Genes involved in convergent evolution of eusociality in bees

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

[a]Program in Ecology, Evolution, and Conservation Biology, Departments of [b]Crop Sciences and [c]Entomology, [e]Institute for Genomic Biology, and [f]Neuroscience Program, University of Illinois, Urbana, IL 61801; and [d]Department 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.

social evolution | social insects | sociogenomics | molecular phylogenetics

The 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 times; 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 intermediately 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. 1*A*). 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

Author contributions: S.H.W., B.J.F., A.G.C., and G.E.R. designed research; S.H.W., B.J.F., A.V., M.E.H., K.V., and S.A.C. performed research; S.H.W., B.J.F., A.V., M.E.H., K.V., S.A.C., A.G.C., and G.E.R. analyzed data; and S.H.W., B.J.F., and G.E.R. wrote the paper.

The authors declare no conflict of interest.

Data deposition: Transcriptome sequences reported in this paper are available at http://insectsociogenomics.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*).

Freely 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.

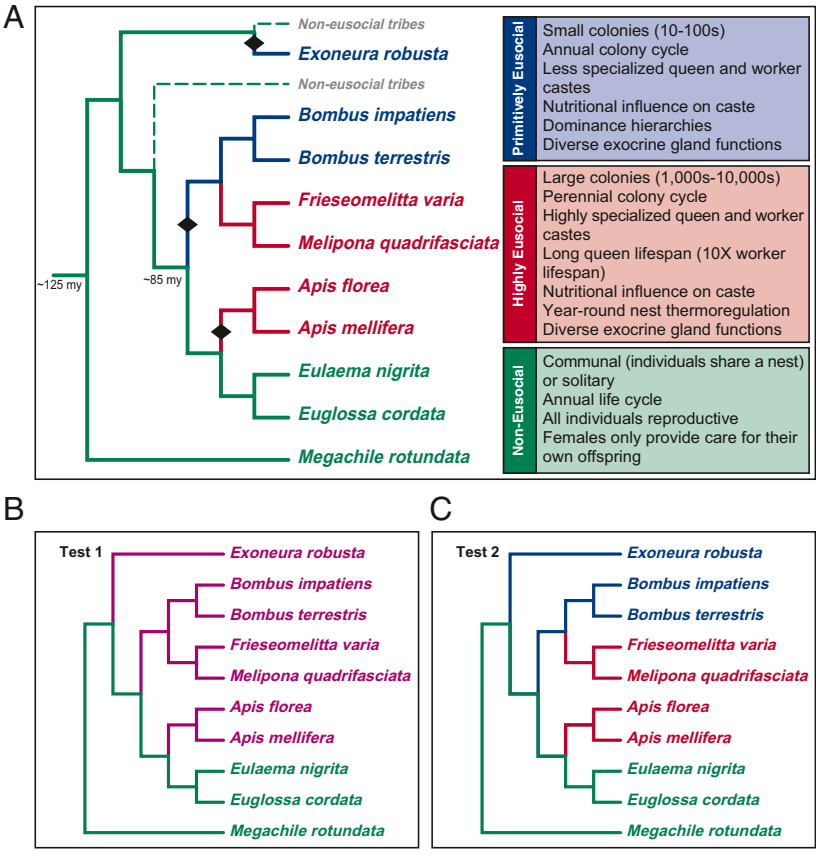

Fig. 1. Bee species and evolutionary models used to identify genes evolving rapidly in eusocial lineages. (*A*) Phylogeny of species in study based on previously published trees (9–11) and EST data (*SI Text*). Some analyses of EST data yielded an alternative topology; molecular evolutionary analyses performed with each topology gave highly similar results (*SI Text* and Table S2). Diamonds represent independent origins of eusociality. Reconstruction of eusocial origins based on phylogenies with greater taxon sampling (refs. 9–11 ; green dashed branches indicate position of non-eusocial lineages not included in the study). Lineages are color-coded by lifestyle: red, highly eusocial; blue, primitively eusocial; and green, non-eusocial. Boxes list key characteristics of each lifestyle (8, 13). (*B* and *C*) Representation of branch models of nonneutral evolution that were compared with null models by using likelihood ratio tests (LRTs). Lineages are color-coded as in *A*, except in test 1, where "All Eusocial" lineages are coded in purple. *B*, Test 1: $\omega_{\text{Eusocial}} \neq \omega_{\text{Non-eusocial}}$; *C*, Test 2: $\omega_{\text{Highly eusocial}} \neq \omega_{\text{Primitively eusocial}} \neq \omega_{\text{Non-eusocial}}$.

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. 1 *B* 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 (d_N/d_S, 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*) d_N 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-

Table 1. Genes evolving more rapidly in eusocial bee lineages

Gene	Function	*A. mellifera* gene	Rank	*P*	Relative ω*
Accelerated evolution in all eusocial lineages (test 1)					
girdin	Actin-binding protein; regulation of cell size	GB14448	4	0.00000	2.78
dihydrolipamide dehydrogenase 1	Enzyme; glycolysis	GB17626	8	0.00006	2.52
la autoantigen-like	Ribonucleoprotein; development	GB14277	11	0.00015	3.51
brahma	Chromatin remodeler; axonogenesis and oogenesis	GB30507	12	0.00015	4.25
syntaxin7	Membrane-bound protein; SNAP receptor activity	GB14433	15	0.00020	5.49
Accelerated evolution in primitively eusocial lineages (test 2)					
dopamine N acetyltransferase	Enzyme; dopamine signaling	GB18080	3	0.00000	24.5
no on or off transient A	mRNA binding protein; courtship song in *Drosophila*	GB18173	8	0.00000	5.74
signal recognition particle 14 kDa	mRNA binding	GB15372	9	0.00000	136.5
no on or off transient A	mRNA binding protein; courtship song in *Drosophila*	GB18173	10	0.00000	5.74
helicase 98B	Enzyme; immune response	GB14810	11	0.00000	4.62
β spectrin	Cytoskeletal protein; nervous system development	GB11407	12	0.00000	1.88
Accelerated evolution in highly eusocial lineages (test 2)					
phosphofructokinase	Enzyme; glycolysis	GB17113	3	0.00000	3.18
enolase	Enzyme; glycolysis	GB15039	4	0.00000	3.35
pelle	Serine/threonine kinase; immune response and axon targeting	GB16397	5	0.00000	2.80
nicotinate phosphoribosyltransferase	Enzyme; nicotinate metabolism	GB15603	24	0.00004	3.28
RhoGAP100F	GTPase; axonogenesis and signal transduction	GB15150	25	0.00005	2.39

Gene rank based on FDR-adjusted *P* values from LRTs. Evolutionary changes in the genes listed here do not appear to be strongly driven by any one lineage, and results do not seem to be affected by removal of any lineage from the analysis (*SI Text*).
*Relative ω is the fold difference compared with the non-eusocial ω. See Dataset S2 for full lists.

