Sociogenomics of social organization: Mechanistic and evolutionary underpinnings of caste development and facial recognition in paper wasps

by

Ali Berens

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Amy L. Toth, Co-major Professor
Peng Liu, Co-major Professor
Karin Dorman
Xiaoqiu Huang
Nicole Valenzuela

Iowa State University
Ames, Iowa
2015

Copyright © Ali Berens, 2015. All rights reserved.
DEDICATION

This work is dedicated to my grandmothers - two strong women whom I greatly admire.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ix</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xi</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Advances in Genomics</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Sociogenomics</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Social Insects</td>
<td>3</td>
</tr>
<tr>
<td>1.4 <em>Polistes</em> Paper Wasp Model System</td>
<td>4</td>
</tr>
<tr>
<td>1.5 Dissertation Organization</td>
<td>6</td>
</tr>
<tr>
<td>CHAPTER 2. COMPARATIVE TRANSCRIPTOMICS OF CONVERGENT EVOLUTION: DIFFERENT GENES BUT CONSERVED PATHWAYS UNDERLIE CASTE PHENOTYPES ACROSS LINEAGES OF EUSOCIAL INSECTS</td>
<td>8</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Results</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Discussion</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Conclusion</td>
<td>21</td>
</tr>
<tr>
<td>2.5 Materials and Methods</td>
<td>22</td>
</tr>
<tr>
<td>2.6 Acknowledgements</td>
<td>28</td>
</tr>
</tbody>
</table>
CHAPTER 3. NOURISHMENT LEVEL AFFECTS CASTE-RELATED GENE EXPRESSION IN *POLISTES* WASPS

3.1 Background ..................................................... 39
3.2 Results ............................................................... 42
3.3 Discussion ........................................................... 48
3.4 Conclusions .......................................................... 50
3.5 Methods ............................................................... 51
3.6 Acknowledgements .................................................. 56

CHAPTER 4. CANDIDATE GENES FOR INDIVIDUAL RECOGNITION IN *POLISTES FUSCATUS* PAPER WASPS

4.1 Introduction ........................................................... 62
4.2 Materials and Methods ............................................. 65
4.3 Results ................................................................. 72
4.4 Discussion ............................................................. 75
4.5 Conclusions ........................................................... 77
4.6 Acknowledgements .................................................. 77

CHAPTER 5. BRAIN GENE EXPRESSION ASSOCIATED WITH FACIAL RECOGNITION IN *POLISTES* PAPER WASPS

5.1 Introduction ........................................................... 84
5.2 Methods ............................................................... 87
5.3 Results ................................................................. 92
5.4 Discussion ............................................................. 95
5.5 Conclusions ........................................................... 98
5.6 Acknowledgements .................................................. 98

CHAPTER 6. GENERAL CONCLUSIONS ............................................ 102

APPENDIX A. ADDITIONAL TABLES AND FIGURES FOR CHAPTER 2 .......................................................... 109
APPENDIX B. ADDITIONAL METHODS, RESULTS, TABLES, AND FIGURES FOR CHAPTER 3 ................................. 121

APPENDIX C. ADDITIONAL METHODS, TABLES, AND FIGURES FOR CHAPTER 4 ........................................... 132

REFERENCES .................................................................................................................................................. 138


**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Outline Of Two Major Hypotheses: Genetic Toolkit And Novel Gene</td>
<td>30</td>
</tr>
<tr>
<td>Table 2.2</td>
<td><em>Polistes metricus</em> Transcriptome Assembly Statistics</td>
<td>31</td>
</tr>
<tr>
<td>Table 2.3</td>
<td><em>Polistes metricus</em> Transcriptome Functional Annotation Statistics And</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>GO Enrichment Statistics</td>
<td></td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Summary Of Transcriptome-Wide Analyses Of Caste Determination Across</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>Three Social Insect Lineages</td>
<td></td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Primer Sequences For Candidate And Control Genes</td>
<td>78</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Classification Of Candidate Genes For Individual Recognition</td>
<td>79</td>
</tr>
<tr>
<td>Table A.1</td>
<td>Directionality Of Differentially Expressed Transcripts Across Social Lineages</td>
<td>110</td>
</tr>
<tr>
<td>Table A.2</td>
<td>Log Fold Correlation Of Differentially Expressed Orthologs Across Social Lineages</td>
<td>111</td>
</tr>
<tr>
<td>Table A.3</td>
<td><em>Polistes metricus</em> RNA-Seq Libraries</td>
<td>112</td>
</tr>
<tr>
<td>Table A.4</td>
<td><em>Polistes metricus</em> Bowtie2 Alignment Results</td>
<td>113</td>
</tr>
<tr>
<td>Table A.5</td>
<td>Number Of Putative Orthologs Across Social Lineages</td>
<td>114</td>
</tr>
<tr>
<td>Table A.6</td>
<td>Number Of Differentially Expressed Transcripts Across Social Lineages</td>
<td>115</td>
</tr>
<tr>
<td>Table A.7</td>
<td>Summary Information About The KEGG Pathway Analysis</td>
<td>116</td>
</tr>
<tr>
<td>Table A.8</td>
<td>Log Fold Correlation Of Differentially Expressed Orthologs Across Social Lineages</td>
<td>117</td>
</tr>
<tr>
<td>Table A.9</td>
<td>Number Of Putative Orthologs Across Social Lineages</td>
<td>118</td>
</tr>
<tr>
<td>Table A.10</td>
<td>Number Of Differentially Expressed Transcripts Across Social Lineages</td>
<td>119</td>
</tr>
<tr>
<td>Table A.11</td>
<td>Summary Information About The KEGG Pathway Analysis</td>
<td>120</td>
</tr>
<tr>
<td>Table A.12</td>
<td>Log Fold Correlation Of Differentially Expressed Orthologs Across Social Lineages</td>
<td>121</td>
</tr>
<tr>
<td>Table A.13</td>
<td>Number Of Putative Orthologs Across Social Lineages</td>
<td>122</td>
</tr>
<tr>
<td>Table A.14</td>
<td>Number Of Differentially Expressed Transcripts Across Social Lineages</td>
<td>123</td>
</tr>
<tr>
<td>Table A.15</td>
<td>Summary Information About The KEGG Pathway Analysis</td>
<td>124</td>
</tr>
<tr>
<td>Table B.1</td>
<td>Nourishment Responsive and Location Dependent Transcripts</td>
<td>124</td>
</tr>
<tr>
<td>Table C.1</td>
<td>Primer Sequences For Candidate And Control Genes</td>
<td>134</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Taxonomic Groupings Of <em>Polistes metricus</em> Transcriptome Annotations</td>
<td>31</td>
</tr>
<tr>
<td>2.2</td>
<td>Heat Map Of <em>Polistes</em> Caste Differentially Expressed Transcripts</td>
<td>33</td>
</tr>
<tr>
<td>2.3</td>
<td>Pairwise Comparisons Of Differentially Expressed Transcripts, KEGG</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Pathways, And Enriched GO Terms Across Three Social Insect Lineages</td>
<td></td>
</tr>
<tr>
<td>2.4</td>
<td>Comparison Of Differentially Expressed Transcripts, KEGG Pathways,</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>And GO Categories Across Three Social Insect Lineages And Differently</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expressed Novel Transcripts Within Polistes Lineage</td>
<td></td>
</tr>
<tr>
<td>2.5</td>
<td>Abridged Glycolysis / Gluconeogenesis Metabolism Pathway</td>
<td>36</td>
</tr>
<tr>
<td>2.6</td>
<td>Conceptual Visualization Of Multilevel Effects Of Transcripts, Metabolic</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Pathways, And Biological Functions</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Heat Map Of Nourishment Responsive Differentially Expressed Transcripts</td>
<td>57</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison of Nourishment- and Caste-biased Differentially Expressed</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Transcripts, KEGG Pathways, and GO Categories</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Bar Chart Of Nourishment Enriched GO Categories</td>
<td>59</td>
</tr>
<tr>
<td>3.4</td>
<td>Comparison Of Nourishment Responsive Transcripts And GO Categories Across</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Species</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Social Interactions Behavior Assay</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>Memory Recall Top Candidates And Additional Candidate Genes</td>
<td>81</td>
</tr>
<tr>
<td>4.3</td>
<td>Manipulation Of Social Interactions Behavioral Assay By Meeting And</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Treatment</td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>Manipulation Of Social Interactions Gene Expression</td>
<td>83</td>
</tr>
</tbody>
</table>
Figure 5.1  Wasp Face And Pattern Training Ability  ........................................ 99
Figure 5.2  Heat Map And Correlation Of Visual Stimuli-Related Expression  ... 100
Figure 5.3  Visual Stimuli Expression Of Individual Memory Candidate Genes  ... 101

Figure A.1 Heat Maps Of The 16 Previously Identified Caste-related Genes ... 109
Figure A.2 Scatterplot Of Log Fold Change Of Differentially Expressed Orthologs Across Social Lineages ................................................................. 117
Figure A.3 Abridged Arginine And Proline Metabolism Pathway .................... 118
Figure A.4 Number Of Modulated KEGG Pathways By Threshold .................... 119
Figure A.5 Number Of Modulated KEGG Pathways Across Social Lineages .... 120

Figure B.1 Fold Change Of The 23 Previously Identified Nourishment-responsive Genes ............................................................ 125
Figure B.2 Heat Maps Of Shared Nourishment And Caste Differentially Expressed Transcripts ................................................................. 126
Figure B.3 Clustering Of The Nourishment-specific Mean Normalized Expression Profiles ................................................................. 127
Figure B.4 Comparison Of The Number Of Transcripts Identified By The Modeling And List Comparison Approaches ................................. 128
Figure B.5 Clustering Of The Interaction-specific Mean Normalized Expression Profiles ................................................................. 129
Figure B.6 Clustering Of Shared Nourishment and Interaction Mean Normalized Expression Profiles ................................................................. 130
Figure B.7 Bar Chart Of Nourishment and Caste Enriched GO Categories .... 131

Figure C.1  Polistes metricus Candidate Gene Expression  ............................. 135
Figure C.2  Polistes fuscatus Candidate Gene Expression  ............................. 136
Figure C.3 Manipulation Of Social Interactions Behavior Assay  ..................... 137
ACKNOWLEDGEMENTS

First and foremost, I am grateful to my advisor, Amy Toth, for taking a chance on me as her first graduate student even though I didn’t know anything about social insects or evolution when I started. Thanks to her support, I have become a better scientist and academic. She challenged me to think about my work in a broader context and to artfully communicate my ideas. Amy has been dedicated to making sure that my pursuits are successful, even going to such lengths as to give up her personal time to focus on my work. This dissertation research is a testament to her efforts and encouragement. I also wish to thank my committee members: Peng Liu, Karin Dorman, Xiaoqiu Huang, and Nicole Valenzuela. Their helpful insights and suggestions were vital to the success of my research.

I want to thank the Toth Lab family with whom I’ve formed a strong relationship that extends beyond the lab and field. Every time that I have popped into his office unannounced, Adam Dolezal has allowed me to discuss problems/challenges or exploit his generosity to get access to supercomputing time and free software trials. In addition to being one of my closest friends, Alex Walton has graciously tolerated the numerous times that I’ve borrowed from his library, most of the times without asking. As one of the early lab members, Amy Geffre has helped me in more ways than I can describe. However, I do need to acknowledge her assistance with learning lab techniques, ordering supplies, and wasp collections. Jenny Jandt always provided phenomenal feedback on manuscripts and presentation while making sure the lab had delicious cupcakes for every meeting. A-Team + Jenny, your enthusiasm for social insects has been infectious, and I can’t imagine having a better lab to work with.

This work could not have been completed without the help of some amazing undergraduate and research assistants in the Toth lab. I would like to thank Cecil Mercado for her assistance with brain dissections. I must acknowledge the dedicated efforts of Eric McCall, Griffin Smith, and Vera Zhao for the many hours that they spent observing wasp behaviors, an unenviable task
that they undertook with enthusiasm. Vera Zhao also helped with wasp rearing and behavioral trials.

I have greatly enjoyed my time at Iowa State thanks to the wonderful BCB and EEOB communities. My tenure at Iowa State would not have gone so smoothly without Trish Stauble who generously drops everything to help with paperwork, logistical questions, or finding funding for computers and conferences. There are a few members of the 2010 BCB cohort who have been exceptionally supportive: Daniel Standage has kindly shared his data with me, Divita Mathur graciously offered advice on projects and grad school life, and Srihari Radhakrishnan has been an amazing partner-in-learning (starting with transcriptome assemblies and 211L TAing). Everyone in the EEOB department goes above and beyond to help in any way they can, and I am indebted to many members of the department. I would especially like to recognize a few of them. Lauren Sullivan has been a terrific mentor with whom I've had many insightful conversations about being a woman scientist. I need to thank Joe Gallagher for being a great friend and someone that I could always turn to for help with computational questions.

My research was funded by the National Science Foundation: IOS-1146410 (to A.L. Toth and J.H. Hunt) and a Doctoral Dissertation Improvement Grant - IOS-1311512 (to A.L. Toth and A.J. Berens). Additional funding was provided by the ISU Ecology, Evolution, and Organismal Biology Department through a graduate student research award.

I would not have been able to accomplish this goal without the unwavering love and support of my family. During this journey, they have graciously sat through practice presentations, discussed wasps and my research, and helped with wasp hunts. Every day I am grateful for the strong work ethic that they instilled in me, which I will draw upon throughout my career as a scientist.
Sociogenomics is an emerging field that employs genomics to address the evolution and underlying mechanisms regulating social living. Studies on highly social (often called eusocial) insects have contributed to advances in sociogenomics. Eusocial insects exhibit extreme cooperation characterized by a division of labor with reproductive queen and non-reproductive worker castes. Eusociality has evolved multiple times within the insect order Hymenoptera (ants, bees, and wasps), and my work represents a major leap forward in the development of genomic resources for wasps, with a focus on the model genus *Polistes*. It has been proposed that there are conserved genetic mechanisms underlying the evolution of eusociality ("genetic toolkit hypothesis"), but this hypothesis has yet to be tested using comparative studies across all three major social lineages. I investigated genetic mechanisms underpinning caste development across three lineages (honey bees, fire ants, and paper wasps) using comparative transcriptomics, and I found greater evidence for a relatively "loose genetic toolkit" of shared molecular pathways and biological functions, rather than specific genes. To provide further insights into how castes evolve, there is also a need for mechanistic studies of caste determination in primitively eusocial insects like *Polistes*, which have morphologically similar behavioral castes. Therefore, I tested nourishment as an environmental factor driving developmental plasticity of paper wasp caste and observed that nourishment manipulation biased gene expression toward worker-like expression patterns. Additionally, new genetically and experimentally tractable model systems for sociogenomics are vital for studying complex social behavioral traits. *Polistes* are poised to be an excellent model for one such trait - facial recognition - since a species of wasps (*P. fuscatus*) is the only insect known to possess this ability. By comparing brain gene expression during staged social interactions for two paper wasp species with (*P. fuscatus*) and without (*P. metricus*) facial recognition, I found that differences in calcium signaling and dominance-related genes may be important for individual memory recall. Memory recall is not necessarily limited to facial
features; so to home in on face-specific recognition, I conducted the first transcriptome-wide analysis of facial learning in any animal. I identified a few genes associated with face learning in *P. fuscatus* that are also known to function in olfactory recognition, which warrants further investigation. My work furthers sociogenomics by illuminating evolutionary and mechanistic underpinnings of caste development and social behaviors in *Polistes*, an important ecological and evolutionary model system.
CHAPTER 1. INTRODUCTION

1.1 Advances in Genomics

Over the past decade, advances in sequencing technology have led to the rapid expansion of the field of genomics, the study of the underlying structure, function, and evolution of genomes to aid in understanding phenotypes. For example, from May 2005 to 2015, there was a 1750% increase in the number of organisms (to over 54,000 organisms) with nucleotide and/or protein sequences available in the open-access database RefSeq (NCBI, 2015, 2005). In order to handle this influx of data, genomics has relied upon advances in the field of bioinformatics for the development of new statistical methods and software tools for understanding these biological data.

One of the fastest growing areas of genomics is transcriptomics, a high-throughput approach for determining global gene-expression patterns. One form of transcriptomics, RNA-Sequencing (RNA-Seq), is especially appealing for the study of non-model organisms due to relatively low cost without the need for reference sequences (Wang et al., 2009). However, the application of transcriptomics to non-model organisms is not without its share of bioinformatics challenges (reviewed in Stapley et al. (2010)). The combination of high sequencing error rates, short read length, and expression variability across genes and tissue types makes it difficult to assemble a complete transcriptome without a reference genome (Garber et al., 2011). Additionally for non-model organisms, gene annotation and functional characterization of assembled transcriptomes is limited to what is already known from model-organisms (Stapley et al., 2010). Despite these challenges, transcriptomics remains a powerful approach for understanding the environmental responsiveness of the genome, even for non-model organisms.
Behavior is especially environmentally responsive; thus, transcriptomics can be a valuable tool for studying behavioral ecology. Behavioral ecology aims to understand the ways that ecological conditions act on a genome to shape the evolution of behavioral phenotypes (Krebs and Davies, 1984). Previously, the molecular underpinnings of behaviors were studied exclusively outside of an ecological context, for example, by screening for mutations that were induced by behavioral abnormalities (Benzer, 1967) or artificial selection of populations for behavioral variants (Erlenmeyer-Kimli.L et al., 1962). With the advances of sequencing technology, it is now possible to study natural variation in behavior within and between species in an ecological context.

### 1.2 Sociogenomics

Sociogenomics is an integrative approach for studying the mechanistic and evolutionary analysis of social behavior which combines the fields of behavioral ecology, molecular biology, genomics, neuroscience, and evolutionary biology (Robinson et al., 2005). Sociogenomics aims to elucidate the molecular basis of social living by testing hypotheses about the evolution and regulation of this lifestyle. Furthermore, this field addresses how sociality influences genes and pathways that regulate development, physiology, and behavior. For example, it has been proposed that the demands of social living, such as flexible social networks, act as selective forces for the evolution of novel cognitive traits (Harpur et al., 2014; Johnson and Tsutsui, 2011).

Despite recent advances in sociogenomics, there are still many challenges to understanding how the genome and its genes influence social behavior and evolution. One such challenge is that behaviors are dynamic – fluctuating in response to (sometimes short-lived) environmental or genetic changes. To understand environmental responsiveness of behaviors, researchers concentrate on the dynamic nature of gene expression. For example, a behavioral change may result from adjusting brain gene expression in response to specific social stimuli. Identifying the link between genotype and phenotype may be especially difficult for behavioral traits because multiple genes often interact to produce complex behaviors (Mackay, 2009). Thus, it is important to be able to study multiple genes simultaneously, so transcriptome-wide approaches
are advantageous for determining the molecular mechanisms underpinning social behaviors. A single gene may control multiple behaviors, i.e. pleiotropy, which may obscure relationships between genotypes and phenotypes. Finally, to truly understand how solitary evolves into social behavior more genomic resources, including a wider taxonomic breadth of species, are necessary (Robinson et al., 2005). Moreover, the classic genetic models (e.g. Drosophila, C. elegans, and mice) are not necessarily the most informative for understanding sociality.

1.3 Social Insects

One of the most important model groups of organisms for both behavioral ecology and genomics are social insects (Zuk and Balenger, 2014). Social insects have been of interest for both behavior and adaptation studies since Darwin recognized their altruism as a challenge for natural selection (1859). Defining traits of eusociality are cooperative brood care, overlapping generations within a colony, and reproductive division of labor (Wilson, 1971). The extreme cooperative living by social insects requires coordination and communication between individuals to organize colony activities (Heinze, 2004; Ratnieks and Reeve, 1992). Thus, their social living may be organized by division of labor and/or dominance hierarchies, which are linked to nutritional resource distribution and reproduction (Jandt et al., 2014). There has been much interest in the genetic control of division of labor, development of worker behavior, and communication between individuals within social insect colonies (reviewed in Rittschof and Robinson (2014)).

There are some doubts that the novelty of sequencing technology is driving ecological genomic studies instead of a more hypothesis driven approach ((Travisano and Shaw, 2013), reviewed in Zuk and Balenger (2014)). Although high-throughput sequencing can illuminate more details about the underlying molecular processes, there is now a bottleneck in new testable hypotheses due in part to the rapid change from too few to an overabundance of data (reviewed in Zuk and Balenger (2014)). Despite these trends, both initial exploratory and hypothesis-driven studies utilizing sequencing technology could potentially advance our knowledge about sociogenomics by identifying functionally important suites of genes for the evolution and regulation of social behavior.
Toth et al. (2007) provides an illustrative example of hypothesis-driven sociogenomics research, where empirical data was used to test a long-standing hypothesis about the evolution of sociality in insects (West-Eberhard, 1989, 1996). Social insects exhibit an extreme form of cooperation where siblings stay on their natal nest to care for their sisters. It has been proposed that sibling care evolved from maternal care and that these two behaviors are regulated by the same gene expression patterns (Linksvayer and Wade, 2005). To test this hypothesis, Toth et al. (2007) examined brain gene expression patterns of *Polistes metricus* paper wasps, where both maternal and sibling care are exhibited within this single species, using next-generation sequencing and observed that individuals who performed brood care (either maternal or sibling care) showed more similar gene expression patterns than other groups of females on the nest. These results advanced the field of sociogenomics by providing empirical data that suggests modulation of gene expression is important for the evolution of novel social behavioral traits.

### 1.4 *Polistes* Paper Wasp Model System

As more genomes and transcriptomes are sequenced and annotated, the relative role of conserved and lineage-specific genes during the evolution of sociality will become clearer, and I provide an important new contribution to this question in Chapter 2. Advances to the field of sociogenomics have also been made by addressing the molecular basis of social evolution and mechanisms underlying social behaviors using *Polistes* paper wasp as a model system. For example, *Polistes* have been used to test the hypothesis that genes important for the evolution of caste systems are conserved across levels and origins of sociality (Toth and Robinson, 2007), and several studies have lent support to this supposition (Hunt et al., 2010; Sumner et al., 2006; Toth et al., 2007). However, there is also evidence in *Polistes* that the expression of lineage-specific (or “novel”) genes is important for the evolution of the worker phenotype, likely due to the fact that workers have greater behavioral complexity (Ferreira et al., 2013).

In paper wasps, individual reproductive fitness and division of labor is controlled by a social dominance hierarchy within the colony (Jandt et al., 2014; Pard, 1948). One form of division of labor between paper wasp workers is foraging behavior, which is regulated by nutritional signaling genes coopted from solitary organisms (Daugherty et al., 2011) and shared across
independent lineages of social insects (Toth et al., 2010). Dominance hierarchies are established through competitive interactions between individuals (Jandt et al., 2014; Pardi, 1948), and it has been shown that genes associated with this dominance behavior are also important for aggressive behavior in honey bees, solitary insects, and mammals (Toth et al., 2014). To avoid costly aggressive interactions and to promote cooperation, some species of paper wasps have evolved the ability to recognize conspecifics and/or their associated rank using variable and distinct facial patterns (Sheehan and Tibbetts, 2010, 2011; Tibbetts, 2002; Tibbetts and Dale, 2004). In most paper wasp species there is little to no variation in facial color and individual recognition abilities are absent (Sheehan and Tibbetts, 2010; Tibbetts, 2004); however at least five times within the Polistes lineage, facial color patterns have independently evolved (Tibbetts, 2004). For species with facial color patterns, these markings may vary with reproductive quality to indicate social status (Tibbetts and Dale, 2004), and for at least one species (P. fuscatus), these variable markings are used for individual recognition (Tibbetts, 2002). In Chapters 4 and 5, I provide the first data on the molecular underpinnings of individual recognition in paper wasps.

As illustrated by the Toth et al. (2007) study, Polistes paper wasps are an excellent system for studying sociogenomics, as they have interesting and well-characterized social behaviors (Gamboa et al., 1986; Pardi, 1948; Pratte and Jeanne, 1984; Tibbetts, 2002; West-Eberhard, 1969), and there are new genomic resources for several species (Berens et al., 2015a; Ferreira et al., 2013; Toth et al., 2007). Paper wasps have plastic female behavioral castes, i.e. genotypic differences do not account for differences between reproductive (queen) and non-reproductive (worker) castes, and these castes remain flexible into adulthood (reviewed in Reeve (1991)). Despite lifelong behavioral plasticity, castes are biased during development due to environmental conditions such as nutrition (Hunt, 2007) and social environment (Jeanne and Suryanarayanan, 2011). The quality of larval nourishment changes throughout the colony cycle according to seasonal changes in the adult-to-offspring ratio, so it has been suggested that nourishment inequalities bias caste development with workers receiving limited nourishment compared to queens (Hunt and Amdam, 2005; Hunt, 2007). Hunt et al. (2010) found that nutrition- and diapause-related genes are important for caste differences, while experimental studies show that
developmental nourishment inequalities produce physiological changes associated with behavioral castes (Judd et al., 2015; Karsai and Hunt, 2002). In Chapter 3, I provide new experimental evidence linking genome-wide caste-related expression changes with nourishment inequalities.

1.5 Dissertation Organization

In this dissertation, I use a sociogenomics framework to test both mechanistic and evolutionary hypotheses related to the social behaviors of Polistes paper wasps. In Chapter 2, I use a comparative transcriptomics approach to investigate the genetic mechanisms underlying the convergent evolution of reproductive castes across three major social insect lineages (ants, bees, and wasps). Moreover, I test two alternative, but not necessarily mutually exclusive, hypotheses that common genetic mechanisms underpin the evolution of caste systems across independent lineages (“genetic toolkit hypothesis”) and lineage-specific genes are important for the evolution of castes in paper wasps (“novel genes hypothesis”). This study entails the most comprehensive examination of caste development across social lineages to date. As part of this study, I develop a new transcriptomic resource for a species of paper wasps, P. metricus, to characterize developmental gene expression differences between behavioral castes in wasps. These observed paper wasp caste-related gene expression patterns laid the groundwork for mechanistic questions asked in Chapter 3.

In Chapter 3 of this dissertation, I examined the role of nourishment as an environmental mechanism driving the plasticity of Polistes behavioral castes. Building off of previous evidence that suggests nourishment inequalities can bias caste development with workers receiving limited nourishment compared to queens (Hunt et al., 2007, 2010), I compare gene expression changes in response to nourishment deprivation in P. metricus larvae with caste-related expression observed in Chapter 2. This work adds to our knowledge about how environmental conditions can lead to differences in complex behavioral traits at the molecular level.

In Chapter 4, I examine the genetic underpinnings of individual recognition in paper wasps using a comparative candidate gene approach. I test the hypothesis that genes associated with learning and dominance have been co-opted to function in individual recognition by comparing brain gene expression during staged social interactions for two sister species where one has
individual recognition (\emph{P. fuscatus}) and the other lacks the ability (\emph{P. metricus}). I then go beyond gene expression – behavior correlations and use pharmacology to manipulate individual recognition behavior by potentially disrupting candidate molecular processes. This is the first work to identify promising candidate genes and molecular processes for individual recognition in any insect.

In Chapter 5, I conduct a genome-wide comparison of gene expression differences between general pattern and facial training in two paper wasp species (\emph{P. fuscatus} and \emph{P. metricus}) that have differing facial learning abilities. Additionally, I tested hypotheses related to the evolution of facial recognition by identifying single nucleotide polymorphisms with this learning ability. This chapter addresses both the mechanistic and evolutionary underpinnings of a highly specialized social behavior.

In summary, this dissertation offers significant advances to sociogenomics by addressing some of the major outstanding questions in the field. How do social behaviors evolve? Are the molecular mechanism associated with social behaviors coopted from other solitary behaviors? Are the same molecular mechanism used in multiple origins of social behaviors? Or are the mechanisms specific to each lineage? Building off of this work, in the next few years the field of sociogenomics will move closer to understanding the molecular basis underlying the origins and maintenance of social living by incorporating broader taxonomic coverage that includes more eusocial lineages with different levels of social complexity.
CHAPTER 2. COMPARATIVE TRANSCRIPTOMICS OF CONVERGENT EVOLUTION: DIFFERENT GENES BUT CONSERVED PATHWAYS UNDERLIE CASTE PHENOTYPES ACROSS LINEAGES OF EUSOCIAL INSECTS

Modified from a paper published in *Molecular Biology and Evolution*

Ali J. Berens, James H. Hunt, Amy L. Toth

Abstract

An area of great interest in evolutionary genomics is whether convergently evolved traits are the result of convergent molecular mechanisms. The presence of queen and worker castes in insect societies is a spectacular example of convergent evolution and phenotypic plasticity. Multiple insect lineages have evolved environmentally induced alternative castes. Given multiple origins of eusociality in Hymenoptera (bees, ants, and wasps), it has been proposed that insect castes evolved from common genetic “toolkits” consisting of deeply conserved genes. Here, we combine data from previously published studies on fire ants and honey bees with new data for *Polistes metricus* paper wasps to assess the toolkit idea by presenting the first comparative transcriptome-wide analysis of caste determination among three major hymenopteran social lineages. Overall, we found few shared caste differentially expressed transcripts across the three social lineages. However, there is substantially more overlap at the levels of pathways and biological functions. Thus, there are shared elements but not on the level of specific genes. Instead, the toolkit appears to be relatively “loose,” that is, different lineages show convergent molecular evolution involving similar metabolic pathways and molecular functions but not the exact same
genes. Additionally, our paper wasp data do not support a complementary hypothesis that “novel” taxonomically restricted genes are related to caste differences.

