4-2011

Draft genome of the red harvester ant *Pogonomyrmex barbatus*

Chris R. Smith
Earlham College

Christopher D. Smith
San Francisco State University

Hugh M. Robertson
University of Illinois Urbana–Champaign

Martin Helmkampf
Arizona State University

Aleksey Zimin
University of Maryland, College Park

See next page for additional authors

Draft genome of the red harvester ant *Pogonomyrmex barbatus*

Chris R. Smith[a,1], Christopher D. Smith[b,1], Hugh M. Robertson[c], Martin Helmkampf[d], Aleksey Zimin[e], Mark Yandell[f], Carson Holt[f], Hao Hu[f], Ehab Abouheif[g], Richard Benton[h], Elizabeth Cash[d], Vincent Croset[h], Cameron R. Currie[i,j], Eran Elhaik[k], Christine G. Elsik[l], Marie-Julie Favé[g], Vilaiwan Fernandes[g], Joshua D. Gibson[d], Dan Graur[m], Wulfila Gronenberg[n], Kirk J. Grubbs[j], Darren E. Hagen[l], Ana Sofia Ibarraran Viniegra[g], Brian R. Johnson[o], Reed M. Johnson[p], Abderrahman Khila[g], Jay W. Kim[b], Kaitlyn A. Mathis[o], Monica C. Munoz-Torres[l], Marguerite C. Murphy[q], Julie A. Mustard[d], Rin Nakamura[b], Oliver Niehuis[r], Surabhi Nigam[q], Rick P. Overson[d], Jennifer E. Placek[b], Rajendhran Rajakumar[g], Justin T. Reese[l], Garret Suen[i,j], Shu Tao[l], Candice W. Torres[o], Neil D. Tsutsui[o], Lumi Viljakainen[s], Florian Wolschin[d,t], and Jürgen Gadau[d,2]

[a]Department of Biology, Earlham College, Richmond, IN 47374; [b]Department of Biology, San Francisco State University, San Francisco, CA 94132; [c]Department of Entomology, University of Illinois Urbana–Champaign, Urbana, IL 61801; [d]School of Life Sciences, Arizona State University, Tempe, AZ 85287; [e]Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742; [f]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112; [g]Department of Biology, McGill University, Montreal, Quebec H3A 1B1, Canada; [h]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland; [i]DOE Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, WI 53706; [j]Department of Bacteriology, University of Wisconsin, Madison, WI 53706; [k]The Johns Hopkins University School of Medicine, Baltimore, MD 21205; [l]Department of Biology, Georgetown University, Washington, DC 20057; [m]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204; [n]Department of Neuroscience, University of Arizona, Tucson, AZ 85721; [o]Department of Environmental Science, Policy and Management, University of California, Berkeley, CA 94720; [p]Department of Entomology, University of Nebraska, Lincoln, NE 68583; [q]Department of Computer Science, San Francisco State University, San Francisco, CA 94132; [r]Center for Molecular Biodiversity, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany; [s]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853; and [t]Department of Biotechnology, Chemistry, and Food Science, Norwegian University of Life Sciences, 1492 Ås, Norway

Edited* by Gene E. Robinson, University of Illinois at Urbana–Champaign, Urbana, IL, and approved November 9, 2010 (received for review June 5, 2010)

We report the draft genome sequence of the red harvester ant, *Pogonomyrmex barbatus*. The genome was sequenced using 454 pyrosequencing, and the current assembly and annotation were completed in less than 1 y. Analyses of conserved gene groups (more than 1,200 manually annotated genes to date) suggest a high-quality assembly and annotation comparable to recently sequenced insect genomes using Sanger sequencing. The red harvester ant is a model for studying reproductive division of labor, phenotypic plasticity, and sociogenomics. Although the genome of *P. barbatus* is similar to other sequenced hymenopterans (*Apis mellifera* and *Nasonia vitripennis*) in GC content and compositional organization, and possesses a complete CpG methylation toolkit, its predicted genomic CpG content differs markedly from the other hymenopterans. Gene networks involved in generating key differences between the queen and worker castes (e.g., wings and ovaries) show signatures of increased methylation and suggest that ants and bees may have independently co-opted the same gene regulatory mechanisms for reproductive division of labor. Gene family expansions (e.g., 344 functional odorant receptors) and pseudogene accumulation in chemoreception and P450 genes compared with *A. mellifera* and *N. vitripennis* are consistent with major life-history changes during the adaptive radiation of *Pogonomyrmex* spp., perhaps in parallel with the development of the North American deserts.

chemoreceptor | de novo genome | eusociality | genomic evolution | social insect

The formation of higher-level organization from independently functioning elements has resulted in some of the most significant transitions in biological evolution (1). These include the transition from prokaryotes to eukaryotes and from uni- to multicellular organisms, as well as the formation of complex animal societies with sophisticated division of labor among individuals. In eusocial insects such as ants, distinct morphological castes specialize in either reproduction or labor (2). Currently, very little is known of the genetic basis of caste and reproductive division of labor in these societies, where individuals follow different developmental trajectories, much like distinct cell lines in an organism (3). The resulting phenotypes, queens and workers, can differ greatly in morphology, physiology, and behavior, as well as in order of magnitude differences in life span and reproductive potential (2). Ants, of all social insects, arguably exhibit the highest diversity in social complexity, such as queen number, mating frequency, and the degree of complexity of division of labor (2), and most social traits have independent origins within the ants, making them well suited to comparative genomic analyses.

The sequencing of the honey bee (*Apis mellifera*) genome marked a milestone in sociogenomics (4, 5), facilitating research on the evolution and maintenance of sociality from its molecular building blocks. Since then, genomes of three closely related species of solitary parasitic hymenopterans, *Nasonia* spp., were published and similarities and differences were extensively discussed in the context of the evolution of eusociality (6). However, *A. mellifera* represents only 1 of at least 10 independent evolutionary origins of eusociality within the order Hymenoptera (7–11), and thus it remains unclear whether differences between the honey bee and *Nasonia* spp. truly reflect differences inherent in sociality. With at least six ant genomes on the horizon (12), among other solitary and social insects, sociogenomic comparisons are likely to yield exciting insights into the common molecular basis for the social lifestyle. Ant genomics will also allow us to gain a better understanding of variation in social organization, of elaborate variations of physical and behavioral divisions of labor, of invasion biology, and of the convergent evolution of life histories and diets. It also remains a major question whether there are many

Author contributions: C.R.S., C.D.S., and J.G. designed research; C.R.S., C.D.S., H.M.R., M.H., A.Z., M.Y., C.H., H.H., M.-J.F., W.G., F.W., and J.G. performed research; C.R.S., C.D.S., H.M.R., M.H., A.Z., M.Y., C.H., H.H., E.A., R.B., E.C., V.C., C.R.C., E.E., C.G.E., V.F., J.D.G., D.G., W.G., K.J.G., D.E.H., A.S.I.V., B.R.J., R.M.J., A.K., J.W.K., K.A.M., M.C.M.-T., M.C.M., J.A.M., R.N., O.N., S.N., R.P.O., J.E.P., R.R., J.T.R., G.S., S.T., C.W.T., N.D.T., L.V., F.W., and J.G. analyzed data; and C.R.S., M.H., and J.G. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Data deposition: The sequences reported in this paper have been deposited in the Hymenoptera Genome Database: http://HymenopteraGenome.org/pogonomyrmex (NCBI Genome Project #45803, Assembly Project ID 45797, Transcriptome Project ID 46577).

See Commentary on page 5477.

[1]C.R.S. and C.D.S. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: jgadau@asu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1007901108/-/DCSupplemental.

evolutionary routes to eusociality, especially at the molecular level, or whether we can extract generalities and rules for the molecular evolution of eusociality (3, 4, 13). Although it is likely that much variation in social structure is due to changes in the regulation of conserved pathways, it is undetermined what, if any, role novel genes or pathways have played in the solitary-to-social transition and diversification of social phenotypes (14).

The genus *Pogonomyrmex* contains species that vary greatly in social organization (15), is among the best studied of ant genera (16, 17), is sister to almost all other genera in the diverse subfamily Myrmicinae (8, 11), and contains species of major ecological importance as granivores in both North and South America (18, 19). Colonies can contain over 10,000 workers and a single multiply mated queen that may live for decades. Some *Pogonomyrmex barbatus* populations have a unique system of genetic queen-worker caste determination (Fig. 1) where individuals are essentially hard-wired to develop as either queens or workers, a contrast to environmentally determined diphenism (20–24) (*SI Appendix*, *Chapter 1*). As a consequence, individuals can be genotyped using genetic markers to determine their caste even before caste differentiation. This unique system of caste determination provides a means of studying the genes and regulatory networks used in caste determination.

Results and Discussion

Genome coverage is 10.5–12× on the basis of the estimates of genome size for *Pogonomyrmex* ants as 250–284 Mb (25). The assembly consists of 4,646 scaffolds (mean contig/scaffold: 7.22) spanning 235 Mb (~88%) of the genome that harbor 220 Mb (~83%) of DNA sequence (15 Mb of which are gaps within scaffolds). The N50 scaffold size of the assembly is 793 kb, and the largest scaffold is 3.8 Mb in length; the N50 contig size is 11.6 kb. The transcriptome assembly yielded 7,400 isogroups with a N50 contig size of 1.3 kb.

The MAKER annotation pipeline predicted 16,331 genes and 16,404 transcripts. InterProScan (26) identified additional genes from the in silico prediction programs, which were added to the MAKER predicted genes. The final official gene set, OGS1.1, which was used for computational analyses, consisted of 17,177 genes encoding 17,250 transcripts. Of these, 7,958 (>46%) had complete or partial EST support from the *P. barbatus* transcriptome assembly. The results of the assembly and annotation of the *P. barbatus* genome are well within the range of other insect genomes (Table 1).

More than 1,200 genes have been manually annotated to improve models generated by MAKER (*SI Appendix*, *Chapter 2*) and were used in gene family-centered analyses (see discussion below and *SI Appendix*, *Chapters 3*, *6–8*, *14*, and *16–29*). There are two fundamentally different reasons for our choice of gene families: One set comprises highly conserved gene families for quality assessment (e.g., sequencing error, genome completeness), whereas the second set is based on biologically interesting functional groups associated with the evolution and regulation of social behavior or adaptations of *P. barbatus* to a desert seed-harvesting lifestyle.

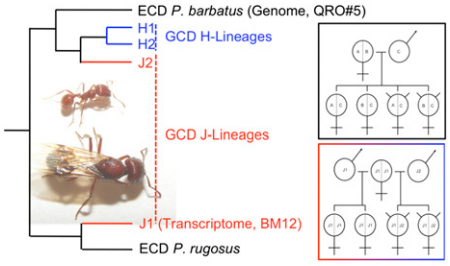


Fig. 1. A pictorial description of the phylogenetic position of the samples used for the genome and transcriptome sequencing, with each put in the context of environmental and genetic caste determination (for a more complete phylogenetic tree, see *SI Appendix*, *Chapter 1*). The dependent lineages (H1/H2 or J1/J2) obligately co-occur because hybridization between them is necessary to produce workers, although within either J or H, the constituent lineages are reproductively isolated because interlineage hybrids cannot become queens (red/blue box). In the boxes to the right, workers are represented by "horned" female symbols. In all *P. barbatus*, the queen mates multiply; polyandry in genetic caste determining (GCD) colonies is obligate to produce both female castes (queens originate from intralineage matings and workers from interlineage matings). In environmental caste determination (ECD), alleles from any father have an equal chance to be in queens or workers (black box). Photo of gyne and worker *P. barbatus* by C. R. Smith.

Quality of Genome Assembly. The core eukaryotic gene-mapping approach (CEGMA) (27) provides a method to rapidly assess genome completeness because it comprises a set of highly conserved, single-copy genes, present in all eukaryotes. In *P. barbatus*, 245 of the 248 (99%) CEGMA genes were found, and 229 of the 248 genes were complete (92%). Cytoplasmic ribosomal protein genes are another highly conserved set of genes that are widely distributed across the physical genome in animals (28, 29). A full complement of 79 proteins was found within the *P. barbatus* genome encoded by 86 genes (*SI Appendix*, *Chapter 6*). Because ribosomal proteins are highly conserved, their manual annotation also provided an estimate of sequencing errors, such as frameshift-inducing homopolymers (a potential problem inherent to pyrosequencing) (30). Six erroneous frameshifts were found in ribosomal protein genes (only one homopolymer); extrapolating from the number of nucleotides encoding the ribosomal genes suggests that 1 in 7,200 coding nucleotide positions (0.014%) may be affected by frameshifts. Analyses of other highly conserved gene families, including the oxidative phosphorylation (31) pathway and the *Hox* gene cluster (32, 33), also suggest high coverage and good genome assembly (*SI Appendix*, *Chapters 7 and 8*). Interestingly, the mitochondrial genome did not auto-assemble into scaffolds greater than 2 kb, but 71% of the mitochondrial genome could be manually assembled with the longest contig containing 5,835 bp (*SI Appendix*, *Chapter 9*, Dataset S1). The largest missing fragment of the mitochondrial genome is typically very high in AT content (96% in *A. mellifera ligustica*) (34) and may not have sequenced due to PCR biases.

In silico-predicted gene models gain significant support through EST sequences. Another way to confirm predicted gene models is a proteomics approach, which has the additional benefit that it demonstrates that a gene is not only transcribed but also translated. A proteomic analysis of the poison gland and antennae confirmed 165 gene and protein models with at least two peptides (*SI Appendix*, *Chapter 10*). It also resulted in the identification of proteins likely associated with nest defense (poison gland) and chemoperception (antenna).

Chromosomal coverage in the current draft assembly was assessed by the identification of telomeres. Most insects outside of the Diptera have telomeres consisting of TTAGG repeats. On the basis of the karyotype data ($n = 16$), we expected 32 telomeres in *P. barbatus* (35). We searched the assembled genome and mate pair reads for TTAGG repeats and extended these where possible (6). In total, 27 of the expected 32 telomeres (88%) were found (*SI Appendix*, *Chapter 11*). These telomeres are even simpler than those of *A. mellifera* (36). Whereas most other insect telomeres commonly include retrotransposon insertions, these seem to be absent from the telomeres of *P. barbatus*.

Genome-Wide Analyses. The mean GC content of the *P. barbatus* genome is 36.5% and the mean ratio of observed-to-expected CpG [CpG(o/e)] is 1.57, both of which are within the ranges reported for other Hymenoptera (5, 6). We define compositional domains as the sequence stretches of variable lengths that differ widely in their GC compositions. A comparison of GC compositional-domain lengths among insects shows that *P. barbatus* and *A. mellifera* have

Table 1. Comparison of metrics for recently sequenced insect genomes

Species	Order/name	Fold coverage	N50 scaffold (kb)	No. of genes	Gene set	Source
Pogonomyrmex barbatus	Hymenoptera (red harvester ant)	12	793	17,177	OGS1.1	This study
Nasonia vitripennis	Hymenoptera (jewel wasp)	6.8	709	18,822	OGS1.2	(6)
Apis mellifera	Hymenoptera (honey bee)	7.5	362	10,156/21,001	OGS1/OGS2	(5)
Acyrthosiphon pisum	Sternorrhyncha (pea aphid)	6.2	88.5	34,604	OGS1	(37)
Tribolium castaneum	Coleoptera (red flower beetle)	7.3	990	16,404	Consensus set	(38)

similar compositional domain-length distributions (*SI Appendix, Chapter 4*). Among the compared insect genomes, the hymenopterans have the smallest proportion (0.1–0.5%) of long compositional domains (>100 kb) as well as the widest range in GC compositional domains. Similar to the other sequenced hymenopteran genomes, but in contrast to other insect orders, genes in *P. barbatus* occur in the more GC-poor regions of the genome. Although the mean CpG(o/e) values of hymenopteran genomes are among the highest observed, species-specific patterns of CpG (o/e) within each genome are not consistent between the hymenopterans studied (Fig. 2). The distribution of CpG(o/e) in *P. barbatus* exons is similar to that in insects without CpG methylation (although with greater variance) (39) and suggests little germline methylation despite the presence of a complete methylation toolkit (see below and *SI Appendix, Chapter 24*). We used an indirect method [single nucleotide polymorphisms (SNP) frequency: CpG – TpG] and a direct method [methylation-sensitive amplified fragment length polymorphism (AFLP) assay; *SI Appendix, Chapter 4*] to determine the presence and frequency of active CpG methylation in *P. barbatus*. We found that CpG/TpG (and vice versa) SNPs constitute 84% of all CpG-to-NpG polymorphisms. This is an indirect measure of CpG methylation because it is has been shown that a methylated cytosine in a CpG has a higher probability to mutate into thymine (*SI Appendix, Chapter 30*). The more direct measure of CpG methylation comes from an AFLP analysis that used methylation-sensitive and -insensitive restriction enzymes. In a comparison of 209 individuals from every female and developmental caste, 33% of all AFLP fragments showed a signature of methylation (*SI Appendix, Chapter 4*). These findings suggest a role of DNA methylation in genome regulation, but additional data are necessary to confirm these predictions and discern the biological role of DNA methylation in *P. barbatus*.

Gene ontology analyses detected significant enrichments in genes associated with sensory perception of smell, cognition, and neurological processes (*SI Appendix, Chapter 5*). These enrichments may reflect the heavy reliance on chemical communication in ants. Consistent with this and detailed analyses of chemosensory and cytochrome P450 gene families (see below), a gene orthology analysis including *Drosophila melanogaster*, *A. mellifera*, and *Nasonia vitripennis* found expansions of genes involved in responses to chemical stimuli and electron transport. The orthology analysis also found a small fraction of genes (3.2% of those in the analysis) common to both social insects studied (*SI Appendix, Chapter 5*); these genes may be important in processes related to the evolution or maintenance of sociality.

Repetitive DNA. Previous results for the *A. mellifera* (5) and *N. vitripennis* (6) genomes illustrate two extreme cases of genomic repeat composition for Hymenoptera: *A. mellifera* is devoid of all except a few *mariner* (40) and rDNA-specific R2 (41) transposable elements whereas *N. vitripennis* has an unusual abundance of repetitive DNA (6). The *P. barbatus* genome assembly contains 18.6 Mb (8% of genome) of interspersed elements (*SI Appendix, Chapter 12*). A total of 9,324 retroid element fragments and 13,068 DNA transposons were identified; however, the majority of interspersed elements (55,373, 8.8 Mb, 3.75% of genome) could not be classified into a specific transposable element family. Gypsy/DIR1 and L2/CR1/Rex elements were the most abundant transposable elements; however, we discovered most families of known insect retrotransposable elements. Nearly 1% (269 loci/1 Mb) of the scaffolded genome is microsatellite DNA (*SI Appendix, Chapter 13*), greater than in most insects (42), which are valuable markers for mapping and population genetic studies.

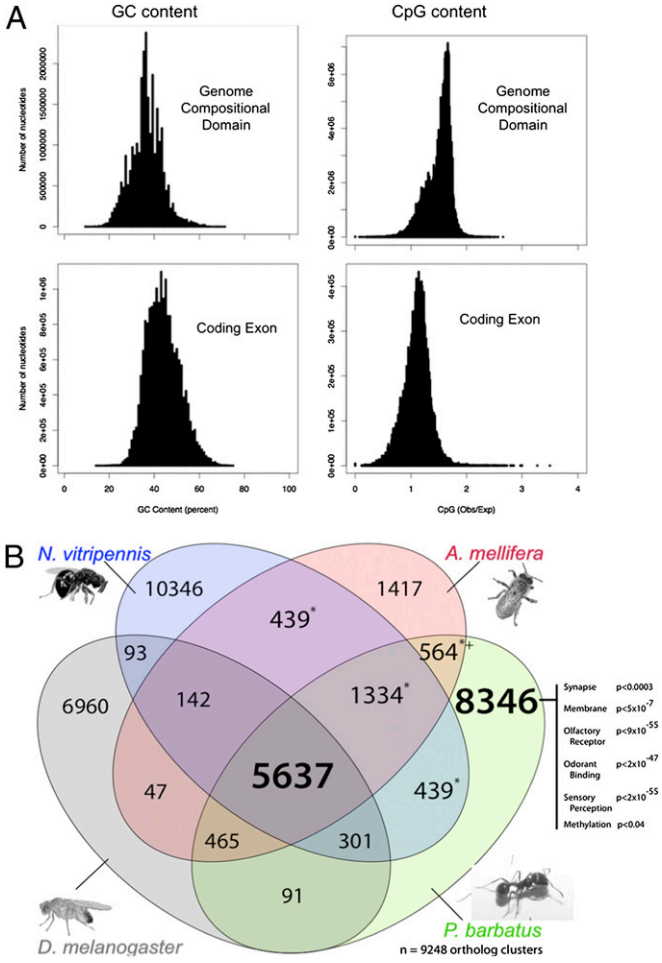


Fig. 2. Genome-wide analyses of nucleotide and relative gene content. (*A*) Synopsis of GC and CpG(o/e) content of the *P. barbatus* genome. (*Upper* panels) Comparison of genome regions with the same GC composition. (*Lower* panels) Comparison of the same features for exons. These distributions are similar to those found in other hymenopterans, except that *P. barbatus* shows no evidence of bimodality in CpG(o/e) for either exons (like *A. mellifera*) or introns (like *N. vitripennis*) (for comparisons, see *SI Appendix, Chapter 4*). (*B*) A Venn diagram displaying overlap in orthologous genes in three hymenopteran and one dipteran insect (for a detailed description of the method, see *SI Appendix, Chapter 5*). A subset of gene ontology terms significantly enriched in *P. barbatus* are displayed at the right. (*) Hymenoptera-specific genes; (+) social Hymenoptera-specific genes.

Chemoreceptor Gene Family Expansions. One special focus of the manual annotation was the proteins involved in chemoperception, which plays an important role in colony communication, a cornerstone of social living. Below we report insights derived from four gene families involved in chemoreception: the ionotropic receptors (IRs), gustatory receptors (Grs), odorant receptors (Ors), and cytochrome P450s.

The IR family in *P. barbatus* consists of 24 genes, compared with 10 in *A. mellifera* and 10 in *N. vitripennis* (43). Phylogenetic analysis and sequence comparison of IRs identified putative orthologs of conserved IRs that are present in other insect genomes and that are expressed in insect antennae (e.g., IR25a, IR8a, IR93a, IR76b) (44), but a number of ant-specific divergent IRs display no obvious orthology to other hymenopteran or insect receptors (*SI Appendix*, *Chapter 14*). Some of these IRs may fulfill contact chemosensory functions by analogy to the gustatory neuron expression of species-specific IRs in *D. melanogaster* (43).

The *P. barbatus* Gr family contains 73 genes compared with just 11 in *A. mellifera* and 58 in *N. vitripennis*. Phylogenetic analysis of the Gr proteins (*SI Appendix*, *Chapter 14*) supports several conclusions about the evolution of this gene family. *A. mellifera* has lost multiple Gr lineages and failed to expand any of them (45, 46), but gene losses are not restricted to *A. mellifera*, with some occurring in *N. vitripennis* and/or *P. barbatus*. The existence of at least 18 Gr lineages is inferred, with *A. mellifera* having lost function in 10 of them, *P. barbatus* in 4, and *N. vitripennis* in 5. *P. barbatus* has expanded two gene lineages independently of the two expansions seen in *N. vitripennis*. Expansion A is considered to be orthologous to the NvGr48-50 gene lineage and a large set of ≈50 highly degraded pseudogenes in *A. mellifera* (represented by AmGrX-Z), and expansion B is somewhat younger. We hypothesize that these are bitter taste receptors that lost function in *A. mellifera* at the time at which they transitioned to nectar feeding, ≈100 Mya (47). Bitter taste perception may be essential for *P. barbatus* to avoid unpalatable seeds (e.g., plant secondary compounds).

The Or family also appears to be considerably expanded in *P. barbatus*, with 344 apparently functional genes among a total of 399 genes (the largest total known for any insect) compared with a total of 166 in *A. mellifera* and 225 in *N. vitripennis* (Dataset S2). We counted 365 ± 10 and 345 ± 10 glomeruli in five queens and five workers, respectively (*SI Appendix*, *Chapter 15*), supporting an ≈1:1 relationship of Or genes to glomeruli resulting from convergence of the axons of all neurons expressing a particular Or on one glomerulus (48, 49). A particularly large expansion of a nine-exon gene subfamily to 169 genes suggests that these genes might comprise the cuticular hydrocarbon receptors (*SI Appendix*, *Chapter 14*). Cuticular hydrocarbons have gained many novel functions important in the context of social behavior, such as colony recognition and queen signaling (50, 51).

P. barbatus has 72 genes in the cytochrome P450 superfamily, compared with 46 in *A. mellifera* and 92 in *N. vitripennis* (5, 6). P450 subfamilies involved in detoxification of xenobiotics show some expansion, whereas those implicated in pheromone metabolism are enigmatically less expanded (*SI Appendix*, *Chapter 16*).

Evolutionary Rate and Pseudogene Accumulation. An evolutionary rate analysis based on amino acid substitutions of the three hymenopteran species with a genome sequence, with *D. melanogaster* as an outgroup, showed that a significant part of the *P. barbatus* genome (4,774 orthologous genes conserved over approximately 350 million y) evolves at a similar rate as the *A. mellifera* genome, and the *A. mellifera* and *P. barbatus* genomes show slightly higher substitution rates than the *N. vitripennis* genome (Fig. 3 and *SI Appendix*, *Chapter 31*). This analysis suggests that the slow evolutionary rate reported for *A. mellifera* may not be associated with sociality, but rather is specific to the Hymenoptera.

A notable feature of *P. barbatus* chemosensory and P450 genes is that the pseudogenes commonly have multiple major mutations suggesting that they are mostly "middle-aged" pseudogenes. Normally a range of pseudogene ages can be inferred in the chemoreceptor gene families from young pseudogenes with single mutations to gene fragments. We estimated the relative ages of the pseudogenes in Ors, Grs, and cytochrome P450s in *P. barbatus*, *A. mellifera*, and *N. vitripennis* by counting the number of obvious pseudogene-causing ("pseudogenizing") mutations per gene (stop codons, intron boundary mutations, small frameshift insertions or deletions, or large insertions or deletions). As shown in Fig. 3, there is a contingent of considerably older pseudogenes in these gene families in *P. barbatus*. The pattern in *P. barbatus* is in contrast to *A. mellifera* and *N. vitripennis*, which have a greater number of young pseudogenes. We hypothesize that the ant lineages that gave rise to *P. barbatus* experienced a major change in chemical ecology ≈10–30 Mya, possibly as a consequence of the increase in elevation of the Sierras and Andes to their present height (52, 53). These western mountain ranges created rain shadows on their eastern sides and spawned the great American deserts. The North American members of the genus *Pogonomyrmex* underwent a significant radiation adapting to these new habitats (16), so the gene expansions in the chemoreceptors and P450s might be adaptations to novel seeds and plant families and their associated toxic components and chemical signatures. Accumulated pseudogenes may therefore reflect a shift toward a more specialized diet concurrent with the adaptive radiation of *Pogonomyrmex* spp. (54).

Innate Immunity Genes. Social insects live in dense groups with high connectivity, putting them at increased risk for disease outbreaks, but they also have social immunity to minimize the introduction and spread of pathogens (55, 56). Very efficient social defenses (e.g., hygienic behaviors) or novel immune pathways were hypotheses put forth to explain the presence of few (roughly half) innate immunity genes in *A. mellifera* compared with *D. melanogaster* (and more recently in the red flour beetle, *Tribolium castaneum*) (5, 38). However, the more recently sequenced genomes of *N. vitripennis* (6) and *Acyrthosiphon pisum* (pea aphid) (37) also have "depauperate" complements of immune genes relative to flies and beetles, which suggests that the gene complement of flies and beetles might be a derived condition within insects. Indeed, the number of innate immune genes in *P. barbatus* is more similar to the other hymenopterans (*SI Appendix*, *Chapter 17*). Although all of the major signaling pathways are present in *P. barbatus* (IMD, Toll, Jak/STAT, and JKN), only a few recognition proteins were identified, which suggests either a highly focused immune system or an alternative unknown pathogen recognition system. Interestingly, we found expansions of antimicrobial peptides relative to *A. mellifera*. These expansions may correspond to a transition to living within the soil and an increased exposure to bacterial and fungal pathogens.

Developmental Networks and Polyphenism. The production of alternative phenotypes during development may occur through the regulation of several key nodes in specific networks during development (57–59). In ant colonies, queens and workers fill divergent adaptive roles—dispersal and reproduction vs. colony maintenance—and their functional differences are reflected in differences in morphology, physiology, and behavior, such as in wings and ovaries. *P. barbatus* workers are completely devoid of wings at the adult stage and have ovaries a fraction of the size of the queen's. In analogy to honey bees (60), we hypothesized that CpG DNA methylation may play a role in the differential regulation of genes in the wing and reproductive development networks of workers and queens. This hypothesis was computationally evaluated by examining the CpG dinucleotide content (39) of wing and reproductive developmental pathway genes relative to the genome (*SI Appendix*, *Chapter 18*). These developmental networks contain significantly fewer CpGs than random genes, suggesting

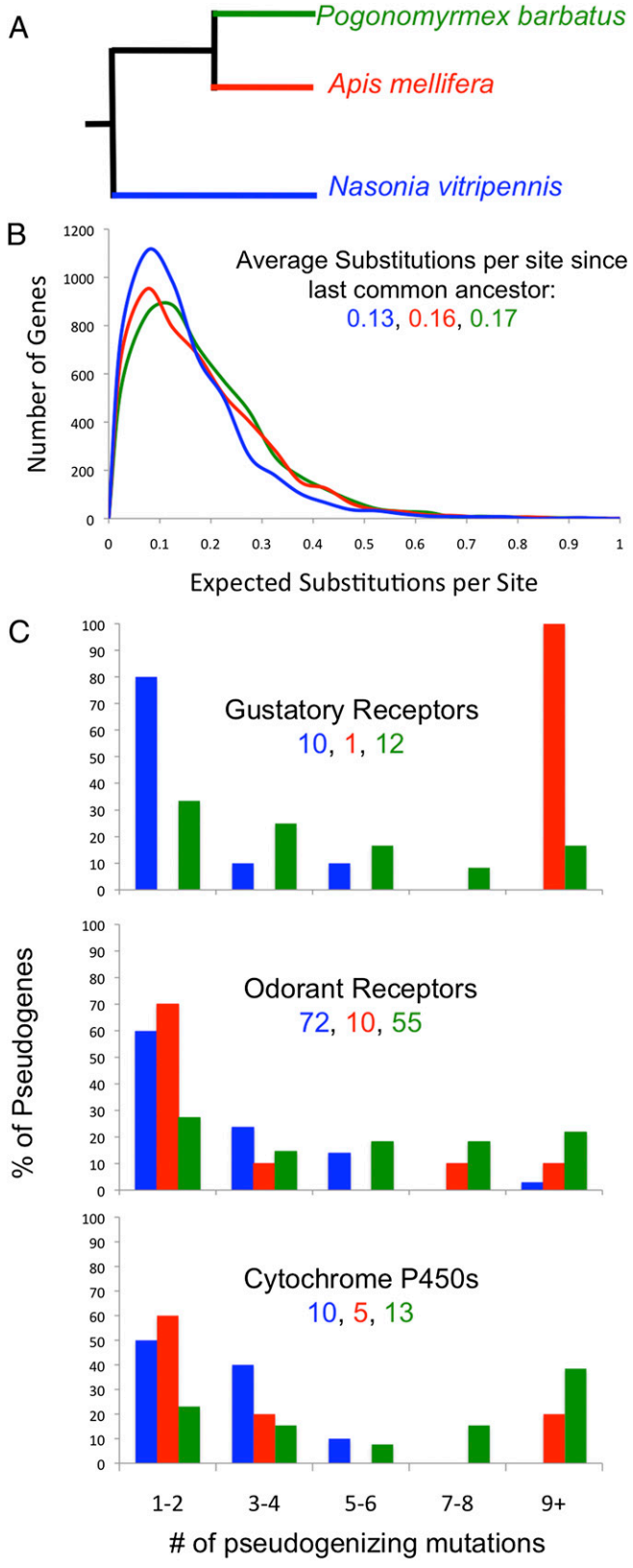


Fig. 3. Evolutionary rate and the accumulation of pseudogene-causing ("pseudogenizing") mutations in three gene families in the ant *P. barbatus* (green), the honey bee *A. mellifera* (red), and the jewel wasp *N. vitripennis* (blue). (*A*) The relationships among analyzed taxa. (*B*) A comparison of the evolutionary rates based amino acid substitutions in a set of 4,774 orthologs shared among the three species and *D. melanogaster* (the outgroup). (*C*) The accumulation of pseudogenizing mutations in three ecologically relevant gene families (Gr, Or, and cytochrome P450s). The number of pseudogenes found in each species is below the gene family name in each panel. Only one gene represents the Grs in *A. mellifera*; all other *A. mellifera* Gr pseudogenes had accrued a very high number of mutations and most are fragments. Of those analyzed here, the pseudogenes in *P. barbatus* tend to be much older than those in *A. mellifera* and *N. vitripennis* (ANOVA: $F_{2,156} = 4.7$, $P = 0.01$).

that they are more methylated than most genes because methylated cytosines are more prone to deamination (6, 39, 61). These results are in contrast to data on *A. mellifera*, where housekeeping genes are the main targets of methylation (39, 61) (which is also in contrast to vertebrates), and suggest a potentially divergent role of methylation in harvester ants compared with honey bees.

Gene Regulation and Reproductive Division of Labor. Various gene families/pathways were specifically targeted for manual annotation because of their known role in queen-worker caste determination (3). These families/pathways included the insulin/TOR-signaling pathway (*SI Appendix*, *Chapter 19*), yellow/major royal jelly genes (*SI Appendix*, *Chapter 20*), biogenic amine receptors (*SI Appendix*, *Chapter 21*), and hexamerin storage proteins (*SI Appendix*, *Chapter 19*). These candidate caste genes will be targeted for studying gene expression differences between castes using RNAi. The RNAi pathway is intact in *P. barbatus* (*SI Appendix*, *Chapter 22*), and RNAi has already been successfully implemented in another ant (62).

Similar to the other sequenced hymenopterans, *P. barbatus* has a full methylation toolkit (*SI Appendix*, *Chapter 24*). All three DNA methyltransferase genes (*Dnmt1–3*) and three methyl-binding proteins (*MBD*) are present in *P. barbatus*, but interestingly there is only a single copy of *Dnmt1* compared with two in *A. mellifera* and three in *N. vitripennis* (6). The loss of multiple copies of maintenance methyltransferase(s) in ants may have implications for the inheritance of epigenetic information.

We analyzed genes within 100 kb of four microsatellite markers diagnostic for the J-lineages (63) with the hypothesis that some genes physically linked to the markers may cause the incompatibility between the lineages that leads to the loss of phenotypic plasticity and genetic caste determination (24) (*SI Appendix*, *Chapter 19*). One interesting candidate from this analysis, *lozenge* (*lz*), has many described mutants in *D. melanogaster*, including sterility due to a loss of oogenesis and a spermathecum (64–67), two traits characteristic of worker ants.

Materials and Methods

Genome Sequencing and Assembly. The genome and transcriptome of *P. barbatus* were sequenced entirely on the 454 XLR titanium platform at SeqWright. Five runs were dedicated to unpaired shotgun reads on DNA isolated from a single haploid male ant, which generated over 6 million reads averaging 370 bp in length (after trimming). Two runs used 8-kb paired-end libraries based on DNA from four brothers of the previous male ant; this yielded a total of nearly 2.9 million reads, each averaging 262 bp in length (after trimming). The assembly presented in this paper was created by a CABOG 5.3 (68) open source assembler. We substituted the OVL overlap module for the recommended MER overlapper for performance reasons (see CABOG documentation at http://sourceforge.net/apps/mediawiki/wgs-assembler).

The transcriptome was sequenced using a single 454 titanium run, which generated 10.4 Mb of sequence across 726,000 reads. The transcriptome was assembled using the Newbler v2.3 assembly software (Roche).

The genome of *P. barbatus* was annotated with the automatic annotation pipeline MAKER (69). The ab initio predictions of MAKER were further refined to produce an official gene set used for computational analyses (*SI Appendix*, *Chapter 2*). This set (OGS1.1) included all nonredundant ab initio predictions from all gene predictors used by MAKER that were supported by an InterProScan domain (26) and excluded any that were flagged as possible repeat elements. A second official gene set (OGS1.2) was produced to include refined

genes on the basis of manual annotation and has been submitted to NCBI. Manual annotations followed a standard methodology described in the *SI Appendix, Chapter 3*. Detailed methods for specific analyses are given in *SI Appendix, Chapters 4–31*.

ACKNOWLEDGMENTS. A very special thanks to S. Pratt for comments on the manuscript. We are thankful to the Earlham College Evolutionary Genomics class, which annotated genes and did preliminary analyses. R. Jones, B. Mott, and T. Holbrook collected specimens. We are grateful for allocated computer time from the Center for High Performance Computing at the University of Utah. We also thank Mike Wong from the Center for Computing for Life Science at San Francisco State University for assistance with custom scripts and hardware configuration. National Science Foundation Grant IOS-0920732 (to J.G. and C.R.S.) funded the sequencing of the genome, and National Institutes of Health Grant 5R01HG004694 (to M.Y.) funded the MAKER annotation.