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. 2*A*). Fifteen of the 26 genes encoding glycolytic enzymes in our dataset showed evidence of accelerated evolution in one or both of these lists (Fig. 2*B*), 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

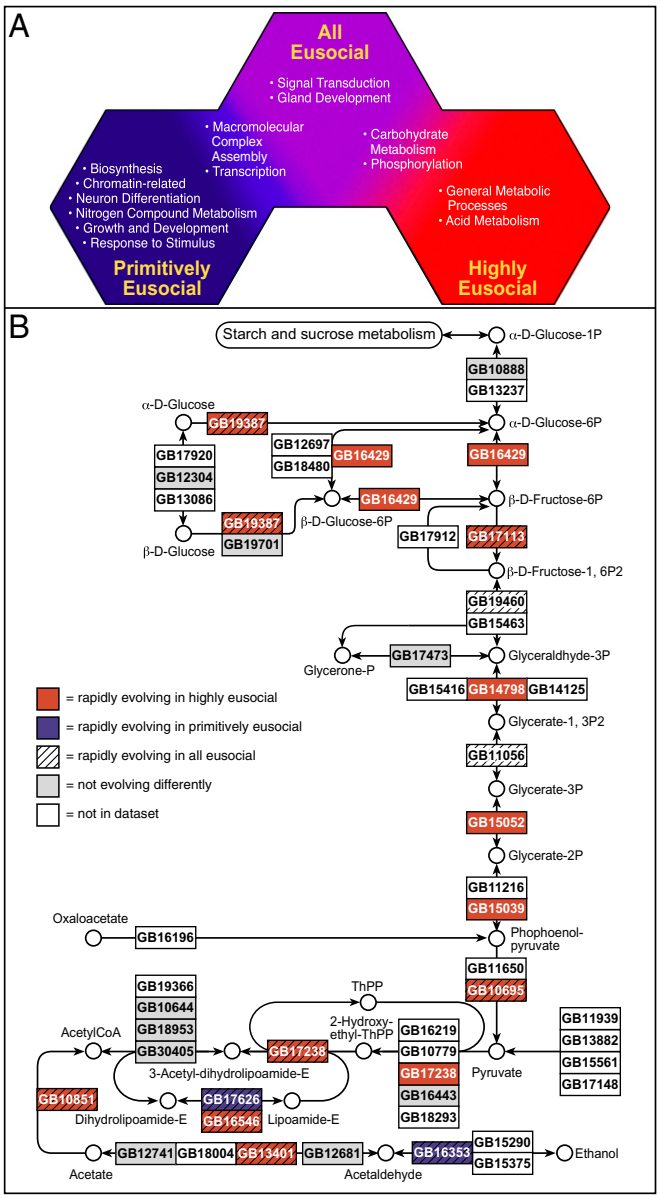

Fig. 2. Biological processes with evidence of accelerated evolution in eusocial lineages. (*A*) Overlap of rapidly evolving biological processes based on GO enrichment analysis of the All Eusocial, Highly Eusocial, and Primitively Eusocial gene lists. Individual GO categories were condensed into "meta-categories" based on related function (Table S3). (*B*) Rapidly evolving genes in the glycolysis pathway. *A. mellifera* genes mapped onto pathway based on orthology to *D. melanogaster* genes in KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/) glycolysis/gluconeogenesis pathway (dme00010). Genes likely evolving most rapidly in *Apis* are GB10695, GB10851, GB13401, GB16546, GB17113, GB17238, and GB19387.

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 processed honeys in their diet than do non-eusocial species (8), although all bees use nectar as their carbohydrate source. In addition, several characteristics shared by all eusocial insects, including worker–queen caste determination and worker–worker division of labor, are influenced by nutrition (6). Transcriptomic analyses have implicated highly conserved molecular pathways associated with metabolism (6), especially insulin signaling (23–25), in both the evolution and function of these traits (6, 26). Our results are consistent with these findings and further suggest that coding sequence changes in carbohydrate metabolism-related genes may have been involved in the evolution of these novel eusocial traits in bees.

Additional changes in carbohydrate metabolism-related genes were also detected in the highly eusocial bee lineages, but not in primitively eusocial bee lineages. This result may be due to the evolution of unique metabolic demands in the highly eusocial lifestyle, such as year-round nest thermoregulation (8), extended lifespan in queens (10-fold longer than workers; refs. 13, 27), and greatly increased foraging activity (14). Nearly half of the genes in the glycolysis pathway that were enriched in the All Eusocial and/or Highly Eusocial gene lists were not robust solely to the removal of the *Apis* lineage from the analysis, suggesting that an abundance of changes in the glycolysis pathway may have occurred in this lineage.

We were initially surprised to detect a signature of accelerated evolution in brain-related GO categories in primitively eusocial bee lineages, but not in highly eusocial bee lineages. The Social

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 regulation of neural plasticity in both invertebrates and vertebrates (30). Recent studies implicate *dunce* and other genes in the cAMP 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 or exhibit consistent model departure stratified by sociality across lineages. 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 paths 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+ females per species. Pooled mRNA (90% brain, 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 ortholog assignment and editing is provided in *SI Text*.

Phylogeny. Nucleotide sequences for 3,647 protein-coding EST gene fragments were aligned (36), edited manually (37), and modified to include fragments containing no gaps for any of the 10 taxa. Gene fragments of 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 MPI (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 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 $d_S > 2$ (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. Wongsiri for bee samples; E. Hadley for help with figures; T. Newman for laboratory assistance; C. Rassmussen for field assistance and bee identification; T. O'Conner for sequence editing; B. Smith for bioinformatic assistance; J. Huelsenbeck, A. Meade, and L. Jermiin 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

Reading Group (UIUC), and members of the Robinson and Cameron laboratories for comments on the manuscript; and P. Bouffard, T. Harkins, T. Jarvie, R. Martone, and J. Osterberger (Roche/454) for expert technical support. This work was supported by 454 Life Sciences (Roche Diagnostics Corporation) via the Roche 1GB contest (G.E.R., S.A.C., A.G.C., M.E.H., and Saurabh Sinha), National Science Foundation (NSF) Grant DEB07-43154 (to G.E.R. and M.E.H.), an NSF Predoctoral Fellowship (to B.J.F.), and National Institutes of Health Training Grant PHS2T32DC006612 (to S.H.W.; A. Feng, Principal Investigator).