### 2.1 Introduction

Convergent evolution of similar phenotypes in distant lineages has historically been viewed as the result of divergent, nonhomologous underlying mechanisms (Arendt and Reznick, 2008; Scotland, 2011). However, some studies of convergent traits such as echolocation in bats and dolphins (Parker et al., 2013) and camera-like eyes in octopi and humans (Ogura et al., 2004) provide evidence that some of the same genes can be recruited multiple times in evolution to regulate the development of convergent traits (Elmer et al., 2010; Scotland, 2011; Stern, 2013). Other studies point to different genetic mechanisms for convergent traits, but importantly, these cases may still involve genes with overlapping molecular functions that can influence the same biochemical pathways (e.g., pigmentation, reviewed in Kronforst et al. (2012); wing loss in ants (Abouheif and Wray, 2002)). These prior studies have focused on fixed phenotypes that characterize a particular species or population. However, many complex traits are plastic, highly responsive to the environment, and the result of the combined expression of many thousands of interacting genes. Are there also convergent molecular mechanisms for plastic traits that result from environmentally responsive gene expression?

Social insects are spectacular examples of both convergent phenotypic evolution and phenotypic plasticity. Across the insects, there have been multiple origins of alternate phenotypes (such as reproductive queen and nonreproductive worker female castes) that are independent of genotype. Environmental factors such as nutritional inequalities during development are associated with caste determination in most social insect species (reviewed in Smith et al. (2008)). Sociality evolved independently in insects multiple times (at least 11 times within the order Hymenoptera (Brady et al., 2006; Cameron and Mardulyn, 2001; Cardinal et al., 2010; Hines et al., 2007; Johnson et al., 2013; Pilgrim et al., 2008)), with convergent evolution of nutritionally dependent caste polyphenisms even in taxa that show large differences in their life history and social biology. For example, advanced social insect societies such as honeybees and fire ants are defined by morphologically distinct queen and worker castes that are established early dur-
ing larval development (Haydak (1970) [Apis mellifera—honey bee]; Petralia and Vinson (1979) [Solenopsis invicta—fire ant]). In such species, larval food provisioning affects a developmental switch to produce striking alternative adult forms differing greatly in morphology, physiology, and behavior ((Corona et al., 1999; Evans and Wheeler, 1999; Haydak, 1970; Hepperle and Hartfelder, 2001; Patel et al., 1960) [A. mellifera—honey bee]; (Linksvayer et al., 2006; Wheeler, 1986); reviewed in Anderson et al. (2008) [Ants]). In primitively social insect societies, such as paper wasps in the genus Polistes, there are no discrete morphological differences between castes, and caste-related behaviors are flexible even in adults (Reeve, 1991). Nonetheless, even in Polistes, castes are biased during early development, such that poorly nourished first generation females are worker destined, whereas well-nourished second generation females are queen destined (Hunt and Dove, 2002; Hunt and Amdam, 2005; Karsai and Hunt, 2002). Thus, there are common nutritional environmental determinants of caste in all three major eusocial lineages.

Examinations of insect social evolution from an evolutionary developmental (evo-devo) perspective led to the proposal that castes are derived from new arrangements of solitary behavioral and physiological modules (Amdam et al., 2004; Hunt and Amdam, 2005; West-Eberhard, 1987, 1996) and, along with these, deeply conserved genetic modules (reviewed in (Page and Amdam, 2007; Toth and Robinson, 2007)). A general hypothesis for convergent evolution of eusociality is that there is a shared “toolkit” of molecular and physiological processes across several independently evolved social insect lineages (Toth and Robinson, 2007; Toth et al., 2010). Because queen and worker castes can be produced from the same genome, the genetic toolkit underlying convergent social caste phenotypes depends on the differential expression (DE) of common genes and/or pathways.

In the study of eusociality, a few previous studies on bees and wasps suggested some overlap in gene expression patterns related to social behavior across lineages. Using relatively small sets of a few dozen candidate genes, Toth et al. (2007) and Daugherty et al. (2011) found between 25% and 60% of genes showed similar expression patterns for aspects of division of labor in P. metricus and honey bees. Further comparative analysis using microarrays between honey bees and P. metricus wasps highlighted commonalities in regulation of foraging/provisioning behavior (Toth et al., 2010) and aggressive behavior (Toth et al., 2014) at the molecular level.
However, the extent of overlap across species in these studies was relatively small, and an additional analysis suggested that reproductive behaviors are associated with nonshared gene expression patterns (Toth et al., 2010). One possible explanation for the relatively small overlap is that these studies focused only on individual genes. Instead, general changes in the modulation of consequential pathways and key biological functions may be more important (Toth and Robinson, 2007). For example, because queens are nearly always larger and better nourished than workers (Wilson, 1971), we would expect important changes in the regulation of metabolic and nutrient signaling pathways (reviewed in Smith et al. (2008)), but the specific genes involved and the direction of their expression may be more evolutionarily labile. Additional research on a genomic scale that includes multiple origins of sociality is needed to advance our understanding of the possible role of shared genes, pathways, and gene networks in social evolution.

In this study, our primary aim was to probe the molecular basis of the convergent evolution of castes by conducting the most comprehensive comparative analysis to date of caste-related global gene expression (Table 2.1). Our study utilizes data from three independent social lineages, represented by a honey bee (A. mellifera), a fire ant (S. invicta), and a temperate paper wasp (Polistes metricus). This analysis was made possible by the release of comparable caste-related developmental transcriptomic resources for these species (Chen et al. (2012b) [A. mellifera]; Ometto et al. (2011) [S. invicta]), including a new P. metricus transcriptome (this article). Alternative queen and worker caste phenotypes in these three species are strongly affected by differential nutrition (Haydak, 1970; Karsai and Hunt, 2002; Wheeler, 1986). Thus, we hypothesized there would be evolutionary convergence across the three lineages in the expression of specific genes, pathways, and/or gene networks associated with caste determination, especially those related to nutrition and metabolism. Because of the fact that the lineages are distant (separated by over 100 My) and caste phenotypes are complex and plastic, we predicted there would be less overlap at the level of specific genes, and more overlap on the levels of pathways and biological functions.

Second, we also use the new P. metricus transcriptome data to address an emerging complementary hypothesis that posits that “novel” genes are important for the evolution of novel caste phenotypes (Table 2.1; (Harpur et al., 2014; Johnson and Tsutsui, 2011)). Novel genes are de-
fined as previously undescribed genes that have no significant homology with known sequences (Ding et al., 2012). Genes that are novel (or “taxonomically restricted”) to Hymenoptera, bees and the genus Apis, are more likely to be overexpressed in honey bee workers (Johnson and Tsutsui, 2011) and show evidence of positive selection (Harpur et al., 2014), suggesting novel genes may play a role in causing caste differences. Similarly, in the tropical paper wasp P. canadensis, there is an overabundance of novel transcripts that show caste DE (Ferreira et al., 2013). However, comparisons of ant genomes suggest an abundance of species-specific novel genes may relate more to derived, species-specific traits possessed by workers, rather than castes differences per se (Simola et al., 2013). Thus, it is important to reassess the hypothesis that novel genes relate to caste differences by examining the association between novel gene expression and caste differences using pairs of more closely related species. In this study, we combined data from the two Polistes species to identify a set of well-supported Polistes-specific transcripts. If novel genes are relevant to the evolution of sociality in Polistes, we predicted Polistes-specific transcripts would be more likely to be caste-differentially expressed in both P. canadensis and P. metricus. The reason for this is that castes evolved only once in the common ancestor of both species, and thus any novel transcripts that contributed to caste evolution in the Polistes lineage should show common, caste-related gene expression patterns.

The objectives of this study were thus 2-fold: 1) investigate transcriptome-wide molecular mechanisms associated with convergent social caste phenotypes across three major hymenopteran social lineages (bees, ants, and wasps) and 2) address the contribution of Polistes-specific transcripts to caste differences in two species of Polistes wasps.

2.2 Results

Transcriptome Assembly and Annotation

We sequenced 16 transcriptomic libraries, representing four biological replicates of P. metricus fifth-instar larvae from two castes (field-collected queen- and worker destined) and two nutritional manipulation levels (laboratory-reared low and high nourishment groups; published separately (Berens et al., 2015b)). After de novo transcriptome assembly using Trinity (Grab-
herr et al., 2011) and filtering out chimeric and contaminant sequences using the mRNA-markup protocol (Brendel V, unpublished data; code available from http://brendelgroup.org/bioinformatics2go/mRNA markup.php, last accessed June 18, 2013) the final \textit{P. metricus} Transcriptome Shotgun Assembly (TSA) consisted of 74,516 transcripts with an N50 of 1,316 base pairs. To address quality and completeness, we assessed the \textit{P. metricus} transcriptome assembly using the Core Eukaryotic Gene Mapping Approach (CEGMA) method (Parra et al., 2007); 93.55\% of the Core Eukaryotic Genes (CEGs) mapped completely and 98.79\% of the CEGs mapped partially to the TSA, thereby indicating a very complete representation of expressed genes. Summary statistics of the TSA are shown in Table 2.2. In total, 16,662 (22.4\%) of the \textit{P. metricus} transcripts are putatively homologous to 13,429 unique sequences in National Center for Biotechnology Information (NCBI)’s nonredundant (NR) database. Ninety-six percent (15,997) of the best hits are from the order Hymenoptera (Figure 2.1 illustrates taxonomic grouping of the \textit{P. metricus} transcriptomic hits to the NR database). With BLAST2GO (Conesa et al., 2005), we made a putative functional annotation of the \textit{P. metricus} TSA based on the best Basic Local Alignment Search Tool (BLAST) hit to \textit{Drosophila melanogaster}. From 12,225 hits to \textit{Drosophila}, 9,736 \textit{P. metricus} transcripts were functionally annotated (Table 2.3 lists the \textit{P. metricus} transcriptome functional annotation and Gene Ontology [GO] enrichment statistics. Online Material—Pmet-tsa-r1.1 Annotation.xlsx, includes \textit{P. metricus} GO annotations).

**Polistes metricus** Caste DE

We identified 736 differentially expressed transcripts (DETs) between \textit{P. metricus} queen and worker-destined larvae using DESeq (Anders and Huber, 2010) (\(n = 4\) per caste, False Discovery Rate (FDR) \(\leq 0.05\) (Benjamini and Hochberg, 1995)). Surprisingly, of these \textit{P. metricus} caste DETs, 91.7\% were upregulated in worker-destined relative to queen-destined larvae (Figure 2.2 is a heat map of the scaled read counts for the \textit{P. metricus} caste DETs by sample; see Online Material—Pmet-tsa-r1.1_Caste.xlsx, for a list of \textit{P. metricus} caste DETs). Interestingly, there is one worker-destined sample that clustered with the queen-destined samples, which highlights the large biological variation that exists among individuals and is not totally surprising given the fact that some early season females can show queen-like phenotypes and even enter early
diapause (reviewed in Hunt (2007)). Strikingly, a previous candidate study (quantitative reverse transcription-polymerase chain reaction [PCR]) identified 16 caste DETs related to lipid metabolism, heat response, and stress response in *P. metricus* (Hunt et al., 2010) and all were downregulated in worker-destined larvae, which is in the opposite direction with respect to the majority of *P. metricus* DETs in this study. A comparison of our RNA-seq data to these published data confirmed the directionality of expression patterns (downregulation in worker-destined larvae, 20% of the significantly differentially expressed between castes for the RNA-seq data) was consistent for these 16 transcripts across the two studies. Appendix A: Figure A.1 is a heat map of the scaled read counts from this study for the 16 caste-related transcripts identified in Hunt et al. (2010). This result suggests a relatively small but important suite of genes including hexamerins, heat shock proteins, and insulin signaling (Hunt et al., 2010) are upregulated in queen-destined larvae, and these results also provide independent validation for our RNA-seq results.

**Comparative Caste Transcriptomics of Wasps, Ants, and Bees**

We compared the results of our *P. metricus* caste DE analysis to the most recent, comparable data sets from other social insects (Chen et al. (2012b) on the honey bee, *A. mellifera* and Ometto et al. (2011) on the fire ant, *S. invicta*). Comparisons were conducted at three levels: The transcript level, at the level of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto, 2000; Kanehisa et al., 2014), and GO categories (Gene Ontology Consortium 2000). The transcript level served as a proxy for genes; transcripts representing longest assembled, expressed sequences were used because many genes of interest did not have full gene sequences available for one or more species. KEGG pathways and GO categories served as rough proxies for genetic pathways and gene networks, respectively. KEGG pathways elucidate information about enzymes that are known to interact as part of well-conserved biochemical pathways, and GO categories provide information about potential broad functional changes (i.e., above the level of pathways) associated with differential gene expression.
DET Analysis

We compared the list of *P. metricus* caste DETs to lists of caste DETs from *A. mellifera* (Chen et al., 2012b) and *S. invicta* (Ometto et al., 2011). Overall, there are 15 shared orthologous DETs (of 2,475 orthologous sequences, E value < 1e-4) across these three social insect species. In pairwise comparisons of orthologous sequences between *A. mellifera* and *P. metricus* and between *A. mellifera* and *S. invicta*, there is an overrepresentation and underrepresentation of shared caste DETs, respectively (Fisher’s exact tests, P values < 0.05). Between *P. metricus* and *S. invicta* orthologous transcripts, the observed number of caste DETs does not deviate significantly from the number of expected DETs (Fisher’s exact test, see Figure 2.3A for the observed and expected number of shared caste DETs by species pairs). In addition to a relatively small overlap in caste DETs, directionality (upregulation in worker- or queen-destined samples) was not well-conserved across species (Figure 2.4A is a heat map of the caste DETs, colored by directionality, for the three species of social insects; Appendix A: Table A.1 summarizes directionality of DETs between pairs of species). In *Polistes* and *S. invicta*, the majority (91.7% and 69%, respectively) of the caste-biased transcripts were upregulated in worker-destined samples, whereas 67.5% of genes were upregulated in queen-destined larvae in *A. mellifera*. Additionally, correlations between expression levels of orthologous DETs across species were weak; log2 fold changes between *P. metricus* and *S. invicta* showed a very weak negative correlation and between *P. metricus* and *A. mellifera* a very weak positive correlation (Pearson’s correlation test, P value < 0.05; Appendix A: Table A.2 and Figure A.2).

KEGG Pathway Analysis

Using the lists of caste DETs, we identified KEGG pathways (Kanehisa and Goto, 2000) from each of the three focal species with large numbers of differentially expressed genes; such pathways were designated as “modulated.” We define a modulated metabolic pathway as a KEGG pathway where, across all enzymes within the pathway, the sum of the proportion of DETs per enzyme is ≥1 (determination of the modulated threshold is described in the Materials and Methods section below). Across all three species, there are eight common modulated KEGG
pathways (see Figure 2.4B for a Venn diagram of the number of modulated KEGG pathways between *A. mellifera*, *P. metricus*, and *S. invicta*, see Online Material—SpeciesComparison.xlsx, for the full list of modified KEGG pathways) including arginine/proline metabolism and glycolysis/gluconeogenesis (see Figure 2.5 for an abridged glycolysis/gluconeogenesis pathway, and Appendix A: Figure A.3 for an abridged arginine/proline metabolism pathway that highlights caste DETs for *A. mellifera*, *P. metricus*, and *S. invicta*). Fisher’s exact tests showed a significant overlap in the number of modulated KEGG pathways between *A. mellifera*–*P. metricus* and *P. metricus*–*S. invicta*. There is no deviation from the expected number of modulated KEGG pathways between *A. mellifera* and *S. invicta* (Figure 2.3B contains the observed and expected number of modulated KEGG pathways between pairs of species).

Although our study focuses on gene expression, previous work in bees identified signatures of selection at the sequence level (Woodard et al., 2011). Woodard et al. (2011) found that genes associated with carbohydrate metabolism, specifically within the glycolysis pathway, showed evidence of positive selection in multiple social lineages of bees. Additionally, they suggest that within the *Apis* lineage, there is an abundance of sequence changes within the glycolysis pathway. Our results suggest evolutionary changes in gene regulation may have also occurred within the glycolysis pathway, because we identified a very large number of differentially expressed enzymes within this pathway in *A. mellifera* (see Figure 2.5 for an abridged glycolysis/gluconeogenesis metabolism pathway). In fact, two-thirds of the glycolysis/gluconeogenesis differentially expressed enzymes are in *A. mellifera* (but note that *A. mellifera* had many more DETs overall). In contrast to the large number of modifications within the *A. mellifera* glycolysis pathway, there are only two enzymes corresponding to *P. metricus* caste DETs—ec:1.2.1.3 (NAD+ dehydrogenase) and ec:1.1.1.2 (NADP+ dehydrogenase). For both *A. mellifera* and *P. metricus*, NAD+ dehydrogenase is upregulated in queen-destined samples, whereas NAD+ dehydrogenase is not differentially expressed in *S. invicta*. NADP+ dehydrogenase is upregulated in *P. metricus* worker-destined larvae, upregulated in *S. invicta* queen-destined samples and not present in *A. mellifera* samples. Again, this indicates that directionality is not conserved across taxa. From Woodard et al. (2011), GB15039 (ec:4.2.1.11, hydratase) was classified as having accelerated evolution within highly social lineages compared with primitively social and
nonsocial lineages. We found hydratase significantly upregulated in workers in highly social honey bees and fire ants but not differentially expressed in the primitively social paper wasp *P. metricus*.

**GO Enrichment Analysis**

For each social insect lineage, we functionally annotated the transcriptomes based on the best BLAST hits to *D. melanogaster* sequences and identified enriched GO categories (FDR \( \leq 0.05\); Online Material — *Pmet-tsa-r1.1_Annotation.xlsx* contains all and significantly enriched GO terms for *A. mellifera* and *S. invicta*. Online Material — *Pmet-tsa-r1.1_Annotation.xlsx* and *Pmet-tsa-r1.1_Caste.xlsx*, include the *P. metricus* functional annotation and significantly enriched GO terms). Pairwise comparisons of caste GO enrichment terms revealed a significant overrepresentation of shared terms for all species pairs (Fisher’s exact tests, P values \( \leq 0.05\); see Figure 2.3C for the pairwise observed and expected number of enriched GO terms). Five significantly enriched caste-biased GO terms (GO:0044424—intracellular part, GO:0005622—intracellular, GO:0009653—anatomical structure morphogenesis, GO:2001141—regulation of RNA biosynthetic process, and GO:0006355—regulation of transcription, DNA dependent) are shared in common between *A. mellifera, P. metricus*, and *S. invicta*, which suggests that similar gene networks may be affected across species (Figure 2.4C is a Venn diagram of the number of overlapping enriched GO categories).

**Polistes-specific Transcripts**

To investigate the role of taxonomically restricted transcripts in caste differences, we examined the expression patterns of transcripts that were specific to the *Polistes* lineage. First, we reanalyzed the raw data from Ferreira et al. (2013) using the same pipeline (Bowtie2-Express-DESeq) used on our own data to control for methodological differences in the identification of DETs. This resulted in the identification of 1,320 DETs (see Online Material — *SpeciesComparison.xlsx*) with our methods compared with 2,543 DETs from (Ferreira et al., 2013). With this more comparable set of DETs, we still identified an overabundance of *P. canadensis* novel transcripts represented within the list of caste DETs (Fisher’s exact test, odds ratio = 1.30, P value =
3.011e-6), as was observed in Ferreira et al. (2013). Next, we identified 57,854 \textit{P. metricus} novel transcripts (without a BLAST hit [E value > 1e-4] to the NCBI NR database), of which 354 were differentially expressed across castes (novel DETs). Forty-one (12\%) of the \textit{P. metricus} novel DETs are upregulated in queen-destined larvae, and the remaining 313 (88\%) \textit{P. metricus} novel DETs are upregulated in worker-destined larvae. These are similar to the proportions for all caste-biased transcripts in \textit{P. metricus} (8\% upregulated in queen destined [61 transcripts] and 92\% upregulated in worker destined [675 transcripts]). Overall, there is an underrepresentation of caste DETs within novel transcripts in \textit{P. metricus} (only 354 of the 57,854 novel transcripts; Fisher’s exact test, odds ratio = 0.26, P-value < 2e-16). This is in stark contrast to the reported overrepresentation of caste DETs in \textit{P. canadensis} novel transcripts (Ferreira et al., 2013) (see Figure 2.4D for a Venn diagram of the number of novel transcripts and DETs for \textit{P. canadensis} and \textit{P. metricus}). Comparing \textit{P. metricus} with \textit{P. canadensis} transcripts, we identified 14,125 shared \textit{Polistes} taxonomically restricted transcripts. Very few (only 2 or 0.01\%) of these showed DE across castes (in both \textit{P. metricus} and \textit{P. canadensis}), and this overlap was not statistically significant (Fisher’s exact test, odds ratio = 1.26, P-value = 0.67; Figure 2.4D).

2.3 Discussion

Comparative Transcriptomics of Convergent Evolution

The genetic toolkit hypothesis for the evolution of sociality predicts that there are common caste-related genes and genetic pathways shared across the order Hymenoptera (Table 2.1). In this analysis, we identified a very small number (15) of common caste DETs between representatives of three hymenopteran social lineages (fire ants, honey bees, and paper wasps). Furthermore, directionality of caste-related gene expression is not conserved across species. This suggests that the same genes are not necessarily involved in caste differences across different origins of eusociality.

Instead, our data suggest that there may be similar molecular functional changes leading to convergent caste phenotypes through modifications within common metabolic pathways.
and gene networks. We identified five common KEGG pathways, including arginine/proline metabolism and glycolysis/gluconeogenesis (Figure 2.5, Online Material—SpeciesComparison.xlsx, Appendix A: Figure A.3) that were modulated in all three species. Previously, Woodard et al. (2011) proposed a link between the glycolysis pathway and multiple origins of sociality within bees based on sequence changes across the *Apis* lineage. The results from our cross-species gene expression analysis suggest that the glycolysis pathway may play a role in caste determination across Hymenoptera.

Even beyond specific metabolic pathways, our data indicate that common molecular functions are associated with caste differences across species, suggesting there may be shared gene networks related to eusocial evolution in the three lineages. We identified a common enrichment of five GO functional categories related to transcription regulation and morphogenesis across fire ants, honey bees, and paper wasps (Figures 2.3B and 2.4B, Online Material - SpeciesComparison.xlsx). This multilevel analysis suggests that caste phenotypes are associated with evolutionary lability at the level of genes and similarity on the level of pathways and GO categories (Figure 2.6). Thus, the genetic toolkit appears to be rather “loose”; social insect caste systems have evolved to utilize metabolic pathways and biological functions rather than the same genes (Table 2.1).

How do these results inform our understanding of convergent evolution? Although debated in the literature (Arendt and Reznick, 2008), some authors (e.g., Scotland (2011)) suggest that focusing on underlying mechanisms can help distinguish between cases of clear convergent evolution (same phenotype, different mechanisms, and often in distant taxa) and cases of parallel evolution (same phenotype, same mechanisms, and often in more closely related taxa). Caste-containing societies of ants, bees, and wasps have convergently evolved queen and worker phenotypes from superficially similar yet distinctly different ancestral solitary states, stemming from various forms of maternal care behavior in each ancestral lineage (Schwarz et al., 2010; West-Eberhard, 1996). Queen and worker phenotypes are multicomponent states that represent a collection of distinct morphological, behavioral, and physiological traits (Amdam and Page, 2010), and they are plastic and highly polygenic (Grozinger et al., 2007). Therefore, our results on gene expression in social insect castes do not fit neatly into prior definitions of convergence
or parallelism that focus more on discrete, fixed traits (Scotland, 2011). However, our results are illuminating in that they suggest that the convergent evolution of complex traits is to some extent associated with repeatable molecular changes, at least on the level of pathways and biological functions.

It is important to note that the fire ant, honey bee, and paper wasp studies were published independently and not designed to be directly comparable, so there are a number of differences between these studies (Table 2.4) in tissue types (whole body vs. head), life stages (fifth-instar larvae vs. pupae), and technologies (microarray vs. RNA-sequencing—both SOLiD and Illumina platforms). The *S. invicta* microarray study may have identified fewer DETs because arrays are limited to a few thousand candidate genes, and in *A. mellifera*, the large number of DETs compared with the two other species may be due in part to a lack of biological replicates. Despite the technical differences (above), ecological differences (Oster and Wilson, 1978; Ross and Keller, 1995), and evolutionary distance (Johnson et al., 2013) between these species, it is noteworthy that we were still able to uncover significant overlaps from our meta-analysis of transcriptomic data. In the future, studies that use more directly comparable approaches (life stages, platforms and tissues) and additional species along the solitary to social spectrum could provide an even more complete assessment of the importance of conserved transcripts and pathways in social evolution.

Our RNA-seq results are robust; we were able to replicate expression patterns from previously published real-time quantitative PCR data on 16 genes associated with insulin signaling pathways and storage proteins (Hunt et al., 2010) (Appendix A: Figure A.1). This verification is an excellent test of the data, because all 16 of the previously identified transcripts have higher expression in queen-destined larvae, whereas the majority of DETs from this study show the opposite pattern. Thus, the specific expression of these 16 genes highlights the consistency between our RNA-Seq and previous quantitative PCR results, despite differences in technologies. Furthermore, these results highlight that certain key processes including those related to insulin signaling pathways and storage proteins are upregulated in queen-destined larvae (Hunt et al., 2003, 2007, 2011; Patel et al., 2007; Wheeler et al., 2006), whereas many other pathways and functions are downregulated.
Novel Genes

An emerging hypothesis for social evolution posits that DE of novel (previously undescribed) transcripts is important for the evolution of novel caste phenotypes (Ding et al., 2012; Ferreira et al., 2013; Harpur et al., 2014; Johnson and Tsutsui, 2011). As a corollary to this hypothesis, we predicted that species within the same social lineage should share common novel DETs. In our analysis of two paper wasps, *P. canadensis* and *metricus*, this prediction was not supported. Unlike a previously published report in which *P. canadensis* showed a significant bias toward novel transcripts being caste related, novel transcripts were actually significantly underrepresented among the caste-related transcripts of *P. metricus*. In addition, there were very few (only two) novel caste-related transcripts shared between the two *Polistes* species (Figure 2.4C). One possible explanation for the difference between *Polistes* species is that the two studies examined different life stages (larval for *P. metricus* compared with adult for *P. canadensis*). As it now stands, there are conflicting reports on the importance of novel genes in social evolution. Studies in honey bees paved the way for studies of novel genes, with data suggesting these genes are particularly important in worker behavior in this advanced eusocial species (Harpur et al., 2014; Johnson and Tsutsui, 2011). On the contrary, a recent report in advanced eusocial fire ants (*S. invicta*) found that colony social organization (queen founding strategy) does not appear to be associated with the expression of novel genes (Manfredini et al., 2013). This collection of results suggests there may be variation in the importance of novel genes for different types of behavior, in specific life stages, or at different levels of sociality. In addition, there are some challenging technical issues associated with defining and detecting novel genes, in particular because genomic resources and gene discovery tools are rapidly advancing and changing (Khalturin et al., 2009).

2.4 Conclusion

Our comparative transcriptomic analysis across eusocial lineages and levels of molecular analysis suggests significant overlap in the types of metabolic pathways and gene functions associated with the convergent evolution of castes (especially those related to carbohydrate and
amino acid metabolism, morphogenesis, oxidation–reduction, and transcriptional regulation). Individual transcripts show less conserved expression patterns and relationships to caste phenotypes across species. This suggests that convergent social behavioral phenotypes across lineages are the product of convergent evolution of molecular mechanisms at the level of metabolic pathways and gene networks, which can be modulated at many different places within such networks (i.e., transcript expression, Figure 2.6). Thus, although evolution may indeed be a “molecular tinkerer” (Jacob, 1977) because different genes can produce similar phenotypes, selection may be constrained to act within a limited set of pathways and gene networks.

2.5 Materials and Methods

We collected four biological replicates of *P. metricus* fifth-instar larvae for each of the following groups: Unmanipulated castes (queen- and worker destined) and two nutritional manipulation levels (low and high). The nutritional manipulation samples are only used for transcriptome assembly and published separately (Berens et al., 2015b). For the unmanipulated caste-destined larvae, we collected samples from eight naturally occurring *P. metricus* nests at the Iowa 4-H Center (Madrid, IA). On the basis of the colony life cycle of paper wasps (Hunt et al., 2011), we collected foundress-reared (worker-destined) and worker-reared (queen-destined) larvae late spring and late summer 2010, respectively. Both males and females may be produced during the late summer. We removed males from the experiment based on the patterns and coloration of the eighth and ninth abdominal segments on the thawed bodies (laboratory protocol based in part on Cotoneschi et al. (2007)). Larvae were flash frozen in liquid nitrogen and stored at -80 °C until RNA extraction.