1. Maynard Smith J, Szathmáry E (1995) *The Major Transitions in Evolution* (W. H. Freeman/Spektrum, New York).
2. Hölldobler B, Wilson EO (1990) *The Ants* (Belknap Press of Harvard University Press, Cambridge, MA).
3. 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.
4. Robinson GE, Grozinger CM, Whitfield CW (2005) Sociogenomics: Social life in molecular terms. *Nat Rev Genet* 6:257–270.
5. Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.
6. Werren JH, et al.; Nasonia Genome Working Group (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. *Science* 327:343–348.
7. Hines HM, Hunt JH, O'Connor TK, Gillespie JJ, Cameron SA (2007) Multigene phylogeny reveals eusociality evolved twice in vespid wasps. *Proc Natl Acad Sci USA* 104:3295–3299.
8. Brady SG, Schultz TR, Fisher BL, Ward PS (2006) Evaluating alternative hypotheses for the early evolution and diversification of ants. *Proc Natl Acad Sci USA* 103:18172–18177.
9. Brady SG, Sipes S, Pearson A, Danforth BN (2006) Recent and simultaneous origins of eusociality in halictid bees. *Proc Biol Sci* 273:1643–1649.
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. Moreau CS, Bell CD, Vila R, Archibald SB, Pierce NE (2006) Phylogeny of the ants: Diversification in the age of angiosperms. *Science* 312:101–104.
12. Smith CD, Smith CR, Mueller U, Gadau J (2010) Ant genomics: Strength and diversity in numbers. *Mol Ecol* 19:31–35.
13. Toth AL, Robinson GE (2007) Evo-devo and the evolution of social behavior. *Trends Genet* 23:334–341.
14. Page RE, Jr., Amdam GV (2007) The making of a social insect: Developmental architectures of social design. *Bioessays* 29:334–343.
15. Johnson RA (2000) Seed-harvester ante (Hymenoptera: Formicidae) of North America: An overview of ecology and biogeography. *Sociobiology* 36:89–122.
16. Taber SW (1998) *The World of the Harvester Ants* (Texas A&M University Press, College Station, TX).
17. Gordon DM (1999) *Ants at Work* (The Free Press, New York).
18. Pirk GI, Lopez de Casenave J (2006) Diet and seed removal rates by the harvester ants *Pogonomyrmex rastratus* and *Pogonomyrmex pronotalis* in the central Monte desert, Argentina. *Insectes Soc* 53:119–125.
19. MacMahon JA, Mull JF, Crist TO (2000) Harvester ants (*Pogonomyrmex* spp.): Their community and ecosystem influences. *Annu Rev Ecol Syst* 31:265–291.
20. Anderson KE, Linksvayer TA, Smith CR (2008) The causes and consequences of genetic caste determination in ants (Hymenoptera: Formicidae). *Myrmecol News* 11:119–132.
21. Helms Cahan S, et al. (2002) Extreme genetic differences between queens and workers in hybridizing *Pogonomyrmex* harvester ants. *Proc Biol Sci* 269:1871–1877.
22. Julian GE, Fewell JH, Gadau J, Johnson RA, Larrabee D (2002) Genetic determination of the queen caste in an ant hybrid zone. *Proc Natl Acad Sci USA* 99:8157–8160.
23. Volny VP, Gordon DM (2002) Genetic basis for queen-worker dimorphism in a social insect. *Proc Natl Acad Sci USA* 99:6108–6111.
24. Cahan SH, et al. (2004) Loss of phenotypic plasticity generates genotype-caste association in harvester ants. *Curr Biol* 14:2277–2282.
25. Tsutsui ND, Suarez AV, Spagna JC, Johnston JS (2008) The evolution of genome size in ants. *BMC Evol Biol* 8:64.
26. Quevillon E, et al. (2005) InterProScan: Protein domains identifier. *Nucleic Acids Res* 33(Web Server issue):W116–W120.
27. Parra G, Bradnam K, Korf I (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061–1067.
28. Uechi T, Tanaka T, Kenmochi N (2001) A complete map of the human ribosomal protein genes: Assignment of 80 genes to the cytogenetic map and implications for human disorders. *Genomics* 72:223–230.
29. Marygold SJ, et al. (2007) The ribosomal protein genes and Minute loci of *Drosophila melanogaster*. *Genome Biol* 8:R216.
30. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 8:R143.
31. Saraste M (1999) Oxidative phosphorylation at the fin de siècle. *Science* 283:1488–1493.
32. Hughes CL, Kaufman TC (2002) Hox genes and the evolution of the arthropod body plan. *Evol Dev* 4:459–499.
33. Gellon G, McGinnis W (1998) Shaping animal body plans in development and evolution by modulation of Hox expression patterns. *Bioessays* 20:116–125.
34. Crozier RH, Crozier YC (1993) The mitochondrial genome of the honeybee Apis mellifera: Complete sequence and genome organization. *Genetics* 133:97–117.
35. Taber SW, Cokendolpher JC, Francke OF (1988) Karyological study of North-American *Pogonomyrmex* (Hymenoptera, Formicidae). *Insectes Soc* 35:47–60.
36. Robertson HM, Gordon KH (2006) Canonical TTAGG-repeat telomeres and telomerase in the honey bee, *Apis mellifera*. *Genome Res* 16:1345–1351.
37. International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS Biol* 8:e1000313.
38. *Tribolium* Genome Sequencing Consortium et al. (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452:949–955.
39. Elango N, Hunt BG, Goodisman MAD, Yi SV (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proc Natl Acad Sci USA* 106:11206–11211.
40. Robertson HM (1993) The mariner transposable element is widespread in insects. *Nature* 362:241–245.
41. Kojima KK, Fujiwara H (2005) Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* 22:2157–2165.
42. Pannebakker BA, Niehuis O, Hedley A, Gadau J, Shuker DM (2010) The distribution of microsatellites in the *Nasonia* parasitoid wasp genome. *Insect Mol Biol* 19(Suppl 1):91–98.
43. Croset V, et al. (2010) Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet* 6:e1001064.
44. Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB (2009) Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell* 136:149–162.
45. Robertson HM, Wanner KW (2006) The chemoreceptor superfamily in the honey bee, *Apis mellifera*: Expansion of the odorant, but not gustatory, receptor family. *Genome Res* 16:1395–1403.
46. Robertson HM, Gadau J, Wanner KW (2010) The insect chemoreceptor superfamily of the parasitoid jewel wasp *Nasonia vitripennis*. *Insect Mol Biol* 19(Suppl 1):121–136.
47. Poinar GO, Jr., Danforth BN (2006) A fossil bee from Early Cretaceous Burmese amber. *Science* 314:614.
48. Mombaerts P (1999) Molecular biology of odorant receptors in vertebrates. *Annu Rev Neurosci* 22:487–509.
49. Gao Q, Yuan B, Chess A (2000) Convergent projections of Drosophila olfactory neurons to specific glomeruli in the antennal lobe. *Nat Neurosci* 3:780–785.
50. Endler A, et al. (2004) Surface hydrocarbons of queen eggs regulate worker reproduction in a social insect. *Proc Natl Acad Sci USA* 101:2945–2950.
51. Hefetz A (2007) The evolution of hydrocarbon pheromone parsimony in ants (Hymenoptera: Formicidae): Interplay of colony odor uniformity and odor idiosyncrasy. *Myrmecol News* 10:59–68.
52. Poulsen CJ, Ehlers TA, Insel N (2010) Onset of convective rainfall during gradual late Miocene rise of the central Andes. *Science* 328:490–493.
53. Cassel EJ, Graham AA, Chamberlain CP (2009) Cenozoic tectonic and topographic evolution of the northern Sierra Nevada, California, through stable isotope paleoaltimetry in volcanic glass. *Geology* 37:547–550.
54. McBride CS (2007) Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. *Proc Natl Acad Sci USA* 104:4996–5001.
55. Walker TN, Hughes WO (2009) Adaptive social immunity in leaf-cutting ants. *Biol Lett* 5:446–448.
56. Fefferman NH, Traniello JFA (2008) Social insects as models in epidemiology: Establishing the foundation for an interdisciplinary approach to disease and sociality. *Insect Sociology*, eds Gadau J, Fewell J (Harvard University Press, Cambridge, MA), pp 545–571.
57. Davidson EH (2006) The sea urchin genome: Where will it lead us? *Science* 314:939–940.
58. Abouheif E, Wray GA (2002) Evolution of the gene network underlying wing polyphenism in ants. *Science* 297:249–252.
59. Khila A, Abouheif E (2008) Reproductive constraint is a developmental mechanism that maintains social harmony in advanced ant societies. *Proc Natl Acad Sci USA* 105: 17884–17889.
60. Kucharski R, Maleszka J, Foret S, Maleszka R (2008) Nutritional control of reproductive status in honeybees via DNA methylation. *Science* 319:1827–1830.
61. Foret S, Kucharski R, Pittelkow Y, Lockett GA, Maleszka R (2009) Epigenetic regulation of the honey bee transcriptome: Unravelling the nature of methylated genes. *BMC Genomics* 10:472.
62. Lu HL, Vinson SB, Pietrantonio PV (2009) Oocyte membrane localization of vitellogenin receptor coincides with queen flying age, and receptor silencing by RNAi disrupts egg formation in fire ant virgin queens. *FEBS J* 276:3110–3123.
63. Schwander T, Cahan SH, Keller L (2007) Characterization and distribution of *Pogonomyrmex* harvester ant lineages with genetic caste determination. *Mol Ecol* 16: 367–387.
64. Anderson RC (1945) A study of the factors affecting fertility of lozenge females of *Drosophila melanogaster*. *Genetics* 30:280–296.
65. Perrimon N, Mohler D, Engstrom L, Mahowald AP (1986) X-linked female-sterile loci in *Drosophila melanogaster*. *Genetics* 113:695–712.
66. Bloch Qazi MC, Heifetz Y, Wolfner MF (2003) The developments between gametogenesis and fertilization: Ovulation and female sperm storage in *Drosophila melanogaster*. *Dev Biol* 256:195–211.
67. Khila A, Abouheif E (2010) Evaluating the role of reproductive constraints in ant social evolution. *Philos Trans R Soc Lond B Biol Sci* 365:617–630.
68. Miller JR, et al. (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 24:2818–2824.
69. Cantarel BL, et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18:188–196.

Appendix

Smith et al.: A draft genome of the red harvester ant *Pogonomyrmex barbatus*

Table of contents:

Contributions and Affiliations

Project Coordination:

Jürgen Gadau[1], Chris R. Smith[2], Christopher D. Smith[3]

Coordination and Writing of Main Text:

Chris R. Smith[2], Jürgen Gadau[1]

Coordination and Editing of Supplementary Information:

Martin Helmkampf[1], Chris R. Smith[2], Jürgen Gadau[1]

Assembly:

Aleksey Zimin[4], Christopher D. Smith[3]

Automated Annotation:

Carson Holt[5], Hao Hu[5], Mark Yandell[5]

Global Compositional Analysis:

Eran Elhaik[6], Justin T. Reese[7], Dan Graur[8], Chris R. Smith[2], Christine G. Elsik[7]

Gene Ontology and Orthology Analysis:

Christopher D. Smith[3]

Cytoplasmic Ribosomal Protein Genes:

Martin Helmkampf[1]

Oxidative Phosphorylation Genes:

Joshua D. Gibson[1]

Hox Genes:

Monica C. Muñoz-Torres[7], Darren E. Hagen[7], Christine G. Elsik[7]

Mitochondrial Genome Assembly:

Joshua D. Gibson[1], Chris R. Smith[2]

Proteomics:

Florian Wolschin[1,9], Martin Helmkampf[1]

Telomeres:

Hugh Robertson[10]

Repetitive Elements:

Jay W. Kim[3], Christopher D. Smith[3]

Microsatellite DNA Abundance and Diversity:

Oliver Niehuis[11]

Chemosensory Genes:

Hugh Robertson[10] (Ors, OBPs and Grs), Vincent Croset[12], Richard Benton[12] (IRs)

Olfactory Glomeruli:

Wulfila Gronenberg[13]

Cytochrome P450 Genes:

Reed M. Johnson[14]

Immune Genes:

Lumi Viljakainen[15], Kirk J. Grubbs[16]

Wing Polyphenism and Reproductive Division of Labor:

Ehab Abouheif[17], Marie-Julie Favé[17], Vilaiwan Fernandes[17], Rajee Rajakumar[17], Ana Sofia Ibarraran Viniegra[17], Abderrahman Khila[17]

Candidate Caste Determination Genes:

Chris R. Smith[2], Rajendhran Rajakumar[17], Martin Helmkampf[1], Garret Suen[16,18], Cameron Currie[16,18]

Yellow / Major Royal Jelly Protein Genes:

Rick P. Overson[1], Martin Helmkampf[1]

Biogenic Amine Receptor Genes:

Julie A. Mustard[1]

RNAi Pathway Genes:

Shu Tao[7], Monica C. Muñoz-Torres[7], Jennifer E. Placek[3], Christopher D. Smith[3], Christine G. Elsik[7]

MicroRNAs:

Darren E. Hagen[7], Christine G. Elsik[7]

DNA Methylation Toolkit:

Chris R. Smith[2]

Delta-9 Desaturase Genes:

Elizabeth Cash[1], Martin Helmkampf[1]

Olfactory Learning and Memory:

Kaitlyn A. Mathis[19], Brian R. Johnson[19]

Opsins and Circadian Genes:

Neil D. Tsutsui[19]

Behavior and Aggression Genes:

Candice W. Torres[19], Rin Nakamura[3]

Earlham College Evolutionary Genomics Class Annotation:

Chris R. Smith et al.[2]

SNP Analysis:

Surabhi Nigam[20], Marguerite C. Murphy[20], Christopher D. Smith[3]

Evolutionary Rate Analysis:

Carson Holt[5], Christopher D. Smith[3]

[1] School of Life Sciences, Arizona State University, Tempe, AZ 85287, United States

[2] Department of Biology, Earlham College, Richmond, IN 47374, United States

[3] Department of Biology, San Francisco State University, San Francisco, CA 94132, United States

[4] Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, United States

[5] Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, United States

[6] Johns Hopkins University School of Medicine, Baltimore, MD 21205, United States

[7] Department of Biology, Georgetown University, Washington, DC 20057, United States

[8] Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, United States

[9] Department of Biotechnology, Chemistry, and Food Science, Norwegian University of Life Sciences, Ås, Norway

[10] Department of Entomology, University of Illinois Urbana-Champaign, Urbana, IL 61801, United States

[11] Center for Molecular Biodiversity, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany

[12] Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

[13] Department of Neuroscience, University of Arizona, Tucson, AZ 85721, United States

[14] Department of Entomology, University of Nebraska, Lincoln, NI 68583, United States

[15] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, United States

[16] Department of Bacteriology, University of Wisconsin, 1550 Linden Dr., Madison, WI, 53706, USA

[17] Department of Biology, McGill University, Montreal, Quebec, Canada

[18] DOE Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, WI 53706, United States

[19] Department of Environmental Science, Policy and Management, University of California-Berkeley, Berkeley, CA 94720, United States

[20] Department of Computer Science, San Francisco State University, San Francisco, CA 94132, United States

Supplemental Acknowledgments

Work highlighted in this supplement was supported by the following grants: Natural Sciences and Engineering Research Council Canada grant to E. Abouheif. Mass spectrometric raw data was acquired by the Arizona Proteomics Consortium supported by NIEHS grant ES06694 to the SWEHSC, NIH/NCI grant CA023074 to the AZCC and by the BIO5 Institute of the University of Arizona. F. Wolschin was supported by the Norwegian Research Council (180504), and the Binational Science Foundation (2007465). We would like to thank Michael Goodisman (Georgia Tech) for discussion, Navdeep Mutti (Airzona State University) for assistance with the msAFLP analysis, Nicolas Dowdy (University of Arizona) for help with brain sectioning and counting of glomeruli, as well as Brendan Hunt (Georgia Tech) and Soojin Yi (Georgia Tech) for sending us the *Drosophila* data to facilitate calculations of the genome-wide CpG[o/e].

Supporting Chapters

1. Source of Sequenced Materials

The source for the genome sequence of *Pogonomyrmex barbatus* was males (brothers) from a single colony collected in Querétaro, México (QRO#5: 20.6663, –100.0706). This population displays environmental caste determination, and mitochondrial sequencing places it as sister to the majority of the genetic caste determining lineages (Fig. S1). DNA extractions were done using the Qiagen DNA blood and tissue kit using the protocol for insects (Qiagen, Inc.). The transcriptome was generated from individuals from a single colony near Portal, New Mexico (BM12: 31.9237, –109.0877). The population from which the transcriptome originates has genetic caste determination. Nuclear and mitochondrial genotyping suggest that the queen of BM12 is of the J1 lineage (Fig. S1). The transcriptome was generated from a combination of whole body RNA extractions using TRIzol reagent (Invitrogen) and the column-based PureLink Total RNA Purification system (Invitrogen). Various life stages (larvae of two stages, pupae, and adult) and castes (queen and worker) were pooled for the RNA isolation.

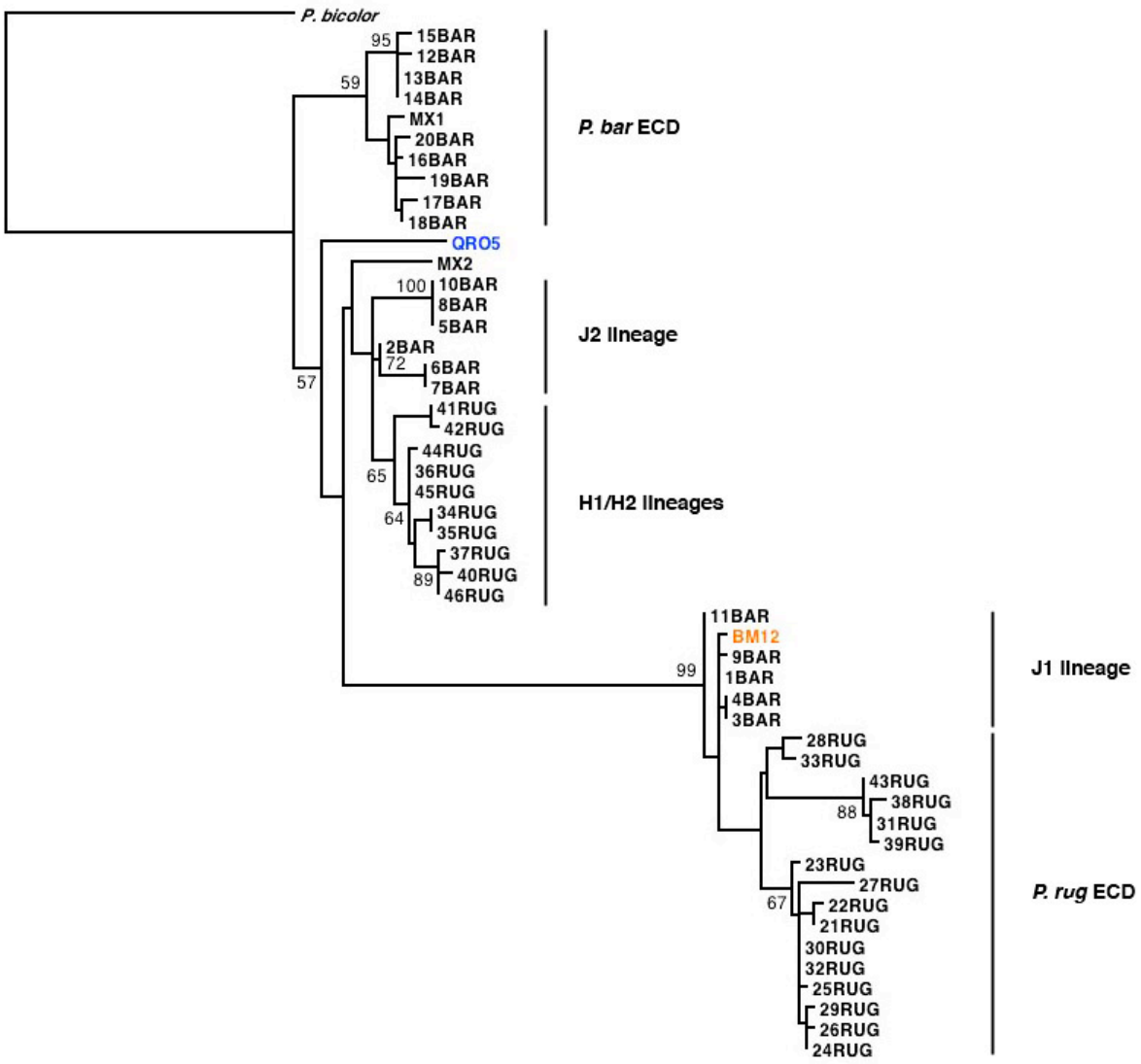


Fig. S1. Maximum likelihood tree depicting the phylogenetic position of the colonies that supplied specimens for the sequencing of the genome (QRO#5) and transcriptome (BM12) within the *Pogonomyrmex barbatus / rugosus* species complex. The tree is based on a 593 bp fragment of the mitochondrial *cytochrome c oxidase* gene obtained from specimens of 50 colonies living in the southwestern USA and Mexico (*P. bicolor* serves as the outgroup). Sequences were taken from the literature (1) or determined for this study (QRO#5, BM12). Calculations were performed with RAxML v7.0.4 (2) under the GTR+G model, employing 100 maximum likelihood searches and 1000 thorough bootstrap replicates (support values > 50 are drawn to the nodes of the best maximum likelihood tree found). This tree confirms results independently obtained from microsatellite genotyping, namely that BM12 belongs to a population characterized by genetic caste determination, while QRO#5 does not.

2. MAKER Annotation

MAKER first identifies repetitive elements using the programs RepeatMasker (www.repeatmasker.org) and RepeatRunner (3). Next MAKER aligns expressed sequence tags (ESTs) from the same organism with blastn, protein evidence with blastx and ESTs from related organisms with tblastx (4). In order to make sure that splice sites are correct, the resulting sequence alignments are refined with Exonerate (5). MAKER then identifies and removes redundant and spurious alignments, and runs a battery of trained gene predictors – Augustus (6), Snap (7), and GeneMark (8) – to produce evidence-based predictions. These predictions are further evaluated by MAKER for consistency with the evidence and other predictions (9), with the best prediction chosen as the final annotation.

The output format of MAKER is a set of GFF3 files (10) (www.gmod.org) containing structural information including UTRs and exon/intron boundaries for each gene, and also the evidence used to generate these annotations, including repeat content, homology evidence and *ab initio* predictions. Each gene annotation also comes with a quality score ranging from 0 to 1 that evaluates the support level.

To identify repetitive DNA, the repeat library for *Pogonomyrmex barbatus* was combined from RepeatMasker's RepBase library and the novel repeats modeled by RepeatModeler (www.repeatmasker.org/RepeatModeler.html) and PILER-DF (11) for *P. barbatus* and *Linepithema humile* genomes (Table S1). For the protein evidence, we combined UniProtKB database (12), *Drosophila melanogaster* proteome from FlyBase (http://flybase.org), annotated *Apis mellifera* and *Nasonia vitripennis* proteins and insect chemosensory proteins from GenBank (13). The N50 of our *P. barbatus* EST collections is large (1192 bp), which is essential for high-quality annotation. We also used the GenBank hymenopteran and *L. humile* ESTs as additional tblastx evidence for annotation.

Prior to running MAKER, we independently trained the three *ab initio* predictors Augustus, Snap, and GeneMark. Augustus was trained with its self-training pipeline autoAug.pl and the *P. barbatus* ESTs, Snap initially with a core eukaryotic gene set predicted by CEGMA (14) and then further with MAKER predictions using its bootstrap functionality (15), and GeneMark with 20 Mb genomic sequence (average contig length 3.2 Mb).

The MAKER annotation was performed on a 24-processor workstation with Intel Xeon X7460 2.66 GHz processors, and took around 8000 CPU hours in total. In sum, 16,331 genes were predicted. Additional 38,260 *ab initio* predictions that overlap no homology evidence were generated. Following automatic annotation with MAKER a further updated *P. barbatus* and *L. humile* repeat library was used to filter out genes that had greater than 50% coverage with a known repeat. This removed 91 of the original MAKER automatic annotations. We then ran InterProScan (16) over the set of *ab initio* gene predictions that did not overlap a MAKER annotation, and found 937 (2.4%) of them contained an InterPro protein domain,

indicating that they are likely to be authentic genes. These genes were added to the final annotation set for a total set of 17,177 predicted genes encoding 17,250 transcripts. The MAKER-generated annotations were then subjected to further human review and curation (see next chapter and chapters on individual gene groups).

Table S1. *De novo* repeat library summary

	No. of RECON predictions	No. of RepeatScout predictions	No. of PILER-DF predictions	Total
Raw output	402	157	84	643
Redundant sequences	19	22	39	70
False positives	8	0	2	10
Final *de novo* repeat library	375	135	43	553

3. Manual Annotation

Manual annotation for specific gene families or functional groups was conducted according to a standardized protocol (for an overview of the gene families, see Table S2). All members of the focal gene families in *Drosophila melanogaster* were selected from FlyBase (http://flybase.org); if focal genes were not present in *D. melanogaster* they were identified from other genome datasets. The BLAST package (17) was used to identify genomic scaffold regions and gene models from the Official Gene Set v1.1 produced by MAKER.

Apollo (18), a sequence annotation editor linked to a Chado database, was employed to confirm and edit the predicted gene models. Among others, the following components were individually evaluated: completeness of the coding domain sequences, untranslated regions, intron-exon boundaries, and sequencing errors resulting in frameshifts. Additional information sources for this process were homologous genes from other holometabolous insects and the *Pogonomyrmex barbatus* EST dataset.

Finally, the homology relation of each manually annotated gene prediction to its reference gene was assessed. For this purpose, the NCBI non-redundant protein database was queried with the predicted protein sequence using the blastp program. Best reciprocal BLAST hits were interpreted as orthologs (19).

Where appropriate, methods that depart from the above are described in detail for specific gene families and functional groups.

Table S2. Manually annotated *Pogonomyrmex barbatus* gene families and functional groups

Gene family / functional group	No. of genes annotated
Aggression	6
Biogenic Amine Receptors	20
Candidate caste determination	24
Chemoreception	
Odorant Receptors	399[a]
Odorant Binding Proteins	15
Gustatory Receptors	73
Ionotropic Receptors	24
Cytoplasmic Ribosomal Proteins	89[b]
Delta-9 Desaturases	10
Developmental pathways	
Hox	10
Wing development	73
Reproductive development	44
DNA CpG Methyltransferases	3
Immune system	97
Ionotropic Glutamate Receptors	10
miRNA	69[c]
Olfactory learning and memory	59
Opsins and circadian rhythm	10
Oxidative Phosphorylation	76
P450 cytochromes	72
RNAi	30
Williams-Beuren Syndrome	17
Yellow / Major Royal Jelly Proteins	16
Other	50
Total	1296

[a] Including putative pseudogenes.

[b] Of these, only 86 are counted among cytoplasmic ribosomal proteins proper (see chapter 6).

[c] Additional genes are currently being annotated.

4. Global Compositional Analysis

Animal genomes are not uniform in their long-range sequence composition, but are composed of a mosaic of compositional domains, i.e., homogeneous and non-homogeneous sequence stretches of variable lengths that differ widely in their GC compositions. Compositionally homogeneous domains are also referred to as "GC-content domains" (20), while a subset of long (≥ 300 kb) compositionally homogeneous domains are traditionally termed "isochores" (21). In all animals studied so far, the distribution of compositional-domain lengths showed an abundance of short domains and a paucity of long ones. The genome of the *Pogonomyrmex barbatus* is no exception in this respect.

A comparison of the distributions of compositional-domain lengths among *P. barbatus* (red harvester ant), *Apis mellifera* (honey bee), *Nasonia vitripennis* (jewel wasp), *Tribolium castaneum* (red-flour beetle), *Anopheles gambiae* (African malaria mosquito), and *Drosophila melanogaster* (fruit fly) shows that *P. barbatus* and *A. mellifera* have similar domain-length distributions (Fig. S2). By contrast, *D. melanogaster* exhibits the lowest abundance of very short domains (< 5 kb) and the highest abundance of medium-long domains (> 10 kb), whereas *T. castaneum* exhibits the opposite pattern. Using a goodness-of-fit test, we determined that none of the above six distributions of domain lengths is similar to any other ($P < 0.05$).

Hymenopterans have the smallest proportion (0.1–0.5%) of long compositional domains (> 100 kb), whereas *T. castaneum* and the dipterans have the largest proportion (> 0.6%) (Table S3). Among the three hymenopterans, *P. barbatus* has the highest proportion of long compositional domains (0.5%). There are six isochoric (≥ 300 kb) domains in the *P. barbatus* genome, compared to 2–4 isochoric domains in other hymenopterans. These isochoric domains cover 0.4–1.5% of the genome in the Hymenoptera. By contrast, 15–18 isochoric domains cover 3–10% of the genome in *T. castaneum* and the two dipteran species.

It has been suggested that the length distribution of compositional domains follows a power-law distribution (22, 23). With the accumulation of complete genomic sequences and the development of unbiased segmentation methods (24, 25), it has became possible to test this hypothesis without *a priori* assumptions. We thus compared the observed domain-length data to data generated from a power-law distribution plotted on a log-log scale (Fig. S3). If domain lengths are truly drawn from a power-law distribution, then the data should fit the power-law model for three or more orders of magnitude. In all the distributions, the cumulative distribution function deviates significantly from a straight line and the *p* value is sufficiently very small (Kolmogorov-Smirnov, $P < 0.01$) so that the power-law model can be ruled out. Previous results indicating a power-law behavior were based on segmentation algorithms that tended to artificially inflate the number of long compositional domains, whereas the present segmentation algorithm has been shown to be unbiased (24).

In insects, the GC contents of compositional domains exhibit non-normal distributions with a mean of 32.7–44.6% and GC-content standard deviations (σ_{GC}) of 7.7–11.1%. The mean GC content of the *P. barbatus* genome is 36.5%, well within the range for hymenopteran insects (32.7–41.7%), with intermediate dispersal (σ_{GC} = 9.8%) compared to values for hymenopterans (8.8% < σ_{GC} < 11.1%).

The range of GC content in hymenopteran compositional domains is the widest among all insects ranging from 3% to 75%, with *A. mellifera* domains setting both upper and lower limits (Fig. S4). Surprisingly, the range of GC content in compositional domains of *P. barbatus* (9–72%) is similar to that of *A. gambiae* (7–71%). Moreover, both the *P. barbatus* and the mosquito genomes contain a large number of short (< 10 kb) GC-rich domains that increase their mean GC content compared to the honey bee and *D. melanogaster*. Interestingly, the *P. barbatus* genome also contains many GC-poor domains. By comparing the lowest tenth percentile of insect GC content distributions, we found that the compositional domains with the lowest GC contents are found in *P. barbatus* and *A. mellifera*.

Comparing the GC content of compositional domains with their lengths provides a general view of insect genomic architecture. We designate long (> 100 kb) compositional domains with GC content above or below the 5% mean genomic GC content as highly GC-rich and highly GC-poor domains, respectively. Overall, long highly GC-poor domains are very rare among insects. By contrast, long highly GC-rich domains are found mostly among hymenopterans, particularly in *A. mellifera* and *N. vitripennis*. Although all genomes in the analysis have a similar number of long domains (72–231), their GC composition varies greatly (Fig. S4). Nearly all long domains in *T. castaneum*, *A. gambiae*, and *D. melanogaster* have GC contents within ± 5% of the genomic mean GC content, whereas in the honey bee and *N. vitripennis*, approximately half of the domains are highly GC-rich. In the *P. barbatus* genome, 75% of the long domains have GC contents within ± 5% of the genomic GC content with only 8% highly GC-rich domains.

We determined the distribution of genes within compositional domains. We previously observed that genes in *A. mellifera* and *N. vitripennis* have a bias toward occurring in the more GC-poor regions of the genome (26, 27). In contrast, the genomes of all other species we have studied (*Saccharomyces cerevisiae, Homo sapiens*, *D. melanogaster*, *A. gambiae*, *Pediculus humanus*, *Strongylocentrotus purpuratus*, *T. castaneum*) showed either no bias at all or a very slight bias toward occurring in more GC-rich regions of the genome (26, 28–30). Similar to the other hymenoptera genomes, genes in *P. barbatus* tend to occur in the more GC-poor regions of the genome, as depicted in the cumulative distribution of GC content in compositional domains containing genes compared to that of all compositional domains (Fig. S5). Therefore, the tendency for genes to occur in the more GC-poor regions of the genome is a characteristic that is shared among and unique to all hymenopteran genomes sequenced to date.

The genomic distributions of GC content (percent GC) was plotted as the number of total number of nucleotides versus percent GC, after concatenating sequences of compositional domains with equivalent GC contents (Fig. S6). The *P. barbatus* genome does not have a pronounced bimodal distribution in GC content,

while *N. vitripennis* has a strong bimodal distribution, and honey bee has a less pronounced bimodal distribution than *N. vitripennis*. A similar analysis performed on exons and introns shows that introns of *P. barbatus* are more AT-rich than the *P. barbatus* genome, while GC contents for introns in *A. mellifera* and *N. vitripennis* are distributed more similarly to their genome distributions (Fig. S6).

Methylation of CpG dinucleotides has been reported from the honey bee (31), *N. vitripennis* (27), and found to be widespread among social hymenoptera, including two ant species (31, 32), but the relationship between the occurrence of CpG methylation and the distribution of CpG dinucleotides within the hymenoptera genomes remains unclear. The mean ratio of observed to expected CpG (CpG[o/e]) of the *P. barbatus* genome is 1.57, intermediate to that of *N. vitripennis* (1.35) and *A. mellifera* (1.66). The genomic distribution of CpG[o/e] was plotted as the number of nucleotides versus CpG[o/e], after concatenating compositional domain sequences with equivalent CpG[o/e] (Fig. S7). There is a clear bimodal distribution of CpG[o/e] in the genome of *N. vitripennis*, but the distributions are only slightly skewed to lower than mean CpG[o/e] in *P. barbatus* and *A. mellifera*. A similar analysis was performed with coding exons and introns (Fig. S7). Introns show a single mode in *A. mellifera*, bimodal distribution in *N. vitripennis*, and a skew to lower than mean CpG[o/e] in *P. barbatus*. The clear bimodal distribution of CpG[o/e] in exons of *A. mellifera* corresponds with the detection of CpG methylation in coding exons of *A. mellifera* (31, 33). *P. barbatus* and *N. vitripennis* do not show bimodal distributions of CpG[o/e] in coding exons, but their distributions are skewed to lower than the mean CpG[o/e]. A comparison of the CpG[o/e] of the coding exons of genes of known methylation status in *A. mellifera* and their putative orthologs in *P. barbatus* (Table S4) shows that the majority of the documented methylated genes in *A. mellifera* do not have similar CpG[o/e] in *P. barbatus*, suggesting that the methylation status is not the same. In total these results suggest a biological change in the function and operation of the methylation system in *P. barbatus* compared to other hymenopterans. Despite this, an assay of whole body and genome wide methylation, using a methylation-sensitive amplified fragment length polymorphism assay (ms-AFLP), detected CpG methylation in a large fraction of the loci scored. On average 33% (± 12%, standard deviation) of all loci scored were methylated in a pool of 209 females comprised of larvae, pupae and adults of both queen and worker castes. The mean and variance of genome-wide methylation are high compared to another study on social insects using the same methodology (32) and suggest that methylation may be a means of regulation though future studies are needed to determine whether castes differ in methylation (especially during critical periods of development) and whether methylation may be a means of regulating expression of housekeeping (as in *A. mellifera*) or tissue-specific genes (as in vertebrates) (33).

Methods

Recursive segmentation procedures that partition genomic sequences into compositional domains were shown to be the most accurate segmentation methods (24, 25). Here, we partitioned the genomic sequences

into compositional domains using IsoPlotter, a segmentation algorithm that employs a dynamic halting criterion (24). IsoPlotter recursively segments the chromosomes by maximizing the difference in GC content between adjacent subsequences. The process of segmentation was terminated when the difference in GC content between two neighboring segments was no longer statistically significant.

We carried four analyses to study genome architecture in insects. In the first analysis, we calculated the distribution of compositional-domain lengths. For convenience, compositional domains were divided by the order of magnitude of their lengths into short (10^3–10^4 bp), medium (10^4–10^5 bp), and long (10^5–10^7 bp). We next tested whether the lengths of compositional domains follow a power-law distribution. The minimum domain length and the power-law exponent were estimated using the method of Clauset et al. (34). To test the power-law hypothesis, the observed data were compared to data generated from a power-law distribution and the similarity between the two distributions was calculated using the Kolmogorov-Smirnov statistic (35). Based on the observed goodness-of-fit, we calculated a *p*-value that quantifies the probability that the data were drawn from the hypothesized distribution. We used the Matlab scripts provided by Clauset et al. (34) at http://www.santafe.edu/~aaronc/powerlaws/. In the third analysis, we compared the distributions of GC contents of compositional domains. Finally, we compared the compositional-domain GC contents versus their lengths in a log scale.

We computed the genomic distribution of the ratio of observed to expected CpG dinucleotides (CpG[o/e]) by computing CpG[o/e] for each compositional domain and then determining the total number of nucleotides for compositional domains with equivalent CpG[o/e]. CpG[o/e] is defined as CpG[o/e] = PCpG/(PC*PG), where PCpG, Pc and PG are the frequencies of CpG dinucleotides, C nucleotides, and G nucleotides, respectively. We also computed the distribution of CpG[o/e] for coding exons and introns, after concatenating coding exons or introns, respectively, for each gene.

In order to evaluate whether genes of known methylation status, via bisulfite sequencing, in *A. mellifera* were also predicted to be methylated in *P. barbatus* they were found in the *P. barbatus* genome and CpG[o/e] scores were compared (Table S4). There was no correlation in predicted CpG[o/e] between orthologs in the two species, suggesting potential differences between *A. mellifera* and *P. barbatus* in the nature of their methylation systems.

To verify that the methylation toolkit present in *P. barbatus* is indeed functional, we evaluated genome-wide methylation using ms-AFLP using the same method as in (32). We used whole body DNA extraction from individuals of all female castes and at all developmental stages (n = 209 individuals); we used the DNeasy extraction kit (Qiagen, Inc.) and the manufacturer's protocol, as modified for insects, for DNA extraction. DNA from each individual was then cut with each pair of enzymes, *EcoRI* and *MspI/HpaII*, and then PCR was done in two rounds, pre-selective and selective; eight selective primer pairs were used giving a total of 76 bands scored per individual.

