Genes involved in convergent evolution of eusociality in bees

S. Hollis Woodard,n,1 Brielle J. Fischman,o,1 Aarti Venkat,b Matt E. Hudson,b Kranthi Varala,b Sydney A. Cameron,a Andrew G. Clark,b and Gene E. Robinsona,b,c,e,f,2

nProgram in Ecology, Evolution, and Conservation Biology, Departments of oCrop Sciences and eEntomology, Institute for Genomic Biology, and fNeuroscience Program, University of Illinois, Urbana, IL 61801; and cDepartment of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853

Contributed by Gene E. Robinson, March 12, 2011 (sent for review February 17, 2011)

Eusociality has arisen independently at least 11 times in insects. Despite this convergence, there are striking differences among eusocial lifestyles, ranging from species living in small colonies with overt conflict over reproduction to species in which colonies contain hundreds of thousands of highly specialized sterile workers produced by one or a few queens. Although the evolution of eusociality has been intensively studied, the genetic changes involved in the evolution of eusociality are relatively unknown. We examined patterns of molecular evolution across three independent origins of eusociality by sequencing transcriptomes of nine socially diverse bee species and combining these data with genome sequence from the honey bee Apis mellifera to generate orthologous sequence alignments for 3,647 genes. We found a shared set of 212 genes with a molecular signature of accelerated evolution across all eusocial lineages studied, as well as unique sets of 173 and 218 genes with a signature of accelerated evolution specific to either highly or primitively eusocial lineages, respectively. These results demonstrate that convergent evolution can involve a mosaic pattern of molecular changes in both shared and lineage-specific sets of genes. Genes involved in signal transduction, gland development, and carbohydrate metabolism are among the most prominent rapidly evolving genes in eusocial lineages. These findings provide a starting point for linking specific genetic changes to the evolution of eusociality.

T

he evolution of eusociality, the phenomenon in which female offspring forgo personal reproduction to care cooperatively for their siblings, is one of the major transitions of life on Earth (1). This evolutionary transition has occurred multiple times, but only in a small number of lineages, primarily in the insects (11 or more sisters; ref. 2). The evolution of eusociality has long fascinated biologists because it requires that the balance between cooperation and conflict shift in favor of cooperation, despite strong selective pressure for individual reproductive success (3).

Despite a rich history of theoretical work on the evolution of eusociality (4, 5), relatively little is known about the molecular changes associated with eusocial evolution (6). These molecular changes have the potential to inform us about the evolutionary processes involved in the evolution of eusociality, such as types and levels of selection (7). Some insights have been gained about molecular mechanisms underlying eusociality in individual eusocial lineages (6), but a broad comparative framework for exploring common principles of the molecular basis of eusocial evolution is lacking. One major unresolved question is whether independent evolutionary trajectories of eusociality involved similar or different genetic changes.

We explored the genetic basis of eusocial evolution in bees, an ideal group for comparative studies of social evolution. There is a wide diversity of social lifestyles within this group, from solitary to intermediate social to elaborate eusociality (8). Additionally, eusociality has been gained independently at least six times (9–12) in the bees, more than in any other group. These features make it possible to compare multiple, independent origins of different social lifestyles among relatively closely related species. Furthermore, the extensive knowledge of bee natural history (8, 13, 14) provides a valuable framework for developing hypotheses about the adaptive significance of genetic changes detected in eusocial bee lineages.

To study patterns of molecular evolution associated with eusociality in bees, we generated ~1 Gbp of expressed sequence tags (ESTs) from a set of nine bee species (Table S1). This set of species reflects the remarkable social diversity in bees by including eusocial and non-eusocial species; three origins of eusociality (9, 10); and two different forms of eusocial lifestyle, “highly eusocial” and “primitively eusocial” (ref. 8; Fig. L1). We combined the ESTs with genome sequence from the highly eusocial honey bee Apis mellifera (15), and created manually curated, 10-species, partial gene sequence alignments. We searched among the alignments for genes with accelerated rates of amino acid substitution in eusocial relative to non-eusocial lineages. Accelerated rates of protein evolution can reflect a molecular signature of positive natural selection (16), and shared patterns of acceleration among lineages can suggest an association between genetic changes and the evolution of shared traits.

Results

Characterization of Alignments. Our alignments corresponded to ~33% of the genes (n = 3,647; 3,638 after removal of alignments showing evidence of saturation) in the A. mellifera Official Gene Set (Dataset S1). To improve the utility of this genomic resource for evolutionary analysis, we used stringent criteria for assessing orthology to minimize misclassification of paralogous sequences within the alignments (SI Text). We also looked for functional biases in the set of genes represented by our alignments by performing Gene Ontology enrichment analysis. We identified biological processes that were overrepresented and underrepresented in our set of genes relative to all genes in the A. mellifera Official Gene Set (Dataset S1).