**RNA Extraction and RNA Sequencing**

We extracted total RNA from individual larval heads using an RNeasy Kit (Qiagen). RNA was quality controlled using spectrophotometry (NanoDrop) and a Bioanalyzer (Agilent). The High-Throughput Sequencing and Genotyping Unit of the W.M. Keck Center (University of Illinois at Urbana-Champaign) prepared mRNA Seq libraries with Illumina’s “TruSeq RNAseq Sample Prep kit” and sequenced the libraries on a HiSeq 2000 (Illumina). This library prepara-
tion included poly(A) RNA purification, fragmentation using sonification, cDNA synthesis from
size selected fragments (270 nucleotides on average) using random primers, and bar coding for
each of the 16 samples. In a balanced incomplete block design (Auer and Doerge, 2010), four
lanes of the sequencer were used to generate over 1 billion 100 base paired-end reads with one
subject from each treatment in each lane (Appendix A: Table A.3 contains read counts for each
sample). Raw sequence data have been deposited to the NCBI's Short Read Archive (BioPro-
ject ID: PRJNA242774, accession numbers: SRX511425, SRX511426, SRX511427, SRX511430,
SRX511432, SRX511433, SRX511434, SRX511435).

Data Preprocessing

Visualization

We visualized the raw reads from each paired end file using FastQC (Andrews S, unpublished
data; code available from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/,
last accessed November 28, 2011) and SolexaQA (Cox et al., 2010) to determine data quality
and identify potential problems with the data. Visualization of all samples using FastQC and
SolexaQA suggested that the overall quality of the data is very high. However, there were bases
of lower quality, especially at the end of the reads; this is typical of Illumina data and does not
indicate a problem if addressed by quality filtering.

Adapter Removal

As part of the library preparation, the High-Throughput Sequencing and Genotyping Unit of
the W. M. Keck Center (University of Illinois at Urbana-Champaign) adhered adapter sequences
to each end (for pair-end sequencing) of the cDNA fragments. We removed the extraneous
sequence before transcriptome assembly using the fastx_clipper tool from the Fastx Toolkit
Version 0.0.13 (Hannon GJ, unpublished data; code available from http://hannonlab.cshl.
edu/fastx_toolkit/, last accessed Dec 20, 2011).
Quality Filtering

We filtered reads for quality (threshold ≥20) using the Trim Perl script (Joshi N, unpublished data; code available from [http://wiki.bioinformatics.ucdavis.edu/index.php/Trim.pl](http://wiki.bioinformatics.ucdavis.edu/index.php/Trim.pl), last accessed November 11, 2011) with a length threshold of 50 bases. After filtering the data, approximately 2% of the reads (600,000–925,000) were removed from each library.

Transcriptome Assembly and Annotation

We assembled the groomed transcriptomic short reads de novo using Trinity ([Grabherr et al., 2011](#)) (Version r2012-06-08). The final trinity-produced TSA was annotated with the mRNA-markup protocol ([Brendel V, unpublished data; code available from [http://brendelgroup.org/bioinformatics2go/mRNAmarkup.php](http://brendelgroup.org/bioinformatics2go/mRNAmarkup.php], last accessed June 18, 2013) mRNAmarkup splits potential chimeric assemblies and discards likely contaminants identified as sequences with strong similarities to UniVec ([http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html](http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html), last accessed June 18, 2013) entries or Escherichia coli genomic sequences. We identified potential full-length mRNAs within the TSA using MuSeqBox ([Xing and Brendel, 2001](#)) as distributed with mRNAmarkup (parameter option –F 5 10 10 90 60). The final TSA was annotated with the most significant BLASTp hit using stringent criteria (MuSeqBox option –A 45 10 75). With this final cleaned-up and annotated TSA assembly, we assessed quality and completeness using the CEGMA (Version 2.4.010312) method and identified putative homologs to NCBI NR databases using BLASTx with an E-value threshold of 1e-4. This TSA project has been deposited at DDBJ/EMBL/GenBank under the accession GBGV00000000. The version described in this article is the first version, GBGV01000000.

Read Mapping, Abundance Estimation, and DE

We aligned the raw paired-end reads to the TSA using Bowtie2 ([Langmead and Salzberg, 2012](#), Version 2.1.0) and quantified the abundances of the transcripts for each library with eXpress ([Roberts and Pachter, 2013](#), Version 1.3.1) (Appendix A: Table A.4 includes read alignment results to the *P. metricus* transcriptome by sample). The raw read counts have been
deposited in NCBI’s Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series accession number GSE61960. We used the R (Team (2013), Version 3.0.1) statistical packages DESeq (Anders and Huber (2010), Version 1.12.0) and EdgeR (Robinson et al. (2010), Version 3.2.3) from the Bioconductor repository Gentleman et al. (2004) to test for DE (FDR $\leq 0.05$, Benjamini and Hochberg (1995)) between castes. Dispersion was computed per-condition with DESeq. For EdgeR, a common dispersion was calculated across all samples, then it was adjusted for each condition. DESeq produced more robust results than EdgeR; thus subsequent analyses utilized the results from DESeq (Online Material—Pmet-tsa-r1.1_Caste.xlsx, contains output from DESeq including fold changes for all *P. metricus* transcripts and a list of *P. metricus* caste-related DETs).

Methods for Comparative Analyses

Choice of Studies for Comparative Analysis

After an extensive review of the literature (including related studies reviewed in Smith et al. (2008)), we focused our comparative analysis on studies that were the most directly comparable to the *P. metricus* data set by the following criteria: 1) the data sets compared late-stage subadult queens and workers (fifth-instar larvae or pupae), 2) the data sets were of sufficiently large scale, that is, they used a genome-wide approach such as microarrays or RNA-seq, and 3) they were the most recent studies on this topic. This led to the choice of Chen et al. (2012b), an RNA-seq study on queen and worker honey bee (*A. mellifera*) fifth-instar larvae (as opposed to an earlier study that used microarrays, Barchuk et al. (2007)), and Ometto et al. (2011), a microarray study on queen and worker fire ant (*S. invicta*) pupae. Numerous other studies have compared gene expression in queen and worker social insects (reviewed in Smith et al. (2008), see also Cameron et al. (2013); Feldmeyer et al. (2014); Weil et al. (2009)), but because of their limited scale (e.g., few differentially expressed genes were assayed or identified) or focus on completely different life stages and tissues, we deemed them suitable for comparative analysis with our data set.
Putative Orthologs

We defined putative orthologous sequences across social insect species as the best BLAST hit (E value ≤ 1e-4) between pairs of species. For this analysis, we used the *A. mellifera* Official Gene Set (OGS) v1.1 transcriptomic sequences and the *S. invicta* OGS v2.2.3 transcriptomic sequences (Appendix A: Table A.5 holds the number of putative homologs between pairs of species). Note that only the hits to the transcriptomic sequences on the *S. invicta* microarray were considered for this analysis.

Transcript DE Analysis

For the transcript DE analysis, we used previously published lists of DE loci between *A. mellifera* fifth-instar queen- and worker-destined larvae (RNA-Seq, single library per caste, P < 0.01, Chen et al. (2012b)) and DE oligos between *S. invicta* queen- and worker-destined pupa (microarray, pools from 6 colonies per caste, FDR ≤ 0.05, http://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE35217, last accessed January 9, 2013, Ometto et al. (2011)) (Appendix A: Table A.6 lists the number of putative homologous DETs between pairs of species). Of the 4,787 DE loci between *A. mellifera* fifth-instar queen- and worker-destined larvae, only 4,363 loci corresponded to *A. mellifera* transcriptomic sequences. The 2,415 DE oligos between *S. invicta* castes corresponded to 854 *S. invicta* transcripts. Table 2.4 provides information about the transcriptome-wide analyses on caste bias for the social insect species: *A. mellifera, P. metricus*, and *S. invicta*. We performed Fisher’s exact tests between pairs of species to determine whether there was an over- or underrepresentation of DETs compared with the background, that is, all shared DETs between all orthologous transcripts (Figure 2.3; Appendix A: Table A.1 summarizes the directionality of DETs between pairs of species; Online Material—SpeciesComparison.xlsx, contains a conversion table of DETs and a list of overlapping DETs between all three social insect species).
KEGG Pathway Analysis

For each species, we used BLAST2GO (Conesa et al., 2005) to annotate enzyme codes and metabolic (KEGG) pathways (Kanehisa and Goto, 2000; Kanehisa et al., 2014) for each transcriptomic sequence. We developed a novel method to focus on the most informative DET containing pathways. This was necessary because of the large number of *A. mellifera* DETs (likely including false positives), resulting in all *A. mellifera* KEGG pathways including at least one DET. With our novel method, we aimed to restrict the number of KEGG pathways to the most informative subset of pathways for which there was more evidence (multiple DETs). If an enzyme was represented by several transcripts, but only one was differentially expressed, it was given less weight (as the evidence for a change in the pathway was less well supported). Appendix A: Table A.7 shows that as intended, mainly poorly supported pathways with only a single DET were removed from the analysis.

To account for multiple transcript mappings to each enzyme, we calculated the proportion of DETs per enzyme. We considered the KEGG pathway to be “modulated” by the DETs if the sum of the proportions of DETs overall all enzymes was greater than a threshold. We investigated modulation thresholds ranging from 0.1 to 3 incremented by 0.1 (Appendix A: Figure A.4 is a plot of the number of KEGG pathways above the modulation threshold). All three species had a large decrease in the number of modulated KEGG pathways from thresholds 1 to 1.1. For more robust results, we set the threshold for modulated KEGG pathways to 1 (see Appendix A: Figure A.5 for a Venn diagram of the number of modulated KEGG pathways above threshold 1.1 between all three species). We performed Fisher’s exact tests between pairs of species to determine whether there was statistically significant number of common modulated KEGG pathways (Fisher’s exact test results summarized in Figure 2.3; see Online Material—SpeciesComparison.xlsx, for *A. mellifera*, *P. metricus*, and *S. invicta* KEGG pathway annotations, modulated KEGG pathways by species, and overlapping modulated KEGG pathways).
GO Enrichment Analysis

We used BLAST2GO (Conesa et al., 2005) for functional annotation of the transcriptomic sequences for each species based on best BLAST hits (E value ≤ 1e-3) to D. melanogaster sequences in NCBI Entrez Protein database. For each species, BLAST2GO was used to assess enrichment of GO terms (FDR ≤ 0.05; two tailed) between caste-biased transcripts and the remaining transcriptome (Online Material—Pmet-tsa-r1.1_Annotation.xlsx and Pmet-tsa-r1.1_Caste.xlsx, include P. metricus GO annotations and Enriched GO terms; Online Material—SpeciesComparison.xlsx, contains A. mellifera and S. invicta GO annotations, A. mellifera- and S. invicta-enriched GO terms, and list of overlapping enriched GO terms between A. mellifera, P. metricus, and S. invicta). We performed Fisher’s exact tests between pairs of species to determine whether there was statistically significant number of shared GO-enriched terms (Fisher’s exact test results summarized in Figure 2.3).

Polistes-specific Transcripts

Polistes metricus novel transcripts were identified as transcripts without a BLAST hit (E value > 1e-4) to the NCBI NR database. Polistes taxonomically restricted transcripts were defined as the best BLAST hit (E value ≤ 1e-4) between P. canadensis novel Isotig sequences (Ferreira et al. (2013), http://genome.crg.es/~pferreira/pcandata/pcan.htm, last accessed April 19, 2013) and P. metricus novel transcripts. We reanalyzed the raw data from Ferreira et al. (2013) using the same pipeline (Bowtie2-Express-DESeq) used for P. metricus to identify DETs between P. canadensis queen and worker adults. We used a Fisher’s exact test to determine whether there was an overrepresentation of DETs within the taxonomically restricted transcripts.

2.6 Acknowledgements

The authors thank Amy Geffre and Cecile Mercado for help with RNA extractions, and Daniel Standage and Volker Brendel for their assistance with the mRNA markup protocol. They thank members of the Toth laboratory, Christina Grozinger, and Gene Robinson for their
insightful comments during the preparation of this manuscript. This work was supported by the National Science Foundation IOS 1146410.
Table 2.1  The Two Major Hypotheses Being Tested in This Study Are Outlined, along with the Level at Which the Analysis Was Conducted (KEGG), the Possible Outcomes from Each Analysis, and the Inferred Relationship from each Possible Outcome.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>At the level of</th>
<th>Possible Outcomes</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic Toolkit</td>
<td>Transcript Expression</td>
<td>Common caste-associated gene expression</td>
<td>Convergent evolution of expression of specific genes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Different caste-associated gene expression</td>
<td>Non-convergent evolution of expression of specific genes</td>
</tr>
<tr>
<td>KEGG Pathways</td>
<td>Gene expression changes in common pathways</td>
<td>Convergent evolution of expression of genetic pathways</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gene expression changes in different pathways</td>
<td>Non-convergent evolution of expression of genetic pathways</td>
<td></td>
</tr>
<tr>
<td>GO Functions</td>
<td>Gene expression changes for common gene functions</td>
<td>Convergent evolution of expression of gene networks</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gene expression changes for different gene functions</td>
<td>Non-convergent evolution of expression of gene networks</td>
<td></td>
</tr>
<tr>
<td>Novel Gene</td>
<td>Novel (i.e. Polistes-specific transcripts) genes important for the evolution of caste system in Polistes lineage</td>
<td>Common caste differentially expressed novel transcripts</td>
<td>Novel genes important for evolution of caste phenotypes in Polistes lineage</td>
</tr>
<tr>
<td>Polistes-specific transcripts</td>
<td>Different caste differentially expressed novel transcripts</td>
<td>No evidence for novel genes in expression of caste phenotypes in Polistes lineage</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.2 *Polistes metricus* Transcriptome Assembly Statistics Assembled De Novo with Trinity and Annotated with the mRNAmarkup Protocol.

<table>
<thead>
<tr>
<th>Number of transcripts</th>
<th>Longest transcript</th>
<th>N50</th>
<th>N90</th>
<th>CEGMA partial</th>
<th>CEGMA complete</th>
</tr>
</thead>
<tbody>
<tr>
<td>74,516</td>
<td>23,236</td>
<td>1,316</td>
<td>279</td>
<td>98.79%</td>
<td>93.55%</td>
</tr>
</tbody>
</table>

Table 2.3 *Polistes metricus* Transcriptome Functional Annotation Statistics and GO Enrichment Statistics. Functional annotation was determined using BLAST2GO based on the best BLAST hit to *Drosophila melanogaster*. In parentheses, the proportion of sequences with BLAST results, mapping results, or annotated compared with the total number of transcripts or the proportion of caste enriched GO terms compared with the total GO terms, respectively.

<table>
<thead>
<tr>
<th>Blast</th>
<th>Mapping</th>
<th>Sequences with Annotation</th>
<th>Annotation and InterPro Scan</th>
<th>GO Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,225 (16.4%)</td>
<td>10,177 (13.7%)</td>
<td>9,532 (12.8%)</td>
<td>9,736 (13.1%)</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>49 (26.2%)</td>
</tr>
</tbody>
</table>

Figure 2.1 Taxonomic grouping of the 16,662 *Polistes metricus* transcripts with hits to the NCBI NR database (E value < 10e-4). Bar chart indicates hits to superfamilies in the order Hymenoptera. Listed next to each taxonomic class is the number of hits; 59 hits (not shown) did not have taxonomic information.
Table 2.4  Summary of Transcriptome-Wide Analyses on Caste Determination in Social Insects: *Apis mellifera*, *Polistes metricus*, and *Solenopsis invicta*.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of DETs between Castes</th>
<th>Life Stage</th>
<th>Tissue</th>
<th>Technology</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. mellifera</em> (honey bee)</td>
<td>4,363</td>
<td>Fifth-instar larvae</td>
<td>Whole body</td>
<td>RNA-Seq (SOLiD)</td>
<td>Chen et al. (2012b)</td>
</tr>
<tr>
<td><em>P. metricus</em> (temperate paper wasp)</td>
<td>736</td>
<td>Fifth-instar larvae</td>
<td>Heads</td>
<td>RNA-Seq (Illumina)</td>
<td>This study</td>
</tr>
<tr>
<td><em>S. invicta</em> (fire ant)</td>
<td>854</td>
<td>Pupae</td>
<td>Whole body</td>
<td>Microarray</td>
<td>Ometto et al. (2011)</td>
</tr>
</tbody>
</table>
Figure 2.2  Heat maps of the scaled sample read counts for the 736 DETs between castes. Groups: QD - queen-destined larvae and WD - worker-destined larvae.
Figure 2.3  Number of observed and expected (in parentheses): (A) Common caste DETs out of total number of shared transcripts, (B) common modulated KEGG pathways out of total number KEGG pathways, (C) common caste GO-enriched terms out of total number GO-enriched terms (FDR < 0.05; two tailed) between pairs of species. The color indicates either a statistically significant overrepresentation (blue) or a statistically significant underrepresentation (yellow) of common DETs, KEGG pathways, or GO-enriched terms (Fisher’s exact test, P value < 0.05).
Figure 2.4  (A) Heat map of the DETs for three species of social insects (Apis mellifera, Polistes metricus, and Solenopsis invicta). For each species, putative homologs of the DETs were determined for all other species by BLAST. White indicates that no putative homolog was found (E value > 1e-4). Based on the log fold change, each transcript was classified as upregulated in queen destined (green), upregulated in worker destined (red), or not differentially expressed between castes (Gray). (B) Venn diagram of the number of interrupted KEGG pathways (threshold 1.0) for A. mellifera, P. metricus, and S. invicta. (C) Venn diagram of the number of GO terms significantly enriched (FDR < 0.05; two-tailed) between differentially expressed caste transcripts and the remaining transcriptome for A. mellifera, P. metricus, and S. invicta. (D) Venn diagram of the number of novel transcripts (transcripts without a hit [E value > 1e-4] to the NCBI NR database) with the number of caste differentially expressed novel transcripts in parentheses for P. canadensis and P. metricus.
Figure 2.5  Abridged glycolysis/gluconeogenesis metabolism pathway (KEGG map:00010), which highlights caste DETs for *Apis mellifera*, *Polistes metricus*, and *Solenopsis invicta*. Based on the log fold change, each transcript was classified as upregulated in queen-destined samples (green, positive direction), upregulated in worker-destined samples (red, negative direction), or not differentially expressed (gray). For some enzymes, there were multiple transcripts that mapped to the enzyme. Thus, the relative proportion of DETs was calculated for each enzyme and displayed as the bar height. White indicates that no putative homolog was found (E value > 1e-4)
Figure 2.6 Conceptual visualization of multilevel effects of transcripts, metabolic pathways, and biological functions. (a) Transcripts were identified as being differentially expressed (filled rectangle) or not significantly different (outlined rectangle), for all putative orthologs (row). Species without a putative ortholog were left blank. (b) KEGG pathways contain multiple, interacting transcripts. Similar phenotypic effects may occur across the three species due to modulation of KEGG pathways at differing transcripts. Both transcripts and KEGG pathways can affect enrichment of (c) GO categories.
CHAPTER 3. NOURISHMENT LEVEL AFFECTS CASTE-RELATED GENE EXPRESSION IN POLISTES WASPS

Modified from a paper published in *BMC Genomics*

Ali J. Berens, James H. Hunt, Amy L. Toth

Abstract

Background

Social insects exhibit striking phenotypic plasticity in the form of distinct reproductive (queen) and non-reproductive (worker) castes, which are typically driven by differences in the environment during early development. Nutritional environment and nourishment during development has been shown to be broadly associated with caste determination across social insect taxa such as bees, wasps, and termites. In primitively social insects such as *Polistes* paper wasps, caste remains flexible throughout adulthood, but there is evidence that nourishment inequalities can bias caste development with workers receiving limited nourishment compared to queens. Dominance and vibrational signaling are behaviors that have also been linked to caste differences in paper wasps, suggesting that a combination of nourishment and social factors may drive caste determination. To better understand the molecular basis of nutritional effects on caste determination, we used RNA-sequencing to investigate the gene expression changes in response to proteinaceous nourishment deprivation in *Polistes metricus* larvae.

Results

We identified 285 nourishment-responsive transcripts, many of which are related to lipid metabolism and oxidation-reduction activity. Via comparisons to previously identified caste-
related genes, we found that nourishment restriction only partially biased wasp gene expression patterns toward worker caste-like traits, which supports the notion that nourishment, in conjunction with social environment, is a determinant of developmental caste bias. In addition, we conducted cross-species comparisons of nourishment-responsive genes, and uncovered largely lineage-specific gene expression changes, suggesting few shared nourishment-responsive genes across taxa.

Conclusion

Overall, the results from this study highlight the complex and multifactorial nature of environmental effects on the gene expression patterns underlying plastic phenotypes.

3.1 Background

Phenotypic plasticity provides an important adaptive mechanism by which morphology, physiology, and/or behavior can be adjusted to biotic and abiotic environmental factors including temperature, nutrition, population density, and predator presence (Whitman and Agrawal, 2009). There are a number of striking examples of phenotypic plasticity in insects: dimorphic horn development in dung beetles (Emlen, 1994), seasonal color polyphenism in butterflies (Shapiro, 1976), and wing polyphenism in aphids (reviewed in Ogawa and Miura (2014)). One of the best-studied models of insect phenotypic plasticity is reproductive castes in social insects, especially social Hymenoptera (bees, ants, and wasps (Wilson, 1971)). In most social insects, genotypic differences do not account for differences between reproductive (queen) and non-reproductive (worker) castes (reviewed in Smith et al. (2012)). Instead, environmental factors induce differences in hormone titers and gene expression (reviewed in Smith et al. (2012)), leading to the development of queens or workers, which vary in physiology and behavior and for advanced social species, in morphology (Wilson, 1971).

One particular environmental factor, food availability, is especially important for caste polyphenism in social insects. Differential nourishment (Hunt, 1991, 1994, 2007; Huber, 1821) and nutrition-related genes and pathways (e.g. storage proteins such as vitellogenin and hexamerin, insulin/insulin-like signaling (IIS) pathways) have been linked to caste differences in honey
bees (Patel et al., 2007; Wheeler et al., 2006), paper wasps (Berens et al., 2015a; Hunt et al., 2010), and termites (Scharf et al., 2005; Zhou et al., 2007). This suggests that the influence of nutrition on caste formation may be broadly shared across diverse taxa (Toth and Robinson, 2007). There have been numerous studies investigating the molecular mechanisms underlying queen-worker caste determination in advanced eusocial species, especially honey bees (Barchuk et al., 2007; Cameron et al., 2013; Chen et al., 2012b; Evans and Wheeler, 1999; Feldmeyer et al., 2014; Ometto et al., 2011; Severson et al., 1989; Weil et al., 2009). In honey bees, nutritional differences for larvae fed either royal jelly or worker jelly precede a developmental switch resulting in alternative caste phenotypes (Winston, 1987). Nutrition is also important for caste differences in primitively eusocial species, which lack morphological castes, but we know much less about the molecular mechanisms that underlie the formation of their more subtle behavioral and physiological castes. In primitively eusocial species such as paper wasps, differential nourishment does not strictly determine caste but can lead to a caste bias, whereby female larvae that are fed larger quantities of food are more likely to be future reproductive queens (called “gynes”) as adults (Hunt and Amdam, 2005). However, the ultimate caste fate of a female is decided during adulthood. First-brood offspring of an established nest are capable of independent reproduction, but instead they perform allomaternal care (worker behavior) as a response to cues emitted by larvae in the nest (Hunt, 2012). Subsequent social reinforcement of worker behavior often occurs via dominance behaviors by the queen or other workers (Pardi, 1948; Reeve, 1991).

Primitively eusocial taxa such as *Polistes* are an informative group for understanding the evolution of eusociality and the origins of castes (Jandt et al., 2014; Starks and Turillazzi, 2006). Across the annual colony cycle of primitively social wasps in the genus *Polistes*, the quantity of larval nourishment changes according to seasonal changes in the adult-to-larva ratio (Hunt and Amdam, 2005; Hunt, 2007). First-brood offspring produced early in the colony cycle have been reared by a single nest-founding queen or few queens, which also perform all colony tasks such as nest building, foraging, egg laying, and brood rearing (Reeve, 1991). At this early time in the colony cycle the adult-to-larva ratio is low (Reeve, 1991; West-Eberhard, 1969), which leads to low feeding rates and more limited larval nourishment compared to offspring reared by workers.
and produced later in the colony cycle (future queens or “gynes”). Physiological evidence of nourishment-related differences between workers and gynes collected from naturally-founded colonies in the field include greater fat body stores in gynes (Eickwort, 1969; Strassmann et al., 1984), greater quantities of the storage protein hexamerin 1 in gynes (Hunt et al., 2003, 2007), and greater quantities of four additional proteins in gynes (Hunt et al., 2010). Experimental studies show that nourishment inequalities that correspond to early-season and late-season larval development contexts are associated with development of offspring having characteristics of worker and gyne phenotypes, respectively (Judd et al., 2015; Karsai and Hunt, 2002).

In addition to the established role of nourishment in caste differences in Polistes, social factors, such as dominance behavior (Jandt et al., 2014; Pardi, 1948) and maternal influences (Jeanne and Suryanarayanan, 2011; Suryanarayanan et al., 2011a,b) also play a role. Jeanne and Suryanarayanan (2011) propose a hypothesis for caste determination in primitively social wasps that incorporates not only nourishment variability, but also social environmental inputs from maternal care, specifically vibrational signals called antennal drumming. Antennal drumming may be an example of a maternal manipulation (Alexander, 1974) that directs larvae toward a worker developmental trajectory, and this effect may interact with nourishment induced changes in caste phenotype. Thus, nourishment is likely to act in concert with social environmental factors in determining differences in gyne and worker caste development in Polistes.

In this study, we investigated the effect of experimental proteinaceous nourishment deprivation during larval development on caste-related gene expression in a primitively eusocial species, the paper wasp Polistes metricus. We had three main goals. First, we explored the transcriptional responses of wasp larvae to high and low nourishment levels during laboratory rearing using RNA-sequencing. Second, we tested the hypothesis that nourishment level relates to caste-related gene expression; specifically that low nourishment is associated with more worker-like gene expression patterns, and high nourishment is associated with more gyne-like expression patterns. To do this, we compared nourishment differential expression to a set of nearly 800 previously identified genes associated with caste development in field-reared P. metricus (Berens et al., 2015a), and we did so on multiple levels: individual transcripts, pathways and biological functions. Third, we tested whether the molecular mechanisms underlying the response
to nourishment are conserved across taxa by comparing our results to two other nourishment deprivation studies in fruit flies (Teleman et al., 2008) and dung beetles (Kijimoto et al., 2014). Our overall goal was to better understand the extent and nature of the role of nourishment and nourishment-related genes in caste development in a primitively social wasp species.

### 3.2 Results

**Differential Expression Analysis**

To assess differential gene expression, we mapped reads to a previously assembled de novo transcriptome for *P. metricus*; the transcriptome was based on both the sequence data described here in conjunction with additional samples described in a previous study (Berens et al., 2015a). By comparing expression patterns from head samples from 8 individual wasp larvae under high and low nourishment, we identified 284 *P. metricus* differentially expressed transcripts (DETs) that differed between low and high nourishment treatments, using DESeq (FDR \( \leq 0.05 \), Benjamini and Hochberg (1995)), heretofore referred to as “*P. metricus* nourishment-responsive DETs”. Of these DETs, 207 (72.9%) were upregulated in low compared to high nourishment larvae (Figure 3.1 is a heat map of the scaled DET read counts across *P. metricus* nourishment samples. Online Material - Additional file 1 includes the list of *P. metricus* nourishment-responsive DETs). Thus, despite the fact that they had less food available, low nourishment led to a majority of genes having higher gene expression, and thus did not simply cause a general shutdown in transcription.

One important feature of these results is the presence of outlier individuals, i.e. one individual low nourishment larva clustered with the high nourishment samples and a high nourishment larva clustered with the low nourishment samples (Figure 3.1). The variability of gene expression amongst biological samples is not completely surprising; our previous work has found high inter-individual variation in gene expression (Berens et al., 2015a), effects of lab rearing (JM Jandt, JL Thomson, AC Geffre, AL Toth: Rearing environment may bias social traits: A case study with *Polistes* wasps, submitted), and variable effects of nourishment level on physiology (Judd et al., 2015).
Validation of select expression patterns via comparison to qRT-PCR data

It is important to validate RNA-Seq data using another method such as quantitative reverse transcription polymerase chain reaction (qRT-PCR). However, for this study, we had limited samples and quantities of RNA (Judd et al., 2015), so we could not perform qRT-PCR validation on actual samples from this experiment. In lieu of the sample limitations, we instead made a comparison to pre-existing data on nourishment-responsive expression patterns from adult brains of *P. metricus* under starved vs. ad lib food conditions (Daugherty et al., 2011). Although not ideal because of differences in tissue type and life stage, this comparison still provides a useful point of comparison to validate whether the expression patterns uncovered in this RNA-Seq study are robust.

