance to caste differences in line with the widely observed queen upregulation of metabolic genes (Cristino et al. 2006; Grozinger et al. 2007; Smith et al. 2008).

A few male-biased genes in our study also had noteworthy functions. Two genes associated with musculature appear to be consistently upregulated in males of vespine wasps: orthologs of *wings up A* and, most strikingly, *tropomyosin 2* (Table 2.1). Remarkably, haploid males in hymenopteran social insects have doubled the nuclear DNA content of muscles to effectively restore diploidy in these tissues (Aron et al. 2005; see above). This important evolutionary innovation, known as endoreduplication, may have occurred in response to the demands of flight for dispersal and mating. Our results are consistent with a scenario in which endoreduplication results in higher levels of transcription (Edgar and Orr-Weaver 2001).

Conclusion

To our knowledge, this is among the first studies to investigate evolutionary variation in gene expression among both castes and sexes. Our results suggest that caste-biased gene expression may be more labile than sex-biased gene expression and demonstrate the utility of social insects in addressing questions of the molecular evolution of phenotypic dimorphisms. Several directions for future investigations are suggested by our study. Firstly, differences in gene regulation that give rise to sexual dimorphism in hymenopteran eusocial insects should be more intensively studied (e.g., Hoffman and Goodisman 2007). Secondly, more detailed comparisons of caste-biased gene expression in taxa with different levels of morphological differentiation between castes (such as vespines and polistines) will provide valuable insight into the evolutionary elaboration of eusociality (Hunt 2007; Toth et al. 2007, 2009). Finally, we recognize that our study is limited in the number of loci examined. With the rapid advancement of comparative genomic resources for eusocial hymenopterans, future studies can test our hypotheses on a transcriptome-wide scale (Smith et al. 2008). Indeed, our hope is that this study will act as an impetus for others to further examine patterns of gene expression in the castes and sexes of vespid wasps and other eusocial insects.

Experimental Procedures

Wasp samples

We sampled adult wasps from colonies collected in the metro Atlanta, Georgia (USA) area from October to November 2007, with the exception of one *D. maculata* colony collected in August 2007. We sampled three males, three pre-reproductive queens (gynes; referred to as 'queens' throughout the manuscript), and three workers from one *V. squamosa* and two *V. maculifrons* colonies. We also sampled individuals from two *D. maculata* colonies. Three workers were sampled from one of these colonies, and three males and three workers were sampled from the other. Finally, we collected two *P. exclamans* colonies and sampled three late-season females and three males from each colony. Live wasps were flash frozen using liquid nitrogen and stored at -80° C.

EST sequencing and primer development

Gene-specific PCR primers were designed to assess gene expression levels. Expressed sequence tags (ESTs) for putative target genes were sequenced from *V. squamosa* cDNA libraries following Hoffman and Goodisman (2007). Chromatogram files were trimmed for quality and vector contamination using Sequencher (Gene Codes, Ann Arbor, MI). Contigs were assembled from a combination of ESTs described in Hoffman and Goodisman (2007; GenBank accession numbers EG325041 - EG327184) and ESTs sequenced thereafter (GenBank accession numbers GW787760 - GW792031) using CAP3 (Huang and Madan 1999) with default settings other than percent identity (set to P = 90) and overlap (set to O = 40). EST sequences used for primer development in this study were chosen based on preliminary analysis according to diverse biological interests, including differential expression between gyne and worker castes, differential expression between larvae and adults, genes upregulated in multiple developmental stages and castes (Hoffman and Goodisman 2007), and those with orthologs previously studied in ants (Goodisman et al. 2005).

Gene-specific primers were designed with the goal of universal amplification among orthologs in vespid wasp taxa. BLAST alignments (Altschul et al. 1997) between *V. squamosa* sequences and the honeybee *A. mellifera* were used to identify putatively conserved regions for primer placement. Primers were designed using Primer3 (Rozen and Skaletsky 2000) after masking repeats with RepeatMasker (Smit et al. 1996-2004). We used primers that amplified in all four taxa during non-quantitative PCR tests.

We annotated EST sequences from *V. squamosa* based on sequence similarity using gene ontology (GO) terms. Each sequence served as a query to search the GO database (release 2009-06-23) using the AmiGO search tool (Carbon et al. 2009) with a similarity threshold of 1e-5 and BLAST filtering enabled. If a hit with *Drosophila melanogaster* was found, its annotation was used. Otherwise, the best hit with GO biological process annotation in another, more distantly related species was used. GO biological process terms and gene names were assigned to each *V. squamosa* EST.

Quantitative real-time reverse transcription PCR

We measured the level of gene expression in each sample using quantitative real-time reverse transcription PCR (QRT-PCR; Heid et al. 1996). First, RNA extraction was performed using whole bodies. Briefly, tissue was homogenized using a liquid nitrogen-chilled mortar and pestle immediately following removal from freezer. RNA was extracted using TRI Reagent (Molecular Research Center, Inc., Cincinnati, OH) following standard RNA isolation protocol. For large individuals, a fraction of the homogenate was used so as not to exceed ~10% of TRI Reagent volume. Total RNA concentration was determined by UV spectrophotometry and normalized across samples. RNA was then treated with DNase to remove genomic DNA contamination using the TURBO DNA-*free* kit (Ambion, Austin, TX). Removal of genomic DNA was verified by PCR and DNase treatment was repeated if necessary.

We synthesized first strand cDNA using SuperScript III reverse transcriptase (Invitrogen, Carlsbad, CA) and a poly $dT_{14\text{-}18}$ primer mix in a 50 µl reaction volume. cDNA concentration was normalized across samples using agarose gel electrophoresis (Goodisman et al. 2005) because no empirically verified exogenous controls exist for vespid wasps. Briefly, we loaded three different volumes of each sample (6 µl, 4 µl, and 2 µl) into a 192-well agarose gel. cDNA intensity was measured from gel images using the ImageJ program (Rasband 1997-2009). A linear regression was generated based on the dilutions of each sample (Mean $R^2 = 0.99$), and the intensity predicted from the linear regression model at a 6 µl volume was used to normalize cDNA concentrations across samples. Nine samples with low cDNA concentrations had to undergo a second cDNA synthesis and subsequent normalization to a subsample of previously normalized cDNA. We used the intensity of size-standard ladders distributed throughout the gel to ensure uniform background intensity. A 1:10 dilution of normalized cDNA was aliquoted, stored at -20 °C, and used for quantitative PCR within 2-3 days of thawing.

We performed QRT-PCR on StepOnePlus and Prism 7000 sequence detection systems from Applied Biosystems (Foster City, CA). All samples of *V. maculifrons* and *V. squamosa* for a given gene were run in triplicate on the same 96-well plate. Samples on every plate were run alongside one species-specific dilution curve consisting of five to six dilutions (1:2, 1:4, 1:8, 1:16, 1:32, 1:64) of cDNA with a higher starting concentration than those used for expression analysis. This curve was used to control for differences in PCR efficiency between species and to generate a standard curve for relative quantification between samples within a species (Applied Biosystems 2001). We did not make direct comparisons in intensity across species or genes. For a given primer set, all samples of *D. maculata* and *P. exclamans* were run in the same manner on a second 96-well plate.

Statistical analysis

The starting transcript abundance for each sample, relative to other samples from the same species for the same gene, was estimated from cycle threshold (Ct) values (calculated by StepOnePlus and Prism 7000 sequence detection system software) and the applicable standard curve (Applied Biosystems 2001). In order to ensure that standard curves were of acceptable technical quality and that cDNA amplification was of acceptable efficiency, only data calibrated with standard curves exhibiting R^2 values > 0.90 are presented.

A nested ANOVA was used to test for significant differences in gene expression between castes and between individuals within castes for each species, treating males as a 'caste' to incorporate sex differences. Relative gene expression levels (the relative Ct values for samples of a given species after making intensity adjustments according to the species- and plate-specific standard curve) were used to calculate least squares means for each caste using a standard least squares model. For visualization and subsequent analysis, least squares means of each caste within a species were normalized such that the highest relative expression level for a given gene was assigned a value of 1.0 and expression levels of other castes became a fraction thereof. For *Vespula* species, which had three sampled castes, a post-hoc Tukey HSD test on least squares means was used to determine which castes differed significantly in expression level for a given gene. ANOVA, standard least squares calculations, and post-hoc tests were performed using the JMP statistical software package (SAS Institute Inc, Cary, NC).

Heatmaps with dendrograms were generated from normalized relative expression levels using R (R Development Core Team 2008). Euclidean distances between rows and columns were used to generate heatmaps, and dendrograms were computed by performing a hierarchical cluster analysis using a set of dissimilarities for the *n* objects being clustered, according to the 'complete' method.

Apis mellifera ortholog analysis

We compared gene expression differences observed between *Vespula* queens and workers to gene expression differences observed between *A. mellifera* queens and workers (Barchuk et al. 2007; Grozinger et al. 2007). *A. mellifera* orthology was determined by sequence similarity; BLASTX was used to determine sequence similarity of *V. squamosa* sequences to those in a database composed of proteins representing the complete *A. mellifera* official gene set (version 1; Honeybee Genome Sequencing Consortium 2006). The best hit for each *V. squamosa* sequence was taken and the corresponding *A. mellifera* gene identifier was used to assign a gene name. We then used *A. mellifera* gene IDs to assign relative gene expression in brains of *A. mellifera* queens and sterile workers (Grozinger et al. 2007) and in *A. mellifera* whole-bodies of queen-destined and worker-destined larvae (Barchuk et al. 2007).

Acknowledgements

This research was supported by the U. S. National Science Foundation (grant DEB 0640690). We thank J.L. Kovacs and S.V. Yi for helpful discussion and two anonymous reviewers for comments that helped to improve the manuscript.

CHAPTER 3

DNA METHYLATION IS WIDESPREAD AND ASSOCIATED WITH DIFFERENTIAL GENE EXPRESSION IN CASTES OF THE HONEYBEE, *APIS MELLIFERA*[2]

Abstract

The recent, unexpected discovery of a functional DNA methylation system in the genome of the social bee *Apis mellifera* underscores the potential importance of DNA methylation in invertebrates. However, the extent of genomic DNA methylation and its role in *A. mellifera* remain unknown. Here we show that genes in *A. mellifera* can be divided into two distinctive classes, one with low-CpG dinucleotide content and one with high-CpG dinucleotide content. This dichotomy is explained by the gradual depletion of CpG dinucleotides, a well-known consequence of DNA methylation. The loss of CpG dinucleotides associated with DNA methylation may also explain the unusual mutational patterns in *A. mellifera* that lead to AT rich regions of the genome. Detailed investigation of this dichotomy implicates DNA methylation in *A. mellifera* development: high CpG genes, which are predicted to be hypo-methylated in germlines, are enriched with functions associated with developmental processes, while low CpG genes, predicted to be hyper-methylated in germlines, are enriched with functions associated with basic biological processes. Furthermore, genes more highly expressed in one caste than another are overrepresented among high CpG genes. Our results highlight the potential

[2] Elango N, Hunt BG, Goodisman MAD, Yi SV. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. Proc Natl Acad Sci USA 106:11206-11211. Navin Elango and Brendan Hunt contributed equally to this manuscript.

significance of epigenetic modifications such as DNA methylation in developmental processes in social insects. In particular, the pervasiveness of DNA methylation in the genome of *A. mellifera* provides fertile grounds for future studies of phenotypic plasticity and genomic imprinting.

Introduction

DNA methylation occurs in the genomes of a wide array of bacteria, plants, fungi and animals (Klose and Bird 2006; Suzuki and Bird 2008). In particular, the methylation of cytosine bases represents an important epigenetic mark that affects gene expression in diverse taxa (Hendrich and Tweedie 2003; Klose and Bird 2006). Despite the phylogenetically widespread and ancient origin of DNA methylation, genomic patterns of methylation show considerable variation (Suzuki and Bird 2008). For example, while vertebrate genomes tend to show extensive levels of DNA methylation, many invertebrate genomes display reduced or minimal levels of methylation (Field et al. 2004; Klose and Bird 2006; Suzuki and Bird 2008). Variation in genome methylation patterns is of great interest because it suggests the role of DNA methylation is not strictly conserved among species. Thus, additional information on the nature and extent of DNA methylation in diverse taxa continues to be a valuable resource for understanding the role of this DNA modification in eukaryotes (Hendrich and Tweedie 2003; Schaefer and Lyko 2007; Suzuki and Bird 2008).

Recent research identified a functional DNA methylation system in a social insect, the honeybee *Apis mellifera* (Wang et al. 2006). Social insects are among the most successful of animal taxa (Wilson 1971; Strassmann and Queller 2007). The success of social insects stems from the cooperative behaviors displayed by society members. In particular, members of social insect colonies belong to different castes, which undertake distinct tasks (Oster and Wilson 1978). For example, the defining feature of hymenopteran social insects (ants, some bees, and some wasps) is a reproductive division of labor, whereby individuals of the queen caste reproduce while members of the worker caste defend the nest, forage, and rear the young. This division of individuals into alternate castes represents a key evolutionary transition that allowed social insects to come to dominate many terrestrial ecosystems (Maynard Smith and Szathmary 1995; Keller 1999).

Remarkably, DNA methylation appears to be directly associated with the differentiation of castes in *A. mellifera* (Kucharski et al. 2008; Maleszka 2008). Kucharski et al (2008) demonstrated that down-regulation of a key DNA methyltransferase (Dnmt3) in developing *A. mellifera* larvae resulted in profound changes in caste developmental trajectories. Accordingly, DNA methylation may represent an important mechanism facilitating the evolution of social systems (Moczek and Snell-Rood 2008).

Despite the potential importance of DNA methylation, the genome-wide patterns of methylation within the *A. mellifera* genome are poorly understood. This is unfortunate, because knowledge of the patterns of DNA methylation in the *A. mellifera* genome is critical for assessing its role and significance in this species (Wang et al. 2006; Kucharski et al. 2008). Moreover, linking molecular changes such as DNA methylation with evolution and development of social phenotypes remains one of the major challenges in understanding sociality (Robinson et al. 2005; Goodisman et al. 2008). In this study, we investigated the nature of DNA methylation in *A. mellifera* by analyzing global patterns of methylation using computational methods and comparing them with experimental results in the honeybee and in other species. We find that DNA methylation is widespread and has played a critical role in *A. mellifera* genome evolution. Methylation is also associated with important developmental processes including caste formation.

Results

Depletion of CpG dinucleotides suggests widespread gene methylation in *A. mellifera*

We used normalized CpG content ($CpG_{O/E}$) to infer the pattern of DNA methylation in *A. mellifera*. $CpG_{O/E}$ is a robust measure of the level of DNA methylation on an evolutionary timescale due to specific mutational mechanisms of methylated cytosines (Saxonov et al. 2006; Suzuki et al. 2007; Weber et al. 2007; Elango et al. 2008). Briefly, methylated cytosines are hypermutable due to their vulnerability to spontaneous deamination, which causes a gradual depletion of CpG dinucleotides from methylated regions over time (Bird 1980). Consequently, genomic regions that are subject to heavy germline DNA methylation (hyper-methylated) lose CpG dinucleotides over time and have lower than expected $CpG_{O/E}$. In contrast, regions that undergo little germline DNA methylation (hypo-methylated) maintain high $CpG_{O/E}$. This measure has been successfully used to indirectly measure historical DNA methylation levels (Bird 1980; Suzuki et al. 2007; Weber et al. 2007; Elango and Yi 2008). In particular, the pattern of DNA methylation inferred from $CpG_{O/E}$ corresponds well to the actual pattern of DNA methylation in diverse taxa such as human and sea squirt (Suzuki et al. 2007; Weber et al. 2007).

We first examined the distribution of $CpG_{O/E}$ in several insect genomes. We focused on analyses of genes, because annotation of other genomic regions (e.g., intergenic regions and non-coding functional elements) in insect genomes other than *Drosophila melanogaster* is far from complete.

The level of methylation in the fly *D. melanogaster* is scarce, and its genome lacks critical DNA methyltransferases (Urieli-Shoval et al. 1982; Suzuki and Bird 2008). Accordingly, $CpG_{O/E}$ in *D. melanogaster* genes follow an approximately normal distribution with a mean around 1 (Fig. 3.1A). Analyses of other published insect genomes, including the beetle *Tribolium castaneum* and the mosquito *Anopheles*

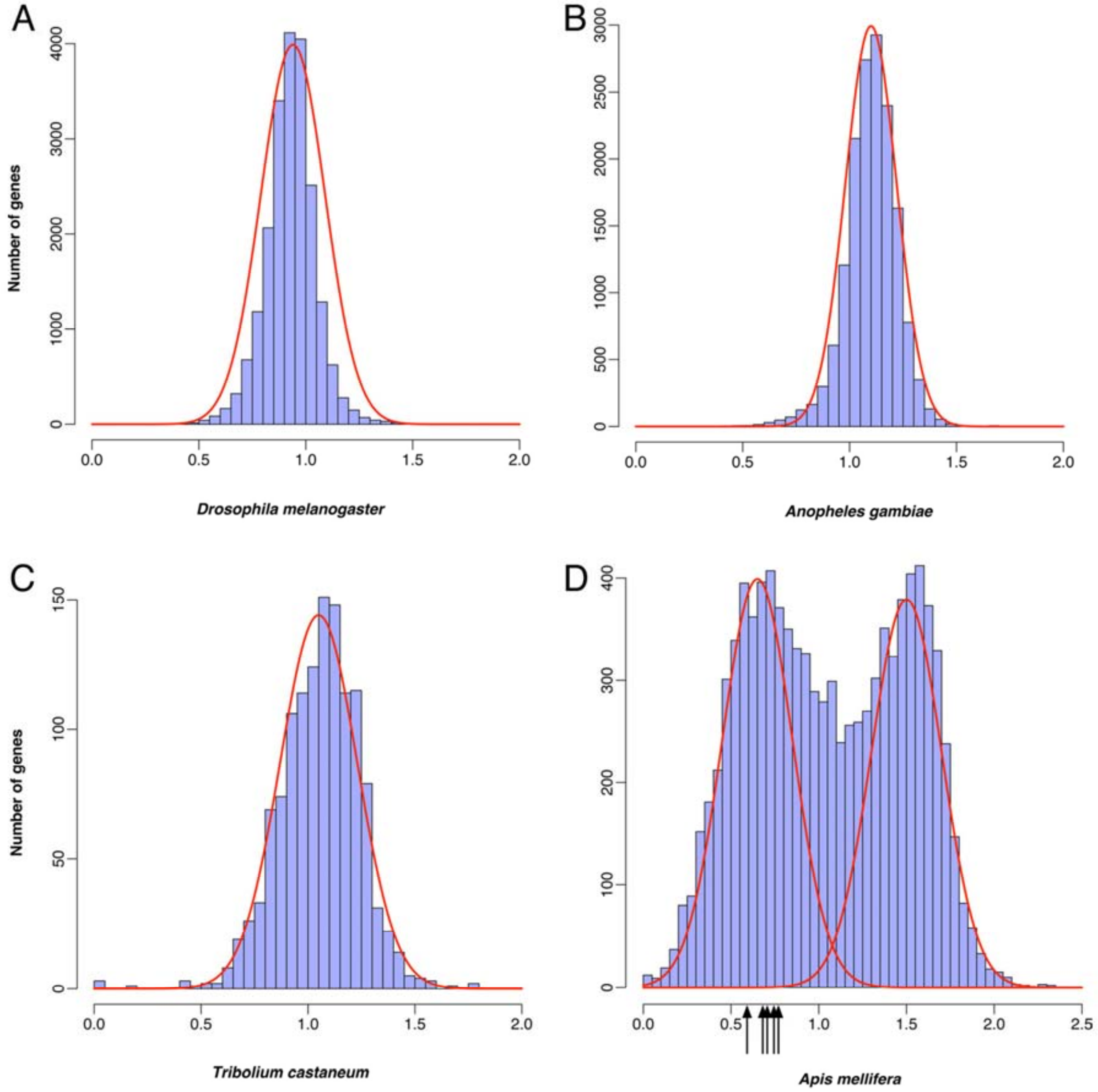

Figure 3.1. Contrasting patterns of DNA methylation in *A. mellifera* genes compared to those in other insects, as measured by normalized CpG dinucleotides contents (*CpG*$_{O/E}$). The Y-axis depicts the number of genes with the specific *CpG*$_{O/E}$ values given on the X-axis. The distribution of *CpG*$_{O/E}$ in A) *D. melanogaster*, B) *A. gambiae*, C) *T. castaneum* genes all show 'unimodal' distributions, reflecting a relative lack of DNA methylation in those species. In contrast, the distribution of *CpG*$_{O/E}$ in D) *A. mellifera* genes is bimodal, likely demonstrating the effects of DNA methylation of CpG dinucleotides (see text). The arrows show the position of the five genes [GB16767 (*CpG*$_{O/E}$=0.56), GB19399 (0.66), GB18099 (0.67), and GB12504 (0.75), XP_001121083 (0.71)] shown to be methylated in a previous study (Wang et al. 2006). Note that the gene GB15223 could not be mapped using our experimental procedure.

gambiae, yield similar patterns (Fig. 3.1 B & C). Thus, genes in these insects exhibit little evidence of DNA methylation according to mutational decay of CpG dinucleotides.

In contrast, we find that $CpG_{O/E}$ of *A. mellifera* genes exhibit a striking bimodal pattern that is best explained by a mixture of two distinct distributions (Fig. 3.1D, see Methods). $CpG_{O/E}$ of approximately half of *A. mellifera* genes falls into a distribution with a remarkably high mean of 1.50 (SD = 0.20), similar to the genomic background (see also The Honeybee Genome Sequencing Consortium 2006). Surprisingly, the other half of honeybee genes formed a distinctive distribution with the mean much lower than the genome average (mean = 0.55, SD = 0.20, Fig. 1D). Low $CpG_{O/E}$ is a signature of DNA methylation, which is the only known mechanism to selectively target CpG dinucleotides in animal genomes. In the remainder of the paper, we will refer to the genes belonging to the first category as 'high CpG' genes and those within the latter category as 'low CpG' genes.

Given that the nucleotide composition of the honeybee genome is extremely heterogeneous (Honeybee Genome Sequencing Consortium 2006; Jorgensen et al. 2007), we tested if the observed bimodality in CpG content arose from a bias in nucleotide composition. Previous studies have shown a positive correlation between GC content and $CpG_{O/E}$ (Duret and Galtier 2000; Fryxell and Zuckerkandl 2000; Elango et al. 2008). We also find that the GC content and CpG content are strongly correlated in honeybee genome (Kendall's correlation coefficient $\tau = 0.32$, $P < 10^{-15}$). Thus it is possible that the distribution of CpG content reflects the influence of GC content. To test this possibility, we investigated the distribution of normalized GpC content ($GpC_{O/E}$). GpC dinucleotides have the same C and G composition as CpG dinucleotides, yet are not targeted by DNA methylation (Razin and Riggs 1980; Wang et al. 2006). For this reason, $GpC_{O/E}$ is often used as an indicator of nucleotide composition bias while controlling for the influence of DNA methylation (Fryxell and Moon 2005; Elango and Yi 2008).

