

Draft genome of the red harvester ant *Pogonomyrmex barbatus*

Chris R. Smith^{a,1}, Christopher D. Smith^{b,1}, Hugh M. Robertson^c, Martin Helmkamp^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^{ij}, Eran Elhaik^k, Christine G. Elsik^l, Marie-Julie Favé^g, Vilaiwan Fernandes^g, Joshua D. Gibson^d, Dan Graur^m, Wulfila Gronenbergⁿ, Kirk J. Grubbs^j, Darren E. Hagen^l, Ana Sofia Ibarra-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. Placek², Rajendran Rajakumar^g, Justin T. Reese^l, Garret Suen^{ij}, Shu Tao^l, Candice W. Torres^o, Neil D. Tsutsui^o, Lumi Viljakainen^s, Florian Wolschin^{d,t}, and Jürgen Gadau^{d,2}

^aDepartment of Biology, Earlham College, Richmond, IN 47374; ^bDepartment of Biology, San Francisco State University, San Francisco, CA 94132; ^cDepartment of Entomology, University of Illinois Urbana-Champaign, Urbana, IL 61801; ^dSchool of Life Sciences, Arizona State University, Tempe, AZ 85287; ^eInstitute for Physical Science and Technology, University of Maryland, College Park, MD 20742; ^fDepartment of Human Genetics, University of Utah, Salt Lake City, UT 84112; ^gDepartment of Biology, McGill University, Montreal, Quebec H3A 1B1, Canada; ^hCenter for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland; ⁱDOE Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, WI 53706; ^jDepartment of Bacteriology, University of Wisconsin, Madison, WI 53706; ^kThe Johns Hopkins University School of Medicine, Baltimore, MD 21205; ^lDepartment of Biology, Georgetown University, Washington, DC 20057; ^mDepartment of Biology and Biochemistry, University of Houston, Houston, TX 77204; ⁿDepartment of Neuroscience, University of Arizona, Tucson, AZ 85721; ^oDepartment of Environmental Science, Policy and Management, University of California, Berkeley, CA 94720; ^pDepartment of Entomology, University of Nebraska, Lincoln, NE 68583; ^qDepartment of Computer Science, San Francisco State University, San Francisco, CA 94132; ^rCenter for Molecular Biodiversity, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany; ^sDepartment of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853; and ^tDepartment 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 po-

tential (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.

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

²To 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 \times 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

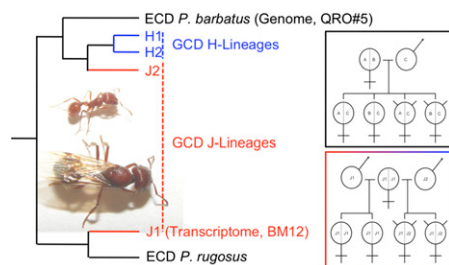


Fig. 1. A pictorial description of the phylogenetic position of the samples used for the genome and transcriptome sequencing, with each put in the context of environmental and genetic caste determination (for a more complete phylogenetic tree, see *SI Appendix, Chapter 1*). The dependent lineages (H1/H2 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.

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.

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 [$\text{CpG}(o/e)$] 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
<i>Pogonomyrmex barbatus</i>	Hymenoptera (red harvester ant)	12	793	17,177	OGS1.1	This study
<i>Nasonia vitripennis</i>	Hymenoptera (jewel wasp)	6.8	709	18,822	OGS1.2	(6)
<i>Apis mellifera</i>	Hymenoptera (honey bee)	7.5	362	10,156/21,001	OGS1/OGS2	(5)
<i>Acyrtosiphon pisum</i>	Sternorrhyncha (pea aphid)	6.2	88.5	34,604	OGS1	(37)
<i>Tribolium castaneum</i>	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 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 retroviral 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.