Table S3. Distribution of compositional-domain lengths

Order	Species	No. of compositional domains				Total number	Assembly size (Mb)[a]
		1–10 kb (%)	10–100 kb (%)	100 kb–1 Mb (%)	1–10 Mb (%)		
Hymenoptera	*P. barbatus*	35,604 (90.3)	3,637 (9.2)	92 (0.5)	0 (0)	39,433	220
	A. mellifera	42,006 (91.1)	3,944 (8.6)	150 (0.3)	0 (0)	46,100	230
	N. vitripennis	51,064 (92.8)	3,870 (7.0)	72 (0.1)	0 (0)	55,006	238
Coleoptera	*T. castaneum*	15,432 (90.0)	1,535 (8.9)	183 (1.1)	3 (0.02)	17,153	131
Diptera	*A. gambiae*	36,941 (91.5)	3,185 (7.9)	231 (0.6)	0 (0)	40,357	223
	D. melanogaster	12,297 (85.3)	1,973 (13.7)	154 (1.1)	0 (0)	14,424	120

[a] Number of non-ambiguous nucleotides in the assembly.

Table S4. CpG[o/e] values for methylated and unmethylated *Apis mellifera* genes and their putative orthologs in *Pogonomyrmex barbatus*

	A. mellifera[a]		*P. barbatus*[b]			
Status in *A. mellifera*	Gene ID	CpG[o/e]	Gene ID	CpG[o/e]	blastp e-value	Source
methylated	GB19036	0.35	PB22047	1.15	–103	(33)
methylated	GB16176	0.47	PB19018	1.03	0	(33)
methylated	GB16767	0.56	PB21292	0.78	0	(31, 36)
methylated	GB19180	0.57	PB21123	0.84	0	(33)
methylated	GB13959	0.58	PB26204	0.96	–149	(33)
methylated	GB12499	0.65	PB21132	1.06	–157	(33)
methylated	GB19399	0.66	PB16047	1.24	0	(31, 36)
methylated	GB18099	0.67	PB19856	1.07	0	(31, 36)
methylated	GB10208	0.68	PB13690	1.01	0	(33)
methylated	XP_001121083	0.67[c]	PB16971	0.73	–178	(33, 36)
methylated	GB12504	0.59[c]	PB15579	0.88	–127	(33, 36)
—	mean	0.58	mean	0.98	—	—
unmethylated	GB19418	1.14	PB24884	1.15	–158	(33)
unmethylated	GB18363	1.17	PB24302	1.08	–11	(33)
unmethylated	GB15796	1.22	PB18937	1.22	–94	(33)
unmethylated	GB13882	1.33	PB13317	0.86	–147	(33)
unmethylated	GB15055	1.39	PB24766	1.21	0	(33)
unmethylated	DB777978	1.96	not found	not found	not found[d]	(33)
—	mean	1.37	—	1.10	—	—

[a] Validated by bisulfite sequencing.

[b] Based on best blastp e-values against OGS1.1.

[c] Different values are reported in either paper, the value here is the average of the two.

[d] Evidence based on *A. mellifera* ESTs. Not found using tblastx or blastn against OGS1.1 peptides, genome scaffolds, or EST isotigs.

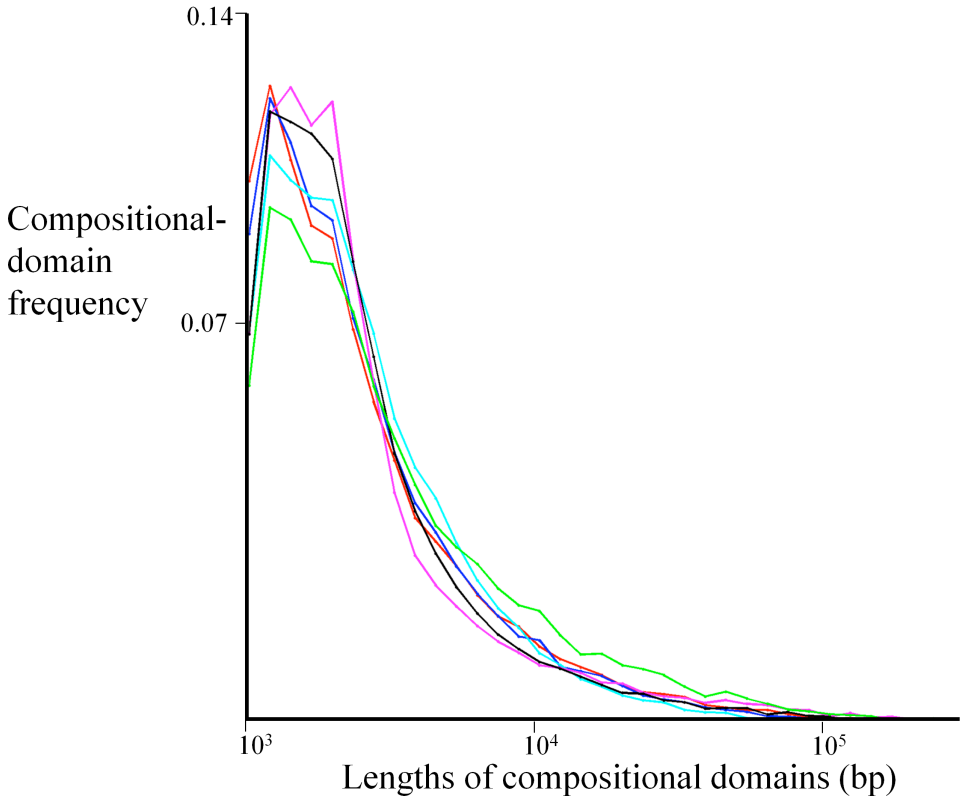


Fig. S2. The frequency of compositional-domain lengths in *Pogonomyrmex barbatus* (red), *Apis mellifera* (blue), *Nasonia vitripennis* (turquoise), *Tribolium castaneum* (purple), *Anopheles gambiae* (black), *Drosophila melanogaster* (green).

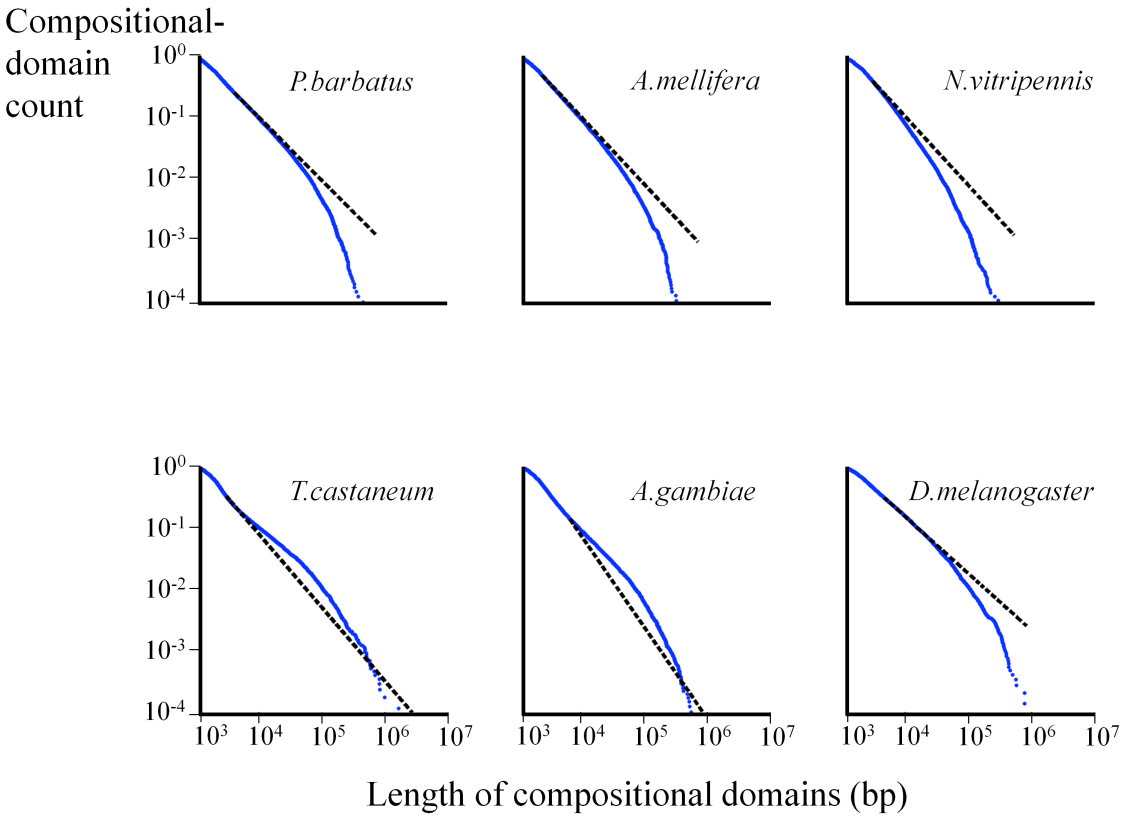


Fig. S3. The cumulative density function of compositional-domain lengths (blue) distributed according to a power-law. The dashed black lines represent best fits to the data.

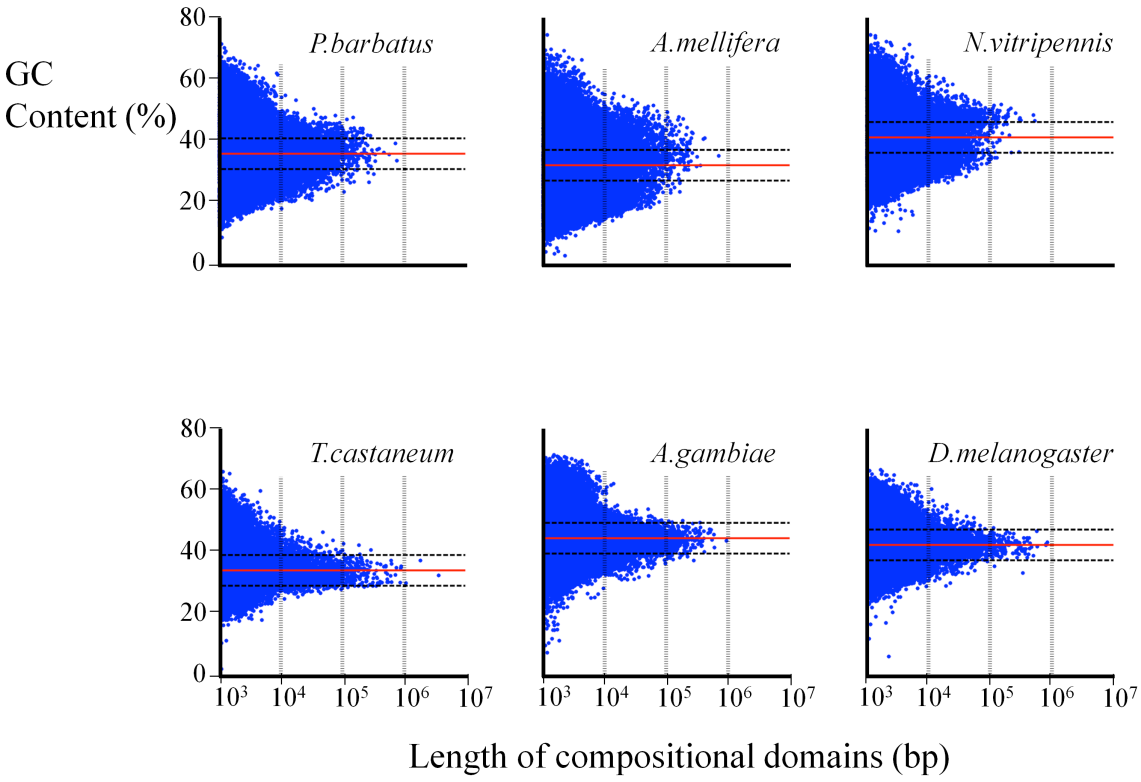


Fig. S4. Domain GC content versus domain lengths on a log scale. The middle horizontal line (solid red) represents the mean genome GC content within margins of ± 5% (dashed black).

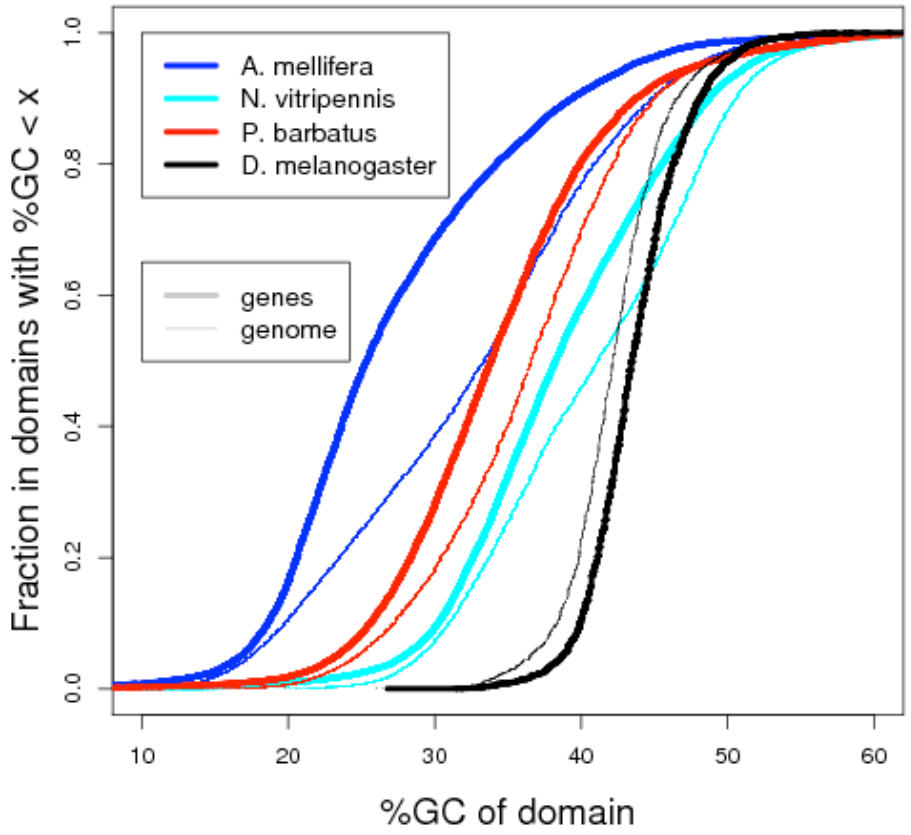


Fig. S5. Comparison of GC content of compositional domains in the insects *Pogonomyrmex barbatus*, *Apis mellifera*, *Nasonia vitripennis* and *Drosophila melanogaster*. Cumulative distributions show the fraction of genes (thick lines) or the entire genome (thin lines) occurring in GC compositional domains (< x %GC). Similar to the other hymenopterans, *P. barbatus* genes tend to occur in the AT-rich parts of the genome.

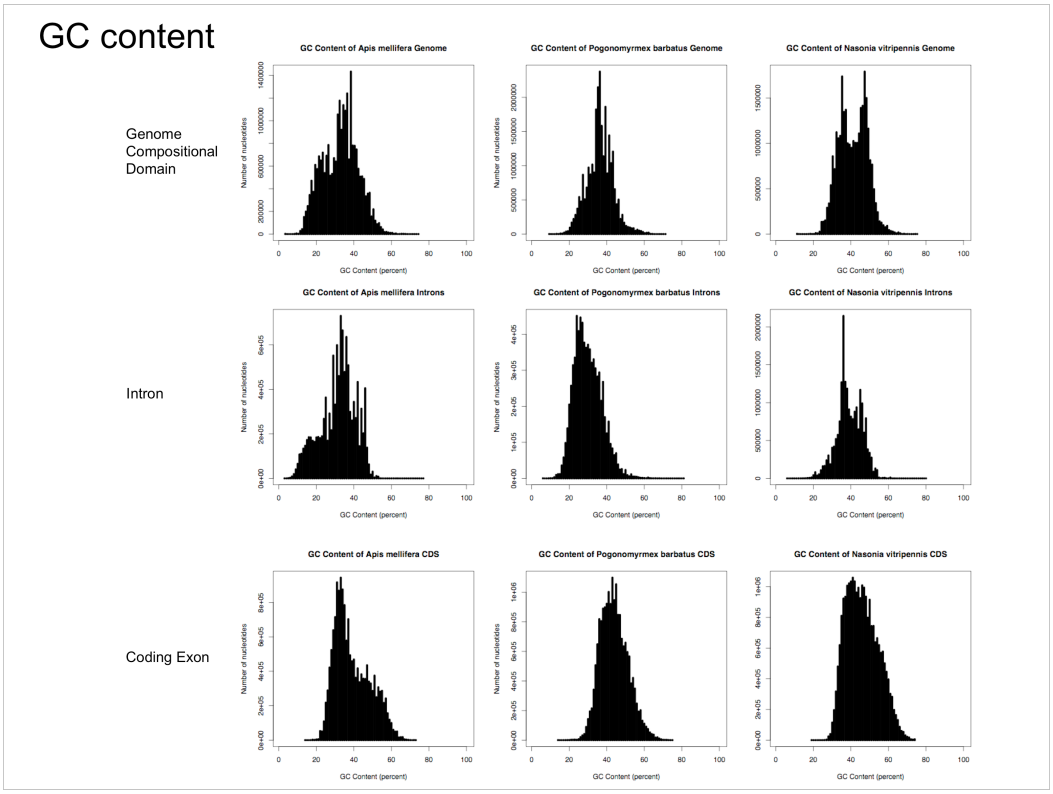


Fig. S6. Distribution of GC content in compositional domains, introns and coding exons of *Apis mellifera*, *Pogonomyrmex barbatus* and *Nasonia vitripennis*.

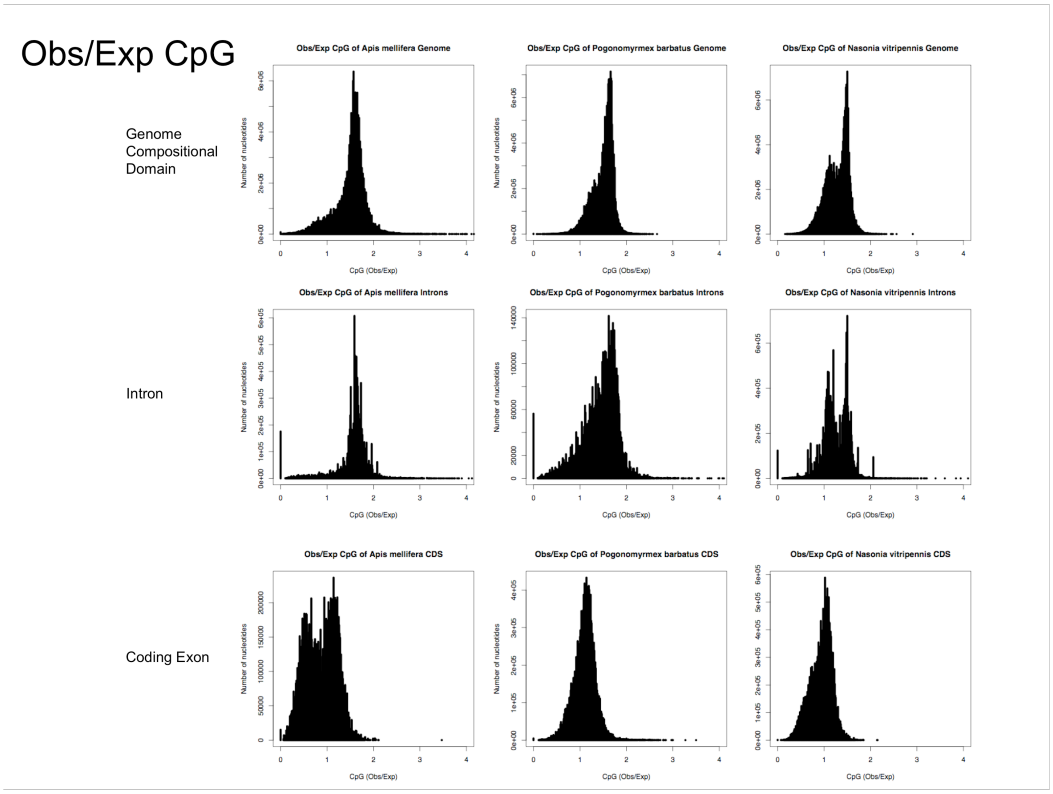


Fig. S7. Distribution of CpG[o/e] in compositional domains, introns and coding exons of *Apis mellifera*, *Pogonomyrmex barbatus* and *Nasonia vitripennis*.

5. Gene Ontology and Orthology Analysis

Gene Ontology

We were interested to see if there were enrichments for genes involved in specific molecular functions, biological processes, or cellular locations in *Pogonomyrmex barbatus* relative to the mostly manually annotated *Drosophila melanogaster* genome or the relatively automated annotations for *Apis mellifera* and *Nasonia vitripennis*. Of 17,250 genes in the OGS1.1, 7,514 genes (44%) were annotated with at least one Gene Ontology term. In total 23,277 GO terms were reported and each GO-annotated gene had, on average, three terms. We looked for specific gene classes enriched in *P. barbatus* compared to *D. melanogaster*, *A. mellifera*, and *N. vitripennis* (Fig. S8–10, Table S5). We found significant enrichment of seven terms related to cellular locations, with most being terms implicating synapse or membrane localization, which is consistent with the expected location of Or genes (37) which have recently expanded in *P. barbatus*. Amongst the 14 enriched genes ($P < 0.05$) associated with molecular functions, six are associated with odorant binding and olfaction and six include cation binding to calcium, zinc, or other ions. Thirteen biological processes were enriched including ones for sensory perception of smell, cognition, and neurological processes, all consistent with gene families that would be required to sense, process, and transduce signals from semiochemicals or other environmental chemical cues (Fig. S9). Future studies will determine whether enrichment of these terms may be associated with the lineage specific expansion of Or genes.

Gene Ontology Methods

We analyzed the complete set of *P. barbatus* MAKER predictions (16,404) and *ab initio* gene predictions (38,260) with InterProScan (16) to identify gene regions with similarity to known function domains. Raw InterProScan results were parsed using custom Perl scripts to generate a Gene Ontology GAF2.0 file (Supp File) that was used as an input to identify enriched Gene Ontology terms using GO-TermFinder v0.86 (38), http://search.cpan.org/dist/GO-TermFinder/). For comparative purposes, we performed a similar GO enrichment analysis on the *D. melanogaster* GO annotations (geneontology.org, (39). Since we were unaware of an existing GO annotation for the honeybee or *N.vitripennis* genomes, we generated one using InterProScan and the preOGS2 and OGS1.2 peptides, respectively (beebase.org). Enrichments were tested for statistical significance using a Fisher exact test implemented in Go-TermFinder.

Identification of Orthology Groups

We identified 9,248 ortholog groups shared between at least two of the *P. barbatus*, *A. mellifera, N. vitripennis*, and *D. melanogaster* datasets, with 5,637 (33%) orthologs common to all four species. There

were 1,334 (7.8%) genes shared between all the hymenopterans and 564 (3.2%) genes common only to the eusocial insects. These latter genes will become increasingly refined in comparison with other ant genomes and may shed significant light on the genetic factors associated with eusociality. *P. barbatus* had 8346 (49%) genes that were not found in any other species and we identified 177 enriched terms for this subset using GO-TermFinder (38). Consistent with our data on Or expansions in *P. barbatus*, several terms for processes or functions (Fig. S9–10) involved in sensory perception of smell, olfaction, G-proteins coupled receptors, odorant binding, and response to chemical stimulus ($P < 0.0001$). As expected, these proteins are enriched in cellular locations (Fig. S8), such as membranes ($P < 4.8\text{x}10^{-7}$), and synaptic regions ($P < 0.0003$). Interestingly, the process of methylation was also weakly enriched ($P < 0.04$), supporting the notion that the complete DNMT system may play a role in gene regulation in *P. barbatus*. Numerous other membrane transporters, electron transport, and peptidase terms were enriched and may represent lineage specific cytochrome P450s and other genes required to meet nutrient loads and synthesize venoms required for adaptation to the unique niche of harvester ants (Fig. S8–10). Future comparisons to other ant and bee genomes will shed considerable light on genes enriched across hymenoptera and help to identify genes specific to the unique biology of *P. barbatus*.

Orthology Methods

OrthoMCL (40) is an algorithm which identifies ortholog groups between two or more species and lineage-specific gene expansion families (inparalogs) based on blastp protein sequence similarity. We used OrthoMCL v2.0 to identify orthologous protein sequences between three Hymenoptera species, *A. mellifera* (preOGS2, (26)), *N. vitripennis* (OGS1.2, (27)), and *P. barbatus* (OGS1.1, this study) as well as *D. melanogaster* (Release 5.27, (41)). To avoid complicating orthology-paralogy results, we first reduced each protein dataset using custom Perl scripts to contain only the single longest isoform when multiple isoforms were present. Next, the results from an all-by-all BLAST were parsed with code provided by OrthoMCL (40) to determine best reciprocal hit orthologs, inparalogs, and co-orthologs. The MCL v09-308 (Markov Clustering algorithm (42), was used to define final ortholog, inparalog, and co-ortholog groupings. We followed OrthoMCL's suggested parameter values and options for all steps in the pipeline.

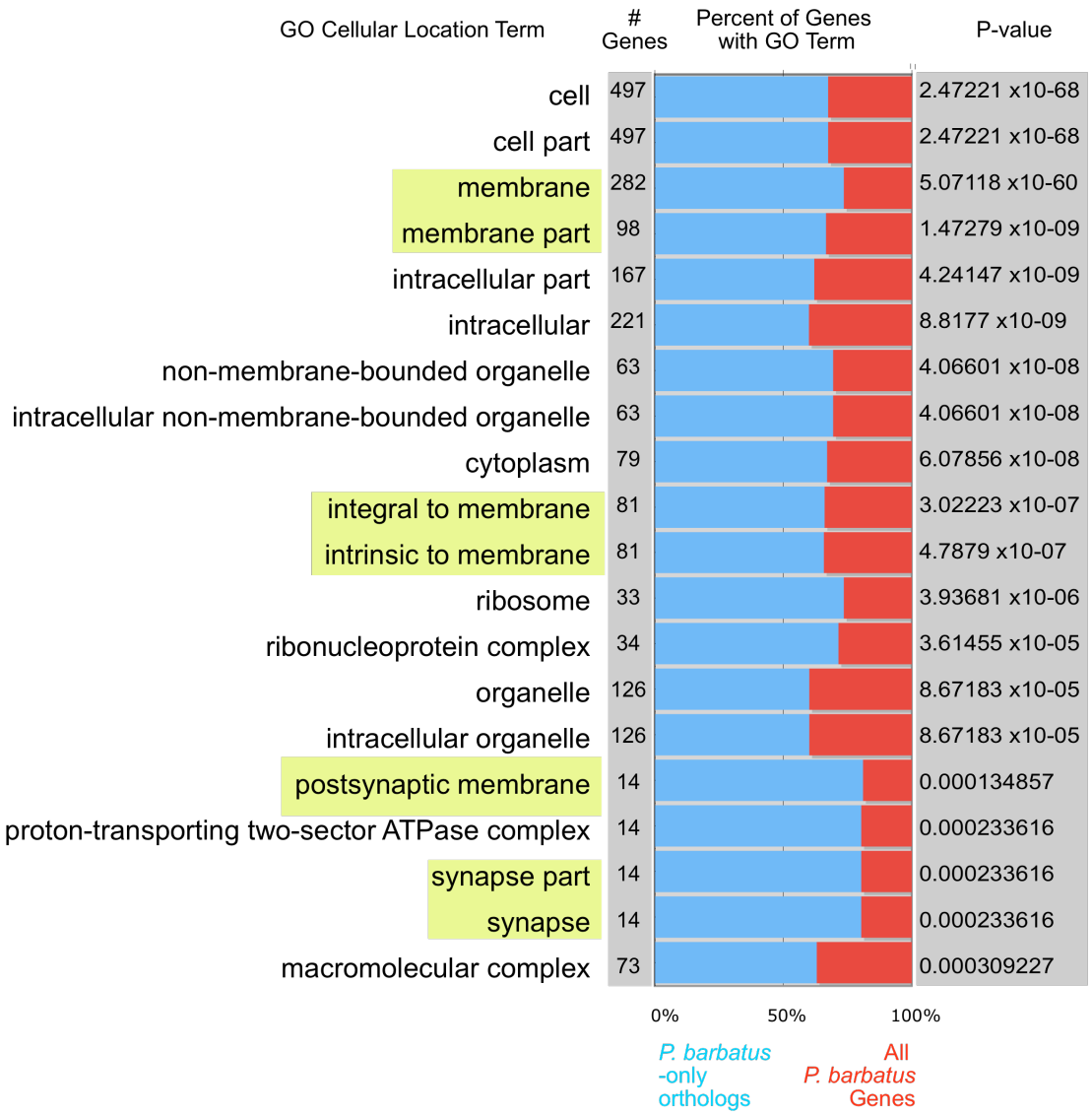


Fig. S8. A list of cellular localization gene ontology terms significantly enriched in *Pogonomyrmex barbatus*. Highlighted terms are consistent with localizations of gene families observed to have expansions in *P. barbatus*. *P*-values are based on Bonferroni-corrected Fisher exact test scores.

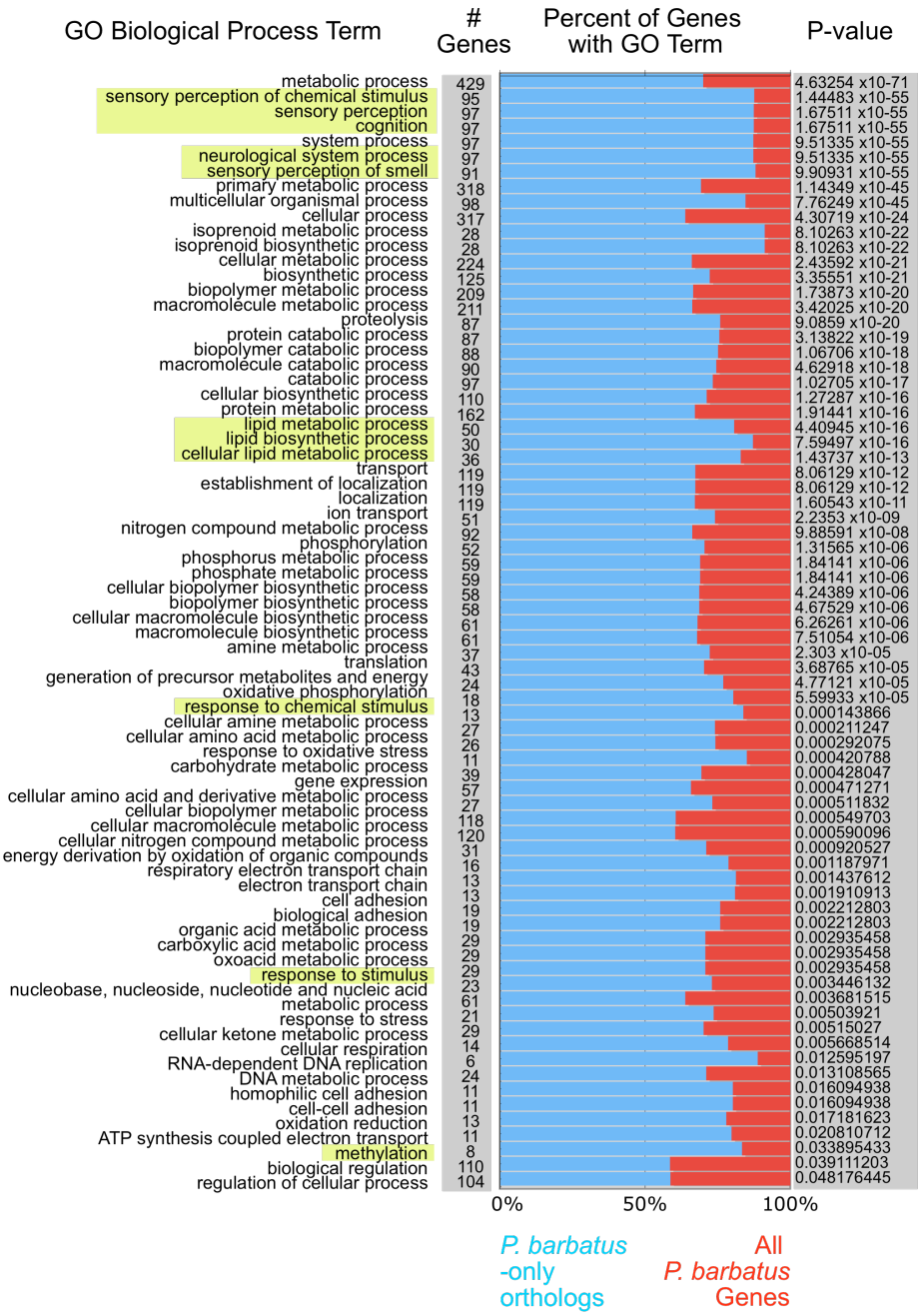


Fig. S9. A list of biological process gene ontology terms significantly enriched in *Pogonomyrmex barbatus*. Highlighted terms are consistent with processes of gene families observed to have expansions in *P. barbatus*. *P*-values are based on Bonferroni-corrected Fisher exact test scores.

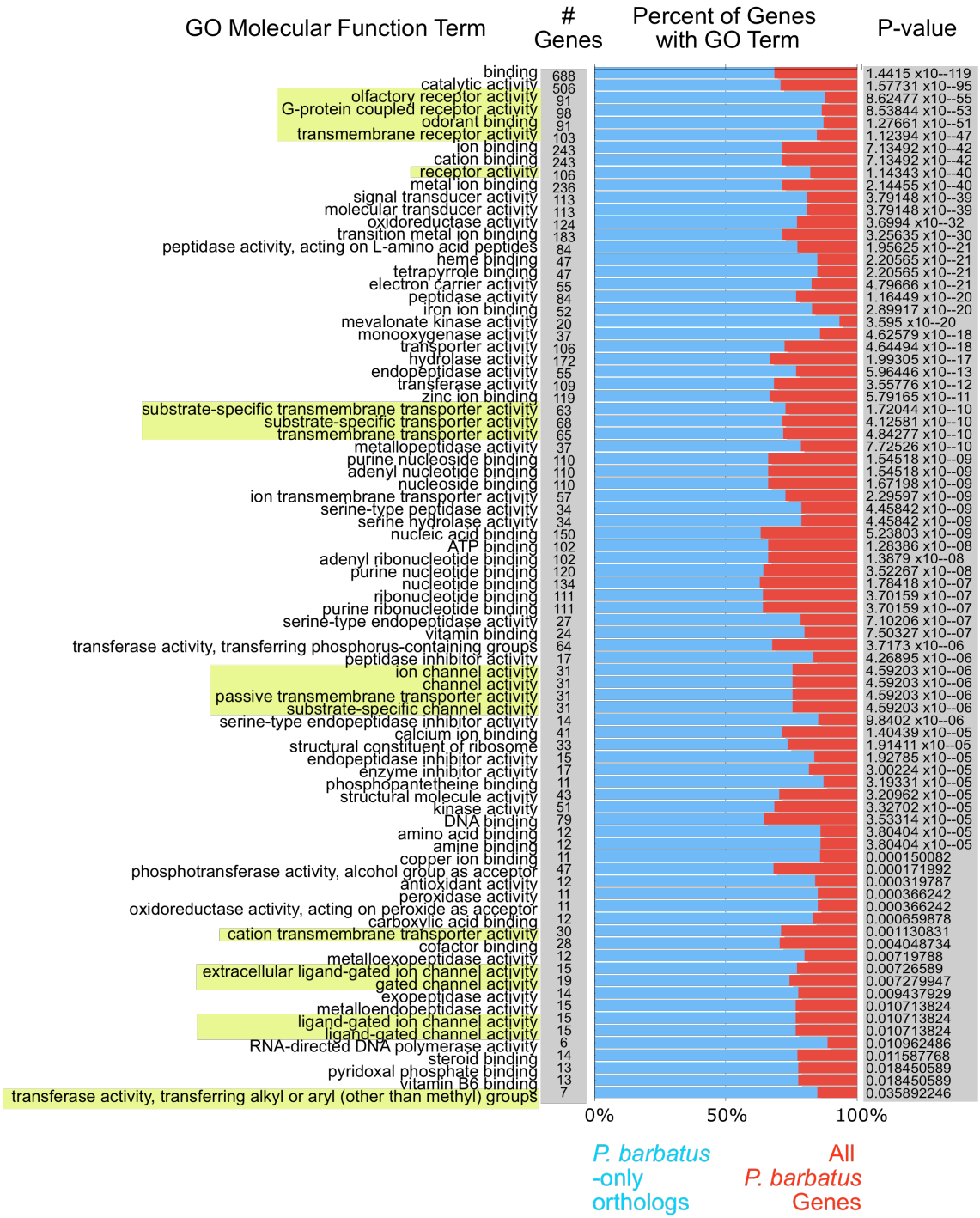


Fig. S10. A list of molecular process gene ontology terms significantly enriched in *Pogonomyrmex barbatus*. Highlighted terms are consistent with gene functions of gene families observed to have expansions in *P. barbatus*. *P*-values are based on Bonferroni-corrected Fisher exact test scores.

Table S5. Gene Ontology terms enriched in *Pogonomyrmex barbatus* relative to *Drosophila melanogaster*, *Apis mellifera*, and *Nasonia vitripennis*

GO ID	GO term	*P*-value	Ontology aspect	No. of genes
GO:0003008	system process	7.23E–11	P	205
GO:0050877	neurological system process	7.23E–11	P	205
GO:0007600	sensory perception	9.44E–11	P	202
GO:0050890	cognition	9.44E–11	P	202
GO:0032501	multicellular organismal process	2.87E–10	P	257
GO:0007606	sensory perception of chemical stimulus	3.66E–10	P	193
GO:0007608	sensory perception of smell	1.35E–08	P	179
GO:0006720	isoprenoid metabolic process	6.45E–05	P	39
GO:0008299	isoprenoid biosynthetic process	6.45E–05	P	39
GO:0007165	signal transduction	1.06E–03	P	487
GO:0007154	cell communication	1.08E–03	P	520
GO:0007264	small GTPase mediated signal transduction	1.23E–03	P	90
GO:0022904	respiratory electron transport chain	1.38E–02	P	43
GO:0042773	ATP synthesis coupled electron transport	2.42E–02	P	40
GO:0007242	intracellular signaling cascade	2.88E–02	P	192
GO:0005623	cell	3.98E–18	C	3432
GO:0044464	cell part	3.98E–18	C	3432
GO:0044456	synapse part	2.37E–06	C	49
GO:0045202	synapse	2.37E–06	C	49
GO:0045211	postsynaptic membrane	5.63E–06	C	47
GO:0005622	intracellular	1.96E–05	C	2114
GO:0016020	membrane	6.69E–05	C	1452
GO:0005509	calcium ion binding	2.57E–14	F	236
GO:0043167	ion binding	4.92E–14	F	1384
GO:0043169	cation binding	4.92E–14	F	1384
GO:0046872	metal ion binding	2.54E–13	F	1346
GO:0004930	G-protein coupled receptor activity	5.68E–09	F	218
GO:0004984	olfactory receptor activity	1.39E–08	F	179
GO:0005549	odorant binding	9.11E–08	F	191

GO:0004496	mevalonate kinase activity	1.03E–07	F	21
GO:0008270	zinc ion binding	5.99E–07	F	857
GO:0004674	protein serine/threonine kinase activity	2.44E–05	F	186
GO:0004888	transmembrane receptor activity	3.31E–05	F	268
GO:0046914	transition metal ion binding	4.74E–04	F	1046
GO:0031177	phosphopantetheine binding	4.20E–03	F	23
GO:0004872	receptor activity	5.18E–03	F	329

P-values are the result of a Bonferroni-corrected Fisher exact test implemented in GO-TermFinder v0.86. Biological process (P), cellular component (C), and molecular function (F) ontology terms are indicated.