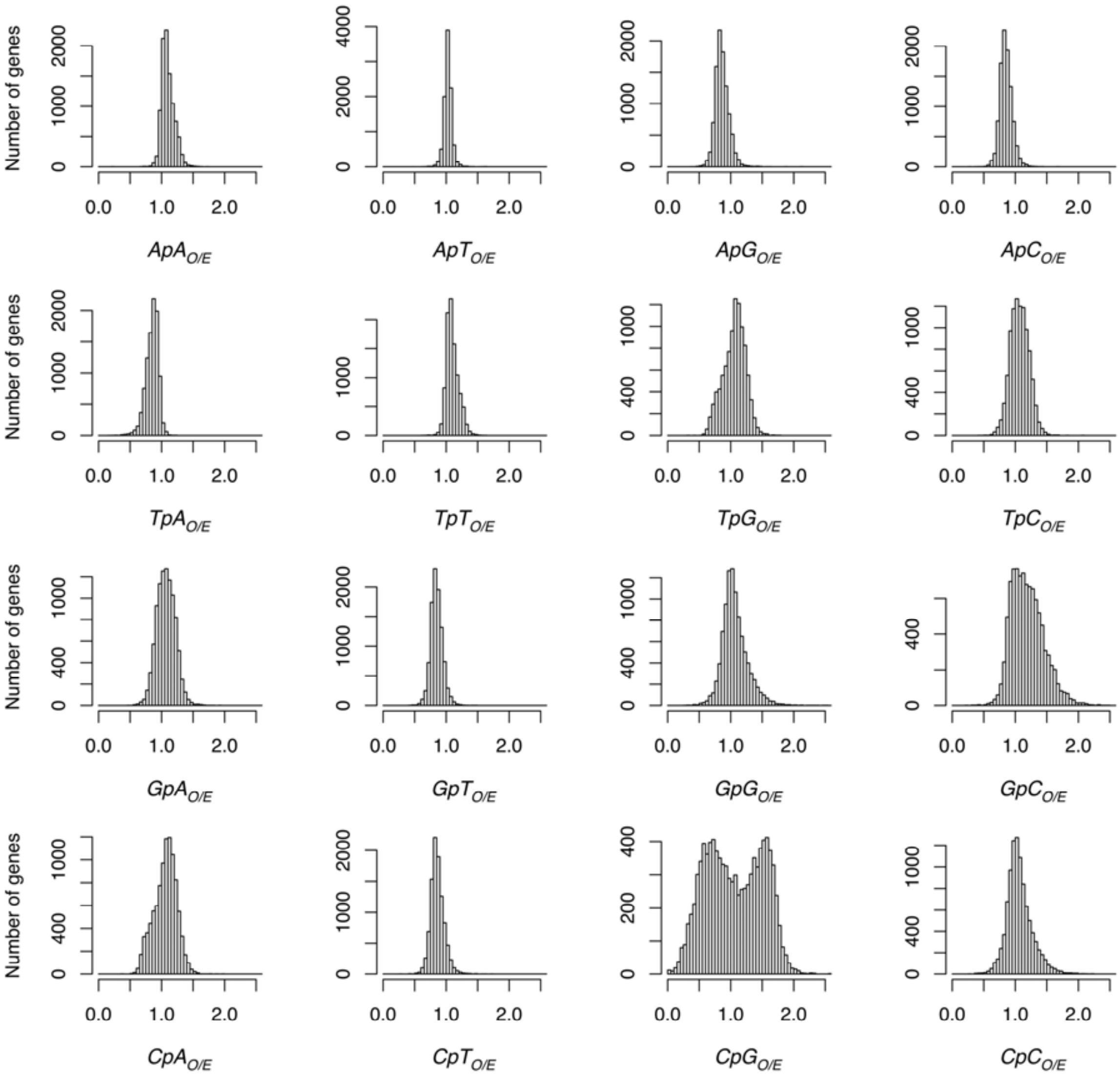

Figure 3.2. Distribution of normalized dinucleotide contents in *A. mellifera* genes. Only $CpG_{O/E}$ exhibits a distinct bimodal distribution, consistent with the mutational processes arising from the action of DNA methylation on CpG dinucleotides.

We find that the distribution of $GpC_{O/E}$ in *A. mellifera* is unimodal (Fig. 3.2). The distribution of normalized GpC content in *D. melanogaster* is also unimodal, as expected (results not shown). Moreover, analyses of all other dinucleotides in *A. mellifera* clearly show that bimodality is exclusive to CpG dinucleotides (Fig. 3.2). These observations indicate that the observed bimodality of $\text{CpG}_{O/E}$ in honeybee genes is caused by the difference in levels of germline DNA methylation on an evolutionary timescale: hyper-methylated genes exhibit CpG depletion, while hypo-methylated genes exhibit high CpG content.

Further support for the link between CpG content and levels of DNA methylation comes from analysis of $CpG_{O/E}$ profiles of genes in the distantly related invertebrate, *Ciona intestinalis*. *C. intestinalis* is the only invertebrate whose genomic pattern of DNA methylation has been experimentally investigated at this time (Simmen et al. 1999; Suzuki et al. 2007), and it has been demonstrated that $CpG_{O/E}$ levels in *C. intestinalis* correspond to actual levels of DNA methylation (Suzuki et al. 2007). Furthermore, *A. mellifera* genes shown to be methylated in a previous study (Wang et al. 2006) are all found in the low CpG class, as predicted under the proposed model (Fig. 3.1D).

To determine if DNA methylation is widespread in genomic regions other than genes, we analyzed the distribution of $CpG_{O/E}$ of the entire *A. mellifera* genome, as well as putative promoter regions (500bps or 1000bps upstream of transcription start sites), untranslated regions, and transposable elements. These analyses demonstrate that the strong bimodality of $CpG_{O/E}$ is unique to amino-acid encoding sequences (Fig. 3.3). Only coding sequences harbor substantial portions of the low CpG class, bearing evolutionary signatures of DNA methylation. These results are consistent with the observation that CpG methylation in *A. mellifera* is predominantly found in exons (Wang et al. 2006). The pattern of CpG depletion in *A. mellifera* introns is also bimodal (Fig. 3.3), suggesting that some introns are methylated. However, the signal of bimodality is clearer when whole gene sequences (exons and introns) are analyzed, as expected if exons are primary

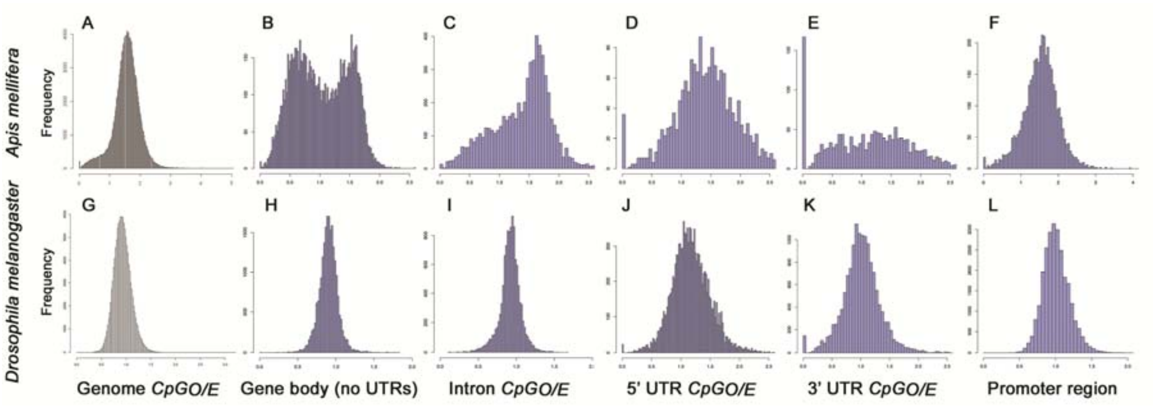

Figure 3.3. Distribution of CpG$_{O/E}$ in the *A. mellifera* genome (*A*), gene bodies without UTRs (coding sequences, exons, and introns) (*B*), introns (*C*), 5’ UTRs (*D*), 3’ UTRs (*E*), and promoters (defined as 1 kb upstream of transcription start sties) (*F*) and in the *D. melanogaster* genome (*G*), gene bodies without UTRs (coding sequences, exons, and introns) (*H*), introns (*I*), 5’ UTRs (*J*), 3’ UTRs (*K*), and promoters (*L*).

targets of DNA methylation (Wang et al. 2006).

Low CpG and high CpG genes are functionally distinct

The observed 'bimodality' of *A. mellifera* genes, which represents an intragenic evolutionary signature of methylation, correlates with gene function; genes found in low CpG and high CpG classes are involved in particular biological processes (Ashburner et al. 2000). Specifically, the low CpG and high CpG classes are enriched with distinct gene ontology categories (Table 3.1). Low CpG genes, predicted to be hyper-methylated in the germlines, are significantly enriched for terms related to metabolism and ubiquitous housekeeping functions of gene-expression and translation (Table 3.1). On the other hand, high CpG genes, which are predicted to be hypo-methylated in the germlines, exhibit a striking and significant enrichment of terms associated with a variety of developmental processes, cellular communication, and adhesion (Table 3.1).

Genes whose expression is strongly biased toward specific castes are enriched in the high CpG class

Social insect development is marked by the remarkable level of phenotypic plasticity. In particular, many hymenopteran social insect females can develop into distinctive queen and worker castes from identical genomes. Recent studies suggest that DNA methylation may regulate caste differentiation in *A. mellifera*, by silencing crucial genes involved in caste formation (Kucharski et al. 2008; Moczek and Snell-Rood 2008). If DNA methylation is truly involved in caste development, then genes that are over-expressed in a specific caste, or 'caste-specific' genes, may show preferential enrichment in low CpG or high CpG genes. We tested this prediction using a data set from a recent study that identified differential gene expression in brains of queens and sterile workers (Grozinger et al. 2007).

We first examined if genes that were identified as 'caste-specific' (at a 5% significance level) tend to be biased towards a specific CpG-content class (low or high CpG). We find that caste-specific genes tend to harbor more high CpG genes than

Table 3.1. Distinctive functional enrichment of low-CpG and high-CpG genes.

CpG class	GO biological process term	Accession	Fold enrichment in class	Significance[a]
LCG	macromolecule metabolic process	GO:0043170	1.13	3.91E-14
LCG	cellular metabolic process	GO:0044237	1.09	1.04E-11
LCG	metabolic process	GO:0008152	1.08	1.20E-10
LCG	primary metabolic process	GO:0044238	1.08	8.05E-09
LCG	cellular process	GO:0009987	1.04	2.83E-08
LCG	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0006139	1.17	2.85E-08
LCG	gene expression	GO:0010467	1.18	3.15E-08
LCG	RNA processing	GO:0006396	1.37	1.05E-07
LCG	biopolymer metabolic process	GO:0043283	1.12	1.42E-06
LCG	RNA metabolic process	GO:0016070	1.19	2.28E-06
HCG	multicellular organismal process	GO:0032501	1.32	1.20E-19
HCG	cell communication	GO:0007154	1.37	4.10E-16
HCG	organ development	GO:0048513	1.41	1.52E-11
HCG	system development	GO:0048731	1.35	1.54E-11
HCG	signal transduction	GO:0007165	1.35	1.71E-11
HCG	multicellular organismal development	GO:0007275	1.28	2.92E-11
HCG	biological adhesion	GO:0022610	1.77	7.40E-11
HCG	cell adhesion	GO:0007155	1.77	7.40E-11
HCG	anatomical structure development	GO:0048856	1.30	9.39E-11
HCG	developmental process	GO:0032502	1.23	1.99E-09

The top 10 significantly enriched terms for low-CpG and high-CpG classes are shown. GO biological process term enrichment is based on 1,781 *D. melanogaster* orthologs of *A. mellifera* high-CpG genes (1,230 with GO annotation) and 2,531 *D. melanogaster* orthologs of *A. mellifera* low-CpG genes (1,713 with GO annotation).

[a] Significance is denoted by a Benjamini correction for multiple testing.

expected based upon the distribution of caste-generic genes (genes that are not differently expressed between the castes; Table 3.2). The enrichment of high CpG genes increases with the bias toward caste-specific expression (Table 3.2, Fig. 3.4). Moreover, the degree of caste-specificity (measured as the absolute value of $\log_2$(queen/worker) gene expression) is significantly positively correlated with $CpG_{O/E}$ (Spearman's rank correlation $r_s = 0.1405$, $P = 2.80e\text{-}09$; Figure 3.4A).

We further expanded our analyses to genes implicated in *A. mellifera* caste differentiation identified by previous studies of gene expression (Corona et al. 1999; Evans and Wheeler 1999; Drapeau et al. 2006; Wheeler et al. 2006; Corona et al. 2007; Patel et al. 2007). Again we found that caste-specific genes overwhelmingly belong to the high CpG class (Table 3.3). Note that caste-specific genes are not necessarily those implicated solely in developmental processes: many of these genes perform basic biological functions (Table 3.3).

Discussion

The genomic distribution of 'normalized' contents of CpG dinucleotides ($CpG_{O/E}$) in *A. mellifera* stands in a sharp contrast to that in *D. melanogaster, T. castaneum* and *A. gambiae* (Fig. 3.1). In particular, approximately half of *A. mellifera* genes belong to a distinctive low CpG class (Fig. 3.1D). Given that (i) methylation in the honeybee is exclusive to CpG dinucleotides (Wang et al. 2006), (ii) only CpG content exhibits bimodal distribution (Fig. 3.2), and (iii) deamination of methylated CpGs to TpG (or CpA in the complementary strand) causes a GC to AT mutational bias in diverse taxa (Bird 1980; Elango and Yi 2008), these observations implicate DNA methylation in the origin of CpG bimodality. As far as we are aware, no other molecular mechanism is known to

Table 3.2. Caste-specific genes, which are differentially expressed between the queen and worker castes, are significantly over-represented in the high CpG class compared to caste-generic genes, whose expression patterns are not significantly different between the two castes.

Gene expression class	Significance threshold[a]	High-CpG class	Low-CpG class	χ^2 *P* value[b]	$\text{CpG}_{O/E}$, Mean ± SEM (Median)	Wilcoxon *P* value[c]
Caste-generic		474	488		1.0895 ± 0.0135 (1.0577)	
Caste-specific	P < 0.05	457	354	0.0034	1.1633 ± 0.0149 (1.2094)	0.0003
Caste-specific	P < 0.01	294	207	0.0008	1.1837 ± 0.0187 (1.2663)	4.39e-05
Caste-specific	P < 0.001	158	75	5.35e-07	1.2439 ± 0.0260 (1.3637)	1.07e-07
Caste-specific	P < 0.0001	75	19	2.96e-08	1.3274 ± 0.0352 (1.4042)	1.07e-07

The significance of the tests increases (i.e., *P* values decrease) as the significance threshold for genes considered caste-specific becomes more stringent.

[a] Significance threshold for caste-specific genes differentially expressed by queens and sterile workers in a pairwise comparison.
[b] *P* values from Pearson's χ^2 test of pairwise comparisons of the distribution among high-CpG and low-CpG classes of caste-specific genes versus the caste-generic class, after Yates's correction.
[c] *P* values of Wilcoxon's rank-sum test with continuity correction from pairwise comparisons of $\text{CpG}_{O/E}$ values for caste-specific genes versus caste-generic genes.

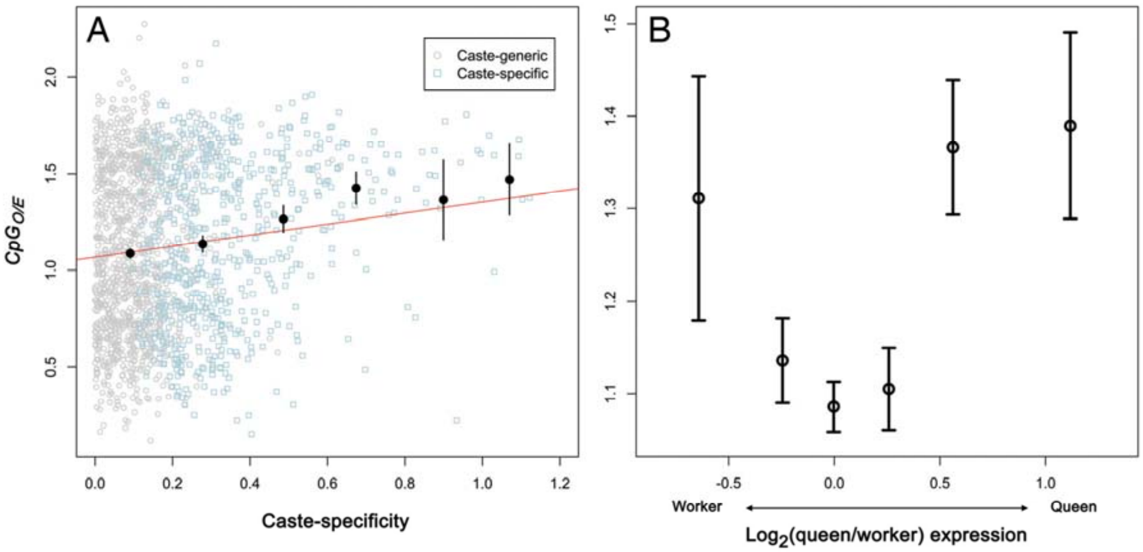

Figure 3.4. Caste-specific genes tend to have high *$CpG_{O/E}$*. (A) Caste-specificity (measured as the absolute value of $\log_2$(queen/worker) gene expression) is correlated with *$CpG_{O/E}$* (Spearman's rank correlation $r_s = 0.1405$, $P = 2.80\text{e-}09$). Mean values of *$CpG_{O/E}$* for equal windows of caste-specificity are shown as black dots with 95% confidence interval error bars. Ten outliers beyond caste-specificity values of 1.2 are excluded in the figure (but are included in calculations of correlation and model fitting). Points in the scatterplot are divided into 'caste-generic' and 'caste-specific' classes according to significant differences in expression between queens and workers (Grozinger et al. 2007). (B) The relationship between the values of $\log_2$-gene expression ratios between castes and *$CpG_{O/E}$* values shows that the enrichment of high CpG genes holds for genes that are either queen- or worker- specific genes. Genes expressed more highly in workers and queens have $\log_2$-ratios less than and greater than 0, respectively. The Y-axis shows the mean and 95% confidence intervals of each group. As the $\log_2$-expression ratios between castes become more extreme (either side of the X-axis), *$CpG_{O/E}$* tends to become more elevated.

influence CpG dinucleotides exclusively and is unique to the honeybee genome compared to other sequenced insect genomes.

Our results suggest a unique influence of DNA methylation in honeybee evolution that may help explain important genome characteristics. For instance, the honeybee genome is known for its overall low and heterogeneous distribution of GC contents (Honeybee Genome Sequencing Consortium 2006; Jorgensen et al. 2007). An earlier study also detected the presence of a mutational bias toward A and T nucleotides (AT) in GC-poor regions of honeybee genes (Jorgensen et al. 2007). However, the nature of such a mutational process has remained unknown. Here we show that $CpG_{O/E}$ exhibits a striking bimodality, and is strongly correlated with GC content in honeybee genes. These observations point to a link between the mutational bias towards AT and the depletion of CpG dinucleotides that occurs as a consequence of DNA methylation.

We also propose that, in addition to the mutational bias decreasing CpG contents in low CpG genes, other molecular mechanisms are operating to *increase* or *maintain* CpG contents in high CpG genes. $CpG_{O/E}$ of high CpG genes is higher than normalized contents of other dinucleotides and greater than the value of 1.0 expected under random association of C and G nucleotides (Fig. 3.2). Thus, a process that conserves or even increases CpG dinucleotides against mutational depletion may exist in the honeybee genome, especially in high CpG genes. The presence and nature of such processes in the *A. mellifera* genome need to be addressed in future studies.

We have demonstrated that a substantial number of *A. mellifera* genes harbor evolutionary signatures of DNA methylation. The question thereby arises, what is the functional significance of DNA methylation in *A. mellifera*? One potential role of DNA methylation is genomic imprinting. Genomic imprinting is an epigenetic mechanism by which the expression of a gene is influenced by the parent from which it is inherited. In mammalian systems, DNA methylation is implicated in genomic imprinting (Li et al. 1993; Jones and Takai 2001; Klose and Bird 2006). Social insects, especially those

Table 3.3. Genes identified as caste-specific from previous studies of gene expression and caste development in *A. mellifera* tend to belong (23 of 28; *P* < 0.005) to the hypo-methylated class (high-CpG).

Gene/gene family	**Function**	**Caste-biased expression**	*CpG*$_{O/E}$ class	Refs
AmIF-2$_{mt}$ translation initiation factor	Translation of mitochondrial-encoded mRNAs	Higher in queen larvae	0/1 High CpG	(Corona et al. 1999)
AmILP-2 insulin-like peptide	Regulation of growth/metabolism	Higher in workers than queens from second instar onward	1/1 High CpG	(Wheeler et al. 2006)
AmInR putative insulin-like peptide receptor family	Regulation of growth/metabolism	Higher in worker adults	2/2 High CpG	(Corona et al. 2007)
amTOR (target of rapamycin)	regulation of growth/metabolism	Higher in queen 3rd instar larvae, but not 5th instar larvae (RNAi linked to worker fate)	0/1 High CpG	(Patel et al. 2007)
Hexamerin family	Storage of amino acids for use in metamorphosis or by adults	Either more highly expressed in queen or worker larvae (based on 2 empirically analyzed genes)	3/4 High CpG	(Evans and Wheeler 1999)
vitellogenin	Yolk protein	Higher in queen adults	1/1 High CpG	(Corona et al. 2007)
Yellow/major royal jelly protein family	Sex-specific reproductive maturity among other functions	Primarily more highly expressed in workers, but some more highly expressed in queens (diverse tissue-dependent expression patterns)	16/18 High CpG	(Drapeau et al. 2006)

belonging to the haplodiploid Hymenoptera (social bees, social wasps, and ants), provide another intriguing context whereby imprinting may play an important role in mediating a wide array of behaviors (Haig 1992; Haig 2000; Queller 2003). We predict that imprinted genes, which should bear epigenetic marks (i.e., methylation) in the germlines, preferentially belong to the hyper-methylated low CpG class. Since our results demonstrate that nearly half of *A. mellifera* genes belong to the low CpG class, many genes are candidates for studies of imprinting in *A. mellifera*. In this respect, it is of great interest to note that DNA methylation is widespread in haplodiploid hymenopteran social insects (Kronforst et al. 2008). Thus, information on CpG depletion for specific sets of genes in social insects provides fertile grounds for future imprinting studies in a comparative context.

Our analyses indicate that methylation primarily targets gene bodies (exons and introns) in the *A. mellifera* genome. Moreover, methylated and non-methylated regions co-exist. Such a pattern is qualitatively similar to that found in echinoderms (e.g., sea urchin) and urochordates (e.g., sea squirt) (Tweedie et al. 1997; Suzuki et al. 2007; Suzuki and Bird 2008). In the sea squirt *C. intestinalis,* where genomic methylation has been examined in detail, it has been proposed that the primary role of DNA methylation is to suppress spurious transcription of genes that are broadly expressed across tissues, with intermediate expression levels (Suzuki et al. 2007; Suzuki and Bird 2008). Our observation that genes that tend to be methylated are involved in basic biological processes (Table 3.1) supports this idea.

We found that low and high CpG classes are populated with genes belonging to distinctive functional categories (Table 3.1). Low CpG genes are often involved in metabolic processes and nucleotide processing, which can be regarded as ‘basic biological processes’. On the other hand, a high proportion of high CpG genes, which are predicted to be hypo-methylated, are involved in development. This finding is particularly intriguing when combined with the results of recent studies implicating DNA methylation in the regulation of phenotypic plasticity in social insects (Kucharski et al. 2008; Moczek and Snell-Rood 2008).

Interestingly, we found that genes that are over-expressed in a specific caste are more frequently found in the hypo-methylated class (Table 3.2 & 3.3, Fig. 3.4). However, it is noteworthy that not all caste-specific genes are found in the high CpG class (Table 3.3); for example, genes associated with metabolism are frequently differentially expressed between castes (Evans and Wheeler 2000; Cristino et al. 2006; Wolschin and Amdam 2007), but are overrepresented in the low CpG class. Thus, the enrichment of caste-specific genes in the high CpG class is particularly striking.

Previous studies in *A. mellifera* have also uncovered associations between cis-regulatory motifs, social behavior, and caste-development (Cristino et al. 2006; Sinha et al. 2006). Thus, cis-regulatory elements represent a putative global control mechanism for caste-specific gene expression. The suggested significance of cis-regulatory elements, coupled with the finding that methylation can regulate caste fate (Kucharski et al. 2008), gives rise to the possibility that methylation interacts with regulatory elements to differentiate developmental pathways. However, methylation of cis-regulatory elements themselves may not be a major mechanism underlying caste differences in *A. mellifera* because our results suggest that methylation is primarily limited to gene bodies (Fig. 3.3).

Why are caste-specific genes preferentially found in the high CpG class? We hypothesize that high CpG genes in *A. mellifera* are generally more prone to epigenetic modulation compared to low CpG genes. Large-scale analyses of methylation patterns in mammals repeatedly show that a subset of high CpG promoters, particularly those associated with developmental processes, exhibit significant epigenetic flexibility, meaning that they are methylated in some tissues or developmental stages while not methylated in others (Illingworth et al. 2008; Meissner et al. 2008). Furthermore, a class of mammalian genes with high CpG promoters achieves complex, tissue-specific gene expression via pliable transcriptional regulation (N. Elango and S. Yi, unpublished data). Our observation that caste-specific genes tend to be enriched in the high CpG class runs parallel to the above findings from mammals, and may share similar underlying molecular mechanisms. Caste-specific genes must be activated or inactivated based on environmental input to proceed along different developmental paths; high CpG content of caste-specific genes may facilitate such modulation, similar to some high CpG promoters in mammalian genomes.