Phylogenetic Tree Inference from EST Data. We used Bayesian inference to estimate the phylogenetic relationships among bee species from our set of 3,638 alignments (SI Text). The phylogenetic tree inferred from third nucleotide positions was identical in structure to trees inferred in published studies that included


The authors declare no conflict of interest.

Data deposition: Transcriptome sequences reported in this paper are available at http://insectomicsociogenomics.illinois.edu and have been deposited in the NCBI Transcriptome Shotgun Assembly (TSA) database, http://www.ncbi.nlm.nih.gov/Genbank/TSA.html (for accession nos. see SI Text).

Frequently available online through the PNAS open access option.

1 S.H.W. and B.J.F. contributed equally to this work.

2 To whom correspondence should be addressed. E-mail: generobi@illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1103457108/-DCSupplemental.
greater taxonomic sampling (9–11; Fig. S1). A single, different topology was obtained by inferring phylogeny from the other nucleotide positions and from amino acid sequences (Fig. S1). We therefore performed all of our molecular evolutionary analyses using both tree topologies. Overall, tree topology had little effect on the results of our molecular evolutionary analyses (Table S2), and the results reported here use the topology in Fig. 1.

**Heterogeneous Patterns of Molecular Evolution Among Bee Lineages.** We searched among the alignments for genes with accelerated rates of amino acid substitution in eusocial relative to non-eusocial lineages. We performed two tests (Fig. 1B and C) that used likelihood ratio tests (LRTs) to compare models of neutral and nonneutral sequence evolution to search for genes in which the ratio of nonsynonymous to synonymous nucleotide substitutions (dN/dS, or ω) is higher in specified groups of eusocial lineages. Test 1 identified genes in which ω is higher in all eusocial lineages as a group relative to non-eusocial lineages and did not discriminate between the highly and primitively eusocial lineages. Test 2 did so discriminate and identified genes in which ω is highest in either all primitively or highly eusocial lineages as a group relative to all other lineages. These tests are not mutually exclusive; a gene may be evolving more rapidly in all eusocial relative to non-eusocial lineages, as well as evolving most rapidly in either the highly or primitively eusocial lineages.

Our tests of heterogeneous rates of protein evolution yielded a number of genes evolving differently between eusocial and non-eusocial lineages, and among eusocial lineages. For test 1, we found 212 out of 3,638 genes (6%) evolving significantly more rapidly in all eusocial lineages relative to non-eusocial lineages (“All Eusocial” gene list). For test 2, we found 173 genes (5%) evolving most rapidly in highly eusocial lineages (“Highly Eusocial” gene list) and 218 genes (6%) in primitively eusocial lineages (“Primitively Eusocial” gene list), relative to other lineages (false discovery rate adjusted P < 0.05 in all three cases; Dataset S2). Table 1 shows the most significant genes (based on P value) on each list. These results demonstrate that the pattern of genetic changes associated with eusocial evolution includes some common changes and some changes that are unique to the different eusocial lifestyles.

**Evaluation of Biases in Data.** We explored the results of our tests to search for potential biases related to nucleotide composition or EST sequence coverage (SI Text). We used Spearman’s rank correlation to determine if the following characteristics of the sequence data were correlated with the P values from the LRTs: (i) average GC content at the third position; (ii) average overall GC content; (iii) transition/transversion ratio (kappa); and (iv) dN tree length. The gappiness of an alignment could introduce potential biases in our results (17, 18), so we also looked for correlations between the P values from the LRTs and two metrics to assess coverage in our alignments: (i) gap percent (gapPCT), or the sum of the number of gaps in each sequence in an alignment divided by the sum of the total number of sites in all of the sequences in an alignment; and (ii) an alignment quality score (described in SI Text). Only a few of these characteristics of the data were significantly correlated (P < 0.05) with the P values of the LRTs, but all correlations were very weak (range of Spearman’s rho = −0.1–0.06, for all tests; Dataset S2).