6. Cytoplasmic Ribosomal Protein Genes

Ribosomal proteins are integral components of ribosomes, the macromolecular machines that govern protein synthesis in all living cells. While the catalytic core of a ribosome is composed of ribosomal RNA (rRNA), ribosomal proteins reside at its surface where they perform many auxiliary functions including assembling and stabilizing the structure of the particle, protecting the rRNA from degradation, and tethering the mRNA to the ribosome during translation. However, their exposed localization also allows ribosomal proteins to mediate the many interactions of the ribosome by serving as binding platforms for other protein factors involved in the process of translation (43). For example, RACK1 – now recognized as an integral ribosomal protein (44) – is linked to cellular signaling pathways and coordinates the regulated translation of specific mRNAs, and may even be involved in the recruitment of ribosomes to sites of localized translation (45). Moreover, some ribosomal proteins are only loosely attached to ribosomes, and serve various extra-ribosomal functions (46).

Ribosomal proteins are of ancient evolutionary origin, which may predate the split between the kingdoms of life. Among eukaryotes, the about 80 cytoplasmic ribosomal proteins (CRPs) are highly conserved both in number and sequence (eukaryotic cells possess two types of ribosomes, cytoplasmic and mitochondrial ones, whose protein components are encoded by two different sets of nuclear genes). Due to the ribosomes' vital role, ribosomal protein genes are also ubiquitously and abundantly expressed. Easily obtained from even small cDNA libraries, they thus represent a valuable resource for studying deep phylogenetic relationships (e.g., 47, 48). The fact that ribosomal protein genes are numerous, well-defined and widely distributed across both the human and the *Drosophila melanogaster* genome (49, 50) also makes them suitable for assessing the coverage and quality of *de novo* genome assemblies.

In *Pogonomyrmex barbatus,* we identified 86 genes encoding 79 cytoplasmic ribosomal proteins (not including RACK1), which is the full set of proteins known from mammalian and insect genomes (49, 50). According to EST evidence, all of these genes are transcriptionally active. In addition, RACK1, two CRP-like genes of unknown function which are present in all eukaryotes and characterized by low sequence similarity to their corresponding CRP genes, as well as several non-processed pseudogenes were found. Consequently, seven CRPs are represented by two distinct genes (RpL11, RpLP0, RpS2, RpS7, RpS10, RpS19 and RpS28). Since the same proteins are encoded by single-copy genes in *Apis mellifera* (as well as in *Nasonia vitripennis* and *D. melanogaster*), the duplicates found in *P. barbatus* must have arisen after the diversification of the aculeate lineages. Indeed, the gene structure and sequence of the duplicates are almost identical, with an average pairwise protein similarity score of 94%, indicating evolutionarily recent duplication events. Although all duplicates are represented by ESTs, it is possible that only one copy is

primarily expressed in the majority of tissues. This is suggested by corresponding findings in *D. melanogaster* (50), which possesses its own set of lineage-specific gene duplicates, and the general belief that each mammalian ribosomal protein is encoded by only a single functional gene despite the existence of thousands of processed pseudogenes (51). The stoichiometrically precise coproduction of all ribosomal components is presumably an essential requirement in all organisms, though the underlying regulatory mechanisms remain to be elucidated. Finally, as in other eukaryotes, RpL40, RpS27A and RpS30 are represented by fusion genes that encode ubiquitin or an ubiquitin-like sequence at the 3' end. Overall, the CRP gene repertoire of *P. barbatus* is highly similar to the one of other insects, both with regard to gene number (88, 80 and 79 in *D. melanogaster*, *A. mellifera* and *N. vitripennis*, respectively) and sequence similarity (79% identity to *D. melanogaster* at the protein level, range 52–100%).

Having found 100% of the generally widely distributed CRPs represented in our assembly indicates that it excellently covers the gene space of the genome. During the annotation process, two cases of putative scaffold misassembly were detected: in one case, a CRP gene model was found to be split across two scaffolds, and in another case a model covered only a fragment of a gene that turned out to be fully represented in the genome raw reads and the EST data. Finally, the high sequence similarity between the CRP gene models and reference genes made it possible to clearly identify sequencing errors that resulted in reading frameshifts or premature stop codons. Six of these instances were uncovered. Unexpectedly, only one of those was clearly associated with a homopolymer run, an error source characteristic for the 454 sequencing technology. Based on the total number of nucleotides coding for CRPs (about 43,000), a minimum estimate for the proportion of positions being affected by sequencing error of 0.014%, or 1 in 7200, can be concluded. While artificial substitutions, which leave the reading frame intact, cannot be detected by this approach, substitution errors have been estimated to contribute less than 20% to the total number of errors made by 454 sequencing on the level of individual reads (52). Even though the error rate is presumably two to three times higher in non-coding regions, this represents a sequence accuracy of the assembly superior to what has been achieved with earlier generations of the 454 technology and comparable to Sanger sequencing (53, compare to chapter 14 of this study). Thus, the coverage and sequence fidelity of the CRP genes can testify to the high overall quality of the *P. barbatus* genome assembly.

7. Oxidative Phosphorylation Genes

The oxidative phosphorylation (OXPHOS) pathway is the major source of cellular energy, in the form of ATP, in most eukaryotes. Four of the OXPHOS protein complexes produce a proton gradient across the inner mitochondrial membrane by harnessing the energy released from electrons traveling through the redox reactions of the electron transport chain. The energy stored in this gradient is then used by ATP synthase to produce ATP. This pathway is unique because it incorporates many nuclear encoded proteins as well as all 13 mitochondrial encoded proteins (54). Mitochondrial genes tend to evolve at a faster rate than their nuclear counterparts, which may result in the interacting nuclear genes evolving quickly to "keep up," a process known as compensatory coadaptation (55). The resultant fast evolution of the nuclear genes may lead to hybrid breakdown in many F2 hybrids because the interacting proteins have potentially evolved in different populations or species (that is, the parental taxa have been separated over long evolutionary distances). The role of the nuclear OXPHOS pathway genes in hybrid breakdown may serve to keep closely related taxa reproductively isolated, ultimately leading to speciation.

We found evidence for 76 nuclear encoded OXPHOS genes in the genome sequence of *Pogonomyrmex barbatus*. There are 81 nuclear OXPHOS genes reported in *Drosophila melanogaster*, of which 14 appear to be duplications (56). We found evidence for two of these duplicated genes in *P. barbatus*, as well as six *P. barbatus* specific duplications that were not found in either *Nasonia vitripennis* or *Apis mellifera*. Additionally, there are two nuclear OXPHOS genes that appear to be absent from the *P. barbatus* genome, but are found in either *N. vitripennis* or *A. mellifera*.

8. Hox Genes

The Hox genes encode transcription factors with a pivotal role in cell-fate determination and embryonic development of the animal body plan. These genes have been identified in all bilateral animal phyla examined and are known to determine the positional specification of the anterior-posterior axis (57). Mutations in Hox genes lead to transformations in body-segments and organ identities along the anterior-posterior axis of the body; these transformations are also known as homeotic mutations. The homeotic capability of the Hox genes is conserved among arthropods and vertebrates, which diverged more than 600 million years ago (57).

The Early Cambrian Ancestor that gave rise to present-day arthropod groups probably had a complex containing ten Hox genes (58). These ten genes are expressed in Hox-like patterns in chelicerates and myriapods. In the insects, however, the closest Hox 3 homologs (*zerknullt –zen, zerknullt2 – zen2, bicoid – bcd*, and *fushi tarazu – ftz*), have novel developmental roles that do not include a Hox-like role in determining segmental identity.

Multiple Hox clusters have been described for several vertebrates including mice, humans and fish. In contrast, single clusters have been identified in a number of invertebrates including amphioxus, sea urchins, and several insects like mosquitoes, beetles and locusts (57).

In *Drosophila melanogaster*, the complement of Hox genes is divided into two clusters, the Antennapedia Complex (ANT-C) and the Bithorax Complex (BX-C), separated by approximately 7.5 Mb on the right (R) arm of chromosome 3. This split is thought to be fairly recent in origin. *D. melanogaster* has eight genes with traditional Hox-like developmental function. The ANT-C contains genes required for proper development of the gnathal and thoracic segments – *labial* (*lab*), *proboscipedia* (*pb*), *Deformed* (*Dfd*), *Sex combs reduced* (*Scr*), and *Antennapedia* (*Antp*) –, while the BX-C genes *Ultrabithorax* (*Ubx*), *Abdominal-A* (*Abd-A*) and *Abdominal-B* (*Abd-B*) are responsible for the development of the abdomen and telson portions of the insect body plan.

Additionally, the *D. melanogaster* ANT-C contains the genes *zen*, *zen2*, *bcd* and *ftz*, all homologs of *Hox-3,* without a Hox-like role. There are also eight cuticle genes, five lysine tRNA genes, and *amalgam* (*Ama*, member of the immunoglobulin superfamily). The complex is contiguous in *Anopheles gambiae*, as well as in *Tribolium castaneum*, *Apis mellifera* and *Nasonia vitripennis*.

Exhaustive computational analyses indicate that the *Pogonomyrmex barbatus* genome does not appear to contain more than one gene associated with each of the ten Hox groups of orthology known amongst arthropods, suggesting a single, compact Hox cluster where all transcription occurs in the same direction (Fig. S11). Three microRNAs are also found within the cluster. *miR-iab-4* and *miR-10* have conserved positions with respect to *A. mellifera* and *D. melanogaster* (59); *miR*-993 is also found in *A.*

mellifera and *N. vitripennis* (27) but not in other arthropod Hox clusters. Additionally, gene model PB20603, which has significant sequence similarity to predicted proteins from *N. vitripennis* and *A. gambiae* (accession numbers XP_001599398.1 and XP_315505.3, respectively), is located within an intron of the *Ultrabithorax* gene. This gene has not been previously reported in this location for any other insects.

Intergenic distances and gene sizes are comparatively smaller in *P. barbatus* relative to those of *N. vitripennis* and *A. mellifera* (59) (Table S6). As a result, the *P. barbatus* cluster is 0.81 Mb in length, comparable to that of *D. melanogaster* and *T. castaneum* (approximately 0.7 Mb), but about half the size of the clusters in the other two hymenopterans sequenced to date (1.68 Mb for *N. vitripennis* and 1.37 Mb in *A. mellifera*).

A number of sequence gaps are found mostly within intronic sequences. Most amino acid sequences are intact except for *Antp*, in which the coding sequence is extended into a gap, rendering it incomplete (marked with an asterisk in Table S6).

Table S6. Individual gene sizes, intergenic distances and total length of the *Pogonomyrmex barbatus* Hox cluster

Gene	Coordinates	Size (kb)	Distance to next gene (bp)
abd-B	1569321–1576452	7	165634
abd-A	1742086–1765527	23	90214
Ubx	1855741–1931242	75	134095
Antp	2065337–2066141	0.8[a]	102375
ftz	2166366–2167712	1	28797
Scr	2177430–2195163	18	67781
dfd	2238979–2245211	6	40754
Hox3-A	2285965–2288454	2	33544
pb	2321998–2346140	24	20363
lab	2366503–2377779	11	—
Total size	—	—	808458

[a] Coding sequence could be determined only partially.

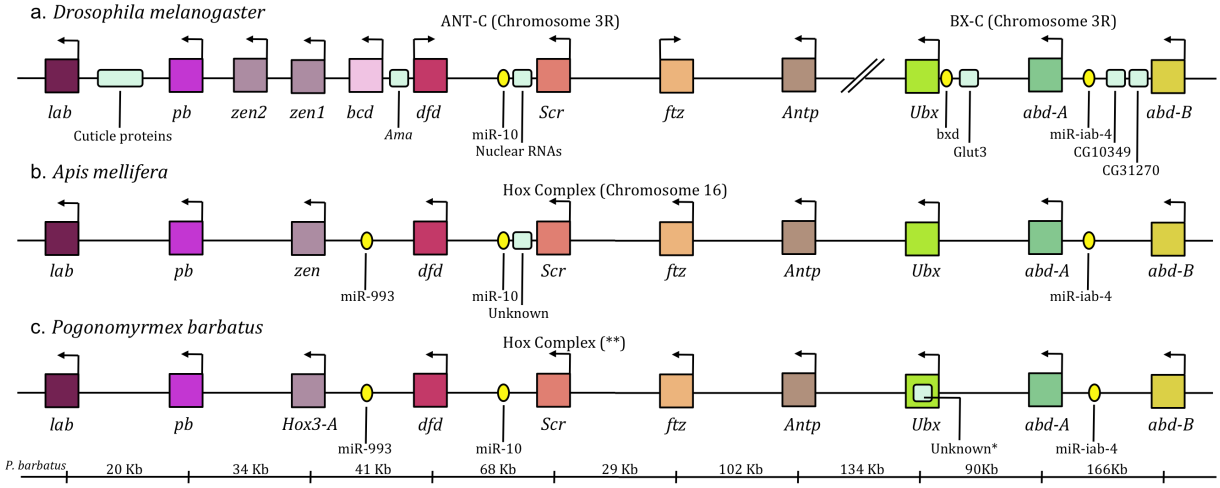


Fig. S11. The *Pogonomyrmex barbatus* (c), *Apis mellifera* (b) and *Drosophila melanogaster* (a) Hox clusters compared (figure modeled after (59)). Arrows indicate the direction of transcription. In *P. barbatus,* the Hox Complex is situated along a genomic region spanning ~ 0.81 Mb (ruler indicating intergenic distances is not drawn to scale). Note that a gene of unknown function (*) is encoded in an intron of the *Ultrabithorax* gene. (**) The *P. barbatus* Hox cluster is located on scaffold number 7180000350303.

9. Mitochondrial Genome Assembly

Animal mitochondrial genomes typically contain 37 genes and are approximately 16 kb in length (60). We could not fully assemble the mitochondrial sequence of *Pogonomyrmex barbatus* in the v03 genome assembly; the longest contiguous sequence of mitochondrial DNA was nearly 6 kb long and included ten genes (one of which is only partially contained). We were able to assemble four scaffolds covering an estimated 71% (11,554 bp) of the mitochondrial genome (calculated using the *Apis mellifera linguistica* mitochondrial genome as a reference, (61)) by using an iterative process of searching the *P. barbatus* v03 scaffolds, transcriptome, raw sequencing reads, and *P. barbatus* mitochondrial sequences deposited at NCBI. Genomic fragments were aligned with the program Sequencher v4.5 (Gene Codes Corp. 2005), using a minimum overlap of 20 bp (Fig. S12). There are 18 genes at least partially covered in this assembly and they include eleven of the 13 protein coding genes, six of 22 tRNAs and one of the two ribosomal RNAs (Fig. S13). To avoid including scaffolds that should be incorporated into the nuclear genome (NuMts) we ensured that the assembled scaffolds covered multiple genes and that all scaffolds had EST support. We estimate that the four scaffolds are separated by three short gaps as well as one large gap that includes the origin of replication (a highly repetitive AT-rich region) (Fig. S13). While we cannot assess large-scale synteny with the *A. mellifera* mitochondrial genome without bridging these gaps, there is complete synteny within each scaffold. The assembled sequences can be found in Dataset S1.

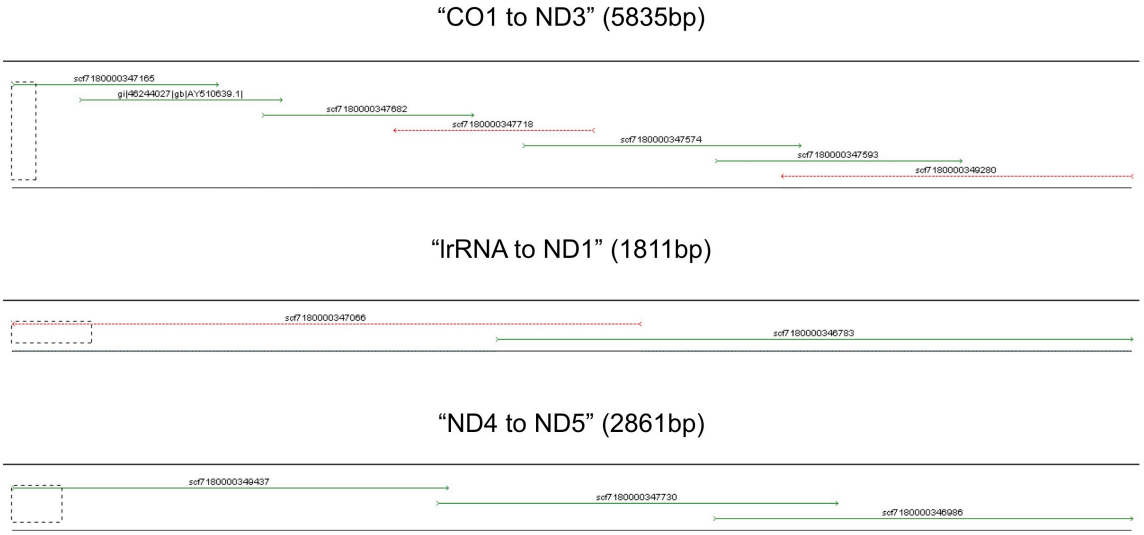


Fig. S12. The assembly of fragments used to make mitochondrial scaffolds (scaffolds in the v03 assembly begin with "scf", and *Pogonomyrmex barbatus* sequences from NCBI begin with "gi"). Forward and reverse complements are indicated by green and red lines, respectively. Scaffold "ND6 to Cyt b" is derived from a single scaffold (scf7180000347661).

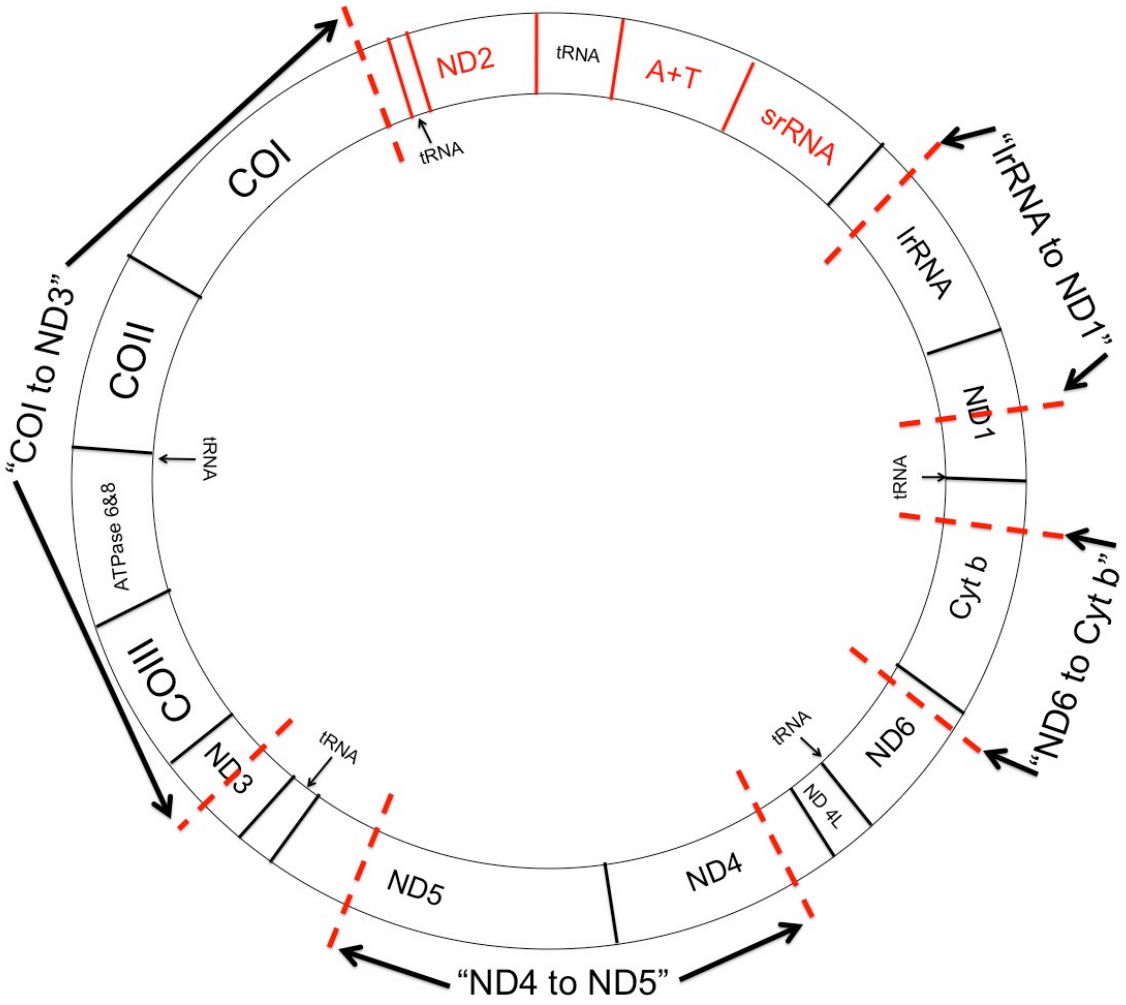


Fig. S13. Representation of the arrangement of the *Pogonomyrmex barbatus* mitochondrial genome based on the mitochondrial genome of *Apis mellifera* (61). The approximate location of the ends of the scaffolds are indicated by dashed red lines. The names of the scaffolds correspond to their sequence names in Dataset S1.

10. Proteomics

The aim of the proteomics analysis was two-fold:

1. To obtain an independent confirmation for gene annotations.
2. To assess the practicability of future proteomics studies on *Pogonomyrmex* addressing nest defense (poison gland) and signal perception (antennae).

Sample preparation

Antennae of *Pogonomyrmex rugosus* specimens were cut off with a scissor, transferred into a centrifuge tube, and immediately frozen in liquid nitrogen. For poison gland collections, individuals were first frozen in liquid nitrogen and then dissected under a drop of double-distilled water. The resulting sample consisted of the poison glands, the venom reservoir and the attached Dufour's glands. After dissection, the sample was transferred into a centrifuge tube, and immediately frozen in liquid nitrogen. Both sample types were stored at –80 °C until further use. Samples from ten individuals (antennae) and 20 individuals (poison gland) were each pooled into two separate samples and the resulting pools were homogenized in 150 µl of protein extraction buffer (50 mM tris pH 8.5, 2% SDS, 5% beta-mercaptoethanol, 0.15 M NaCl, 30% glycerol). Further protein extraction and quantification was performed as previously described (62). A total of 30 µg per sample were subjected to digestion over night at 30 °C with 1 µg of trypsin in digestion buffer (50 mM tris pH 8.5, 0.15 M NaCl, 1 mM $CaCl_2$). Peptide desalting was performed the next day as described before (63, 64).

LC-MS/MS Analysis

Peptides were dissolved in H_2O / 0.1% TFA to a concentration of 1 µg/µl. The equivalent of 5 µg of protein per sample was used for analysis. LC-MS/MS was carried out using a linear quadrupole ion trap ThermoFinnigan LTQ mass spectrometer (San Jose, CA) equipped with a Michrom Paradigm MS4 HPLC, a SpectraSystems AS3000 autosampler, and a nanoelectrospray source. Peptides were eluted from a 15 cm pulled tip capillary column (100 um I.D. x 360 µm O.D; 3–5 um tip opening) packed with 7 cm Vydac C18 (Hesperia, CA) material (300 Å pore size), using a gradient of 0–90% solvent B (98% methanol / 2% water / 0.5% formic acid / 0.01% triflouroacetic acid) over a 90 min period at a flow rate of ~ 350 nl/min. Total run time was 125 min. Further settings were: LTQ electrospray positive mode spray voltage 1.6 kV, capillary temperature 180 °C, isolation window 3 m/z, collision energy 35, and activation time 30 ms. MS^2 spectra were recorded for the ten most abundant peaks in each MS survey spectrum. Using the open source search tool OMSSA (version 2.1.7) (65) the spectra were matched against a database containing: 1. the Official Gene Set v1.1; 2. manually annotated sequences of chemosensory and other proteins (Odorant Binding

Proteins, Gustatory Receptors, Odorant Receptors, Ionotropic Receptors, and P450 Cytochromes; see Dataset S1); 3. trypsin and keratin sequences; 4. the reverse sequences to all aforementioned sequences. The following filtering criteria were used for the analysis: 0.8 Da fragment tolerance, 0.8 Da precursor tolerance, maximum of two missed cleavages, only tryptic sequences allowed, initially eleven possible peptide hits per spectrum reported then filtered to one peptide hit per spectrum, variable modifications: methionine oxidation, deamidation of N and Q. Acceptance threshold: $e \leq 0.1$. A protein was only reported if at least two peptide hits matched to the respective sequence. This meant a false positive rate of 0% on the level of reported proteins (see (62) for further details). Database hits were tentatively identified by employing a BLAST search against NCBI's non-redundant nucleotide database.

General results

The analysis resulted in the detection of 165 proteins, 48 of which were identified in both sample types, 98 (out of 146) were unique to the antennae and 19 (out of 67) were unique to the poison gland and venom (Fig. S14). These numbers, which were generated without extensive pre-fractionation of proteins or peptides, are a good indicator that future in-depth studies of poison gland and antenna samples can reveal important metabolic processes involved in nest defense and chemoperception. In fact, the present study already resulted in the identification of proteins that are involved in these processes. We proceed to discuss some of the proteins that were only found in one of the two tissues.

Antennae

Despite the fact that antennae are needed for perception of the environment, which includes social interactions, only very few studies have reported on the protein complement of insect antennae (66, 67) and none have been conducted on a social insect.

Our antenna-specific results included proteins involved in chemoperception (OBP12), neuronal proteins (contactin, neuroglian), defense against reactive oxygen species (e.g., thioredoxin, superoxide dismutase, peroxiredoxin, glutathione peroxidase), stress response (heat shock proteins), nutrient transport and storage (vitellogenin, apolipophorin-III), detoxification (glutathione-S-transferase, cytochrome P450), glucose metabolism (glycoytic and citric acid cycle enzymes), signal transduction (14-3-3 protein, calcyphosine), and others.

Overall, this analysis thus allowed for a first glance at the biological processes that can be represented in antennal cells of a social insect and provides an encouraging outlook for future in-depth proteomics studies.

Venom

Harvester ants of the genus *Pogonomyrmex* are notorious for their potent sting, which they employ to deter potential predators. Their venom is, like most hymenopteran venoms, particularly effective against

vertebrates, and ranges among the most lethal venoms against mice of any arthropod known (68). Its components possess hemolytic, neurotoxic and algogenic properties, and can biochemically be separated into enzymes, peptides, and smaller molecules (69). Although important hemolytic components like barbatolysin fall into the second category, peptides are difficult to identify by the present method due to their short size. We therefore focus on venomous enzymes, many of which have been described for other hymenopteran species.

We identified 19 proteins unique to the poison glands or the venom of *P. rugosus*, five of which are most likely venom components. The remaining ones are mainly metabolic enzymes present in a wide range of tissues, and could therefore stem from the cellular material in the sample. The venom candidates are Dipeptidyl peptidase IV (PB13449), a hyaluronidase (PB18076), an acid phosphatase (PB24267), and two proteins similar to allergens known from *Solenopsis invicta*, Allergen 4 (PB26316) and Antigen 3 (PB25344) (70).

The serine peptidase Dipepdidyl peptidase IV is known to process mellitin, a main component of honeybee venom by cleaving its precursor (71), and has also been identified in crude *Nasonia vitripennis* venom extracts (72). Hyaluronidases, which have been found in high concentrations in *P. badius* venom (73, 74) are nontoxic agents acting as a "spreading factor" by hydrolyzing hyluronic acid in animal connective tissue. Acid phosphatases have also been described for *N. vitripennis* (72) and *P. badius* venom, where it is highly active and possesses a wide range of substrates (73, 74). Since little is know about the function of the *Solenopsis invicta* allergens, the role of the last two candidate proteins remains obscure (although PB26316 displays partial similarity to vespid Phospholipase A1).

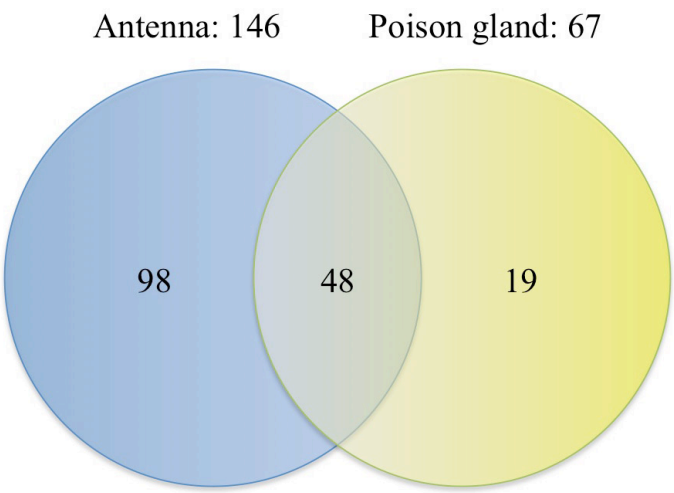


Fig. S14. Venn diagram illustrating the number of total and unique proteins identified by MALDI-TOF in proteome samples isolated from *Pogonomyrmex rugosus* antennae and poison glands, as well as those found in both samples.

11. Telomeres

Most insects outside of the Diptera have telomeres consisting of TTAGG repeats, which are assumed to be added by a canonical telomerase. The *Pogonomyrmex barbatus* genome encodes such a candidate telomerase (PB26363), so we searched the assembled genome for TTAGG repeats and found two long sets on the ends of two long scaffolds of 170 and 967 kb, as well as one short scaffold of 8 kb (Table S7). To find additional putative telomeres, of which we expect 32 given the karyotype of 16 chromosomes (75), the 8 kb paired ends reads were searched for TTAGG repeats. This strategy has been successfully used before to identify telomeres in the honey bee *Apis mellifera* and the flour beetle *Tribolium castaneum* (30, 76), albeit with longer fosmid mate pairs. Using the mate pairs of such reads we repeatedly re-identified all three of the assembled telomeres (see below for details), and 24 more at the ends of scaffolds ranging in size from 1 kb to 1.5 Mb (Table S7). Two of these could be manually extended to reach the telomeric repeats, and along with the three assembled telomeres allow recognition of what is likely a similar structure in all of these telomeres. Unlike most eukaryotic telomeres, except the placozoan *Trichoplax adhaerens* (77), there is only a short shared subtelomeric region of ~ 100 bases that starts with a long T homopolymer. Inside of this shared region each telomere has a unique region of several kb containing only divergent copies of repeats also present elsewhere in the genome and rarely shared by other telomeres, followed by a unique gene in either strand (Table S7).

Detailed Methods for Telomere Identification

The first 200 blastn matches to a string of 500 bases of TTAGG repeats amongst the 1,323,577 unique 8 kb PE reads, after filtering for redundant reads, were manually examined in detail. Five such reads had mate pairs that were too short to yield productive matches, while another six had no matches in the assembly or in the unassembled contigs and no additional matches in the raw reads. Seven mate pairs matched short contigs that could not reliably be identified as candidate telomeres (see below). Twenty-seven mates matched highly repetitive regions, typically in short contigs and typically satellites, so could not be convincingly matched to particular scaffolds. A rather high number of mate pairs, 57, each matched uniquely to a different location well within a large scaffold, in positions that are not candidates for telomeres. Because these are each unique matches, unlike the truly telomeric matches discussed below, we conclude that these result from chimeric molecules produced during the initial ligation step in the mate pair library construction during linker addition to the gel-fractionated fragments. This rate of clearly chimeric clones, over one quarter of the total examined, is of some concern, however as they will inevitably all be unique in connecting disparate parts of the genome, they did not likely cause serious problems with the assembly because the criterion for joining

contigs into scaffolds is conservative, as long as redundant reads are eliminated from the database first. The only problem they might cause is to prevent valid connections of some contigs into scaffolds.

The remaining 97 mate pairs repeatedly match in the 5' or 3' ends of long and sometimes short scaffolds, in the appropriate orientation and within 8 kb of the end of the scaffolds. By the end of the 200 TTAGG-containing matches examined, the TTAGG repeats were approximately 140 bases long, so this dataset has not necessarily been exhausted, however since all but two scaffolds were hit more than once, a Poisson distribution suggests this approach has identified almost all telomeric scaffolds that can be found using this methodology. The missing five telomeres presumably are amongst those detailed above with mate pairs with no matches in the existing assembly or highly repetitive matches. That the identified scaffolds are truly telomeric was confirmed by the fact that not only were the three scaffolds already assembled all the way to the telomeric repeats repeatedly re-identified this way, but all but two of the additional candidate telomeres were hit at least twice, confirming that these matches are not the result of chimeric molecules. For the two scaffolds with single hits, and all others examined as well, reversing the search and querying the raw 8 kb paired end reads with the last 8 kb of the matching scaffold yielded at least one additional mate pair linking appropriately into the TTAGG repeats, confirming them as telomeric scaffolds (Table S7). The three short scaffolds of 1–2 kb might be questioned as unique telomeres, because they might belong within gaps in the longer identified scaffolds, however two of them could be repeatedly connected using 8 kb mate pairs back through at least two "stepping stone" short unassembled contigs, indicating that they derive from telomeres that were poorly sequenced (the third has no 8 kb mate pairs connecting backwards from the telomeric end).

To examine the subtelomeric regions of these scaffolds, in addition to the three already assembled, attempts were made to manually extend the ends of the scaffolds towards the telomeres, however these were only successful for two that already reached almost to the subtelomeric region (Table S7). For all remaining 22 instances such manual assembly was prevented by long regions of simple sequence, typically A or T homopolymers or AT microsatellites. Most reads would terminate in these, preventing extension of the scaffolds. The shared subtelomeric region of these five completely assembled telomeres described above is ~100 bases including a long T homopolymer, one sequence of which is TTTCTTTTTTTTTTTTTTTTTTGCTTGTCGGTTGTGTTTTGGGTAACCTAATTGACTCGTCCTAACCAAATTGATGGAAAGTTAGGACGACATGG, followed by TTAGG repeats. The distance to the unique neighboring gene is only known for certain for four of these five completely assembled telomeres, ranging from 1–6 kb, while in the others it can range beyond 15 kb (Table S7). The assembled sequence beyond the end of each flanking unique gene is essentially unique to each telomere. The only repeats are generally not high identity, and almost all are present elsewhere in the genome. The greatest resemblance between these subtelomeric regions is that three telomeres share a ~ 500 bp repeat of 80–90% sequence identity that is only present elsewhere in the genome once.

The availability of multiple TTAGG-containing mate pair reads for several of these telomeres, especially the five that are completely assembled or manually assembled, allowed recognition that the TTAGG repeats are several kb long at most telomeres. Although nominally called 8 kb paired end reads, the inter-mate pair distance is actually generally around 6 kb, ranging from 5–9 kb (manual assessment from these and other regions examined). Sometimes mate pairs from TTAGG repeat regions matched from 1–6 kb inside the scaffold, indicating that the TTAGG repeats are around 6 kb long. They are not much longer than that, however, because a search of the 8 kb paired end reads with 100 bases of TTAGG repeats on either side of the 42 base linker yielded only three matches for both mates, indicating that few telomeres are much longer than 6 kb.

Table S7. Features of the identified telomeres of *Pogonomyrmex barbatus* in assembly v03

Scaffold	Size [kb]	End	Hits	Nearest gene: distance [kb], orientation, gene ID, protein
350310	1510	3’	8	5, –, PB20894, similar to ACYPI010120 (*Acyrthosiphon pisum*)
350194	1382	5’	5	0, +, PB16302, predicted protein
350382[a]	967	5’	3	4, +, PB25597, ATP-binding cassette subfamily E, member 1
350231[c]	727	3’	2	4, –, PB17746, similar to Y38F2AL.2 (*Apis mellifera*), B9 superfamily
350180	678	5’	3	6, +, PB15841, hunchback
350327[c]	634	5’	2	1, +, PB22133, alpha globin regulatory element containing gene
350322	602	3’	5	4, –, PB21954, ketohexokinase
350315[b]	578	3’	5	1, +, PB21422, phosphoribosylformylglycinamidine synthase
350363	544	3’	7	2, +, PB24332, WD40-repeat protein
349973	421	3’	4	2, –, PB13083, dishevelled 3
350236	358	3’	5	1, –, PB17870, putative ATP-dependent RNA helicase
349926	280	5’	2	2, –, PB11703, conserved hypothetical protein
350121	280	3’	4	1, +, PB14936, pre–rRNA-processing protein TSR1
350297[c,d]	262	3’	1	7, +, PB20197, ribosome maturation protein SBDS
350118[a]	170	5’	5	5, +, PB14828, mitogen activated protein kinase kinase 2
350047[c]	150	3’	1	0, +, PB14253, autophagy-specific gene 13
350352[b]	140	3’	3	6, –, PB23927, transmembrane protein 70
347531[c]	121	3’	1	3, –, PB10326, DNA replication complex GINS protein PSF1
346833	106	3’	3	3, +, PB10223, similar to CG1066-PA
349883	93	5’	4	0, +, PB11228, ATP-dependent RNA helicase
350100	57	5’	5	1, –, PB14750, transmembrane protein 20
346568	20	5’	6	5, –, PB10142, tolkin, tolloid-like
349382	15	5’	4	none
346727[a]	8	3’	3	none
346716	2	3’	2	none
350567	1	3’	2	none
349330[c,d]	1	3’	2	none

[a] Already assembled telomeres.