Of 24 candidate genes, Daugherty et al. (2011) identified 10 genes with differential expression in adult *P. metricus* brains reared under low and high nourishment conditions using qRT-PCR. We tested for a correlation (Spearman) in the log2 fold changes between the 24 candidate genes from Daugherty et al. (2011) and the orthologous transcripts in this study, i.e. best BLAST hits between the primer sequences from Daugherty et al. (2011) and the *P. metricus* transcriptome (Appendix B: Figure B.1 is a visual representation of the log2 fold changes for both studies; Online Material - Additional file 3 lists the log2 fold change for each gene/transcript and the directionality for differentially expressed genes). For some genes, there is more than one best BLAST hit to the *P. metricus* transcriptome, so all transcripts were used for the correlation analysis. The gene *PmTOR* is absent from the *P. metricus* transcriptome, so this gene was removed from the analysis. Although none of the orthologous transcripts are significantly differentially expressed in the RNA-seq study, we nonetheless identified a significant positive correlation in log2 fold changes between these two studies (Spearman $\rho$: 0.54, p-value = 0.001), which provides support for our observed RNA-Seq results.

One reason we may not have observed statistically significant differential expression of these candidate genes in the RNA-Seq study, despite the strong correlation with the qRT-PCR data, is because of the limited statistical power. Differential expression calls are more stringent in this RNA-Seq study compared to the qRT-PCR analysis because familywise error correction is more
severe due to the larger number of genes (over 75,000 compared to 24 genes). Furthermore, the qRT-PCR study has a larger sample size (n = 47) compared to the RNA-Seq study (n = 8), which provides greater statistical power for detecting differential expression.

**Comparison to caste-related gene expression**

Next, we investigated how *P. metricus* nourishment-responsive DETs compared to caste-related gene expression. A previous study (Berens et al., 2015a) compared gene expression in field-collected, early season (worker-destined) larvae to late season (gyne-destined) larvae, and identified 736 caste-related DETs. Both low nourishment (described above) and worker-destined larvae (Berens et al., 2015a) show a pronounced bias towards upregulated gene expression (72.9% upregulated in low nourishment and 91.7% upregulated in worker-destined larvae). This pattern generally agrees with the prediction that worker-destined and nourishment-deprived larvae have similar transcriptional states.

There was a statistically significant but relatively small overlap (43 transcripts) between the nourishment-responsive and caste-related DETs (Chi-square with Yates’ correction: \( p < 0.0001 \)). Of these shared DETs, the directionality of gene expression change was unexpected: eight (18.6%) were upregulated in low nourishment larvae compared to high nourishment larvae, whereas 38 (88.4%) were upregulated in worker-destined larvae relative to queen-destined larvae (see Figure 3.2a for a Venn diagram of the number of unique and overlapping caste and nourishment-responsive DETs, Appendix B: Figure B.2 for a heatmap of the overlapping caste and nourishment-responsive DETs, and Online Material - Additional file 4 for the list of overlapping DETs). To further explore these data beyond examining an overlap of gene lists, we performed a combined statistical analysis of data from both the nourishment level and caste contrasts, and the results also indicate some overlap in gene expression patterns across the two studies (see Methods, with complete description of the approach and results in Appendix B and Online Material - Additional file 5). Taken together, these results partially support our prediction of gene expression similarity between nourishment-responsive and caste-related gene expression but suggest some unanticipated dissimilarities between them.
Pathway level analysis

Our previous study suggested gene expression similarity across studies is more pronounced on the level of pathways and gene functional categories, rather than specific genes or transcripts (Berens et al., 2015a). Therefore, we also examined our data at the level of pathways using the Kyoto Encyclopedia of Genes and Genomes (KEGG) Kanehisa and Goto (2000); Kanehisa et al. (2014). Using best BLAST hits to D. melanogaster, very few transcripts (799 = 1.0% of the transcriptome) were annotated with enzyme codes for the KEGG analysis with Blast2GO (Conesa et al., 2005) (see Online Material - Additional file 1 for a list of transcripts annotated with enzyme codes by KEGG pathway). The presence of many transcripts without homology is a shortcoming of the dataset, but is expected and standard for Illumina-based data from non-model species (Ferreira et al., 2013). Only 7% (20) of the nourishment DETs were known members of KEGG pathways, therefore we identified very few (7 of 118) KEGG pathways with at least one DET. We then looked to see if these seven pathways were also related to caste differences (Berens et al., 2015a). Four of the seven pathways also had caste-related DETs (including glycerolipid metabolism, nitrogen metabolism, and purine metabolism; listed in Online Material - Additional file 4), but this amount of overlap was not statistically significant (Fisher’s exact test, p-value = 0.12). The majority of the DETs are upregulated in both low nourishment and worker-destined larvae for all four of these shared KEGG pathways (Figure 3.2b is a Venn diagram of the number of KEGG pathways with nourishment and/or caste DETs), which is in agreement with the prediction that nourishment-responsive biochemical pathways are regulated in a similar direction to what is found between castes in paper wasps.

Gene Ontology (GO) enrichment analysis

On the level of gene functional categories, fifty-two GO terms were significantly enriched within the P. metricus nourishment-responsive DETs compared to the remainder of transcriptome (Figure 3.3; Online Material - Additional file 1 includes the list of enriched GO terms). These included functions related to lipoprotein metabolism, oxidation reduction activity, and polysaccharide metabolism. More than 30% of GO terms (including terms related to oxidore-
ductase activity and carbohydrate metabolism) were common to both caste- and nourishment manipulation-related DETs, representing a significant overlap (Chi-square test with Yates’ correction, $p < 0.0001$; Figure 3.2c; Online Material - Additional file 4 lists the shared caste and nourishment enriched GO terms). However, if we examine the direction of differential expression of DETs associated with these GO terms, we do not consistently see the predicted pattern of the same directional bias to both low nourishment larvae and worker-destined larvae (Figure 3.2c; Online Material - Additional file 4 lists the direction).

**Cross-species comparisons**

Because nourishment is an important driver of phenotypic plasticity in many species, we were interested in determining whether the molecular mechanisms underlying the response to nourishment identified in paper wasps are conserved in other taxa. After an extensive literature search, we identified two studies from other insects that also examined transcriptional responses to nutritional stress: 1) a microarray study of tissue-specific (fat body and muscle) gene expression derived from fruit fly larvae (Teleman et al., 2008) and 2) a microarray study of thoracic horn development in female dung beetle pupae (Kijimoto et al., 2014). Although not ideal comparisons to our paper wasp dataset because of differences in life stages and sampled tissues, these studies still provide useful preliminary comparisons to begin addressing whether there are any conserved of nourishment-responsive transcripts.

When comparing *P. metricus* nourishment-responsive transcripts with nourishment-responsive transcripts in D. melanogaster fat body and muscle tissues (Teleman et al., 2008), we identified small, non-significant overlaps: only 18 and 13 common DETs, respectively (Figure 3.4a and c; Chi-square tests with Yates’ correction; fat body: $p$-value = 0.84; muscle: $p$-value = 0.19). For both tissue types in fruit flies, most transcripts are down-regulated with low nourishment (Online Material - Additional file 6 lists the common DETs and directionality for each species). Of these shared DETs, the majority are expressed in the same direction in both species: 56% (10 DETs for the fat body dataset; Fisher’s Exact Test, $p$-value = 0.676) and 54% (7 DETs for the muscle; Fisher’s Exact Test, $p$-value = 0.730).
At the level of GO categories, there is a statistically significant overlap (six GO terms including carbohydrate metabolic processes [GO:0005975] and oxidation-reduction process [GO:0055114]; see Online Material - Additional file 6 for the complete list of overlapping GO categories) in the GO terms associated with nourishment-responsive transcripts for *P. metricus* and *D. melanogaster* fat bodies (Figure 3.4b; chi-squared test with Yates’ correction, p-value = 9.557e-11). The consistency in directionality between shared *P. metricus* and *D. melanogaster* fat body GO categories (i.e. up-regulation in low nourishment samples for all GO categories, except carbohydrate metabolism with up-regulation in high nourishment samples) further supports common functional changes related to nourishment manipulation in both species. However, this signal is not observed in the comparison between *P. metricus* nourishment-enriched GO terms and *D. melanogaster* nourishment-enriched GO terms in the muscle tissue. There is only one common nourishment-enriched GO term (oxidation-reduction process [GO:0055114]; Chi-square test with Yates’ correction, p-value = 0.4235), which is upregulated in the *P. metricus* low nourishment larvae but upregulated in the high nourishment samples for the *D. melanogaster* muscle tissue (Figure 3.4d).

Out of 18,061 homologous transcripts between *P. metricus* and a dung beetle *Onthophagus taurus* (Kijimoto et al., 2014), we again identified a very small overlap (4 transcripts; not significant; Chi-square test, p-value = 0.57) in the nourishment-responsive transcripts for *P. metricus* female larvae and nourishment-responsive transcripts in *O. taurus* female pupal thoracic horns. However, this overlap between paper wasps and dung beetles is again consistent in the directionality of nourishment-responsiveness with shared DETs being upregulated in low nourishment individuals (Figure 3.4e; Online Material - Additional file 6). Comparing enriched GO terms between paper wasps and dung beetles, we found only two shared GO categories: aminoglycan metabolic process (GO:0006022) and chitin metabolic process (GO:0006030) (Chi-square test with Yates’ correction, p-value = 0.08; Figure 3.4f), which suggests few common functional changes related to nourishment in samples from these two species.
3.3 Discussion

In this study, we provide the first genome-wide transcriptional profiling of nourishment response during development in the genus *Polistes*, a model for understanding the evolution of social castes. We identified 285 nourishment differential expression transcripts (DETs) between larvae raised experimentally on low vs. high nourishment, many of which are associated with lipid metabolism and oxidation-reduction activity. Most (73%) of the *Polistes* nourishment-responsive transcripts are upregulated in larvae with low nourishment, which is opposite to the pattern observed in other insects including fruit flies (Teleman et al., 2008) and dung beetles (Kijimoto et al., 2014). There were few conserved nourishment-responsive transcripts across species, and these were related to aminoglycan metabolism, carbohydrate metabolism, and oxidation-reduction activity. Among transcripts corresponding to those functions, our data show some cross-species consistency in the direction of expression in response to nourishment deprivation. The overall picture from our preliminary cross-species comparisons is that it is largely different genes that show transcriptional responses to nourishment stress across insect orders. However, it is important to note that transcriptional similarities across these three systems may be underestimated from our analysis because of inconsistencies in the datasets such as analyzing different life stages and tissue types. Further work with directly comparable datasets is needed in order to better understand the extent of conservation of nourishment-responsive gene expression across taxa.

Nourishment inequalities have long been considered the most important environmental determinant of castes in *Polistes* (Hunt, 1991; Hunt and Amdam, 2005; Hunt et al., 2010; Karsai and Hunt, 2002). If nourishment plays a major role in caste bias in *P. metricus*, we predicted that worker-destined larvae from natural nests in the field would have similar transcript expression patterns as experimental low nourishment larvae. With respect to directionality of overall transcript expression, this was the case - most DETs were upregulated in worker-destined larvae (92.1%) and in low nourishment larvae (72.9%). This work agrees with previous work in honey bees that also showed more genes with worker-biased expression in larvae, but more genes with queen-biased expression later in development (pupal stage) (Barchuk et al., 2007; Evans
and Wheeler, 1999). We also found a small but statistically significant overlap (43 common transcripts) between nourishment-responsive and caste-related DETs in *Polistes* (Chi-square with Yates' correction: $p < 0.0001$; Figure 3.2; Appendix B: Figure B.2). Focusing on the 43 common DETs, however, shows that the majority of these were upregulated in worker-destined larvae but down-regulated in low nourishment larvae, which is opposite to our prediction and to the pattern of all DETs. This suggests that while nourishment restriction in the laboratory may lead to upregulation of transcript expression in some of the same pathways that are related to worker development, this manipulation did not succeed in causing a full shift to worker-like transcript expression patterns.

Overall, our data suggest that nourishment level caused a partial shift in gene expression, with low nourishment individuals being more worker-like and high nourishment individuals being more gyne-like. However, it is pertinent to note that the nourishment manipulation may not have resulted in a strong nutritional stress; i.e., wasps may have compensated for low proteinaceous food availability by consuming more sugar. In a recent study examining adult wasps reared from the same nests as the larvae analyzed here, the nourishment manipulation caused only a partial shift in caste-related physiology (Judd et al., 2015). In that study, total lipid and protein hemolymph levels in adults were not affected by low nourishment, whereas adults that had been reared with low nourishment showed greater ovary development after two weeks of being fed while in isolation. This counterintuitive observation corresponds to previous studies showing that adult workers collected in the field have greater ovary development than better-nourished gynes (Toth et al., 2009) and substantial evidence that workers are in a physiological state ready to reproduce upon emergence (Judd et al., 2015), whereas gynes are in reproductive diapause until after the overwintering period (Hunt, 2007). Taken together, the physiological data (Judd et al., 2015) and transcriptomic data (this study) indicate that nourishment alone did not completely shift the developmental trajectory of larvae towards one caste or another.

One important caveat of our study is that the effects of the nourishment treatment on gene expression and physiology may be influenced by laboratory rearing, which may produce conflicting results compared to a natural field setting (JM Jandt, JL Thomson, AC Geffre,
AL Toth: Rearing environment may bias social traits: A case study with *Polistes* wasps, submitted). Lab-reared wasps typically have higher lipid stores perhaps due to both overfeeding and inactivity (Daugherty et al., 2011), and lab-rearing can perturb caste-related gene expression (JM Jandt, JL Thomson, AC Geffre, AL Toth: Rearing environment may bias social traits: A case study with *Polistes* wasps, submitted). Two genes (*inositol oxygenase* and *Hsp90alpha*) known to exhibit caste-specific gene expression are perturbed due to lab rearing, but neither gene was differentially expressed in the current study. In addition, the low and high nourishment samples in our study were reared by single foundresses with unrestricted access to sucrose as opposed to the two feeding levels for caterpillars that they fed to larvae. Diets with a high carbohydrate to protein ratio can cause an increase in lipid levels in the insect fat body (Musselman et al., 2011; Skorupa et al., 2008; Warbrick-Smith et al., 2006). The effects of lab-rearing could thus be a factor in some of the non-overlap we observed between nourishment-responsive and caste-related DETs in the current study. Nonetheless, our data do suggest that low nourishment can trigger expression of some of the same genes and pathways that are associated with worker caste development in *Polistes metricus*.

Our results provide insight into the role of nourishment in the differential gene expression patterns that lead to different castes in *Polistes*. Our results also support the notion that nourishment inequalities alone cannot explain all caste variability in gene expression. Instead, caste-related gene expression bias is likely to be additionally influenced by social factors such as dominance behavior (Jandt et al., 2014; Pardi, 1948), vibrational communication (Jeanne and Suryanarayanan, 2011; Suryanarayanan et al., 2011b,a), and/or epigenetically mediated environmental influences (Weiner and Toth, 2012). Further work on gene expression in relation to nourishment, other influences on caste determination in *Polistes*, and interactions among them can further advance our growing understanding of caste determination in primitively social wasps.

### 3.4 Conclusions

This study provides new data on nourishment-responsive gene expression in the context of caste development for *Polistes metricus*, a model for studying the evolution of social insect
castes. We identified suites of nourishment-responsive transcripts in developing *P. metricus* larvae. Interestingly, most transcripts were upregulated when larvae experienced proteinaceous nourishment deprivation; thus, reduced food level did not shut down gene expression but instead resulted in active transcription of many genes including several involved in lipid metabolism, carbohydrate metabolism, and oxidation-reduction processes. By comparing to previously reported caste-related gene expression patterns from the same species, we uncovered some similarity in transcripts, pathways, and gene functions related to both nourishment deprivation and worker caste-biased expression. However, many caste-associated genes were not found to be nourishment-responsive, so there are additional factors (likely including many social environmental factors) that influence caste-related gene expression. These results support the notion that nourishment level during development can somewhat bias development into queen or worker caste as adults, but leave room for other factors, and thus underscore the complex and multifactorial nature of caste development.

### 3.5 Methods

**Samples**

We collected adult female *P. metricus* wasps in early March 2009 near Raleigh, NC, as they were exiting the attic of a house in which they had passed the winter in behavioral quiescence and reproductive diapause. Single wasps were placed in cages 30 cm in length, width, and height constructed of clear plastic with plastic screen on the top and two sides. An opening in the top was covered with a piece of hardboard 10 cm square with a nest from a previous season attached to the underside. Nests were trimmed to seven cells ca. 0.5 cm deep, and meconia (larval feces) of the original nest occupants were removed. Each seven-cell nest therefore served as a “starter” nest on which the newly-captured wasps could initiate construction and then expand as their own. Construction paper was the pulp source for nest construction.

Caged wasps were placed on wire racks in a growth chamber $1.2 \text{ m} \times 2.5 \text{ m} \times 2.1 \text{ m}$ in the North Carolina State University Phytotron. The chamber contained incandescent lights on a 16 L/8D cycle; fluorescent lights came on 2 h following and went off 2 h before the incandescent
lights. Light intensity was 21 micromoles/sec/m² in the incandescent-only morning and evening periods and 225 micromoles/sec/m² during the both-lights midday. Chamber temperature was 20°C at night and 30°C during full light, with a gradual ramp-up in the morning and ramp-down in the evening. Each day, we repositioned cages in a rotation pattern to distribute any light or temperature variation in the chamber across all colonies over the course of the experiment.

Each cage was provided with ad libitum sucrose (rock candy) and a water source (15 ml tube plugged with cotton). Both high and low nourishment groups had equal access to sucrose, but differed in the amount of proteinaceous food provided. Proteinaceous nourishment was provided in the form of late 3rd or early 4th-instar *Manduca sexta* larva ca. 2.0 cm in length. Low nourishment foundresses received one caterpillar every fourth day. After May 22, the feeding schedule for low nourishment foundresses was accelerated to one caterpillar every three days. In the high nourishment treatment, foundresses were provided with caterpillars ad lib, adjusted each day to one caterpillar above the number that had been consumed since the previous day’s provisioning. For this experiment, we collected the largest final instar larva in each nest immediately following spinning of a cocoon by the third larva in the nest to do so. Those three larvae were allowed to develop into adults that then were analyzed for developmental and physiological characteristics. These results were published separately (Judd et al., 2015). For our study, four larvae per treatment were flash frozen in liquid nitrogen and stored at -80°C prior to RNA extraction.

**RNA-extraction and RNA-sequencing**

*Polistes* possess behaviorally distinct, not morphologically distinct castes, thus we were especially interested in developmental expression changes in larval brains. This led us to choose larval heads as the target tissue for our analysis. Whole heads were used rather than whole brains because of concerns about larval brain dissection quality and low RNA yields. We acknowledge that it is likely that the nourishment manipulation caused additional gene expression changes in other tissues we did not sample. However, we did not use whole bodies for two reasons: 1) our primary interest in brains and behavioral castes, and 2) the fact that brain gene expression differences in social insects can be subtle (Whitfield et al., 2003), meaning that high levels of
expression from the fat body and other tissues could have drowned out expression patterns from the brain.

To preserve RNA during dissection, we removed the head region (indicated by dark coloration) of each individual larva, while kept on dry ice, using a sterilized razor blade. From these individual larval heads, we extracted total RNA using an RNeasy Mini Kit (Qiagen), which was quality controlled using spectrophotometry (NanoDrop 2000) and a Bioanalyzer (Agilent). The High-Throughput Sequencing and Genotyping Unit of the W.M. Keck Center (University of Illinois at Urbana-Champaign) prepared 16 mRNA Seq libraries (n = 4 per group, high and low nourishment level) derived from two experimental nourishment levels (high and low; this study) and unmanipulated castes (queen- and worker-destined; Berens et al. (2015a)) using the “TruSeq RNAseq Sample Prep kit” (Illumina). These prepared libraries were sequenced on a HiSeq 2000 (Illumina) to generate over 1 billion 100 base paired-end reads, which we assembled de novo into a transcriptome (deposited at DDBJ/EMBL/GenBank under the accession GBGV00000000; this study used the first version GBGV01000000). Raw sequence data has been deposited to the National Center for Biotechnology Information’s (NCBI) Short Read Archive (BioProject ID: PRJNA242774). For a more details about the RNA-Sequencing and the transcriptome assembly, see Berens et al. (2015a).

Read mapping, abundance estimation, and differential expression analysis

As described in Berens et al. (2015a), we aligned the raw paired-end reads to the reference transcriptome (GBGV01000000) using Bowtie 2 (Version 2.1.0) (Langmead and Salzberg, 2012) with default settings. From these alignments, we quantified transcript abundances for each library using eXpress (Version 1.3.1) (Roberts and Pachter, 2013), which have been deposited in NCBI’s Gene Expression Omnibus (Edgar et al., 2002) and are accessible through GEO Series (NCBI:GSE61960). From these raw read counts, we identified differentially expressed transcripts (FDR ≤ 0.05, Benjamini and Hochberg (1995)) between nourishment manipulation levels (high and low) using the R (Version 3.0.1) (Team, 2013) statistical package DESeq (Version 1.12.0) (Anders and Huber, 2010) downloaded from the Bioconductor repository (Gentleman
et al., 2004). Raw read counts were normalized by the effective library size, and dispersion factors were determined based on the per-condition method.

**Kyoto encyclopedia of genes and genomes (KEGG) and gene ontology (GO) analyses**

From functional annotations based on best BLASTx hit (E-value ≤ 1e-3) to *Drosophila* melanogaster sequences, we used Blast2GO (Version 2.6.5) (Conesa et al., 2005) to assign enzyme codes, metabolic (KEGG) pathways, and GO terms to *P. metricus* transcriptomic sequences (see Berens et al. (2015a)). Based on these assignments, we identified the KEGG pathways that contain nourishment DETs and assessed enrichment of GO terms (FDR ≤ 0.05; one-tailed, Benjamini and Hochberg (1995)) between the nourishment DETs compared to the background – the complete *P. metricus* transcriptome.

**Caste and nourishment manipulation comparison**

To assess the extent that differential nourishment biased gene expression patterns toward a particular caste phenotype in *Polistes* paper wasps, we compared molecular signatures of nourishment manipulation and caste determination (Berens et al., 2015a) at three levels: transcript differential expression, KEGG pathways with DETs, and GO enrichment. At the transcript level, we tested for significance in overlap between nourishment and caste DETs given the number of DETs per each dataset and the total number of *P. metricus* transcripts by using a Chi-square test with Yates’ correction. For both the nourishment and caste datasets, we identified which KEGG pathways had at least one DET, and we then performed a Fisher exact test to determine whether there is a statistically significant number of shared KEGG pathways with DETs given the number of unique pathways with DETs per each dataset and all pathways without DETs. Finally, we assessed significance in the overlap between nourishment and caste GO enriched categories in relation to the total number of annotated, not enriched, GO terms and unique nourishment/caste enriched GO terms using a Fisher exact test.
Alternative statistical approach to address overlap across nourishment and caste studies

We investigated an alternative approach (beyond comparisons of DET lists) to test for similarity in the transcript expression patterns associated with nourishment restriction and worker-caste development. In some sense, across the two studies we have a replicated experiment with respect to nourishment levels, which utilizes both the lab and field setting. As described in the introduction, worker-destined typically larvae receive more limited nourishment compared to queen-destined larvae (Hunt and Amdam, 2005; Hunt, 2007; Reeve, 1991; West-Eberhard, 1969). Therefore, we modeled these datasets together while controlling for location (lab or field) to identify nourishment-responsive transcripts. Overall, our results suggest some consistency in the modeling and list comparison approaches. For full details and results using this approach, please see Appendix B for Supplemental Methods, Results, Table B.1, and Figures B.3-B.7 and Online Material -Additional file 5 for lists of nourishment-responsive transcripts.

Cross-species comparison

Based on our interest in whether nourishment differences induce similar molecular responses in transcript expression and biological processes, we reviewed the literature and searched data repositories (NCBI Gene Expression Omnibus and EMBL ArrayExpress) to identify the most directly comparable studies to the *P. metricus* nourishment manipulation dataset based on the following criteria: 1) the datasets compared late-stage subadult insects reared on two nourishment levels and 2) the studies used a transcriptome-wide approach, i.e. microarray or RNA-seq. We identified two studies that meet both of these criteria: one in fruit flies (*D. melanogaster*, Teleman et al. (2008)) and the other in dung beetles (*Onthophagus Taurus*, Kijimoto et al. (2014)). For this analysis, we compared lists of putative orthologous nourishment-biased DETs and nourishment enriched GO terms between paper wasps and fruit flies or dung beetles using Chi-square tests with Yates’ correction. We defined putative orthologous sequences between *P. metricus* and either *D. melanogaster* or *O. taurus* as the best BLAST hit (E-value ≤ 1e-3) between pairs of species. For the fruit fly comparison, we only used the wild-type fruit fly data,
and we compared the paper wasp dataset to both tissue types in fruit flies (adipose and muscle). The GO terms listed in the publication of Teleman et al. (2008) were not consistent with the current Gene Ontology database, so we identified fruit fly nourishment enriched using Blast2GO (FDR ≤ 0.05; one-tailed, Benjamini and Hochberg (1995)). For the dung beetle comparison, we focused our study on a comparison with female thoracic horn dataset, which we felt was the most akin to the paper wasp dataset because samples were derived from tissues in female insects during the pupal stage of development.

**Availability of supporting data**

Raw sequence data has been deposited to the National Center for Biotechnology Information’s (NCBI) Short Read Archive (BioProject: PRJNA242774). The transcript abundance data is available in the NCBI’s Gene Expression Omnibus repository (Accession number: GSE61960). All other data sets supporting the results of this article are included within the article and its additional files.

### 3.6 Acknowledgements

Talbia Choudhury assisted with the wasp rearing at North Carolina State University. Carole Saravitz and Janet Shurtleff accommodated our work at the North Carolina State University Phytotron, including complimentary use of the growth chamber. *Manduca sexta* caterpillars were provided by the North Carolina State University Insectary, Beverley Pagura Director. Amy Geffre and Cecile Mercado extracted RNA for sequencing. Members of the Toth laboratory provided insightful comments during the preparation of the manuscript. This work was supported by the National Science Foundation IOS 1146410.
Figure 3.1  Heat map of the relative expression (sample read counts scaled by library size then across each gene) for the 284 differentially expressed transcripts between nourishment manipulation levels by sample. Samples are clustered based on relative expression across all differentially expressed transcripts (top), and transcripts are clustered based on relative expression across all samples (left). Transcripts that are upregulated in high nourishment (H) are highlighted in yellow, and low nourishment (L) upregulated transcripts are highlighted in blue.
Figure 3.2  Number of A) differentially expressed transcripts (DETs), B) KEGG pathways with at least one DET, and C) enriched GO term unique to or shared between nourishment and caste datasets. Directionality, i.e. upregulated treatment group, is indicated for each dataset, where, for KEGG pathways and GO, terms, directionality is defined as the treatment group with the greater number of upregulated DETs per pathway or category, respectively. * indicates statistically significant overlap between nourishment and caste DETs (Chi-square test with Yates’ correction, p-value < 0.0001). ** indicates statistically significant overlap between nourishment and caste enriched GO terms (Fisher’s exact test, p-value = 4.4e-6). The fatty acid biosynthesis pathway had an equal number of DETs upregulated in the high and low nourishment groups (one per each group), so the directionality of this pathway is counted as one half for each group (indicated by "+")
Figure 3.3 Bar chart of GO categories significantly enriched (FDR < 0.05; one-tail) between nourishment differentially expressed transcripts (DETs) and remaining transcriptome. All significantly enriched GO categories were over represented in the nourishment-responsive DETs compared to the rest of the transcriptome. 17 significantly enriched GO categories were shared in common for both caste and nourishment. Directionality is indicated for the enriched GO categories and defined as the treatment group with the greater number of upregulated DETs per category.
**Figure 3.4** Number and overlap of nourishment-responsive differentially expressed transcripts (DETs) between *Polistes metricus* (paper wasp) and A) *Drosophila melanogaster* (fruit fly) fat body, C) fruit fly muscle, or E) *Onthophagus taurus* (dung beetle) female thoracic horn. Number and overlap of enriched GO terms between paper wasps and B) fruit fly fat body, D) fruit fly muscle, or F) dung beetle thoracic horn. Directionality, i.e. which nourishment treatment group showed upregulation, is indicated for each dataset. There is only a statistically significant overlap between paper wasp and fruit fly wild-type fat body enriched GO terms (indicated by *, Chi-squared test with Yates' correction, p-value = 9.557e-11). All other comparisons were not significant (Chi-squared tests with Yates' corrections, p-values > 0.05).
CHAPTER 4. CANDIDATE GENES FOR INDIVIDUAL RECOGNITION IN POLISTES FUSCATUS PAPER WASPS

Modified from a paper to be submitted to Genes, Brain, and Behavior

Ali J. Berens, Elizabeth A. Tibbetts, Amy L. Toth

Abstract

Few animals are known to individually recognize conspecifics, i.e. learning the unique features of conspecifics and recalling these individuals during subsequent encounters, and nearly all of them are social vertebrates. Remarkably, the social paper wasp Polistes fuscatus has recently been discovered to possess this ability, which appears to be useful for remembering individual identity during competitive social interactions while wasps are initiating their nests. Polistes provide a unique opportunity to dissect the mechanistic basis of this specialized form of memory. We analyzed brain gene expression in staged encounters between pairs of individuals in the laboratory to explore potential mechanisms underlying wasps' ability to recall familiar individuals using real time quantitative reverse transcription polymerase chain reaction. We identified two top candidate genes (Nckx30C and Su(var)1-10, previously shown to be correlated with wasp social dominance status) that were down-regulated in the presence of familiar individuals compared to solitary wasps and wasps meeting for the first time. These candidate genes are related to calcium signaling, therefore, we treated wasps with lithium chloride, a pharmacological agent known to inhibit calcium signaling in neurons. This treatment decreased aggression in paper wasps, but did not affect expression of genes related to calcium signaling. Overall, the results from this study suggest that differences in calcium signaling may be related to individual memory recall in wasps, and we present two promising candidate genes for future study with
functions in calcium signaling. These data suggest genes associated with dominance behavior may have been coopted for individual recognition, but further work is needed to address whether these genes have a causal association with the behavior.