Conclusions

The pattern of DNA methylation in insect genomes varies greatly (Field et al. 2004). Here we show that the genome of *A. mellifera* can be divided into two distinct classes based upon the level of CpG depletion. Several pieces of evidence suggest that DNA methylation is the causative mechanism for the observed bimodality. In particular, our prediction correctly assigns all genes that were shown to be methylated in a previous study (Wang et al. 2006) to the low CpG class. Our results suggest that DNA methylation regulates development, as seems to be the case in numerous other taxa (Jones and Takai 2001; Bird 2002; Li 2002; Klose and Bird 2006). In fact, DNA methylation is believed to play a critical role in caste differentiation (Kucharski et al. 2008). Our analyses of caste-specific genes provide support to this idea, but demand future studies and experimental verification of caste-specific gene expression and DNA methylation.

The social Hymenoptera are ideal for studying the evolution and development of phenotypic plasticity because they comprise diverse taxa with multiple independent evolutionary origins of specialized queen and worker castes (Hughes et al. 2008). The honeybee provides an important first look into the genome of a social hymenopteran insect (Honeybee Genome Sequencing Consortium 2006), but the genomes of an estimated 10-20 social insects and related species are likely to be sequenced in the next 10 years (Smith et al. 2008). Comparative genomic analyses of evolutionary methylation signatures and experimental verification will more fully elucidate the evolutionary history and functional roles of DNA methylation in this important group.

Materials and Methods

Genome sequences and annotations

Genome sequences and gene annotations of *A. mellifera, A. gambiae* and *D. melanogaster* were downloaded from the UCSC genome browser (genome builds *apimel2, anoGam1* and *dm3*). The genome sequence and gene annotation of *T. castaneum* was downloaded from BeetleBase (www.beetlebase.org). Repetitive elements were annotated using the RepeatMasker program (Smit et al. 1996-2004).

Measurement of normalized CpG content and tests for bimodality

The 'normalized CpG content' ($CpG_{O/E}$) is a metric of depletion of CpG dinucleotides, normalized by G and C nucleotide content (GC content) of the specific region of interest. $CpG_{O/E}$ for each gene is defined as

$$CpG_{O/E} = \frac{P_{CpG}}{P_C * P_G}$$

where $\boldsymbol{P}_{CpG}$, $\boldsymbol{P}_c$ and $\boldsymbol{P}_G$ are the frequencies of CpG dinucleotides, C nucleotides, and G nucleotides, respectively, estimated from each gene. The normalized content of other dinucleotides were measured in a similar manner. We defined a gene as all exons (both coding sequences and untranslated exons) and introns.

The unimodality or bimodality of normalized CpG content distributions was tested using the NOCOM software package (Ott 1992). Briefly, the software uses an expectation maximization algorithm to fit the data to both unimodal and bimodal distribution models, and finds the maximum likelihood values (L_0 and L_1 for unimodal and bimodal models, respectively). The statistic $G^2 = 2\ [\ln(L_1) – \ln(L_0)]$, which approximately follows a Chi-square distribution with two degrees of freedom, can be used to test if a bimodal distribution provides a better fit to the data than a unimodal distribution. The cutoff value between high CpG genes and low CpG genes was determined by plotting curves based on the NOCOM means of 0.55 (SD =0.20) and 1.50 (SD = 0.20) and determining their point of intersection (1.08; Figure 1D).

Gene ontology biological process term enrichment

Because gene ontology annotation (Ashburner et al. 2000) is limited in *A. mellifera*, annotations of orthologs in *D. melanogaster* were used for GO term analysis. To identify orthologous proteins between *A. mellifera* and *D. melanogaster*, Refseq RNA nucleotide accessions for *A. mellifera* sequences were converted to protein GI identifiers using the gene2refseq database from the NCBI FTP site (http://www.ncbi.nlm.nih.gov/Ftp/) and *D. melanogaster o*rthologs of *A. mellifera* genes were downloaded from the Roundup database of orthology (DeLuca et al. 2006), which uses the reciprocal smallest distance algorithm. A divergence threshold of 0.8 and BLAST E-value cutoff of 1e-10 were used for ortholog identification. A total of 4312 orthologous gene pairs between *A. mellifera* and *D. melanogaster* were obtained for further analysis.

GO biological process term enrichment was determined by comparing orthologs of low CpG and high CpG genes separately to a background composed of both low CpG and high CpG orthologs using the DAVID bioinformatics database functional annotation tool (Dennis et al. 2003). A Benjamini multiple testing correction of the EASE Score (a modified Fisher Exact *P*-value) was used to determine statistical significance of gene-enrichment (Hosack et al. 2003).

Differential gene expression between honeybee queen and worker castes

Differential gene expression in brains of *A. mellifera* adult queens and sterile workers was determined using cDNA microarray analyses by Grozinger et al. (2007). A list of BAGEL normalized expression levels (Townsend and Hartl 2002) and *P*-values for expression differences between queens and sterile workers was obtained from CM Grozinger. Gene identifiers for microarray data were converted to RNA nucleotide accessions using the gene_info and gene2refseq databases from the NCBI FTP site (http://www.ncbi.nlm.nih.gov/Ftp/).

Acknowledgements

This study is supported by the Alfred P. Sloan Research Fellowship to SY and a NSF grant (DEB 0640690) to MG and SY. We thank Tim Nowack for analyzing whole genome data and CM Grozinger for readily corresponding and providing data from prior research.

CHAPTER 4

FUNCTIONAL CONSERVATION OF DNA METHYLATION IN THE PEA APHID AND THE HONEYBEE[3]

Abstract

DNA methylation is a fundamental epigenetic mark known to have wide-ranging effects on gene regulation in a variety of animal taxa. Comparative genomic analyses can help elucidate the function of DNA methylation by identifying conserved features of methylated genes and other genomic regions. In this study, we used computational approaches to distinguish genes marked by heavy methylation from those marked by little or no methylation in the pea aphid, *Acyrthosiphon pisum.* We investigated if these two classes had distinct evolutionary histories and functional roles by conducting comparative analysis with the honeybee, *Apis (Ap.) mellifera*. We found that highly methylated orthologs in *A. pisum* and *Ap. mellifera* exhibited greater conservation of methylation status, suggesting that highly methylated genes in ancestral species may remain highly methylated over time. We also found that methylated genes tended to show different rates of evolution than unmethylated genes. In addition, genes targeted by methylation were enriched for particular biological processes that differed from those in relatively unmethylated genes. Finally, methylated genes were preferentially ubiquitously expressed among alternate phenotypes in both species, while genes lacking signatures of methylation were preferentially associated with diverse functional roles and condition-

[3] Hunt BG, Brisson JA, Yi SV, Goodisman MAD. 2010a. Functional conservation of DNA methylation in the pea aphid and the honeybee. Genome Biol Evol 2:719-728.

specific gene regulation. Overall, our analyses support a conserved role for DNA methylation in insects with comparable methylation systems.

Introduction

DNA methylation is an important epigenetic modification that plays a role in gene regulation in many organisms (Wolffe and Matzke 1999; Jaenisch and Bird 2003; Weber et al. 2007). Although DNA methylation occurs in all three domains of life, its genomic patterns show considerable variation among taxa (Hendrich and Tweedie 2003; Field et al. 2004; Suzuki and Bird 2008). For example, vertebrate genomes exhibit global patterns of methylation, but invertebrate genomes tend to display reduced or minimal levels of methylation (Suzuki and Bird 2008). Moreover, methylation of gene promoter regions in vertebrates leads to transcriptional repression (Wolffe and Matzke 1999; Jaenisch and Bird 2003; Weber et al. 2007; Zemach et al. 2010), but this relationship has not been observed in invertebrates. Instead, methylation primarily targets invertebrate gene bodies (Suzuki and Bird 2008; Xiang et al. 2010; Zemach et al. 2010). These contrasting patterns and effects have traditionally enforced the view that DNA methylation plays a fundamentally different role in in vertebrate and invertebrate genomes.

The arrival of genome sequences from multiple insects now makes a greater understanding of the patterns and phenotypic consequences of DNA methylation more tangible (Honeybee Genome Sequencing Consortium 2006; Wang et al. 2006; International Aphid Genomics Consortium 2010; Nasonia Genome Working Group 2010; Walsh et al. 2010). Specifically, comparative genomic analysis can be used to determine whether targets of DNA methylation are conserved between taxa. Moreover, the inferred patterns of methylation can be used to test current hypotheses explaining the evolutionary persistence of DNA methylation (Yi and Goodisman 2009). For example, it has been hypothesized that gene body methylation may act to minimize spurious transcription patterns (Suzuki et al. 2007; Maunakea et al. 2010), which could explain observations of dense methylation in functionally conserved genes and genes with ubiquitous expression among tissues in invertebrates (Suzuki et al. 2007; Foret et al. 2009; Xiang et al. 2010). It has also been suggested that DNA methylation persists in animals for genomic defense against transposable elements (Yoder et al. 1997, but see Regev et al. 1998, Simmen et al. 1999, Suzuki et al. 2007, and Xiang et al. 2010). DNA methylation may also act as an important mechanism for genomic imprinting, which results in the differential expression of parental alleles (Reik and Walter 2001). Finally, *de novo* DNA methylation is hypothesized to play an important role in developmental responsiveness to environmental factors and the regulation of phenotypic plasticity, as is apparently the case in the honeybee (Jaenisch and Bird 2003; Kucharski et al. 2008; Maleszka 2008).

The purpose of this study was to determine whether DNA methylation plays a conserved role in divergent insects with comparable DNA methylation systems. We provided insight into this question by comparing and contrasting the evolutionary signatures of DNA methylation in the genomes of the pea aphid, *Acyrthosiphon pisum*, and the honeybee, *Apis (Ap.) mellifera*.

A. pisum diverged from *Ap. mellifera* more than 300 million years ago (Gaunt and Miles 2002; Honeybee Genome Sequencing Consortium 2006), a time-frame roughly equivalent to the divergence of modern birds and mammals (Kumar and Hedges 1998). Developmentally, *Ap. mellifera* undergoes full metamorphosis and possesses morphologically distinct larval, pupal, and adult stages. In contrast, *A. pisum* develops gradually and does not undergo metamorphosis. However, *A. pisum* and *Ap. mellifera* both serve as important models for understanding the evolution and development of phenotypic plasticity (Evans and Wheeler 2001; Brisson and Stern 2006; Honeybee Genome Sequencing Consortium 2006; Brisson 2010; International Aphid Genomics Consortium 2010).

Specifically, aphids have a complex life cycle that alternates between asexual and sexual development. Asexual females exhibit a wing polyphenism in which they produce either winged or unwinged morphs depending on environmental cues (reviewed in Müller et al. 2001). During the sexual portion of the life cycle, males also produce winged or unwinged morphs. However, morph determination is genetic in males, and thus male wing dimorphism is referred to as a polymorphism (Smith and MacKay 1989). Honeybees, on the other hand, are highly social and dwell in large, predominantly female, colonies (Wilson 1971). Individuals partake in a remarkable division of labor, with a single queen typically dominating reproduction and workers engaged in tasks related to brood rearing, foraging, and colony defense (Wilson 1971). Queen and worker castes are developmentally determined by nutritional factors and exhibit dramatically different anatomy and behavior (Wheeler 1986; Evans and Wheeler 2001).

Importantly, both *Ap. mellifera* and *A. pisum* show evidence of widespread DNA methylation that is predominantly targeted to genes (Wang et al. 2006; Elango et al. 2009; Walsh et al. 2010). Consequently, patterns of genome methylation in *A. pisum* and *Ap. mellifera* can provide considerable insight into the function of gene methylation in insects, in particular, and invertebrates, in general.

In this study, we investigated the conservation of DNA methylation patterns in *A. pisum* and *Ap. mellifera* by first testing whether genes with similar functions are targeted by DNA methylation in both species. To achieve this aim, we examined patterns of functional enrichment among genes marked by relatively dense methylation and relatively sparse methylation. We further tested whether shared patterns of functional enrichment among DNA methylation targets are associated with conservation at the sequence level (Suzuki et al. 2007). Next, we examined whether *A. pisum* provided support for the hypothesis that genes with sparse methylation exhibit condition-specific gene expression (Elango et al. 2009; Foret et al. 2009). Finally, we synthesized our results with those from other recent investigations to advance a more comprehensive understanding of DNA methylation in insects. Overall, our results provide support for a remarkable level of conservation in gene methylation status and function over evolutionary time.

Materials and Methods

Gene Sequences

Analyses were conducted on mRNA transcript sequences because evidence suggests DNA methylation preferentially targets exons in insects and other invertebrates (Wang et al. 2006; Suzuki et al. 2007; Elango et al. 2009; Xiang et al. 2010; Zemach et al. 2010). For *A. pisum*, the 'ACYPmRNA' and the 'ACYPproteins' official genes consensus sets were obtained from AphidBase (http://www.aphidbase.com). For *Ap. mellifera*, the 'Amel_pre_release2' amino acid sequence official gene set was obtained from BeeBase (http://www.beebase.org), and model RefSeq transcripts were downloaded from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/Ftp). *Ap. mellifera* official gene set IDs were converted to RefSeq accessions using the 'gene_info' and 'gene2refseq' databases, also available from NCBI. For *Drosophila melanogaster*, 'Release_5.21' transcript and protein sequence sets were obtained from flybase (http://flybase.org).

Normalized CpG Dinucleotide Content (CpG$_{O/E}$)

We used CpG$_{O/E}$ as a measure of the level of DNA methylation of genes (Saxonov et al. 2006; Suzuki et al. 2007; Weber et al. 2007; Yi and Goodisman 2009). CpG$_{O/E}$ acts as a metric of levels of DNA methylation because methylation occurs predominantly on CpG dinucleotides in animals and methylated cytosines are hypermutable due to spontaneous deamination. This deamination causes a gradual depletion of CpG dinucleotides from methylated regions over time (Bird 1980). Consequently, genomic regions with relatively dense germline methylation have low CpG$_{O/E}$ and regions with little or no germline methylation maintain high levels of

$CpG_{O/E}$. It is important to note that $CpG_{O/E}$ could be influenced by either the number of methylated CpG sites or the proportion of cells incurring methylation at a given locus. In addition, somatic mutations are not transmitted to progeny and therefore cannot influence $CpG_{O/E}$ in and of themselves. However, $CpG_{O/E}$ *has* been linked to empirically-determined levels of DNA methylation in somatic tissues in insects, suggesting that many genes are universally methylated in germlines and soma (Foret et al. 2009; Xiang et al. 2010). Nevertheless, instances of *de novo* methylation in somatic tissues are not expected to directly influence $CpG_{O/E}$ measures.

$CpG_{O/E}$ was calculated as described previously (Elango et al. 2009), from the gene sets above. Only RefSeq model sequences were used for analyses involving $CpG_{O/E}$ in *A. pisum* (except in the case of gene expression analysis, described below) because RefSeq models were used for *Ap. mellifera* in our analysis. Sequences with $CpG_{O/E}$ values of 0 were removed from further analysis.

Bimodal distributions of $CpG_{O/E}$ have previously been reported in both *Ap. mellifera* (Elango et al. 2009; Foret et al. 2009; Wang and Leung 2009) and *A. pisum* (Walsh et al. 2010). In this study, we used the NOCOM software package (Ott 1979) to estimate means, standard deviations, and proportions of two components of the mixture of normal distributions of $CpG_{O/E}$ for both *A. pisum* and *Ap. mellifera*. These distributions were plotted using R (R Development Core Team 2010), and their intersections were used as cutoffs to divide low $CpG_{O/E}$ and high $CpG_{O/E}$ gene classes.

Orthology

Three-way orthologs between *A. pisum*, *Ap. mellifera*, and *D. melanogaster* were identified by first performing pairwise BLASTp comparisons of complete protein sequence sets with a cutoff of 1e-5, next identifying pairwise reciprocal best hits, and finally identifying orthologs with shared best hits among all pairwise comparisons (Altschul et al. 1997; Stajich et al. 2002). Orthologs determined in this manner were used for comparisons of $\mathrm{CpG}_{O/E}$ and evolutionary distance between orthologs from *A. pisum* and *Ap. mellifera*.

Pairwise orthologs shared between *A. pisum* and *D. melanogaster* were identified by performing BLASTp comparisons of complete protein sequence sets with a cutoff of 1e-5 and identifying reciprocal best hits. Only orthologs with RefSeq model proteins in *A. pisum* were retained.

Sequence Divergence

In order to compare the evolutionary divergence of low $\mathrm{CpG}_{O/E}$ and high $\mathrm{CpG}_{O/E}$ orthologs between *A. pisum* and *Ap. mellifera*, a total of 2,222 orthologous protein sequences were first aligned using ClustalW (Thompson et al. 1994). Confidently aligned gap-free columns were then extracted using Gblocks with default settings (Castresana 2000), and only long alignments (≥ 100 amino acids) were kept for analysis. PAL2NAL was used to convert protein sequence alignments to corresponding codon alignments (Suyama et al. 2006). Finally, PAML was used to calculate rates of synonymous (dS) and nonsynonymous (dN) substitution with the 'codeml' method (Yang 2007). Because synonymous substitution rates were predominantly saturated (dS > 2), measures of dN and DNA sequence percent identity were used to assess sequence divergence.

Gene Ontology

Gene ontology (GO) annotations for *D. melanogaster* orthologs of *A. pisum* proteins were used to analyze enrichment of biological process terms (Ashburner et al. 2000). GO biological process term enrichment was determined by comparing orthologs of low $\mathrm{CpG}_{O/E}$ and high $\mathrm{CpG}_{O/E}$ genes separately to a background composed of both low $\mathrm{CpG}_{O/E}$ and high $\mathrm{CpG}_{O/E}$ orthologs using the DAVID bioinformatics database functional annotation tool (Dennis et al. 2003). A Benjamini multiple-testing correction of the EASE score (a modified Fisher exact P value; Hosack et al. 2003) was used to determine statistical significance of GO term enrichment.

EST Mapping

A. pisum ESTs, previously used to characterize differential gene expression underlying developmental differences, sex differences, female wing polyphenism, and wing morph differences (Brisson et al. 2007), were mapped to the *A. pisum* official genes consensus set (OGS) to aid in assessing the relationship between the degree of differential gene expression among phenotypic classes and $\mathrm{CpG}_{O/E}$. EST sequences were compared to all OGS mRNA sequences by BLASTn (Altschul et al. 1997). To be considered a match, EST query sequences were required to have > 50% sequence alignment to an OGS hit, > 95% identity of the aligned sequence, and reciprocal best hits resulting from BLASTn analysis of the OGS query against an EST database. GLEAN as well as RefSeq gene models were accepted in this case to map a greater proportion of microarray data.

Gene Expression

Brisson et al. (2007) previously examined the gene expression differences underlying distinct phenotypes in *A. pisum* using cDNA microarrays (Wilson et al. 2006). Specifically, microarrays were utilized to determine the degree of differential gene expression in comparisons of (i) fourth instar juveniles versus adults (compared within unwinged males, within winged males, within unwinged asexual females, and within winged asexual females), (ii) males versus asexual females (compared within winged fourth instars, within unwinged fourth instars, within winged adults, and within unwinged adults), (iii) polyphenic winged versus unwinged females (compared within fourth instars and within adults), and finally, polymorphic winged versus unwinged males (compared within fourth instars and within adults).

For the present study, we calculated the mean of the absolute value of $\log_2$-transformed ratios across multiple comparisons to measure the degree of differential gene expression. In this manner, we combined data from all pairwise comparisons of (i) development, (ii) sex, (iii) female wing polyphenism, and (iv) male wing polymorphism.

The mean of $\log_2$-transformed gene expression ratios across all 12 pairwise comparisons was also calculated. We further divided each of these measures into two bins at a mean $|\log_2$ expression ratio$|$ value of 0.5, with genes below this threshold roughly corresponding to genes with similar expression between groups and genes above this value roughly corresponding to genes with differential expression between groups.

We also re-visited analysis previously described and published by Elango et al. (2009) which demonstrated that high $\mathrm{CpG}_{O/E}$ genes were overrepresented among genes that were differentially expressed between queen and worker castes (Grozinger et al. 2007). For the present manuscript we analyzed NCBI transcript sequences, rather than introns and exons combined, to remain consistent with our analyses of aphid gene expression.

Finally, Foret et al. (2009) previously used an oligonucleotide microarray representing the honeybee official gene set (Honeybee Genome Sequencing Consortium 2006) to assess the expression breadth of genes among the following tissues in *Ap. mellifera*: antenna, brain, whole-body larva, hypopharyngeal gland, ovary, and thorax. They further demonstrated that low $\mathrm{CpG}_{O/E}$ genes were vastly overrepresented among genes with ubiquitous expression (Foret et al. 2009). We expanded upon their analysis by splitting genes into six classes based upon the number of tissues with observed expression. To do so, we utilized lists of genes expressed in each tissue, along with a fasta file of sequences used to design the array. To map sequences with generic microarray identifiers to honeybee model RefSeq transcripts, we compared the sequences using BLASTn (Altschul et al. 1997). To be considered a match, array query sequences were required to have > 50% sequence alignment to a model RefSeq transcript hit and > 98% identity for the aligned sequence. We then generated a numeric count of the number of tissues in which each gene was expressed (integers from 1-6) and recorded the $\mathrm{CpG}_{O/E}$ for each associated model RefSeq transcript. Data for expression breadth and $\mathrm{CpG}_{O/E}$ was obtained in this manner for a total of 7,576 *Ap. mellifera* genes.

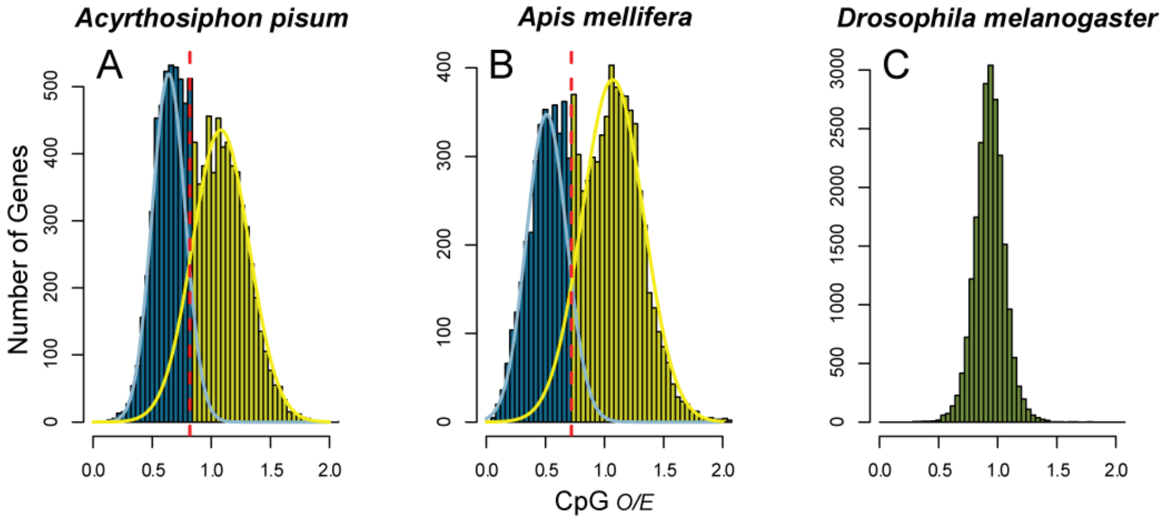

Figure 4.1. Distributions of normalized CpG dinucleotide content (CpG$_{O/E}$). (A) *A. pisum* and (B) *Ap. mellifera* exhibit bimodal distributions of CpG$_{O/E}$ among genes, signifying variation in germline DNA methylation levels. Dashed red lines indicate cutoffs used to divide low CpG$_{O/E}$ genes (blue) from high CpG$_{O/E}$ genes (yellow). In contrast to *A. pisum* and *Ap. mellifera*, (C) *D. melanogaster* has a unimodal distribution of CpG$_{O/E}$ and does not exhibit substantial levels of CpG methylation.

Additional Analysis

Statistical tests (rank sum tests and correlations) were performed using either R (R Development Core Team 2010) or the JMP statistical software package (SAS Institute Inc, Cary, NC). Proportional Venn diagrams were generated using the Venn Diagram Plotter available from Pacific Northwest National Laboratory (http://omics.pnl.gov).