1. Maynard Smith J, Szathmary E (1995) *The Major Transitions in Evolution* (Oxford Univ. Press, New York).
2. Wilson EO, Hölldobler B (2005) Eusociality: Origin and consequences. *Proc Natl Acad Sci USA* 205:13367–13371.
3. Darwin C (1859) *On the Origin of Species* (Murray, London).
4. Queller DC (1992) A general model for kin selection. *Evolution* 46:376–380.
5. Nowak MA, Tarnita CE, Wilson EO (2010) The evolution of eusociality. *Nature* 466: 1057–1062.
6. Smith CR, Toth AL, Suarez AV, Robinson GE (2008) Genetic and genomic analyses of the division of labour in insect societies. *Nat Rev Genet* 9:735–748.
7. Wong A (2010) Testing the effects of mating system variation on rates of molecular evolution in primates. *Evolution* 64:2779–2785.
8. Michener CD (1974) *The Social Behavior of the Bees* (Harvard Univ. Press, Cambridge, MA).
9. Cameron SA, Mardulyn P (2001) Multiple molecular data sets suggest independent origins of highly eusocial behavior in bees (Hymenoptera:Apinae). *Syst Biol* 50: 194–214.
10. Schwarz MP, Richards MH, Danforth BN (2007) Changing paradigms in insect social evolution: Insights from halictine and allodapine bees. *Annu Rev Entomol* 52: 127–150.
11. Cardinal S, Straka J, Danforth BN (2010) Comprehensive phylogeny of apid bees reveals the evolutionary origins and antiquity of cleptoparasitism. *Proc Natl Acad Sci USA* 107:16207–16211.
12. Danforth BN, Sipes S, Fang J, Brady SG (2006) Recent and simultaneous origins of eusociality in halictid bees. *Proc Natl Acad Sci USA* 103:15118–15123.
13. Wilson EO (1971) *The Insect Societies* (Harvard Univ. Press, Cambridge, MA).
14. Roubik DW (1992) *Ecology and Natural History of Tropical Bees* (Cambridge Univ. Press, Cambridge, UK).
15. HBGSC (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.
16. Yang ZH, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 15:496–503.
17. Schneider A, et al. (2010) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol* 1:114–118.
18. Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. *Science* 319:473–476.
19. Wong WSW, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051.
20. Dean AM, Thornton JW (2007) Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet* 8:675–688.
21. Ben-Shahar Y, Robichon A, Sokolowski MB, Robinson GE (2002) Influence of gene action across different time scales on behavior. *Science* 296:741–744.
22. Root CM, et al. (2008) A presynaptic gain control mechanism fine-tunes olfactory behavior. *Neuron* 59:311–321.
23. Ament SA, Corona M, Pollock HS, Robinson GE (2008) Insulin signaling is involved in the regulation of worker division of labor in honey bee colonies. *Proc Natl Acad Sci USA* 105:4226–4231.
24. Corona M, et al. (2007) Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc Natl Acad Sci USA* 104:7128–7133.
25. Patel A, et al. (2007) The making of a queen: TOR pathway is a key player in diphenic caste development. *PLoS ONE* 2:e509.
26. Toth AL, Robinson GE (2009) Evo-devo and the evolution of social behavior: Brain gene expression analyses in social insects. *Cold Spring Harb Symp Quant Biol* 74: 419–426.
27. Finkel T, Holbrook N (2000) Oxidants, oxidative stress and the biology of ageing. *Nature* 405:239–247.
28. Dunbar RIM (2009) The social brain hypothesis and its implications for social evolution. *Ann Hum Biol* 36:562–572.
29. Gronenberg W, Riveros AJ (2009) *Organization of Insect Societies: From Genome to Sociocomplexity*, eds Gadau J, Fewell J (Harvard Univ. Press, Cambridge, MA), pp 377–401.
30. Silva A, et al. (1998) CREB and memory. *Annu Rev Neurosci* 21:127–148.
31. Sokolowski MB (2010) Social interactions in "simple" model systems. *Neuron* 65: 780–794.
32. Nozawa M, Suzuki Y, Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci USA* 106:6700–6705.
33. Suzuki Y, Nei M (2002) Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. *Mol Biol Evol* 19:1865–1869.
34. Wolf JBW, Kunstner A, Nam K, Jakobsson M, Ellegren H (2009) Nonlinear dynamics of nonsynonymous (d(N)) and synonymous (d(S)) substitution rates affects inference of selection. *Genome Biol Evol* 1:308–319.
35. Lein ES, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445:168–176.
36. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30: 3059–3066.
37. Drummond AJ, et al. (2010) Geneious v5.1. Available at http://www.geneious.com.
38. Ronquist F, Huelsenbeck JP (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 19:1572–1574.
39. Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
40. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57.

Supporting Information

Woodard et al. 10.1073/pnas.1103457108

SI Text

Bees. Taxonomy and sources of the bee collections are listed in Table S1. For *Apis mellifera*, genome sequence was used, so collection and expressed sequence tag (EST) information are not applicable. "Purchased" colonies were purchased from rearing facilities that breed the bees for pollination. For brain tissue dissections, whole heads were partially lyophilized, and dissections were performed over dry ice to prevent RNA degradation. For abdomen tissue dissections, whole abdomens were removed from bodies on dry ice. RNA from pooled brain and abdomens was isolated by using RNeasy Mini Kits (QIAGEN).

Sequencing and EST Assembly. mRNA from each species was sequenced on the GS-FLX platform (454/Roche Life Sciences). Approximately 75–100 million bp of sequence were obtained in reads of average length of 240 bp (results of sequencing in Table S1). The ESTs from each species were masked to remove overrepresented oligos, as identified by Roche gsAssembler software, and assembled by using Phrap (1) to generate a nonredundant set of sequences. Phrap (Version 1.080721) was used with the following parameters: -ace, -max_group_size 0, and -vector_bound 0. Removal of clonal reads reduced the time required to assemble by one to a few orders of magnitude. The assemblies reduced the number of unique sequences to ~50,000 across the species. Many mRNA species in this assembly are represented by multiple contigs and/or singletons; thus, we used a reference-guided assembly as the next step.

Putative Ortholog Assignment and Alignment. For each species' EST set, we attempted to identify orthologs for all of the gene models (n ~ 11,062) from the *A. mellifera* genome (http://genomes.arc.georgetown.edu/Amel_pre_release2_OGS_cds.fa). It is possible that low-quality assemblies or alignments could influence the results of selection analysis (2, 3). For this reason, we took great care with our assemblies, alignments, and ortholog identification. Putative orthologs were identified by stringent reciprocal BLAST criteria to minimize misclassification of paralogs as follows. Each *A. mellifera* gene model was blasted against each of the species-specific, nonredundant EST databases b7 using blastn (E < 1e-6). All of the blast hits that were within 10% identity from the top hit were then blasted back against the *A. mellifera* gene models, and only those ESTs whose top hit was the same *A. mellifera* gene were called orthologs.

We concatenated all of the orthologous ESTs in the order in which they occur on the reference *A. mellifera* gene model to generate Gapped Ortholog-reference-guided Transcript Assemblies (GOTAs) for each species for each gene model in which orthologous sequence was identified. Short overlaps were trimmed to remove redundancy. GOTAs cover a given bee gene model to varying degrees between different species. There are also large differences in coverage among genes, as is expected due to random sampling from a pool of transcripts with diverse expression levels. We found putative orthologous sequences, counting partial and complete coverage, in all nine species for 3,647 (~33%) of the *A. mellifera* gene models, whereas 10% of the gene models did not have orthologs in any of the species. We used the E-INS-i alignment strategy of the Multiple Sequence Alignment by Fast Fourier Transform (MAFFT) algorithm (4) to align the orthologs thus obtained from the computational pipeline. The ortholog assignment pipeline was coded in PHP (hyper text preprocessor scripting language).