4.1 Introduction

The social complexity hypothesis posits that the demands of social living, such as flexible social networks, act as selective forces for the evolution of novel cognitive traits (Byrne and Whiten, 1989; Humphrey, 1976; Jolly, 1966). One such proposed cognitive trait is individual recognition, the ability of an organism to remember a conspecific based on individually distinctive characteristics. Individual recognition is prevalent in many species with complex social systems such as birds (Boeckle and Bugnyar, 2012; Bond et al., 2003; Godard, 1991), dolphins (Bruck, 2013; Sayigh et al., 1999), elephants (McComb et al., 2000; Rasmussen, 1995), and primates (Boysen and Berntson, 1989; Matthews and Snowdon, 2011; Murai et al., 2011; Smuts, 1987). In order to recognize conspecifics, individuals within a population must have variable and distinctive characteristics to be used as individual identity signals and flexible learning and memory capabilities. These distinctive characteristics can be chemical (including olfactory chemical profiles (Mateo, 2006; Rasmussen, 1995)), vocal (including voice biometrics (Boughman and Moss, 2003; Fitch and Hauser, 2003)), and/or visual (including facial features (Leopold and Rhodes, 2010)).

Individual recognition can serve multiple functions in animals: recognition of kin, recognition of aggressive conspecifics, and maintenance of dominance hierarchies. Kin recognition may be important for avoiding inbreeding (Pusey and Wolf, 1996) and when performing altruistic acts by receiving indirect fitness benefits (Gouzoules, 1984; Komdeur and Hatchwell, 1999). Many species of birds, fish, frogs, lizards, and mammals are known to use individual recognition to identify rivals from bordering territories (reviewed in Tibbetts and Dale (2007)). Ranks established through dominance hierarchies allow for cooperation among group-mates trying to gain access to limited resources or mating opportunities, whereby additional (and often costly) aggressive interactions can be avoided. Recognition of individuals and their hierarchical rank can facilitate this cooperation.
Outside of vertebrates, some of the most extreme examples of complex social organization occur in eusocial insects (e.g. ants, bees, wasps, and termites). Eusocial insects are characterized by having colonies of individuals from overlapping generations that undertake cooperative brood care based on a division of labor into reproductive (e.g. female queen) and non-reproductive (e.g. female worker) castes (Wilson, 1971). Chemical cues are common recognition tools used by social insects to differentiate groups of individuals (e.g. nest mates from non-nestmates, age classes, or reproductive castes (Hölldobler and Wilson, 1990)). Until recently, social insects were not considered to recognize individuals, perhaps because of the perceived high cognitive demands of such a behavior. However, two ant species and a wasp species have been shown to exhibit individual recognition of conspecifics during competitive social interactions while initiating a nest (D’Ettorre and Heinze, 2005; Dreier et al., 2007; Sheehan and Tibbetts, 2008; Tibbetts, 2002), a particularly tumultuous phase when dominance hierarchies are being established. Multiple individuals may cooperate to build, defend, and provision for a nest, which increases likelihood of nest survival, but may result in reduced reproductive opportunities for some individuals. In establishing a dominance (and reproductive) rank, queens will have intense battles. The ability to recognize individuals appears to be useful to avoid additional costly dominance contests (D’Ettorre and Heinze, 2005; Dreier and D’Ettorre, 2009; Sheehan and Tibbetts, 2008).

Polistes paper wasps are an excellent system for investigating mechanisms of individual recognition, as they have the best-characterized individual recognition of any insect (Tibbetts and Sheehan, 2013), and there are new genomic resources for several species (Berens et al., 2015a; Ferreira et al., 2013; Toth et al., 2007). Polistes spp. vary in terms of nest founding strategies; they may have a single founding queen, multiple founding queens, or flexibility in number of founding queens (reviewed in Tibbetts (2004)). Species with more flexible nesting strategies tend to have highly variable facial markings, which have independently evolved at least five times within the Polistes genus (Tibbetts, 2004). In some species, such as Polistes metricus, there is little to no variation in facial color patterns and individual recognition abilities are absent (Sheehan and Tibbetts, 2010). In other species, facial markings vary with reproductive quality and are used as signals of social status, without individual recognition (Tibbetts and
Dale, 2004). Finally, in other species of paper wasps, such as our focal species *Polistes fuscatus*, a close relative of *P. metricus*, variable markings are used for individual recognition (Tibbetts, 2002). *P. fuscatus* facial markings are highly variable (Sheehan and Tibbetts, 2010) and used for robust, long-term memory of conspecifics (Sheehan and Tibbetts, 2008, 2011; Tibbetts, 2002). In species of paper wasps with multiple founding queens, such as *P. fuscatus*, queens may interact with many potential partners at multiple nest sites before finally settling and founding a nest (Reeve, 1991).

Although most work has focused on nest founding queens, workers also have highly variable markings (Tibbetts, 2002) and some level of individual recognition (Injaian and Tibbetts, 2014). *Polistes* exhibit social systems that remain flexible into adulthood such that workers can become queens (Reeve, 1991), so it may be important for both *P. fuscatus* queens and workers to recognize individuals to maintain dominance hierarchies (Injaian and Tibbetts, 2014).

To our knowledge, nothing is known about the molecular underpinnings of individual recognition in any insect. Of the little that is known about the underlying genetic and molecular mechanisms of individual recognition in vertebrates, the primary focus has been on olfactory recognition during courtship or parent-offspring interactions (reviewed in Brennan and Kendrick (2006)). Based on studies implicating oxytocin receptors in olfactory recognition of conspecifics in mice (Ross and Young, 2009), researchers have focused on oxytocin as a potential candidate for individual recognition in humans (Rimmele et al., 2009; Skuse et al., 2014). Skuse et al. (2014) identified a single common polymorphism (rs237887) in the oxytocin receptor of humans that is associated with decreased visual facial recognition memory in children with autism spectrum disorders. This study highlights the evolutionary conservation of genetic mechanisms underpinning social cognitive traits across distantly related taxa that use olfactory and/or visual mechanisms to recognize individuals.

Our overall objective is to identify the first candidate genes for insect individual recognition using *Polistes* paper wasps as a model. *Polistes* are a good comparative model for individual recognition because the genus includes a pair of closely related species with and without individual recognition. Furthermore, the *Polistes* system provides a unique opportunity to understand the mechanistic basis of this cognitive ability due to the availability of gene sequence informa-
tion from recent transcriptome studies (Berens et al., 2015a; Ferreira et al., 2013; Toth et al., 2007). We hypothesize that genes associated with individual recognition have been co-opted from genes involved in learning and aggressive behavior in other species of Polistes and other insects. We compared brain gene expression for 18 candidate genes during staged laboratory social encounters from two sister species with differing recognition abilities – *P. fuscatus* wasps recognize individuals while *P. metricus* lack this ability – in order to hone in on promising candidate genes for individual recognition. Following a lead from these expression data that suggested a role of genes related to calcium signaling, we aimed to then address a potential causal role of calcium signaling using a pharmacological treatment, and examined effects on recognition of individual wasps in a laboratory setting.

### 4.2 Materials and Methods

**Social Interactions**

**Behavioral Assay**

Solitary *P. fuscatus* and *P. metricus* foundresses were collected from field sites in Michigan soon after nest founding in May 2011. Foundresses were housed in isolation in individual cages using established methods (Daugherty et al., 2011) for one week. To initially investigate gene expression changes associated with social interactions, we performed behavioral assays on three groups for each species: 1) “single” wasps placed in isolation in a small “trial arena” (8x8x2cm) and videotaped for 3 hours, 2) “first meeting”, pairs of unfamiliar wasps placed in trial arena and videotaped for 3 hours, and 3) “second meeting”, unfamiliar wasp pairs placed in trial arena for two sequential meetings (together for 24 hours (2 hours videotaped), separated for 24 hours hours, and reintroduced in trial arena for 2 hours (videotaped)). To ensure unfamiliarity, randomly assigned wasps pairs were collected at least 3 km apart. For each group, the wasps were freeze-killed on dry ice after the behavioral assay and stored at -80°C.
Behavioral Scoring

Observers watched videotapes of the social interactions to quantify the number of aggressive contacts, non-aggressive contacts, and time spent engaged in non-aggressive behavior. Observers were blind to the meeting number and treatment group. Non-aggressive contacts are defined as both wasps being less than one wasp length away from each other but not performing aggressive behaviors for more than ten seconds. Aggressive contacts included darts (quick lunge movement with or without mandibles open), bites, mounts (where one wasp stand on top of the other), attempted mounts, antennal boxing events (one or both wasps strike each other’s antennae with their own), and grapples (both wasps are entangled and often bite, sting, or exhibit quick rolling movements around the arena). To compare aggression between social interaction trials, an aggression index was calculated between paired wasps (based on the index used in Sheehan and Tibbetts (2008)) that accounts for the number and intensity of each aggressive behavior. This aggression index was calculated by assigning behaviors a score and dividing the summed aggression score by the total number of contacts (aggressive or non-aggressive). Behaviors were scored as follows: non-aggressive contact (0), antennal box (1), dart (2), dart with mandible open (3), bite or attempted mount (4), and grapple or mount (5).

Statistics - Behavioral Assay

For the social interactions behavioral assay, we performed paired t-tests between meeting one and two for the amount of non-aggressive time, number of aggressive contacts, and aggressive index. All statistical analyses were performed in R.

Social Interaction Gene Expression

We compared brain expression between the social groups (single, first meeting, and second meeting) within and between the two Polistes species to test 18 candidate genes for associations with memory and/or social interactions. We selected these candidate genes because of previously known correlations with reproductive dominance in P. metricus paper wasps (Toth et al., 2014) and/or learning/memory in Drosophila (Table 4.1). Of these candidates, we identified
memory and/or social interaction candidates (listed below and summarized in Table 4.6) based on expression patterns between social interaction groups and paper wasp species using real-time quantitative reverse transcription polymerase chain reaction (qRT-PCR).

1. **Individual Memory Recall Candidate Genes:** Unique expression patterns (up- or down-regulation) in second meeting wasps in *P. fuscatus* (and not *P. metricus*, which does not have individual recognition) are top candidate genes for memory recall of individual identity.

2. **Memory Recall / Habituation Candidate Genes:** However, we cannot rule out genes with expression differences between second meeting wasps and all other groups in both species. *P. metricus* does not have individual recognition, but aggression decreases after multiple social interactions regardless of familiarity between wasps (Sheehan and Tibbetts, 2010) suggesting that *P. metricus* wasps become habituated to aggressive social competitions. In contrast, *P. fuscatus* wasps decrease aggression only during interactions with familiar wasps, i.e. social interactions remain aggressive even after multiple competitions with new unfamiliar wasps (Sheehan and Tibbetts, 2008, 2010). Thus, genes with unique expression during the second meeting in both *P. fuscatus* and *P. metricus* could be coopted between species for different behavioral functions (habituation and individual memory recall). So to distinguish them from the top individual recognition memory candidate genes, we refer to these genes as “memory recall / habituation related” candidates.

3. **Memory Formation Candidate Genes:** Unique expression patterns (up- or down-regulation) in first meeting wasps in *P. fuscatus* (and not *P. metricus*) are candidate genes for individual memory formation.

4. **Social Interaction Candidate Genes:** Genes with expression differences in either species between single wasps vs. first and second meeting wasps are potentially associated generally with social interactions.
Statistics – Gene Expression

For the social interaction gene expression experiment, we initially compared expression across social interaction groups (single, first meeting, and second meeting) of each gene normalized to RP49 using a one-way ANOVA (see Additional Methods in Appendix C for selection of control gene). Then for genes with significant social interaction effects (p-value < 0.05), we made pairwise comparisons between the social interaction groups using a t-test.

Memory Recall Manipulation

Some of the memory recall candidate genes (identified during the social interaction assay) are associated with calcium signaling in Drosophila (Table 4.1). So we attempted to manipulate conspecific memory recall by artificially disrupting calcium signaling using the pharmacological agent lithium chloride (LiCl). LiCl treatments have been previously shown to disrupt calcium signaling by blocking calcium receptors (Wallace, 2014) or inhibiting second messengers (Berridge, 1993). To explore the role of calcium signaling during individual recognition we staged social encounters between pairs of P. fuscatus wasps, then tested whether there were associated behavioral and gene expression changes. If calcium signaling is important for individual recognition, we predicted changes in memory recall (indicated by changes in aggressive behavior) for only the wasps treated with LiCl, not the two controls (DMF or MgCl dissolved in DMF, which should not disrupt calcium signaling). Because some calcium-signaling related candidate genes showed reduced expression during social memory recall, we predicted that LiCl treatment would reduce social memory in wasps.

Behavioral Assay

In May 2013, 72 P. fuscatus foundresses were collected from field sites in Minnesota and Iowa. These foundresses were initially isolated in individual cages for one week. Randomly assigned unfamiliar wasp pairs, consisting of both an Iowan and a Minnesotan wasp, were initially placed in small trial arenas (9x9x2cm) for 24 hours (videotaped for the first 30 minutes – Meeting 1). After 24 hours separation following the initial meeting, these same pairs were
reintroduced in the trial arena for 24 hours (videotaped for the first 30 minutes – Meeting 2). Again these pairs were separated for 24 hours.

We randomly assigned wasp pairs to either a lithium chloride treatment (LiCl dissolved in dimethylformamide [DMF] to final 1M concentration; n = 11 pairs), a magnesium chloride control (MgCl dissolved in DMF to final 1M concentration; n = 11 pairs), or DMF control group (n = 12 pairs). To our knowledge, nothing is known about the acute toxicity of lithium in paper wasps or any other insect. We attempted to determine the lethal dose of a single topical treatment of lithium; however the highest soluble concentration of LiCl in DMF (1M) did not result in any mortality after 5 days (data not shown). We thus selected this highest soluble concentration for both the LiCl and MgCl treatments in order to provide the strongest possible treatment and increase chances that the pharmacological agents would reach the brain.

In honey bees, topical application of octopamine dissolved in DMF to the thorax reached the brain within 15 minutes of application, yet within 60 minutes the octopamine was rapidly lost from the brain and accumulated in the abdomen (Barron et al., 2007). Therefore, we selected an intermediate time point (30 minutes) to test the effects of the treatment groups on memory recall in paper wasps. Thirty minutes prior to a third and final interaction, we topically applied 1 µL of assigned treatment/control to the base of each wasp’s head. These wasp pairs met for a final one hour interaction in the trial arena (videotaped for the first 30 minutes – Meeting 3), then after the behavioral assay the wasps were freeze-killed in liquid nitrogen and stored at -80°C.

Statistics – Behavioral Assay

We performed the same behavioral scoring for the memory recall behavioral assay as social interaction behavioral assay (described above). Because there was no treatment difference between groups during meetings one and two, we performed paired t-tests between meeting one and two for the amount of non-aggressive time, number of aggressive contacts, and aggressive index with all wasp pairs to initially confirm individual recognition. Additionally, for each topical treatment group, we compared the amount of non-aggressive time, number of aggres-
sive contacts, and aggressive index pairwise between meetings using paired t-tests with holm’s adjustments.

Memory Recall Manipulation Gene Expression

We investigated whether the LiCl treatment (as compared to MgCl and DMF controls) affected brain gene expression of five genes related to calcium signaling (IP3K, IP3R, mGluR, Nckx30C, and Nmdar1; Table 4.1) using qRT-PCR. Specifically, we predicted that the expression of the Nckx30C would increase in the presence of LiCl compared to controls, as observed in Drosophila (W. Liang, unpublished, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8330).

Statistics – Gene Expression

To identify whether there were significant expression differences between topical treatments for the memory recall manipulation experiment, we performed a one-way ANOVA for each calcium signaling gene normalized to the control gene RP49.

Brain Dissections

Heads were removed from the body in a small petri dish on dry ice and stored in 1.5mL microcentrifuge tubes at -80°C. Prior to dissections, all tools (small petri dish, castro-viejo scissors, dissecting tweezers, and a dissecting needle with spear head) were cleaned with RNase Away solution. When ready to dissect, the brain was submerged in 100% ethanol solution with dry ice in a small petri dish placed in a larger container of dry ice. From the back of the head, scissors were used to cut away the cuticle while tweezers stabilized the head. We removed the remaining cuticle, muscle tissue, ocelli, compound eyes, glandular tissue, and subesophageal ganglion to dissect out the whole wasp brain.

RNA Extraction

We extracted total RNA from individual whole wasp brains using an RNeasy Mini Kit (Qiagen) with 350 µL of Buffer RLT + β-mercaptoethanol, as recommended for tissue size
< 20mg. As part of the extraction, a DNase I treatment (Qiagen) was applied to remove DNA contamination. We performed the final elution in two steps with 30 µL of RNase-free water per step. To verify the quality of the extracted RNA, we analyzed the samples using a spectrophotometer (NanoDrop 2000). Based on concentration, 260/280 and 260/230 ratios, and the shape of the curve from the NanoDrop, eight samples were selected per group per species.

**Real-time quantitative reverse-transcription polymerase chain reaction (qRT-PCR)**

We reverse transcribed 200 ng of RNA using Superscript II (Invitrogen), then performed thermal cycling in a PCR mix that includes dithiothreitol (DTT), random primers, RNase inhibitor, deoxynucleotides and buffer to amplify the cDNA. PCR conditions were: 25°C for 10 min, 42°C for 50 min, 70°C for 10 min, then gradually lowered to 4°C. We performed quantitative real-time PCR using SYBR green (Applied Biosystems) on a Roche Light Cycler at the Iowa State University Genomic Technologies Facility or BioRad CFX384, with all samples in triplicate reactions and with a standard curve of genomic DNA. Primers were designed based on EST sequences for *P. metricus* (Toth et al., 2007) using Primer Quest Software (Integrated DNA Technologies; sequences given in Appendix C: Table C.1). We verified that the primers were also effective for amplification in *P. fuscatus*, as previously demonstrated for other genes (Jandt et al., 2015). Lab rearing environment perturbs social traits: A case study with *Polistes* wasps. Accepted. *Behavioral Ecology*. qPCR conditions were: 95°C for 10 min, 40 cycles of 95°C for 15 sec followed by 65°C for 1 min (for quantification), then gradual decrease from 65°C to room temperature over 10 min (for melting peak analysis). The concentration for each biological sample was determined to be the mean concentration of the technical replicates excluding any outliers. Technical replicate outliers were identified using a modified Grubb’s test with the additional criteria that the greatest deviation between replicate critical threshold value and mean critical threshold value was greater than 0.25 cycles. We normalized expression to *RP49* (see Additional Methods in Appendix C for selection of control gene).
4.3 Results

Social Interaction Behavioral Assay

For *P. fuscatus*, we observed that the amount of non-aggressive time decreased, the number of aggressive contacts increased, and the aggressive index increased during the second meeting compared to the first meeting (paired t-tests, p-value < 0.1; Figures 4.1a, c, and e). These results are consistent with previous *P. fuscatus* behavioral studies (Sheehan and Tibbetts, 2008, 2010) and verify that individual recognition occurred between *P. fuscatus* wasp pairs during the behavioral trials. There was no significant difference in the amount of non-aggressive time, number of aggressive contacts, or aggressive between the first and second meeting for *P. metricus* (paired t-tests, p-value > 0.1; Figures 4.1b, d, and f). However, there was a large, yet not significant decrease in the number of aggressive contacts for *P. metricus* wasp pairs (paired t-test, p-value = 0.176; Figure 4.1d). This may indicate that some habituation occurred for *P. metricus* wasps; although, the habituation behavior was not confirmed by additional social interactions with unfamiliar wasps.

Social Interaction Gene Expression

We selected 18 candidate genes (see Table 4.1) based on known associations with dominance in paper wasps (Toth et al., 2014) and learning/memory in *Drosophila*, then we examined brain gene expression using qRT-PCR. We compared brain expression for these genes between the groups (single, first meeting, and second meeting) within and between the two *Polistes* species to test candidate genes for associations with memory and/or social interactions.

Individual Memory Recall Candidate Genes

Two genes (*Nckx30C*, and *Su(var)2-10*) are top candidates for memory recall of individual identities because these genes showed unique expression during the second meeting compared to the other groups in only *P. fuscatus* (Figure 4.2a,b; pairwise t-test with Benjamini-Hochberg correction, FDR < 0.05). For all *P. fuscatus* memory recall candidates, gene expression was down-regulated during the second meeting of wasp pairs.
Memory Recall / Habituation Candidate Genes

Two (11%) of the eighteen candidate genes (*IP3K* and *IP3R*) were down-regulated in the second meeting group compared to all other groups in both *P. fuscatus* and *P. metricus* (pairwise t-test with Benjamini-Hochberg correction, FDR < 0.05; Figure 4.2c,d and Appendix C: Figure C.1). All other candidate genes were not differentially expressed between groups for *P. metricus* (pairwise t-test with Benjamini-Hochberg correction, FDR > 0.05; Appendix C: Figure C.1).

Memory Formation Candidate Genes

Five (28%) of the potential candidate genes (*Bap60, Gug, mGluR, Nmdar1, PsGEF*) were up-regulated in the brains of *P. fuscatus* during the first meeting of unfamiliar wasps compared to single wasps with no social interactions or the second meeting of wasp pairs (Appendix C: Figure C.2; pairwise t-test with Benjamini-Hochberg correction, FDR < 0.05). None of the candidate genes showed unique expression in the first meeting wasps for *P. metricus* (pairwise t-test with Benjamini-Hochberg correction, FDR > 0.05; Appendix C: Figure C.1).

Social Interaction Candidate Genes

In *P. fuscatus, Sca* was up-regulated during social interactions (both first and second meeting) compared to no social interaction (single); whereas *IP3K* gene expression was highest with no social interaction, slightly lower during the first meeting of unfamiliar wasps, and lowest during the second meeting of wasp pairs (Figure 4.2c, Appendix C: Figure C.2; pairwise t-test with Benjamini-Hochberg correction, FDR < 0.05).

Memory Recall Manipulation Behavioral Assay

Three of the four memory recall candidate genes are associated with calcium signaling in *Drosophila* (Table 4.1). *Nckx30C* is a potassium, sodium, and calcium exchanger that functions in maintenance of calcium signaling homeostasis (Haug-Collet et al., 1999). *IP3K* and *IP3R* mediate calcium activity by regulating the availability of the second messenger inositol trisphosphate to initiate calcium signaling within a cell (Berridge, 1993). Based on the results
from our previous social interaction assay (above), we focused on calcium signaling as a candidate mechanism for individual memory recall and used LiCl as a pharmacological agent for manipulating calcium signaling.

First, we verified individual recognition had occurred in *P. fuscatus* wasp pairs, as indicated by increased non-aggressive time (paired t-test, p-value=0.054; Appendix C: Figure C.3a), decreased number aggressive contacts (paired t-test, p-value = 0.0014; Appendix C: Figure C.3b), and decreased aggressive index (paired t-test, p-value = 0.054; Appendix C: Figure C.3c) during a second social interaction compared to the first interaction of wasp pairs. We did not observe differences in non-aggressive time or number of aggressive contacts between the second and third meetings (after pharmacological agents were topically applied) for the treatment or control groups (paired t-tests with holm’s adjustment, p-values > 0.1; Figures 4.3a-f). In contrast to both control groups where no difference in aggressive index (accounts for number of contacts and intensity) between meetings two and three was observed (paired t-tests with holm’s adjustment, p-values > 0.1; Figures 4.3g and h), we detected a lower aggressive index in meeting three compared to meeting two for the LiCl treatment (paired t-test with holm’s adjustment, p-value = 0.089; Figure 4.3i). We also observed a significant reduction in aggression index between meeting one and three (paired t-test with holm’s adjustment, p-value = 0.044), but not from meeting one to two (paired t-test with holm’s adjustment, p-value = 0.623), for the LiCl treatment group, which suggests a greater decrease in aggression after the LiCl treatment.

**Memory Recall Manipulation Gene Expression**

We investigated whether the LiCl treatment affected brain gene expression of the memory recall top candidate *Nckx30C* gene and other genes related to calcium signaling (*IP3K*, *IP3R*, *mGluR*, and *Nmdar1*; Table 4.1) using qRT-PCR. None of the memory recall candidate genes, including top candidate *Nckx30C*, associated with calcium signaling showed any change in expression after LiCl treatment compared to the control groups (Figure 4.4). This suggests that expression-mediated changes in these genes did not produce the change in aggression index and memory recall. However, calcium signaling activity was likely still affected by the LiCl, thus it
is possible that LiCl disrupted calcium signaling in a way that reduced aggression, independent of differences in gene expression.

4.4 Discussion

In this study, we identify the first candidate genes for individual recognition (both during memory formation and recall) in any insect. Five genes (Bap60, Gug, mGluR, Nmdar1, and PsGEF) may be important for the formation of memories in P. fuscatus that will be used for individual recognition during subsequent social interactions. The top individual memory recall candidate genes (Nckx30C and Su(var)2-10) had been previously identified as associated with dominance interactions in P. metricus (Table 4.1, Toth et al. (2014)) suggesting genes associated with dominance behavior in this closely related species may have been co-opted for individual recognition in P. fuscatus. Two additional individual memory candidate genes (IP3K and IP3R) were differentially expressed in both Polistes species suggesting that these genes may regulate habituation of competitive social interactions in paper wasps and may also be co-opted for memory recall of known social partners in P. fuscatus. Additional differences in IP3K gene expression between first and second meetings suggest that IP3K may also be important for other social interaction regulation within P. fuscatus.

Based on the work in Drosophila and our gene expression results in paper wasps, we suggest calcium signaling may be important for individual recognition during memory formation and recall in paper wasps. Several of the genes we found to be differentially expressed have roles in calcium signaling. For example, one of our top individual memory formation candidates is mGluR, which is important for courtship interactions and memory of these interactions in Drosophila (McBride et al., 2005; Schoenfeld et al., 2013). Reduction of mGluR activity via lithium treatment (which affects calcium signaling) has been shown to rescue decreased memory in fly mutants that exhibit reduced social interactions (Bolduc et al., 2008; McBride et al., 2005; Pan et al., 2008). Another top individual memory formation candidate is Nmdar1, a glutamate-gated ion channel that functions in associative memory (Qin and Dubnau, 2010), long-term memory (Chen et al., 2012a; Miyashita et al., 2012), and olfactory learning (Xia et al., 2005). Calcium influx can be blocked by antagonizing NMDA receptors with lithium treatment,
which may improve memory (Wallace, 2014). A top memory recall candidate gene is Nckx30C, a potassium-dependent sodium and calcium exchanger that functions in the maintenance of calcium homeostasis during neural signaling in Drosophila adults (Haug-Collet et al., 1999). In Drosophila, Nckx30C appears to be lithium-responsive with greater brain gene expression in the presence of LiCl (W. Liang, unpublished, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8330). Finally, a second memory recall candidate from our study is IP3R, which is an inositol trisphosphate receptor. Lithium decreases IP3R-mediated calcium signaling by reducing available inositol trisphosphate (reviewed in McBride et al. (2010)).