Results

We divided genes into low and high CpG$_{O/E}$ classes based on the bimodal distributions of CpG$_{O/E}$ observed in *A. pisum* (CpG$_{O/E}$ cutoff = 0.82; Fig. 4.1A) and *Ap. mellifera* (CpG$_{O/E}$ cutoff = 0.72; Fig. 4.1B). These two classes of genes roughly correspond to genes incurring relatively dense versus relatively sparse methylation (Saxonov et al. 2006; Suzuki et al. 2007; Weber et al. 2007; Elango et al. 2009; Foret et al. 2009; Wang and Leung 2009; Yi and Goodisman 2009; Xiang et al. 2010).

To gain insight into the evolutionary maintenance of genes with different levels of methylation, we first investigated whether genes belonging to distinct CpG$_{O/E}$ classes showed differences in their conservation of CpG$_{O/E}$ status over evolutionary time. A total of 2,339 three-way orthologs were identified with non-zero CpG$_{O/E}$ values in *A. pisum*, *Ap. mellifera*, and *D. melanogaster*. By comparing the CpG$_{O/E}$ classification of orthologs in *A. pisum* and *Ap. mellifera* from this data, we found that genes with high CpG$_{O/E}$ exhibited considerably less conservation of CpG$_{O/E}$ status than genes with low CpG$_{O/E}$ (Fig. 4.2, Table 4.1; Pearson's Chi-squared test with Yates' continuity correction *P* = 0.0075). Thus, patterns of dense DNA methylation have been more conserved over evolutionary time than patterns of sparse DNA methylation in *A. pisum* and *Ap. mellifera*.

We next determined whether the differential conservation of low CpG$_{O/E}$ and high CpG$_{O/E}$ status was associated with differential conservation of nucleotide and amino acid sequence. We found that genes from the low CpG$_{O/E}$ class in *A. pisum* and *Ap. mellifera* both harbored significantly greater proportions of genes with detectable three-way

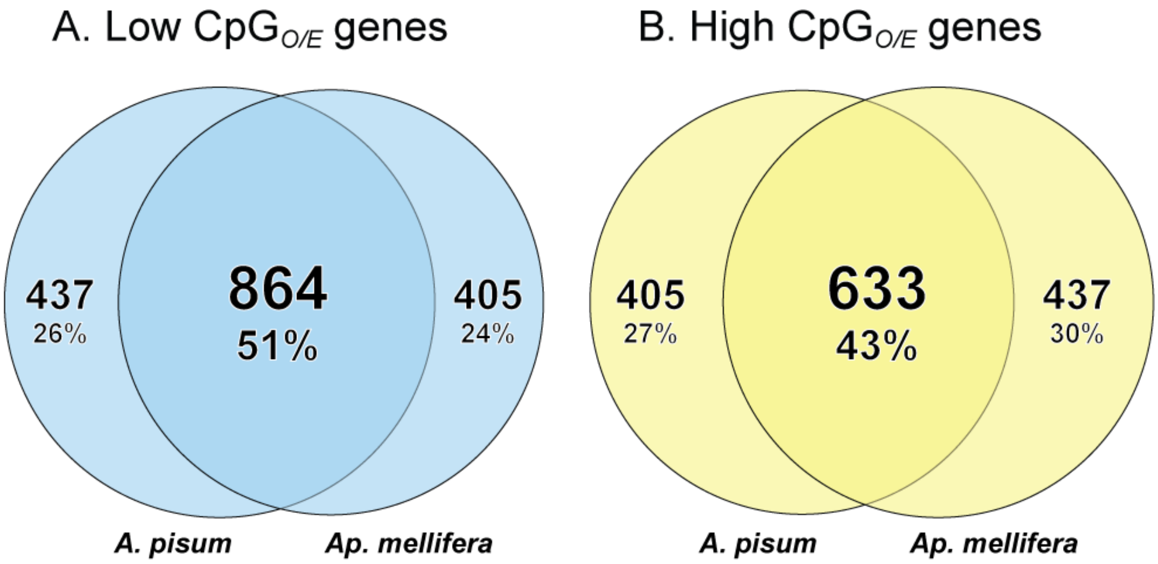

Figure 4.2. Pan-genomic high CpG$_{O/E}$ status is less conserved than low CpG$_{O/E}$ status. Analysis of orthologs in *A. pisum* and *Ap. mellifera* show that a higher proportion of (A) low CpG$_{O/E}$ genes are conserved with respect to normalized CpG content than (B) high CpG$_{O/E}$ genes. Each circle represents the number of genes from one species belonging to the designated CpG$_{O/E}$ class; overlap designates the number of orthologs with agreement in CpG$_{O/E}$ classification in both species.

Table 4.1. Contingency table of $\mathrm{CpG}_{O/E}$ conservation between *A. pisum* and *Ap. mellifera*.

	Conserved $\mathrm{CpG}_{O/E}$ status with *Ap. mellifera*	**Non-conserved $\mathrm{CpG}_{O/E}$ status with *Ap. mellifera***	**Proportion conserved**
***A. pisum* low $\mathrm{CpG}_{O/E}$ genes**	864	437	66.4 %
***A. pisum* high $\mathrm{CpG}_{O/E}$ genes**	633	405	61.0 %

Conservation differs significantly between low $\mathrm{CpG}_{O/E}$ genes and high $\mathrm{CpG}_{O/E}$ genes (Pearson's Chi-squared test with Yates' continuity correction $P = 0.0075$).

orthologs than genes from the high CpG$_{O/E}$ class (Table 4.2; Pearson's Chi-squared test with Yates' continuity correction *P* < 1E-15). We also found that DNA sequence conservation was significantly higher between *A. pisum* and *Ap. mellifera* orthologs from the low CpG$_{O/E}$ class than orthologs from the high CpG$_{O/E}$ class (Kruskal-Wallis rank sum test *P* = 0.0003; Fig. 4.3A). Both of these results suggested that densely methylated genes, as a whole, were considerably more conserved at the sequence level than sparsely methylated genes. However, in contrast to the results obtained from analysis of ortholog loss and DNA sequence identity, amino acid substitution rates among genes with detectable three-way orthologs were slightly higher among low CpG$_{O/E}$ genes than high CpG$_{O/E}$ genes (Kruskal-Wallis rank sum test *P* = 0.0012; Fig. 4.3B).

To investigate whether genes with different levels of methylation were associated with specific functions, we next tested for enrichment of gene ontology biological process terms in 4,404 *A. pisum* genes with *D. melanogaster* orthologs. We found that functions related to cellular metabolic processes were overrepresented among low CpG$_{O/E}$ genes (Table 4.3). In contrast, functions associated with cellular signaling, behavior, and environmental stimulus were overrepresented among high CpG$_{O/E}$ genes (Table 4.3).

We also found that six of the top ten enriched functional terms for *A. pisum* low CpG$_{O/E}$ genes were among the top ten enriched functional terms in *Ap. mellifera* low CpG$_{O/E}$ genes (Table 4.3; Elango et al. 2009). In contrast, only two of the top ten high CpG$_{O/E}$ functional enrichment terms were in agreement between *A. pisum* and *Ap. mellifera* (Table 4.3; Elango et al. 2009). Thus the function of low CpG$_{O/E}$ genes appears to be relatively conserved over evolutionary history.

Finally, we investigated whether CpG$_{O/E}$ measures were associated with patterns of gene expression among distinct phenotypic groups in *A. pisum* using microarray data for 1,347 genes (Brisson et al. 2007). We analyzed the degree of differential gene expression between developmental stages (development; 4th instar versus adult), between

Table 4.2. Ortholog detection among low CpG$_{O/E}$ and high CpG$_{O/E}$ genes.

	Acyrthosiphon pisum			***Apis mellifera***		
	Three-way orthology	**No three-way orthology**	**Proportion with three-way orthology**	**Three-way orthology**	**No three-way orthology**	**Proportion with three-way orthology**
Low CpG$_{O/E}$	1301	3309	28.2 %	1269	2331	35.3 %
High CpG$_{O/E}$	1038	4818	17.7 %	1070	4790	18.3 %

Ortholog detection differs significantly between low CpG$_{O/E}$ genes and high CpG$_{O/E}$ genes (Pearson's Chi-squared test with Yates' continuity correction P < 1E-15 for both *A. pisum* and *Ap. mellferia*, each analyzed separately).

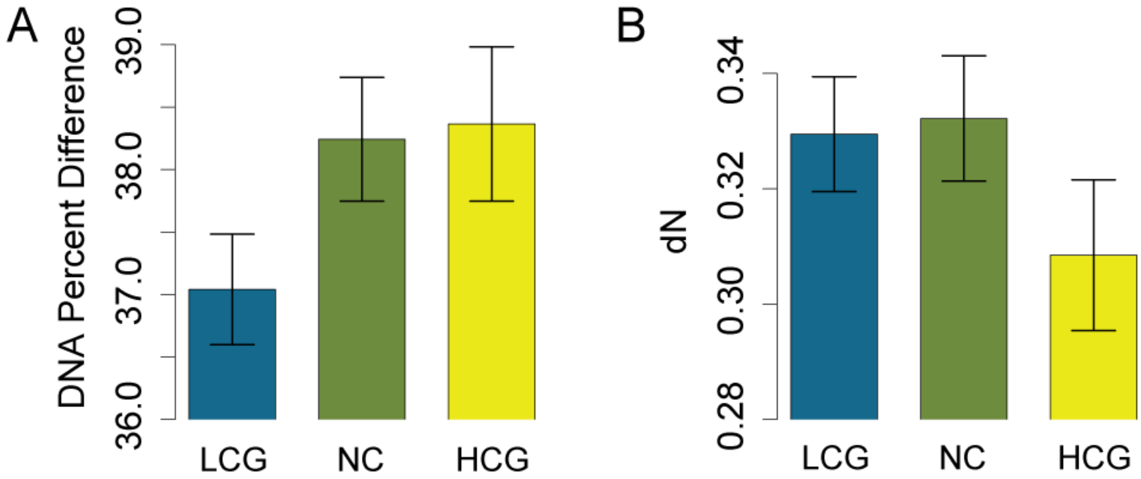

Figure 4.3. High CpG$_{O/E}$ genes exhibit significantly greater nucleotide divergence, but lower amino acid divergence, when compared to low CpG$_{O/E}$ genes with three-way orthology. (A) DNA percent difference is significantly higher between *A. pisum* and *Ap. mellifera* for conserved high CpG$_{O/E}$ orthologs (HCG) and orthologs with non-conserved CpG$_{O/E}$ status (NC) than those with conserved low CpG$_{O/E}$ status (LCG; Kruskal-Wallis rank sum test *P* = 0.0003). (B) In contrast, the nonsynonymous substitution rate (dN) is lower for conserved high CpG$_{O/E}$ orthologs compared to orthologs with non-conserved CpG$_{O/E}$ status or low CpG$_{O/E}$ status (Kruskal-Wallis rank sum test *P* = 0.0012). Means with 95% confidence intervals are plotted.

sexes (sex; male versus asexual female), between environmentally-sensitive asexual female wing phenotypes (female wing polyphenism; winged versus unwinged), and between genetically-determined male wing phenotypes (male wing polymorphism; winged versus unwinged).

Our results suggested that genes with low levels of DNA methylation exhibited complex, condition-specific regulation of gene expression: differential gene expression, when combined for all pairwise comparisons of alternate phenotypes, displayed a significant positive correlation with CpG$_{O/E}$ in *A. pisum* (Pearson Product-Moment correlation $P < 0.001$; Table 4.4, Fig. 4.4A). This signal was primarily driven by development, sex, and female wing polyphenism, which each demonstrated that differential gene expression was significantly associated with high CpG$_{O/E}$ (Table 4.4; Fig. 4.4A). Differential gene expression between male wing morphs was not significantly associated with CpG$_{O/E}$ in *A. pisum*, although the trend was in the same direction as the other tests (Table 4.4, Fig. 4.4A).

We also reanalyzed data linking gene expression to methylation levels in *Ap. mellifera* to illustrate that differential gene expression between caste phenotypes (Elango et al. 2009) and gene expression breadth (Foret et al. 2009) were also each associated with CpG$_{O/E}$ (Fig. 4.4B, Fig. 4.4C). Specifically, genes with differential expression between *Ap. mellifera* queens and workers, and those expressed in few *Ap. mellifera* tissues, preferentially exhibited high CpG$_{O/E}$. Overall, our results reveal that genes with condition-specific regulation are associated with higher CpG$_{O/E}$, and lower levels of DNA methylation, than ubiquitously expressed genes in both *A. pisum* and *Ap. mellifera*.

Discussion

Gene Evolution and DNA Methylation

We have reported distinct levels of conservation of DNA methylation status for orthologs with heavy methylation (low CpG$_{O/E}$) and sparse methylation (high CpG$_{O/E}$) in the pea

Table 4.3. Top 10 enriched gene ontology biological process terms by CpG$_{O/E}$ class for *A. pisum*.

CpG$_{O/E}$ Class	Accession	GO biological process term	Fold enrichment in class	Top 10 in *Ap. mellifera*[a]	Significance[b]
Low	GO:0044260	cellular macromolecule metabolic process	1.15	no	1.72E-10
	GO:0044237	cellular metabolic process	1.11	yes	1.53E-09
	GO:0016070	RNA metabolic process	1.32	yes	5.81E-09
	GO:0008152	metabolic process	1.09	yes	1.66E-08
	GO:0043170	macromolecule metabolic process	1.12	yes	3.65E-08
	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.20	yes	4.72E-08
	GO:0009987	cellular process	1.06	yes	3.62E-07
	GO:0009057	macromolecule catabolic process	1.45	no	3.83E-07
	GO:0044265	cellular macromolecule catabolic process	1.46	no	4.63E-07
	GO:0030163	protein catabolic process	1.47	no	4.58E-06
High	GO:0007186	G-protein coupled receptor protein signaling pathway	1.72	no	2.48E-05
	GO:0007165	signal transduction	1.28	yes	0.0035
	GO:0007610	Behavior	1.40	no	0.0074
	GO:0003008	system process	1.30	no	0.0179
	GO:0050890	Cognition	1.43	no	0.0267
	GO:0050877	neurological system process	1.29	no	0.0279
	GO:0032501	multicellular organismal process	1.12	yes	0.0280
	GO:0009581	detection of external stimulus	1.77	no	0.0492
	GO:0009582	detection of abiotic stimulus	1.77	no	0.0492
	GO:0006811	ion transport	1.39	no	0.0565

[a] According to Elango et al. (2009)

[b] Benjamini multiple-testing correction of the EASE score (a modified Fisher exact *P*-value)

aphid, *A. pisum*, and the honeybee, *Ap. mellifera* (Fig. 4.2, Table 4.1). In particular, a greater proportion of orthologs maintain low CpG$_{O/E}$ status than high CpG$_{O/E}$ status over evolutionary time. Thus, genes that were presumably densely methylated in the ancestor of *A. pisum* and *Ap. mellifera* were more likely to remain methylated through evolutionary time, whereas genes with sparse methylation were less likely to maintain their low methylation status.

Furthermore, we found that heavily methylated genes had a greater number of detectable orthologs and exhibited greater DNA sequence conservation than genes with sparse methylation (Table 4.2; Fig. 4.3A). In line with these results, a prior study also found that genes with signatures of methylation were enriched among orthologs that could be identified between distantly related taxa (Suzuki et al. 2007). Thus heavily methylated genes, overall, appear to be more conserved at the sequence level than sparsely methylated genes. This observation is particularly striking because DNA methylation increases the occurrence of mutations at CpG sites and might be expected to lead to rapid DNA sequence divergence (Elango et al. 2008). One possible explanation for the observed trend, however, is that orthologs with consistently low CpG$_{O/E}$ over evolutionary history have fewer total CpG dinucleotides than methylated genes with intermediate CpG$_{O/E}$, and thus do not incur new mutations at a comparable rate (Suzuki et al. 2009). Another possibility is that genes targeted by DNA methylation may be under greater functional constraint, as a class, than unmethylated genes.

Surprisingly, in contrast to our results from analysis of DNA sequence identity, we found that densely methylated genes with detectable orthologs may be under less constraint at the amino acid level than their sparsely methylated counterparts (Figs. 4.3B). Apparently, *A. pisum* and *Ap. mellifera* high and low CpG$_{O/E}$ genes that *do not* retain detectable orthologs in *D. melanogaster* differ more from each other, in terms of evolutionary constraint at the protein level, than do high and low CpG$_{O/E}$ genes *with* detectable orthologs (Table 4.2 and Fig. 4.3). It remains unclear why this may be the case, but our results suggest that different classes of genes may behave differently with respect to the interaction between selective constraints or mutability and methylation status.

Gene Expression and DNA Methylation

In the present study, we add to the emerging view that genes with ubiquitous expression in insects are preferentially targeted by DNA methylation (Elango et al. 2009; Foret et al. 2009; Xiang et al. 2010). Specifically, genes with similar expression levels among phenotypic groups exhibit evolutionary signatures of significantly higher levels of DNA methylation than genes with differential expression between phenotypes in both *A. pisum* and *Ap. mellifera* (Fig. 4.4A, Fig. 4.4B; Elango et al. 2009). Genes with ubiquitous expression among tissues are also preferentially targeted by DNA methylation in both *Ap. mellifera* (Fig. 4.4C; Foret et al. 2009) and the silkworm, *Bombyx mori*, even though *B. mori* possesses only a partial complement of DNA methylation enzymes (Xiang et al. 2010). By comparison, genes with tissue-specific expression in *Ap. mellifera* (Fig. 4.4C; Foret et al. 2009) and *B. mori* (Xiang et al. 2010), with caste-specific expression in *Ap. mellifera* (Fig. 4.4B; Elango et al. 2009), and with differential expression between developmental stages, sexes, and polyphenic wing morphs in *A. pisum*, all exhibit lower levels of DNA methylation than their ubiquitously expressed counterparts (Fig. 4.4A). Thus, sparse levels of DNA methylation are associated with flexibility in gene expression, either between polyphenic forms or different tissues.
Our results reveal that complex gene regulation is associated with low levels of DNA methylation in disparate insects. This finding may appear to contrast with the idea that DNA methylation plays an important role in the epigenetic regulation of phenotypic plasticity (Jaenisch and Bird 2003; Kucharski et al. 2008; Maleszka 2008). Indeed, our observations suggest that the primary targets of DNA methylation are those genes least likely to be implicated as leading to phenotypic variation. However, we cannot rule out the cooption of DNA methylation for complex regulatory roles operating on a smaller

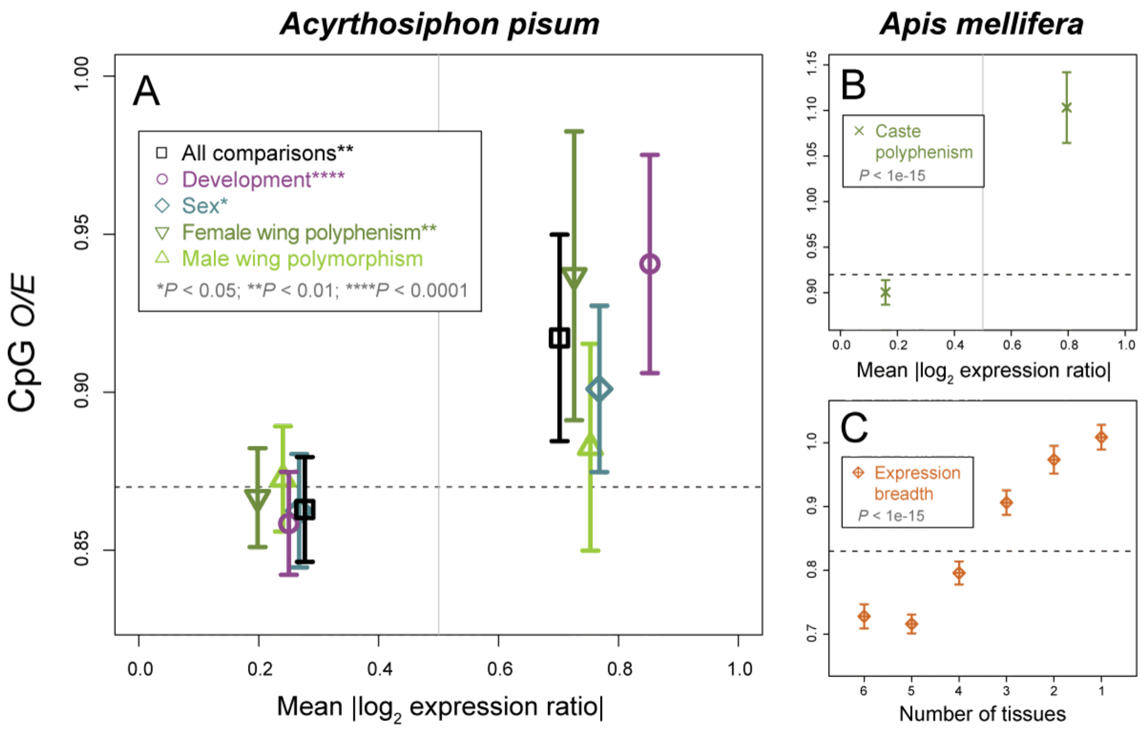

Figure 4.4. Ubiquitously expressed genes exhibit higher levels of DNA methylation than genes with condition-specific expression. (A) Genes with a high degree of differential expression between groups exhibit significantly higher CpG$_{O/E}$ than genes with ubiquitous expression in *A. pisum*. This relationship also holds true for (B) differential expression between *Ap. mellifera* queen and worker castes (adapted from Elango et al. 2009). (C) Similarly, genes with a high degree of tissue-specificity exhibit significantly higher CpG$_{O/E}$ than genes with ubiquitous expression among tissues in *Ap. mellifera* (adapted from Foret et al. 2009). Significance values represent Wilcoxon signed-rank tests in panels A and B, and a Kruskal-Wallis rank sum test in panel C. Means and 95% confidence intervals are plotted. Horizontal dashed lines represent the mean CpG$_{O/E}$ for all genes in a given panel. Vertical grey lines represent bin cutoffs for classification of genes according to mean |log$_2$ expression ratio|.

Table 4.4. Correlations between *A. pisum* differential gene expression and CpG$_{O/E}$.

	Pearson Product-Moment correlation with CpG$_{O/E}$
Mean |log$_2$ expression ratio| for all comparisons	0.0996***
Mean |log$_2$ expression ratio| for developmental stages	0.1091****
Mean |log$_2$ expression ratio| for female wing polyphenism	0.0905***
Mean |log$_2$ expression ratio| for sexes	0.0660*
Mean |log$_2$ expression ratio| for male wing polymorphism	0.0144

* $P < 0.05$; *** $P < 0.001$; **** $P < 0.0001$

number of loci.

Steps toward a Unified View of Intragenic Methylation

Recently, a unified view of the functional role of intragenic (versus intergenic or promoter) DNA methylation in vertebrates and invertebrates has begun to emerge. For example, it has recently been found that methylation of gene bodies in many vertebrates and invertebrates is associated with moderate gene expression levels (Zemach et al. 2010). Our data, obtained from microarray analyses, do not directly address overall levels of gene expression, but instead address expression breadth among tissues or alternate phenotypic classes. We find that genes with high CpG$_{O/E}$ measures possess an enriched aptitude for conditional expression associated with distinct tissues or alternate phenotypes. In contrast, genes with dense methylation exhibit a greater propensity for static levels of expression.

A recent mammalian study revealed that intragenic methylation limits the generation of alternate gene transcripts by masking intragenic promoters (Maunakea et al. 2010). This mechanism may explain why broadly expressed genes are subject to the highest levels of methylation in invertebrates: broadly expressed genes may be preferentially targeted by DNA methylation due to the enhanced negative effects associated with alternate promoters at such loci. Importantly, the proposed link between intragenic methylation and the regulation of alternate transcription (Maunakea et al. 2010) suggests that different levels of methylation in distinct tissues or developmental stages could have important phenotypic consequences.

Finally, we note that our results do not apply to insect taxa which have heavily diminished methylation systems (Urieli-Shoval et al. 1982; Field et al. 2004). Instead, we suggest that DNA methylation is one of many tools that can be co-opted for the purposes of gene regulation in organisms that have retained a complete enzymatic toolkit for mediating DNA methylation.

Acknowledgments

This work was supported by the U. S. National Science Foundation (grant numbers MCB-0950896 to SVY and MADG, DEB-0640690 to MADG and SVY, and DEB-1011349 to BGH, SVY, and MADG) and by the National Institute of Environmental Health Sciences (grant number 4R00ES017367 to JAB). We thank Sylvain Foret and Christina Grozinger for sharing data from previous investigations and Jungsun Park from the Yi lab for sharing code used to analyze CpG depletion. We also thank three anonymous reviewers for comments which helped to improve this manuscript.