Manual Editing of Alignments. Sequencing and alignment errors can be a problem for large-scale, comparative genomic analyses, particularly in scans for signatures of selection that are based on identifying variable nucleotide positions among sequences (2, 3). To ensure that our alignments were of appropriately high quality, we visually inspected all 3,647 alignments in our dataset in Geneious (5) and manually edited the sequences, if necessary, to remove sequencing and alignment errors. The most frequent sequencing errors encountered were homopolymers, which were identified as a string of the same nucleotide base repeated several times in a row that leads to downstream stop codons in the sequence. All of the EST sequences came from expressed transcripts, so it was considered highly unlikely that the transcripts would contain more than one stop codon or that stop codons would be located anywhere but at the end of sequences. When homopolymers were corrected (either by adding or removing a base), stop codons were removed from the sequence, and the correct frame was restored. If editing the homopolymer did not eradicate the frameshift, that region of sequence was masked from further analysis. For ambiguous regions (e.g., a 2-base insertion leading to a frameshift, but not a homopolymer error), entire codons were masked from further analysis. Nucleotides at the beginning and ends of gaps were trimmed if they did not exactly match the bases seen in other species at those positions, as these were considered to be unresolvable artifacts of the alignment process.

Tree Inference. We analyzed both nucleotide and amino acid sequences of the same alignments that contained no gaps for any taxa. To account for uncertainty in the estimate of phylogeny based on the EST fragments, we used Bayesian inference, which, rather than estimating a single tree, estimates a posterior distribution of trees from which a majority-rule consensus tree can be estimated. For the amino acid sequence analyses (717 gene fragments, 66.601 characters), we used the fixed-rate JTT model (6) implemented in MrBayes 3.1.2 (7). The MCMC was run twice with four chains for 1,000,000 generations (SD of split frequencies for both runs went to 0.00 within <3,000 generations, indicating rapid convergence), with trees sampled every 1,000 generations. The first 10,000 trees were discarded as burn-in, and we constructed a consensus of the remaining trees, with all clades displaying a ≥98% posterior probability (Fig. S1).

For the nucleotide sequences (717 gene gapless fragments), we implemented multiple analyses: (*i*) concatenated sequences (208,383 characters), using the GTR + I + G model (most general model), 4 chains, 5,000,000 generations (SD of split frequencies went to 0.00 within <7,000 generations), trees sampled every 1,000 generations, discarded the first 25,000 trees as burn-in, and constructed a consensus tree displaying clades with ≥98% posterior probability (Fig. S1); (*ii*) first codon positions only, GTR + I + G model, 4 chains, 1,000,000 generations (SD of split frequencies went to 0.00 within <7,000 generations), trees sampled every 1,000 generations, discarded the first 5,000 trees as burn-in, and constructed a consensus tree displaying clades with a ≥98% posterior probability (Fig. S1); (*iii*) second codon positions, analyzed under same conditions as analysis *ii*, resulted in the same tree topology (Fig. S1); and (*iv*) third codon positions analyzed under same conditions as *analyses ii* and *iii*, estimated the tree topology of Fig. 1 (Fig. S1).

We performed all of our molecular evolutionary analyses using both inferred tree topologies. The effect of tree topology on enrichment of biological processes is presented in Table S2.

Evolutionary Tests in PAML. We used branch models in the program codeml in the package PAML (8) to fit the alignments to models

of codon substitution by maximum likelihood. For each test, we used a likelihood ratio test (LRT) to compare the likelihoods of two models: (*i*) a null model of neutral evolution in which ω is the same for all lineages; and (*ii*) an alternative model of nonneutral evolution in which ω differs between specified lineages (e.g., eusocial vs. non-eusocial lineages). The null model was rejected for genes in which the LRT was significant at an FDR-corrected $P < 0.05$. For all models, the control files used in codeml had the following settings: Codonfreq = 2, kappa = 3, initial omega = 0.2, fix_alpha = 1. For the null model control file, model = 0; for the alternative model control file, model = 2. codeml estimates the value of ω for each specified group of lineages. Our gene lists (All Eusocial, Highly Eusocial, Primitively Eusocial, and the lineage-specific gene lists) were created by comparing the estimated ω values of significant genes and grouping genes according to which specified group of lineages had the highest ω value.

Evaluation of Biases in Data. The alignment quality scoring system was devised to account for breadth and depth of sequence coverage in an alignment and was calculated as follows: Each codon in the alignment was binned based on the number of species that had sequence at that codon position (1–10); the number of codons in each bin was summed and then multiplied by the number of species in that bin; and the values of all bins were summed then normalized to the total alignment length by dividing this sum by the total sequence length for which there is sequence from at least two species. The higher the alignment score, the better the quality of the alignment (e.g., best alignment score = 10). The results of this analysis are provided in Dataset S2. The four alignment composition metrics and the two alignment quality metrics used for our correlation analysis (calculated for each alignment) are included in Dataset S1.

Gene Ontology Enrichment Analysis. A subset of genes did not have *D. melanogaster* orthologs and was excluded from the GO analysis; these genes are identified in Dataset S1. To test for enrichment, we compared each gene list to the background gene dataset ($n = 3{,}638$) at the GO fat level. The enriched biological processes for the gene lists from tests 1 and 2 are presented in Table S3.

Lineage-Specific Tests. In addition to our main tests (tests 1 and 2), we performed four lineage-specific tests, in which we compared the likelihoods of the null model of neutral evolution and an alternative model of nonneutral evolution in which ω for a specified eusocial lineage differs from all other lineages. For these tests, the same general methodologies we described in *Evolutionary Tests in PAML* and *Gene Ontology Enrichment Analysis* were used. These lineage-specific tests were performed on the following lineages: (*i*) Apini-specific; (*ii*) Bombini-specific; (*iii*) *E. robusta*-specific; and (*iv*) Meliponini-specific. The significant gene lists with functional annotation are included in Dataset S2.

Robustness of Results. We performed a series of "exclusion tests" to investigate whether the removal of any of the four eusocial lineages (Apini, Bombini, *E. robusta*, and Meliponini) from the alignments would result in a loss of significance (i.e., FDR adjusted $P > 0.05$ for LRT) for any genes. We removed each lineage, one at a time, and reran tests 1 and 2.