Our data also suggest that aggressive behavior in P. fuscatus was reduced during social encounters with familiar wasps by calcium signaling disrupting treatment (LiCl) and potentially indicate that memory recall was enhanced. Further work showing sustained decrease in aggression after longer periods of time is needed in order to substantiate memory recall enhancement after LiCl treatment. Expression of candidate calcium signaling-related genes (IP3R, IP3K, mGluR, Nckx30C, and Nmdar1) were not perturbed by the LiCl treatment, thus we were unable to conclusively test for a causal association between these genes and memory recall in P. fuscatus. It is possible that the treatment disrupted calcium signaling without affecting the expression of our candidate genes. In our study, calcium activity was not directly measured, thus the role of calcium signaling in memory recall stands as an intriguing, but as of yet inconclusive, possibility that deserves further work in the future.

The fact that we did not observe changes in the expression of candidate calcium signaling genes (IP3K, IP3R, mGluR, Nckx30C, and Nmdar1) after the LiCl treatment could be attributable to several possible different explanations. First, there is evidence that lithium may affect protein levels without affecting mRNA levels for target genes. Lithium treatment rescued locomotory defects and increased lifespan in Arctic mutant amyloid beta 42 (Aβ42) Drosophila, a model for adult-onset Alzheimer’s disease, by reducing Aβ protein levels while mRNA levels remained constant (Sofola et al., 2010; Sofola-Adesakin et al., 2014). Although not tested in our experiment, it is plausible levels of calcium signaling proteins were changed by the lithium treatment which did not cause subsequent changes in gene expression. Second, calcium signaling is used extensively in neurological signaling (as observed in honeybees (Ellen and Mercer, 2012;
Perisse et al., 2009) and fruit flies (Chorna and Hasan, 2012; Ng et al., 2011)), thus the lithium effects may be widespread and not necessarily targeted to our genes of interest. Finally, a calcium imbalance induced by the lithium treatment may have been too short-lived to influence gene or protein expression, but strong enough to produce changes in behavioral aggression.

4.5 Conclusions

The ability to recognize conspecific individuals has evolved many times in highly social species, including a species of Polistes paper wasps. Our results suggest that four genes (IP3K, IP3R, Nckx30C, and Su(var)2-10) previously known to be associated with learning and dominance behavior may be related to memory recall of individual identities with Nckx30C and Su(var)2-10 being top candidates. All of these genes, except Su(var)2-10, have known functions related to calcium signaling. Also, decreased aggressive behavior potentially indicates enhanced individual recognition after disruption of calcium signaling by lithium treatment. These data provide an important first step in understanding the molecular underpinnings of memory of individual identity in a social insect. In the future, targeted approaches for manipulating expression of candidate genes, as well as transcriptome-wide comparative analyses between multiple Polistes species can help to further our understanding of the molecular mechanisms underlying individual recognition.

4.6 Acknowledgements

Taylor Forrest collected wasps in Michigan. Cecil Mercado helped with brain dissections. Vera Zhao assisted with wasp rearing and behavioral trials in Iowa. Erin McCall, Griffin Smith, and Vera Zhao watched many hours of videos to obtain behavioral observations of wasp social interactions. A.J.B. received funding support for this project from the Iowa State University Ecology, Evolution and Organismal Biology (EEOB) Department through an EEOB Graduate Student Research Award and from the National Science Foundation (NSF) through a Doctoral Dissertation Improvement Grant (IOS-1311512).
Table 4.1 Candidate genes for individual recognition based on associations with dominance behavior in *P. metricus* (Toth et al., 2007) or learning/memory in *Drosophila*, some of which are associated with calcium signaling in *Drosophila*. See Table 2 for definition of candidate genes identified by the study based on expression patterns between social interaction groups and between species.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Gene Abbreviation</th>
<th>Dominance in <em>P. metricus</em></th>
<th>Learning / memory in <em>Drosophila</em></th>
<th>Calcium signaling in <em>Drosophila</em></th>
<th>Candidate gene for (this study)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetylcholinesterase</td>
<td>Ace</td>
<td>X</td>
<td></td>
<td></td>
<td>Memory formation</td>
</tr>
<tr>
<td>Brahma associated protein 60kD</td>
<td>Bap60</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choline acetyltransferase</td>
<td>Cha</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep orange</td>
<td>dor</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ebony</td>
<td>e</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Golden goal</td>
<td>gogo</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grunge</td>
<td>Gug</td>
<td>X</td>
<td></td>
<td></td>
<td>Memory formation</td>
</tr>
<tr>
<td>Inositol 1,4,5-trisphosphate kinase 1</td>
<td>IP3K</td>
<td>X McBride et al. (2010)</td>
<td>X Berridge (1993)</td>
<td>Social interaction, Memory recall / Habituation Related</td>
<td></td>
</tr>
<tr>
<td>Metabotropic Glutamate Receptor</td>
<td>mGluR</td>
<td>X McBride et al. (2005); Schoenfeld et al. (2013) X McBride et al. (2010)</td>
<td>Memory formation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Na+/Ca2+-K+ exchanger 30C</td>
<td>Nckz30C</td>
<td>X</td>
<td></td>
<td>X Haug-Collet et al. (1999); Webel et al. (2002)</td>
<td>Individual memory recall</td>
</tr>
<tr>
<td>N-methyl-D-aspartate receptor 1</td>
<td>Nmdar1</td>
<td>X</td>
<td></td>
<td>X Xia et al. (2005)</td>
<td>Memory formation</td>
</tr>
<tr>
<td>Notch</td>
<td>N</td>
<td>X</td>
<td></td>
<td>X Ge et al. (2004); Matsuno et al. (2009); Zhang et al. (2013)</td>
<td></td>
</tr>
<tr>
<td>Protostome-specific guanine nucleotide exchange factor</td>
<td>PsGEF</td>
<td>X</td>
<td></td>
<td></td>
<td>Memory formation</td>
</tr>
<tr>
<td>Scabrous</td>
<td>sca</td>
<td>X</td>
<td></td>
<td></td>
<td>Social interaction</td>
</tr>
<tr>
<td>Suppressor of variegation 2-10</td>
<td>Su(var)2-10</td>
<td>X</td>
<td></td>
<td></td>
<td>Individual memory recall</td>
</tr>
<tr>
<td>Staufen</td>
<td>stau</td>
<td>X Dubnau et al. (2003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synaptotagmin 7</td>
<td>Syt7</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4.2  Classification of candidate genes based on comparisons between social interaction groups (single, first meeting, and second meeting) and between species (*P. fuscatus* and *P. metricus* - with and without individual recognition respectively).

<table>
<thead>
<tr>
<th>Candidate gene for</th>
<th>Species with pattern</th>
<th>Unique expression in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Memory Recall</td>
<td><em>P. fuscatus</em></td>
<td>Second meeting</td>
</tr>
<tr>
<td>Memory Recall / Habituation Related</td>
<td><em>P. fuscatus</em> and <em>P. metricus</em></td>
<td>Second meeting</td>
</tr>
<tr>
<td>Memory Formation</td>
<td><em>P. fuscatus</em></td>
<td>First meeting</td>
</tr>
<tr>
<td>Social Interaction</td>
<td><em>P. fuscatus</em> and/or <em>P. metricus</em></td>
<td>Single</td>
</tr>
</tbody>
</table>
Figure 4.1 Mean (a,b) non-aggressive time, (c,d) number of aggressive contacts, and (e,f) aggressive index with standard error by *Polistes* species (n = 8 for each species, *P. fuscatus* and *P. metricus*) during the social interactions behavioral assay. Bars are colored by meeting number (light grey - first meeting, dark grey - second meeting). There are significant differences in the amount of non-aggressive time, number of aggressive contacts, and aggressive index between the first and second meeting for *P. fuscatus* (paired t-test; p-values < 0.1), not *P. metricus* (paired t-test; p-values > 0.1).
Figure 4.2 Memory recall top candidates and additional candidate genes for *Polistes fuscatus*. Mean (a) *Nckx30C*, (b) *Su(var)2-10*, (c) *IP3K*, and (d) *IP3R* brain gene expression with standard error shown across social interaction types: single (white) - wasps with no interactions, first meeting (light grey) - unfamiliar wasps after first interaction, second meeting (dark grey) - unfamiliar wasps after second interaction (n=8 per social interaction type). Memory recall candidate gene expression was normalized to *RP49* brain gene expression. After normalization, gene expression was scaled with single interaction group set to 1. For these four memory recall candidate genes, there is a significant difference between single (*), first meeting (**), or second meeting (***)) and other social interaction groups in *P. fuscatus* (pairwise t-test with Benjamini-Hochberg correction; FDR < 0.05).
Figure 4.3  Mean (a-c) non-aggressive time, (d-f) number of aggressive contacts, and (g-i) aggressive index with standard error during manipulation of social interactions behavioral assay. Bars are colored by meeting number (white - first meeting, light grey - second meeting, and dark grey - third meeting). Pharmacological agents were applied between the second and third meeting (n = 12 wasp pairs for DMF control (a,d,g), n = 11 for MgCl + DMF control (b,e,h), and LiCl + DMF treatment (c,f,i)), so there are no treatment differences between the first and second (see Supplemental Figure 3 for results the first and second meeting with all wasp pairs combined). P-values provided for significant pairwise differences between meetings by treatment (paired t-test with holm adjustment; p-values < 0.1).
Figure 4.4  Mean *Polistes fuscatus* IP3K, IP3R, mGluR, Nckx30C, and Nmdar1 brain gene expression with standard error across treatments: DMF control (white), 1M MgCl dissolved in DMF (light grey), and 1M LiCl dissolved in DMF (dark grey; n=8 per treatment group). Candidate gene expression was normalized to RP49 brain gene expression, which was not significantly different between treatment groups. After normalization, gene expression was scaled with DMF control group set to 1. No significant treatment effect detected for any of the candidate calcium signaling genes (one-way ANOVA; p-values > 0.05).
CHAPTER 5. BRAIN GENE EXPRESSION ASSOCIATED WITH FACIAL RECOGNITION IN POLISTES PAPER WASPS

Ali J. Berens and Amy L. Toth

Abstract

A few social vertebrates have the ability to learn and recall individuals during social encounters based on distinct facial features, but how is this specialized form of memory accomplished? Social paper wasps in the genus Polistes are the only insects known to possess individual facial recognition. An invertebrate model for facial recognition affords us a unique comparative opportunity to investigate the mechanistic basis of this behavior. We analyzed genome-wide brain gene expression during facial and pattern learning for two sister species of paper wasps (P. fuscatus – has facial recognition, P. metricus – does not possess this ability) using RNA-Seq. We identified expression differences in a relatively small number of genes associated with face vs. pattern learning that were unique to each species of paper wasp. None of these differentially expressed genes were shared between species. Some of the P. fuscatus differentially expressed genes are known to function in olfactory response in organisms without the ability to recognize faces. This result suggests facial recognition may be functionally related on the molecular level to olfactory recognition, but further work is needed to address this connection between recognition systems.

5.1 Introduction

Individual facial recognition, the ability of an organism to remember a conspecific based on individually distinctive facial characteristics, is critical for appropriate social interactions in
some animals including humans, macaques, and sheep. However, very little is known about the underlying mechanisms controlling the recognition of facial identities (reviewed in Tate et al. (2006)). Most of the work has focused on brain region activity associated with facial recognition. There appears to be hemispheric asymmetry with humans (Kanwisher and Yovel, 2006) and sheep (Da Costa et al., 2004; Peirce and Kendrick, 2002) having greater responsiveness in the right hemisphere and left asymmetry in macaques (Perrett et al., 1988). In humans, the inferior occipital and lateral fusiform gyri are activated during facial identity recognition (Haxby et al., 2000). The temporal cortex has been implicated in individual facial recognition for both macaques (specifically the inferior temporal cortex; (Eifuku et al., 2004; Sigala, 2004)) and sheep (along with the frontal cortex and basal amygdala; (Broad et al., 2002; Kendrick, 1991; Kendrick and Baldwin, 1987)).

An important open question relates to the genetic mechanisms underlying facial recognition. Using sheep as a model, there is also some limited evidence for brain gene expression differences associated with facial recognition. Sheep presented with familiar sheep face pictures compared with goat faces or simple pattern pictures had increased expression of neuronal activity-dependent genes (\textit{c-fos} and \textit{zif268}) in specialized face-processing brain regions (Da Costa et al., 2004). In a less direct study on olfaction recognition, Broad et al. (2002) found increased mRNA levels of \textit{brain-derived neurotrophic factor} (BDNF) and its receptor \textit{(tyrosine receptor kinase, trk-B)} in not only brain regions involved in olfaction (pyriform and entorhinal cortices), but also in areas of the brain that function in visual memory (temporal cortex, entorhinal cortex, and basolateral amygdala) and integration (medial frontal and anterior cingulate cortices and diagonal band). Based on these results, Broad et al. (2002) suggest that there is integration between olfactory and visual information such that initial olfactory neural circuits are reorganized for visual recognition.

There are logistical and ethical limitations with these vertebrate models of individual facial recognition. In this study, we begin an exploration of the genetic mechanisms of facial recognition in a more experimentally tractable invertebrate system – \textit{Polistes} paper wasps. While most paper wasps have little variation in facial coloration, facial markings have evolved independently at least five times in the genus \textit{Polistes} (Tibbetts, 2004). For a species with highly variable
and distinct facial markings, *P. fuscatus*, these characteristics can be used for individual facial recognition (Sheehan and Tibbetts, 2008, 2011). In another wasp species with facial markings, the markings only signify social status, not individual identity (Sheehan and Tibbetts, 2010; Tibbetts and Dale, 2004). For species with little to no variation in facial color pattern such as *P. metricus* (a sister species to *P. fuscatus*), the ability to recognize faces is limited (Sheehan and Tibbetts, 2011). *P. metricus* wasps can be trained to learn faces, but their ability to learn is significantly lower than *P. fuscatus* wasps, which more readily learn faces compared to general patterns or prey (Sheehan and Tibbetts, 2011). Therefore, *P. fuscatus* have a cognitive specialization for faces, a complex behavioral repertoire that is shared with only a few other social species such as humans. Comparisons between *P. fuscatus* and *P. metricus* are a great opportunity to study the mechanisms of facial recognition due to the close evolutionary relationship and contrasting abilities of these two species.

We previously identified candidate genes for individual recognition in *P. fuscatus* compared to *P. metricus* based on staged social encounters (Chapter 4). Individual recognition requires both memory formation (encoding perceived initial social interactions into a construct that can be stored in the brain) and memory recall (retrieval of stored memories in the brain during subsequent social interactions). Results from our previous study suggest that differences in brain calcium signaling may be related to individual memory recall in wasps. However, the study in Chapter 4 only tested a limited number (18) of candidate genes for brain expression differences, and there are likely to be many other genes involved in facial recognition besides what we uncovered in that initial study.

That study examined individual memory recall – the retrieval of previously learned individual identities – which is not necessarily specific to facial features. Instead, there could be other distinctive characteristics (such as chemical signals or pheromones) that are used for identifying individuals. With the social encounters assay, wasps recalled not only identity but also other social information such as relative rank. However, we are also interested in homing in on face-specific recognition, so we staged facial learning using an assay that required wasps to indicate whether they have encountered a face before.
The main objective of this study is to investigate the molecular basis of individual facial recognition by conducting the first genome-wide analysis of facial learning-related gene expression. First using RNA-sequencing, we identify genome-wide brain gene expression differences during facial learning compared to pattern learning for *P. fuscatus* and *P. metricus*. This comparative approach affords us the opportunity to determine unique expression associated with the cognitive specialization of individual facial recognition in *P. fuscatus*. We also examine whether calcium signaling and individual memory recall candidate genes (from Chapter 4) are important for facial compared to pattern learning.

Unlike the candidate gene approach in Chapter 4, a genome-wide analysis affords us an opportunity to explore more than just conserved candidate genes. The “novel” taxonomically restricted gene hypothesis posits that genes underlying novel social traits are newly evolved within the social lineage and not present in solitary ancestors (Harpur et al., 2014; Johnson and Tsutsui, 2011; Simola et al., 2013). Since sociality evolved only once in a common ancestor to *Polistes* wasps, we might expect (under the novel taxonomically restricted gene hypothesis) that genes important for facial recognition – a novel social behavior in *P. fuscatus* are related to paper wasp-specific genes. Using four available *Polistes* transcriptomes, we identify a set of well-supported *Polistes*-specific transcripts that can be used to test this novel taxonomically restricted gene hypothesis. If novel genes are important for the evolution of facial recognition in facial recognition, we predict that *Polistes*-specific transcripts would be more likely to be differentially expressed between face and pattern learning in *P. fuscatus*.

### 5.2 Methods

#### Visual Training Experiments

**Training Assay**

We collected 42 *P. fuscatus* and 20 *P. metricus* foundresses near Ann Arbor, MI in early spring 2012. Foundresses were housed in isolation in individual cages using established methods (Daugherty et al., 2011) for one to three weeks with access to water and sugar ad libitum prior to visual training. We used a negative reinforced T-maze (previously described in Sheehan and
Tibbetts (2011)) to train wasps on two different sets of stimuli: simple patterns (n = 12 for *P. fuscatus*, n = 10 for *P. metricus*) or *P. fuscatus* faces (n = 30 for *P. fuscatus*, n = 10 for *P. metricus*). An electric current (~2-4 V) ran throughout the entire length of the T-maze except for a safety chamber, which was associated with one of two images. The height of the T-maze was very low, so that the wasp had to walk through the maze and could not fly to get away from the electric shock. After we placed the wasp in the T-maze, we opened a sliding door that allowed the wasp to enter a decision chamber. In order to avoid being shocked, the wasp had to learn the image associated with the safety chamber. Learning was measured by recording the first choice chamber that an individual wasp enters. Each wasp was trained in 40 sequential trials performed on the same day. Thirty minutes after the 40th test, the wasps were freeze-killed in liquid nitrogen and stored at -80°C.

**Statistics**

We tested whether each wasp species was able to learn each visual stimulus by determining if the total number of stimuli correctly identified out of all 40 trials was greater than expected by random chance (20 correct) using the 't.test' function from the 'stats' package in R (Team, 2013). Additionally, we tested for differences in the total number of stimuli correctly identified (out of all 40 trials) between species and stimuli interactions using the 'HSD.test' function from the 'agricolae' package in R. For the Tukey Honest Significant Difference (HSD) test, we first modeled the total correct by species, stimuli and their interaction using a generalized linear model with a poisson family.

**Brain Dissections**

We removed the heads from the wasp body in a small petri dish on dry ice, and then stored the head in 1.5 mL microcentrifuge tubes at -80°C. We cleaned all tools (small petri dish, castroviejo scissors, dissecting tweezers, and a dissecting needle with spear head) with RNase Away prior to dissections. During the dissection, the brain was submerged in 100% ethanol solution with dry ice in a small petri dish placed in a larger container of dry ice. We dissected out the
brain by removing surrounding cuticle, muscle tissue, ocelli, compound eyes, glandular tissue, and subesophageal ganglion.

**RNA Extraction**

We extracted total RNA from individual whole wasp brains using an RNeasy Mini Kit (Qiagen) with 350 μL of Buffer RLT + β-mercaptoethanol, as recommended for tissue size < 20mg. As part of the extraction, a DNase I treatment (Qiagen) was applied to remove DNA contamination. We performed the final elution in two steps with 15 μL of RNase-free water per step for a final elution volume of 30 μL. To verify the quality of the extracted RNA, we analyzed the samples using a spectrophotometer (NanoDrop 2000) and Bioanalyzer (RNA 6000 Nano Kit, Agilent). We selected representative wasps for RNA sequencing based on RNA concentration, RNA integrity number (RIN), and training score (n = 4 per visual stimuli per species, except n = 5 for *P. fuscatus* trained to faces).

**RNA Sequencing**

BGI Americas at the University of California – Davis prepared mRNA Seq libraries for the 17 foundress and 16 additional worker (8 per species) samples with Illumina’s “TruSeq RNAseq Sample Prep kit”, which included Poly(A) RNA purification, fragmentation using sonification, cDNA synthesis using random primers, and bar-coding for each of the 33 samples. For the purposes of the current study, the worker samples were only used for transcriptome assembly and a more detailed description of these data will be published separately (A.J. Berens et al. unpublished data). Four lanes of the sequencer (HiSeq 2000, Illumina) were used to generate over 1.5 billion 100 base paired-end reads with one sample from each treatment and species in each lane (except lane had 2 samples of *P. fuscatus* trained to faces). Raw sequence data has been deposited to the National Center for Biotechnology Information’s (NCBI) Short Read Archive (BioProjects: PRJNA287145 (*P. fuscatus*) and PRJNA287152 (*P. metricus*)).
Transcriptome Assembly

Before transcriptome assembly, we removed extraneous adapter sequences and filtered for quality using Trimmomatic (v. 0.32; ILLUMINACLIP:TruSeq2-SE.fa:2:40:15 LEADING:3 TRAILING:3 SLIDINGWINDOW:5:20 MINLEN:40) (Bolger et al., 2014). For each species, we assembled the groomed transcriptomic short reads de novo using Trinity (v. r20140717, Grabherr et al. (2011)) to produce separate P. fuscatus and P. metricus transcriptomes. To assemble the most complete possible P. metricus transcriptome, we included all 16 foundress and worker visual training libraries (described above or to be published separately (A.J. Berens et al. unpublished) as well as 16 additional P. metricus libraries previously described in (Berens et al., 2015a,b). After assembly, the final Trinity-produced Transcriptome Shotgun Assemblies (TSA) were annotated with the mRNAmarkup protocol (Volker Brendel, unpublished; code available from http://brendelgroup.org/bioinformatics2go/mRNAMarkup.php). mRNAmarkup splits potential chimeric assemblies and discards likely contaminants identified as sequences with strong similarities to UniVec (http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html) entries or E. coli genomic sequences. We identified potential full-length mRNAs within each TSA using MuSeqBox (Xing and Brendel, 2001) as distributed with mRNAmarkup (parameter option –F 5 5 10 10 90 60). Each species final TSA was annotated with the most significant BLASTp hit using stringent criteria (MuSeqBox option –A 45 10 75). For each of these two final cleaned-up and annotated TSA assemblies, we assessed quality and completeness using the CEGMA (Core Eukaryotic Gene Mapping Approach, Version 2.5) method (Parra et al., 2007). These Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accessions: GDFS00000000 (P. fuscatus, first version: GDFS01000000) and GDHQ00000000 (P. metricus, first version: GDHQ01000000).

Read Mapping, Abundance Estimates, and Differential Expression

For each species, we aligned the paired-end reads to the corresponding TSA using Bowtie2 (v. 2.2.4, Langmead and Salzberg (2012)) and quantified the abundances of the transcripts for each library using eXpress (v. 1.5.1, Roberts and Pachter (2013)). The raw read counts
have been deposited in NCBI’s Gene Expression Omnibus (Edgar et al., 2002). We used the R statistical program (v. 3.0.1, Team (2013)) to test for differential brain gene expression between pattern and face (FDR < 0.1, Benjamini and Hochberg (1995)) trained groups for each species separately using the DESeq2 (v. 1.6.3, Love et al. (2014)) statistical package from the Bioconductor repository (Gentleman et al., 2004). Resulting differentially expressed transcripts (DETs) between face and pattern groups are hereinto referred to as “visual stimuli-related DETs”.

Transcripts with low read counts (less than one read per sample on average) or outlier samples (Cook’s distance in the 99% quantile of $F(p,p-m)$ distribution where $p =$ number of parameters, $m =$ number of samples) were removed from the differential expression analysis. For P. fuscatus, 69,092 (44.5% of the assembled transcriptome) transcripts were tested for differential expression between pattern and face trained wasps using DESeq2. The rest of the transcriptome was not tested because of low read counts across samples (54.5% of the assembled transcriptome) or outlier samples (1% of the assembled transcriptome). However, most of the reads (96.8%) mapped to the transcripts tested for the P. fuscatus differential expression analysis. Sixty percent (76,146 transcripts) of the transcriptome was excluded from the P. metricus differential expression analysis because of low read counts across samples (59.7%) or outlier samples (0.3%). Yet, almost all (99.2%) of the reads mapped to the 50,613 transcripts included in the P. metricus differential expression analysis.

Gene Ontology (GO)

We used Blast2GO (Conesa et al., 2005) to assign GO terms to P. fuscatus and P. metricus transcriptomic sequences based on best BLASTx hit (E-value < 1e-3) to Drosophila melanogaster sequences in the sequences in NCBI Entrez Protein database. For each species, Blast2GO was used to assess enrichment of GO terms using a Fisher’s Exact Test, with a FDR < 0.1, (Benjamini and Hochberg, 1995) for differentially expressed transcripts (DETs) compared to the set of all transcripts tested in the differential expression analysis.
Candidate Individual Memory Recall Genes

We were interested in determining whether individual recognition candidate genes (Chapter 4) were also differentially expressed between visual stimuli groups in the current study. We identified these candidate genes in the \textit{P. fuscatus} and \textit{P. metricus} transcriptomes using Blastn with an E-value < 1e-4.

\textbf{Polistes-specific Transcripts}

We defined \textit{Polistes}-specific transcripts for each species as transcripts that did not have a Blastx hit (E-value < 1e-4) to any Hexapoda proteins (Taxon id: 6960) from the NCBI RefSeq database but had a Blastn hit (E-value < 1e-4) to at least one other \textit{Polistes} transcriptome (\textit{P. canadensis} \cite{Ferreira2013}, \textit{P. dominula} (D.S. Standage et al., unpublished), \textit{P. fuscatus} (this study), and \textit{P. metricus} (this study)). The reason for doing so was to identify transcripts that are restricted to the \textit{Polistes} social lineage, which may be important for the evolution of novel social traits as proposed by the novel taxonomically restricted genes hypothesis. Under this hypothesis, we predicted that \textit{Polistes}-specific transcripts would be more likely to be differentially expressed between visual stimuli groups for \textit{P. fuscatus}. So for each species, we performed a Chi-squared test with Yates’ correction to test for an over-abundance of \textit{Polistes}-specific DETs given the number of DETs identified, the number of \textit{Polistes}-specific transcripts, and the total number of transcripts tested for differential expression.

5.3 Results

\textbf{Visual Training Experiments}

\textit{P. fuscatus} wasps were equally able to distinguish between pattern and facial stimuli in the negatively reinforced T-maze (Tukey HSD post-hoc test, p-value > 0.05; Figure 5.1), which does not agree with previous studies on facial cognitive specialization for \textit{P. fuscatus} \cite{Sheehan2011}. However, we did confirm that \textit{P. fuscatus} has better facial recognition abilities than \textit{P. metricus} \cite{Sheehan2011}, since \textit{P. metricus} training scores are significantly lower than \textit{P. fuscatus} for both sets of stimuli (Tukey HSD post-hoc tests, p-values
Amongst training stimuli, *P. metricus* wasps correctly identified more patterns than faces (Tukey HSD post-hoc test, p-values < 0.05). Both species showed signs of learning the visual stimuli from the negatively reinforced T-maze, except *P. metricus* wasps trained to facial stimuli which did not perform significantly better than expected by random chance (20 stimuli correctly identified; t-test, p-value = 0.365). Thus, we are comparing between visual pattern vs. face learning for *P. fuscatus* (hereinto referred to as “visual stimuli-related” learning) and between visual pattern learning vs. no learning for *P. metricus* (hereinto referred to as “visual learning”).

**Transcriptome Assemblies**

After *de novo* transcriptome assembly using Trinity and filtering out chimeric and contaminant sequences using the mRNAmarkup protocol, the final *P. fuscatus* Transcriptome Shotgun Assembly (TSA) consisted of 157,691 transcripts with an N50 of 1982 base pairs. To address quality and completeness, we assessed the *P. fuscatus* transcriptome assembly using the CEGMA (Core Eukaryotic Gene Mapping Approach) method (Parra et al., 2007). 87.50% of the CEGs mapped completely and 93.15% of the CEGs mapped partially to the *P. fuscatus* TSA, suggesting a very complete representation of expressed genes. The final *P. metricus* TSA consisted of 127,674 transcripts with an N50 of 1829 base pairs, and was also very complete with 84.27% CEGs mapping completely and 94.35% of the CEGs mapping partially.

**Differential Expression**

We identified 64 *P. fuscatus* brain “visual stimuli-related” DETs between pattern vs. face trained groups of wasps (FDR < 0.1, Benjamini and Hochberg (1995)), of which 44% (28 transcripts) were up-regulated in the face trained group (Figure 5.2a). For *P. metricus*, we identified 270 transcripts that were differentially expressed between visual pattern trained and not trained (face stimuli) wasps. Sixty-four percent (172 transcripts) of these *P. metricus* “visual learning-related” DETs were up-regulated in the not trained (face stimuli) wasps compared to visual pattern-trained wasps (Figure 5.2b).
None of the DETs are shared between *P. fuscatus* and *P. metricus*. There is a low, yet significant, correlation (Pearson correlation = 0.0326, df = 45,198, p-value = 4.06e-12) between expression patterns in the two species (log fold change between pattern vs. face stimuli) for all shared transcripts between these two species (based on putative homologs, best tBLASTx hit with E-value < 1e-4; Figure 5.2c). Examining only the transcripts that are differentially expressed in either species, there is no significant correlation in the log fold changes between *P. fuscatus* and *P. metricus* (Pearson correlation = -0.069, df = 274, p-value = 0.250; Figure 5.2c).