CHAPTER 5

SOCIALITY IS LINKED TO RATES OF PROTEIN EVOLUTION IN A HIGHLY SOCIAL INSECT[4]

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 non-biased 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.

Introduction

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

[4] Hunt BG, Wyder S, Elango N, Werren JH, Zdobnov EM, Yi SV, Goodisman MAD. 2010b. Sociality is linked to rates of protein evolution in a highly social insect. Mol Biol Evol 27:497-500.

castes. Specifically, queens and males reproduce while 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 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.

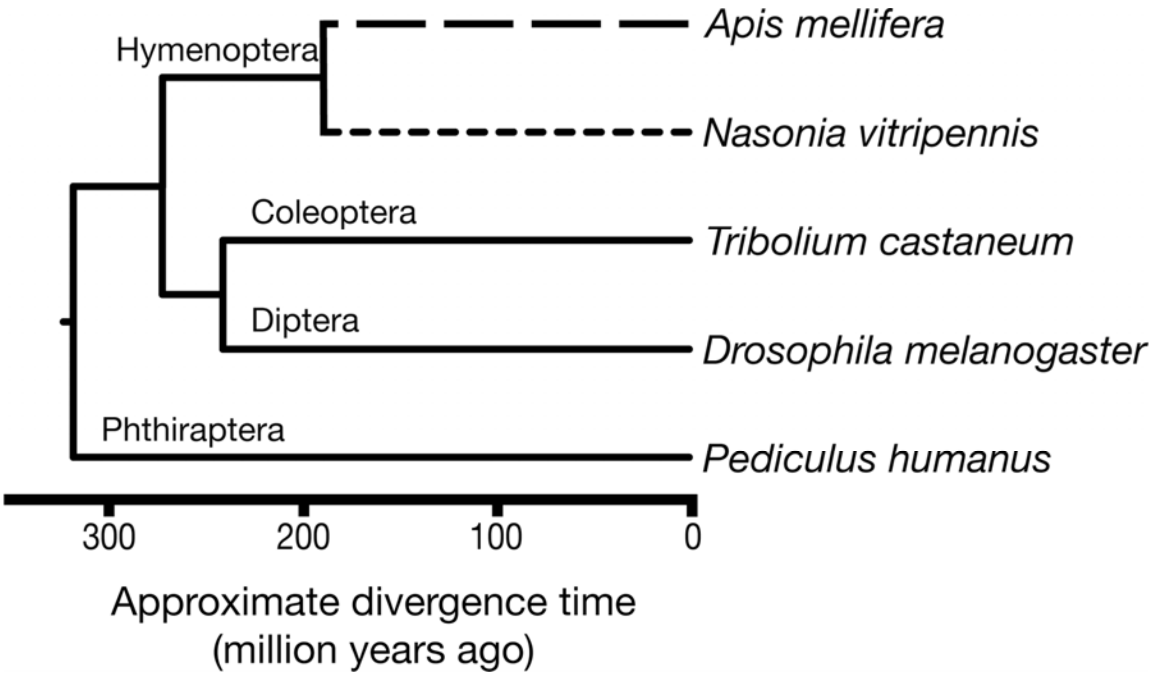

Figure 5.1. Species tree with approximate divergence times. Dotted lines depict terminal branch lengths for *A. mellifera* and *N. vitripennis* used in this study. Insect orders are labeled on internal branches.

Results and Discussion

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 non-eusocial parasitoid wasp *Nasonia vitripennis* and 1-3 additional insect species (Fig. 5.1). 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 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. 5.2).

Proteins with queen-biased expression in *A. mellifera* exhibited significantly higher evolutionary rates than proteins with nonsignificant bias or worker bias (Fig. 5.2B). In contrast, worker-biased proteins did not evolve at elevated rates (Fig. 5.2B). In fact, worker-biased proteins had the lowest rates of amino acid substitution of the three gene expression classes, suggesting there is not an overall positive relationship between caste-specificity and evolutionary rate.

Protein evolutionary rate was significantly correlated with the ratio of queen to worker gene expression, but was also strongly associated with several other factors (Table 5.1; Pal et al. 2006). For example, there was a strong positive correlation between evolutionary rates in *A. mellifera* and non-eusocial *N. vitripennis*. This suggests that protein evolution is heavily influenced by selective pressures shared with a non-eusocial common ancestor (Table 5.1, Table 5.2) 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

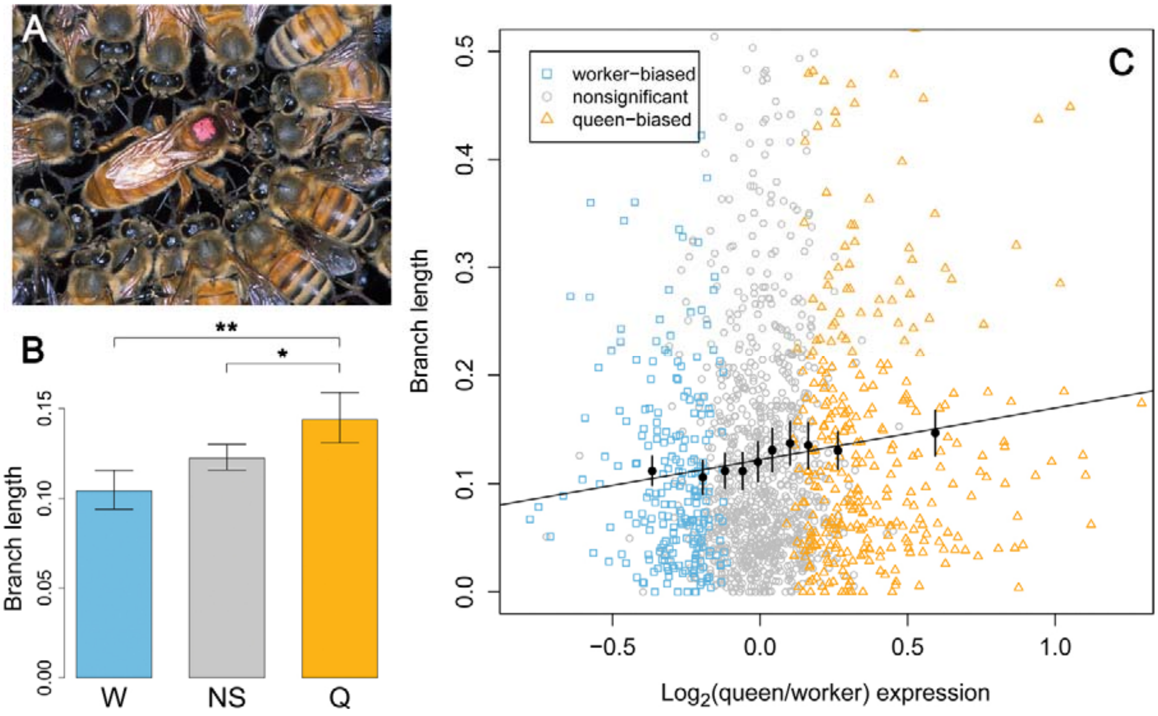

Figure 5.2. Caste-biased gene expression is linked to protein evolutionary rate in *Apis mellifera*. **(A)** *A. mellifera* workers surround a queen. **(B)** *A. mellifera* evolutionary rates (branch lengths in amino acid substitutions per site) differ significantly among genes with worker-biased expression (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; ** *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, USDA/ARS.

Table 5.1. Spearman's rank correlation coefficients (r_s) and partial correlations between *Apis mellifera* evolutionary rates[a] and selected gene attributes.

Variable (*X*)	Correlation $r_{s\ X,\ A.\ mellifera\ \text{branch length}}$	Partial correlation $r_{s\ X,\ A.\ mellifera\ \text{branch length}\mid\text{all other variables}}$
***N. vitripennis* branch length[a]**	0.706*****	0.702*****
Synonymous 3rd 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

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

[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 *Nasonia 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 5.2. 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[c]					
Log$_2$(queen/worker) gene expression	0.2	**26.8**	**19.6**	**51.5**	
N. vitripennis branch length	1.0	1.0	**66.1**	**31.8**	
Effective number of codons	**44.2**	3.0	0.3	3.3	
Synonymous 3rd 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	

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

[a] Principal components are numbered in 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. *N. 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).

non-biased proteins (Table 5.3). 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; Kim and Yi 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 associated with evolutionary rates in other taxa (Table 5.1). 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 5.2; 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 (Table 5.4). 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

Table 5.3. 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] ± SEM			**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

[a] See Wolf, Carmel, and Koonin (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.

* P-value < 0.05

Table 5.4. 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)

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

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. 5.1). 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. 5.1) were taken from the literature (Grimaldi and Engel 2005; Nasonia Genome Working Group 2010). 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.

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 et al. (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 et al. 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 et al. 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 2010). 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 (Ashburner et al. 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 5.4).

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.

Acknowledgements

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 U. S. 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.).

CHAPTER 6

RELAXED SELECTION IS A PRECURSOR TO THE EVOLUTION OF PHENOTYPIC PLASTICITY[5]

Abstract

Phenotypic plasticity represents one of the most important ways that organisms adaptively respond to environmental variation. Alternate phenotypes produced through phenotypic plasticity generally arise through conditional gene expression (Gibson 2008; Smith et al. 2008; Ayroles et al. 2009), which is predicted to result in relaxed selective constraint in conditionally expressed genes (Barker et al. 2005; Linksvayer and Wade 2009; Snell-Rood et al. 2010; Van Dyken and Wade 2010). However, ancestral relaxation of selection may also act as a predecessor to the evolution of conditional gene expression (Mank and Ellegren 2009). Thus, whether relaxed selection acts primarily as a cause or consequence of conditional gene expression remains unclear. Here we show that genes with diminished selective constraints have a higher propensity for evolving conditional expression associated with phenotypic plasticity. By analyzing gene expression patterns associated with specialized castes, sexes, and developmental stages in the fire ant *Solenopsis invicta* (Ometto et al. in press), we show that the rate of molecular evolution is higher for conditionally expressed genes. Surprisingly, we also find that orthologs of genes with caste-biased expression in either *S. invicta* or the honeybee *Apis mellifera* exhibit heightened lineage-specific rates of amino acid substitution in taxa

[5] Hunt BG, Ometto L, Wurm Y, Shoemaker D, Keller L, Yi SV, Goodisman MAD. in preparation. Relaxed selection is a precursor to the evolution of phenotypic plasticity.

lacking castes. Our results indicate that relaxed selection precedes differential gene expression and plays an underappreciated role in the evolution of phenotypic plasticity.

Introduction

The efficiency of natural selection is influenced fundamentally by molecular, developmental, and population-level constraints on protein evolution (Pal et al. 2006; Drummond and Wilke 2008; Koonin and Wolf 2010; Snell-Rood et al. 2010). For example, a reduction in selective constraint may be associated with functional specialization, as demonstrated by increased rates of evolutionary change in protein-coding genes with tissue-specific (Duret and Mouchiroud 2000) and sex-biased (Ellegren and Parsch 2007) expression. This phenomenon has been explained as a product of conditional gene expression, which is theorized to resolve intralocus conflict through a reduction in pleiotropy (Chippindale et al. 2001; Bonduriansky and Chenoweth 2009) and to weaken the efficiency of natural selection (Barker et al. 2005; Brisson and Nuzhdin 2008; Van Dyken and Wade 2010). However, relaxed selection on proteins may itself facilitate the evolution of conditional gene expression if the costs of changes in gene expression are low for such loci (Castillo-Davis et al. 2004; Mank and Ellegren 2009).

Hymenopteran eusocial insects (ants, social bees, and social wasps) represent ideal subjects for investigating the evolutionary implications of conditional gene expression arising from phenotypic plasticity (Hunt et al. 2010b) because specialized female queen and worker castes are often genetically indistinguishable and exhibit extreme polyphenisms in many taxa (Hölldobler and Wilson 1990). Hymenopteran larval and adult stages also differ dramatically in phenotype. In this study, we first determined whether relaxed selective constraint is associated with differential gene expression between *S. invicta* sexes, developmental stages, and castes. We then tested if relaxed selective constraint on proteins followed from differential gene expression or if differential gene expression among polyphenic forms preceded changes in protein evolution.

Results and Discussion

Analyses of adult and pupal males, queens, and workers (Ometto et al. in press) (Fig. 6.1A) using *S. invicta* cDNA microarrays (Wang et al. 2007) revealed that many genes are differentially expressed between sexes, female castes, and developmental stages (Ometto et al. in press) (Fig. 6.1B). Analyses of relative molecular evolutionary rates, based primarily on protein coding sequences from the ants *S. invicta* (Wurm et al. in press), *Pogonomyrmex barbatus* (Smith et al. in press-b), and *Linepithema humile* (Smith et al. in press-a), provided insight into variation in protein evolution in the context of *S. invicta* conditional gene expression.

Variation in rates of protein evolution may reflect differences in the strengths of either positive selection or purifying selection. Although adaptive changes in protein sequence clearly play an important evolutionary role (Sella et al. 2009), positive selection is not dominant in shaping overall rates of protein evolution in the 1,104 *S. invicta* genes we examined. Only three genes exhibited clear evidence of positive selection in *S. invicta* when averaged across all codons (dN/dS > 1; mean dN/dS ratio of 0.12 (SEM ± 0.0034) for the remaining proteins; Table A.1). Four additional genes exhibited lineage-specific positive selection in *S. invicta* according to branch-site tests (Zhang et al. 2005) among four ant taxa using highly stringent criteria (Fletcher and Yang 2010) (Table A.2). Together, these results suggest that correlations observed between conditional gene expression and rates of protein evolution are shaped largely by variation in purifying selection in *S. invicta* (Barker et al. 2005; Wall et al. 2005; Brisson and Nuzhdin 2008; Mank and Ellegren 2009; Snell-Rood et al. 2010; Van Dyken and Wade 2010).

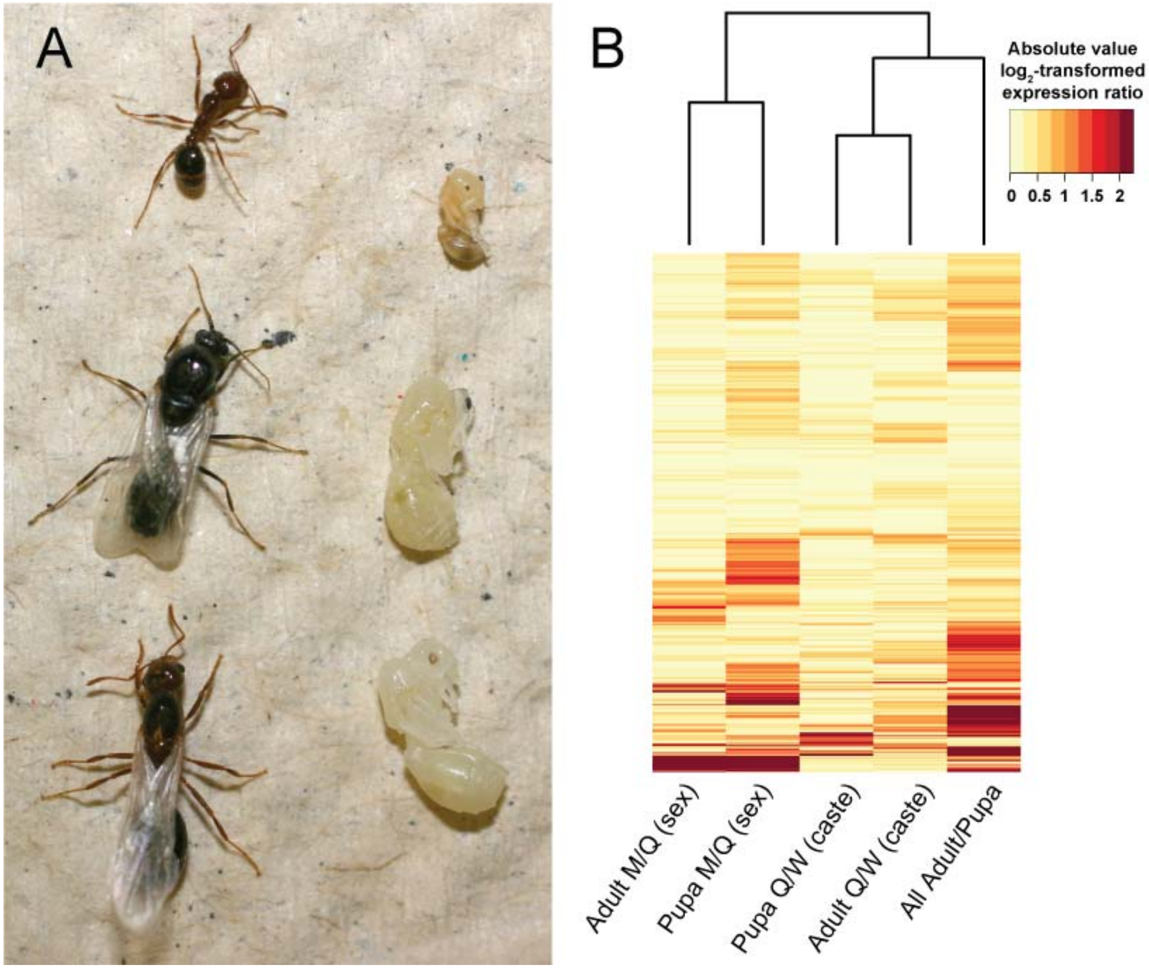

Figure 6.1. Distinct gene expression profiles associated with dramatic phenotypic differences in the fire ant *S. invicta*. (A) Adults (left) and pupae (right) are pictured from top to bottom as follows: worker, male, and queen. (B) Widespread gene expression differences between pupal and adult queens (Q), workers (W), and males (M) for 1,101 genes, where a $\log_2$-transformed expression ratio of one corresponds to a twofold difference in expression.

Our analyses revealed that the rate of protein evolution (dN/dS) in *S. invicta* is significantly correlated with two principal components (PC 1 and PC 3) representing the pooled effects of conditional gene expression and gene expression level, respectively (Fig. 6.2, Fig. A.1). The strengths of correlations with PC 1 and PC 3 were similar, but in opposite direction, indicating that conditional gene expression and gene expression level bear similarly important, but independent, associations with protein evolution (Fig. 6.2B). Correlations between conditional gene expression and protein evolutionary rate were significant for sex, caste, and developmental stage in *S. invicta* (Fig. A.2), and the rate of protein evolution was roughly proportional to the degree of differential expression among phenotypes (Snell-Rood et al. 2010) (Fig. 6.3A). Furthermore, this association was roughly additive in nature (i.e., larger for genes differentially expressed across sexes, castes, and developmental stages than for genes with differential expression in only a single context; Figs. 6.3B and 6.3C).

Interestingly, the association between protein evolution and conditional gene expression differed according to caste (Fig. 6.3D). Queen-biased genes, unlike worker-biased genes, exhibited significantly higher rates of protein evolution when compared to unbiased genes (Fig. 6.3D). This result is consistent with prior analysis of the honeybee *A. mellifera* (Hunt et al. 2010b), which found that queen-biased genes evolve rapidly relative to unbiased or worker-biased genes. Thus, in contrast with theoretical predictions (Linksvayer and Wade 2009), selective constraints on non-reproductive worker phenotypes may be as strong or stronger than those acting on reproductive queen phenotypes. Adult-biased genes also disproportionately contributed to the higher protein evolutionary rates of developmentally-biased genes (as compared to unbiased genes; Fig. 6.3D).

Our results are in general agreement with theory (Linksvayer and Wade 2009; Snell-Rood et al. 2010; Van Dyken and Wade 2010), suggesting that conditional gene expression is associated with relaxed selection. Nevertheless, we wished to gain further

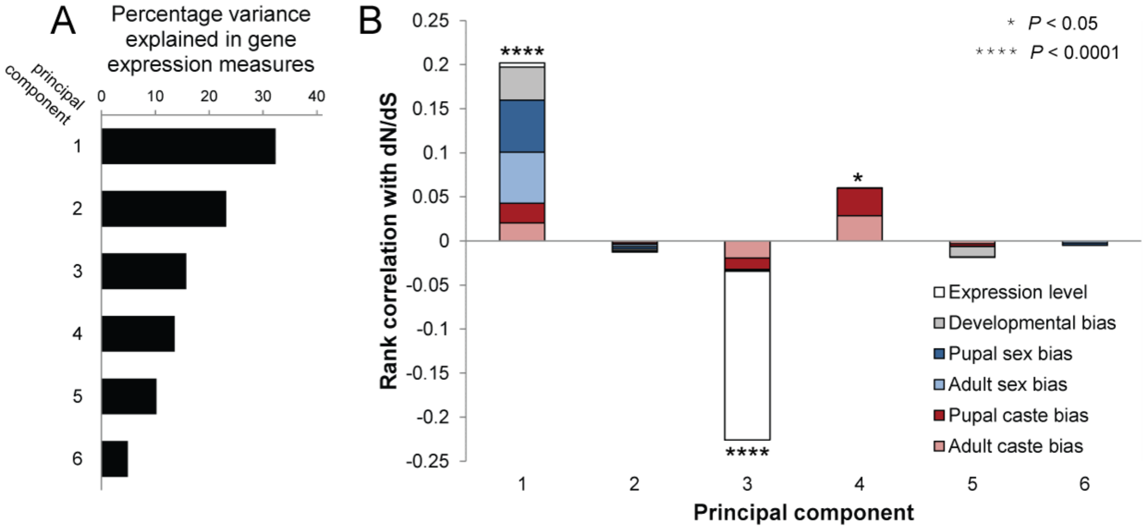

Figure 6.2. Associations between gene expression and protein evolution in the fire ant *S. invicta*. Principal component analysis of *S. invicta* gene expression measures. (A) The percentage variance explained by six principal components composed of expression level, developmental bias, pupal sex expression bias, adult sex expression bias, pupal caste expression bias, and adult caste expression bias. (B) The relative contribution of each gene expression measure to each principal component when normalized by the Spearman's rank correlation coefficient of each principal component with *S. invicta* evolutionary rate (dN/dS).

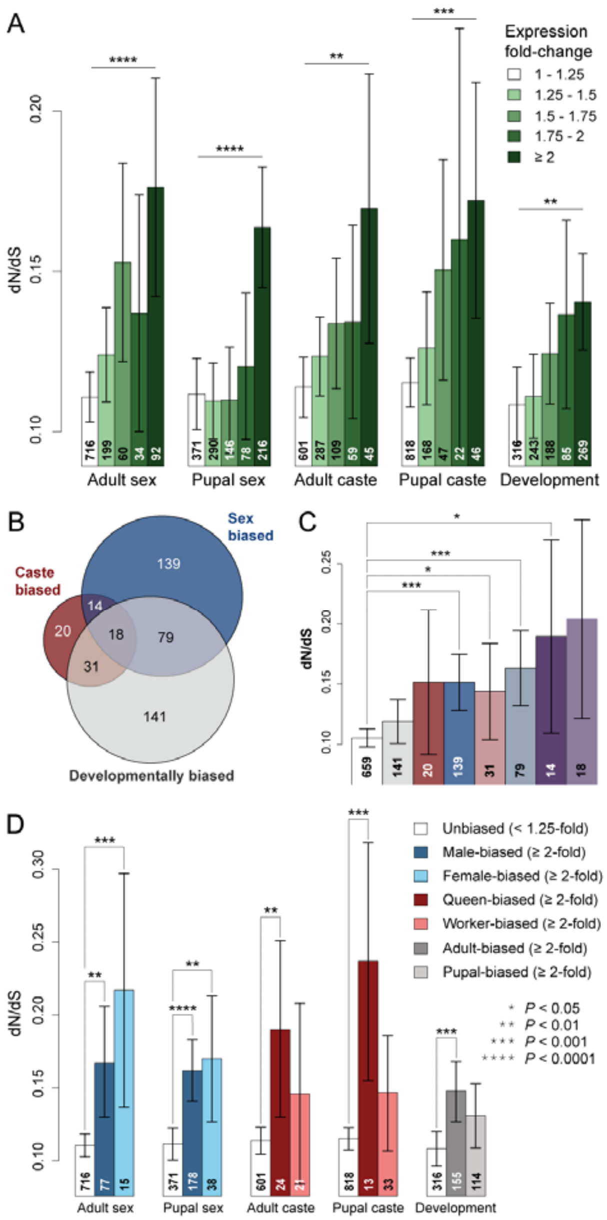

Figure 6.3. Relaxed selection is ubiquitously associated with conditional gene expression in the fire ant *S. invicta*. (A) Protein evolutionary rate (dN/dS) for genes showing varying levels of differential expression between sexes, castes, and developmental stages (significance determined by Kruskal-Wallis tests). (B) The overlap of genes with greater than twofold difference in expression between sexes, castes, and developmental stages. (C) Protein evolutionary rate for each section of the Venn diagram in part B (significance determined by Bonferroni-corrected Wilcoxon signed-rank tests in comparisons with all genes exhibiting less than twofold expression difference among phenotypic groups). (D) Genes with greater than twofold expression difference, categorized according to the phenotypic group (significance determined by Bonferroni-corrected Wilcoxon signed-rank tests). Means with 95% confidence intervals are plotted in panels A, C, and D; text in bars denotes the number of genes in each bin.

insight into the timescales at which rates of protein evolution vary with respect to conditional gene expression. Eusocial insect castes arose recently relative to sexes and distinct holometabolous developmental stages. Thus, we analyzed the evolutionary histories of caste-biased genes and their orthologs in lineages with common ancestry predating the origin of castes by comparing lineage-specific rates of amino acid substitution (dN) among the ants *S. invicta* and *P. barbatus*, and two non-eusocial insects, the jewel wasp *Nasonia vitripennis* and *Drosophila melanogaster* (Fig. 6.4A). We discovered that the correlation between caste-biased gene expression and the rate of molecular evolution was strongest in the *S. invicta* lineage (Fig. 6.4B), as predicted if conditional gene expression itself results in relaxed selection. Surprisingly, the non-eusocial wasp *N. vitripennis* also exhibited a significant positive correlation between amino acid substitution and *S. invicta* caste-biased gene expression (Fig. 6.4B), indicating that relaxed selection on orthologs of genes with differential expression between *S. invicta* castes predated the evolutionary origin of castes.