17.5% ($n = 37$) of genes on the All Eusocial gene list were robust in all four exclusion tests (Fig. S2). These genes have the strongest support for evolving rapidly in all eusocial lineages in the analysis. The majority of genes on the All Eusocial gene list (55%, $n = 117$) were not robust in two or more exclusion tests, suggesting that the loss of information resulting from the removal of species from the analysis resulted in a loss of statistical power to detect patterns of molecular changes among sequences. Approximately one-third (27%, $n = 58$) of the genes on the All Eusocial gene list were not robust only in one exclusion test. It seems likely that for this set of genes, the lineage that was removed may have been driving the original results of the LRTs. Fig. S2 shows, for this set of genes, which of the four eusocial lineages appears to have been driving the result. That the smallest number of genes appear to have been driven by *E. robusta* ($n = 5$) is not surprising, given that our analyses included only one species from this lineage vs. two species for all other lineages.

We performed a similar analysis for the Highly Eusocial and Primitively Eusocial gene lists. For Highly Eusocial, a smaller proportion of genes on the list (5%, $n = 9$) were robust in all four exclusion tests, and a larger proportion (60%, $n = 104$) were not robust in two or more exclusion tests (Fig. S2). For all 60 genes on the Highly Eusocial gene list that appear to have been driven by one lineage, one of the two highly eusocial lineages appears to have been driving the result (Fig. S2). For the Primitively Eusocial gene list, 5% of genes on the list ($n = 10$) were robust in all four exclusion tests, and 84% of the genes ($n = 183$) were not robust in two or more exclusion tests (Fig. S2). For the 11% of genes ($n = 25$) that appear to have been driven by one lineage, all lineages except the stingless bees appear to have been driving results (Fig. S2).

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. Specifically, the original gene lists were filtered to remove genes that were not robust in a single exclusion test. Our reasoning for not removing genes that were not robust in two or more exclusion tests from our original gene lists is that for these genes, it seems more likely that the loss of significance was due to a lack of statistical power.

We reran GO enrichment analyses on these three new gene lists and compared the results of these more stringent analyses with the results of our original analysis, which included all species. Many of the GO categories enriched in our original gene lists were also enriched in the new gene lists, although some GO categories did lose enrichment. We consider the GO categories whose enrichment was robust to these exclusion tests to be our strongest results, but we do not believe that the loss of enrichment of some GO categories in these exclusion tests is sufficient to invalidate the enrichment results obtained from analyses using the full set of species. The gene lists and GO results are provided in Dataset S3.

For our tests of whether artificial groupings of lineages would result in similar enriched GO categories as with our actual eusocial groupings, we made three artificial groupings to compare with test 1 and three artificial groupings to compare with test 2, as follows. Test 1: Grouping 1, (Meliponini, *M. rotundata*; "group A") vs. (Euglossini, Apini, Bombini, *E. robusta*; "group B"); Grouping 2, (*M. rotundata*, Meliponini, Bombini; "Group A") vs. (*E. robusta*, Apini, Euglossini; "group B"); Grouping 3, (*M. rotundata*, Euglossini, Bombini, Meliponini; "group A") vs. (Apini, *E. robusta*; "group B"). Test 2: Grouping 1, (*M. rotundata*, Apini; "group A") vs. (Euglossini, Bombini; "group B") vs. (Meliponini, *E. robusta*; "group C"); Grouping 2, (*M. rotundata,* Bombini; "group A") vs. (*E. robusta*, Apini; "group B") vs. (Euglossini, Meliponini; "group C"); Grouping 3, (Euglossini, *M. rotundata*; "group A") vs. (Bombini, Apini, "group B") vs. (Meliponini, *E. robusta*; "group C"). The GO results of this analysis are included in Dataset S3.

Data Deposition. Raw sequences have been deposited at: http://trace.ncbi.nlm.nih.gov/Traces/home with the following accession numbers (Species, Project ID, SRA accession numbers): *Apis florea*, 62549, SRR098291.1, SRR098292.1; *Bombus impatiens*, 62671, SRR098293.1, SRR098294.1; *Bombus insularis*, 62673, SRR098295.1, SRR098296.1; *Bombus terrestris*, 62675, SRR098297.1; *Centris flavifrons*, 62677, SRR098298.1, SRR098299.1; *Eulaema nigrita*, 62679, SRR098300.1, SRR098301.1; *Euglossa cordata*, 62681, SRR098302.1; *Exoneura*

robusta, 62683, SRR098303.1; *Frieseomelitta varia*, 62685, SRR098304.1; *Melipona quadrifasciata*, 62691, SRR098313.1; *Megachile rotundata*, 62687, SRR098305.1, SRR098306.1, SRR098307.1, SRR098308.1, SRR098309.1, SRR098310.1; *Megalopta genalis*, 62689, SRR098311.1, SRR098312.2. Assemblies have been deposited at: http://www.ncbi.nlm.nih.gov/genbank/TSA.html with the following accession numbers (Species, Project ID, TSA accession numbers): *Apis florea*, 62549, HP823158–HP849658; *Bombus impatiens*, 62671, JI092751–JI122887; *Bombus insularis*, 62673, JI045409-JI062504; *Bombus terrestris*, 62675, JI025924–JI045408; *Centris flavifrons*, 62677, HP985225–HP999999, JI000001–JI025923; *Eulaema nigrita*, 62679, HP931237–HP959564; *Euglossa cordata*, 62681, HP959565–HP985224; *Exoneura robusta*, 62683, HP893798–HP931236; *Frieseomelitta varia*, 62685, HP873735–HP893797; *Megachile rotundata*, 62687, JI122888–JI136238; *Megalopta genalis*, 62689, JI136239–JI148409; *Melipona quadrifasciata*, 62691, HP849659–HP873734.

1. Green P (1999) Phrap, SWAT, Crossmatch (Univ of Washington, Seattle).
2. Schneider A, et al. (2010) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol* 1:114–118.
3. Wong KM, Suchard MA, Huelsenbeck JP (2008) Alignment uncertainty and genomic analysis. *Science* 319:473–476.
4. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30: 3059–3066.
5. Drummond AJ, et al. (2010) Geneious v5.1. Available at http://www.geneious.com.
6. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8:275–282.
7. Ronquist F, Huelsenbeck JP (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 19:1572–1574.
8. Yang ZH (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.