**Biological Processes Evolving More Rapidly in Eusocial Relative to Non-Eusocial Bees.** We performed Gene Ontology (GO) enrichment analyses based on orthology to *Drosophila melanogaster* to identify biological processes that were overrepresented on the All Eusocial, Highly Eusocial, and Primitively Eusocial gene lists. GO enrichment analysis accounts for the overrepresentation of cate-
gories present in our set of 3,638 genes, but the underrepresentation of some categories in this set is one explanation for why these categories may not have been enriched in our gene lists (Dataset S1). “Gland development” and “cell surface receptor-linked signal transduction” were among the terms overrepresented exclusively in the All Eusocial gene list (P < 0.05, all GO results; Dataset S2 and Table S3).

Carbohydrate metabolism-related categories were enriched in both the All Eusocial and Highly Eusocial gene lists, suggesting that these genes are evolving both more rapidly in eusocial relative to non-eusocial lineages and also most rapidly in highly eusocial lineages (Fig. 24). Fifteen of the 26 genes encoding glycolytic enzymes in our dataset showed evidence of accelerated evolution in one or both of these lists (Fig. 2B), including enzymes that play a key regulatory role (phosphofructokinase) or are involved in glycolytic flux (hexokinase, pyruvate kinase). Subsequent analyses (see Robustness of Results) revealed that 7 out of these 15 genes appear to be evolving most rapidly in honey bees (genus Apis; SI Text). Two of the most rapidly evolving genes on the Highly Eusocial gene list encode glycolytic proteins (Table 1).

Transcription-related categories were enriched in both the All Eusocial and Primitively Eusocial gene lists, but not in the Highly Eusocial gene list. This enrichment exclusively in the All Eusocial and Primitively Eusocial gene lists suggests a similar pattern to that seen with carbohydrate-metabolism related genes in the All Eusocial and Highly Eusocial gene lists, only here with an emphasis in primitively eusocial lineages.

Lifestyle- and Lineage-Specific Patterns of Molecular Evolution. Some biological processes were enriched exclusively in either the Highly Eusocial or Primitively Eusocial gene lists and were not enriched in the All Eusocial gene list (Dataset S2 and Table S3). For example, we detected a signature of accelerated evolution in brain-related functional categories in primitively eusocial bees, but not in highly eusocial bees.

We performed an additional series of “lineage-specific” tests to identify genes evolving more rapidly in any individual eusocial lineages relative to all other lineages in our study (SI Text). We were specifically interested in whether lineages with the same eusocial lifestyle showed similar biological processes undergoing accelerated evolution, but via changes in unique sets of genes. We did find evidence for this pattern in some lineages. For example, genes related to reproduction are rapidly evolving in both primitively eusocial lineages, Bombini and Exoneura robusta, relative to all other lineages, but the actual genes in Bombini and E. robusta are largely different (Dataset S2).

Robustness of Results. We performed an additional set of analyses to explore whether specific lineages may have contributed disproportionately to some of the results reported above. We performed “exclusion tests” in which we removed eusocial lineages from our alignments, one at a time, and reran tests 1 and 2 to look for genes for which one species may have driven the pattern of accelerated evolution that we had detected previously (SI Text and Fig. S2). Given that the removal of lineages can also affect statistical power to detect accelerated evolution in a gene (19), we consider this analysis to be useful for highlighting our strongest results, but we do not believe that this analysis is sufficient to invalidate the results obtained using the full set of species.

We created three new gene lists by removing the genes from the original All Eusocial, Highly Eusocial, and Primitively Eusocial gene lists whose significance appeared to have been driven by one eusocial lineage (SI Text and Dataset S3). GO enrichment analysis revealed that some of the trends identified in our analysis using all species (Fig. 2 and Dataset S2) were not robust to the removal of lineages (Table S3 and Dataset S3), including the enrichment of “gland development” in the All Eusocial gene list and the enrichment of transcription-related categories in the All Eusocial and Primitively Eusocial gene lists. Many biological processes were robust to the removal of lineages, including “cell surface receptor-linked signal transduction” in the All Eusocial gene list, carbohydrate metabolism-related categories in the All Eusocial and Highly Eusocial gene lists, and neuron differentiation-related categories in the Primitively Eusocial gene list.

We performed an additional analysis to determine whether artificial groupings of species would lead to the same enriched biological processes as our groupings of eusocial and non-eusocial
lineages (SI Text). GO terms were enriched in the lists of significant genes derived from each artificial grouping (Dataset S3), but these terms were largely different from those obtained from our eusocial groupings. The finding of different enriched terms in this artificial grouping analysis provides additional support that our results truly relate to eusocial evolution.

Discussion

We identified several hundred genes with a molecular signature of accelerated evolution, including some with a signature in all eusocial lineages in our study and some with a signature that was specific to a certain type of eusocial lifestyle or specific to individual lineages. Together, these results demonstrate that convergent evolution can involve a mosaic pattern of molecular changes in both shared and lineage-specific sets of genes.