Rates of evolution of caste-biased genes in *A. mellifera* (Grozinger et al. 2007) exhibited the same trend as caste-biased genes in *S. invicta*: *A. mellifera* caste-biased gene expression was most strongly associated with dN for the *A. mellifera* lineage, but was also significantly correlated with dN in *S. invicta* and *N. vitripennis* (Figs. 6.4C and 6.4D). Furthermore, the ratios of queen-to-worker gene expression we used from *A. mellifera* (brain tissue) and *S. invicta* (whole-bodies) were not significantly correlated (Spearman's $\rho = 0.0321$, $P = 0.5558$, $n = 339$). Thus, our analysis of *A. mellifera* provides further evidence that relaxed selection on proteins is a precursor to conditional gene expression.

Our results suggest that a relaxation of purifying selection increases a gene's likelihood of adopting conditional expression and contributing to phenotypic plasticity. Thus, genes with relatively high dispensability (Wall et al. 2005; Mank and Ellegren 2009) may be best preconditioned to resolve phenotypic antagonism (Bonduriansky and

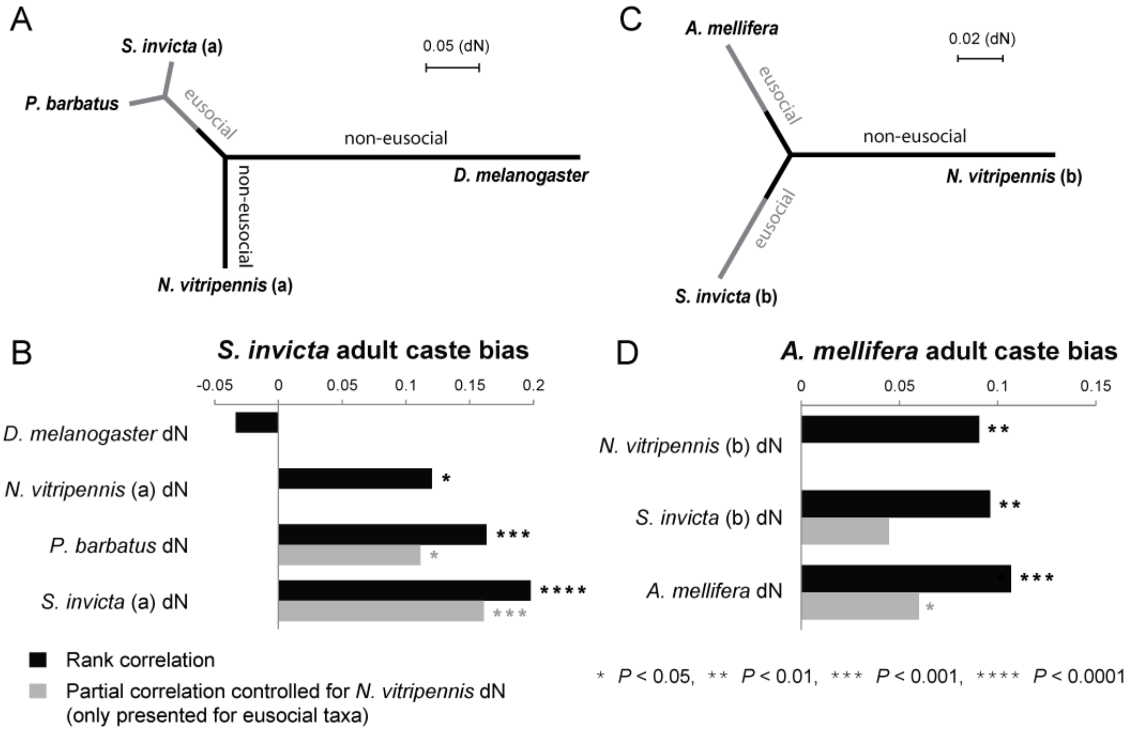

Figure 6.4. Evolutionary histories of conditionally-expressed genes. (A) A phylogeny based on mean nonsynonymous substitution rate (dN) for 447 ortholog groups from the ants *Solenopsis invicta* and *Pogonomyrmex barbatus*, the non-eusocial wasp *Nasonia vitripennis* and *Drosophila melanogaster*. (B) Correlations between *S. invicta* adult caste bias and dN for terminal branches in part A. (C) A phylogeny based on mean dN for 1,152 ortholog groups including the honeybee *Apis mellifera*, *S. invicta*, and *N. vitripennis*. (D) Correlations between *A. mellifera* adult caste bias and dN for branches from part C. Black bars in B and D represent Spearman's rank correlations between branch-specific dN and gene expression measures. Grey bars in B and D represent Spearman's rank partial correlations controlled for *N. vitripennis* dN to illustrate trends observed within eusocial taxa when controlling for the coincidence of associations between dN and caste bias in a non-eusocial relative.

Chenoweth 2009) between environmental or developmental contexts through differential expression, a process which may play an important role in evolutionary innovation (Pfennig et al. 2010).

Methods

Gene sequences

The following official gene sets (OGS) were used in our analyses: *Solenopsis invicta* OGS 2.2.0 (Wurm et al. in press), *Pogonomyrmex barbatus* OGS 1.1 (in press-b), *Linepithema humile* OGS 1.1 (in press-a), *Harpegnathos saltator* OGS 3.3 (Bonasio et al. 2010), *Apis mellifera* OGS 1 (Honeybee Genome Sequencing Consortium 2006), *Nasonia vitripennis* OGS 1.2 (Nasonia Genome Working Group 2010), and *Drosophila melanogaster* flybase release 5.21 (Crosby et al. 2007). *S. invicta* sequences are available from Fourmidable (Wurm et al. 2009) (http://fourmidable.unil.ch). *P. barbatus*, *L. humile*, *A. mellifera*, and *N. vitripennis* sequences are available from the Hymenoptera Genome Database (Munoz-Torres et al.) (http://hymenopteragenome.org). *D. melanogaster* sequences are available from flybase (Crosby et al. 2007) (http://flybase.org).

Ortholog determination

Pairwise orthology between proteins from each species was assigned based on reciprocal blastp (Altschul et al. 1997) hits with an E-value cutoff of E < 1e-10. Blastp output was then parsed with a custom bioperl script (Stajich et al. 2002). Pairwise reciprocal best hits and orthologous groups with three-way shared orthology between (i) *S. invicta*, *P. barbatus*, and *L. humile* and (ii) *S. invicta*, *A. mellifera*, and *N. vitripennis*, and four-way shared orthology between (iii) *S. invicta*, *P. barbatus*, *N. vitripennis*, and *D. melanogaster* and (iv) *S. invicta*, *P. barbatus*, *L. humile*, and *H. saltator* were assigned using custom perl scripts. We identified 6,900 three-way orthologous groups between *S. invicta*, *P. barbatus*, and *L. humile* with this method.

Sequence alignment

MUSCLE 3.8.31 (Edgar 2004) was used to generate protein alignments for each orthologous group with default settings. Pal2nal v13 (Suyama et al. 2006) was then used to back-translate codon alignments from protein alignments with the 'nomismatch' setting. Gblocks 0.91b (Talavera and Castresana 2007) was used to extract confidently aligned regions. Resulting alignments were converted to PHYLIP interleaved format using Readseq 2.1.27 (http://iubio.bio.indiana.edu/soft/molbio/readseq/java/).

To confirm branch-site tests of positive selection, described below, a separate alignment procedure, which has been shown to result in fewer false positive results (Fletcher and Yang 2010), was undertaken. In this case, PRANK (Löytynoja and Goldman 2005) (release 100802) was used to generate codon alignments with default settings. PRANK output was then run through Gblocks and converted to PHYLIP interleaved format as described above.

Evolutionary rates

PhyML 3.0 (Guindon and Gascuel 2003) was used to generate maximum likelihood phylogenies from codon alignments for each orthologous group with the 'HKY85' nucleotide substitution model, a minimum parsimony starting tree, transition/transversion ratio estimation, and subtree pruning and regrafting tree topology estimation. PAML 4.4b (Yang 2007) was then used to estimate synonymous and nonsynonymous substitution rates (dS and dN) and ω (dN/dS) for each orthologous group using a phylogenetic tree produced by PhyML as input and the 'F3x4' codon frequency model.

For three-way orthologs among *S. invicta*, *P. barbatus*, and *L. humile* and three-way orthologs among *S. invicta*, *A. mellifera*, and *N. vitripennis*, substitution rates were averaged across all codons for a given protein (NSsites = 0) with free dN/dS ratios for each branch (model = 1). For four-way orthologs among *S. invicta*, *P. barbatus*, *N. vitripennis*, and *D. melanogaster*, only orthologous groups with PhyML phylogenies matching the most common unrooted topology were used. In this case, substitution rates were again averaged across all codons for a given protein with free dN/dS ratios for each branch. Three genes with evidence of positive selection (dN/dS > 1) were identified using this approach. In each case, dS was zero and dN/dS values were infinite. These three genes were eliminated from further analysis to avoid skewing means, but are summarized in Table A.1.

To test for further evidence of positive selection, we used the branch-site test of positive selection (branch-site model A, test 2) (Zhang et al. 2005), which uses likelihood ratio tests to detect positive selection on a small number of sites along a specific lineage. For this test, we used MUSCLE alignments of four-way orthologs among the ants *S. invicta*, *P. barbatus*, *L. humile*, and *H. saltator* to construct phylogenetic trees using PhyML as before. *S. invicta* was set as the foreground branch, with two dN/dS ratios for branches (model = 2). For the alternative model, ω_2 was estimated from the data (fix_omega = 0, omega = 1), while the null model fixes ω_2 at one (fix_omega = 1, omega = 1) as described in PAML documentation and sample files (Yang 2007). The log-likelihoods for the null and alternative models were used to calculate a likelihood ratio test statistic, which was then compared against the χ^2 distribution (degrees of freedom = 1, with a critical value of 3.84 at a 5% significance level) (Yang 2007). *P*-values were Bonferroni-corrected according to the number of tests of significance performed (n = 861; α/n = 5.8e$^-$5 used as the threshold for significance). This highly conservative threshold was implemented because of the documented occurrence of false-positive results in branch-site tests of positive selection as a result of alignment errors (Fletcher and Yang 2010). To better control for the influence of alignment errors, PRANK codon alignments were used as input in an alternative analysis of positive selection (Fletcher and Yang 2010). For analysis of PRANK codon alignments, a bonferroni cutoff of 1.4e-4 was used (n =365). The consensus set of significant genes was then taken. 40 of 861 orthologs with MUSCLE alignments in our dataset retained a significant signal of positive selection on the *S. invicta* branch, while 5 of 365 orthologs with PRANK alignments in our dataset retained a significant signal of positive selection on the *S. invicta* branch. The consensus set consisted of four proteins (Supplementary Table 2). PRANK codon alignments were not used in other analyses due to the extensive computation time required.

In each analysis of evolutionary rates, only genes with a length of ≥ 100 codons analyzed by PAML were used in our results. Phylogeny topology, foreground branch specification, branch-specific substitution rates, and log-likelihoods were processed and parsed using custom perl scripts.

S. invicta gene expression

Gene expression measures were obtained using custom cDNA microarrays (Wang et al. 2007; Wurm et al. 2010). We estimated the relative expression of each clone for whole-body adult and pupal queens, workers, and males using the Bayesian approach implemented in the program Bayesian Analysis of Gene Expression Levels (BAGEL) (Townsend and Hartl 2002) as described previously (Ometto et al. in press). This BAGEL expression level was used to determine all $\log_2$-transformed ratios of gene expression between phenotypic groups. We used the absolute value of $\log_2$-transformed pairwise BAGEL expression ratios to measure the degree of sex-bias in expression (male vs. queen) and caste-bias in expression (queen vs. worker) for adults and pupae separately. To measure developmental bias, the absolute value of the $\log_2$-transformed ratio of summed adult male, worker, and queen BAGEL values versus summed pupal male, worker, and queen BAGEL values was taken.

The overall gene expression level for each of the array clones was calculated as follows. For each hybridization experiment, gene expression level was estimated for each clone as the ratio between its net signal intensity and the average net signal intensity across clones. In particular, for each hybridization experiment *h*, we assigned to each clone *c* a relative expression

$$E_c^h = \frac{\left(F_c^h - B_c^h\right)}{\sum_{i=1}^{n} \frac{\left(F_i^h - B_i^h\right)}{n}},$$

where n is the number of *S. invicta* good quality clones (as flagged in Genepix), and F and B are respectively the foreground and background signal intensities for either the green or red channel, depending on the labeling of the sample. Then the overall gene expression level for each clone was calculated across all k hybridizations as

$$\overline{E}_c = \sum_{i=1}^{k} E_c^i.$$

S. invicta array probe to genome mapping

Sanger EST sequences of array clones were mapped to *S. invicta* genome scaffolds (assembly Si_gnF) using GMAP version 2010-03-09 (Wu and Watanabe 2005). The best hit for each EST sequence was taken according to the genomic coordinates with the highest percent identity and only ESTs with ≥ 95% identity and 50% coverage were retained. Genome coordinates and strand information were extracted for each *S. invicta* gene from the OGS 2.2.0 GFF file, which was subsequently used to map ESTs to genes using a custom perl script. Only ESTs which mapped to genome coordinates that overlapped the coding sequence of a gene on the correct strand were considered representative probes for a given gene. 12 ESTs mapped to multiple genes and were removed from further analysis.

The mean relative BAGEL expression level and the overall gene expression level of duplicate spots on the array were first calculated and then the R aggregate function was used to take the mean of expression values for all clones that mapped to a given gene. This procedure resulted in gene expression measures for 2,088 unique genes (represented by 4,624 array clones). We filtered our dataset to include only *S. invicta* genes with proteins represented by a single cDNA and the highest possible quality rank (as indicated by the tag 'quality=5'). We thus obtained a total of 1,104 genes with measures of both evolutionary rates in three-species analysis and gene expression levels (1,101 following removal of genes with infinite dN/dS values, which were used in all statistical analyses).

S. invicta gene characteristics

Normalized CpG dinucleotide content (CpG o/e) acted as a metric of levels of DNA methylation and was calculated as described previously (Elango et al. 2009).

Synonymous 3rd codon position GC content (GC3s) and the frequency of optimal codons (Fop) were calculated using CodonW (http://codonw.sourceforge.net). Fop was calculated with automatic detection implemented to use the top 5% of genes (in terms of bias in codon usage). As is the case with the honeybee (Jorgensen et al. 2007), GC content and codon usage were tightly linked in *S. invicta* (GC3s vs. Fop Spearman rank correlation = -0.996; Supplementary Table 2).

Coding sequence length and intron counts were calculated from the OGS 2.2.0 GFF file using a custom perl script.

A. mellifera gene expression

Data on gene expression bias between brains of queens and workers in *A. mellifera* were obtained from a study by Grozinger et al. (2007). Data on expression breadth among tissues in *A. mellifera* were obtained from a study by Foret et al. (2009) and processed as described previously (Hunt et al. 2010a). RNA-Seq data on expression levels in *A. mellifera* whole-body workers were obtained from a study by Zemach et al. (2010).

Miscellaneous statistical tests

The JMP 8.0.2 statistical software package (SAS Institute Inc, Cary, NC) was used to calculate 95% confidence intervals, Spearman rank correlations, Pearson product-moment correlations, Kruskal-Wallis tests, and principal component analysis. R 2.10.1 (R Development Core Team 2010) was used to calculate partial correlations, as described previously (Kim and Yi 2006), to generate a heatmap with dendrograms of gene expression profiles, and to calculate Wilcoxon signed-rank tests. The heatmap was generated using Euclidean distances between rows and columns, with dendrograms computed by performing a hierarchical cluster analysis using a set of dissimilarities for the *n* objects being clustered, according to the ‘complete’ method. Proportional Venn diagrams were generated using the Venn Diagram Plotter available from Pacific Northwest National Laboratory (http://omics.pnl.gov).

Acknowledgements

This research was supported by the US National Science Foundation (grant DEB 0640690 to MG and SY), the Swiss National Science Foundation, an ERC Advanced Grant and the Rectorate of the University of Lausanne (to LK), and by USDA AFRI (award no. 2009-35302-05301 to DS). We thank Karl Glastad for helpful discussion and Amanda Palaski and Ariel Thompson for the photograph used in Fig. 6.1.

CHAPTER 7

CONCLUSIONS

This dissertation encompasses five studies aimed at understanding molecular evolution in social insects. These studies focused on evolutionary contrasts of gene expression patterns, DNA methylation, and protein-coding sequences. From these analyses, a common theme has emerged: genes with conditional expression tend to be less important to fitness than ubiquitously expressed genes, as indicated by a reduction in selective constraint.

The results from chapter two, *Evolutionary variation in gene expression is associated with dimorphism in eusocial vespid wasps*, suggest that caste-biased gene expression may be more labile than sex-biased gene expression and demonstrate the utility of social insects in addressing questions of the molecular evolution of phenotypic dimorphisms. This study was among the first to investigate evolutionary variation in gene expression among both castes and sexes. The high turnover of genes associated with caste differences between closely related taxa provides support to the idea that caste-biased genes, as a class, are subject to reduced selective constraints.

The results in chapter three, *DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera*, represented an important step in establishing the honeybee as a model for studying the epigenetics of behavior and for understanding the role of DNA methylation in insects. In particular, this study highlighted the pervasiveness of DNA methylation in the genome of *A. mellifera*, providing fertile grounds for studies of phenotypic plasticity and genomic imprinting. This study was particularly well-timed, with rapid, subsequent growth in the field of social insect genomics (Bonasio et al. 2010; Smith et al. in press-a; Smith et al. in press-b; Wurm et al. in press) and epigenomics (Lyko et al. 2010). It is important to note, however, that the vast attention being paid to epigenetic causes of behavioral variation, in social insects as in humans, continues to be rooted by relatively few empirical examples (Miller 2010). Time will tell if DNA methylation is as important to caste differences in the honeybee and other social insects as is presently hypothesized (Kucharski et al. 2008; Maleszka 2008; Moczek and Snell-Rood 2008).

The study comprising chapter four, *Functional conservation of DNA methylation in the pea aphid and the honeybee*, expanded upon our understanding of DNA methylation in insects. In response to the newfound prominence of the study of behavioral epigenomics among social insect biologists, this manuscript made an effort to present a clear picture of the function of DNA methylation in insects at large. In fact, a unity of function for intragenic methylation in diverse animal taxa has recently been emerging (Feng et al. 2010; Zemach et al. 2010). We found that broadly expressed genes among tissues and alternative phenotypes are subject to the highest levels of methylation in invertebrates. As in mammals, intragenic methylation in insects may limit alternative splicing by masking intragenic promoters (Maunakea et al. 2010). Thus, broadly expressed genes may be preferentially targeted by DNA methylation due to enhanced negative effects associated with spurious transcription at loci with greater importance to fitness (Duret and Mouchiroud 2000). These data are consistent with a primary role of intragenic methylation in providing robustness to broadly expressed genes. Thus, the relative scarcity of DNA methylation in conditionally-expressed genes supports the hypothesis that genes with conditional expression have lower impacts on overall fitness relative to ubiquitously expressed genes.

The study comprising chapter five, *Sociality is linked to rates of protein evolution in a highly social insect*, examined the interaction between highly social behavior and dynamics of protein evolution for the first time. As a collaborative effort with several key members of the *Nasonia vitripennis* genome consortium (Nasonia Genome Working Group 2010), this study examined rates of protein evolution over the course of hundreds of millions of years. At this time scale, genes with relative upregulation of expression in brains of *A. mellifera* queens exhibited higher rates of protein evolution compared with ubiquitously expressed genes, as predicted if queen-biased genes are subject to diminished selective constraints relative to ubiquitously expressed genes. However, worker-biased genes did not exhibit a similar signal.

The study comprising chapter six, *Relaxed selection is a precursor to the evolution of phenotypic plasticity*, improved upon the analyses in chapter five in several important ways and benefitted from the arrival of several new ant genome sequences (Bonasio et al. 2010; Smith et al. in press-a; Smith et al. in press-b; Wurm et al. in press). First, microarray data from coauthors of this chapter (Ometto et al. in press) encompassed analyses of developmental stages, sexes, and castes. Second, the synonymous substitution rate in protein-coding genes (dS) was not saturated in the phylogenetic analysis underlying this study. Controlling for increased mutation rates, as measured by dS, in worker-biased genes relative to unbiased genes revealed that worker-biased genes contributed to the signal of faster protein evolution for conditionally-expressed genes in *S. invicta*. Indeed, the rate of molecular evolution was universally higher (among sexes, castes, and developmental stages) for conditionally expressed genes in *S. invicta*, as predicted by theory (Barker et al. 2005; Linksvayer and Wade 2009; Snell-Rood et al. 2010; Van Dyken and Wade 2010). Surprisingly, we also found that orthologs of genes with caste-biased expression exhibit heightened lineage-specific rates of amino acid substitution, even in taxa lacking castes. These results suggested that a relaxation of purifying selection increases a gene's likelihood of adopting conditional expression and contributing to phenotypic plasticity. Thus, genes with relatively low impacts on fitness (Wall et al. 2005; Mank and Ellegren 2009) may be best preconditioned to resolve phenotypic antagonism (Bonduriansky and Chenoweth 2009) between environmental or developmental contexts through differential expression, a process which may play an important role in evolutionary innovation (Pfennig et al. 2010).

My research has demonstrated that conditionally expressed genes are subject to relaxed purifying selection relative to ubiquitously expressed genes, with relaxed selection acting as both a precursor and consequence of conditional gene expression. This dissertation suggests that diminished selective constraint on protein sequence corresponds with increased rates of evolution in gene expression (Castillo-Davis et al. 2004; Zhang et al. 2007). It is an intriguing possibility that the majority of adaptive evolutionary changes associated with phenotypic plasticity occur at the level of gene expression (Wray 2007; Fraser et al. 2010). With rapid advances in sequencing technology fueling the burgeoning field of social insect comparative genomics, future studies will be able to fully disentangle the relative contributions of neutral and adaptive processes to evolutionary changes in proteins and gene expression patterns associated with phenotypic plasticity.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 6

Supplementary Text

Positive selection in the fire ant *Solenopsis invicta*

To better understand the evolutionary processes giving rise to higher rates of protein evolution for conditionally expressed genes compared with unbiased genes, we investigated the prevalence of positive selection in our data. Our analysis confirmed that changes in few genes were subject to positive selection at the scale of the length of an entire protein (Koonin and Wolf 2010) (Table A.1). Indeed, only three of 1,104 genes had dN/dS values > 1 in the *S. invicta* lineage, and the signal in one of those genes was driven by only two substitutions (Table A.1). Our stringent criteria (Fletcher and Yang 2010) for branch-site detection of positive selection on specific codons in *S. invicta* revealed four genes that have undergone site-specific positive selection (Table A.2). Of all seven genes putatively subject to positive selection, three exhibited upregulation in pupal males (Tables A.1 and A.2). None of the other genes exhibited twofold differences in gene expression among alternate phenotypes.