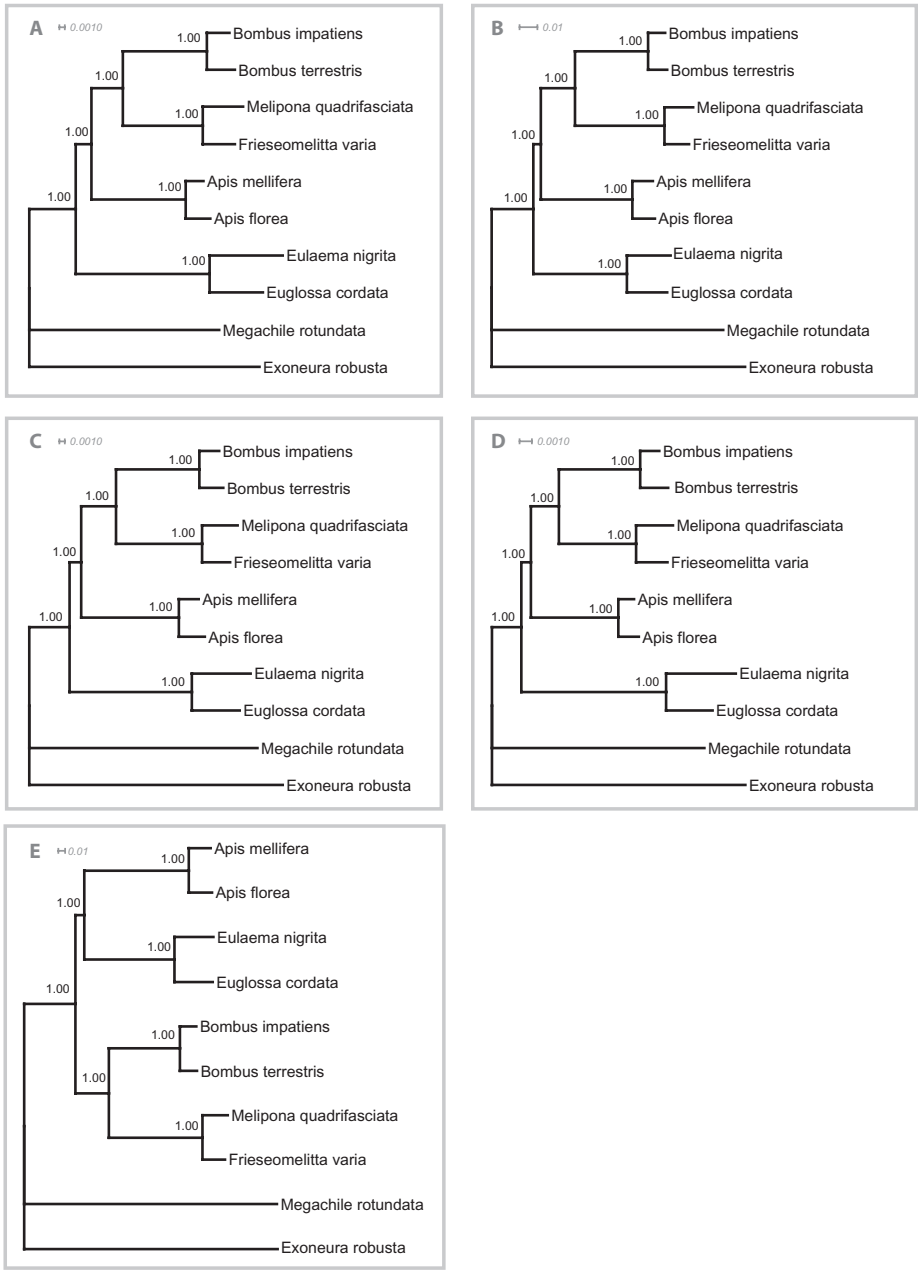

Fig. S1. Tree inferences. (*A*) Tree based on amino acid sequence. (*B*) Tree based on all nucleotide sites. (*C*) Tree based on first codon position nucleotide sites. (*D*) Tree based on second codon position nucleotide sites. (*E*) Tree based on third codon position nucleotide sites. This topology was used for Fig. 1 and is identical in structure to trees inferred in published studies that included greater taxonomic sampling.

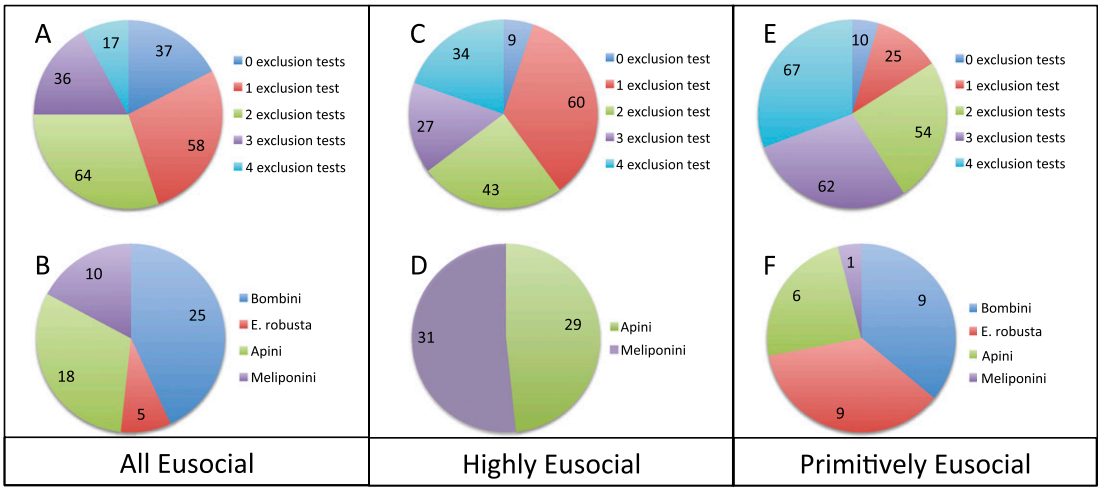

Fig. S2. Summary of eusocial lineage exclusions tests. For each of the three original gene lists (All Eusocial, Highly Eusocial, and Primitively Eusocial), one pie chart shows the number of genes for which the LRT was no longer significant in 0–4 exclusion tests (*A*, *C*, and *E*, respectively), and another shows, only for the genes no longer significant in 1 exclusion test, the identity of the eusocial lineage (*B*, *D*, and *F*, respectively). From these results, it can be inferred which eusocial lineage may have been contributing most strongly to the original results of the LRTs.

Table S1. Collection information and sequencing results for species in study

Species	Family, subfamily, tribe	Collection type (location)	Total bases	Nonredundant sequences	Nonredundant bases	*A. mellifera* orthologs
Apis mellifera	Apidae, Apinae, Apini	Genome sequence available	NA	NA	NA	NA
Apis florae	Apidae, Apinae, Apini	Wild caught (Thailand)	72,105,391	59,010	19,005,995	8,179
Bombus impatiens	Apidae, Apinae, Bombini	Purchased (USA)	98,630,488	54,542	22,439,256	8,013
Bombus terrestris	Apidae, Apinae, Bombini	Purchased (Israel)	76,405,196	42,816	15,221,160	6,768
Euglossa cordata	Apidae, Apinae, Euglossini	Wild caught (Brazil)	76,603,013	49,830	19,862,919	7,364
Eulaema nigrita	Apidae, Apinae, Euglossini	Wild caught (Brazil)	89,954,584	56,689	21,505,840	7,443
Frieseomelitta varia	Apidae, Apinae, Meliponini	Wild caught (Brazil)	74,352,793	50,809	16,214,275	6,438
Melipona quadrifasciata	Apidae, Apinae, Meliponini	Wild caught (Brazil)	77,477,792	54,525	19,198,721	7,161
Exoneura robusta	Apidae, Xylocopinae, Allodapini	Wild caught (Australia)	116,660,239	60,647	26,104,224	6,680
Megachile rotundata	Megachilidae, Megachilinae, Megachilini	Purchased (USA)	48,306,504	44,670	15,133,664	6,747