[b] Manually extended to the subtelomeric region or telomeric repeats.

[c] Additional mate pairs have been found that link to telomere repeats.

[d] A “no matches” mate pair has been extended to connect to these scaffolds.

Scaffolds are ordered by decreasing size (scaffold IDs are preceded by pbar_scf7180000). The distance to the nearest gene is from the end of the current assembled scaffold and does not account for the gap. The ID of the nearest gene is from the automated annotation, with approximate distance from the telomeric end of the assembled scaffold, orientation, and encoded protein identification.

12. Repetitive Elements

The *Pogonomyrmex barbatus* genome assembly spans 235 Mb of the expected 245–280 Mb genome, with the missing 10–45 Mb presumably comprising of satellite and simple repeats that could not be assembled. Furthermore, 15 Mb of the 235 Mb assembly are composed by N residues, which are also likely to represent repetitive sequences. Thus, we estimate that 25–85 Mb (9–18%) of the *P. barbatus* assembly is likely to consist of repetitive sequences, even before accounting for transposable element (TE) predictions.

One difficulty in new metazoan genome projects is that repeat libraries from other species are poor at identifying repetitive regions due to the extremely high sequence divergence and fast evolution of interspersed elements. TEs from even closely related species fail to identify repetitive regions, necessitating the creation of *de novo* repeat libraries. We generated *de novo* repeat libraries for *P. barbatus* using several methods. RepeatModeler is a tool that integrates RECON (78), TRF (79), and RepeatScout (80) data and classifies repeats with a RepBase RepeatMasker library. We also used PILER-DF (11) to identify regions present three or more times in PALS whole genome self-alignments. RepeatModeler identified 559 repeats (402 RECON, 157 Repeatscout), only 21% of which could be classified. PILER-DF identified 84 repeat regions. We 'downsized' the *de novo* repeat libraries by removing any element more than 80% similar over 80% of the length, resulting in 563 predictions. Of the 115 classified predictions, 40 were retrotransposons, 43 were DNA TEs, and 27 were other simple repeats. We screened out potential false positive by aligning our predictions with blastx to *Drosophila melanogaster* genes and UniProt proteins and removing any predictions with a bit score of 100 or an alignment over 50% of their length with 50% or more sequence similarity.

We generated a whole genome repeat annotation of the *P. barbatus* genome using RepeatMasker (http://www.repeatmasker.org, version open-3.0) and the RepeatRunner (3) subroutine that is integrated into the MAKER annotation pipeline (15) (Table S8). While highly fragmented, the 553-element *P. barbatus de novo* repeat library improved masking considerably (+13Mb, 6.25% vs. 11.52%) compared to using generic insects repeats from RepBase.

Viral sequences

Viruses specifically infecting hymenopterans have been reported for the red imported fire and *Solenopsis invicta* (81, 82) and *Apis mellifera* (83) and may play a significant role in colony success. We screened the *P. barbatus* genome for the presence of 1778 sequenced virus and viroid genomes. We report all tblastx hits that have bit scores greater than 100 or have more than 50% of the virus aligned in the genome with 50% or greater sequence identity. This analysis yielded ~ 300 significantly aligning regions spanning 620 kb of the genome (Table S9). Previous studies in *Nasonia vitripennis* identified poxvirus-associated PRANC domains

in the genome that appeared to be laterally transferred from *Wolbachia* endosymbionts. We downloaded *Nasonia*-defined PRANC domains from treebase.org (Study # S10521, (27)) and used them to train a custom hidden Markov model using HHMER 3.0 (84). We then scanned the *P. barbatus* genome using this HMM, but could not identify statistically significant PRANC domains.

Table S8. Summary of repetitive elements found in the *Pogonomyrmex barbatus* genome

Repeat type	No. of elements	Length occupied (bp)	% of sequence
Retroelements	9324	3962661	1.69
SINEs	923	94445	0.04
Penelope	395	193243	0.08
LINEs	2687	600216	0.26
CRE/SLACS	0	0	0
L2/CR1/Rex	1008	77345	0.03
R1/LOA/Jockey	981	251074	0.11
R2/R4/NeSL	24	2863	0
RTE/Bov-B	127	61270	0.03
L1/CIN4	0	0	0
LTR elements	5714	3268000	1.39
BEL/Pao	469	197603	0.08
Ty1/Copia	679	170128	0.07
Gypsy/DIRS1	4360	2863052	1.22
Retroviral	183	35596	0.02
DNA transposons	13068	5873276	2.5
hobo-Activator	673	95757	0.04
Tc1-IS630-Pogo	3693	1818005	0.77
En-Spm	529	64844	0.03
MuDR-IS905	2	479	0
PiggyBac	8	3533	0
Tourist/Harbinger	12	1053	0
Other	91	13570	0.01
Rolling-circles	0	0	0
Unclassified	55373	8819904	3.75
Total interspersed repeats	77765	18655841	7.93
Small RNA	305	52858	0.02
Satellites	12	878	0
Simple repeats	75045	4018862	1.71
Low complexity	66329	4626778	1.97

Table S9. Virus and viroid sequences found in the *Pogonomyrmex barbatus* genome

Virus family	No. of bases
Polydnaviridae	319,880
Baculoviridae	102,563
Caulimoviridae	82,407
Poxviridae	41,007
Phycodnaviridae	26,009
Mimiviridae	9,053
Unclassified dsDNA viruses	8,084
Herpesvirales	6,916
Iridoviridae[a]	6,034
Caudovirales	4,524
Ascoviridae	3,927
Cocaviroid	2,595
Apscaviroid	2,346
Hostuviroid	1,584
Nimaviridae	1,257
Pospiviroid	579
Coleviroid	504
Pelamoviroid	414
Parvoviridae	348
Unclassified phage	264
Asfarviridae	222
Total	620,517

Classes are ranked by total number of bases aligned by tblastx. Viruses and viroids mainly restricted to insects (orange), plants or algae (green), microbes (purple), vertebrates (grey), or unknown (white) are indicated ([a] Iridoviridae can also infect fish, amphibians, and reptiles).

13. Microsatellite Abundance and Diversity

The microsatellite DNA content of the *Pogonomyrmex barbatus* genome is 0.99%, higher than that of most other insects including the honey bee, *Apis mellifera* (0.77%), but is comparable with that of the parasitoid wasp *Nasonia vitripennis* (0.96%) (85). Counting 63,240 microsatellite loci, we estimate 269 microsatellite loci per Mb; a plethora of potentially informative length-polymorphic markers that can be exploited for mapping purposes. Dinucleotide repeats (7.80 kb per Mb genome sequence) are by far the most abundant motif type, accounting for 79% of the ascertained microsatellite DNA. The relative proportion of the remaining motif types decreases with increasing motif length: tri- (12.5%), tetra- (6.3%), penta- (1.6%), hexa-nucleotides (0.6%). The high dinucleotide microsatellite DNA content of the ant genome falls between that of *A. mellifera* (61%) and *N. vitripennis* (89%) and nourishes the idea that a high dinucleotide microsatellite DNA content and/or a high microsatellite DNA content in general could be a derived feature of the Hymenoptera or of a subordinated taxon within this insect order (e.g., the Apocrita) (85).

The *P. barbatus* genome assembly was scanned for microsatellite DNA with the aid of the software Msatfinder 2.0.9 (86) and using the same parameters as applied by Pannebakker et al. (85). Specifically, we searched for di-, tri-, tetra-, penta-, and hexa-nucleotides with a minimum of eight (dinucleotides) and five (remaining motive types) repeats and we considered interrupted microsatellites (details are given by (85)). The 4646 analyzed genome scaffolds spanned a total of 235 Mb.

14. Chemosensory Genes

Odorant Receptors

The Odorant Receptor (Or) family of seven-transmembrane proteins in insects mediates most of insect olfaction (e.g., 87, 88), with additional contributions from a subset of the distantly related Gustatory Receptor (Gr) family, for example, the carbon dioxide receptors in flies (89, 90, 91), and a subset of the recently described and unrelated Ionotropic Receptors (IRs) (92). The Or family ranges in size from a low of ten genes in the human body louse (28), to between 50 and 100 receptors in *Drosophila* flies (93, 94), the mosquitoes *Anopheles gambiae* and *Aedes aegypti* (95, 96), the silk moth *Bombyx mori* (97, 98), and the pea aphid *Acyrthosiphon pisum* (e.g., 99), to between 100 and 300 in the beetle *Tribolium castaneum* (100), the honey bee *Apis mellifera* (101), and *Nasonia* wasps (102). Although most genes in the *Drosophila* flies are scattered around the genome, with only a few in small tandem arrays, tandem arrays are more typical of the other species, especially those with large repertoires, from which it is inferred that these larger repertoires partly result from retention of gene duplicates generated in these tandem arrays by unequal crossing over.

Ants are expected to have a large Or gene family. Their sensory ecology and social behavior are largely dependent on chemical information, and several species have been shown to have ~400 glomeruli in the antennal lobes of their brains (e.g., 103, 104), with *Pogonomyrmex rugosus* workers having 365 ± 10 glomeruli (see main text and chapter 15). Assuming that ants are like flies in usually having one specific Or (plus the obligate heterodimer DmOr83b ortholog) per neuron type, with all neurons expressing a particular Or converging on a single glomerulus in the antennal lobe, the so-called one receptor-one neuron-one glomerulus hypothesis, we anticipated approximately 400 Ors. This assumes that an unknown subset of the 73 Grs and the 24 IRs in *P. barbatus* (see below) are also expressed in discrete olfactory sensory neurons that send axons to glomeruli in the antennal lobe.

Or Annotation

The *P. barbatus* Or family (PbOr) was manual annotated using methods employed before for the *Drosophila*, mosquito, moth, beetle, bee, wasp, aphid, and louse genomes. Briefly, tblastn searches were performed using *A. mellifera* Ors (AmOr), and sometimes *N. vitripennis* Ors (NvOr), as queries, and gene models were manually assembled in the text editor of PAUP* v4.0b10 (105), using the gene structures of the bee and wasp relatives to inform the ant genes. Iterative searches were also conducted with each new ant protein as query until no new genes were identified in each major subfamily or lineage. Occasionally the gene structures of ant genes were useful in informing improved gene models for some bee genes, specifically the 9-exon subfamily which is highly expanded in ants and wasps. The bee relatives are scattered throughout the AmOr naming system, because their relationship was not properly understood when they were annotated

in 2006. A short exon was missed from several of them, specifically *AmOr122–139*. In addition, recognition of the conserved 9-exon structure of these genes allowed refinement of the *AmOr172–174* genes, which were only recognized in light of the NvOr genes, but could not be completely built at that time (102). In addition, *AmOr175–177* were newly built in this large subfamily, and there are additional fragments of related genes in the poorly assembled AT-rich regions of the bee genome that might represent additional genes. All of the PbOr genes and encoded proteins are detailed in Dataset S2. In addition, the protein sequences of the PbOrs and AmOrs are provided in Dataset S1.

As described in the main text, the *P. barbatus* genome assembly v03 suffers from homopolymer length errors inherent to the 454 sequencing technology. No effort was made to correct these in introns or intergenic areas, however whenever a frameshift mutation appeared in a homopolymer in an exon, the raw reads were inspected and almost always contained additional reads without the frameshifts. In these cases the assembled sequence was fixed and these problems are noted in Dataset S2. In addition, as is typical of draft genome assemblies, gaps between contigs often interrupt gene models, especially when very similar genes are found in tandem arrays. These were repaired as best possible using the raw reads and are similarly noted in Dataset S2. There were several gene fragments resulting from assembly gaps that encoded less than half the typical length of an insect Or (200 amino acids), and these were not included in the analysis, although some likely represent intact genes.

Pseudogenes were translated as best possible to provide an encoded protein that could be aligned with the intact proteins for phylogenetic analysis, and particular attention was paid to the precise number of pseudogenizing mutations in each pseudogene. These were also newly reassessed in the honey bee and wasp genes to ensure comparable analyses for them (see main text). Again a 200 amino acid minimum was enforced for including pseudogenes in the analysis, with the exception of *PbOr150PSE*, which encodes only 159 amino acids, but represents a divergent lineage related to *AmOr160*. All ant, bee, and wasp Ors were aligned in ClustalX v2.0 (106) using default settings and problematic gene models and pseudogenes were refined in light of these alignments.

Or Phylogenetic Analysis

For phylogenetic analysis, the poorly aligned and variable length N-terminal and C-terminal regions were excluded (specifically before the conserved GhWP motif in the N-terminus and after the conserved SYFT motif in the C-terminus), as were major internal regions of length differences, especially a long length difference region between the longer *DmOr83b* orthologs (*PbOr1*, *AmOr2*, and *NvOr1*) and most of the other Ors. Other regions of potentially uncertain alignment between these highly divergent proteins were retained, because while potentially misleading for relationships of the subfamilies (which are poorly supported anyway), they provide important information for relationships within subfamilies.

Phylogenetic analysis of this large set of 877 proteins is difficult, but was successfully carried out in

the same fashion as for previous Or analyses (e.g., 101, 102). This involved a combination of model-based correction of distances between each pair of proteins, and distance-based phylogenetic tree building. Pairwise distances were corrected for multiple changes in the past using the BLOSUM62 amino acid exchange matrix in the maximum likelihood phylogenetic program TREE-PUZZLE v5.2 (107). These corrected distances were fed into PAUP* v4.0b10 where a full heuristic distance search was conducted with tree-bisection-and-reconnection branch swapping to search for the shortest tree. Given the large number of proteins, this search was unlikely to end and was terminated after two days with ~18 million trees examined. The resultant tree is shown in Fig. S15 (found at the very end of this documentd due to its size). Unfortunately this large number of proteins precludes distance-based bootstrap analysis to assess the confidence of major branches in the tree, but likely orthologs and obvious gene losses and subfamily expansions are noted on the right margin of the tree. The tree was manually colored and labels attached to lineages and subfamilies in Adobe Illustrator.

Or Results and Discussion

The PbOr gene set herein consists of 399 models. Of these, 55 (14%) are apparent pseudogenes, 35 apparent 454-caused frameshifts were corrected (for a rate of approximately 35 / (399 × 1200 bp) or 1 in 13,000 bases, compare chapter 6 of this study), and 70 gene models required repair of assembly gaps. The result is 344 apparently intact Or proteins, although 52 of these are still missing N-terminal, C-terminal, or internal regions, so their functionality remains uncertain (excluding the set of Or70–103 which have short N-terminal exons that are hard to recognize with confidence). Less obvious pseudogenes (for example with small in-frame deletions or insertions, crucial amino acids changes, or promoter defects) would not be recognized, so this total might be high. Approximately ten gene fragments remain so short and incomplete they were not included, but some might represent intact genes.

The automated gene modeling process had access to all available AmOrs and NvOrs, as well as other insect Ors in GenBank, for comparative information, and succeeded in building at least partial gene models for 255 of these 399 genes. However, as has been true for most other insect genome projects, just seven of these are precisely correct. Most others require multiple changes, while many instances of concatenated gene models were observed (Dataset S2), resulting in a total of 190 automated models representing these Ors (the most extreme was PB26716, which includes parts of ten genes and spans 15 genes over 86 kb on scaffold 7180000350254). Unfortunately because these genes are typically expressed at very low levels in only a few cells, they are seldom represented by ESTs in the whole body 454-sequencing project employed for this genome, indeed just eight genes had one to three useful ESTs representing them (Dataset S2), hence there is little experimental support for most gene models. Nevertheless, there is EST support for representatives of most NvOr subfamilies and many AmOr subfamilies (102) so these manually build gene models are highly confident. This situation again reveals the importance of manual annotation for

these rapidly evolving and highly divergent genes. Manual annotation was also obviously essential for detailed analysis of pseudogenes.

As expected there is a single highly conserved ortholog of the DmOr83b protein, named *PbOr1* in hopes of encouraging this convention for this gene and protein in other species, and sharing 77% amino acid identity with *AmOr2*, 76% with *NvOr1*, and 61% with *DmOr83b*. Only two other possible examples of simple orthology across these three hymenopteran genomes were observed, those of *PbOr145* with *AmOr161* (57% identity) and *NvOr296* (and *297PSE*) (45% identity), and those of *PbOr176* with *AmOr142* (50% identity) and *NvOr44* (35% identity), although the latter is not very confident. Other relatively simple relationships include *PbOr2/3*, which are clearly orthologous to *AmOr1/3* (67% identity) and *NvOr2* (60% identity), at the base of a large expansion in both bees and ants which include several complicated relationships as well as the only hymenopteran Or whose ligand is known, AmOr11 perceiving the major queen pheromone 9-ODA in bees (108).

There are many instances of differential gene lineage or subfamily expansions, as previously seen for the bee/wasp comparison (102), including differential expansions in the ant, for example an expansion of 28 ant genes related to *AmOr121* in the middle of the tree figure. The largest ant gene subfamily expansions, however, have occurred in a subfamily of 9-exon genes at the top of the tree. This subfamily consists of several discrete lineages in the bee totaling 43 genes, including *AmOr98–105*, *106–113*, *122–139*, *159*, *162*, *172–174*, and *175–177* (*AmOr140* might also belong in this subfamily as it has the same gene structure, but did not tree with it). This subfamily is expanded in *Nasonia*, where it consists of 90 genes in three small lineages (*NvOr210/211*, *191–196*, and *198–205*) and a large expansion of 74 genes (*NvOr129–190*, *197*, *206–209*, *212–217*, and *301*). Having recognized this distinctive subfamily while annotating this ant repertoire, they are numbered consecutively from *PbOr231–399*, a total of 169 genes.

This major 9-exon subfamily expansion in the ant is of particular interest as candidates for the cuticular hydrocarbon receptors in ants. The details have only been established for one ant, *Campanotus japonicus* (104, 109), in which females have a distinctive set of sensilla that house 150–200 neurons, each of which is presumed to express a particular Or, sending their neurons to a distinctive set of 150–200 glomeruli in the antennal lobe. Cuticular hydrocarbons are long non-volatile chemicals of enormous variety (e.g., 110), and it is not obvious which receptors likely perceive them. In *Drosophila melanogaster* two related lineages of Gustatory Receptors or Grs have been implicated in the perception of female cuticular hydrocarbons by males, but the exact ligand-receptor identification has yet to be made, and these are expressed in contact chemosensilla on the male foretarsi, and their neurons send axons to the sub-oesophageal ganglion instead of the antennal lobe (111, 112). As described below, *P. barbartus* has two expansions of Grs, but neither is large enough to encode such a repertoire of CHC receptors, and at least one is likely to encode receptors for bitter plant defensive compounds, therefore this expansion of ant Ors appears to be the strongest candidate for CHC receptors.

These species-specific expansions have typically occurred in large tandem arrays, some of which are evidently very old because they are shared with bee and even wasp, and commonly the genes within an array are so divergent they barely find each other in tblastn searches. For example, *PbOr2–51* is a 50-gene tandem array spanning 150 kb in 1,252 Mb scaffold 7180000349920 (Dataset S2), and is related to a 60-gene tandem array in bee described in detail previously (101), although the nine related genes in the wasp (*NvOr2–10*) are split on three scaffolds (102). Indeed the first and/or second gene in this array appear to be orthologous (noted above – *PbOr2/3*, *AmOr1/3*, and *NvOr2*), while the remainder form multiple species-specific gene lineage expansions (Fig. S15). Similar arrays characterize much of the large 9-exon subfamily, including two that span scaffolds (Dataset S2). *PbOr237–268* are 32 genes spanning ~100 kb at the 3' end of 1,806 Mb scaffold 7180000350284 and the 5' end of 546 kb scaffold 7180000350355, albeit not all in the same orientation. *PbOr235–367* are a perfect tandem array of 33 genes spanning ~212 kb on the reverse strand at the 5' end of 1,646 Mb scaffold 7180000350254 and the 5' end of 1,084 Mb scaffold 7180000350207.

Finally, the Or family reveals many instances of apparent gene loss, with some lineages completely absent from one or more of these three hymenopterans. In the absence of bootstrap analysis the numbers of these losses in each species cannot be confidently determined, and the uncertain orthology of several subfamily lineages also makes it difficult to determine the number of losses, but obvious examples are noted in Fig. S15. Separate subfamily tree analysis confirms all of these, and adds many more, confirming the dynamic gene family evolution known already from comparisons of other species Or repertoires.

Odorant Binding Proteins

We identified 15 genes in the *P. barbatus* genome encoding odorant binding proteins (OBPs), which are short secreted proteins typically containing six highly conserved cysteines that form three disulfide bonds (although some have lost two of these cysteines, in this set only PbOBP2). This is a low total compared with 21 in *A. mellifera* and 90 N. vitripennis. The genes and their encoded proteins are summarized in Dataset S2. There were at least partial automated gene models for all but one OBP (no. 12 has a internal gap in the assembly). Only four of these were perfect, however, with others requiring fixes of assembly gaps, correction of a frameshifting homopolymer, joining across scaffolds, or addition of missing exons. Only one gene could not be fully built, that for OBP9 is missing the expected N-terminal exon that typically encodes the signal sequence at the start of these secreted proteins, and unfortunately there are no ESTs for it. Like OBPs in other insects, most of these genes are highly expressed enough to have ESTs in whole body EST projects like that undertaken for this ant. These ranged from zero ESTs for three genes, to fewer than ten for five more, and up to about 1000 for OBP3.

There are simple apparent orthologs for a subset of the 21 OBPs known from the honey bee genome, some of which are conserved throughout endopterygote insects. These are AmOBP 1, 5, 6/8, 9, 10, and 11. The apparently orthologous ant OBPs were given the same names. This ant also has a small expansion of

OBPs distantly related to AmOBPs 7 and 12, given the numbers 2, 3, 4, 8, 12, 13, 14, 15. There are no obvious ant relatives of the AmOBPs 2, 3, 4 and 7. Finally, this ant has a single OBP7 with a possible relationship to the bee OBP expansion of 13–21. Thus ants and bees share a core set of six conserved OBPs that are probably involved in multiple functions, some not even related to odorant binding, given their expression in other tissues. On the other hand, they have differentially expanded different lineages of species-specific OBPs, which are more likely to be involved in olfaction. One of these is the fire ant *Solenopsis invicta* Gp-9 protein, which is an OBP implicated in regulation of queen numbers in colonies. The closest relative in P. barbatus are OBPs 13 and 14, sharing only 40% amino acid identity.

Gustatory Receptors

The Gustatory Receptor (Gr) family consists of 73 genes (Dataset S2), of which eight were successfully repaired, while three are still missing C-terminal exons in gaps. Six genes spanned scaffolds and twelve are apparent pseudogenes. Like *A. mellifera* and *N. vitripennis*, there are no obviously alternatively spliced genes like those seen commonly in the Gr family in flies and *Tribolium castaneum*, and no candidate carbon dioxide receptors were identified. Most genes are simple orthologs of the *A. mellifera* and *N. vitripennis* Grs (Fig. S16), and there are at least partial automated gene models for all of these (*PbGr1–11*), while the two large subfamily expansions only have occasional gene models which commonly fuse genes (Dataset S2). With the exception of *PbGr3*, these genes are essentially unrepresented in the available EST set.

Ionotropic Receptors

The Ionotropic Receptor (IR) family consists of 24 genes, compared with ten in *A. mellifera* and ten in *N. vitripennis* (113). These genes encode members diverged from the ancestral Ionotropic Glutamate Receptor (iGluR) family of neurotransmitter receptors, which itself comprises ten genes in *P. barbatus* including members of each of the principal subfamilies of animal iGluR (AMPA, Kainate and NMDA). At least two IRs and three iGluRs are predicted pseudogenes, containing one or two internal frameshift and/or nonsense mutations. Phylogenetic analysis and sequence comparison of ant IRs identified putative orthologs to many members of the repertoire of conserved IRs that are present in most or all sequenced insect genomes and shown to be expressed in the antennae of *D. melanogaster* and *A. mellifera* (e.g., *IR25a*, *IR8a*, *IR93a*, *IR76b*) (92) (Fig. S17). In addition, there are a number of ant-specific divergent IRs, which display no obvious orthology to other hymenopteran or insect receptors.

Most genes described in this chapter are not accurately represented in the Official Gene Set v1.1. The manually curated protein sequences can be found in Dataset S1.

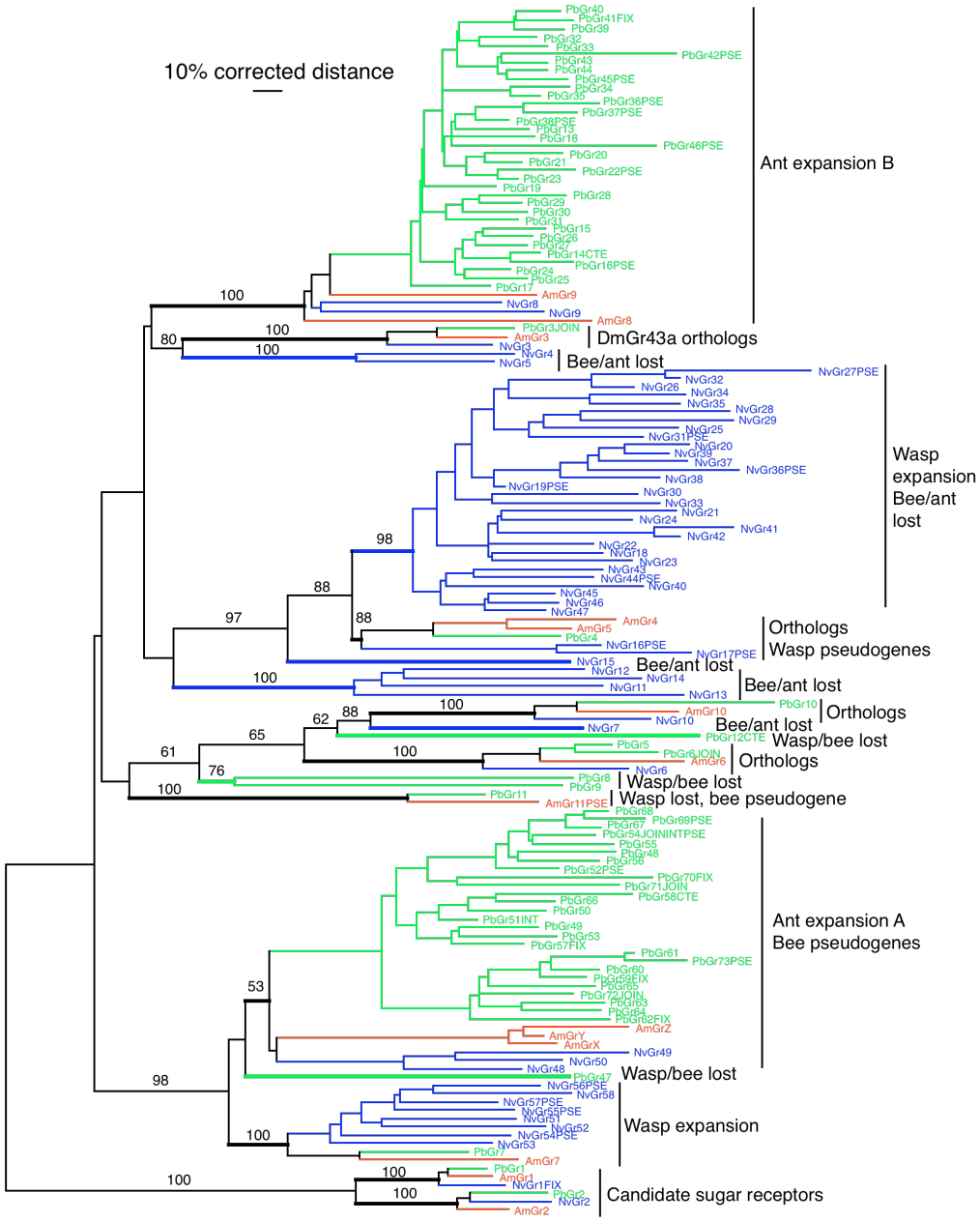


Fig. S16. Phylogenetic tree of the hymenopteran Gustatory Receptor (Gr) genes. This is a corrected distance tree generated as in Robertson et al. (102). The two candidate sugar receptors were defined as the outgroup to root the tree, based on the highly divergent sequence and gene structure of this gene subfamily (93, 114). The *Pogonomyrmex barbatus* ("ant"), *Apis mellifera* ("bee"), and *Nasonia vitripennis* ("wasp") gene / protein names are highlighted in green, red, and blue, respectively, as are the branches leading to them to emphasize gene lineages. Numbers above branches are percentage support from 1000 bootstrap replications of uncorrected distance analysis. Double thickness branches indicate inferred independent Gr lineages.

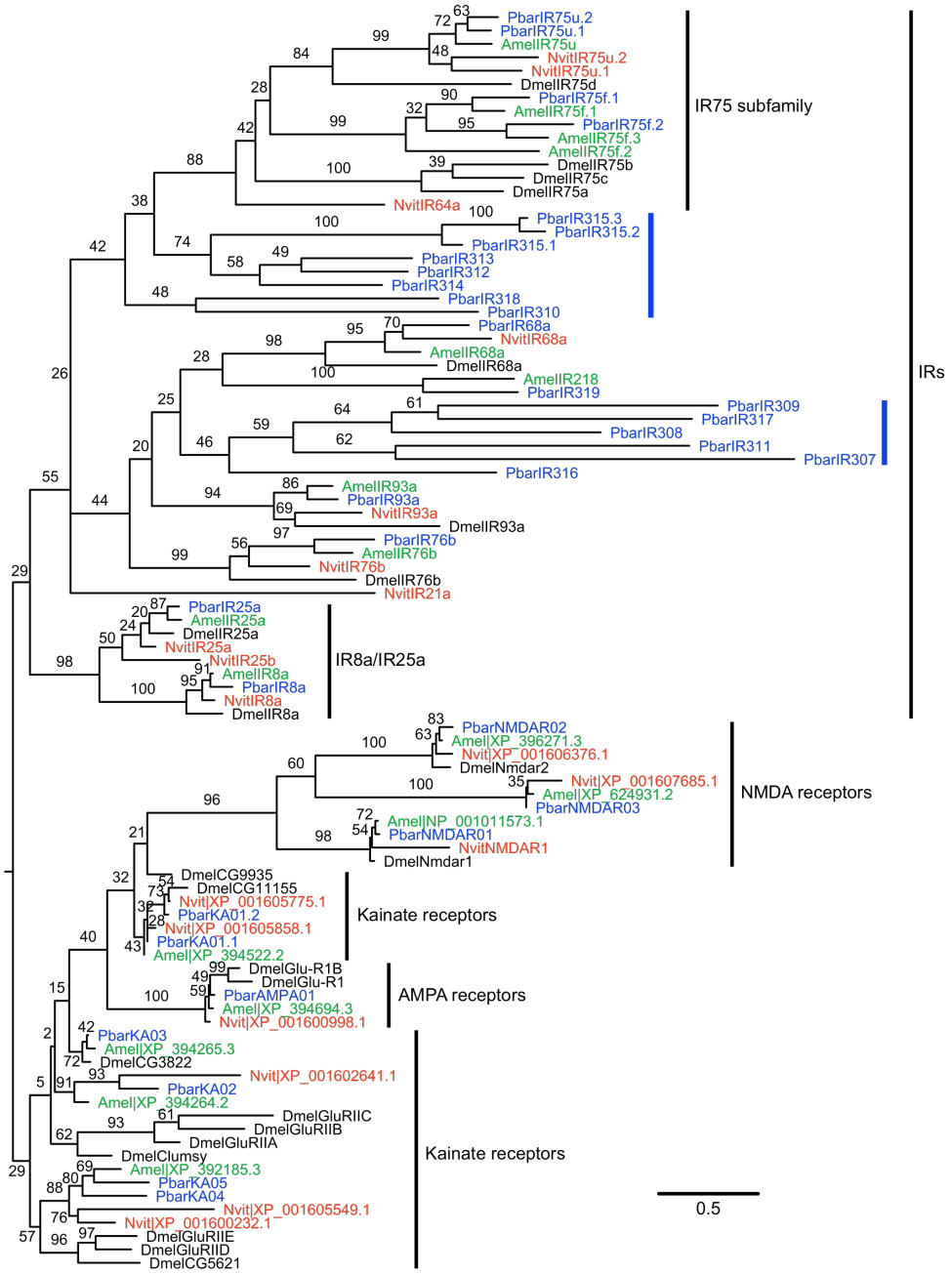


Fig. S17. Phylogenetic tree of the *Pogonomyrmex barbatus* (blue), *Apis mellifera* (green) and *Nasonia vitripennis* (red) Ionotropic Glutamate Receptor (iGluR) and Ionotropic Receptor (IR) genes, as well as *Drosophila melanogaster* (black) orthologs. Two *P. barbatus*-specific expansions of IRs are highlighted with a blue vertical line. Protein sequences were aligned with ProbCons, and the tree was built with RAxML under the WAG model of substitution with 1000 bootstrap replicates. Bootstrap values for each branch are indicated on the tree. The scale bar represents the number of substitutions per site.

15. Olfactory Glomeruli

Across the animal kingdom, odors are perceived by receptor neurons that come in contact with odorants in the environment (e.g., in a human's nose or an insect's antennal sensilla). The dendritic membranes of these sensory cells comprise odorant receptor (Or) molecules that bind a more or less narrow range of chemical compounds. Via a cascade of cellular processes, odorant binding leads to electrical activity in the receptor cells (115). In many (but not all) animal phyla, each receptor neuron expresses only one kind of Or protein, which thus determines a particular neuron's odor specificity. Each Or molecule is coded by a specific Or gene and the number of these genes has been established for some animal model systems (e.g., less than 100 in some fish, less than 400 in humans and Chimpanzees, about 1000 in mice, 54–71 in different *Drosophila* species, about 80 in mosquitoes and about 170 in honeybees (116)). The total number of Or genes in a genome probably indicates the range and precision of different odors that a particular animal species can discriminate.

Peripheral olfactory neurons from the nose (vertebrates) or antenna (insects) send their nerve fibers into a primary olfactory center in the brain, referred to as 'olfactory bulb' (vertebrates) or 'antennal lobe' (insects). Across phyla, these primary olfactory centers are organized in a strikingly similar way where all the (hundreds to thousands of) neurons that express a particular Or protein converge onto a common target region, referred to as an olfactory glomerulus. The presence of any odor is represented by the simultaneous activity of many olfactory receptor neurons. At the level of the antennal lobe (or the vertebrate olfactory bulb), different odors are represented by different, overlapping sets of activated glomeruli, giving rise to odor specific spatial maps of active glomeruli (115).

From this brief description of primary olfactory systems follows that the number of different Or molecules, and of the Or genes by which they are coded, should be strongly correlated with the number of olfactory glomeruli. In adult *Drosophila melanogaster*, the best studied system, 47 glomeruli and 62 olfactory receptor genes have been described (117, 118), the difference resulting from a few cases of Or co-expression and Or genes expressed in larval but not adult olfactory systems (119). Honey bees feature 160–165 glomeruli (120) and 166 functional Or genes (101, see also chapter 14), the mosquito *Anopheles gambiae* 60 glomeruli (121) and 79 Or genes (95), and the parasitic wasp *Nasonia vitripennis* 259 glomeruli and 225 functional Or genes (102, see also chapter 14).

Ants generally feature high numbers of antennal lobe glomeruli: ca. 466 in *Camponotus japonicus* workers (104), 434 in *C. floridanus* (122), 340–492 in different worker morphs of *C. sericeus* and 408–501 in different worker morphs of *C. compressus* (123), 396–442 in the leafcutting ant *Atta vollenweideri* (124) and up to 630 glomeruli in other Attini (125). One would therefore expect ants to also feature high numbers of Or genes, but the genomes of none of these ants have been sequenced yet. Here, we therefore establish the

number of glomeruli for *Pogonomyrmex rugosus* harvester ants, the closest relative of *P. barbatus*, whose genome is the topic of the current study.

We found an average of 365 ± 10 antennal lobe glomeruli for workers (n = 5) and 354 ± 10 for virgin queens (n = 5). The difference between workers and virgins is small yet almost significant (t-test; P = 0.053). Smaller numbers of glomeruli have also been described for virgins of different *Camponotus* species (104, 123) and for *A. vollenweideri* (124). Importantly, the number of glomeruli we found in our samples is in the same range as the number of annotated Or genes for *P. barbatus* in the present study, 344 (see chapter 14). The general 'rule' that one olfactory glomerulus corresponds with one Or gene therefore seems to apply to ants too, despite their overall very high number of olfactory glomeruli. Most ants have a wide range of diets, often including plant and animal matter, and heavily rely on olfaction for foraging as well as social interaction. The high number of olfactory receptor genes and olfactory glomeruli provides a perfect base for the olfactory lifestyle of ants.

Materials and Methods

P. rugosus workers and virgin queens were taken from laboratory colonies originating from collections in Maricopa County, Arizona, USA. Ants were decapitated, the head capsule cut open frontally, and the brain dissected out under fixative (4% formaldehyde in phosphate buffer, pH 6.8) and fixed over night. Brains were rinsed in four repeated changes of buffer and then stained in 1% aqueous osmiumtetroxide for 2 hours at 4 °C and for an additional 30 minutes at room temperature. Brains were then rinsed, dehydrated, plastic-embedded (Spurr's low viscosity medium) and polymerized at 65 °C. Brains were sectioned on a sliding microtome at 7 μm thickness, assuring that each glomerulus (average diameter more than 20 μm) was represented in at least three consecutive sections. Each section of the mounted and cover-slipped brains that contained parts of the antennal lobes (on average about 30 sections) was photographed (SPOTflex digital camera, Zeiss Axioplan microscope) and images were manually aligned (Adobe Photoshop CS3). For counting glomeruli (Fig. S18), each glomerulus' cross-section was marked and compared with the previous and subsequent section to assure that each glomerulus was only counted once.