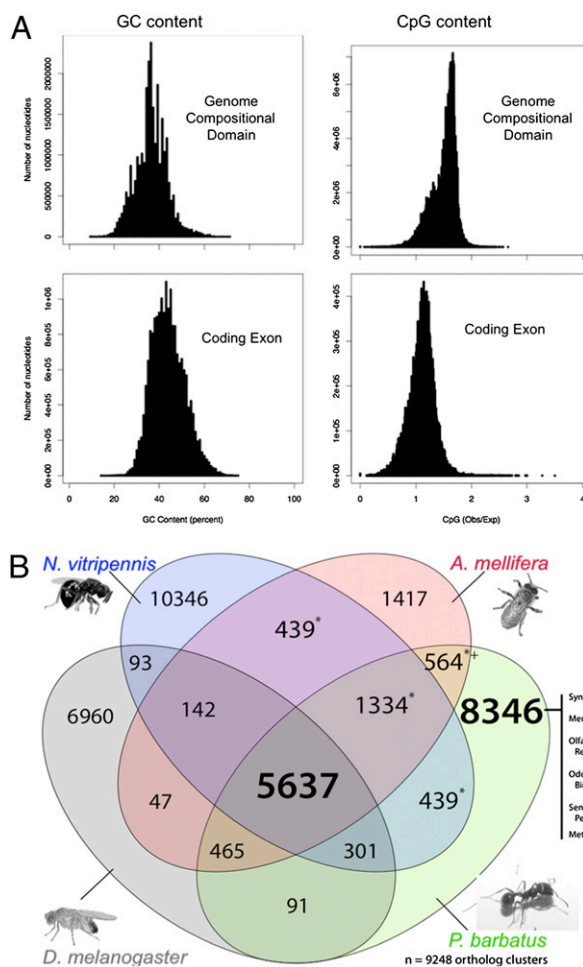


Fig. 2. Genome-wide analyses of nucleotide and relative gene content. (A) Synopsis of GC and CpG(o/e) content of the *P. barbatus* genome. (Upper panels) Comparison of genome regions with the same GC composition. (Lower panels) Comparison of the same features for exons. These distributions are similar 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 $\approx 1:1$ 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 frame-shift 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 *Acyrtosiphon 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