Table S2. Effect of alternative tree topologies on results

Biological process	Tree topology 1	Tree topology 2
Accelerated evolution in all eusocial lineages (test 1)		
Alcohol catabolic process	*	*
Carbohydrate catabolic process	*	*
Cell surface receptor linked signal transduction	*	*
Cellular carbohydrate catabolic process	*	*
Gland development	*	*
Glucose catabolic process	*	*
Glucose metabolic process	*	*
Glycolysis	*	*
Hexose catabolic process	*	*
Hexose metabolic process	*	*
Macromolecular complex assembly	*	*
Macromolecular complex subunit organization	*	*
Monosaccharide catabolic process	*	*
Monosaccharide metabolic process	*	*
Protein complex assembly	*	*
Protein complex biogenesis	*	*
Cellular protein complex assembly	*	
ncRNA processing	*	
Nucleoplasm part	*	
Protein amino acid phosphorylation	*	
RNA processing	*	
Transcription	*	
Regulation of developmental growth		*
Regulation of growth		*
Accelerated evolution in highly eusocial lineages (test 2)		
Alcohol catabolic process	*	*
Carbohydrate catabolic process	*	*
Carboxylic acid biosynthetic process	*	*
Cellular carbohydrate catabolic process	*	*
Generation of precursor metabolites and energy	*	*
Glucose catabolic process	*	*
Glucose metabolic process	*	*
Glycolysis	*	*
Hexose catabolic process	*	*
Hexose metabolic process	*	*
Monosaccharide catabolic process	*	*
Monosaccharide metabolic process	*	*
Organic acid biosynthetic process	*	*
Oxidation reduction	*	*
Phosphorylation	*	
Protein amino acid phosphorylation	*	
Fatty acid metabolic process		*
Lipid catabolic process		*
Nitrogen compound biosynthetic process		*
Ovarian follicle cell development		*
Accelerated evolution in primitively eusocial lineages (test 2)		
Axonogenesis	*	*
Cell morphogenesis	*	*
Cell morphogenesis involved in differentiation	*	*
Cell morphogenesis involved in neuron differentiation	*	*
Cell part morphogenesis	*	*
Cell projection morphogenesis	*	*
Cell projection organization	*	*
Cellular component morphogenesis	*	*
Chromatin modification	*	*
Covalent chromatin modification	*	*
Growth	*	*
Histone modification	*	*
Macromolecular complex assembly	*	*
Macromolecular complex subunit organization	*	*
Muscle organ development	*	*
Neuron development	*	*

Table S2. Cont.

Biological process	Tree topology 1	Tree topology 2
Neuron differentiation	*	*
Neuron projection development	*	*
Neuron projection morphogenesis	*	*
Photoreceptor cell differentiation	*	*
Positive regulation of biosynthetic process	*	*
Positive regulation of cellular biosynthetic process	*	*
Positive regulation of gene expression	*	*
Positive regulation of macromolecule biosynthetic process	*	*
Positive regulation of nitrogen compound metabolic process	*	*
Positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	*	*
Positive regulation of transcription	*	*
Protein complex assembly	*	*
Protein complex biogenesis	*	*
Regulation of transcription	*	*
Regulation of transcription, DNA-dependent	*	*
Response to endogenous stimulus	*	*
Response to hormone stimulus	*	*
Response to organic substance	*	*
RNA biosynthetic process	*	*
Transcription	*	*
Transcription initiation	*	*
Transcription initiation from RNA polymerase II promoter	*	*
Transcription, DNA-dependent	*	*
Postembryonic development	*	
Cell growth		*
Compound eye morphogenesis		*
Eye morphogenesis		*
Positive regulation of transcription from RNA polymerase II promoter		*
Positive regulation of transcription, DNA-dependent		*
Regulation of growth		*
Regulation of RNA metabolic process		*
Regulation of transcription from RNA polymerase II promoter		*

Biological processes are enriched terms from GO enrichment analysis. Tree topology 1: [(Apini, Euglossini)(Bombini, Meliponini)]. Tree topology 2: {[(Apini)(Bombini, Meliponini)](Euglossini)}.

*Term significant in list at $P < 0.05$. See *SI Text* for details on GO enrichment and molecular support for phylogenies.

Table S3. Biological processes enriched in the All Eusocial, Primitively Eusocial, Highly Eusocial, and Non-Eusocial gene lists

Biological process	No. of genes	Fold enrichment	*P*
Accelerated evolution in all eusocial lineages (test 1)			
Carbohydrate metabolism			
Glycolysis*	5	6.66	0.00475
Cellular carbohydrate catabolic process	6	4.57	0.00776
Glucose metabolic process	6	4.57	0.00776
Hexose metabolic process	7	3.61	0.01037
Monosaccharide metabolic process	7	3.29	0.01626
Carbohydrate catabolic process	6	3.84	0.01658
Glucose catabolic process	5	4.70	0.01793
Hexose catabolic process	5	4.70	0.01793
Monosaccharide catabolic process	5	4.70	0.01793
Alcohol catabolic process	5	4.21	0.02652
Gland development			
Gland development	6	3.00	0.04442
Signal transduction			
Cell surface receptor linked signal transduction*	12	1.88	0.04765
Transcription			
RNA processing	16	1.73	0.03671
ncRNA processing	8	2.41	0.04255
Transcription	15	1.70	0.04985
Phosphorylation			
Protein amino acid phosphorylation	10	2.08	0.04569
Macromolecular complex assembly			
Protein complex assembly*	13	2.85	0.00139
Protein complex biogenesis*	13	2.85	0.00139
Macromolecular complex subunit organization	15	1.95	0.01751
Cellular protein complex assembly	6	2.91	0.04983
Accelerated evolution in non-eusocial lineages (test 1)			
Protein metabolism			
Proteolysis involved in cellular protein catabolic process	15	2.62	0.00114
Cellular protein catabolic process	15	2.62	0.00114
Protein catabolic process	15	2.39	0.00281
Modification-dependent protein catabolic process	12	2.27	0.01336
Ubiquitin-dependent protein catabolic process	8	2.60	0.02994
Protein amino acid dephosphorylation	5	3.79	0.03818
Neurotransmission			
Neurotransmitter secretion	8	2.60	0.02994
Regulation of neurotransmitter levels	8	2.60	0.02994
Neurotransmitter transport	8	2.47	0.03852
General metabolism			
Macromolecule catabolic process	18	2.32	0.00120
Cellular macromolecule catabolic process	16	2.37	0.00212
Modification-dependent macromolecule catabolic process	12	2.27	0.01336
Secretion			
Secretion	9	2.34	0.03378
Secretion by cell	9	2.34	0.03378
Translation			
Translational initiation	7	3.74	0.00882
Cell signaling			
Generation of a signal involved in cell-cell signaling	8	2.55	0.03264
Protein modification			
Protein ubiquitination	5	5.35	0.01147
Proteolysis	18	1.81	0.01622
Protein modification by small protein conjugation	5	3.95	0.03318
Transport and movement			
Vesicle-mediated transport	22	1.88	0.00423
Microtubule-based process	17	1.83	0.01827
Microtubule-based movement	6	3.76	0.01828
Protein localization	18	1.77	0.01984
Cellular macromolecule localization	13	2.06	0.01997
RNA localization	8	2.55	0.03264
Protein transport	14	1.76	0.04770
Cell division			

Table S3. Cont.