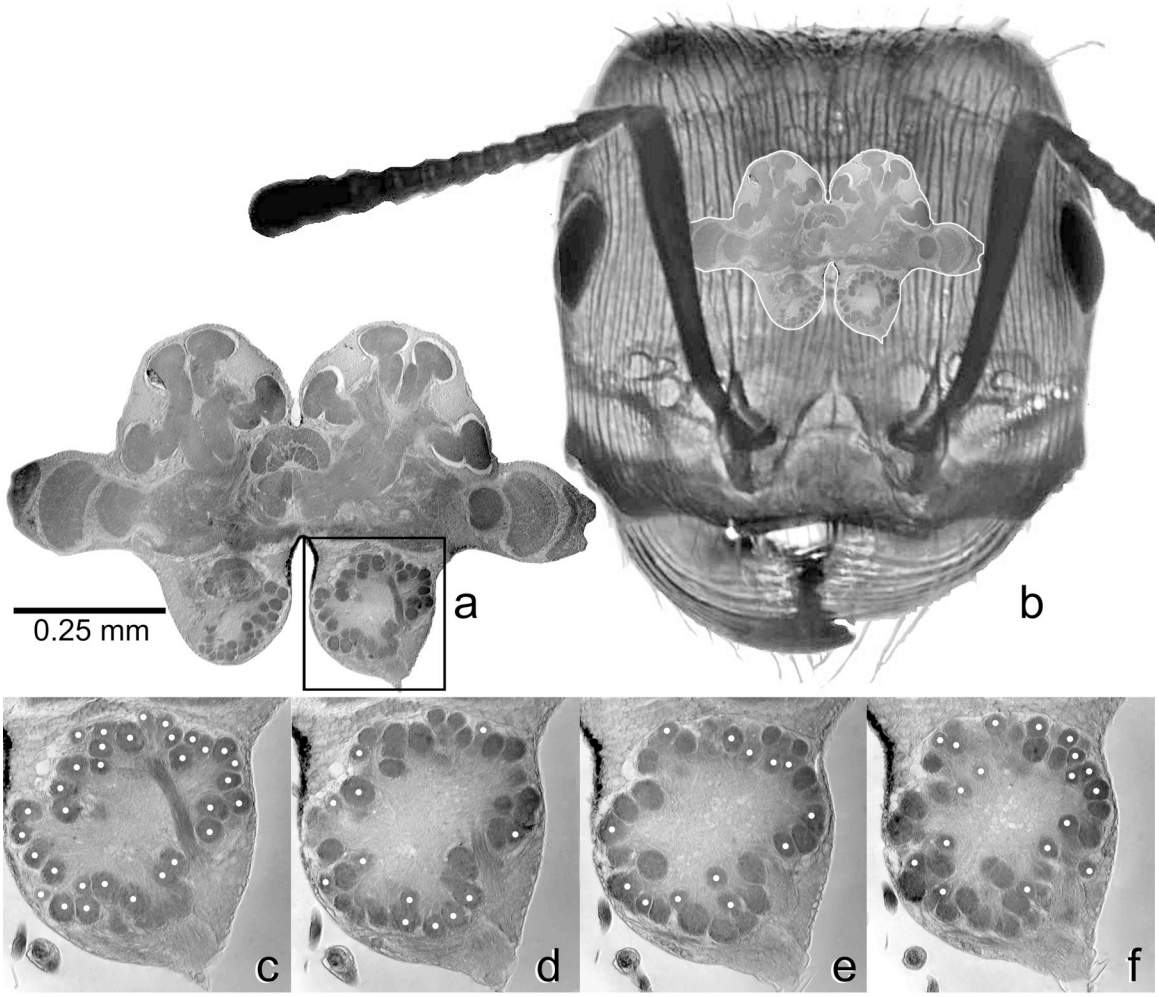


Fig. S18. *Pogonomyrmex rugosus* worker brain (a); approximate size and position of the brain with respect to the head capsule indicated in (b); four consecutive antennal lobe sections enlarged (c–f; enlarged area boxed in a); all glomeruli marked in (c); 'new' glomeruli (not present in previous sections) marked in (d–f), respectively. Scale bar refers to (a).

16. Cytochrome P450 Genes

With 72 genes in the cytochrome P450 superfamily (P450s), the *Pogonomyrmex barbatus* genome encodes more P450s than the *Apis mellifera* genome (126) with its 46 members, but fewer than *Nasonia vitripennis* with 92 genes (127), *Tribolium castaneum* with 123 genes (30), *Drosophila melanogaster* with 85 genes (128) and *Acyrthosiphon pisum* with 83 genes (129).

The *P. barbatus* genome includes orthologs of highly conserved P450s classified in the CYP2 and mitochondrial clans that are involved in ecdysteroid metabolism (130). *P. barbatus* encodes 40 P450s in the CYP3 clan, a group of P450s associated with detoxification of xenobiotics (131), which is intermediate between the gene counts in *A. mellifera*, 28, and *N. vitripennis*, 49, and is consistent with the number of CYP3 P450s in other insect genomes. The *A. mellifera* genome encodes just four P450s in the CYP4 clan, which is far fewer than *A. pisum*, 32, *D. melanogaster*, 32, or *T. castaneum*, 41. With 18 CYP4 clan P450s *P. barbatus* is intermediate between the highly reduced set in *A. mellifera* and the 29 in *N. vitripennis.* The function of CYP4 P450s is not clear, though some CYP4s are associated with pheromone metabolism (132). Given the importance of chemical communication in coordinating social behavior in bees and ants it is surprising that the genomes of *P. barbatus* and *A. mellifera* encode fewer of the putatively pheromone-related CYP4 P450s than non-social insects.

The *P. barbatus* genome includes 13 pseudogenes (> 50% the length of a full gene) which is more than the five pseudogenes in *A. mellifera* and ten in *N. vitripennis* (Fig. 3, main text). Most hymenopteran P450 pseudogenes are similar to CYP3 P450s, except for a single CYP4 pseudogene in *N. vitripennis* and five CYP4 pseudogenes in *P. barbatus.*

Most genes described in this chapter are not represented in the Official Gene Set v1.1. The manually curated sequences can be found in Dataset S1.

17. Immune Genes

Innate immune response is the most important defense against pathogens in insects. Social insects are exceptional in having diverse repertoire of social defenses, e.g., hygienic behavior (133) and antimicrobial secretions (134). In this context the importance of individual defenses in social insects have been questioned and indeed, honey bees (*Apis mellifera*) seem to have fewer immune genes compared with the fruit fly (*Drosophila melanogaster*) and mosquito (*Anopheles gambiae*) (135). However, it is not clear whether the paucity of immune genes is attributable to sociality since two solitary insects, the parasitic wasp *Nasonia vitripennis* (27), and the pea aphid, *Acyrthosiphon pisum* (136), also have fewer immune genes when compared with dipteran insects.

Manual annotation of immune genes in the red harvester ant (*Pogonomyrmex barbatus*) established the presence of both the classical signaling pathways IMD, Toll, Jak/STAT and JNK and chitinases, which are beginning to be recognized as major effectors of the immune response (137). The classical pathways consist of recognition of pathogens followed by intracellular signaling and expression of effector proteins such as antimicrobial peptides. In *P. barbatus*, the recognition protein repertory is similar to *A. mellifera*, which is about half the number of recognition proteins compared to dipterans (*A. gambiae* and *D. melanogaster*) (135). This pattern of reduction and similarity to *A. mellifera* continues in the chitinases. The signaling genes are mostly present as single copies in insect genomes and this holds true also in *P. barbatus*. Comparison of effector proteins across insects is challenging as they tend to show lineage specific expansions and losses (138–140). While comprehensive characterization of antimicrobial peptides (AMPs) in ants requires further computational and experimental analysis, the initial annotation in this study found orthologs for *hymenoptaecin*, *defensin*, *abaecin* and *naickin*. All except *hymenoptaecin* were found in multiple copies and the total number of these AMPs clearly exceeded the corresponding AMP group in the honey bee.

The small number of recognition proteins in the *P. barbatus* genome could indicate that they are infected by a narrow set of pathogens and thus do not require a wide array of recognition proteins. On the other hand the diverse and duplicated AMPs highlight the importance of the physiological immune system in *P. barbatus*. It is possible that ants use additional, yet to be characterized, proteins in recognizing pathogens. Immunological assays should help to clarify this issue.

Specific Methods for Immune Gene Annotation

A non-overlapping set of *A. mellifera* and *D. melanogaster* immune genes were used as a query against the *P. barbatus* genome scaffolds using blastn in standalone BLAST. The hit scaffold regions were used in reciprocal blastx against the honey bee Official Gene Set pre-release 2 or *D. melanogaster* protein database downloaded from NCBI. A subset of the candidate immune gene loci were used in manual annotations.

These scaffold regions were used in blastx against the *P. barbatus* Official Gene Set v1.1 (OGS1.1) and the resulting gene predictions were blasted against NCBI's non-redundant protein database to verify orthology, and further manually annotated in Apollo (18). In addition, a set of *N. vitripennis* immune proteins were separately blasted against the *P. barbatus* OGS1.1 either because *A. mellifera* and *D. melanogaster* queries did not have matches or these proteins were only found in the *N. vitripennis* genome. The resulting hits were blasted against NCBI's non-redundant protein database and the reciprocal best hits were used in manual annotation. For some gene families, protein sequences of several insects were aligned with ClustalW2 (106) and a profile of the alignment was made using profile hidden Markov models (141) in HMMER3 (84). This profile was used in a HMMER3 search against the *P. barbatus* OGS1.1 in order to find homologs that belong to the gene family.

18. Wing Polyphenism and Reproductive Division of Labor

Wing polyphenism and reproductive division of labor between queens and workers are two major and universal features of eusociality in ants (142). Both of these features evolved approximately 150 million years ago (143, 144), and have been key to their amazing evolutionary success – wing polyphenism was key for allowing ants to colonize the ground, while reproductive division of labor was key for their organization into eusocial colonies (142). The gene networks that underlie wing polyphenism (Fig. S19*A*) and reproductive division of labor (Fig. S19*B*) are generally conserved between ants and the model fruit fly *Drosophila melanogaster* (145, 146). In ants, however, these networks have evolved the ability to be differentially expressed between winged reproductive castes and wingless sterile worker castes in response to either environmental or genetic factors (145, 146). In response to these factors, these networks must simultaneously produce fully functional wings and reproductive organs in the queen and male castes, but interrupt the expression of specific genes in the network to halt the development of wings and constrain reproduction in worker castes. While we have cloned and identified just a few candidate genes that are differentially expressed between queens and workers in these networks (145, 146), our ability to understand the evolutionary and developmental dynamics of these genes both within and between species has been limited by the absence of an ant genome.

We therefore annotated genes in the networks that underlie wing polyphenism and reproductive division of labor in the ant *Pogonomyrmex barbatus*. Wing polyphenism in *P. barbatus* occurs between queens and workers, and even occurs between queens in other *Pogonomyrmex* species (147). Reproduction is also highly regulated because queens perform all the reproduction and workers are functionally sterile (148, 149), even though they possess ovaries. In addition, because this species determines its castes genetically (150), these networks must respond to both genetic and the environmental factors. Together, this suggests that the networks underlying these processses require more complex regulation than other genes in the genome. We therefore assessed whether the networks underlying critical processes in *P. barbatus* are putatively more or less methylated than the genome as a whole.

We followed the same method of Elango et al. (36) to assess whether or not the genes we annotated show signatures of putative methylation relative to the rest of *P. barbatus* genome. For the coding region (exons and introns) of each annotated gene we calculated the frequency of the observed number of CpG dinucleotides using a custom Perl script. We calculated the observed over expected 'CpG[o/e]' values using the formula CpG[o/e] = (P(N1N2)/P(N1)*P(N2)), where N1 and N2 indicate the two DNA bases present in the CpG dinucleotide. We then calculated the mean of CpG[o/e] values for three sets of genes: genes underlying wing polyphenism, genes underlying reproductive division of labor, and genes known to control apoptosis (Fig. S19, Dataset S2). Genes that control apoptosis are intimately linked to the networks that

control wing development (151, 152) (Fig. S19) and oogenesis (153), and thus we included these genes in our analysis.

In order to compare the mean CpG[o/e] values for these three sets of genes to a genome-wide mean CpG[o/e], we segmented the scaffolds from the draft assembly of the *P. barbatus* genome into 1 kb non-overlapping fragments using custom Perl scripts. We measured the frequency of CpG dinucleotides and calculated the CpG[o/e] values for each 1 kb fragment using custom Perl scripts based on the same equation as above. We then calculated a genome-wide mean CpG[o/e] by taking the mean CpG[o/e] of all 1 kb fragments. Although there are alternative methods for generating a genome-wide mean CpG[o/e], e.g., see Genome Compositional Analyses above, we used this specific method because it was the only way we could perform the equivalent analyses in *P. barbatus* and compare them to *D. melanogaster*, an insect species which lacks a CpG methylation system.

To test whether or not there are any significant differences in the mean CpG[o/e] values between our three sets of annotated genes and the genome-wide mean CpG[o/e], we performed a statistical randomization procedure as follows: first, we generated a random distribution of CpG[o/e] values by randomly selecting 50 CpG[o/e] values from the genome-wide distribution. We randomly selected 50 because this is approximately the same number of genes as that contained within each of the three sets of genes we manually annotated. We then calculated the mean CpG[o/e] for this random distribution. Second, we repeated this first step 10,000 times, and then ploted all 10,000 randomly generated mean CpG[o/e] values (x-axis representing the mean CpG[o/e] values and y-axis representing the frequency). Third, we then determined where the observed mean CpG[o/e] for each of the three sets of genes we annotated fall with respect the randomly generated mean CpG (O/E) values. If it falls within the top or bottom 5% of the distribution of randomly generated mean CpG[o/e] values, then the observed mean CpG[o/e] values are significantly different than the genome-wide mean CpG[o/e]. We performed the same statistical analyses in *D. melanogaster* using orthologs of the genes we annotated in *P. barbatus.*

We discovered that the mean CpG[o/e] for genes (coding regions) in the network underlying reproductive division of labor ($n = 37$; mean = 1.18; $P < 0.00$) and apoptosis ($n = 18$; mean = 1.39; $P < 0.00$) are significantly less (Fig. S20*A*) than the genome-wide mean CpG[o/e] (mean = 1.73). The genes (coding regions) in the network underlying wing polyphenism is also less (Fig. S20*A*) than the genome-wide mean CpG[o/e] and is only marginally non-significant ($n = 41$; mean = 1.47; $P = 0.06$). The mean CpG[o/e] of the *D. melanogaster* orthologs (coding regions) that underlie wing development (mean = 0.95; $P = 0.86$), reproduction (mean = 0.98; $P = 0.98$), and apoptosis (mean = 1.00; $P = 0.99$) are not significantly different (Fig. S20*B*) than the genome-wide mean CpG[o/e]. Together, these results indicate that developmental genes in the network underlying wing polyphenism, reproductive division of labor, and apoptosis have a distinct methylation signature relative to the rest of the *P. barbatus* genome.

According to Elango et al. (36), genes that are methylated in the germline should exhibit a mean CpG[o/e] that is under 1.0, while genes that are methylated in the soma should exhibit a mean CpG[o/e] that is over 1.0. The fact that mean CpG[o/e] values of the three sets of devlopmental genes are greater than 1.0, but significantly less than the genome-wide mean CpG[o/e], may indicate that they are still methylated in the soma, but have a different methylation signature than the rest of the genes in the genome. The high level of significance for genes underlying reproductive divison of labor and apoptosis may be due to their dramatic regulation between the two castes. This is partly because many of the genes underlying reproductive division of labor are germline specific, and partly because apoptosis is a major mechanism by which they are differentiating castes. The marginal non-significance of the genes underlying wing polyphenism may be due to the fact that they are used so broadly and in so many different structures during development. Although these intriguing results await empircal validation, they open many avenues for future research.

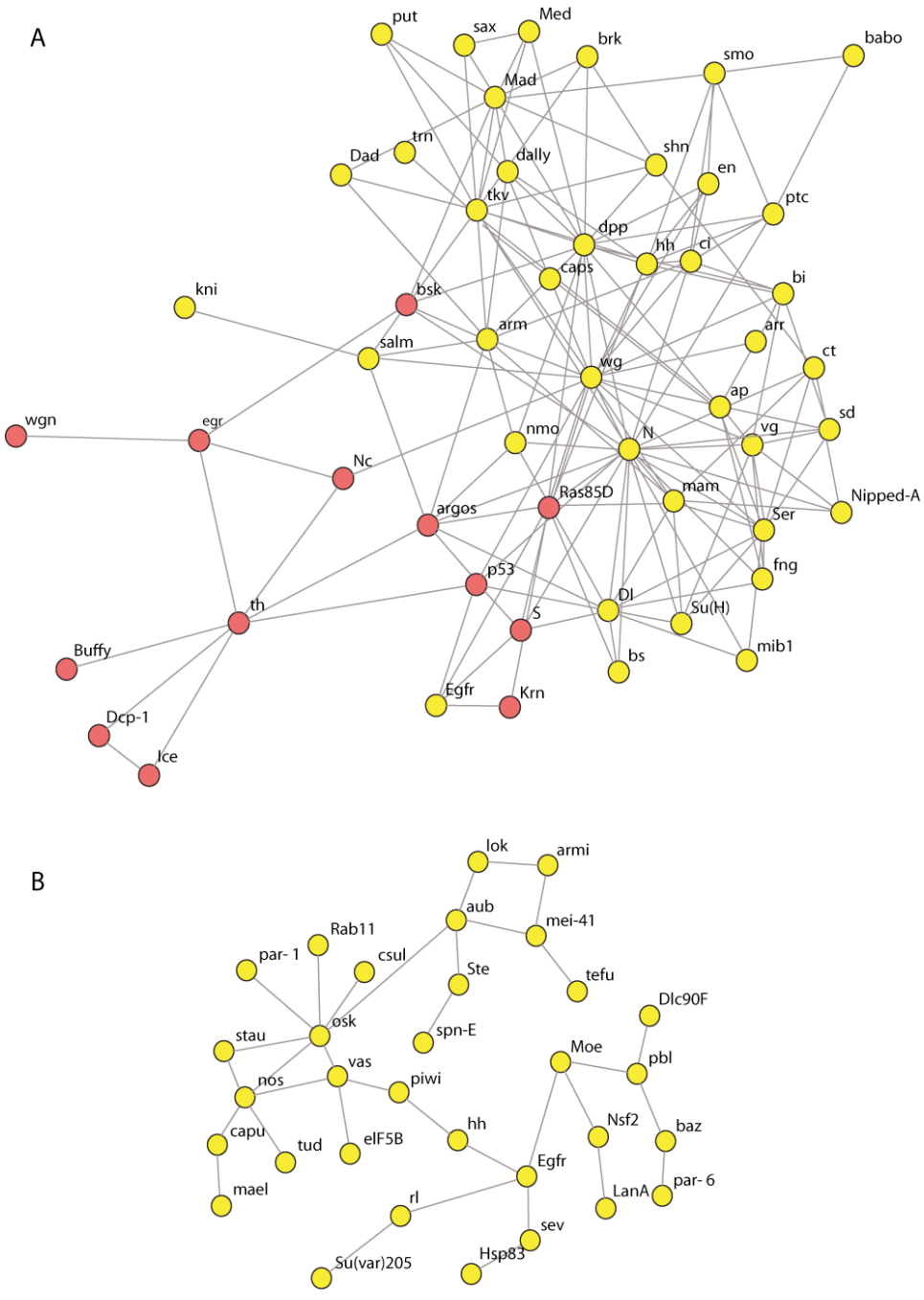
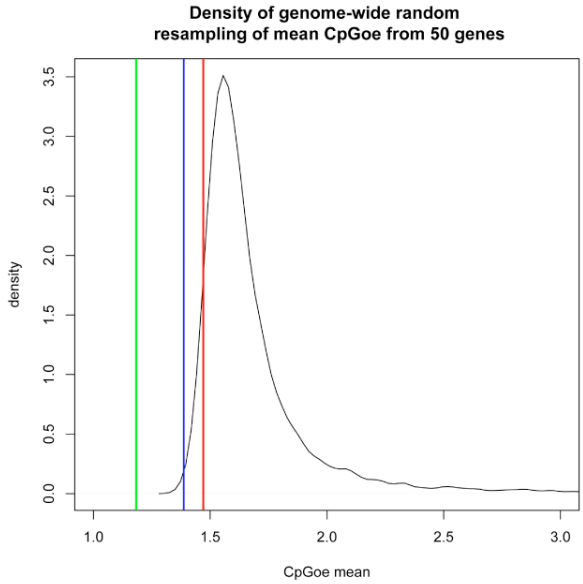


Fig. S19. Gene networks underyling (A) wing polyphenism and (B) reproductive division of labor in *Pogonomyremx barbatus*. Yellow-filled circles represent genes, while the letters beside each gene represent the abbreviated name of the genes. Light gray lines indicate a genetic interaction between two genes, and can be either an activation or suppression. (A) Red-filled circles represent 'apoptosis' genes. All genetic interactions are based on experimentally-validated interactions known from FlyBase (41), and were reconstructed in the IM Browser in DroID (154, 155). Full gene names are found in Dataset S2.

A



B

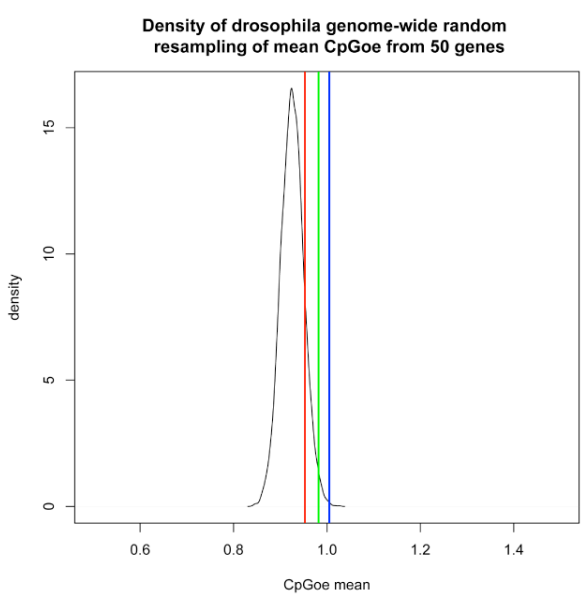


Fig. S20. Density plot of frequency (y-axis) versus mean CpG[o/e] (x-axis) for (A) *Pogonomyremx barbatus* and (B) *Drosophila melanogaster*. The observed mean for genes in the networks underlying wing polyphenism (red), reproductive division of labor (green), and apoptosis (blue) are plotted relative to the distribution of CpG[o/e] values for all genes in the genome.

19. Candidate Caste Determination Genes

Caste determination is the hallmark of eusocial species, where females differentiate into worker or queen castes (or discrete forms of specialized worker) during larval development (156). Work on the honey bee (*Apis mellifera*) model system has elucidated the importance of specific genes and networks in the process of caste determination (157). Evidence of gene involvement in caste determination is typically differential gene expression between developing queen and worker larvae, and in some cases includes gene knock-down via RNAi (158). Multiple gene networks have been implicated in caste determination, but much attention has focused on genes associated with nutrient signaling and nutrient use / storage because of their association with differential growth and the historic knowledge that differential nutrition is a sufficient signal to alter caste fate (see (156) for a review). For this reason, the insulin / TOR signaling pathway has been a particular focus and was specifically selected for annotation in the harvester ant. Other candidate "caste genes" annotated were the hexamerins; these are storage proteins that are differentially expressed in workers and queens at both larval and adult stages in ants, bees and wasps (159–161) and are associated with variation in colony founding strategies in ants of the genus *Pogonomyrmex* (162). The list of candidate caste genes annotated and described here is by no means exhaustive; for example, mitochondrial genes associated with differential metabolism are known to be differentially expressed in developing larvae (163). Other gene groups previously associated with caste differentiation are being annotated and are described in other chapters, among these are: P450 genes, yellow and major royal jelly proteins, and methyltransferase genes. Furthermore, while candidate caste genes are those thought to be far upstream in the caste determination cascade, many genes downstream are likely regulated by intercellular signaling molecules such as hormones and biogenic amines (see chapter 21).

Molecular markers have been used to study the genetic caste determination system of the J-lineage *P. barbatus*. While many markers have been assayed, only few have alleles that segregate between the J1 and J2 lineages (164), making them informative for determining whether an undifferentiated larva will develop as a queen or worker. Three of these markers are microsatellites, L18 (165), Myrt3 (166), and Pb8 (167), while another is the allozyme locus phosphoglucoisomerase (PGI) (168). We identified the genomic regions of these markers and searched 100 kb around each locus with the hope of identifying candidate genes involved in causing incompatibilities, and the loss of phenotypic plasticity, between the two lineages (169). Table S10 shows the gene models and their distances from each locus. Any of these genes may play a causal role in generating genetic caste determination, but one stood out as an interesting candidate, *lozenge* (*lz*). *lozenge* mutants in *Drosophila melanogaster* are most noted for their eye phenotypes, but are often sterile. There have been many studies of these mutants and *lz* was one of the first genes fine mapped due to high recombination in its region. Some *lz* mutants are sterile, and in females this is due to a loss of oogenesis

and/or a loss of a spermathecum (170–172). As these two traits are diagnostic of differences between queen and worker ants *lz* is a viable candidate for affecting caste determination.

Table S10. Gene models lying within 100 kb of genetic loci associated with genetic caste determination

Locus	Gene ID	Distance [kb]	Putative identity
L18	PB12735	6	lozenge (fragmentary)
	PB12734	22	runt
	PB12741	23	lozenge
	PB12736	27	–
	PB12744	43	–
	PB12739	47	lozenge
Myrt3	PB11346	7	– (contains repetitive DNA)
	PB11339	7	– (contains repetitive DNA)
	PB11345	11	hypothetical protein (model problematic)
	PB11349	17	hypothetical protein
	PB11348	19	RNA helicase
	PB11347	21	eukaryotic translation initiation factor (fragmentary)
	PB11344	24	budding uninhibited by benzimidazoles (Bub3)
	PB11350	27	alternative testis transcripts ORF (transcript variant)
	PB11350	28	alternative testis transcripts ORF (transcript variant)
Pb8	PB24573	4	– (contains repetitive DNA)
	PB24577	16	jagged
	PB24579	28	jagged
	PB24580	34	–
	PB24576	37	jagged
	PB24578	42	hypothetical protein
	PB24574	44	ADP-ribosylation factor
	PB24575	46	Hspb associated protein
	PB24565	48	NEDD8-conjugating enzyme
PGI	PB11744	3	RAD1
	PB11747	6	CG8311

PB11746	6	fatty acid binding protein (Fabp)
PB11742	16	dorsal interacting protein 3 (Dip3)
PB11735	18	Synaptotagmin IV
PB11748	20	CG7264
PB11749	23	protofilament ribbon protein
PB11743	26	hypothetical protein
PB11737	29	CG1105

20. Yellow / Major Royal Jelly Protein Genes

Evidence suggests that the yellow/major royal jelly protein gene family is ancient but has been lost in many lineages, as yellow-like proteins have been found in species of bacteria, fungi and insects (173, but see also 174). Although their role in microbial organisms is unclear (27), all insects with sufficient sequence data investigated to date possess yellow-like genes (149). Interestingly, they have not been detected in non-insect arthropods such as *Daphnia pulex* and *Ixodes scapularis* (175). Yellow proteins function in a diversity of processes including development, locomotion, melanization, immune response, and mating and courtship behavior. In *Apis mellifera* rapid duplications of an ancestral *yellow* gene similar in structure to the extant *yellow-e3* have led to the expansion of the major royal jelly protein subfamily (MRJP), which functions in a nutritive role relevant to caste determination. Proteins from the MRJP subfamily also have age, sex, and caste specific expression including expression in the brain implicating a role in behavior (173). A similar but apparently independent expansion of MRJP-like genes has occurred in *Nasonia vitripennis*, a solitary and parasitoid Hymenopteran.

A total of 16 yellow and MRJP genes were detected and annotated in the *Pogonomyrmex barbatus* genome assembly using the BLAST strategy described above (chapter 3). Six of these genes are shorter than the average length for yellow genes, appear to be fragmentary, and/or lack an open reading frame and thus likely represent pseudogenes. Of the ten complete genes, seven have direct similarity to yellow genes of *Drosophila melanogaster* (*Pbar_Y-b*, *-c*, *-e3*, *-g*, *-g2, -h* and *y*) and three others to yellow genes found in *N. vitripennis* and *A. mellifera* (*Pbar_Y-x1a*, *-x1b*, and *-x2*). The remaining six genes seem to have been fragmented and possess no EST support. Two of them are yellow-like genes with no clear orthologs in other insects studied so far (*Pbar_Y-1* and *Pbar_Y-2*), and four genes share striking similarities to the MRJP and MRJP-like genes of *A. mellifera* and *N. vitripennis*. The fact that all detected MRJP-like genes were fragmentary suggests that these genes may have been pseudogenized, and have lost their function in *P. barbatus*.

Phylogenetic Analyses

To elucidate the homology relations of members of the yellow family, we performed a phylogenetic analysis. First, genes homologous to the *D. melanogaster* reference gene set were retrieved from the genomes of *A. mellifera*, *N. vitripennis* and *Tribolium castaneum* by BLAST. Amino acid sequences of these genes and those from *P. barbatus* and *D. melanogaster* (84 genes in total; putative *P. barbatus* pseudogenes were not included) were aligned using MAFFT v6 and the E-INS-i algorithm (176). Ambiguously aligned positions were automatically removed by Gblocks (177) using low stringency parameters, which resulted in a final dataset containing 172 amino acid positions. The evolutionary model with the best fit to this dataset,

WAG+G+F, was determined by ProtTest (178) according to the Akaike Information Criterion corrected for small sample size. Based on this model, a maximum likelihood tree was reconstructed using RAxML v7.0.4 (2). Nodal support values were obtained by the rapid bootstrap algorithm as implemented in RAxML (500 replicates). We used a yellow gene sequence from the bacterium *Dienococus radiodurans* as the outgroup for the analysis (*DR_1790*).

The phylogenetic tree (Fig. S21) provides strong support for several yellow gene clades that contain genes with the same letter designations across the five taxa. We thus assigned corresponding labels to orthologous *P. barbatus* genes. Most of these clades are characterized by single-copy genes, although moderate expansions (e.g., *N. vitripennis* and *T. castaneum* genes in clade *yellow x1*) and infrequent losses have occurred in individual taxa. *P. barbatus* is represented by single copy genes in all clades except *yellow x1*, where two copies are found, and *yellow e* and *yellow f*, which seem to have been lost in *P. barbatus* (the latter is restricted to non-hymenopterans). The tree also shows that although in the *D. melanogaster* genome the genes *Dmel_Y-e*, *-e2*, and *-e3* lie adjacent to one another, the *Dmel_Y-e* gene falls into a clade separate from the *Dmel_Y-e2* and *-e3* genes (which we choose to call the *yellow e* and *yellow e3* clades, respectively), demonstrating that they might not be as closely related as previously suspected (179). Although the nodal support values for many yellow clades are strong, only a few inter-clade relationships could be resolved. Notably, the clades *yellow b, c, f, h* and *y* form a well supported monophylum we chose to call the yellow core group, as it contains the originally described yellow gene of *D. melanogaster* (*Dmel_Y-y*). Further, the *yellow x2* genes, which are restricted to the hymenopteran taxa, seem to be the closest relatives of the ancestral genes that gave rise to the independent MRJP expansion in *A. mellifera* and *N. vitripennis*. Although four putative MRJP pseudogenes were found in *P. barbatus*, they were too fragmentary to include in the phylogenetic analysis.

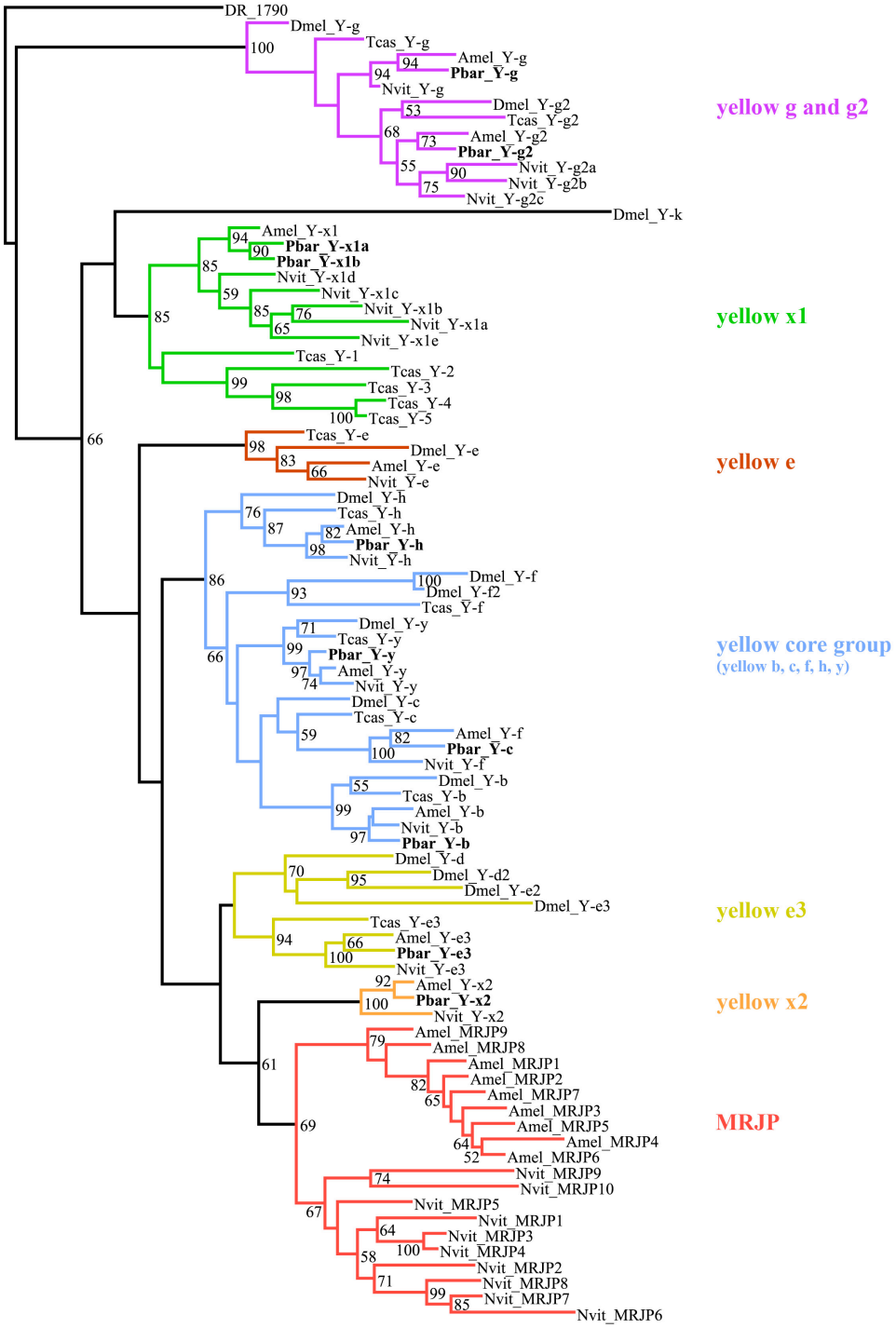


Fig. S21. Maximum likelihood tree of the yellow / MRJP genes found in the genomes of *Pogonomyrmex barbatus* (Pbar, in bold), *Apis mellifera* (Amel), *Nasonia vitripennis* (Nvit), *Drosophila melanogaster* (Dmel) and *Tribolium castaneum* (Tcas). Support values > 50 based on 500 rapid bootstrap replicates are shown at the nodes of the tree.

21. Biogenic Amine Receptor Genes

The biogenic amines are small signaling molecules derived from amino acids, and which act as neurotransmitters, neuromodulators and neurohormones. The biogenic amines act upon target cells by binding to specific G protein coupled receptors. Activation of receptors then leads to changes in second messenger levels such as cAMP and intracellular Ca^{2+}, phosphorylation of proteins and changes in gene expression. Signaling via biogenic amines such as dopamine and serotonin is found throughout the animal kingdom and the evolution of subtypes of dopamine or serotonin receptors predates the vertebrate – invertebrate split (180–182). In insects, the biogenic amines modulate a number of processes including learning and memory (183–185), sensory processing (186, 187), locomotion (188, 189) and metabolism (190). In addition, the mechanisms underlying signaling via biogenic amines is of particular interest for those working on social insects, as they have been implicated in division of labor (191), responses to pheromonal cues (192), nestmate recognition (193) and reproductive dominance (194, 195). The number and type of biogenic amine receptors found in the genome of the red harvester ant, *Pogonomyrmex barbatus*, are similar to those found in the honey bee, *Apis mellifera*. Both *P. barbatus* and *A. mellifera* appear to have one less tyramine receptor and one more octopamine receptor than *Drosophila melanogaster*. The increase in octopamine receptor number may reflect octopamine's role in processes such as division of labor (191) and social trophallaxis between nestmates (196).

22. RNAi Pathway Genes

A recent study silencing the vitellogenin receptor gene (*VgR*) in fire ant (*Solenopsis invicta*) virgin queens first demonstrated the existence of RNAi in ants (197). In our study, a full repertoire of RNAi pathway genes (*Drosha, Pasha, Exportin 5, Dicer, Loquacious, AGO, R2D2*) has been manually annotated, suggesting the existence of RNAi pathway in *P. barbatus* (Fig. S22). Noticeably, all genes were found as a single copy except *Loquacious*, which has two copies. Domain analysis of these two proteins by InterProScan (198) indicates that Loquacious 1 has three DRSM (Double-stranded RNA binding motif), which are located at the N-terminal, middle and C-terminal, while Loquacious 2 has only one such motif at the N-terminal. The unambiguous identification of Loquacious 2 needs further experimental validation, although it has almost full coverage of EST evidence. A very recent study indicates that Loquacious is required for miRNA biogenesis as well as for processing of dsRNA into mature siRNA duplexes by Dicer-2 (199). Therefore, it is tempting to study whether the difference in the number of DRSM can affect each Loquacious' involvement in either pathway. Besides, several important additional genes are also present in our annotation. These include genes encoding CRM1, a protein that mediates the import of miRNA guide sequences to the nucleus (200) (where they possibly play a role in chromatin remodeling), and C3PO. C3PO, the third component of RISC in addition to Dicer 2 and R2D2, is a complex of Translin and Trax, which were recently found to play a role in activating RISC by removing cleavage products of the siRNA passenger strand in *D. melanogaster* (201). Furthermore, key proteins (AGO3, Aubergine and Piwi) involved in the biogenesis and function of piRNAs (piwi-interacting RNAs) (202) are also present in our annotation. Other miRNA and RNAi pathway related genes annotated are summarized in Table S11.