Correlates of protein evolutionary rate (dN/dS) in *S. invicta*

Analysis of potential evolutionary rate determinants (Lemos et al. 2005; Pal et al. 2006; Hunt et al. 2010a) in *S. invicta* revealed that gene expression level was negatively correlated with dN/dS in *S. invicta* (Fig. A.2F) as is prevalent in diverse taxa (Drummond and Wilke 2008). Coding sequence length was positively correlated with dN/dS (Lemos et al. 2005) and intron number was negatively correlated with dN/dS (Carmel et al. 2007). Interestingly, CpG o/e, a measure of CpG depletion which is negatively correlated with methylation in *S. invicta* (Wurm et al. in press), was positively correlated with dN/dS (Fig. A.2F) and was largely decoupled from GC content and expression level

(Table A.3). Thus, methylated genes may be under greater functional constraint in *S. invicta* compared to unmethylated genes (Hunt et al. 2010a; Lyko et al. 2010).

Synonymous substitution rates (dS) and caste in *S. invicta*

In the case of sex differences and developmental differences, dN was the primary driver of signal for increased dN/dS relative to unbiased genes (Fig. A.3). Although dN is also associated with caste differences, synonymous substitution (dS) is significantly lower for caste-biased genes relative to unbiased genes at both the pupal and adult stages (Fig. A.3). Thus, caste-biased genes may have lower mutation rates or be subject to increased selection on synonymous codon usage relative to unbiased genes.

The relationship between caste bias and distinct forms of conditional gene expression

To determine whether the association between relaxed selection and caste-biased gene expression is driven by other forms of conditional gene expression, we first analyzed tissue-specificity of gene expression in *A. mellifera* (Foret et al. 2009). Our analyses revealed that the association between rates of protein evolution and caste-biased gene expression in *A. mellifera* was not weakened when controlling for tissue-specificity among six tissue types (Spearman's partial correlation between *A. mellifera* dN from Fig. 6.4C and caste bias, when controlling for tissue-specificity = 0.1137, P = 0.0001; Fig. A.4). Furthermore, the association between rate of protein evolution and caste-biased gene expression in *S. invicta* was not weakened when controlling for sex bias and developmental bias (Spearman's partial correlation between *S. invicta* dN from Fig. 6.4A and adult caste bias, when controlling for adult sex bias and developmental bias = 0.1797, P = 0.0001). These results are consistent with a scenario in which the observed correlations between caste bias and the rate of protein evolution are not explained by other forms of conditional gene expression.

Supplementary Tables and Figures

Table A.1. Genes under positive selection according to analysis of substitution rates averaged over all codons in a given gene.

S. invicta **gene ID**	**N*dN**	**S*dS**	**dN/dS**	**> 2-fold expression**	**AmiGO[a] BLAST protein similarity**	**Gene ontology biological process**
SI2.2.0_02232	28.8	0	∞	male pupa upregulation vs. queen pupa	no hits (P < 1e-5)	n/a
SI2.2.0_12283	17.7	0	∞	None	*Mus musculus* 1110059E24Rik (P = 2.9e-21)	no biological process annotation
SI2.2.0_09898[b]	2	0	∞	None	*Drosophila melanogaster* mago nashi (P = 1.7e-71)	cell-cell signaling (GO:0007267); nuclear mRNA splicing, via spliceosome (GO:0000398); oogenesis (GO:0048477); 11 other terms

[a] Results generated using AmiGOn(Carbon et al. 2009) gene ontology database release 2010-11-20.

[b] Note that the low number of substitutions in this gene severely limits the inference of positive selection.

Table A.2. Positive selection according to branch-site analysis.

***S. invicta* gene ID**	**Branch-site MUSCLE *P*-value**	**Branch-site PRANK *P*-value**	**> 2-fold expression**	**AmiGO[a] BLAST protein similarity**	**Gene ontology biological process**
SI2.2.0_02629	1.88e-11	1.19e-11	male adult upregulation vs. queen adult; male pupa upregulation vs. queen pupa	no hits ($P < 1e\text{-}5$)	n/a
SI2.2.0_05545	9.09e-7	8.07e-7	male pupa upregulation vs. queen pupa	*Drosophila melanogaster* Bifunctional Phosphopantetheine adenylyltransferase - Dephospho-CoA kinase ($P = 1.5e\text{-}95$)	coenzyme A biosynthetic process (GO:0015937); imaginal disc-derived wing morphogenesis (GO:0007476); ovarian follicle cell migration (GO:0007297); 2 other terms
SI2.2.0_11797	4.39e-6	1.94e-5	none	*Gallus gallus* Uncharacterized protein ($P = 3.6e\text{-}249$)	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (GO:0006139)
SI2.2.0_12264	4.90e-10	1.37e-4	none	*Pan troglodytes* Dyslexia susceptibility 1 candidate gene 1 protein homolog ($P = 8.7e\text{-}56$)	neuron migration (GO:0001764); regulation of estrogen receptor signaling pathway (GO:0033146); regulation of proteasomal protein catabolic process (GO:0061136)

[a] Results generated using AmiGO (Carbon et al. 2009) gene ontology database release 2010-12-04.

Table A.3. Associations between *S. invicta* codon usage, CpG depletion, expression level, and evolutionary rates.

	Principal component[b]		
	1	**2**	**3**
Percent variance explained in CpG o/e, expression level, Fop, and GC3s[a]	60.923	24.715	14.265
Rank correlation with *S. invicta* dN/dS	-0.02	-0.19****	0.18****
Rank correlation with *S. invicta* dN	-0.09**	-0.17****	0.18****
Rank correlation with *S. invicta* dS	-0.17****	0.03	0.01
Percent contributions to principal component			
CpG o/e [a]	22.4	2.7	74.9
Expression Level	1.3	97.1	1.6
Fop [a]	38.2	0.1	11.8
GC3s [a]	38.2	0.1	11.7

[a] CpG o/e is normalized CpG dinucleotide content, Fop is the frequency of optimal codons, GC3s is 3rd codon position synonymous GC content

[b] Principal component 4 explains 0.10 % of the variance in CpG o/e, expression level, Fop, and GC3s and is not presented

** $P < 0.01$; **** $P < 0.0001$

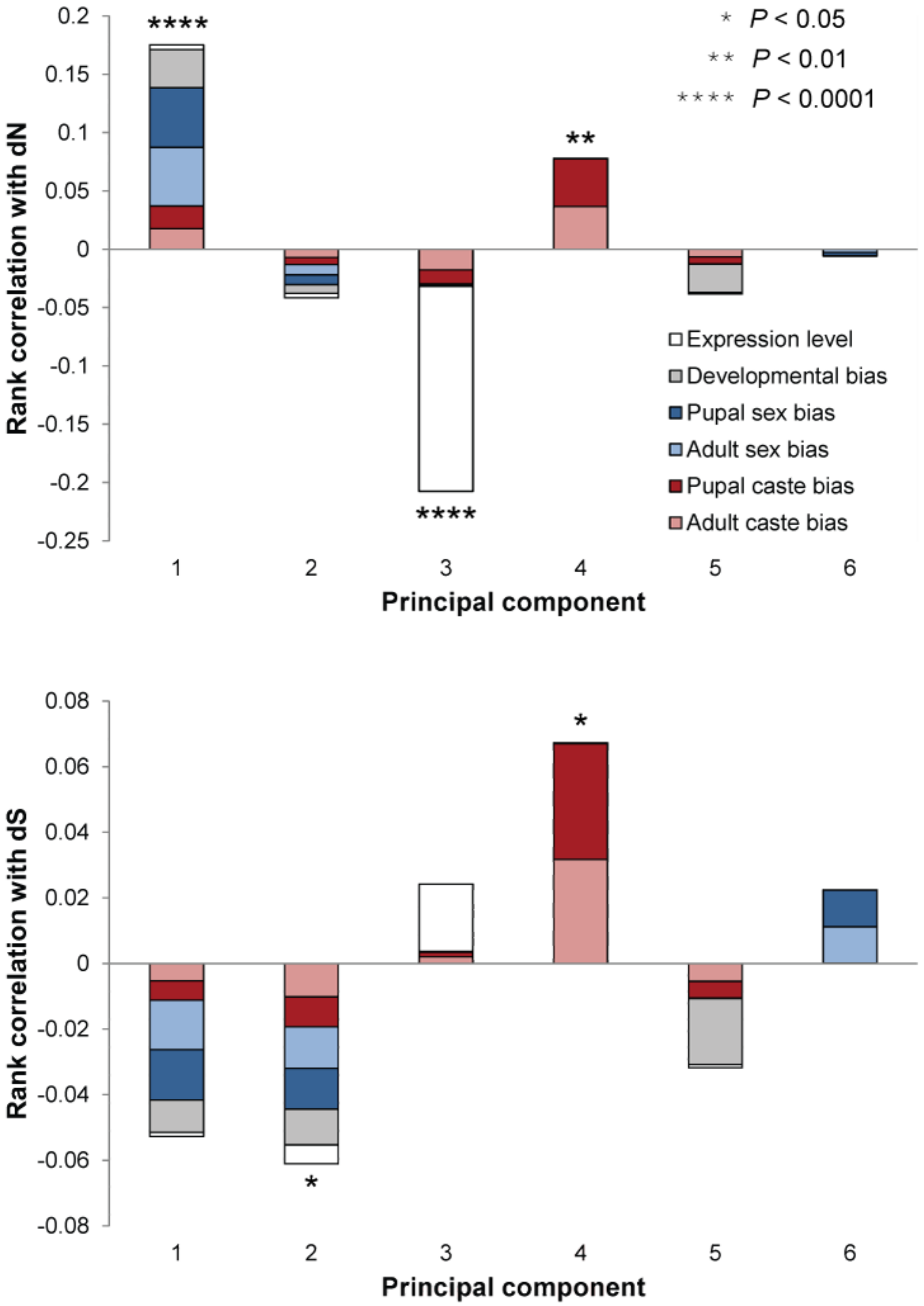

Figure A.1. Principal component analysis of gene expression measures and evolutionary rates in *S. invicta*. Bars represent the relative contribution of each gene expression measure to each principal component when normalized by the Spearman's rank correlation coefficient of each principal component with *S. invicta* nonsynonymous substitution (dN) and synonymous substitution (dS).

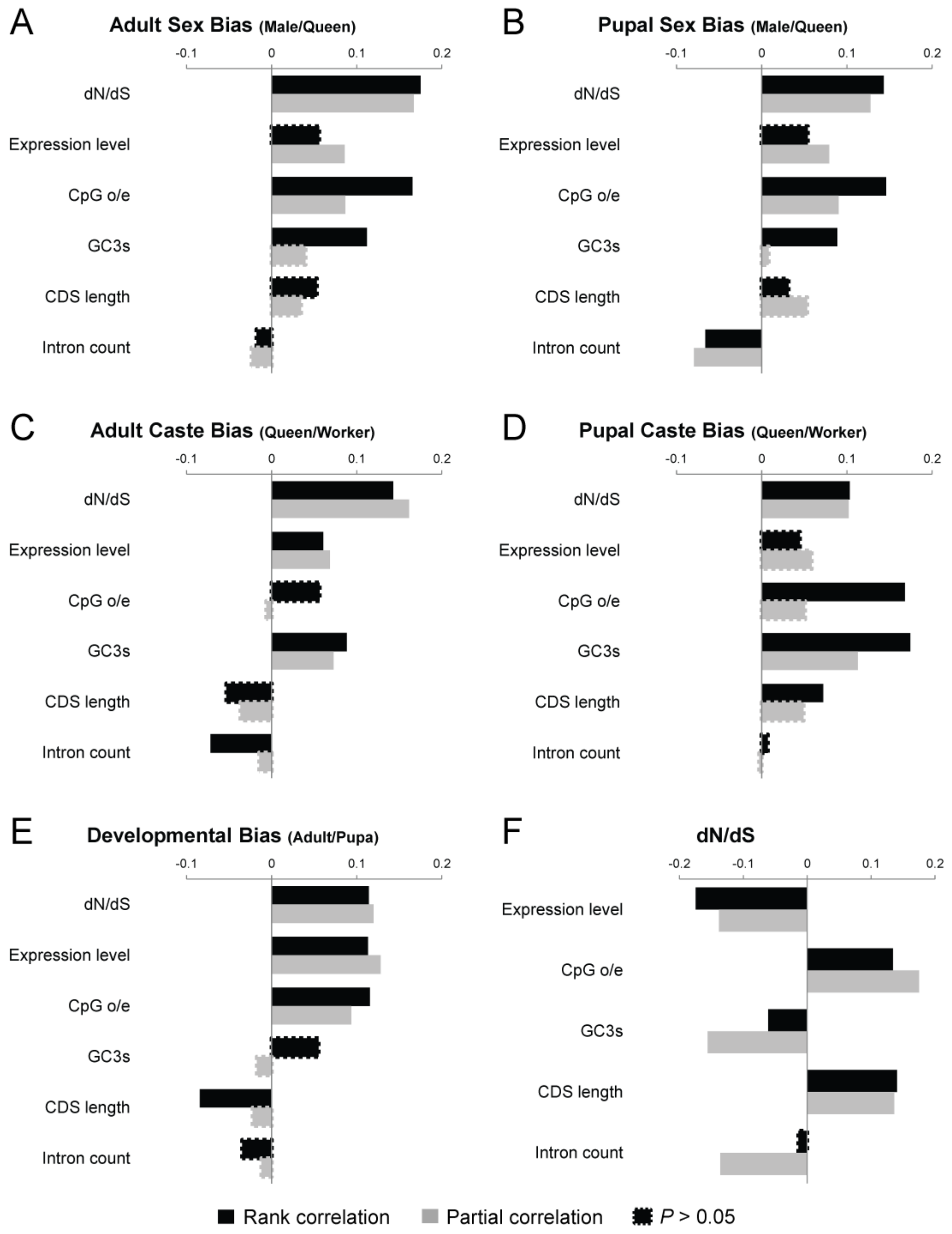

Figure A.2. Correlations between protein evolutionary rate (dN/dS), conditional gene expression, and other evolutionary rate determinants in *S. invicta*. Gene expression level, CpG o/e (a negative correlate of DNA methylation), 3rd codon position synonymous GC content (GC3s), and coding sequence (CDS) length are each significantly correlated with evolutionary constraint (dN/dS) in *S. invicta*. Partial rank correlations with dN/dS for each aforementioned variable and intron count are significant when controlling for all other variables. The relative gene expression biases between sexes, between castes, and between developmental stages are each associated with dN/dS when controlling for all other variables. Black bars represent Spearman's rank correlations. Grey bars represent partial correlations controlling for all other variables.

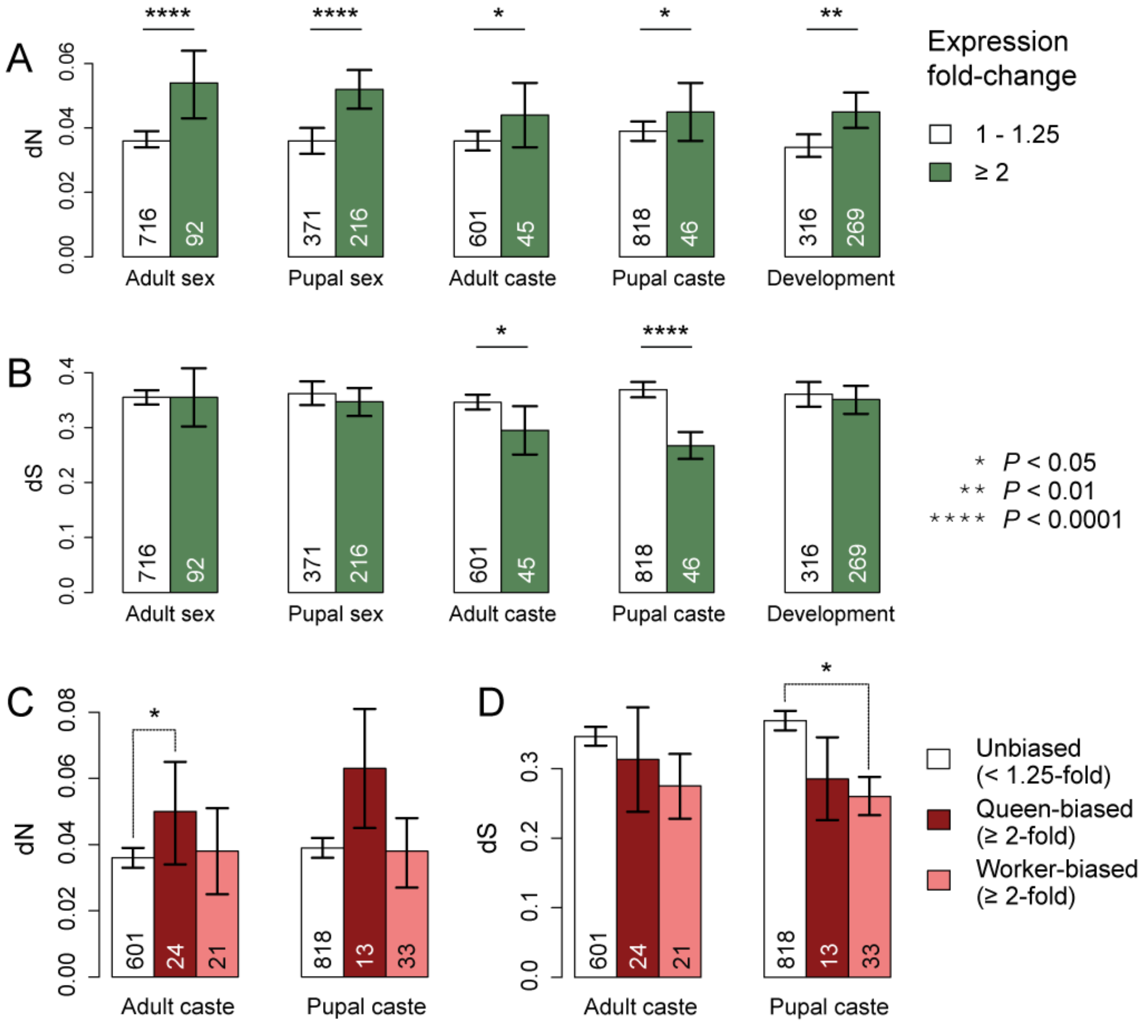

Figure A.3. Synonymous and nonsynonymous substitution rates in relation to conditional gene expression. Means with 95% confidence intervals are plotted. (A) dN denotes rates of nonsynonymous substitution. (B) dS denotes rates of synonymous substitution. (C) Dissection of the unique properties of caste-biased genes with respect to dN and dS. Significance denotes results of Wilcoxon signed-rank tests, subject to Bonferroni-correction in C and D.

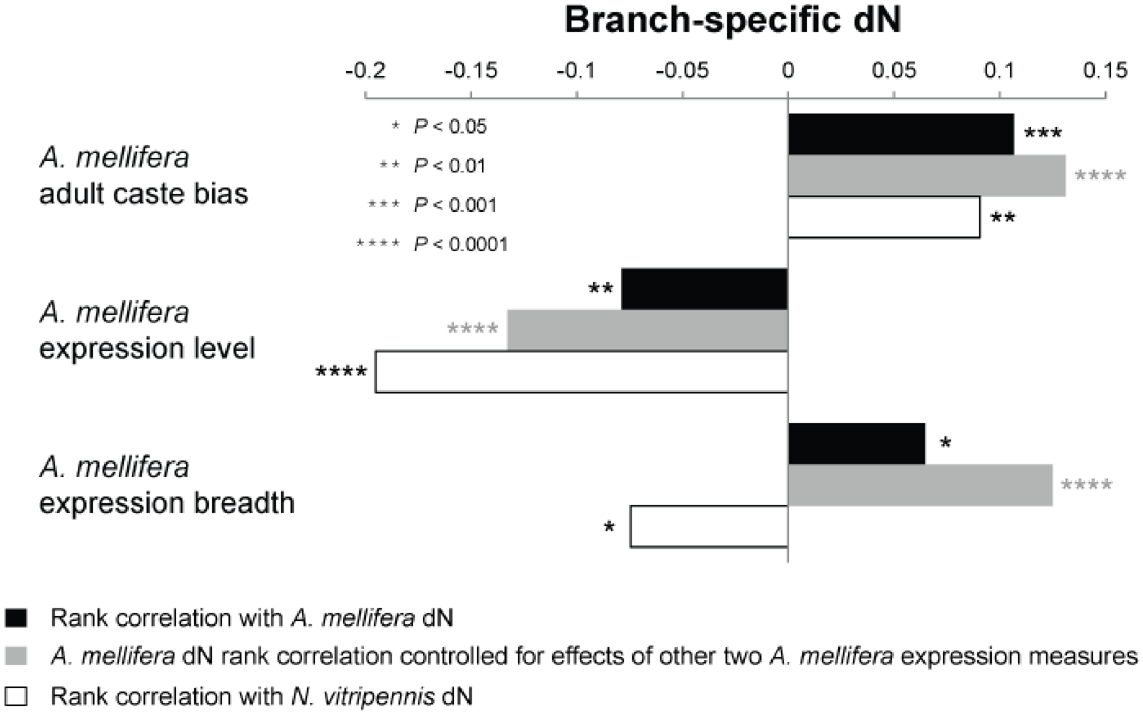

Figure A.4. Protein evolutionary rate, caste bias, gene expression level, and gene expression breadth in *A. mellifera*. Bars represent Spearman's rank correlations and partial correlations. Branch-specific dN was determined for a three-species phylogeny among *A. mellifera*, *S. invicta*, and *N. vitripennis*, as in Fig. 6.4C. 1,152 ortholog groups were analyzed with brain gene expression measures among castes (Grozinger et al. 2007), worker whole-body gene expression levels according to RNA-seq analysis (Zemach et al. 2010), and the number of tissues with observed expression (ranging from 1-6) (Foret et al. 2009).

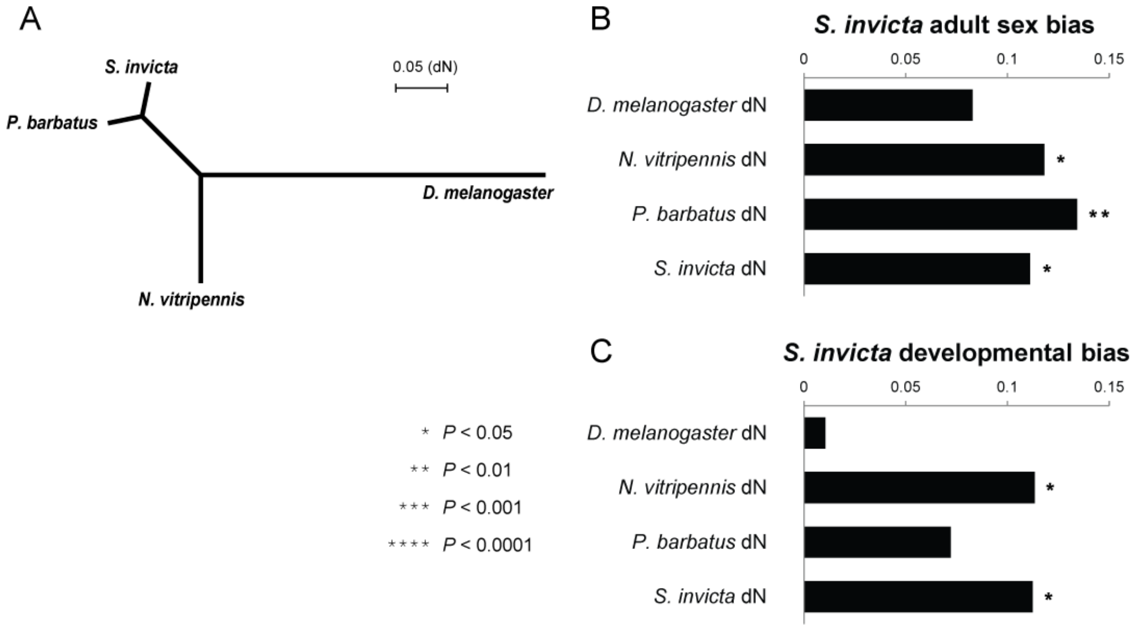

Figure A.5. Evolutionary histories of sex-biased and developmentally-biased genes. (A) A phylogeny based on mean nonsynonymous substitution rate (dN) for 447 ortholog groups from the ants *Solenopsis invicta* and *Pogonomyrmex barbatus*, the non-eusocial wasp *Nasonia vitripennis* and the fruit fly *Drosophila melanogaster*. Black bars represent Spearman's rank correlations between dN for terminal branches from part A and (B) *S. invicta* adult sex bias or (C) *S. invicta* developmental bias.

