SOCIAL EVOLUTION

Genomic signatures of evolutionary transitions from solitary to group living

Karen M. Kapheim,[1,2,3*†] Hailin Pan,[4*] Cai Li,[4,5] Steven L. Salzberg,[6,7] Daniela Puiu,[7] Tanja Magoc,[7] Hugh M. Robertson,[1,2] Matthew E. Hudson,[1,8] Aarti Venkat,[1,8,9] Brielle J. Fischman,[1,10,11] Alvaro Hernandez,[12] Mark Yandell,[13,14] Daniel Ence,[13] Carson Holt,[13,14] George D. Yocum,[15] William P. Kemp,[15] Jordi Bosch,[16] Robert M. Waterhouse,[17,18,19,20] Evgeny M. Zdobnov,[17,18] Eckart Stolle,[21,22] F. Bernhard Kraus,[21,23] Sophie Helbing,[21] Robin F. A. Moritz,[21,24] Karl M. Glastad,[25] Brendan G. Hunt,[26] Michael A. D. Goodisman,[25] Frank Hauser,[27] Cornelis J. P. Grimmelikhuijzen,[27] Daniel Guariz Pinheiro,[28,29] Francis Morais Franco Nunes,[30] Michelle Prioli Miranda Soares,[28] Érica Donato Tanaka,[31] Zilá Luz Paulino Simões,[28] Klaus Hartfelder,[32] Jay D. Evans,[33] Seth M. Barribeau,[34] Reed M. Johnson,[35] Jonathan H. Massey,[2,36] Bruce R. Southey,[37] Martin Hasselmann,[38] Daniel Hamacher,[38] Matthias Biewer,[38] Clement F. Kent,[39,40] Amro Zayed,[39] Charles Blatti III,[1,41] Saurabh Sinha,[1,41] J. Spencer Johnston,[42] Shawn J. Hanrahan,[42] Sarah D. Kocher,[43] Jun Wang,[4,44,45,46,47†] Gene E. Robinson,[1,48†] Guojie Zhang[4,49†]

The evolution of eusociality is one of the major transitions in evolution, but the underlying genomic changes are unknown. We compared the genomes of 10 bee species that vary in social complexity, representing multiple independent transitions in social evolution, and report three major findings. First, many important genes show evidence of neutral evolution as a consequence of relaxed selection with increasing social complexity. Second, there is no single road map to eusociality; independent evolutionary transitions in sociality have independent genetic underpinnings. Third, though clearly independent in detail, these transitions do have similar general features, including an increase in constrained protein evolution accompanied by increases in the potential for gene regulation and decreases in diversity and abundance of transposable elements. Eusociality may arise through different mechanisms each time, but would likely always involve an increase in the complexity of gene networks.

The evolution of eusociality involves changes in the unit of natural selection, from the individual to a group (*1*). Bees evolved eusociality multiple times and are extremely socially diverse (*2*) (Fig. 1), but all pollinate angiosperms, including many crops essential to the human diet (*3*). Simple eusociality may be facultative or obligate, and both forms are characterized by small colonies with a reproductive queen and one or more workers that, due to social and nutritional cues, forego reproduction to cooperatively care for their siblings (*2*). Further evolutionary elaborations have led to complex eusociality, "superorganisms" with colonies of several thousand individuals, sophisticated modes of communication, and morphological specializations for division of labor (*4*).

Theory predicts that the evolution of simple eusociality involves increased regulatory flexibility of ancestral gene networks to create specialized reproductive and nonreproductive individuals, and the evolution of complex eusociality requires genetic novelty to coordinate emergent properties of group dynamics (*5*). To test these predictions, we analyzed five de novo and five publicly available draft genome sequences of 10 bee species from three families, representing two independent origins of eusociality in Apidae and Halictidae and two independent elaborations of simple to complex eusociality in two apid tribes [Apini (honeybees) and Meliponini (stingless bees); Fig. 1]. The draft genomes were of comparable, high quality (supplementary materials).

We found that the transition from solitary to group life is associated with an increased capacity for gene regulation. We scanned the promoter regions of 5865 single-copy orthologs among the 10 species to calculate a motif score [representing the number and binding strength of experimentally characterized transcription factor binding sites (TFBSs)] for 188 *Drosophila melanogaster* TFs (*6*) with at least one ortholog in each of the 10 bees, and correlated motif score with social complexity, using phylogenetically independent contrasts (*7*). Of 2101 significantly correlated motif-gene pairs, 89% were positive and 11% negative, showing that TFs tend to have increased capacity to regulate genes in eusocial species of bees, relative to solitary species (Fig. 2A, supplementary materials).

Further evidence for increased capacity for gene regulation throughout social evolution is a positive ranked correlation between social complexity and the number of genes predicted to be methylated (*7*) (Spearman's rho = 0.76, P = 0.01; phylogenetically corrected Spearman's rho = 0.64, P = 0.06; Fig. 2B; bioinformatics predictions validated with bisulfite sequencing data for three invertebrate species; supplementary materials). DNA methylation affects gene expression in a variety of ways (*8*). Thus, this result suggests an expansion in regulatory capacity with increasingly sophisticated sociality.