**Gene Ontology**

On the level of gene functional categories, only 4 GO terms [olfactory receptor activity (GO:0004984), sensory perception of smell (GO:0007608), G-protein coupled receptor signaling pathway (GO:0007187), and sensory perception of chemical stimulus (GO:0007606)] were significantly enriched in the *P. fuscatus* visual stimuli-related DETs. Interestingly, all of the *P. fuscatus* DETs associated with these GO terms were up-regulated in face compared to pattern trained groups. For *P. metricus* visual learning-related DETs, there are 25 significantly enriched GO terms [including gene expression (GO:0010467), ribosome biogenesis (GO:0042254), and translation (GO:0006412)]. For all of these *P. metricus* enriched GO terms, the majority (between 85-100% per GO term) of the associated DETs were down-regulated during visual pattern training. None of the enriched GO terms are shared between *P. fuscatus* and *P. metricus*.

**Candidate Individual Memory Recall Genes**

In Chapter 4, we identified four genes (IP3K, IP3R, Nckx30C, and Su(var)2-10) that were significantly down-regulation during memory recall of staged social encounters in *P. fuscatus* and were considered to be candidates for individual memory recall. Of these candidates, only *Nckx30C* was significantly differentially expressed in the current study between visual pattern learning vs. no learning (face stimuli) for *P. metricus*. Yet, this gene was not differentially expressed during the staged social encounters (Chapter 4) for this species (only for *P. fuscatus*).
None of the memory recall candidate genes were classified as differentially expressed for *P. fuscatus* in this study. In *P. metricus*, IP3K and IP3R had lower expression during memory recall and no learning (Figure 5.3). Only IP3K and Nckx30C showed consistent expression directionality (but still not significant in RNA-Seq study) across studies for *P. fuscatus* with down-regulation in individual memory recall and facial learning.

**Polistes-specific Transcripts**

We were interested in testing whether genes underlying the complex social behavior of facial recognition have evolved within the *Polistes* social lineage as proposed by the novel taxonomically restricted genes hypothesis. We identified 23,737 *Polistes*-specific transcripts in the *P. fuscatus* transcriptome, of which only 26 were differentially expressed between visual stimuli groups. We did not detect a significant number of *Polistes*-specific transcripts that were differentially expressed between face and pattern groups in *P. fuscatus* (Chi-squared test with Yates’ correction, p-value = 0.354). For *P. metricus*, 35% (95) of the 270 DETs were classified as *Polistes*-specific (total of 18,098 *Polistes*-specific transcripts), which was not different than expected (Chi-squared test with Yates’ correction, p-value = 0.888).

### 5.4 Discussion

This study is the first genome-wide assessment of gene expression associated with facial recognition in any animal. This complex social behavior has been challenging for mechanistic analysis, and we provide valuable data on the molecular correlates of this behavior using comparative transcriptomics of *Polistes* paper wasps, the only known insect taxon to possess facial recognition. In this study, we identified brain gene expression changes between wasps in the process of facial learning compared to pattern learning for two species of paper wasps (*P. fuscatus* – has facial recognition and *P. metricus* – does not have facial recognition). For *P. fuscatus*, there were relatively few differences in brain gene expression (64 DETs) related to visual stimuli learning, and surprisingly, several transcripts have putative functions related to olfaction. In *P. metricus*, there were slightly more differentially expressed transcripts correlated with visual learning (270 DETs) and several of these are associated with ribosomal biosynthesis
and translation. Of the *P. fuscatus* visual stimuli-related DETs, few are *Polistes*-specific transcripts, which does not support the idea that novel transcripts are important for the evolution of the facial recognition in paper wasps. Instead, our data suggest conserved genetic mechanisms, possibly some with ancestral functions in olfaction, may have been recruited for the purpose of recognizing individual faces in *P. fuscatus*.

The main reasons we suggest that genes associated with insect olfactory response (i.e. based on known functions from *Drosophila*) may have been co-opted for facial learning in *P. fuscatus* are as follows: 1) the visual stimuli enriched GO terms are related to olfaction function and 2) all of the DETs associated with these olfactory GO were up-regulated during facial compared to pattern learning. Intriguingly, evidence from other organisms also indicates a connection between visual recognition and olfaction. A few hours after giving birth, sheep ewes are able to recognize their lambs based solely on olfactory cues (Ferreira et al., 2000; Terrazas et al., 1999). Within several weeks, ewes are also able to recognize their lambs’ faces (Kendrick et al., 1996), although visual connections are already forming within the first 12 hours (Ferreira et al., 1999, 2000; Terrazas et al., 1999). During olfactory recognition a few hours postpartum, the sheep’s brain activity increases not only in regions of the brain associated with olfaction but also visual memory centers, which suggests some integration between olfactory and visual recognition (Broad et al., 2002). Potentially, these results indicate that neural circuitry already formed for olfactory recognition also function in visual recognition (Broad et al., 2002). The gene expression data from this study suggests that there may also be a similar connection between visual and olfactory recognition in paper wasps, and this open question deserves further investigation in the future.

The behavioral results from this study provide some confirmation of the results from previous studies – most importantly, they verified that *P. fuscatus* has a superior ability to recognize faces compared to *P. metricus*. However, the results from this training assay did not completely confirm previous observations from Sheehan and Tibbetts (2011) for both species; therefore, we discuss caveats and possible limitations of our study here. Overall, the training scores were lower than previously documented (Sheehan and Tibbetts, 2011), and this difference may be the result of changes to the floor matting of the T-maze, which was less conductive of the
electrical current in this study (E.A. Tibbetts, unpublished data). We will discuss the results for each of the two species in turn. First, for *P. fuscatus*, the wasps were able to learn both types of visual stimuli, but *P. fuscatus*’ learning of faces in the current study was not better than learning non-face patterns as in previous studies (Sheehan and Tibbetts, 2011), so we could not confirm that facial learning was a cognitive specialization. Thus, expression differences we describe for *P. fuscatus* are still valid as candidates for learning of facial stimuli, but the fact that facial learning was not as strong as in previous experiments suggests that we may have missed some expression differences associated with the cognitive specialization of facial recognition in this species. Second, for *P. metricus*, even though this species does not naturally use facial features to recognize conspecifics due to little facial color variation (Sheehan and Tibbetts, 2010), *P. metricus* have previously been shown to have some, but more limited, ability to learn the variable faces of *P. fuscatus* (Sheehan and Tibbetts, 2011). In the current study, it appears that *P. metricus* did not learn faces at all – that is, they did not perform significantly better than expected by random chance. In the current study, we were able to confirm that *P. metricus* were able to learn non-facial pattern stimuli. Thus, the gene expression results for *P. metricus* may not reflect mechanisms associated with facial learning *per se*, but rather a more generalized form of visual learning. The fact that we found different learning patterns in the two species may also help explain why we found no overlap in gene expression patterns or gene functions (Gene Ontology) between the two species, i.e. we observed gene expression related to different visual learning phenomena in the two species.

Individual facial recognition is a complex behavior that requires multiple cognitive tasks. In this study, we assayed the facial learning component of facial recognition – the point when individuals recall whether they recognize previously encountered facial stimuli. Previously, we looked at individual memory recall (Chapter 4), which is the process of individual identity retrieval from brain storage. This identity information is not necessarily specific to faces, but could consist of other variable features. Previous results in Chapter 4 suggest that calcium signaling may be important for individual facial recognition in *P. fuscatus*; however, we did not observe any brain gene expression differences of individual memory recall or calcium signaling candidate genes during visual training. Additionally, there was no enrichment of calcium-related
GO terms based on the *P. fuscatus* visual stimuli-related DETs. Thus, these results suggest that these two processes of individual recognition utilize different molecular mechanisms.

5.5 Conclusions

Only a few highly social species, including a species of *Polistes* paper wasps, possess the ability to recognize conspecific individuals using faces, and little is known about the molecular underpinnings of this complex social behavior. This is the first study to look at genome-wide gene expression associated with facial learning in any species that has facial recognition. Our results suggest that there are few neural expression changes during facial learning compared to pattern learning. Some of these genes with expression changes are known to function in olfaction in *Drosophila*, which does not have individual facial recognition. Thus, there appears to be a functional connection between olfaction and visual facial recognition. This work highlights the power of the *Polistes* system as a comparative model for understanding the molecular underpinnings of individual facial recognition.

5.6 Acknowledgements

We would like to thank Elizabeth Tibbetts for her collaboration on this project. She graciously sent the wasp samples for this gene expression study. Her students Mike Sheehan collected the wasps and Andrew Madagame performed the training assays. She also provided valuable insights into the interpretation of the behavioral data. Funding for this work was provided by the National Science Foundation (NSF) through a Doctoral Dissertation Improvement Grant (IOS-1311512).
Figure 5.1 Training ability for wasps based on visual stimuli (general patterns and *P. fuscatu*s faces), species (*P. fuscatu*s and *P. metricu*s), and their interaction. There is no difference between stimuli for *P. fuscatu*s, but there is a difference between stimuli for *P. metricu*s (Tukey HSD posthoc tests, p-values < 0.05). *P. metricu*s training ability is lower than *P. fuscatu*s for both stimuli.
Figure 5.2  Heat map of the relative expressions (sample read counts scaled by library size then across each gene) for the A) 64 *P. fuscatus* and B) 270 *P. metricus* differentially expressed transcripts between pattern and face stimuli groups by sample (DESeq2, FDR < 0.1). Pattern training up-regulated transcripts are highlighted in yellow and transcripts up-regulated during face training (*P. fuscatus*)/no training to face stimuli (*P. metricus*) are indicated by the blue color. None of the differentially expressed transcripts are shared between *Polistes* species. C) Log fold changes of pattern vs. face stimuli groups for putative homologs between *P. fuscatus* and *P. metricus* (best tBLASTx hit with E-value < 1e-4). *P. fuscatus* differentially expressed homologs are indicated by orange triangles, *P. metricus* differentially expressed are shown as blue squares, and all other homologs are not differentially expressed (grey circles).
**Figure 5.3** Individual recognition candidate genes from Chapter 4 with corresponding mean log2 fold changes during visual stimuli-related learning (*P. fuscatus*) / visual learning (*P. metricus*) from this Chapter (negative fold change (red): higher expression in face group, positive fold change (green): higher expression in pattern group). Four genes (*IP3K, IP3R, Nckx30C, and Su(var)2-10*) are candidates for individual memory recall, and all showed down-regulation during memory recall (green) compared to no social interaction and initial memory formation. None of these genes were differentially expressed between face vs. pattern groups in either paper wasp species (*P. fuscatus* or *P. metricus*), except *Nckx30C* (indicated by *) which was up-regulated in face group (no learning) compared to pattern learning for *P. metricus*. For most genes, there were multiple transcripts with Blast hits [E-value < 1e-4] to the gene sequences, so the log2 fold changes were averaged across all transcripts. *mGluR* was not found in the *P. metricus* transcriptome.

<table>
<thead>
<tr>
<th>Gene</th>
<th><em>P. fuscatus</em> Individual Memory Recall Candidate log2 Fold Change</th>
<th><em>P. metricus</em> Individual Memory Recall Candidate log2 Fold Change</th>
<th><em>P. fuscatus</em> Visual Stimuli Mean log2 Fold Change</th>
<th><em>P. metricus</em> Visual Stimuli Mean log2 Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ace</td>
<td>-0.063</td>
<td>0.162</td>
<td>0.287</td>
<td>0.009</td>
</tr>
<tr>
<td>Bap60</td>
<td>0.072</td>
<td>0.009</td>
<td>0.236</td>
<td>0.173</td>
</tr>
<tr>
<td>Cha</td>
<td>-0.100</td>
<td>0.162</td>
<td>-0.109</td>
<td>0.053</td>
</tr>
<tr>
<td>dor</td>
<td>0.030</td>
<td>-0.109</td>
<td>0.287</td>
<td>-0.050</td>
</tr>
<tr>
<td>gogo</td>
<td>-0.069</td>
<td>0.016</td>
<td>0.028</td>
<td>0.016</td>
</tr>
<tr>
<td>Gug</td>
<td>0.099</td>
<td>-0.050</td>
<td>-0.028</td>
<td>0.053</td>
</tr>
<tr>
<td>IP3K</td>
<td>Down regulated</td>
<td>0.282</td>
<td>Down regulated</td>
<td>0.516</td>
</tr>
<tr>
<td>IP3R</td>
<td>Down regulated</td>
<td>-0.014</td>
<td>Down regulated</td>
<td>0.251</td>
</tr>
<tr>
<td>mGluR</td>
<td>0.008</td>
<td>N/A</td>
<td>0.089</td>
<td>0.105</td>
</tr>
<tr>
<td>N</td>
<td>0.089</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nckx30C</td>
<td>Down regulated</td>
<td>0.156</td>
<td></td>
<td>0.527*</td>
</tr>
<tr>
<td>Nmdar1</td>
<td>0.073</td>
<td></td>
<td></td>
<td>0.234</td>
</tr>
<tr>
<td>PsGEF</td>
<td>0.066</td>
<td></td>
<td></td>
<td>0.173</td>
</tr>
<tr>
<td>sca</td>
<td>0.171</td>
<td></td>
<td></td>
<td>0.036</td>
</tr>
<tr>
<td>Stau</td>
<td>-0.142</td>
<td></td>
<td></td>
<td>0.035</td>
</tr>
<tr>
<td>Su(var)2-10</td>
<td>Down regulated</td>
<td>-0.030</td>
<td></td>
<td>-0.026</td>
</tr>
<tr>
<td>Syt7</td>
<td>-0.194</td>
<td></td>
<td></td>
<td>0.244</td>
</tr>
</tbody>
</table>
CHAPTER 6. GENERAL CONCLUSIONS

Sociogenomics is an emerging field that harnesses the power of genomics to understand the evolution and regulation of social living. My dissertation work is an important contribution to the field of sociogenomics for several reasons: 1) it has transformed an important ecological and evolutionary model system – *Polistes* paper wasps – into a viable system for studying genomics, 2) it exemplifies the power of a comparative approach to address major outstanding sociogenomics questions, 3) by studying complex behavior in an ecological context, this work emphasizes the multifactorial effects of the environment on developmental plasticity of social phenotypes, and 4) this research demonstrates the interconnectedness of genetic mechanisms, not single genes, to produce complex behavioral phenotypes. In the following paragraphs, I summarize the contributions of my thesis to each of the four aforementioned themes.

Genomics resource development for a classic behavioral model system

Insects in the order Hymenoptera (ants, bees, and wasps) have moved to the forefront of sociogenomics research. Nearly all previous sociogenomics studies in social insects have focused on ants and bees, leaving a large gap for the third major social hymenopteran lineage, wasps in the family Vespidae (reviewed in (Jandt and Toth, 2015; Rehan and Toth, 2015)). Yet, vespid wasps are an excellent group for studying transitions during the evolution of social behavior because species of these wasps exhibit a wide range of social behaviors: from solitary to primitively social to highly eusocial (reviewed in Jandt and Toth (2015)). Of all vespids, paper wasps in the genus *Polistes*, which have an intermediate form of sociality where reproductive castes remain flexible into adulthood, are one of the best studied (Jandt and Toth, 2015). Due to this caste flexibility, paper wasps societies are characterized by both cooperation and competition; however to reduce the cost of conflict between nestmates, paper wasps establish social dom-
inance hierarchies that dictate ranking order for individual reproductive rights (Jandt et al., 2014; Pardi, 1948). Since the discovery of social dominance hierarchies in paper wasps by Pardi (1948), *Polistes* have been and remain an important model for understanding the maintenance of dominance behavior (reviewed in Jandt et al. (2014)). As an important ecological and evolutionary model system, applying sociogenomics approaches to *Polistes* paper wasps provides new opportunities for studying complex social behaviors and social evolution.

As part of my dissertation work, I have developed new transcriptomic resources for two paper wasp species, which were used to test hypotheses about the evolution of social behavior and mechanisms underpinning complex social behaviors. For the temperate paper wasp species *P. metricus*, I produced two *de novo* transcriptome assemblies: the first for larval head tissue and second for the brains of adult wasps. With the *P. metricus* larval transcriptome, I identify caste developmental gene expression (Chapter 2) that was used to examine 1) genetic mechanisms important for the evolution of insect castes (Chapter 2) and 2) nourishment as an environmental factor affecting caste (Chapter 3). From RNA-sequencing reads of individual adult *P. fuscatus* brains, I also assembled a *P. fuscatus de novo* transcriptome. Comparing adult brain gene expression between the two *Polistes* species, I studied the genetic mechanisms underpinning facial recognition (Chapter 5).

**Insights from comparative sociogenomics**

I conducted the most comprehensive examination to date of conserved and novel genetic mechanisms regulating caste development across independent evolutionary lineages of social behavior in Chapter 2. The results from the comparison of three major lineages of sociality (fire ants, honey bees, and paper wasps) support the hypothesis that the convergently evolved and quintessential eusocial trait – queen and worker castes – involves some conserved molecular mechanisms that are shared across lineages (Toth and Robinson, 2007). These results bear on the “genetic toolkit” hypothesis from evolutionary developmental biology (evo-devo) by suggesting that “tools” are relatively loose, i.e. convergent phenotypes are produced by different, lineage-specific modifications within conserved genetic networks (Abouheif and Wray, 2002; Rudel and Sommer, 2003). I also tested an alternative, but not necessarily mutually exclu-
sive, hypothesis that “novel” taxonomically restricted genes are important for caste differences within the paper wasp lineage (Ferreira et al., 2013; Johnson and Tsutsui, 2011). Comparing two species of paper wasps, I did not find support for this hypothesis; thus, the role of “novel” genes during the evolution of sociality remains inconclusive. One plausible explanation for these seemingly contradictory results is that there were slight modifications in the expression of genetic “tools” during the origins of sociality, but with the evolution of lineage specific adaptations, novel genes play an increasingly important role for derived social phenotypes (Rehan and Toth, 2015). Chapter 2 is an illustrative example of how new genomic resources, in a comparative perspective, can be a powerful tool for studying the evolution of sociality and other outstanding sociogenomics questions.

In Chapter 5, I am the first to use genome-wide transcriptional profiling to study facial recognition, a very complex social behavior that has previously evaded mechanistic analysis. Using a comparative transcriptomics approach, I identify brain gene expression changes associated with learning faces compared to patterns in species with and without the ability to recognize faces (P. fuscatus and P. metricus respectively). As mentioned above, “novel” taxonomically restricted genes have been hypothesized to be important for the evolution of novel social traits (Harpur et al., 2014; Johnson and Tsutsui, 2011; Simola et al., 2013), but I do not find support for this idea based on gene expression changes associated with facial recognition in these two paper wasps. Instead the results suggest that conserved genetic mechanisms, known to function during olfactory recognition in Drosophila, have been recruited for the purpose of recognizing individual faces in P. fuscatus. Chapter 5 is an important first step in pinpointing the genetic mechanisms underlying facial recognition, and illustrates the power of comparative sociogenomics with the Polistes system for advancing our understanding of molecular mechanisms regulating complex social behavior.

The dynamic genome: environmental factors regulating developmental plasticity of social behaviors

Developmental plasticity is an important adaptive mechanism by which phenotype within a single genotype can be adjusted to environmental conditions during development. Most ge-
omics studies have focused on developmental plasticity of morphological traits such as seasonal variation in butterfly eyespot size (Oliver et al., 2013) and horn development in beetles (Kijimoto et al., 2014); while few have looked at how early developmental changes influence adult behavior. I address this inquiry in Chapter 3 by testing the hypothesis that nutritional environment drives behavioral caste bias in paper wasps (Chapter 2, Hunt et al. (2007, 2010)). In Chapter 3, using a comparative transcriptomics approach between natural caste and nourishment-manipulated development, I found that limited nourishment partially biases genes expression toward worker-like expression patterns (identified in Chapter 2), but does not account for all variability. Thus, it is likely that nourishment, in conjunction with other social inputs, is an important determinant of developmental caste bias in paper wasps. Chapter 3 illustrates the multifactorial nature of the environmental effects on the developmental plasticity of social phenotypes.

Beyond single genes: the importance of gene interactions

There has been a long history of interest in understanding the genetic underpinnings of behavior. Until recently, behavioral genetics studies have focused on mutations or expression differences in single genes. For example, allelic variation of the foraging (for) gene in Drosophila influences larval foraging behavior such that rovers (for\textsuperscript{R}) traverse longer paths in search of food compared to sitters (for\textsuperscript{s}) (Osborne et al., 1997; Sokolowski, 1980). Analysis of the for gene in honey bees (Ben-Shahar et al., 2002) and C. elegans (Fujiwara et al., 2002) shows a conserved relationship between this gene and food-related behaviors across species. In honey bees, for brain gene expression increases during the ontogenic transition of adult worker division of labor from performing hive tasks (nurses) to foraging outside the nest for pollen and nectar (foragers) (Ben-Shahar et al., 2002). However, when taking a genome-wide approach, for is just one of almost three thousand genes differentially expressed between the nurse and forager brains (Whitfield et al., 2003), which suggests that complex behavioral traits are produced by multigenic changes where each individual gene has a relatively small effect.

My dissertation provides additional support for multigenic associations with complex social behaviors, and also suggests that these genes are connected through a few key pathways. From a comparative analysis of queen/worker caste determination across three social lineages in Chapter
I identified greater similarity on the level of metabolic pathways and biological functions than at the gene level. These results highlight the flexible nature of genetic mechanisms to produce strikingly similar behavioral caste phenotypes. My results from Chapter 4 also support this notion. In this chapter, the results point toward calcium signaling as being an important biological process regulating individual memory recall in paper wasps. Thus, despite the fact that individual candidate genes were selected, a biological process, rather than any single gene, emerged as being important in this complex form of social behavior. Thus, these results suggest that we need to move beyond single genes in order to understand how interactions of many genes influence social behaviors.

**Current Sociogenomics Challenges**

From my dissertation work, I see a number of challenges that still remain for the field of sociogenomics. One challenge will be the development of new social genetic model organisms. Currently, functions of genetic mechanisms of social behaviors are inferred by comparisons to solitary organisms such as *Drosophila*. However, to truly understand these molecular mechanisms in a social context, it is important that functional information is derived from a variety of other social organisms. This is now becoming more feasible with sociogenomics, advances in gene silencing and germ line transformation techniques, and their application to a wider variety of species.

Another challenge will be to better characterize and understand the role of “novel” genes, i.e. without homologues. The “novel gene” hypothesis for the evolution of sociality is an emerging hypothesis with increasing evidential support (Ferreira et al., 2013; Harpur et al., 2014; Johnson and Tsutsui, 2011) and contradiction (Chapter 2, Simola et al. (2013)). Unfortunately, novel genes are poorly defined as previously undescribed genes that have no significant homology with known sequences (Ding et al., 2012). Even though improved taxonomic sampling has reduced the overall percentage of novel genes identified per genome, there is still a growing number novel genes, especially as new taxonomic branches are sequenced (Wilson et al., 2005). Typically homology is detected using Blast; however, taxonomically restricted genes are likely to be fast-evolving leading to difficulties detecting homology with blast and misclassification as novel (Albà
and Castresana, 2007; Elhaik et al., 2006). Thus, as genomic resources and gene discovery tools are rapidly advancing and changing, the definition and detection of novel genes also changes (Khalturin et al., 2009). This ever changing classification of novel genes is problematic for testing their importance in social evolution; thus, a more systematic approach for identifying all novel genes within a lineage is needed, such as phylostratigraphy which groups genes based on their phylogenetic origin in order to uncover important adaptive events in evolution (Domazet-Lošo et al., 2007). However, there is new evidence that phylostratigraphy underestimates the age of many genes, especially rapidly-evolving or short genes (Moyers and Zhang, 2015), so this approach needs to be revised for proper evolutionary inferences.

Another challenge will be to determine how to use transcriptomics to capture functionally relevant expression associated behavioral changes. As of now, it is unclear the extent to which gene expression changes influence behavior (vs. the result of the changes in behavior) and how long behavior-related gene expression changes are maintained. It is likely that some behaviors are influenced more by fast acting genes (such as immediate early genes) while others behaviors are sustained by late response genes (Clayton, 2000; Fernald and Maruska, 2012; Robinson et al., 2008). Therefore, it is important for researchers to consider this dynamic nature of gene expression when designing experiments and making inferences about behaviors.

Another major challenge I encountered was the highly variable nature of transcriptome data. Inter-individual variation in gene expression – which arises from differences in genetic background (including regulatory elements)(Øverli et al., 2007), epigenetic factors (Turan et al., 2010), and ecological context (Rey et al., 2013) – is found in every genome-wide expression study. Traditional approaches try to minimize this variation by using controlled genetic backgrounds; however, this could result in misleading conclusions (Crawford and Oleksiak, 2007). Instead the inter-individual variation needs to be accounted for during differential expression analyses, but this is not the focus of statistical models for differential expression analyses because individuals within a group are considered as replicates (Crawford and Oleksiak, 2007; Whitehead and Crawford, 2006). The identification of biologically relevant variation in genome-wide expression studies remains challenging because there are typically too few individuals per group and variation is context dependent (e.g. tissue-specific (Whitehead and Crawford, 2005) or
species-specific (Whitehead and Crawford, 2006)). Thus, significant correlations between gene and phenotype are only discernable within a specific context (Crawford and Oleksia, 2007).

**Sociogenomics Future Directions**

The insights from my work warrant additional studies in several areas, such as the following: 1) the interactions between different types of genetic mechanisms in influencing behavioral variation (e.g. regulatory vs. coding variants), 2) the multifactorial effects of the environment on gene expression and phenotypes, and 3) the evolution of sociality and derived social behaviors using comparative approaches. For example, there is great deal of variability in individual facial learning abilities even within the same species, so a crucial next step will be to determine the complex genetic mechanisms controlling this variability. One possible source of variability is single nucleotide polymorphism (SNPs) between individuals of the same species. It would also be interesting to compare SNPs across paper wasps species, in order to determine whether there are genes under positive selection in *P. fuscatus* compared to all other *Polistes* species, which could be important for facial learning. To help illuminate the multifaceted environmental mechanisms regulating caste bias in paper wasps, work to investigate the interaction of nourishment and the social environment is needed and already underway in the Toth laboratory. More comparative work is needed to illuminate the relative roles of conserved and novel mechanisms underpinning social evolution. Future studies on conserved genetic mechanisms should incorporate a greater number of (possibly all) social lineages and close solitary relatives. At the same time, examining multiple species within the same lineage is needed to address the alternative hypothesis that novel mechanisms are important for the evolution of the worker caste (Rehan and Toth, 2015).

In summary, my dissertation research has advanced the field of sociogenomics by developing new genomic resources for an ecological and evolutionary model system – *Polistes* paper wasps – to address questions about genetic and environmental mechanisms underpinning social behaviors and the evolution of sociality.
Figure A.1  Heat maps of the scaled sample read counts with best BLAST hit to the 16 caste-related genes previously identified in Hunt et al. (2010). Samples listed as GroupID-SampleID. GroupIDs: FR - foundress-reared larvae, and WR - worker-reared larvae. * indicates significant differential expression between castes using DESeq.
Table A.1  Directionality (up-regulation in queens [Q] or workers [W]) of differentially expressed orthologs between pairs of species. 0 indicates that the ortholog was not differentially expressed between castes.

<table>
<thead>
<tr>
<th>Species 1</th>
<th>Species 2</th>
<th>$P.\ metricus$ - $S.\ invicta$</th>
<th>$P.\ metricus$ - $A.\ mellifera$</th>
<th>$S.\ invicta$ - $A.\ mellifera$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>4156</td>
<td>8217</td>
<td>1319</td>
</tr>
<tr>
<td>0</td>
<td>Q</td>
<td>124</td>
<td>4169</td>
<td>998</td>
</tr>
<tr>
<td>0</td>
<td>W</td>
<td>288</td>
<td>2361</td>
<td>486</td>
</tr>
<tr>
<td>Q</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>43</td>
</tr>
<tr>
<td>Q</td>
<td>Q</td>
<td>0</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>Q</td>
<td>W</td>
<td>0</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>125</td>
<td>197</td>
<td>96</td>
</tr>
<tr>
<td>W</td>
<td>Q</td>
<td>0</td>
<td>52</td>
<td>69</td>
</tr>
<tr>
<td>W</td>
<td>W</td>
<td>7</td>
<td>86</td>
<td>37</td>
</tr>
</tbody>
</table>
Table A.2  Correlation between log2 fold change of differentially expressed orthologs between pairs of species. Note that log2 fold changes were only available for *A. mellifera* differentially expressed transcripts (Chen et al. 2012. Insect Biochem and Mol Biol.), so analyses were restricted to differentially expressed orthologs.