Table S11. Summary of other RNAi pathway related genes

Name	Potential function	Reference
Aubergine and *Spindle E*	Required for activating RNAi pathway during *D. melanogaster* oocyte maturation	(203)
Aubergine and *Vasa RNA helicase*	Retrotransposon silencing in the female germline of *D. melanogaster*	(204)
Sid1	Encoding a transmembrane protein involved in the widely conserved systemic RNAi pathway	(205)
Elp1	Protein interacts with Dicer 2 and participates in RNAi; also plays a role in transposon suppression	(206)
Vig and *Fmr1*	Each encoding a putative RNA-binding protein identified as a RISC component	(207)
Belle, Pros45 and *Chc*	Chc, a component of the endocytic machinery, might participate in RNAi in *D. melanogaster* S2 cells through the uptake of dsRNA, while Belle and Pros45 may act at later steps of the silencing process	(208)

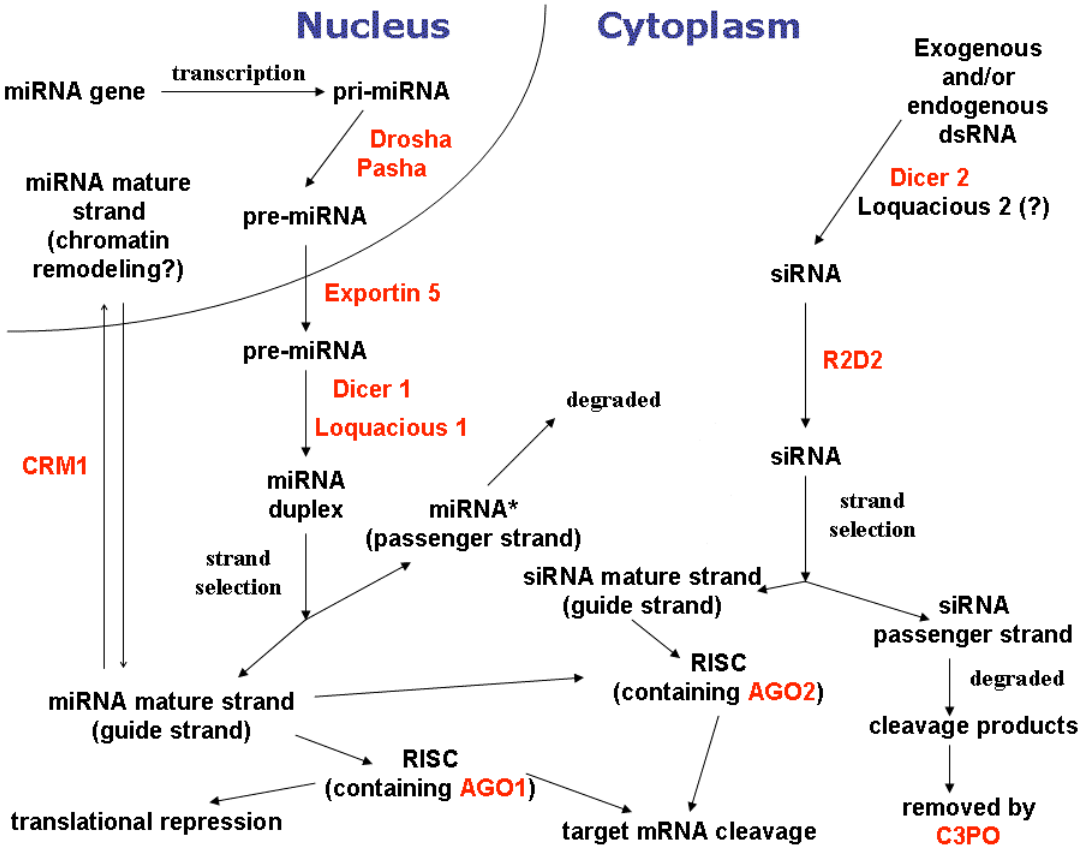


Fig. S22. A proposed schema of the RNAi pathway in *P. barbatus.* Genes in red have been annotated manually.

23. MicroRNAs

Our first strategy for identifying miRNAs invoked blastn searches of known miRNAs from miRBase release 14.0 (209–211) against the *Pogonomyrmex barbatus* genome assembly using word size 7 and E-score threshold ≤ 0.01. These searches identified approximately 100 candidate *P. barbatus* miRNAs with significant matches to miRNAs from other species. Sequences including 75 nucleotides (nt) upstream and downstream of the match were extracted from the genome. Nucleotide sequence alignments were performed using ClustalW (106), aligning the putative miRNA sequence with known miRNAs from the honey bee, *Apis mellifera*, the jewel wasp, *Nasonia vitripennis*, and the fruit fly, *Drosophila melanogaster*. *P. barbatus* miRNA candidates were trimmed leaving only the most likely pre-miRNA sequence. We used RNAfold (212) to score the folding energy (minimum 20 Kcal/mol) and assess the structure of the pre-miRNA candidate. This analysis resulted in the identification of 69 conserved miRNAs in *P. barbatus*.

The second strategy for miRNA identification uses three-way genome comparison between *P. barbatus*, *A. mellifera*, and *N. vitripennis*, for the identification of micro-conserved sequence elements (MCEs). MCEs are typically 20–29 nt in length and have previously been exploited to identify miRNAs (213, 214). The identification of three-way genome intersections results in hundreds of thousands of MCEs across a rather large evolutionary distance (approximately 190 million years). MCEs representing simple sequence repeats were excluded and the remaining sequences were clustered to reduce redundancy before being mapped back to the genome. Approximately 75 nucleotides of sequence flanking the MCEs were extracted. We mapped the extended *P. barbatus* sequences to the *N. vitripennis* and *A. mellifera* genomes and retained only those sequences with identifiable homology. Work in this area is currently ongoing. With just over 6,000 sequences remaining in our study, we expect to identify additional miRNA candidates by scoring, and folding the extended sequences, similar to the methods described above.

24. DNA Methylation Toolkit

CpG DNA methylation is an important regulator of gene expression in many animal taxa (215–218); this includes the inhibition of individual gene transcription to the silencing of entire chromosomes, as in mammal X chromosome inactivation (219). Three DNA methyltransferase genes, *Dnmt1–3*, are involved in the methylation of the cytosine in CpG dinucleotides, but each *Dnmt* functions in a different context (217). *Dnmt1* is involved in the maintenance of CpG methylation in the germ line, ensuring consistent methylation from parent to offspring; this gene is thus implicated in genomic imprinting. *Dnmt2* is a $tRNA^{Asp}$ methyltransferase and its function is poorly understood, but appears to be involved in the silencing of transposable elements in the *Drosophila melanogaster* (220). Interestingly, *Dnmt2* is the only DNA methyltransferase present in the Diptera, and in *D. melanogaster* it predominantly methylates CpT and CpA dinucleotides instead of CpG, which are those predominantly methylated by *Dnmt2* in vertebrates (216). *Dnmt3* is the *de novo* methyltransferase and methylates DNA in response to environmental stimuli. RNAi knock-down of *Dnmt3* in *Apis mellifera* larvae was sufficient to alter patterns of caste determination; knocked-down individuals tended to develop as queens (221). This study suggests that *Dnmt3* is an upstream regulator of many caste-related genes and that nutritional stimuli (differential diet) alter its expression. Furthermore, the results of Kucharski et al. (221) suggest that *Dnmt3* represses the transcription of genes associated with queen development in larvae that develop as workers.

The common ancestor of the arthropods and chordates likely had all three *Dnmt* genes, but there have been various duplications and deletions within the arthropods (patterns of gain and loss summarized in (27)). Only *Dnmt2* is present in all sequenced arthropods. *Dnmt1* and *Dnmt3* have been lost in several lineages, but both are present in arachnids, crustaceans and hemimetabolous insects (though *Dnmt3* was lost in the louse). *Dnmt3* occurs in triplicate in humans, but is either single copy, or lost completely in the insects; the moths, beetles and flies lack *Dnmt3*. *Dnmt1*, on the other hand, has been both duplicated and lost. The hemimetabolous insects have it in duplicate, as does *Apis mellifera*. *Nasonia vitripennis* has gained a third copy. The Coleoptera and Lepidoptera have only one copy, and the Diptera have lost it altogether.

The genome of the red harvester ant, *Pogonomyrmex barbatus*, has a complete DNA methylation toolkit, which is predicted based on the findings in *A. mellifera* (31) and *N. vitripennis* (27) along with the finding that CpG methylation is present across multiple origins of social insects (32). We found only a single copy of *Dnmt1* in *P. barbatus*, compared with two and three in *A. mellifera* and *N. vitripennis*, respectively. The additional copies in *A. mellifera* and *N. vitripennis* are due to lineage specific duplications. The role of *Dnmt1* duplications is unclear, but may play a role in the different pattern of distribution of CpG[o/e] in *P. barbatus* compared to *A. mellifera* (see CpG dinucleotide analysis in chapter 4 above, Fig. S7, Table S10). Both *Dnmt2* and *Dnmt3* exist as single copies. All three *Dnmt* genes have conservation in the expected

conserved domains. We also found a complement of three methyl binding proteins (MBD) which function in gene silencing via the recruitment of additional proteins (222).

25. Delta-9 Desaturase Genes

A distinguishing feature of social insects, in fact all animal societies, is the development of a complex communication system. Implicit to the ability to communicate is the ability to differentiate between individuals as 'self' (i.e., recognition of individuals within one's social group) or 'other' (i.e., recognition, or lack-there-of, of individuals belonging to another group or species) (223). With the exception of unicolonial species, all social insects possess some form of colony recognition that deters members of nearby colonies from entering a foreign colony (142). The evidence to date shows that colony recognition in many ant species, including *Pogonomyrmex barbatus*, is predominantly based on signals produced by cuticular hydrocarbons (CHC) (224–226). The resulting colony recognition signals are mixtures of the innate CHC profiles of all workers, the queen, and the environment (227).

A recent review of ant CHCs shows that nearly 1000 CHC compounds have been found across 78 species studied (228). Within these, two particular biochemical pathways are used to alter *n*-alkanes – the addition of double bonds and methyl braches – which suggests that these two compound-groups likely contribute substantially to colony recognition (228). Until now, no genes have been identified or isolated that are known to influence the CHC patterns of any social insect, however, previous studies of CHC components in *Drosophila melanogaster* show that carboxylases, elongases and desaturases each play important roles in CHC biosynthesis (229, 230). The best studied of these are the desaturases, which create carbon-carbon double bonds in *n*-alkanes forming monoenes and dienes. Currently, only three desaturase genes, *desat1*, *desat2*, and *desatF* (syn. *Fad2*), are known to contribute specifically to *D. melanogaster* alkene synthesis and phenotypic variation of cuticular hydrocarbons (230, 231). Because *P. barbatus* queens and workers produce variable quantities of five CHC alkenes, these three desaturase genes make excellent genome query candidates for the study of *P. barbatus* CHC alkene genes.

Annotation Analyses

Manual annotation was carried out as described in chapter 3 above. The three *D. melanogaster* query genes, *desat1*, *desat2*, and *desatF*, produced the same eleven candidate genes in the *P. barbatus* genomes: ten predicted functional Δ9 desaturase genes and one fragmentary desaturase gene, nine of which are supported by ESTs, and nine of which group together along a 90 kb region of the genome (Fig. S23). Reciprocal BLAST analyses found that four of these *P. barbatu*s genes are most similar to *D. melanogaster desat1*, four to the *D. melanogaster* Δ9 desaturase gene CG9747, one to the *D. melanogaster* Δ9 desaturase gene CG9743, one to the *D. melanogaster* Δ9 desaturase gene CG15531, and the fragment to the *D. melanogaster* Δ9 desaturase gene CG8630. Notably, only the previously mentioned four *P. barbatus desat1*-like genes returned results most similar to the original query set, i.e., no *desat2* or *desatF*-like genes were found. A

comparative analysis of the desaturase genes of two other hymenoptera genomes, the parasitic wasp *Nasonia vitripennis,* and the honey bee *Apis mellifera*, identified a total of 16 and seven respective predicted Δ9 desaturase genes. In *N. vitripennis*, five of the 16 total desaturase genes were found to be most similar to *D. melanogaster desat1* according to reciprocal BLAST (one more than found in *P. barbatus*), and in *A. mellifera* three of the seven total desaturase genes were found to be most similar to *D. melanogaster desat1* (one less than found in *P. barbatus*).

Phylogenetic Analyses

To further understand the relationships between the Δ9 desaturase genes of *P. barbatus* and other insects, we performed a phylogenetic analysis using the genes found in *P. barbatus* and six other insect taxa with completed genomes: *A. mellifera*, *N. vitripennis*, *Tribolium castaneum*, *D. melanogaster*, *Anopheles gambiae*, and *Acyrthosiphon pisum*. For that purpose, the amino acid sequences of 71 homologous genes were aligned using the L-INS-i algorithm implemented in MAFFT v6 (176) (note that partial gene sequences were removed to improve the final length of the trimmed dataset: a *P. barbatus* fragment similar to CG8630, an *A. mellifera* partial gene similar to CG15531, and an *A. mellifera* partial gene similar to CG8630). Ambiguously aligned positions were eliminated by Gblocks (177) set to low stringency parameters, resulting in a final dataset comprising 225 amino acid positions. The evolutionary model with the best fit to this dataset, CpREV+G, was determined using ProtTest (178) according to the Akaike Information Criterion corrected for small sample size (the LG model was not considered since it is not implemented in the phylogenetic software used). Based on this model, a maximum likelihood tree was reconstructed using RAxML v7.0.4 (2), obtaining nodal support values by a rapid bootstrap analysis of 500 replicates (BS).

The phylogenetic analysis reveals the existence of five major clades within the Δ9 desaturase gene family in insects (Fig. S24). The strongly supported (BS = 100) clades D and E are comprised of single-copy genes, although some members have evidently been lost in specific lineages. Clade C (moderately supported with BS = 75) is notable for a gene expansion in *N. vitripennis* that contrasts with apparent gene loss in the aculeate lineages (*P. barbatus* and *A. mellifera*). Since the respective gene is also missing in *D. melanogaster*, reciprocal BLAST searches based on this taxon erroneously identify members of this clade as orthologs of other genes, which demonstrates that this method alone can be misleading when assessing homology relations. Further, all CG9747 desaturases form a well supported (BS = 98) monophyletic group (clade B) that is characterized by multiple rounds of gene expansion in *P. barbatus* and *N. vitripennis*, which seem to have occurred both before and after the split of these lineages. Strikingly, this group is not represented in the honey bee. All the remaining Δ9 desaturases are found in clade A (sub-divided into clades A1 and A2), a large and weakly supported (BS < 50) group that contains multiple members in all represented taxa, although internal resolution is largely lacking. This group contains *D. melanogaster desat1*, *desat2* and *desatF* (clade A1), which arose from dipteran or *Drosophila* specific gene duplications (the timing of these

events relative to the split between the lineages leading to *Drosophila* and *Anopheles* remains unclear). Another *D. melanogaster* gene, CG8630, falls into clade A2, although its orthologs cannot be identified reliably: with the exception of two *A. pisum* and one *T. castaneum* gene, all genes in clade A2 are more closely related to CG8630 according to the phylogenetic reconstruction, but are deemed orthologous to *desat1* according to the best reciprocal BLAST criterion. This contradiction also applies to the three *P. barbatus* genes in clade A2. The fourth *P. barbatus* gene in clade A, *Pbar_desat1*, can be considered an ortholog of the *D. melanogaster desat1* according to both the phylogenetic and reciprocal BLAST analysis, and therefore makes a prime candidate gene for experimental study regarding its contribution to CHC alkene biosynthesis and recognition cues in *P. barbatus*. Interestingly, all Δ9 desaturase genes – with the exceptions of two members of clade B – are closely linked along a 90 kb region of the *P. barbatus* genome (Fig. S23), whereas the seven genes in *D. melanogaster* are distributed across four regions of chromosome 3.

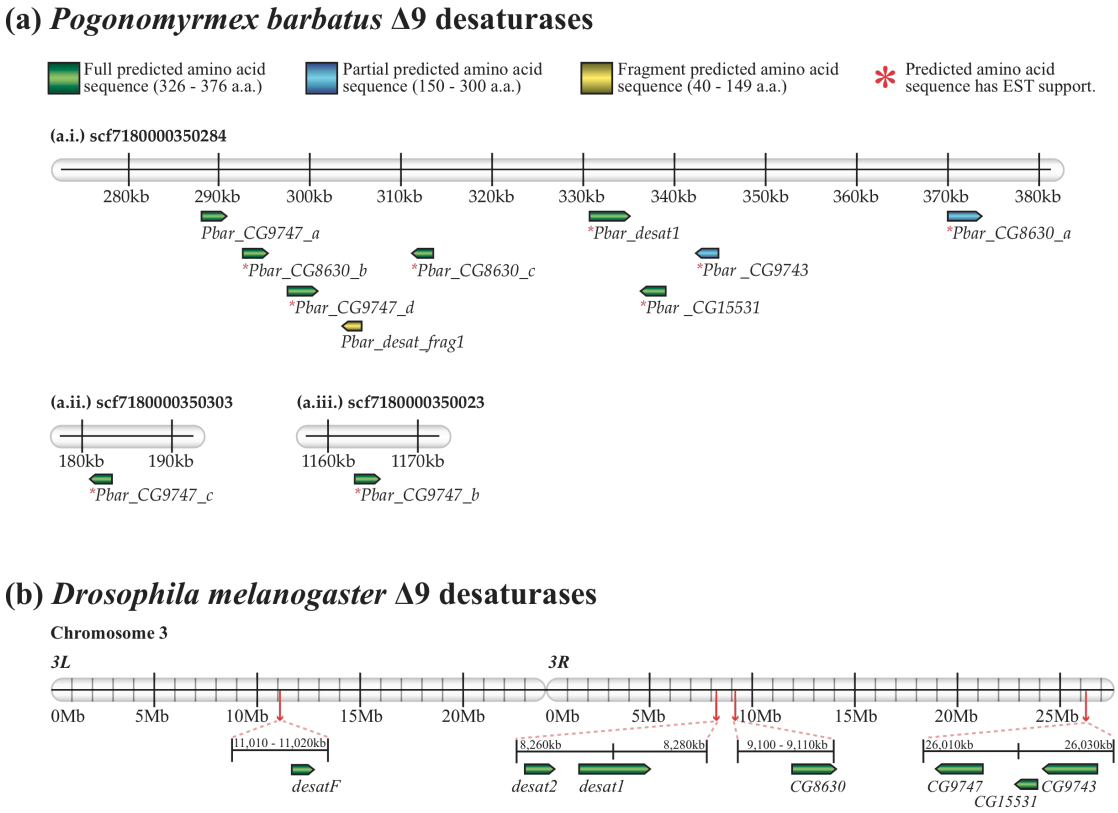


Fig. S23. Arrangement of the ten Δ9 desaturase genes and one fragmentary desaturase gene found in *Pogonomyrmex barbatus*, of which nine are grouped along a ~90 kb region of the genome (a). The seven Δ9 desaturase genes of *Drosophila melanogaster* are situated on one chromosome, but spread out across multiple chromosomal regions (b). Note that both figures are drawn to different scale.

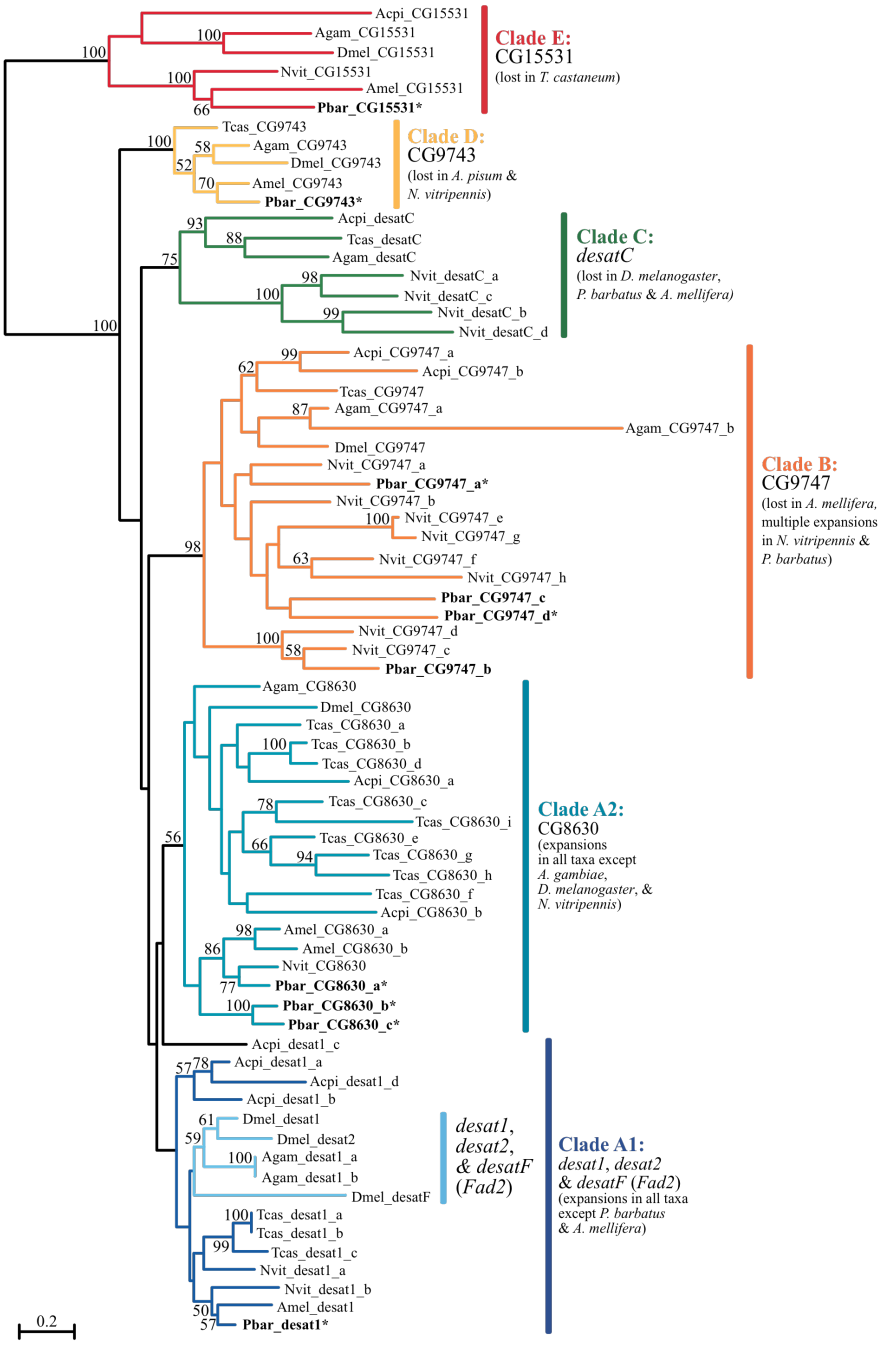


Fig. S24. Unrooted maximum likelihood tree of all Δ9 desaturase genes found in the genomes of *Pogonomyrmex barbatus* (Pbar, in bold), *Nasonia vitripennis* (Nvit), *Apis mellifera* (Amel), *Acyrthosiphon pisum* (Acpi), *Tribolium castaneum* (Tcas), *Drosophila melanogaster* (Dmel), and *Anopheles gambiae* (Agam). Gene labels reflect phylogenetic results based on respectively named *D. melanogaster* genes, except for clade C genes, which were given a new label "desatC" due to the absence of *D. melanogaster* genes in this clade. Support values ≥ 50 based on 500 rapid bootstrap replicates are shown at the nodes of the tree. Asterisks indicate *P. barbatus* genes closely linked together (see Fig. S23).

26. Olfactory Learning and Memory

The ability to gain and process information about the environment (learning) and the ability to store and retrieve information over time (memory) is widespread through all animal species (232). As insects often rely on olfaction for a variety of behaviors (such as mating, foraging, and predator avoidance), they have become a widely used model for studying olfactory learning and memory (233). *Drosophila melanogaster* and *Apis mellifera* have been particularly well studied in the last four decades. A variety of gene types have been implicated in learning and memory function including those coding for cAMP signaling cascade molecules (234, 235), CaMKII proteins (236–238), cell adhesion molecules (239, 240), RNA transport and translation molecules (241), and neurotransmitters like Dopa decarboxylase and tyramine beta-hydroxylase (242–244), to name a few.

Of the 69 learning and memory genes found in *D. melanogaster* that we investigated, 59 genes – 35 involved in olfactory learning and 23 memory function genes – were found and manually annotated in *Pogonomyrmex barbatus*. Six of these genes we identified in the Red harvester ant were not found in *Nasonia vitripennis* or *A. mellifera*. As learning and memory genes are generally discovered using behavioral assays of *D. melanogaster* mutants, the potential number of *P. barbatus* specific learning and memory genes remains to be elucidated. This data set will provide a rich resource for further studies in learning and memory within ant species.

27. Opsins and Circadian Genes

As diurnal foragers, red harvester ants (*Pogonomyrmex barbatus*) rely on their vision for foraging, evasion of parasites and predators, territorial interactions, and myriad other everyday activities. Unlike the honey bee (*Apis mellifera*), however, *P. barbatus* is unlikely to require high-resolution color vision to localize floral resources or for flight navigation, except by the reproductive castes during short nuptial flights (245).

Analysis of the *P. barbatus* genome reveals that they possess a complement of opsin genes similar to *A. mellifera*. *P. barbatus* possesses apparent orthologs of the *ultraviolet opsin*, *blue opsin*, *long wavelength opsin 1* and *long wavelength opsin 2*. In addition, the harvester ant genome contains a vertebrate-like, non-visual *pteropsin* (as does *A. mellifera*) that possibly plays a role in the regulation of circadian cycles (246).

P. barbatus also exhibits a strong diurnal activity pattern, foraging almost exclusively in the heat and light of the daylight hours (247, 248). This daily cycle is likely regulated by a combination of light and thermal cues, and endogenous hormonal and physiological cycles (142). The genetic architecture underlying insect circadian cycles is well-studied (249), but also an active arena of ongoing research.

Our analysis of a subset of the circadian gene network in *P. barbatus* revealed a complement of genes similar that reported in *Nasonia vitripennis* (27) and *A. mellifera* (26). The *P. barbatus* genome contains copies of the *cryptochrome 2* (*cry2*) gene, as well as *cycle*, *timeless*, *clock*, and *period.* Like *N. vitripennis* and *A. mellifera*, *P. barbatus* does not possess the *cryptochrome 1* gene, which is the sole cryptochrome in *D. melanogaster*, and which is present (along with *cry2*) in butterflies and mosquitoes (250).

28. Behavior and Aggression Genes

The widespread occurrence of aggression across the animal kingdom underscores the importance of this behavior in defending and obtaining resources necessary for survival and reproduction. In social insects, aggressive behavior plays a large role in colony defense and exclusion of alien individuals from access to colony resources. For ants, aggression is particularly well studied in the context of nestmate recognition (251). Aggressive behavior is a complex phenotype involving the direct action and regulation of several genes. Most genetic studies on aggression have focused on genes that control or are involved in neurological pathways including bioamines (see previous chapter), substances that have been shown to have clear effects on aggression in both mammals and invertebrates (252, 253). More recent studies on *Drosophila melanogaster*, however, have indicated that other genes that carry out basic biological and molecular functions also play a role in aggression (254–256). Such genes include but are not limited to those involved in cell communication, electron transport, and metabolic processes. Although these genes have been discovered in the context of intraspecific male aggression in *D. melanogaster*, similar gene categories have also been implicated in the honey bee colony defense (257). Additionally, some of the genes associated with aggressive behavior in *D. melanogaster* appear evolutionary conserved with orthologs found in humans (258).

Here, we identified six genes in *Pogonomyrmex barbatus* that are similar to those involved in interspecific male aggression in *D. melanogaster* including *ade5*, *eclair*, *echinoid*, *Laminin A*, *no occelli*, and *sugarless*. The ontology of these genes varies from intracellular protein transport and signaling pathways to central nervous system development. Interestingly, all of these genes have been implicated as having pleiotropic effects on other male *D. melanogaster* phenotypes such as number of sensory bristles, sleep, and starvation stress resistance (256). EST evidence was found for the genes *eclair* and *Laminin A* examined in *P. barbatus*. All six of the aggression genes examined in *P. barbatus* had significant matches to *A. mellifera* indicating possible orthology while only three gene similarities were found for *Nasonia vitripennis*. Though these six genes examined in *P. barbatus* represent few of many genes involved in aggression in *D. melanogaster*, the findings here provide a basis for testing whether these potentially homologous genes also affect aggression in *P. barbatus*, particularly with regard to nestmate recognition (226, 259).

Additionally, using KEGG analysis (260) we identified 726 genes involved in 30 different human disease pathways ranging from Alzheimer's disease to cancer. We annotated 17 genes involved in the human social interaction disease Williams-Beuren Syndrome (WBS, reviewed in (261). Hemizygous deletion of ~28 genes in a 1.5 Mb region result in hypergregarious social behavior in WBS-affected humans along with other physical and neurological phenotypes. While the mechanism of how gene dosage results in altered social interaction are poorly understood, the annotation orthologs for 61% of the WBS genes in *P. barbatus* offers a

new model system to test the genetic component of complex group behaviors for this and other human diseases.

29. Earlham College Evolutionary Genomics Class Annotation

Undergraduate students manually annotated 60 genes as part of an upper level course, Evolutionary Genomics, at Earlham College. The genes annotated were student-chosen, but independent of those annotated as parts of larger pathways or functional groups. Genes ranged in function from involvement in the cell cycle and cell structure, to some implicated in caste determination (e.g., hexamerins, see chapter 19). The workflow of these annotations involved a quality control step where students turned in an assignment detailing the evidence used in making changes to the MAKER gene predictions. Evidence included alignments to genes in well-curated genomes and related species (e.g., *Drosophila melanogaster*, *Apis mellifera*, and *Nasonia vitripennis*) as well as the presence of conserved functional domains. After passing quality control the students uploaded their annotations to the Apollo genome server. Student annotations followed the basic process of manual annotation described in the Supplementary Information (above).

List of students in the class that annotated genes: H. Albers, M. Bahnick, T. Carter, K. Clay, P. Hallowell, J. Hood, S. McGuire, A. Miller, M. Naughton, K. O'Rourke-Owens, K. Paine, J. Pillow, P. Raines, and C. Wertman

30. SNP Analysis

Since the genomic reads used for the *Pogonomyrmex barbatus* assembly were derived from multiple males, it was possible to identify single nucleotide polymorphisms (SNPs) from the natural genetic diversity captured in the raw reads. Although such diversity may only give insight into genes that vary amongst males, this might be informative to identify gene classes that vary within the species in general.

We identified 241,067 total SNPs (59,842 A, 59,504 C, 58,084 T, 59,374 G, 4263 N), 230,279 (96%) of which were present in at least 10% of the overlapping reads. The 3,870 N SNPs (0.3%) and might indicate regions undergoing rapid evolution. Overall, 4.4% of SNPs were found in exons, while 8% were intronic (87.5% intergenic). We also manually investigated the genes with the highest number of SNPs. Further analyses will be required to ascertain SNPs associated with regulatory regions, transposable elements, and other genomic features.

Since we identified all three major DNA methyltransferase protein families, we looked specifically for SNP signatures potentially associated with this process. It is well documented that methylated cytosine residues spontaneously mutate into thymine bases. For example, genomic imprinting can occur when distinct males differentially methylate a specific position. We observed 17,044 cases (937 exons, 1087 introns, seven both intronic and exonic, 15,013 intergenic) where a CG < > TG mutation occurred, compared to only 1316 CG < > AG or 2047 CG < > GG mutations. In total, 7247 (44%) of genes had at least one site with a CG-TG polymorphism. Mutation rates for other SNPs in the context of dinucleotide pairs were comparable. Thus, mutations associated with sites of possible biological methylation were present at least eight times as often as at other non-CG sites, suggesting a bias in this specific mutation that could be related to DNA methylation. Amongst other genes, the 'Major facilitator superfamily MFS-1 (IPR011701)' protein had a large number of SNPs (>10). These membrane transporters are involved in multidrug resistance and sugar transport and could represent a rapidly evolving class of genes. Strikingly, male sterility proteins also ranked high in genes with SNPs along with RING/U-box zinc finger transcription factors, NAD binding proteins, and several cell adhesion proteins implicated in neuronal development. While experimental validation is required to verify methylation differences in these targets genes, they all represent classes that could be imprinted by males to affect success of their patriline.

Methods

We used the Roche gsMapper tool and custom Perl scripts to identify SNPs in the v03 Celera assembly of the *P. barbatus* genome from unpaired and paired-end 454 reads. We intentionally only evaluated cases where a single nucleotide in one read has another single base transition or transversion mutation and omitted cases of insertions and deletions. Custom Perl scripts were used to extract SNPs that were present in 10% or

more of the reads and to identify C > T and T > C SNPs followed by a G. All SNP data was converted to GFF3 that was then loaded into the Chado database to determine intersections with InterProScan and other results.

31. Evolutionary Rates Analysis

To analyze the rate of evolution of *Pogonomyrmex barbatus*, we compared the amino acid composition of *P. barbatus* proteins to orthologous proteins in related insects: *Apis mellifera*, *Nasonia vitripennis*, and *Drosophila melanogaster*. Because of the long separation between these species (>> 100 million years), performing a standard K_a/K_s rate analysis would be difficult since K_s is likely greater than 1. Instead we focused primarily on non-synonymous substitutions in highly conserved regions of orthologous proteins.

Orthologous proteins among all species were identified using OrthoMCL (40) to group annotated proteins into putative orthologous sets. Each set was required to have a single gene copy from each species with no species being unrepresented in the set. In total 4774 orthologous sets were identified that met this criteria. Proteins from of each set were then individually aligned to each other using ClustalW (106). These alignments were further processed using Gblocks (177) to extract only conserved blocks from each alignment that were found in all four organisms used. The final alignments were then processed individually to estimate the distribution of amino acid substitution rates for each gene set. We then concatenated the multiple alignments from each orthologous set together to estimate the average substitution rate for the genome as a whole.

The amino acid substitution rate was estimated using the program Proml available in the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html). Proml was provided with a constrained tree topology corresponding to the known phylogenetic relationship of the organisms used along with multiple alignments for each orthologous gene set (both individually and concatenated). Branch lengths for the trees were then allowed to vary to best fit the multiple alignments to the constrained tree topology. The units of branch lengths produced by Proml are in expected amino acid substitutions per alignment site, which provides a simple means to calculate the substitution rates between different nodes of the tree. Each tree was rooted using *D. melanogaster* as the outgroup, which allowed us to calculate the number of amino acid substitutions per site occurring in the remaining three species relative to their last shared common ancestor. The amino acid substitution rate for the concatenated alignment and the distribution of substitution rates among all orthologous sets were calculated in this manner.

References

1. Anderson KE, et al. (2006) Distribution and evolution of genetic caste determination in Pogonomyrmex seed-harvester ants. *Ecology* 87:2171–2184.
2. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
3. Smith CD, et al. (2007) Improved repeat identification and masking in dipterans. *Gene* 389:1–9.
4. Korf I, Yandell M, Bedell J (2003) BLAST (O'Reilly, Associates, Sebastopol, CA)
5. Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31.
6. Stanke M, Tzvetkova A, Morgenstern B (2006) AUGUSTUS at EGASP: Using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* 7 Suppl 1:S11.1–8.
7. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59.
8. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res* 26:1107–1115.
9. Eilbeck K, Moore B, Holt C, Yandell M (2009) Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics 10:67.
10. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M (2005) The sequence ontology: A tool for the unification of genome annotations. Genome Biol 6:R44.
11. Edgar RC, Myers EW (2005) PILER: Identification and classification of genomic repeats. Bioinformatics 21 Suppl 1:i152–8.
12. UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38:D142–8.
13. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank. Nucleic Acids Res 34:D16–20.
14. Parra G, Bradnam K, Korf I (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067.
15. Cantarel BL, et al (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18:188–196.
16. Zdobnov EM, Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848.
17. Altschul SF, et al (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

18. Lewis SE, et al (2002) Apollo: A sequence annotation editor. Genome Biol 3:research 0082.1–0082.14
19. Wall DP, Fraser HB, Hirsh AE (2003) Detecting putative orthologs. Bioinformatics 19:1710–1711.
20. Lander ES, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921.
21. Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17.
22. Li W, Kaneko K (1992) Long-range correlation and partial 1/fα spectrum in a noncoding DNA sequence. Europhyics Letters 17:655–660.
23. Bernaola-Galván P, Roman-Roldán R, Oliver JL (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 53:5181–5189.
24. Elhaik E, Graur D, Josić K (2010) Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm. Nucleic Acid Res in press
25. Elhaik E, Graur D, Josić K (2010) Comparative testing of DNA segmentation algorithms using benchmark simulations. Mol Biol Evol 27:1015–1024.
26. Honey Bee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honey bee Apis mellifera. Nature 443:931–948.
27. Werren JH, et al (2010) Functional and evolutionary insights from the genomes of three parasitoid Nasonia species. Science 327:343–348.
28. Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, et al. (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci USA 107:12168–12173.
29. Sea Urchin Genome Sequencing Consortium (2006) The genome of the sea urchin strongylocentrotus purpuratus. Science 314:941–952.
30. Tribolium Genome Sequencing Consortium (2008) The genome of the model beetle and pest Tribolium castaneum. Nature 452:949–955.
31. Wang Y, et al (2006) Functional CpG methylation system in a social insect. Science 314:645–647.
32. Kronforst MR, Gilley DC, Strassmann JE, Queller DC (2008) DNA methylation is widespread across social hymenoptera. Curr Biol 18:R287-R288.
33. Foret S, Kucharski R, Pittelkow Y, Lockett GA, Maleszka R (2009) Epigenetic regulation of the honey bee transcriptome: Unravelling the nature of methylated genes. BMC Genomics 10:472.
34. Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Review 51:661–703.
35. Sokal RR, Rohlf FJ (1995) Biometry, (Freeman and Company, New York).
36. Elango N, Hunt BG, Goodisman MAD, Yi SV (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. Proc Natl Acad Sci USA 106:11206–11211.