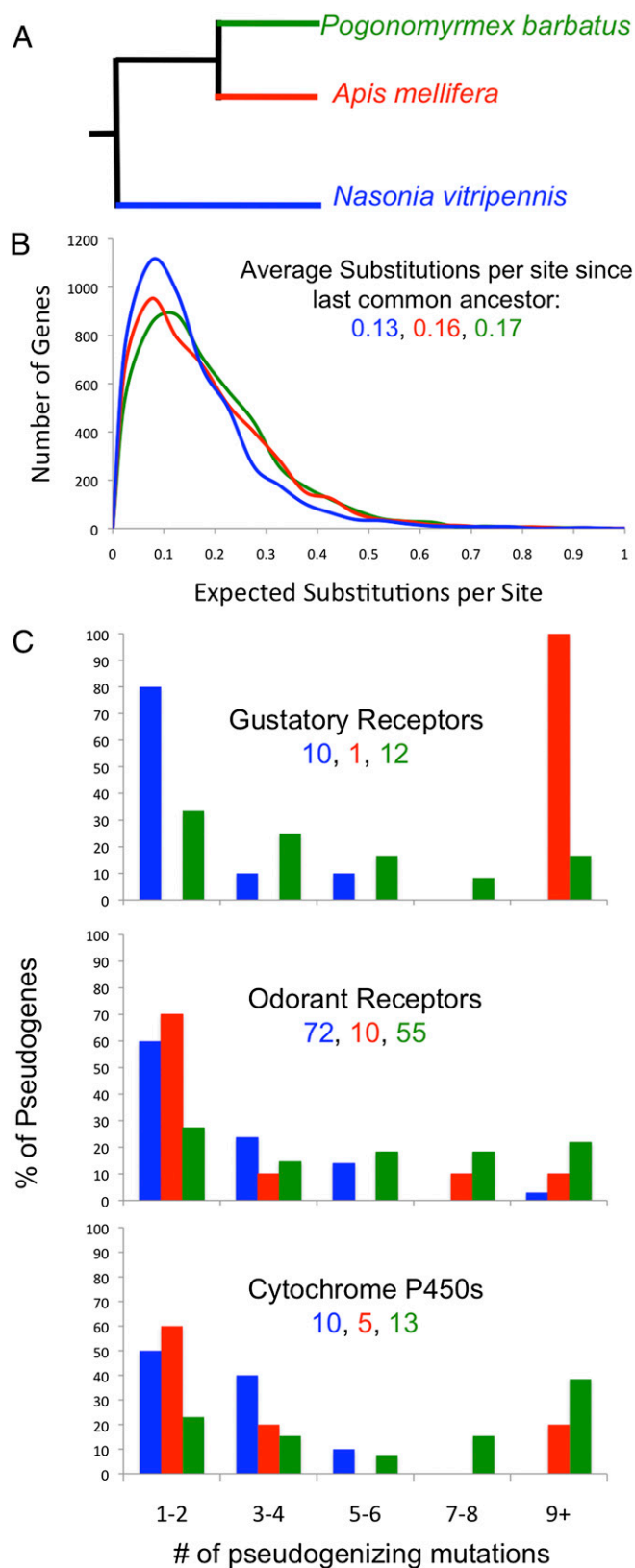


Fig. 3. Evolutionary rate and the accumulation of pseudogene-causing ("pseudogenizing") mutations in three gene families in the ant *P. barbatus* (green), the honey bee *A. mellifera* (red), and the jewel wasp *N. vitripennis* (blue). (A) The relationships among analyzed taxa. (B) A comparison of the 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

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

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: $F_{2,156} = 4.7$, $P = 0.01$).

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.