<table>
<thead>
<tr>
<th></th>
<th><em>P. metricus - S. invicta</em></th>
<th><em>P. metricus - A. mellifera</em></th>
<th><em>S. invicta - A. mellifera</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson's correlation</td>
<td>-0.1560027</td>
<td>0.0873281</td>
<td>-0.01776526</td>
</tr>
<tr>
<td>95% confidence interval</td>
<td>[-0.2374062, -0.0724233]</td>
<td>[0.0633334, 0.1112218]</td>
<td>[-0.0661127, 0.0306654]</td>
</tr>
<tr>
<td><em>t</em></td>
<td>-3.6565</td>
<td>7.1202</td>
<td>-0.7191</td>
</tr>
<tr>
<td>df</td>
<td>536</td>
<td>6597</td>
<td>1638</td>
</tr>
<tr>
<td>p-value</td>
<td>0.0002809</td>
<td>1.19E-12</td>
<td>0.4722</td>
</tr>
</tbody>
</table>
Table A.3  RNA-Seq libraries listed by sample ID with treatment group, lane of the sequencer, and read count per end of paired end (i.e. there are twice as many reads for each sample corresponding to each end of the paired end reads). FR—Foundress Reared larvae, WR-Worker Reared larvae.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Group</th>
<th>Lane</th>
<th>Read Count (per end)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4H7L4</td>
<td>FR</td>
<td>3</td>
<td>38,913,850</td>
</tr>
<tr>
<td>4H8L2</td>
<td>FR</td>
<td>4</td>
<td>34,615,040</td>
</tr>
<tr>
<td>4H11L2</td>
<td>FR</td>
<td>1</td>
<td>38,528,539</td>
</tr>
<tr>
<td>4H18L1</td>
<td>FR</td>
<td>2</td>
<td>37,804,818</td>
</tr>
<tr>
<td>4HW3L1</td>
<td>WR</td>
<td>3</td>
<td>40,442,586</td>
</tr>
<tr>
<td>4HW10L2</td>
<td>WR</td>
<td>1</td>
<td>42,931,136</td>
</tr>
<tr>
<td>4HW11L2</td>
<td>WR</td>
<td>4</td>
<td>47,477,856</td>
</tr>
<tr>
<td>4HW12L4</td>
<td>WR</td>
<td>2</td>
<td>36,285,428</td>
</tr>
</tbody>
</table>
Table A.4  Bowtie2 alignment results by sample ID

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Aligned Mates Concordantly exactly 1 time</th>
<th>Aligned Mates Concordantly &gt; 1 times</th>
<th>Aligned Mates Discordantly 1 time</th>
<th>One Mate Aligned exactly 1 time</th>
<th>One Mate Aligned &gt; 1 time</th>
<th>Total Reads Aligned</th>
<th>Overall Alignment Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>4H7L4</td>
<td>58,967,922</td>
<td>4,905,010</td>
<td>2,782,074</td>
<td>3,593,562</td>
<td>441,285</td>
<td>70,689,853</td>
<td>90.83%</td>
</tr>
<tr>
<td>4H8L2</td>
<td>50,990,578</td>
<td>6,019,122</td>
<td>2,924,836</td>
<td>3,199,371</td>
<td>601,932</td>
<td>63,735,839</td>
<td>92.06%</td>
</tr>
<tr>
<td>4H11L2</td>
<td>58,173,758</td>
<td>7,280,168</td>
<td>2,506,218</td>
<td>3,098,928</td>
<td>571,589</td>
<td>71,630,661</td>
<td>92.96%</td>
</tr>
<tr>
<td>4H18L1</td>
<td>55,136,232</td>
<td>8,191,046</td>
<td>2,986,294</td>
<td>3,327,456</td>
<td>674,859</td>
<td>70,315,887</td>
<td>93.00%</td>
</tr>
<tr>
<td>4HW3L1</td>
<td>59,329,176</td>
<td>8,303,770</td>
<td>2,863,700</td>
<td>3,527,490</td>
<td>666,410</td>
<td>74,690,546</td>
<td>92.34%</td>
</tr>
<tr>
<td>4HW10L2</td>
<td>63,185,430</td>
<td>8,209,268</td>
<td>3,236,190</td>
<td>3,916,023</td>
<td>699,587</td>
<td>79,246,498</td>
<td>92.29%</td>
</tr>
<tr>
<td>4HW11L2</td>
<td>67,054,862</td>
<td>8,607,818</td>
<td>5,045,772</td>
<td>4,652,144</td>
<td>1,066,198</td>
<td>86,426,794</td>
<td>91.02%</td>
</tr>
<tr>
<td>4HW12L4</td>
<td>54,071,366</td>
<td>6,410,612</td>
<td>2,844,160</td>
<td>3,208,246</td>
<td>594,925</td>
<td>67,129,309</td>
<td>92.50%</td>
</tr>
<tr>
<td>Species 1</td>
<td>Species 2</td>
<td>Species 1 Hits</td>
<td>Species 2 Hits</td>
<td>Species 1 Number of Sequences</td>
<td>Species 2 Number of Sequences</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. mellifera</td>
<td>P. canadensis</td>
<td>7043</td>
<td>11062</td>
<td>12163</td>
<td>45326</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. mellifera</td>
<td>P. metricus</td>
<td>8495</td>
<td>11062</td>
<td>15167</td>
<td>74517</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. mellifera</td>
<td>S. invicta</td>
<td>2833</td>
<td>11062</td>
<td>3097</td>
<td>3773</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. canadensis</td>
<td>P. metricus</td>
<td>16799</td>
<td>45326</td>
<td>74517</td>
<td>3773</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. canadensis</td>
<td>S. invicta</td>
<td>2366</td>
<td>45326</td>
<td>3773</td>
<td>3773</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. metricus</td>
<td>S. invicta</td>
<td>2691</td>
<td>4730</td>
<td>3773</td>
<td>3773</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table A.6: Number of differentially expressed transcripts with putative orthologs between pairs of species.

<table>
<thead>
<tr>
<th>Species 1</th>
<th>Number of Hits</th>
<th>Species 2</th>
<th>Number of Hits</th>
<th>Differentially Expressed Hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. mellifera</td>
<td>3304</td>
<td>P. canadensis</td>
<td>4363</td>
<td>426</td>
</tr>
<tr>
<td></td>
<td>3598</td>
<td>P. metricus</td>
<td>4363</td>
<td>348</td>
</tr>
<tr>
<td></td>
<td>1477</td>
<td>S. invicta</td>
<td>1477</td>
<td>190</td>
</tr>
<tr>
<td></td>
<td>187</td>
<td>P. canadensis</td>
<td>2543</td>
<td>587</td>
</tr>
<tr>
<td></td>
<td>136</td>
<td>P. metricus</td>
<td>2543</td>
<td>658</td>
</tr>
<tr>
<td></td>
<td>136</td>
<td>S. invicta</td>
<td>136</td>
<td>854</td>
</tr>
</tbody>
</table>
Table A.7  Summary information about the KEGG pathway analysis by species including: the number of KEGG pathways, the average number of differentially expressed transcripts (DETs) per KEGG pathway, and median number of DETs per KEGG pathway. Data is included for all KEGG pathways with at least one DET, all modulated KEGG pathways (at the 1.0 threshold), and all KEGG pathways with at least one DET that are not modulated (i.e. Removed).

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of KEGG Pathways</th>
<th>Average number of DETs</th>
<th>Median number of DETs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>At least 1 DET</td>
<td>Modulated</td>
<td>Removed</td>
</tr>
<tr>
<td><em>A. mellifera</em></td>
<td>113</td>
<td>100</td>
<td>13</td>
</tr>
<tr>
<td><em>P. metricus</em></td>
<td>36</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td><em>S. invicta</em></td>
<td>43</td>
<td>20</td>
<td>23</td>
</tr>
</tbody>
</table>
117

Figure A.2 Scatterplot of log2 fold change for differentially expressed orthologous sequences between a) *P. metricus* and *S. invicta*, b) *P. metricus* and *A. mellifera*, and c) *S. invicta* and *A. mellifera*. Color indicates the directionality (up-regulation in queen [Q], worker [W], or not differentially expressed [0]) for each species (listed by species 1 [x-axis], species 2 [y-axis] as indicated above). Note that log2 fold changes were only available for *A. mellifera* differentially expressed transcripts (Chen et al. 2012. Insect Biochem and Mol Biol.), so points were restricted to differentially expressed orthologs. Positive fold change indicates up-regulation in queen-destined samples.
Figure A.3  Abridged Arginine and Proline metabolism pathway (KEGG map:00330) which highlights caste differentially expressed transcripts for *A. mellifera*, *P. metricus*, and *S. invicta*. Based on the log fold change, each transcript was classified as up-regulated in queen-destined (Green, positive direction), up-regulated in worker-destined (Red, negative direction), or not differentially expressed (Gray). For some enzymes, there were multiple transcripts that mapped to the enzyme. Thus, the relative proportion of differentially expressed transcripts was calculated for each enzyme and displayed as the bar height. White indicates that no putative homolog was found (E-value > 1e-4).
Figure A.4  Number of modulated KEGG pathways by species for thresholds ranging from 0.1 to 3. KEGG pathways were considered to be "modulated" if the sum of the proportion of differentially expressed transcripts over all enzymes was greater than the threshold.
Figure A.5  Venn diagram of the number of modulated KEGG pathways (threshold 1.1) for *A. mellifera*, *P. metricus*, and *S. invicta*. Fisher Exact Tests were used to determine that there is significant overlap in number of modulated KEGG pathways between all pairs of species.
APPENDIX B. ADDITIONAL METHODS, RESULTS, TABLES, AND FIGURES FOR CHAPTER 3

Additional Methods

In addition to comparing lists of differentially expressed transcripts between nourishment treatment (this study) and caste (Berens et al., 2015a) separately as described in the main text, we also modeled the two datasets together. Because worker-destined larvae are known to have lower levels of nourishment than gyne-destined larvae, we actually have the opportunity to model the effects of differential nourishment in both settings by combining the data from both studies. From our previous study, which used samples collected directly from the field, four samples were collected early in the season when the larval nourishment is low due to small adult-to-larvae ratio and four samples collected late in the season when larval nourishment is high because of a high adult-to-larvae ratio (Berens et al., 2015a). Then, we can use the location of the samples (lab or field) as the blocking factor where each block contains eight samples with four assigned to each nourishment level. In the model, we will control for the blocking factor in order to identify transcripts with significant nourishment effects and nourishment by location effects. The model is described below.

Model

Let \( Y_{ijkt} \) denote the read counts of block (location) \( i = \{field, lab\} \), nourishment level \( j = \{low, high\} \), and replicate \( k = \{1, 2, 3, 4\} \) for transcript \( t = \{1, 2, \ldots, 74516\} \). Then the full main effects and interaction model for \( Y_{ijkt} \) is:

\[
Y_{ijkt} = \mu_{..t} + \beta_{i..t} + \tau_{.jt} + (\beta\tau)_{ij.t} + \epsilon_{ijkt}
\]
where $\mu_{...t}$ is the mean read count (expression), $\beta_{i..t}$ is the main effect of the $i$th block (location), $\tau_{.j.t}$ is the main effect of the $j$th nourishment level, $(\beta\tau)_{ij.t}$ is the interaction effect for the combination of location $i$ and nourishment level $j$, and $\epsilon_{ijkt}$ is the random residual effect for the $k$th replicate of the $j$th nourishment level in the $i$th location for transcript $t$. The model is over-parameterized, so we impose the follow constraints: $\beta_{i=field..t} = 0$, $\tau_{.j=high.t} = 0$, $(\beta\tau)_{i=field,j.t} = 0$ for all $j$, and $(\beta\tau)_{i=high,j.t} = 0$ for all $i$.

**Test for Nourishment Effect**

The nourishment-responsive transcripts are identified as the transcripts where the null hypothesis that expression under low and high nourishment is equivalent:

$$H_0 : \tau_{.j=low.t} = 0$$

is rejected (FDR < 0.05, Benjamini and Hochberg (1995)) in favor of the alternative hypothesis that there is differential expression between nourishment levels:

$$H_0 : \tau_{.j=low.t} \neq 0$$

To do this, we contrast the reduced location only effects model:

$$Y_{ijkt} = \mu_{...t} + \beta_{i..t} + \epsilon_{ijkt}$$

with the reduced location and nourishment effects model:

$$Y_{ijkt} = \mu_{...t} + \beta_{i..t} + \tau_{.j.t} + \epsilon_{ijkt}$$

using the generalized linear model test in DESeq (Anders and Huber, 2010) with pooled dispersion. From this test, we have a list of nourishment-responsive transcripts across locations.

These models were run in R (Team, 2013) following the DESeq (Anders and Huber, 2010) workflow described for multi-factor designs using a pooled dispersion for all samples and performing two generalized linear model tests to identify 1) nourishment-responsive transcripts across locations and 2) location-dependent nourishment-responsive transcripts. We identified clusters of nourishment-responsive transcripts across locations and location-dependent using
K-means clustering with the Mfuzz package (Kumar and M, 2007) from the Bioconductor repository (Gentleman et al., 2004). The lists of differentially expressed transcripts from each test and approach were then compared and visualized using Venn diagrams.

**Additional Results**

Of the 74,516 *P. metricus* transcripts, very few (60 = 0.08%) were identified to have nourishment main effects, whereas, more than twice that number of transcripts (132 = 0.18%) have a nourishment by location interaction effect (see Table B.1 and Online Material - Additional file 4 for the list of transcripts with nourishment effects and nourishment by location interaction effects). There were fewer (total of 192 transcripts) nourishment-responsive transcripts across location and location-dependent nourishment-responsive transcripts identified with this combined data approach compared to the number of nourishment-responsive transcripts identified with just the lab samples (described in Chapter 3, 284 transcripts). This is likely due to differences in variance – the dispersion of lab samples was calculated “per condition” whereas the dispersion of the lab and field samples combined was a “pooled” calculation.

The 33 (55%) transcripts with a significant nourishment effect are transcripts that share nourishment-response regardless of location (herein referred to as nourishment-specific transcripts). Figure B.3 displays the expression profiles of these 33 nourishment-specific transcripts, shown as clusters of genes with similar expression patterns (identified with the K-means approach). Most of the clusters show higher expression in the low nourishment samples compared to the high nourishment samples. Comparing these nourishment-specific transcripts to the list of 43 overlapping nourishment and caste DETs, we did not find any shared transcripts (Figure B.4), which may not be too surprising considering the majority of these overlapping nourishment and caste DETs were in the opposite directionality than predicted. Instead, we find that most (25 DETs, or 75.8%) of the nourishment-specific transcripts from this analysis overlap with the caste-related DETs identified by Berens et al. (2015a). Thus, our results from this alternative statistical analysis do still provide some support for the observation that genes related to nourishment level may also be related to caste differences.
For the 105 interaction-specific transcripts (i.e. the transcripts that were identified to have nourishment by location interaction effects but not nourishment effects) three of the four clusters shared the same strong pattern: higher expression in the low nourishment (worker-destined) compared to the high nourishment (gyne-destined) samples in the field (see Figure B.5 for the expression profiles of these interaction-specific transcripts by K-means clustering). If we focus on the lab-based contrast, three of the four clusters had higher expression in the high nourishment vs. the low nourishment group. These data show there are many DETs that show different directional patterns of expression in the two studies, similar to what we reported based on the DET list comparison in the main text. Of these interaction-specific transcripts, 48 transcripts overlap with the caste DETs and 20 are shared with the nourishment-responsive DETs, where eight are common to both the nourishment and caste DETs (Figure B.4).

Finally, there are 27 transcripts with both nourishment main effects and nourishment by location interaction effects in this model (Figure B.6), which display similar profiles to the nourishment-specific transcripts with most (four of six) clusters having higher mean expression in the low nourishment field samples. There is one cluster with higher expression in the high field samples compared to all other groups, and the other cluster has different directionality depending on the location (up-regulation in low nourishment in the field and high nourishment in the lab). Almost all (26 of 27) of the interaction transcripts are shared with the caste DETs, but there are only four common to both nourishment and caste DETs (Figure B.4).

Table B.1  Number of differentially expressed transcripts between nourishment levels across and dependent on location identified by DESeq.

<table>
<thead>
<tr>
<th></th>
<th>Nourishment-responsive across location</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Location-dependent</td>
<td>Yes</td>
<td>27</td>
<td>105</td>
<td>132</td>
</tr>
<tr>
<td>nourishment-responsive</td>
<td>No</td>
<td>33</td>
<td>74352</td>
<td>74384</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>60</td>
<td>74456</td>
<td>74516</td>
</tr>
</tbody>
</table>
Figure B.1 Bar charts of the log2 fold changes for 23 genes from Daugherty et al. (2011) and orthologous transcripts (best BLAST hits to genes in Daugherty et al. (2011)) in this study. A positive log2 fold change indicates higher expression in low nutrition samples. For some genes, there were multiple best BLAST hits to transcripts in this study, so all were used for downstream analyses. From Daugherty et al. (2011), six genes (ILPs, Kul, SPARC, tun, usp, Vg1) were identified as up-regulated in high nutrition samples, and 4 genes (CG11971, InR2, sNPFR, Tachykinin) were identified as up-regulated in low nutrition. None of the orthologous transcripts in this study are differentially expressed between nutrition levels; however, there is a significant positive correlation in log2 fold changes between these two studies (Spearman rho: 0.54, p-value = 0.001).
Figure B.2 Heat map of relative expressions (sample read counts scaled by library size then across each gene) for the 43 transcripts that were differentially expressed for both caste and nutrition. Transcripts are clustered based on relative expression across all samples (left). Transcripts that are up-regulated in high nutrition are highlighted in yellow, and low nutrition up-regulated transcripts are indicated by the blue color. Up-regulation in queen caste is indicated by the orange color, and transcripts that are up-regulated in worker caste is highlighted by the purple color.
Figure B.3  Clustering of the nourishment-specific mean normalized expression profiles by transcript for wasps raised under high nourishment in the field, high nourishment in the lab, low nourishment in the field, and low nourishment in the lab. The number of transcripts per cluster is listed next to each of the six clusters. For each cluster, transcript normalized expression profiles are displayed as grey lines, and the mean expression profile for all transcripts within the cluster is the thick colored line.
Figure B.4  Venn diagram of the number of transcripts identified by the modeling and list comparison approaches. For the modeling approach, all data from the lab and field were used to identify nourishment main effects and nourishment by location interaction effects using a generalized linear modeling test in DESeq. With the list comparison approach, we identified nourishment-responsive and caste-related transcripts using either the lab or field dataset, respectively.
Figure B.5 Clustering of the interaction-specific mean normalized expression profiles by transcript for wasps raised under high nourishment in the field, high nourishment in the lab, low nourishment in the field, and low nourishment in the lab. The number of transcripts per cluster is listed next to each of the six clusters. For each cluster, transcript normalized expression profiles are displayed as grey lines, and the mean expression profile for all transcripts within the cluster is the thick colored line.
Figure B.6  Clustering of the mean normalized expression profiles of the 27 overlapping nourishment and interaction significant transcripts for wasps raised under high nourishment in the field, high nourishment in the lab, low nourishment in the field, and low nourishment in the lab. The number of transcripts per cluster is listed next to each of the six clusters. For each cluster, transcript normalized expression profiles are displayed as grey lines, and the mean expression profile for all transcripts within the cluster is the thick colored line.
Figure B.7  Bar chart of GO categories significantly enriched (FDR < 0.05; one-tail) between caste differentially expressed transcripts (DETs) and remaining transcriptome. Note that all GO categories were over expressed in the caste DETs compared to the rest of the transcriptome. 17 significantly enriched GO categories were shared in common for both caste and nutrition. Directionality is indicated for the enriched GO categories and defined as the treatment group with the greater number of up-regulated DETs per category. All shared enriched GO terms are up-regulated in the worker caste.
APPENDIX C. ADDITIONAL METHODS, TABLES, AND FIGURES
FOR CHAPTER 4

Additional Methods

Selection of control gene

Quantitative reverse transcription polymerase chain reaction data are routinely normalized to control or “housekeeping” genes in order to account for difference between samples such as variation in RNA levels and reverse-transcription efficiency. Mounting evidence in model organisms suggest that there are no true “housekeeping” genes with expression that remains constant under all condition in a given organism (Savonet et al., 1997). Thus, determining such appropriate control genes in emerging model organisms, like Polistes paper wasp, can be challenging.

Current approaches typically select control genes based on previous studies of the same or somewhat closely-related organism, despite differences in tissue type and experimental conditions, and are not guaranteed to yield viable controls (Daugherty et al., 2011). Based on previous control genes from honey bees (Ament et al., 2008), we tested two ribosomal protein genes (RP49 and RPS8) as internal control genes. RPS8 was up-regulated during the second meeting of wasp pairs for both Polistes fuscatus and metricus, whereas RP49 initially showed no expression differences between social interaction groups for both species (A.J. Berens, unpublished data). Thus, we selected RP49 for normalization across samples.

For three candidate genes (IP3K, IP3R, and mGluR), we performed qPCR for the social interaction gene expression experiment at a later date than all other candidate genes, so we re-ran the RP49 control gene at that time. In these later runs, RP49 gene expression was significantly different between groups in P. fuscatus (ANOVA, p-value = 7.08e-6). To investigate
whether changes in \textit{RP49} gene expression were caused by cDNA degradation, we performed a correlation test of the \textit{RP49} expression from the initial and later qPCR runs. We identified high correlation between the two dates (Pearson correlation coefficient = 0.86, p-value = 7.154e-8), which suggests that there was not degradation of the samples. Although we observed some change of the level of \textit{RP49} expression (fold changes are 1.3, 1.8, and 2.5 for pairwise comparisons between social interaction groups), but these fold changes fall within an acceptable range of variability for control genes (DeRisi et al., 1997).

Therefore, \textit{RP49} was still used for normalization of \textit{IP3K}, \textit{IP3R}, and \textit{mGluR} for the social interaction gene expression experiment so that these results would be comparable to the other 15 candidate genes and comparable across species.
Table C.1  Primers for candidate and control genes developed from previously published *Polistes metricus* sequences (Toth et al., 2007)

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Gene Abbreviation</th>
<th><em>P. metricus</em> contig</th>
<th>Forward Primer Sequence (5' to 3')</th>
<th>Reverse Primer Sequence (5' to 3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetylcholinesterase</td>
<td>Ace</td>
<td>cn49499</td>
<td>CCGTTCCAGGTTCTCAATGGGTAAT</td>
<td>GTCTCGTACGTTGAGTTCCAGTCGA</td>
</tr>
<tr>
<td>Brahma associated protein 60kD</td>
<td>Bap60</td>
<td>cn7090</td>
<td>CACAATGGGATCAGGTGGATGCAA</td>
<td>GGAACAAATCTTTGCTTGTCCAG</td>
</tr>
<tr>
<td>Choline acetyltransferase</td>
<td>Cha</td>
<td>cn32427</td>
<td>GTCACACGTTCATACAGGTCGTCG</td>
<td>GGACAGGTCCGAAATGGGAATCGATAAAGG</td>
</tr>
<tr>
<td>Deep orange</td>
<td>dor</td>
<td>cn40356</td>
<td>CCTATGTAATATAATATTCTGGTGAGG</td>
<td>AGATCTCTTATAGCTATCTTTGTT</td>
</tr>
<tr>
<td>Ebony</td>
<td>e</td>
<td>cn11524</td>
<td>TTCCGTTGCTCGGCCATGATCAA</td>
<td>AGTTGGAGAATGGGTTTTTATGAA</td>
</tr>
<tr>
<td>Golden goal</td>
<td>gogo</td>
<td>cn45874</td>
<td>AATTAGAGGTGGAGGTGGTGCAA</td>
<td>ACTTTAAGGTGATCGTGCCGAGT</td>
</tr>
<tr>
<td>Grunge</td>
<td>Gig</td>
<td>cn13769</td>
<td>CTCTCTCTGCTTTCTTTCTTTCTG</td>
<td>TACGACATGGAAATCGAGGCGGA</td>
</tr>
<tr>
<td>Inositol 1,4,5-triphosphate kinase 1</td>
<td>IP3K</td>
<td>cn34626</td>
<td>CTGATCTCGCCATGGTTCTTT</td>
<td>GAGACAGGTCATCGAGTTACA</td>
</tr>
<tr>
<td>Inositol 1,4,5-triphosphate receptor</td>
<td>IP3R</td>
<td>cn43881</td>
<td>CGCACCAGCTCTAATCTACTTAAC</td>
<td>GATACCGGGTGAGCAAGTTAA</td>
</tr>
<tr>
<td>Metabotropic Glutamate Receptor</td>
<td>mGlur</td>
<td>cn30490</td>
<td>GTGTTCGCCCTTACCAATTAATA</td>
<td>AGTGGTCGTTCCTTCTCTTTTCT</td>
</tr>
<tr>
<td>Na+/Ca2+-K+ exchanger 30C</td>
<td>Nckx30C</td>
<td>cn44473</td>
<td>AAGTAGATCAGGGTGGATGCTTTGC</td>
<td>TCCGCTACGTTATTATCCCTACAA</td>
</tr>
<tr>
<td>N-methyl-D-aspartate receptor 1</td>
<td>Nmdar1</td>
<td>cn54652</td>
<td>TGATGGAATATGGCCGGAGCGTT</td>
<td>GTATCTCTTCCGGTGACTCAGCAT</td>
</tr>
<tr>
<td>Notch</td>
<td>N</td>
<td>cn15158</td>
<td>ACGACCGTTGATATCAGGGTGCCTT</td>
<td>AGAACACCTTCTCCATTCCAGCCGT</td>
</tr>
<tr>
<td>Protostome-specific guanine nucleotide exchange factor</td>
<td>PsGEF</td>
<td>cn20219</td>
<td>TCCCGTTGCAAAAGTTGTACGCAG</td>
<td>CGTCTCTGGGAGACGACGTACCATATT</td>
</tr>
<tr>
<td>Ribosomal protein 49 (control)</td>
<td>RP49</td>
<td>cn13098</td>
<td>TTCTTGGTTTCGCCGTAACCAA</td>
<td>TGTAACAGAGTTCTCGGAGGTTTT</td>
</tr>
<tr>
<td>Scabrous</td>
<td>sca</td>
<td>cn38025</td>
<td>CCTGCTGTGTAATCCGGTTCAAAGGTCG</td>
<td>GCAACCGTTACTACGTCGCTGATAA</td>
</tr>
<tr>
<td>Suppressor of variegation 2-10</td>
<td>Su(var)2-10</td>
<td>cn21226</td>
<td>CTGGAAGCTTCTCTACAGATACGC</td>
<td>GCTCTCTGGTACGGAATCAGATTTCA</td>
</tr>
<tr>
<td>Staufen</td>
<td>stau</td>
<td>cn8504</td>
<td>ATATGAAAGGCGGTTGAAACCGGA</td>
<td>TGTCGAAACGTTGAACGTAACGGCTC</td>
</tr>
<tr>
<td>Synaptotagmin 7</td>
<td>Syt7</td>
<td>cn11653</td>
<td>CATCAGAGAAGAACTCTCTGATCTTT</td>
<td>GTGCACTCTCAATCCTGTTATCAACGG</td>
</tr>
</tbody>
</table>
Figure C.1 *Polistes metricus* mean candidate gene expression with standard error across social interaction types: single (white) - wasps with no interactions, first meeting (light grey) - unfamiliar wasps after first interaction, second meeting (dark grey) - unfamiliar wasps after second interaction (n=8 per social interaction type). Candidate gene expression was normalized to *RP49* brain gene expression, which was not significantly different between social interaction types. After normalization, gene expression was scaled with single interaction group set to 1. For two of the candidate genes, there is a significant difference between second meeting (***), and other social interaction groups in *P. metricus* (pairwise t-test with Benjamini-Hochberg correction; FDR < 0.05).
Polistes fuscatus mean candidate gene expression with standard error across social interaction types: single (white) - wasps with no interactions, first meeting (light grey) - unfamiliar wasps after first interaction, second meeting (dark grey) - unfamiliar wasps after second interaction (n=8 per social interaction type). Candidate gene expression was normalized to RP49 brain gene expression (see Supplemental Methods above). After normalization, gene expression was scaled with single interaction group set to 1. For eight of the candidate genes, there is a significant difference between single (*) or first meeting (**) and other social interaction groups in P. fuscatus (pairwise t-test with Benjamini-Hochberg correction; FDR < 0.05).
Figure C.3  Mean a) non-aggressive time, b) number of aggressive contacts, and c) aggressive index with standard error between the first meeting (light grey) and second meeting (dark grey) during manipulation of social interactions behavioral assay (n = 34 wasp pairs per meeting). Test statistics and p-values provided for each pairwise t-test.


Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis,
B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry,
R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G.,


Gouzoules, S. (1984). Primate mating systems, kin associations, and cooperative behavior:

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis,

reveals differences in brain gene expression patterns associated with caste and reproductive

87, 104, 106


pression supports an evolutionary link between maternal behavior and eusociality. *Science*, 318(5849):441–444. 3, 4, 5, 10, 63, 65, 71