The potential for increased regulatory capacity was further revealed at the protein-coding level. Increased social complexity also is associated with rapid evolution of genes involved in coordinating gene regulation. A Bayesian phylogenetic covariance analysis (*9*) of 5865 single-copy orthologs identified 162 genes with accelerated evolution in species with increased social complexity (*7*) (additional data table S3). These rapidly evolving genes were significantly enriched (P < 0.05) for Gene Ontology (GO) terms related to regulation of transcription, RNA splicing, ribosomal structure, and regulation of translation (supplementary text and tables S11 and S12). Similar results have been reported for bee and ant species (*10*–*13*); our findings reveal the underlying causes. Approximately two-thirds of these genes are under stronger directional selection in species with increasingly complex eusociality, but we also detected nonadaptive evolution. One-third of the rapidly evolving genes are under relaxed purifying selection in species with complex eusociality, possibly due to reduced effective population sizes (*14*).

We also found an additional 109 genes, significantly enriched (P < 0.05) for functions related to protein transport and neurogenesis, which evolve slower with increased social complexity (supplementary text, table S13, and additional data table S3). This includes orthologs of *derailed 2* and *frizzled*, which function as Wnt signaling receptors in *Drosophila* synaptogenesis (*15*), and *rigor mortis*, a nuclear receptor involved in hormone signaling (*16*). A similar pattern of reduced evolutionary rate has been described for genes expressed in human and honey bee brains, potentially due to increasing pleiotropic constraint in complex gene networks (*17*, *18*). Constrained protein evolution of neural and endocrine-related genes seems at odds with the evolution of complexity, but this constraint appears to be compensated for, or perhaps driven by, increased capacity for gene regulation.

We next investigated whether these molecular evolution patterns involve similar sets of genes and cis-regulatory elements among the early (facultative and obligate simple eusociality) and advanced (complex eusociality) stages of independent social transitions. We identified lineage-specific differences in coding sequences and promoter regions of 1526 "social genes" for which evolutionary rate (dN/dS) is faster or slower with increased social complexity in two independent origins and two independent elaborations of eusociality (*7*)

Downloaded from www.sciencemag.org on July 9, 2015

(Fig. 1). Among these lineage-specific social genes, we found common patterns of cis-regulatory evolution: gains of TFBSs in the promoters of genes that evolve slower with increasing social complexity (Fig. 2C and supplementary text). This suggests that a shared feature of both independent origins and elaborations of eusociality is increasingly constrained protein evolution with increasing potential for novel gene expression patterns. The TFs responsible for this pattern were different for each social transition, even though our analysis was limited to highly conserved TFs (Table 1). Several function in neurogenesis or neural plasticity, or are prominent regulators of endocrine-mediated brain gene expression in honeybees (*19*, *20*).

We found further lineage-specific differences among the rapidly evolving "social genes" themselves. Genes undergoing accelerated evolution at the origins of eusociality were significantly enriched for GO terms related to signal transduction in both Apidae and Halictidae, but they shared only six genes (6 out of 354 and 167 genes, respectively; hypergeometric test, $P = 0.82$; Fig. 2D and additional data tables S5 and S6). Rapid evolution of signal transduction pathways may be a necessary step in all origins of eusociality to mediate intracellular responses to novel social and environmental stimuli (*10*), but selection appears to have targeted different parts of these pathways in each independent transition. Caste-specific expression and other analyses of these genes are needed to determine their function in eusociality.

Genes showing signatures of rapid evolution with the elaborations of complex eusociality were also highly disparate between honeybees and stingless bees, with only 43 shared genes and no shared enriched GO terms (43 out of 625 and 512 genes, respectively; hypergeometric test, $P = 0.70$; Fig. 2D and additional data tables S5 and S6). In addition, only 2 out of 5865 single-copy orthologs showed a signature of convergent evolution by fitting a dendrogram based on social complexity significantly better than the accepted molecular phylogeny (*7*) (supplementary text and fig. S21). Similarly, families of major royal jelly protein genes, sex-determining genes, odorant receptors, and genes involved in lipid metabolism expanded in some, but not all, lineages of complex eusocial bees (*7*) (Table 2 and supplementary text). These results suggest that gene family expansion is associated with complex eusociality as predicted (*5*), but involves different genes in each case. Despite striking convergence of social traits among the superorganisms (*4*), the final stages of transformation to this level of biological organization do not necessarily involve common molecular pathways.