Biological process	No. of genes	Fold enrichment	*P*
Cell division	9	2.60	0.01903
Accelerated evolution in highly eusocial lineages (test 2)			
Carbohydrate metabolism			
Glycolysis*	8	14.81	0.00000
Hexose catabolic process*	8	10.45	0.00000
Glucose catabolic process*	8	10.45	0.00000
Monosaccharide catabolic process*	8	10.45	0.00000
Alcohol catabolic process*	8	9.35	0.00001
Cellular carbohydrate catabolic process*	8	8.46	0.00002
Glucose metabolic process*	8	8.46	0.00002
Hexose metabolic process*	9	6.45	0.00004
Carbohydrate catabolic process*	8	7.11	0.00007
Monosaccharide metabolic process*	9	5.88	0.00008
General metabolism			
Generation of precursor metabolites and energy*	11	2.98	0.00283
Oxidation reduction*	16	2.10	0.00645
Acid metabolism			
Organic acid biosynthetic process	5	4.27	0.02600
Carboxylic acid biosynthetic process	5	4.27	0.02600
Phosphorylation			
Protein amino acid phosphorylation	9	2.60	0.01914
Phosphorylation	13	1.94	0.02978
Accelerated evolution in primitively eusocial lineages (test 2)			
Transcription			
Transcription initiation from RNA polymerase II promoter	6	2.50	0.00050
Transcription initiation	7	2.27	0.00077
Positive regulation of transcription	7	2.55	0.00150
Positive regulation of gene expression	7	1.90	0.00377
RNA biosynthetic process	9	2.16	0.00383
Transcription, DNA-dependent	9	2.63	0.00443
Regulation of transcription, DNA-dependent	15	2.63	0.00443
Transcription	20	1.99	0.00665
Regulation of transcription	22	2.24	0.00743
Chromatin-related			
Histone modification	5	2.43	0.00817
Chromatin modification	8	2.31	0.00822
Covalent chromatin modification	5	2.37	0.00985
Response to stimulus			
Response to hormone stimulus	5	2.34	0.01079
Response to endogenous stimulus	5	2.29	0.01286
Response to organic substance	6	2.77	0.01309
Neuron differentiation			
Photoreceptor cell differentiation*	6	5.00	0.01435
Axonogenesis	9	5.00	0.01435
Cell morphogenesis involved in neuron differentiation*	12	3.29	0.01619
Neuron projection morphogenesis*	12	2.21	0.01652
Neuron projection development*	12	1.95	0.01738
Neuron development*	15	3.20	0.01857
Growth and development			
Growth	6	1.94	0.01857
Muscle organ development	7	2.78	0.02108
Cell morphogenesis involved in differentiation*	12	2.78	0.02108
Cell projection morphogenesis*	12	3.11	0.02118
Cell part morphogenesis*	13	3.11	0.02118
Postembryonic development	13	2.02	0.02287
Cell projection organization*	14	2.95	0.02712
Cell morphogenesis*	17	2.95	0.02712
Cellular component morphogenesis*	18	3.31	0.03027
Macromolecular complex assembly			
Protein complex assembly	12	3.20	0.03457
Protein complex biogenesis	12	2.29	0.03798
Macromolecular complex assembly	13	2.29	0.03798

Table S3. Cont.

Biological process	No. of genes	Fold enrichment	*P*
Macromolecular complex subunit organization	15	2.46	0.03870
Biosynthesis			
Positive regulation of macromolecule biosynthetic process	7	1.86	0.04088
Positive regulation of cellular biosynthetic process	8	2.67	0.04210
Positive regulation of biosynthetic process	8	3.00	0.04427
Nitrogen compound metabolism			
Positive regulation of nitrogen compound metabolic process	7	2.91	0.04967
Positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	7	3.48	0.04990
Accelerated evolution in non-eusocial lineages (Test 2)			
Protein metabolism			
Cellular protein catabolic process	16	2.80	0.00035
Proteolysis involved in cellular protein catabolic process	16	2.80	0.00035
Protein catabolic process	16	2.55	0.00096
Modification-dependent protein catabolic process	14	2.65	0.00163
Ubiquitin-dependent protein catabolic process	9	2.92	0.00960
Neurotransmission			
Regulation of neurotransmitter levels	8	2.60	0.02994
Neurotransmitter secretion	8	2.60	0.02994
Neurotransmitter transport	8	2.47	0.03852
General metabolism			
Macromolecule catabolic process	18	2.32	0.00120
Modification-dependent macromolecule catabolic process	14	2.65	0.00163
Cellular macromolecule catabolic process	16	2.37	0.00212
Secretion			
Secretion	9	2.34	0.03378
Secretion by cell	9	2.34	0.03378
Translation			
Translational initiation	7	3.74	0.00882
Cell signaling			
Generation of a signal involved in cell-cell signaling	8	2.55	0.03264
Protein modification			
Protein modification by small protein conjugation	6	4.75	0.00673
Proteolysis	19	1.91	0.00750
Protein ubiquitination	5	5.35	0.01147
Transport and movement			
Microtubule-based movement	7	4.39	0.00389
Monovalent inorganic cation transport	7	3.18	0.01928
RNA localization	8	2.55	0.03264
Microtubule-based process	16	1.72	0.03721
Metal ion transport	6	3.12	0.03856
Phosphorylation			
Protein amino acid dephosphorylation	7	5.31	0.00138
Dephosphorylation	7	4.11	0.00552

Biological processes are enriched GO categories or KEGG pathways from enrichment analysis, grouped into "metacategories". *P* values are uncorrected. See *SI Text* for details of GO enrichment analysis.
*Categories that were robust to removal of lineages (*SI Text*).

Dataset S1. List of 3,647 genes in dataset, including information about saturation, alignment characteristics, annotation information, and overrepresentation and underrepresentation of GO biological processes

Dataset S1

Dataset S2. All Eusocial, Primitively Eusocial, Highly Eusocial, and Lineage-Specific gene lists, including annotation information, results of GO analysis, and evaluation of biases

Dataset S2

Dataset S3. Results of robustness analyses, including exclusion tests and artificial groupings

Dataset S3