REFERENCES

Abouheif E, Wray GA. 2002. Evolution of the gene network underlying wing polyphenism in ants. Science 297:249-252.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Applied Biosystems. 2001. Relative quantitation of gene expression. User Bulletin #2: ABI Prism 7700 Sequence Detection System.

Archer ME. 2006. Taxonomy, distribution and nesting biology of species of the genus *Dolichovespula* (Hymenoptera, Vespidae). Entomol Sci 9:281-293.

Aron S, de Menten L, Van Bockstaele DR, Blank SM, Roisin Y. 2005. When hymenopteran males reinvented diploidy. Curr Biol 15:824-827.

Ashburner M, et al. 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25:25-29.

Ayroles JF, et al. 2009. Systems genetics of complex traits in *Drosophila melanogaster*. Nat Genet 41:299-307.

Barchuk AR, Cristino AS, Kucharski R, Costa LF, Simoes ZLP, Maleszka R. 2007. Molecular determinants of caste differentiation in the highly eusocial honeybee Apis mellifera. BMC Dev Biol 7:70.

Barker MS, Demuth JP, Wade MJ. 2005. Maternal expression relaxes constraint on innovation of the anterior determinant, bicoid. PLoS Genet 1:527-530.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. Trends Genet 21:219-226.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes Dev 16:6-21.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499-1504.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185-193.

Bonasio R, et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science 329:1068-1071.

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

Brisson JA. 2010. Aphid wing dimorphisms: linking environmental and genetic control of trait variation. Philosophical Transactions of the Royal Society B-Biological Sciences 365:605-616.

Brisson JA, Davis GK, Stern DL. 2007. Common genome-wide patterns of transcript accumulation underlying the wing polyphenism and polymorphism in the pea aphid (*Acyrthosiphon pisum*). Evol Dev 9:338-346.

Brisson JA, Nuzhdin SV. 2008. Rarity of males in pea aphids results in mutational decay. Science 319:58-58.

Brisson JA, Stern DL. 2006. The pea aphid, *Acyrthosiphon pisum*: an emerging genomic model system for ecological, developmental and evolutionary studies. BioEssays 28:747-755.

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, the AmiGO Hub, the Web Presence Working Group. 2009. AmiGO: Online access to ontology and annotation data. Bioinformatics 25:288-289.

Carmel L, Rogozin IB, Wolf YI, Koonin EV. 2007. Evolutionarily conserved genes preferentially accumulate introns. Genome Res 17:1045-1050.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. Cell 134:25-36.

Castillo-Davis CI, Hartl DL, Achaz G. 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res 14:1530-1536.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540 - 552.

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.

Corona M, Estrada E, Zurita M. 1999. Differential expression of mitochondrial genes between queens and workers during caste determination in the honeybee Apis mellifera. J Exp Biol 202:929-938.

Corona M, Velarde RA, Remolina S, Moran-Lauter A, Wang Y, Hughes KA, Robinson GE. 2007. Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. Proc Natl Acad Sci USA 104:7128-7133.

Cristino AS, et al. 2006. Caste development and reproduction: a genome-wide analysis of hallmarks of insect eusociality. Insect Mol Biol 15:703-714.

Crosby M, Goodman J, Strelets V, Zhang P, Gelbart W. 2007. FlyBase: Genomes by the dozen. Nucleic Acids Res 35:D486-D491.

Cui X, Kerr MK, Churchill GA. 2003. Transformations for cDNA microarray data. Stat Appl Genet Mol Biol 2:Article4.

Cutter AD, Ward S. 2005. Sexual and temporal dynamics of molecular evolution in *C. elegans* development. Mol Biol Evol 22:178-188.

DeLuca TF, Wu IH, Pu J, Monaghan T, Peshkin L, Singh S, Wall DP. 2006. Roundup: a multi-genome repository of orthologs and evolutionary distances. Bioinformatics 22:2044-2046.

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4:R60.

Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. 2005. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet 37:544-548.

Drapeau MD, Albert S, Kucharski R, Prusko C, Maleszka R. 2006. Evolution of the Yellow/Major Royal Jelly Protein family and the emergence of social behavior in honey bees. Genome Res 16:1385-1394.

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

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134:341-352.

Duret L, Galtier N. 2000. The covariation between TpA deficiency, CpG deficiency, and G + C content of human isochores is due to a mathematical artifact. Mol Biol Evol 17:1620-1625.

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

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol Biol Evol 17:68-74.

Edgar BA, Orr-Weaver TL. 2001. Endoreplication cell cycles: More for less. Cell 105:297-306.

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

Elango N, Hunt BG, Goodisman MAD, Yi SV. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. Proc Natl Acad Sci USA 106:11206-11211.

Elango N, Kim SH, Vigoda E, Yi SV, NISC Comparative Sequencing Program. 2008. Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. PLoS Comput Biol 4:e1000015.

Elango N, Yi SV. 2008. DNA methylation and structural and functional bimodality of vertebrate promoters. Mol Biol Evol 25:1602-1608.

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

Evans JD, Wheeler DE. 1999. Differential gene expression between developing queens and workers in the honey bee, Apis mellifera. Proc Natl Acad Sci USA 96:5575-5580.

Evans JD, Wheeler DE. 2000. Expression profiles during honeybee caste determination. Genome Biol 2:1-6.

Evans JD, Wheeler DE. 2001. Gene expression and the evolution of insect polyphenisms. BioEssays 23:62-68.

Feng SH, et al. 2010. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci USA 107:8689-8694.

Field LM, Lyko F, Mandrioli M, Prantera G. 2004. DNA methylation in insects. Insect Mol Biol 13:109-115.

Fittkau EJ, Klinge H. 1973. On Biomass and Trophic Structure of the Central Amazonian Rain Forest Ecosystem. Biotropica 5:2-14.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27:2257-2267.

Foret S, Kucharski R, Pittelkow Y, Lockett G, Maleszka R. 2009. Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. BMC Genomics 10:472.

Foster KR, Ratnieks FLW, Gyllenstrand N, Thoren PA. 2001. Colony kin structure and male production in *Dolichovespula* wasps. Mol Ecol 10:1003-1010.

Fraser HB, Moses AM, Schadt EE. 2010. Evidence for widespread adaptive evolution of gene expression in budding yeast. Proc Natl Acad Sci USA 107:2977-2982.

Fryxell KJ, Moon WJ. 2005. CpG mutation rates in the human genome are highly dependent on local GC content. Mol Biol Evol 22:650-658.

Fryxell KJ, Zuckerkandl E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Mol Biol Evol 17:1371-1383.

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

Gaunt MW, Miles MA. 2002. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol Biol Evol 19:748-761.

Gibson G. 2008. The environmental contribution to gene expression profiles. Nat Rev Genet 9:575-581.

Gilbert SF. 2001. Ecological developmental biology: Developmental biology meets the real world. Dev Biol 233:1-12.

Goodisman MAD, Isoe J, Wheeler DE, Wells MA. 2005. Evolution of insect metamorphosis: A microarray-based study of larval and adult gene expression in the ant *Camponotus festinatus*. Evolution 59:858-870.

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.

Graff J, Jemielity S, Parker JD, Parker KM, Keller L. 2007. Differential gene expression between adult queens and workers in the ant *Lasius niger*. Mol Ecol 16:675-683.

Greene A. 1991. *Dolichovespula* and *Vespula*. In: KG Ross, RW Matthews, editors. The Social Biology of Wasps. Ithaca: Cornell University Press. p. 263-305.

Grimaldi D, Engel M. 2005. Cambridge: Cambridge University Press.

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.

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

Haerty W, et al. 2007. Evolution in the fast lane: Rapidly evolving sex-related genes in *Drosophila*. Genetics 177:1321-1335.

Haig D. 1992. Intragenomic conflict and the evolution of eusociality. J Theor Biol 156:401-403.

Haig D. 2000. The kinship theory of genomic imprinting. Annu Rev Ecol Syst 31:9-32.

Hamilton WD. 1964. The genetical evolution of social behaviour. II. The journal of theoretical biology 7:17-52.

Heid CA, Stevens J, Livak KJ, Williams PM. 1996. Real time quantitative PCR. Genome Res 6:986-994.

Heimpel GE, de Boer JG. 2008. Sex determination in the Hymenoptera. Annu Rev Entomol 53:209-230.

Hendrich B, Tweedie S. 2003. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. Trends Genet 19:269-277.

Hines HM, Hunt JH, O'Connor TK, Gillespie JJ, Cameron SA. 2007. Multigene phylogeny reveals eusociality evolved twice in vespid wasps. Proc Natl Acad Sci USA 104:3295-3299.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: Evo devo and the genetics of adaptation. Evolution 61:995-1016.

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

Hölldobler B, Wilson EO. 1990. The Ants. Cambridge, MA: The Belknap Press of Harvard University Press.

Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443:931-949.

Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. 2003. Identifying biological themes within lists of genes with EASE. Genome Biol 4:P4.

Hughes WOH, Oldroyd BP, Beekman M, Ratnieks FLW. 2008. Ancestral monogamy shows kin selection is key to the evolution of eusociality. Science 320:1213-1216.

Hunt BG, Brisson JA, Yi SV, Goodisman MAD. 2010a. Functional conservation of DNA methylation in the pea aphid and the honeybee. Genome Biol Evol 2:719-728.

Hunt BG, Goodisman MAD. 2010. Evolutionary variation in gene expression is associated with dimorphism in eusocial vespid wasps. Insect Mol Biol 19:641-652.

Hunt BG, Ometto L, Wurm Y, Shoemaker D, Keller L, Yi SV, Goodisman MAD. in preparation. Relaxed selection is a precursor to the evolution of phenotypic plasticity.

Hunt BG, Wyder S, Elango N, Werren JH, Zdobnov EM, Yi SV, Goodisman MAD. 2010b. Sociality is linked to rates of protein evolution in a highly social insect. Mol Biol Evol 27:497-500.

Hunt JH. 2007. The Evolution of Social Wasps. New York: Oxford University Press.

Illingworth R, et al. 2008. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol 6:e22.

International Aphid Genomics Consortium. 2010. Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol 8:e1000313.

Jablonka E, Raz G. 2009. Transgenerational epigenetic inheritance: Prevalence, mechanisms, and implications for the study of heredity and evolution. Q Rev Biol 84:131-176.

Jaenisch R, Bird A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 33:245-254.

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet 29:389-395.

Jirtle RL, Skinner MK. 2007. Environmental epigenomics and disease susceptibility. Nat Rev Genet 8:253-262.

Jones PA, Takai D. 2001. The role of DNA methylation in mammalian epigenetics. Science 293:1068-1070.

Jorgensen FG, Schierup MH, Clark AG. 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.

Keller L, editor. 1999. Levels of Selection in Evolution. Princeton: Princeton University Press.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Pääbo S. 2004. A Neutral Model of Transcriptome Evolution. PLoS Biol 2:e132.

Kijimoto T, Costello J, Tang ZJ, Moczek AP, Andrews J. 2009. EST and microarray analysis of horn development in *Onthophagus* beetles. BMC Genomics 10:504.

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.

Klose RJ, Bird AP. 2006. Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci 31:89-97.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat Rev Genet 11:487-498.

Kovacs JL, Goodisman MAD. 2007. Irregular brood patterns and worker reproduction in social wasps. Naturwissenschaften 94:1011-1014.

Krauss V, Thummler C, Georgi F, Lehmann J, Stadler PF, Eisenhardt C. 2008. Near Intron Positions Are Reliable Phylogenetic Markers: An Application to Holometabolous Insects. Mol Biol Evol 25:821-830.

Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM. 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res 36:D271-275.

Kronforst MR, Gilley DC, Strassmann JE, Queller DC. 2008. DNA methylation is widespread across social Hymenoptera. Curr Biol 18:R287-R288.

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

Kumar S, Hedges SB. 1998. A molecular timescale for vertebrate evolution. Nature 392:917-920.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 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.

Li E. 2002. Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet 3:662-673.

Li E, Beard C, Jaenisch R. 1993. Role for DNA methylation in genomic imprinting. Nature 366:362-365.

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.

Linksvayer TA, Wade MJ. 2009. Genes with social effects are expected to harbor more sequence variation within and between species. Evolution 63:1685-1696.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci USA 102:10557-10562.

Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R. 2010. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. PLoS Biol 8:e1000506.

Macdonald JF, Matthews RW. 1981. Nesting biology of the eastern yellowjacket, *Vespula maculifrons* (Hymenoptera, Vespidae). J Kans Entomol Soc 54:433-457.

Macdonald JF, Matthews RW. 1984. Nesting biology of the southern yellowjacket, *Vespula squamosa* (Hymenoptera, Vespidae): Social parasitism and independent founding. J Kans Entomol Soc 57:134-151.

Maleszka R. 2008. Epigenetic integration of environmental and genomic signals in honey bees. Epigenetics 3:188-192.

Mank JE. 2009. The W, X, Y and Z of sex-chromosome dosage compensation. Trends Genet 25:226-233.

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.

Maunakea AK, et al. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466:253-257.

Maynard Smith J, Szathmary E. 1995. The Major Transitions in Evolution. New York: Oxford University Press.

Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. Proc Natl Acad Sci USA 100:9894-9899.

Meissner A, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454:766-770.

Mileyko Y, Joh RI, Weitz JS. 2008. Small-scale copy number variation and large-scale changes in gene expression. Proc Natl Acad Sci USA 105:16659-16664.

Miller G. 2010. The Seductive Allure of Behavioral Epigenetics. Science 329:24-27.

Moczek AP. 2007. Developmental capacitance, genetic accommodation, and adaptive evolution. Evol Dev 9:299-305.

Moczek AP, Andrews J, Kijimoto T, Yerushalmi Y, Rose DJ. 2007. Emerging model systems in evo-devo: horned beetles and the origins of diversity. Evol Dev 9:323-328.

Moczek AP, Snell-Rood EC. 2008. The basis of bee-ing different: the role of gene silencing in plasticity. Evol Dev 10:511-513.

Müller CB, Williams IS, Hardie J. 2001. The role of nutrition, crowding and interspecific interactions in the development of winged aphids. Ecol Entomol 26:330-340.

Munoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, Childs KL, Anzola JM, Milshina N, Elsik CG. 2011. Hymenoptera Genome Database: Integrated community resources for insect species of the order Hymenoptera. Nucleic Acids Res 39:D658-D662.

Nasonia Genome Working Group. 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 327:343-348.

Nijhout HF. 1999. When developmental pathways diverge. Proceedings of the national academy of sciences 96:5348-5350.

Nijhout HF. 2003. Development and evolution of adaptive polyphenisms. Evol Dev 5:9-18.

Oleksiak MF, Churchill GA, Crawford DL. 2002. Variation in gene expression within and among natural populations. Nat Genet 32:261-266.

Ometto L, Shoemaker D, Ross KG, Keller L. in press. Evolution of gene expression in fire ants: The effects of developmental stage, caste, and species. Mol Biol Evol.

Orr WC, Sohal RS. 1994. Extension of life-span by overexpression of superoxide dismutase and catalase in *Drosophila melanogaster*. Science 263:1128-1130.

Oster GF, Wilson EO. 1978. Caste and ecology in the social insects. Princeton: Princeton University Press.

Ott J. 1979. Detection of rare major genes in lipid-levels. Hum Genet 51:79-91.

Ott J. 1992. NOCOM and COMPMIX programs release. New York: Rockefeller University.

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

Parker JD, Parker KM, Sohal BH, Sohal RS, Keller L. 2004. Decreased expression of Cu-Zn superoxide dismutase 1 in ants with extreme lifespan. Proc Natl Acad Sci USA 101:3486-3489.

Patel A, Fondrk MK, Kaftanoglu O, Emore C, Hunt G, Frederick K, Amdam GV. 2007. The Making of a Queen: TOR Pathway Is a Key Player in Diphenic Caste Development. PLoS One 2:e509.

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

Pfennig DW, Wund MA, Snell-Rood EC, Cruickshank T, Schlichting CD, Moczek AP. 2010. Phenotypic plasticity's impacts on diversification and speciation. Trends Ecol Evol 25:459-467.

Pigliucci M, Murren CJ, Schlichting CD. 2006. Phenotypic plasticity and evolution by genetic assimilation. J Exp Biol 209:2362-2367.

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

Queller DC. 2003. Theory of genomic imprinting conflict in social insects. BMC Evol Biol 3:23.

R Development Core Team. 2010. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science 300:1742-1745.

Rasband WS. 1997-2009. ImageJ. Bethesda, Maryland, USA: U. S. National Institutes of Health.

Razin A, Riggs AD. 1980. DNA methylation and gene function. Science 210:604-610.

Regev A, Lamb MJ, Jablonka E. 1998. The role of DNA methylation in invertebrates: Developmental regulation or genome defense? Mol Biol Evol 15:880-891.

Reik W, Walter J. 2001. Genomic imprinting: parental influence on the genome. Nat Rev Genet 2:21-32.

Rifkin SA, Houle D, Kim J, White KP. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature 438:220-223.

Robinson GE, Fernald RD, Clayton DF. 2008. Genes and Social Behavior. Science 322:896-900.

Robinson GE, Grozinger CM, Whitfield CW. 2005. Sociogenomics: Social life in molecular terms. Nat Rev Genet 6:257-U216.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-386.

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

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci USA 103:1412-1417.

Schaefer M, Lyko F. 2007. DNA methylation with a sting: an active DNA methylation system in the honeybee. BioEssays 29:208-211.

Scharf ME, Wu-Scharf D, Zhou X, Pittendrigh BR, Bennett GW. 2005. Gene expression profiles among immature and adult reproductive castes of the termite *Reticulitermes flavipes*. Insect Mol Biol 14:31-44.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? PLoS Genet 5:e1000495.

Simmen MW, Leitgeb S, Charlton J, Jones SJM, Harris BR, Clark VH, Bird A. 1999. Nonmethylated transposable elements and methylated genes in a chordate genome. Science 283:1164-1167.

Sinha S, Ling X, Whitfield CW, Zhai CX, Robinson GE. 2006. Genome scan for cis-regulatory DNA motifs associated with social behavior in honey bees. Proc Natl Acad Sci USA 103:16352-16357.

Smit A, Hubley R, Green P. 1996-2004. RepeatMasker Open-3.0.

Smith CD, et al. in press-a. The draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). Proc Natl Acad Sci USA.

Smith CR, et al. in press-b. A draft genome of the red harvester ant *Pogonomyrmex barbatus*. Proc Natl Acad Sci USA.

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.

Smith MAH, MacKay PA. 1989. Genetic variation in male alary dimorphism in populations of pea aphid, *Acyrthosiphon pisum*. Entomol Exp Appl 51:125-132.

Snell-Rood EC, Van Dyken JD, Cruickshank T, Wade MJ, Moczek AP. 2010. Toward a population genetic framework of developmental evolution: the costs, limits, and consequences of phenotypic plasticity. BioEssays 32:71-81.

Stajich JE, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12:1611-1618.

Stearns SC. 1989. The evolutionary significance of phenotypic plasticity - phenotypic sources of variation among organisms can be described by developmental switches and reaction norms. Bioscience 39:436-445.

Strassmann JE, Queller DC. 2007. Insect societies as divided organisms: The complexities of purpose and cross-purpose. Proceedings of the national academy of sciences 104:8619-8626.

Sultan SE. 2007. Development in context: the timely emergence of eco-devo. Trends Ecol Evol 22:575-582.

Sumner S. 2006. Determining the molecular basis of sociality in insects: progress, prospects and potential in sociogenomics. Ann Zool Fenn 43:423-442.

Sumner S, Pereboom JJM, Jordan WC. 2006. Differential gene expression and phenotypic plasticity in behavioural castes of the primitively eusocial wasp, *Polistes canadensis*. Proc R Soc B 273:19-26.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609-W612.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9:465-476.

Suzuki MM, Kerr ARW, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res 17:625-631.

Suzuki Y, Gojobori T, Kumar S. 2009. Methods for incorporating the hypermutability of CpG dinucleotides in detecting natural selection operating at the amino acid sequence level. Mol Biol Evol 26:2275-2284.

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

Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

Toth AL, Bilof KBJ, Henshaw MT, Hunt JH, Robinson GE. 2009. Lipid stores, ovary development, and brain gene expression in *Polistes metricus* females. Insectes Soc 56:77-84.

Toth AL, Robinson GE. 2007. Evo-devo and the evolution of social behavior. Trends Genet 23:334-341.

Toth AL, et al. 2007. Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. Science 318:441-444.

Townsend J, Hartl D. 2002. Bayesian analysis of gene expression levels: Statistical quantification of relative mRNA level across multiple strains or treatments. Genome Biol 3:research0071-research0071.0016.

True JR, Haag ES. 2001. Developmental system drift and flexibility in evolutionary trajectories. Evol Dev 3:109-119.

Turillazzi S, West-Eberhard MJ. 1996. Natural History and Evolution of Paper-Wasps. New York: Oxford University Press.

Tweedie S, Charlton J, Clark V, Bird A. 1997. Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol Cell Biol 17:1469-1475.

Urieli-Shoval S, Gruenbaum Y, Sedat J, Razin A. 1982. The absence of detectable methylated bases in *Drosophila melanogaster* DNA. FEBS Lett 146:148-152.

Van Dyken JD, Wade MJ. 2010. The genetic signature of conditional expression. Genetics 184:557-570.

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci USA 102:5483-5488.

Walsh TK, Brisson JA, Robertson HM, Gordon K, Jaubert-Possamai S, Tagu D, Edwards OR. 2010. A functional DNA methylation system in the pea aphid, *Acyrthosiphon pisum*. Insect Mol Biol 19:215-228.

Wang J, Jemielity S, Uva P, Wurm Y, Graff J, Keller L. 2007. An annotated cDNA library and microarray for large-scale gene-expression studies in the ant *Solenopsis invicta*. Genome Biol 8:R9.

Wang Y, Jorda M, Jones PL, Maleszka R, Ling X, Robertson HM, Mizzen CA, Peinado MA, Robinson GE. 2006. Functional CpG methylation system in a social insect. Science 314:645-647.

Wang Y, Leung F. 2009. In silico prediction of two classes of honeybee genes with CpG deficiency or CpG enrichment and sorting according to gene ontology classes. J Mol Evol 68:700-705.

Weaver ICG, Cervoni N, Champagne FA, D'Alessio AC, Sharma S, Seckl JR, Dymov S, Szyf M, Meaney MJ. 2004. Epigenetic programming by maternal behavior. Nat Neurosci 7:847-854.

Weber M, Hellmann I, Stadler MB, Ramos L, Paabo S, Rebhan M, Schubeler D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39:457-466.

Weil T, Korb J, Rehli M. 2009. Comparison of queen-specific gene expression in related lower termite species. Mol Biol Evol 26:1841-1850.

West-Eberhard MJ. 2003. Developmental Plasticity and Evolution. New York: Oxford University Press.

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

Wheeler DE, Buck N, Evans JD. 2006. Expression of insulin pathway genes during the period of caste determination in the honey bee, Apis mellifera. Insect Mol Biol 15:597-602.

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 A, Dunbar H, Davis G, Hunter W, Stern D, Moran N. 2006. A dual-genome microarray for the pea aphid, *Acyrthosiphon pisum*, and its obligate bacterial symbiont, *Buchnera aphidicola*. BMC Genomics 7:50.

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.

Wolffe AP, Matzke MA. 1999. Epigenetics: regulation through repression. Science 286:481-486.

Wolschin F, Amdam GV. 2007. Comparative proteomics reveal characteristics of life-history transitions in a social insect. Proteome Science 5.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206-216.

Wray GA, Abouheif E. 1998. When is homology not homology? Curr Opin Genet Dev 8:675-680.

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

Wu TD, Watanabe CK. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859-1875.

Wurm Y, Uva P, Ricci F, Wang J, Jemielity S, Iseli C, Falquet L, Keller L. 2009. Fourmidable: A database for ant genomics. BMC Genomics 10:5.

Wurm Y, Wang J, Keller L. 2010. Changes in reproductive roles are associated with changes in gene expression in fire ant queens. Mol Ecol 19:1200-1211.

Wurm Y, et al. in press. The genome of the fire ant *Solenopsis invicta*. Proc Natl Acad Sci USA.

Xiang H, et al. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat Biotechnol 28:516-U181.

Yang ZH. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yi SV, Goodisman MAD. 2009. Computational approaches for understanding the evolution of DNA methylation in animals. Epigenetics 4:551-556.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13:335-340.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328:916-919.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472-2479.

Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B. 2007. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. Nature 450:233-238.