37. Bush CF, Hall RA (2008) Olfactory receptor trafficking to the plasma membrane. Cell Mol Life Sci 65:2289–2295.
38. Boyle EI, et al (2004) GO::TermFinder – open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics 20:3710–3715.
39. Ashburner M, et al (2000) Gene ontology: Tool for the unification of biology. the gene ontology consortium. Nat Genet 25:25–29.
40. Li L, Stoeckert CJ,Jr, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189.
41. Tweedie S, et al (2009) FlyBase: Enhancing drosophila gene ontology annotations. Nucleic Acids Res 37:D555–9.
42. van Dongen S (2000) Graph clustering by flow simulation. PhD thesis Universiteit Utrecht.
43. Brodersen DE, Nissen P (2005) The social life of ribosomal proteins. FEBS J 272:2098–2108.
44. Sengupta J, et al (2004) Identification of the versatile scaffold protein RACK1 on the eukaryotic ribosome by cryo-EM. Nat Struct Mol Biol 11:957–962.
45. Nilsson J, Sengupta J, Frank J, Nissen P (2004) Regulation of eukaryotic translation by the RACK1 protein: A platform for signalling molecules on the ribosome. EMBO Rep 5:1137–1141.
46. Wool IG, Chan YL, Gluck A (1995) Structure and evolution of mammalian ribosomal proteins. Biochem Cell Biol 73:933–947.
47. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.
48. Helmkampf M, Bruchhaus I, Hausdorf B (2008). Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the Lophotrochozoa concept. Proc R Soc London B 275:1927–1933.
49. Uechi T, Tanaka T, Kenmochi N (2001) A complete map of the human ribosomal protein genes: Assignment of 80 genes to the cytogenetic map and implications for human disorders. Genomics 72:223–230.
50. Marygold SJ, Roote J, Reuter G, Lambertsson A, Ashburner M, Millburn GH, Harrison PM, Yu Z, Kenmochi N, Kaufman TC, et al. (2007) The ribosomal protein genes and minute loci of Drosophila melanogaster. Genome Biol 8:R216.
51. Zhang Z, Harrison P, Gerstein M (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 12:1466–1482.
52. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007). Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8:R143.

53. Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE (2006). Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biology 6:17.
54. Porcelli D, Barsanti P, Pesole G, Caggese C (2007) The nuclear OXPHOS genes in insecta: A common evolutionary origin, a common cis-regulatory motif, a common destiny for gene duplicates. BMC Evol Biol 7:215.
55. Rand DM, Haney RA, Fry AJ (2004) Cytonuclear coevolution: The genomics of cooperation. Trends Ecol Evol 19:645–653.
56. D'Elia D, et al (2006) The MitoDrome database annotates and compares the OXPHOS nuclear genes of Drosophila melanogaster, Drosophila pseudoobscura and Anopheles gambiae. Mitochondrion 6:252–257.
57. Lemons D, McGinnis W (2006) Genomic evolution of hox gene clusters. Science 313:1918–1922.
58. Hughes CL, Kaufman TC (2002) Hox genes and the evolution of the arthropod body plan. Evol Dev 4:459–499.
59. Dearden PK, et al (2006) Patterns of conservation and change in honey bee developmental genes. Genome Res 16:1376–1384.
60. Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780.
61. Crozier RH, Crozier YC (1993) The mitochondrial genome of the honeybee Apis mellifera: Complete sequence and genome organization. Genetics 133:97–117.
62. Wolschin F, Gadau J (2009) Deciphering proteomic signatures of early diapause in nasonia. PLoS One 4:e6394.
63. Rappsilber J, Ishihama Y, Mann M (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Anal Chem 75:663–670.
64. Wolschin F, Amdam GV (2007) Plasticity and robustness of protein patterns during reversible development in the honey bee (Apis mellifera). Anal Bioanal Chem 389:1095–1100.
65. Geer LY, et al (2004) Open mass spectrometry search algorithm. J Proteome Res 3:958–964.
66. Dani FR, et al (2008) Exploring proteins in Anopheles gambiae male and female antennae through MALDI mass spectrometry profiling. PLoS One 3:e2822.
67. Anholt RR, Williams TI (2010) The soluble proteome of the Drosophila antenna. Chem Senses 35:21–30.
68. Schmidt PJ, Sherbrooke WC, Schmidt JO (1989) The detoxification of ant (Pogonomyrmex) venom by a blood factor in horned lizards (Phyrnosoma). Copeia 603–607.
69. Schmidt JO (1982) Biochemistry of insect venoms. Annu Rev Entomol 27:339–368.
70. Hoffman DR (1993) Allergens in Hymenoptera venom XXIV: The amino acid sequences of imported fire ant venom allergens sol i II, sol i III, and sol i IV. J Allergy Clin Immunol 91:71–78.

71. Kreil G, Haiml L, Suchanek G (1980) Stepwise cleavage of the pro part of promelittin by dipeptidylpeptidase IV. evidence for a new type of precursor--product conversion. Eur J Biochem 111:49–58.
72. de Graaf DC, et al (2010) Insights into the venom composition of the ectoparasitoid wasp Nasonia vitripennis from bioinformatic and proteomic studies. Insect Mol Biol 19 Suppl 1:11–26.
73. Schmidt JO, Blum MS (1978) The biochemical constituents of the venom of the harvester ant, Pogonomyrmex badius. Comp Biochem Physiol C 61C: 239–247.
74. Schmidt JO, Blum MS (1978) Pharmacological and toxicological properties of harvester ant, Pogonomyrmex badius, venom. Toxicon 16:645–651.
75. Taber SW, Cokendolpher JC, Francke OF (1988) Karyological study of north-american Pogonomyrmex (Hymenoptera, Formicidae). Insectes Soc 35:47–60.
76. Robertson HM, Gordon KH (2006) Canonical TTAGG-repeat telomeres and telomerase in the honey bee, Apis mellifera. Genome Res 16:1345–1351.
77. Robertson HM (2009) Simple telomeres in a simple animal: Absence of subtelomeric repeat regions in the placozoan trichoplax adhaerens. Genetics 181:323–325.
78. Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12:1269–1276.
79. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res 27:573–580.
80. Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1:i351–8.
81. Valles SM, Strong CA, Hashimoto Y (2007) A new positive-strand RNA virus with unique genome characteristics from the red imported fire ant, Solenopsis invicta. Virology 365:457–463.
82. Valles SM, et al (2004) A picorna-like virus from the red imported fire ant, Solenopsis invicta: Initial discovery, genome sequence, and characterization. Virology 328:151–157.
83. Genersch E, Aubert M (2010) Emerging and re-emerging viruses of the honey bee (Apis mellifera L.). Vet Res 41:54.
84. Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. (Cambridge University Press, Cambridge, UK)
85. Pannebakker BA, Niehuis O, Hedley A, Gadau J, Shuker DM (2010) The distribution of microsatellites in the Nasonia parasitoid wasp genome. Insect Mol Biol 19 Suppl 1:91–98.
86. Thurston MI, Field D (2005) Msatfinder: Detection and characterization of microsatellites.http://www.genomics.ceh.ac.uk/msatfinder/
87. Su CY, Menuz K, Carlson JR (2009) Olfactory perception: receptors, cells, and circuits. Cell 139:45–59.

88. Touhara K, Vosshall LB (2009) Sensing odorants and pheromones with chemosensory receptors. Annu Rev Physiol 71:307–332.
89. Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB (2007) Two chemosensory receptors together mediate carbon dioxide detection in Drosophila. Nature 445:86–90.
90. Kwon JY, Dahanukar A, Weiss LA, Carlson JR (2007) The molecular basis of CO2 reception in Drosophila. Proc Natl Acad Sci USA 104:3574–3578.
91. Lu T, Qiu YT, Wang G, Kwon JY, Rutzler M, Kwon HW, Pitts RJ, van Loon JJ, Takken W, Carlson JR, Zwiebel LJ (2007) Odor coding in the maxillary palp of the malaria vector mosquito Anopheles gambiae. Curr Biol 17:1533–1544.
92. Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB (2009) Variant ionotropic glutamate receptors as chemosensory receptors in Drosophila. Cell 136:149–162.
93. Robertson HM, Warr CG, Carlson JR (2003) Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. Proc Natl Acad Sci USA 100 Suppl 2:14537–14542.
94. Nozawa M, Nei M (2007) Evolutionary dynamics of olfactory receptor genes in Drosophila species. Proc Natl Acad Sci USA 104:7122–7127.
95. Hill CA, Fox AN, Pitts RJ, Kent LB, Tan PL, Chrystal MA, Cravchik A, Collins FH, Robertson HM, Zwiebel LJ (2002) G protein-coupled receptors in Anopheles gambiae. Science 298:176–178.
96. Bohbot J, Pitts RJ, Kwon HW, Rützler M, Robertson HM, Zwiebel LJ (2007) Molecular characterization of the Aedes aegypti odorant receptor gene family. Insect Mol Biol 16:525–537.
97. Wanner KW, Anderson AR, Trowell SC, Theilmann DA, Robertson HM, Newcomb RD. (2007) Female-biased expression of odourant receptor genes in the adult antennae of the silkworm, Bombyx mori. Insect Mol Biol 16:107–119.
98. Tanaka K, Uda Y, Ono Y, Nakagawa T, Suwa M, Yamaoka R, Touhara K (2009) Highly selective tuning of a silkworm olfactory receptor to a key mulberry leaf volatile. Curr Biol 19:881–890.
99. Smadja C, Shi P, Butlin RK, Robertson HM (2009) Large gene family expansions and adaptive evolution for odorant and gustatory receptors in the pea aphid, Acyrthosiphon pisum. Mol Biol Evol 26:2073–2086.
100. Engsontia P, Sanderson AP, Cobb M, Walden KK, Robertson HM, Brown S (2008) The red flour beetle's large nose: an expanded odorant receptor gene family in Tribolium castaneum. Insect Biochem Mol Biol 38:387–397.
101. Robertson HM, Wanner KW (2006) The chemoreceptor superfamily in the honey bee, Apis mellifera: expansion of the odorant, but not gustatory, receptor family. Genome Res 16:1395–1403.
102. Robertson HM, Gadau J, Wanner KW (2010) The insect chemoreceptor superfamily of the parasitoid jewel wasp Nasonia vitripennis. Insect Mol Biol 19 Suppl 1:121–136.

103. Kelber C, Rössler W, Kleineidam CJ (2010) Phenotypic plasticity in number of glomeruli and sensory innervation of the antennal lobe in leaf-cutting ant workers (A. vollenweideri). Dev Neurobiol 70:222–234.
104. Nakanishi A, Nishino H, Watanabe H, Yokohari F, Nishikawa M (2010) Sex-specific antennal sensory system in the ant Camponotus japonicus: Glomerular organizations of antennal lobes. J Comp Neurol 518:2186–2201.
105. Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
106. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.
107. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504.
108. Wanner KW, Nichols AS, Walden KK, Brockmann A, Luetje CW, Robertson HM (2007) A honey bee odorant receptor for the queen substance 9-oxo-2-decenoic acid. Proc Natl Acad Sci USA 104:14383–14388.
109. Ozaki M, Wada-Katsumata A, Fujikawa K, Iwasaki M, Yokohari F, Satoji Y, Nisimura T, Yamaoka R (2005) Ant nestmate and non-nestmate discrimination by a chemosensory sensillum. Science 309:311–314.
110. Brandt M, van Wilgenburg E, Sulc R, Shea KJ, Tsutsui ND (2009) The scent of supercolonies: the discovery, synthesis and behavioural verification of ant colony recognition cues. BMC Biol 7:71.
111. Bray S, Amrein H (2003) A putative Drosophila pheromone receptor expressed in male-specific taste neurons is required for efficient courtship. Neuron 39:1019–1029.
112. Ebbs ML, Amrein H (2007) Taste and pheromone perception in the fruit fly Drosophila melanogaster. Pflugers Arch 454:735–747.
113. Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R (2010) Ancient Protostome Origin of Chemosensory Ionotropic Glutamate Receptors and the Evolution of Insect Taste and Olfaction. PLoS Genet 6:e1001064.
114. Kent LB, Walden KK, Robertson HM (2008) The gr family of candidate gustatory and olfactory receptors in the yellow-fever mosquito Aedes aegypti. Chem Senses 33:79–93.
115. Hildebrand JG, Shepherd GM (1997) Mechanisms of olfactory discrimination: Converging evidence for common principles across phyla. Annu Rev Neurosci 20:595–631.
116. Nei M, Niimura Y, Nozawa M (2008) The evolution of animal chemosensory receptor gene repertoires: Roles of chance and necessity. Nat Rev Genet 9:951–963.

117. Laissue PP, et al (1999) Three-dimensional reconstruction of the antennal lobe in Drosophila melanogaster. J Comp Neurol 405:543–552.
118. Fishilevich E, Vosshall LB (2005) Genetic and functional subdivision of the Drosophila antennal lobe. Curr Biol 15:1548–1553.
119. Vosshall LB, Stocker RF (2007) Molecular architecture of smell and taste in Drosophila. Annu Rev Neurosci 30:505–533.
120. Galizia CG, Menzel R (2001) The role of glomeruli in the neural representation of odours: Results from optical recording studies. J Insect Physiol 47:115–130.
121. Ghaninia M, Hansson BS, Ignell R (2007) The antennal lobe of the african malaria mosquito, Anopheles gambiae – innervation and three-dimensional reconstruction. Arthropod Struct Dev 36:23–39.
122. Zube C, Kleineidam CJ, Kirschner S, Neef J, Rossler W (2008) Organization of the olfactory pathway and odor processing in the antennal lobe of the ant Camponotus floridanus. J Comp Neurol 506:425–441.
123. Mysore K, et al (2009) Caste and sex specific olfactory glomerular organization and brain architecture in two sympatric ant species Camponotus sericeus and Camponotus compressus (fabricius, 1798). Arthropod Struct Dev 38:485–497.
124. Kuebler LS, Kelber C, Kleineidam CJ (2010) Distinct antennal lobe phenotypes in the leaf-cutting ant (Atta vollenweideri). J Comp Neurol 518:352–365.
125. Kelber C, Rossler W, Roces F, Kleineidam CJ (2009) The antennal lobes of fungus-growing ants (Attini): Neuroanatomical traits and evolutionary trends. Brain Behav Evol 73:273–284.
126. Claudianos C, et al (2006) A deficit of detoxification enzymes: Pesticide sensitivity and environmental response in the honeybee. Insect Mol Biol 15:615–636.
127. Oakeshott JG, et al (2010) Metabolic enzymes associated with xenobiotic and chemosensory responses in Nasonia vitripennis. Insect Mol Biol 19 Suppl 1:147–163.
128. Tijet N, Helvig C, Feyereisen R (2001) The cytochrome P450 gene superfamily in Drosophila melanogaster: Annotation, intron-exon organization and phylogeny. Gene 262:189–198.
129. International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biol 8:e1000313.
130. Rewitz KF, O'Connor MB, Gilbert LI (2007) Molecular evolution of the insect halloween family of cytochrome P450s: Phylogeny, gene organization and functional conservation. Insect Biochem Mol Biol 37:741–753.
131. Li X, Schuler MA, Berenbaum MR (2007) Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annu Rev Entomol 52:231–253.

132. Maibeche-Coisne M, Jacquin-Joly E, Francois MC, Nagnan-Le Meillour P (2002) cDNA cloning of biotransformation enzymes belonging to the cytochrome P450 family in the antennae of the noctuid moth Mamestra brassicae. Insect Mol Biol 11:273–281.
133. Rothenbuhler WC (1964) Behavior genetics of nest cleaning in honey bees. iv. responses of F1 and backcross generations to disease-killed blood. Am Zool 4:111–123.
134. Beattie AJ, Turnbull C, Hough T, Knox RB (1986) Antibiotic production: A possible function of the metapleural glands of ants (hymenoptera: Formicidae). Ann Entomol Soc Am 79:448–450.
135. Evans JD, et al (2006) Immune pathways and defence mechanisms in honey bees Apis mellifera. Insect Mol Biol 15:645–656.
136. Gerardo NM, et al (2010) Immunity and other defenses in pea aphids, Acyrthosiphon pisum. Genome Biol 11:R21.
137. Lee CG, Da Silva CA, Lee JY, Hartl D, Elias JA (2008) Chitin regulation of immune responses: An old molecule with new roles. Curr Opin Immunol 20:684–689.
138. Christophides GK, et al (2002) Immunity-related genes and gene families in Anopheles gambiae. Science 298:159–165.
139. Sackton TB, et al (2007) Dynamic evolution of the innate immune system in Drosophila. Nat Genet 39:1461–1468.
140. Tian C, Gao B, Fang Q, Ye G, Zhu S (2010) Antimicrobial peptide-like genes in Nasonia vitripennis: A genomic perspective. BMC Genomics 11:187.
141. Eddy SR (1998) Profile hidden markov models. Bioinformatics 14:755–763.
142. Hölldobler B, Wilson EO (1990) The ants, (Belknap Press of Harvard University Press, Cambridge, MA).
143. Moreau CS, Bell CD, Vila R, Archibald SB, Pierce NE (2006) Phylogeny of the ants: Diversification in the age of angiosperms. Science 312:101–104.
144. Brady SG, Schultz TR, Fisher BL, Ward PS (2006) Evaluating alternative hypotheses for the early evolution and diversification of ants. Proc Natl Acad Sci USA 103:18172–18177.
145. Abouheif E, Wray GA (2002) Evolution of the gene network underlying wing polyphenism in ants. Science 297:249–252.
146. Khila A, Abouheif E (2008) Reproductive constraint is a developmental mechanism that maintains social harmony in advanced ant societies. Proc Natl Acad Sci USA 105:17884–17889.
147. Johnson RA (2010) Independent colony founding by ergatoid queens in the ant genus Pogonomyrmex: Queen foraging provides an alternative to independent colony founding. Insectes Soc 57:169–176.
148. Suni SS, Gignoux C, Gordon DM (2007) Male parentage in dependent-lineage populations of the harvester ant Pogonomyrmex barbatus. Mol Ecol 16:5149–5155.

149. Smith CR, Schoenick C, Anderson KE, Gadau J, Suarez AV (2007) Potential and realized reproduction by different worker castes in queen-less and queen-right colonies of Pogonomyrmex badius. Insectes Soc 54:260–267.
150. Anderson KE, Linksvayer TA, Smith CR (2008) The causes and consequences of genetic caste determination in ants (Hymenoptera: Formicidae). Myrmecol News 11:119–132.
151. Gibson MC, Perrimon N (2005) Extrusion and death of DPP/BMP-compromised epithelial cells in the developing Drosophila wing. Science 307:1785–1789.
152. Sameshima SY, Miura T, Matsumoto T (2004) Wing disc development during caste differentiation in the ant Pheidole megacephala (Hymenoptera : Formicidae). Evol Dev 6:336–341.
153. Klattenhoff C, et al (2007) Drosophila rasiRNA pathway mutations disrupt embryonic axis specification through activation of an ATR/Chk2 DNA damage response. Dev Cell 12:45–55.
154. Yu J, Pacifico S, Liu G, Finley RL,Jr (2008) DroID: The Drosophila interactions database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics 9:461.
155. Pacifico S, et al (2006) A database and tool, IM browser, for exploring and integrating emerging gene and protein interaction data for Drosophila. BMC Bioinformatics 7:195.
156. 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.
157. Barchuk AR, et al (2007) Molecular determinants of caste differentiation in the highly eusocial honeybee Apis mellifera. BMC Devel Biol 7:70–70.
158. Patel A, et al (2007) The making of a queen: TOR pathway is a key player in diphenic caste development. PLoS One e509.
159. Evans JD, Wheeler DE (1999) Differential gene expression between developing queens and workers in the honey bee, Apis mellifera. Proc Natl Acad Sci USA 96:5575–5580.
160. Hoffman EA, Goodisman MAD (2007) Gene expression and the evolution of phenotypic diversity in social wasps. BMC Biology 5:23–23.
161. Hunt JH, et al (2007) A diapause pathway underlies the gyne phenotype in polistes wasps, revealing an evolutionary route to caste-containing insect societies. Proc Natl Acad Sci USA 104:14020–14025.
162. Hahn DA, Johnson RA, Buck NA, Wheeler DE (2004) Storage protein content as a functional marker for colony-founding strategies: A comparative study within the harvester ant genus Pogonomyrmex. Physiol Biochem Zool 77:100–108.
163. Corona M, Estrada E, Zurita M (1999) Differential expression of mitochondrial genes between queens and workers during caste determination in the honeybee Apis mellifera. J Exp Biol 202:929–938.
164. Schwander T, Cahan SH, Keller L (2007) Characterization and distribution of Pogonomyrmex harvester ant lineages with genetic caste determination. Mol Ecol 16:367–387.

165. Foitzik S, Haberl M, Gadau J, Heinze J (1997) Mating frequency of Leptothorax nylanderi ant queens determined by microsatellite analysis. Insectes Soc 44:219–227.
166. Evans JD (1993) Parentage analyses in ant colonies using simple sequence repeat loci. Mol Ecol 2:393–397.
167. Volny VP, Gordon DM (2002) Characterization of polymorphic microsatellite loci in the red harvester ant, Pogonomyrmex barbatus. Mol Ecol Notes 2:302–303.
168. Helms Cahan S, et al (2002) Extreme genetic differences between queens and workers in hybridizing Pogonomyrmex harvester ants. Proc R Soc B 269:1871–1877.
169. Cahan SH, et al (2004) Loss of phenotypic plasticity generates genotype-caste association in harvester ants. Curr Biol 14:2277–2282.
170. Anderson RC (1945) A study of the factors affecting fertility of lozenge females of Drosophila melanogaster. Genetics 30:280–296.
171. Perrimon N, Mohler D, Engstrom L, Mahowald AP (1986) X-linked female-sterile loci in Drosophila melanogaster. Genetics 113:695–712.
172. Bloch Qazi MC, Heifetz Y, Wolfner MF (2003) The developments between gametogenesis and fertilization: Ovulation and female sperm storage in Drosophila melanogaster. Dev Biol 256:195–211.
173. Drapeau MD, Albert S, Kucharski R, Prusko C, Maleszka R (2006) Evolution of the Yellow/Major royal jelly protein family and the emergence of social behavior in honey bees. Genome Res 16:1385–1394.
174. Makarova KS, et al (2001) Genome of the extremely radiation-resistant bacterium deinococcus radiodurans viewed from the perspective of comparative genomics. Microbiol Mol Biol Rev 65:44–79.
175. Arakane Y, et al (2010) Identification, mRNA expression and functional analysis of several yellow family genes in Tribolium castaneum. Insect Biochem Mol Biol 40:259–266.
176. 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.
177. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552.
178. Abascal F, Zardoya R, Posada D (2005) ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.
179. Drapeau MD (2001) The family of yellow-related Drosophila melanogaster proteins. Biochem Biophys Res Commun 281:611–613.
180. Le Crom S, Kapsimali M, Barome PO, Vernier P (2003) Dopamine receptors for every species: Gene duplications and functional diversification in craniates. J Struct Funct Genomics 3:161–176.
181. Walker RJ, Brooks HL, Holden-Dye L (1996) Evolution and overview of classical transmitter molecules and their receptors. Parasitology 113 Suppl: S3–33.

182. Peroutka SJ (1994) 5-hydroxytryptamine receptors in vertebrates and invertebrates: Why are there so many? Neurochem Int 25:533–536.
183. Hammer M (1997) The neural basis of associative reward learning in honeybees. Trends Neurosci 20:245–252.
184. Schwaerzel M, et al (2003) Dopamine and octopamine differentiate between aversive and appetitive olfactory memories in Drosophila. J Neurosci 23:10495–10502.
185. Bicker G, Menzel R (1989) Chemical codes for the control of behaviour in arthropods. Nature 337:33–39.
186. Kloppenburg P, Mercer AR (2008) Serotonin modulation of moth central olfactory neurons. Annu Rev Entomol 53:179–190.
187. Scheiner R, Pluckhahn S, Oney B, Blenau W, Erber J (2002) Behavioural pharmacology of octopamine, tyramine and dopamine in honey bees. Behav Brain Res 136:545–553.
188. Fussnecker BL, Smith BH, Mustard JA (2006) Octopamine and tyramine influence the behavioral profile of locomotor activity in the honey bee (Apis mellifera). J Insect Physiol 52:1083–1092.
189. Mustard JA, Pham PM, Smith BH (2010) Modulation of motor behavior by dopamine and the D1-like dopamine receptor AmDOP2 in the honey bee. J Insect Physiol 56:422–430.
190. Libersat F, Pflueger HJ (2004) Monoamines and the orchestration of behavior. Bioscience 54:17–25.
191. Schulz D.J., Robinson GE (1999) Biogenic amines and division of labor in honey bee colonies: Behaviorally related changes in the antennal lobes and age-related changes in the mushroom bodies. J Comp Physiol A 184:481–488.
192. Beggs KT, et al (2007) Queen pheromone modulates brain dopamine function in worker honey bees. Proc Natl Acad Sci USA 104:2460–2464.
193. Vander Meer RK, Preston CA, Hefetz A (2008) Queen regulates biogenic amine level and nestmate recognition in workers of the fire ant, Solenopsis invicta. Naturwissenschaften 95:1155–1158.
194. Bloch G, Hefetz A, Hartfelder K (2000) Ecdysteroid titer, ovary status, and dominance in adult worker and queen bumble bees (Bombus terrestris). J Insect Physiol 46:1033–1040.
195. Cuvillier-Hot V, Lenoir A (2006) Biogenic amine levels, reproduction and social dominance in the queenless ant Streblognathus peetersi. Naturwissenschaften 93:149–153.
196. Boulay R, Soroker V, Godzinska EJ, Hefetz A, Lenoir A (2000) Octopamine reverses the isolation-induced increase in trophallaxis in the carpenter ant Camponotus fellah. J Exp Biol 203:513–520.
197. Lu H, Vinson SB, Pietrantonio PV (2009) Oocyte membrane localization of vitellogenin receptor coincides with queen flying age, and receptor silencing by RNAi disrupts egg formation in fire ant virgin queens. FEBS Journal 276:3110–3123.
198. Quevillon E, et al (2005) InterProScan: Protein domains identifier. Nucleic Acids Res 33:W116–20.

199. Marques JT, et al (2010) Loqs and R2D2 act sequentially in the siRNA pathway in drosophila. Nat Struct Mol Biol 17:24–30.
200. Castanotto D, Lingeman R, Riggs AD, Rossi JJ (2009) CRM1 mediates nuclear-cytoplasmic shuttling of mature microRNAs. Proc Natl Acad Sci USA 106:21655–21659.
201. Liu Y, et al (2009) C3PO, an endoribonuclease that promotes RNAi by facilitating RISC activation. Science 325:750–753.
202. Aravin AA, Hannon GJ, Brennecke J (2007) The piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 318:761–764.
203. Kennerdell JR, Yamaguchi S, Carthew RW (2002) RNAi is activated during drosophila oocyte maturation in a manner dependent on aubergine and spindle-E. Genes Dev 16:1884–1889.
204. Vagin VV, et al (2004) The RNA interference proteins and vasa locus are involved in the silencing of retrotransposons in the female germline of Drosophila melanogaster. RNA Biol 1:54–58.
205. May RC, Plasterk RH (2005) RNA interference spreading in C. elegans. Methods Enzymol 392:308–315.
206. Lipardi C, Paterson BM (2009) Identification of an RNA-dependent RNA polymerase in Drosophila involved in RNAi and transposon suppression. Proc Natl Acad Sci USA 106:15645–15650.
207. Caudy AA, Myers M, Hannon GJ, Hammond SM (2002) Fragile X-related protein and VIG associate with the RNA interference machinery. Genes Dev 16:2491–2496.
208. Ulvila J, et al (2006) Double-stranded RNA is internalized by scavenger receptor-mediated endocytosis in drosophila S2 cells. J Biol Chem 281:14370–14375.
209. Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Res 32:D109–11.
210. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: MicroRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–4.
211. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: Tools for microRNA genomics. Nucleic Acids Res 36:D154–8.
212. Hofacker IL, et al (1994) Fast folding and comparison of RNA secondary structures. Monatsh Chem 125:167–188.
213. Tran T, Havlak P, Miller J (2006) MicroRNA enrichment among short 'ultraconserved' sequences in insects. Nucleic Acids Res 34:e65.
214. Weaver DB, et al (2007) Computational and transcriptional evidence for microRNAs in the honey bee genome. Genome Biol 8:R97.
215. Razin A, Shemer R (1995) DNA methylation in early development. Human Mol Genet 4:1751–1755.
216. Field LM, Lyko F, Mandrioli M, Prantera G (2004) DNA methylation in insects. Insect Molecular Biology 13:109–115.
217. Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. Ann Rev Biochem 74:481–514.

218. Jaenisch R, Bird A (2003) Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. Nat Genet Supp 33:245–254.
219. Okamoto I, Otte AP, Allis CD, Reinberg D, Heard E (2004) Epigenetic dynamics of imprinted X inactivation during early mouse development. Science 303:644–649.
220. Phalke S, et al (2009) Retrotransposon silencing and telomere integrity in somatic cells of drosophila depends on the cytosine-5 methyltransferase DNMT2. Nat Genet 41:696–702.
221. Kucharski R, Maleszka J, Foret S, Maleszka R (2008) Nutritional control of reproductive status in honeybees via DNA methylation. Science 319:1827–1830.
222. Hendrich B, Tweedie S (2003) The methyl-CpG binding domain and the evolving role of DNA methylation in animals. Trends Genet 19:269–277.
223. Payne CM, Tillberg CV, Suarez AV (2004) Recognition systems and biological invasions. Ann Zool Fenn 43:843–858.
224. Hefetz A, Errard C, Chambris A, Lenegrate A (1996) Postpharyngeal gland secretion as a modifier of aggresive behavior in the myrmicine ant Manica rubida. J Insect Behav 9:709–717.
225. Lahav S, Soroker V, Hefetz A, Vander Meer RK (1999) Direct behavioral evidence for hydrocarbons as ant recognition discriminators. Naturwissenschaften 86:246–249.
226. Wagner D, Tissot M, Cuevas W, Gordon DM (2000) Harvester ants utilize cuticular hydrocarbons in nestmate recognition. J Chem Ecol 26:2245–2257.
227. Carlin NF, Hölldobler B (1987) The kin recognition system of carpenter ants (Camponotus spp.): II. larger colonies. Behav Ecol Sociobiol 20:209–217.
228. Martin S, Drijfhout F (2009) A review of ant cuticular hydrocarbons. J Chem Ecol 35:1151–1161.
229. Gleason JM, Jallon JM, Rouault JD, Ritchie MG (2005) Quantitative trait loci for cuticular hydrocarbons associated with sexual isolation between Drosophila simulans and D. sechellia. Genetics 171:1789–1798.
230. Dallerac R, et al (2000) A delta 9 desaturase gene with a different substrate specificity is responsible for the cuticular diene hydrocarbon polymorphism in Drosophila melanogaster. Proc Natl Acad Sci U S A 97:9449–9454.
231. Legendre A, Miao XX, Da Lage JL, Wicker-Thomas C (2008) Evolution of a desaturase involved in female pheromonal cuticular hydrocarbon biosynthesis and courtship behavior in Drosophila. Insect Biochem Mol Biol 38:244–255.
232. Skoulakis EM, Davis RL (1996) Olfactory learning deficits in mutants for leonardo, a Drosophila gene encoding a 14–3-3 protein. Neuron 17:931–944.
233. Liu X, Davis RL (2006) Insect olfactory memory in time and space. Curr Opin Neurobiol 16:679–685.
234. Dudai Y, Jan YN, Byers D, Quinn WG, Benzer S (1976) Dunce, a mutant of Drosophila deficient in learning. Proc Natl Acad Sci USA 73:1684–1688.

235. Aceves-Pina EO, et al (1983) Learning and memory in drosophila, studied with mutants. Cold Spring Harb Symp Quant Biol 48 Pt 2:831–840.
236. Mehren JE, Griffith LC (2004) Calcium-independent calcium/calmodulin-dependent protein kinase II in the adult Drosophila CNS enhances the training of pheromonal cues. J Neurosci 24:10584–10593.
237. Joiner M, Griffith LC (1997) CaM kinase II and visual input modulate memory formation in the neuronal circuit controlling courtship conditioning. J Neurosci 17:9384–9391.
238. Griffith LC, et al (1993) Inhibition of calcium/calmodulin-dependent protein kinase in Drosophila disrupts behavioral plasticity. Neuron 10:501–509.
239. Grotewiel MS, Beck CD, Wu KH, Zhu XR, Davis RL (1998) Integrin-mediated short-term memory in drosophila. Nature 391:455–460.
240. Cheng Y, et al (2001) Drosophila fasciclinII is required for the formation of odor memories and for normal sensitivity to alcohol. Cell 105:757–768.
241. Dubnau J, et al (2003) The staufen/pumilio pathway is involved in Drosophila long-term memory. Curr Biol 13:286–296.
242. Waddell S, Quinn WG (2001) Flies, genes, and learning. Annu Rev Neurosci 24:1283–1309.
243. Schwaerzel M, et al (2003) Dopamine and octopamine differentiate between aversive and appetitive olfactory memories in Drosophila. J Neurosci 23:10495–10502.
244. Tempel BL, Livingstone MS, Quinn WG (1984) Mutations in the dopa decarboxylase gene affect learning in Drosophila. Proc Natl Acad Sci U S A 81:3577–3581.
245. Hölldobler B (1976) The behavioral ecology of mating in harvester ants (hymenoptera: Formicidae: Pogonomyrmex). Behav Ecol Sociobiol 1:405–423.
246. Velarde RA, Sauer CD, Walden KK, Fahrbach SE, Robertson HM (2005) Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. Insect Biochem Mol Biol 35:1367–1377.
247. Johnson RA (2000) Seed-harvester ante (Hymenoptera : Formicidae) of North America: An overview of ecology and biogeography. Sociobiology 36:89–122.
248. Gordon DM (1999) Ants at work, (The Free Press, New York, NY).
249. Giebultowicz JM (2000) Molecular mechanism and cellular distribution of insect circadian clocks. Annu Rev Entomol 45:769–793.
250. Zhu H, et al (2005) The two CRYs of the butterfly. Curr Biol 15:R953–4.
251. Kravitz EA, Huber R (2003) Aggression in invertebrates. Curr Opin Neurobiol 13:736–743.
252. Edwards AC, Mackay TF (2009) Quantitative trait loci for aggressive behavior in Drosophila melanogaster. Genetics 182:889–897.
253. Nelson RJ ed. (2006) Biology of aggression (Oxford University Press, New York, NY).
254. Dierick HA, Greenspan RJ (2006) Molecular analysis of flies selected for aggressive behavior. Nat Genet 38:1023–1031.

255. Edwards AC, et al (2009) A transcriptional network associated with natural variation in Drosophila aggressive behavior. Genome Biol 10:R76.
256. Edwards AC, Zwarts L, Yamamoto A, Callaerts P, Mackay TF (2009) Mutations in many genes affect aggressive behavior in Drosophila melanogaster. BMC Biol 7:29.
257. Alaux C, et al (2009) Honey bee aggression supports a link between gene regulation and behavioral evolution. Proc Natl Acad Sci USA 106:15400–15405.
258. Edwards AC, Rollmann SM, Morgan TJ, Mackay TF (2006) Quantitative genomics of aggressive behavior in Drosophila melanogaster. PLoS Genet 2:e154.
259. Adler FR, Gordon DM (2003) Optimization, conflict, and nonoverlapping foraging ranges in ants. Am Nat 162:529–543.
260. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acid Res 35:W182–185.
261. Francke U (1999) Williams-Beuren syndrome: genes and mechanisms. Hum Mol Genet 8:1947–1954.

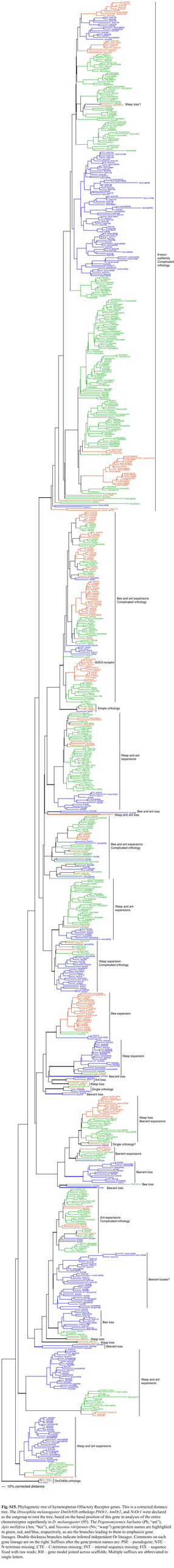


Fig. S15. Phylogenetic tree of hymenopteran Olfactory Receptor genes. This is a corrected distance tree. The *Drosophila melanogaster DmOr83b* orthologs *PbOr1*, *AmOr2*, and *NvOr1* were declared as the outgroup to root the tree, based on the basal position of this gene in analyses of the entire chemoreceptor superfamily in *D. melanogaster* (95). The *Pogonomyrmex barbatus* (Pb, "ant"), *Apis mellifera* (Am, "bee"), and *Nasonia vitripennis* (Nv, "wasp") gene/protein names are highlighted in green, red, and blue, respectively, as are the branches leading to them to emphasize gene lineages. Double thickness branches indicate inferred independent Or lineages. Comments on each gene lineage are on the right. Suffixes after the gene/protein names are: PSE – pseudogene; NTE – N-terminus missing; CTE – C-terminus missing; INT – internal sequence missing; FIX – sequence fixed with raw reads; JOI – gene model joined across scaffolds; Multiple suffixes are abbreviated to single letters.