Genes involved in gland development, signal transduction, and carbohydrate metabolism were among the most rapidly evolving genes identified in this study. These findings provide a starting point for linking specific genetic changes to the evolution of eusociality in bees, which will be an important challenge for the future. Major steps in this endeavor include determining the consequences of the changes in amino acid sequence for protein function, learning how changes in protein function affect a particular biological process, and then understanding how evolutionary changes in a particular biological process might affect traits associated with eusociality (20). Below, we provide some speculation for the biological processes highlighted in our findings.

Genes associated with gland development appear to have been a strong target of selection during eusocial bee evolution. This is not surprising, because, relative to solitary insects, eusocial insects have remarkably diverse exocrine gland functions and produce many novel glandular secretions, including pheromones, brood food, and antimicrobial compounds (8, 13, 14). Chemical signaling is a vital mechanism used to coordinate the behavior and physiology of colony members, and it is possible that at least some of the protein-coding sequence changes identified here are related to the evolution of advanced systems of chemical communication found in social bees.

Another category of genes that appear to have been a strong target of selection during eusocial bee evolution is genes involved in signal transduction. Signal transduction has been implicated in the regulation of behavior across disparate taxa (21), and several genes on the All Eusocial gene list in this category have known roles in behavior and neuronal function, including metabotropic GABA-B receptor subtype 1 (22). Our results provide further evidence that signal transduction may be a general target of selection during behavioral evolution.

Genes associated with carbohydrate metabolism appear to have been a particularly strong target of selection during eusocial bee evolution. Our finding of a shared pattern of accelerated evolution across all eusocial lineages in our study may reflect the fact that many eusocial bees rely more heavily on highly pro-"
Brain Hypothesis, developed to explain primate brain evolution, posits that the cognitive demands of social life are a strong selective force in brain evolution (28). It might be assumed that these demands are greater in the larger and more complex colonies of the highly eusocial bees, and thus a stronger signature of rapid evolution in brain-related genes would be found in highly eusocial relative to primitively eusocial lineages (29). However, perhaps it is the primitively eusocial society members that face greater sociocognitive challenges, because social roles are more fluid and the balance between cooperation and competition is more dynamic in primitively eusocial colonies relative to the more structured, highly eusocial colonies (8, 13, 29).

One rapidly evolving gene in the Primitively Eusocial gene list, *dunce*, was originally identified as a *Drosophila* learning and memory mutant, and it has emerged as an important gene in the *dunce* pathway in social learning (31). Both the lineage-specific and robustness analyses suggest that, of the taxa studied here, *dunce* is evolving most rapidly in bumble bees. This finding of accelerated evolution in brain-related genes exclusively in primitively eusocial bees might eventually help us understand more about the evolution of behavioral differences that exist between primitively and highly eusocial species.

In addition to positive natural selection, nonadaptive phenomena such as relaxed constraint may contribute to the pattern of heterogeneous nucleotide substitution among sequences that we observed (16). Whether a gene is exposed to increased positive selection in eusocial lineages or to less purifying selection relative to non-eusocial lineages is a distinction that we cannot formally establish. In both cases, a difference in selective regime between the eusocial and non-eusocial lineages resulted in an increased rate of protein evolution in the eusocial lineages. Other issues have been raised regarding the reliability of the statistical methods we used for detecting adaptive molecular changes in individual genes (32–34). However, our focus on identifying biological processes represented by groups of genes, rather than individual genes, ameliorates these concerns. It is unlikely that so many genes in a single functional GO category, particularly those involved in basic “housekeeping” processes (e.g., carbohydrate metabolism), have been under relaxed constraint. However, the extent to which nonselective forces contributed to a differential rate of molecular evolution between eusocial and non-eusocial lineages is poorly understood.

The results we present motivate further investigation into differences in the functioning of these biological processes between eusocial and non-eusocial species and the functional effects of the specific genetic changes identified.

A key finding in this study is that convergent evolution of eusociality in bees involves both shared and lineage-specific sets of genes. The lineage-specific findings suggest that the multiple, independent evolutionary pathways to eusociality may have each been shaped by different combinations of extrinsic and intrinsic factors, and perhaps also via different forces of selection. In the future, it may be possible to use molecular signatures of selection on different functional classes of genes to identify which forces of selection were important in eusocial evolution. Recent evidence suggests that reproductive protein evolution can be driven by sexual selection (7), but it is not yet known if there are similar connections between other selective forces and functional classes of genes.