- Maynard Smith J, Szathmáry E (1995) *The Major Transitions in Evolution* (W. H. Freeman/Spektrum, New York).
- Hölldobler B, Wilson EO (1990) *The Ants* (Belknap Press of Harvard University Press, Cambridge, MA).
- 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.
- Robinson GE, Grozinger CM, Whitfield CW (2005) Sociogenomics: Social life in molecular terms. *Nat Rev Genet* 6:257–270.
- Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949.
- 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.
- 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.
- 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.
- Brady SG, Sipes S, Pearson A, Danforth BN (2006) Recent and simultaneous origins of eusociality in halictid bees. *Proc Biol Sci* 273:1643–1649.
- 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.
- 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.
- Smith CD, Smith CR, Mueller U, Gadau J (2010) Ant genomics: Strength and diversity in numbers. *Mol Ecol* 19:31–35.
- Toth AL, Robinson GE (2007) Evo-devo and the evolution of social behavior. *Trends Genet* 23:334–341.
- Page RE, Jr., Amdam GV (2007) The making of a social insect: Developmental architectures of social design. *Bioessays* 29:334–343.
- Johnson RA (2000) Seed-harvester ante (Hymenoptera: Formicidae) of North America: An overview of ecology and biogeography. *Sociobiology* 36:89–122.
- Taber SW (1998) *The World of the Harvester Ants* (Texas A&M University Press, College Station, TX).
- Gordon DM (1999) *Ants at Work* (The Free Press, New York).
- 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.
- MacMahon JA, Mull JF, Crist TO (2000) Harvester ants (*Pogonomyrmex* spp.): Their community and ecosystem influences. *Annu Rev Ecol Syst* 31:265–291.
- Anderson KE, Linksvayer TA, Smith CR (2008) The causes and consequences of genetic caste determination in ants (Hymenoptera: Formicidae). *Myrmecol News* 11:119–132.
- Helms Cahan S, et al. (2002) Extreme genetic differences between queens and workers in hybridizing *Pogonomyrmex* harvester ants. *Proc Biol Sci* 269:1871–1877.
- 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.
- Volny VP, Gordon DM (2002) Genetic basis for queen-worker dimorphism in a social insect. *Proc Natl Acad Sci USA* 99:6108–6111.
- Cahan SH, et al. (2004) Loss of phenotypic plasticity generates genotype-caste association in harvester ants. *Curr Biol* 14:2277–2282.
- Tsutsui ND, Suarez AV, Spagnac J, Johnston JS (2008) The evolution of genome size in ants. *BMC Evol Biol* 8:64.
- Quevillon E, et al. (2005) InterProScan: Protein domains identifier. *Nucleic Acids Res* 33(Web Server issue):W116–W120.
- Parra G, Bradnam K, Korf I (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061–1067.
- 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.
- Marygold SJ, et al. (2007) The ribosomal protein genes and Minute loci of *Drosophila melanogaster*. *Genome Biol* 8:R216.
- Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 8:R143.
- Saraste M (1999) Oxidative phosphorylation at the fin de siècle. *Science* 283:1488–1493.
- Hughes CL, Kaufman TC (2002) Hox genes and the evolution of the arthropod body plan. *Evol Dev* 4:459–499.
- Gellon G, McGinnis W (1998) Shaping animal body plans in development and evolution by modulation of Hox expression patterns. *Bioessays* 20:116–125.
- Crozier RH, Crozier YC (1993) The mitochondrial genome of the honeybee *Apis mellifera*: Complete sequence and genome organization. *Genetics* 133:97–117.
- Taber SW, Cokendolpher JC, Francke OF (1988) Karyological study of North-American *Pogonomyrmex* (Hymenoptera, Formicidae). *Insectes Soc* 35:47–60.
- Robertson HM, Gordon KH (2006) Canonical TTAGG-repeat telomeres and telomerase in the honey bee, *Apis mellifera*. *Genome Res* 16:1345–1351.
- International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrtosiphon pisum*. *PLoS Biol* 8:e1000313.
- Tribolium* Genome Sequencing Consortium et al. (2008) The genome of the model beetle and pest *Tribolium castaneum*. *Nature* 452:949–955.
- 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.
- Robertson HM (1993) The mariner transposable element is widespread in insects. *Nature* 362:241–245.
- Kojima KK, Fujiwara H (2005) Long-term inheritance of the 28S rDNA-specific retrotransposon R2. *Mol Biol Evol* 22:2157–2165.
- 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.
- 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.
- Benton R, Vannice KS, Gomez-Diaz C, Voshall LB (2009) Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell* 136:149–162.
- 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.
- 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.
- Poinar GO, Jr., Danforth BN (2006) A fossil bee from Early Cretaceous Burmese amber. *Science* 314:614.
- Mombaerts P (1999) Molecular biology of odorant receptors in vertebrates. *Annu Rev Neurosci* 22:487–509.
- 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.
- 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.
- 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.
- Poulsen CJ, Ehlers TA, Insel N (2010) Onset of convective rainfall during gradual late Miocene rise of the central Andes. *Science* 328:490–493.
- 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.
- 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.
- Walker TN, Hughes WO (2009) Adaptive social immunity in leaf-cutting ants. *Biol Lett* 5:446–448.
- 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.
- Davidson EH (2006) The sea urchin genome: Where will it lead us? *Science* 314:939–940.
- Abouheif E, Wray GA (2002) Evolution of the gene network underlying wing polyphenism in ants. *Science* 297:249–252.
- 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.
- Kucharski R, Maleszka J, Foret S, Maleszka R (2008) Nutritional control of reproductive status in honeybees via DNA methylation. *Science* 319:1827–1830.
- 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.
- 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.
- Schwander T, Cahan SH, Keller L (2007) Characterization and distribution of *Pogonomyrmex* harvester ant lineages with genetic caste determination. *Mol Ecol* 16:367–387.
- Anderson RC (1945) A study of the factors affecting fertility of lozenge females of *Drosophila melanogaster*. *Genetics* 30:280–296.
- Perrimon N, Mohler D, Engstrom L, Mahowald AP (1986) X-linked female-sterile loci in *Drosophila melanogaster*. *Genetics* 113:695–712.
- 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.
- 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.
- Miller JR, et al. (2008) Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 24:2818–2824.
- Cantarel BL, et al. (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18:188–196.