Our finding of shared sets of rapidly evolving genes across three independent lineages that gave rise to eusociality in bees suggests that there might also be some common molecular roots for eusocial evolution, despite the incredible social diversity found among bees. Among the biological processes that appear to have been under selection across all eusocial lineages in our study, carbohydrate metabolism stands out. Insulin signaling, which is involved in carbohydrate metabolism, has been broadly implicated in the regulation of several eusocial traits, as mentioned above (6). It has been suggested that there is a “genetic toolkit” for eusociality, a set of highly conserved genes and molecular pathways that were co-opted for novel, social functions during eusocial evolution (26). Our results provide additional support for the possibility that genes related to carbohydrate metabolism are key components of this putative toolkit (6, 26).

The existence of a genetic toolkit for eusociality can be rigorously tested because there are at least another eight independent gains of eusociality in the bees, ants, wasps, and termites (2). The insect societies provide rich material to explore how changes in DNA sequence are associated with the evolution of social life.

Materials and Methods

Bee Collection and Sequencing. Bees used for sequencing were free-flying or collected from nests. They were placed directly into liquid nitrogen for RNA preservation. Different ages, behavioral groups, and castes (when applicable) were used to maximize transcript diversity. RNA was extracted from brains and abdomens of 50+ *A. mellifera* (30) and from 10% abdomen was sequenced by 454 Life Science/Roche on the GS-FLX platform. Most transcripts in the genome are expressed in the brain (35); abdomen tissue was added to enhance transcript discovery for reproduction-related processes. Additional information about collections, RNA extractions, and sequencing is provided in SI Text.

EST and Alignment Assembly. EST reads were assembled by using Phrap to generate species-specific, nonredundant contigs and singletons. *A. mellifera* gene models were obtained from BeeBase (Official Honey Bee Gene Set; http://genomes.arc.georgetown.edu/drupal/beebase/). For each species, the assembled ESTs were matched to the *A. mellifera* gene models. Orthology was determined by using the reciprocal best BLAST hit. Gapped ortholog-reference-guided transcript assemblies (GOTAs) were created by concatenating the top reciprocal hits and trimming the overlaps. Multiple sequence alignments were then created by using MAFFT software (36). All alignments were manually inspected in Geneious (37), and ambiguous regions were masked from further analyses. Additional information about orthology assignment and editing is provided in SI Text.

Phylogeny. Nucleotide sequences for 3,647 protein-coding EST gene fragments were aligned (38), edited manually (37), and modified to include fragments containing no gaps for any of the 10 taxa. Genes >300 nucleotides in length >100 bp were concatenated, and the resulting inframe nucleotide alignment (n = 717 gene fragments; 69,461 bp total) was analyzed with Bayesian inference in MrBayes (v3.1.2 MP) (parallel version for unix clusters) (38) under the substitution model GTR + I + G, amino acid translation analyses were run by using the JTT fixed-rate model. Fig. 1 shows the consensus tree of the Bayesian posterior distribution of phylogenetic trees from analysis of third codon positions (Fig. S1). The consensus trees based on all, first, and second position nucleotide sites and amino acid sequence are reported in Fig. S1.

Evolutionary Tests. We used the program codeml in the PAML package (39) to fit our alignment data to branch models of codon substitution by maximum likelihood to identify differences in ω within the tree. For each test, the likelihoods of two models of evolution (neutral and nonneutral) were compared by using an LRT. Any genes with one or more branches with ω > 2 (n = 9) were considered to be saturated and were excluded from further analyses. To correct for multiple tests, we performed an FDR correction on nominal P values obtained from the LRTs.

GO Enrichment Analysis. For functional analyses, we used a preexisting list of *A. mellifera*–D. melanogaster orthologs (15). Orthologous fly sequences with annotation information were available for most (n = 3,451) genes in our dataset. Our GO analyses were performed by using the functional annotation tool on DAVID (40). Additional information about GO analysis is provided in SI Text.

Acknowledgments. We thank G. Bloch, C. Chanchao, K. Kapheim, M. de Lama, K. Hartfelder, M. P. Schwarz, T. Pitts-Singer, and S. Wernig for bee samples; E. Hadley for help with figures; T. Newman for laboratory assistance; C. Rasmussen for field assistance and bee identification; T. O’Connor for sequence editing; B. Smith for bioinformatic assistance; J. Huelsenbeck, A. Meade, and L. Jermin for substitution model discussion; T. D. Seeley for helpful discussion; members of the Clark laboratory for helpful suggestions on the analyses; S. A. Ament, A. L. Toth, J. B. Whitfield, the Social Insect