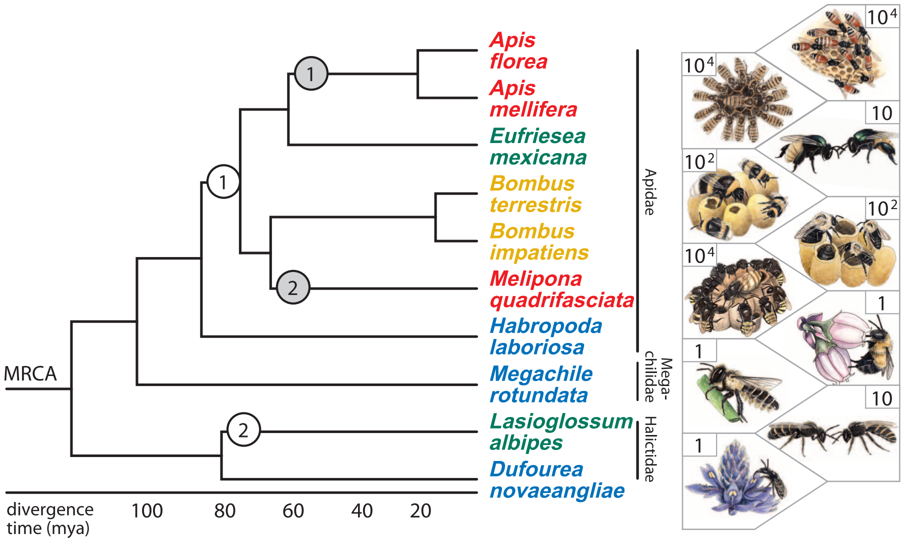

Fig. 1. Phylogeny and divergence times (*28*) of bees selected for genome analysis. We analyzed two independent origins of simple eusociality from a solitary ancestor, one each in Apidae (white circle 1) and Halictidae (white circle 2), and two independent elaborations of complex eusociality in honeybees (gray circle 1) and stingless bees (gray circle 2). Most bees mate once, but honeybees mate with multiple males. All bees eat pollen and nectar from flowering plants. Species names are colored according to degree of social complexity: blue: ancestrally solitary; green: facultative simple eusociality; orange: obligate simple eusociality; red: obligate complex eusociality. The social biology of *E. mexicana* is unknown, but is representative of the facultative simple eusocial life history (*29*). Numbers in each box are approximate colony size on a log scale. MRCA, most recent common ancestor; mya, millions of years ago.

[1]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [2]Department of Entomology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [3]Department of Biology, Utah State University, Logan, UT 84322, USA. [4]China National GeneBank, BGI-Shenzhen, Shenzhen, 518083, China. [5]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, 1350, Denmark. [6]Departments of Biomedical Engineering, Computer Science, and Biostatistics, Johns Hopkins University, Baltimore, MD 21218, USA. [7]Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA. [8]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [9]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA. [10]Program in Ecology and Evolutionary Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [11]Department of Biology, Hobart and William Smith Colleges, Geneva, NY 14456, USA. [12]Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [13]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA. [14]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112, USA. [15]U.S. Department of Agriculture–Agricultural Research Service (USDA-ARS) Red River Valley Agricultural Research Center, Biosciences Research Laboratory, Fargo, ND 58102, USA. [16]Center for Ecological Research and Forestry Applications (CREAF), Universitat Autonoma de Barcelona, 08193 Bellaterra, Spain. [17]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [18]Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland. [19]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [20]The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [21]Institute of Biology, Department Zoology, Martin-Luther-University Halle-Wittenberg, Hoher Weg 4, D-06099 Halle (Saale), Germany. [22]Queen Mary University of London, School of Biological and Chemical Sciences Organismal Biology Research Group, London E1 4NS, UK. [23]Department of Laboratory Medicine, University Hospital Halle, Ernst Grube Strasse 40, D-06120 Halle (Saale), Germany. [24]German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, 04103 Leipzig, Germany. [25]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA. [26]Department of Entomology, University of Georgia, Griffin, GA 30223, USA. [27]Center for Functional and Comparative Insect Genomics, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [28]Departamento de Biologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, 14040-901 Ribeirão Preto, SP, Brazil. [29]Departamento de Tecnologia, Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista (UNESP), 14884-900 Jaboticabal, SP, Brazil. [30]Departamento de Genética e Evolução, Centro de Ciências Biológicas e da Saúde, Universidade Federal de São Carlos, 13565-905 São Carlos, SP, Brazil. [31]Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900 Ribeirão Preto, SP, Brazil. [32]Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, 14049-900 Ribeirão Preto, SP, Brazil. [33]USDA-ARS Bee Research Lab, Beltsville, MD 20705 USA. [34]Department of Biology, East Carolina University, Greenville, NC 27858, USA. [35]Department of Entomology, Ohio Agricultural Research and Development Center, Ohio State University, Wooster, OH 44691, USA. [36]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA. [37]Department of Animal Sciences, University of Illinois, Urbana, IL 61801, USA. [38]Department of Population Genomics, Institute of Animal Husbandry and Animal Breeding, University of Hohenheim, Germany. [39]Department of Biology, York University, Toronto, ON M3J 1P3, Canada. [40]Janelia Farm Research Campus, Howard Hughes Medical Institue, Ashburn, VA 20147, USA. [41]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [42]Department of Entomology, Texas A&M University, College Station, TX 77843, USA. [43]Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138, USA. [44]Department of Biology, University of Copenhagen, 2200 Copenhagen, Denmark. [45]Princess Al Jawhara Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah 21589, Saudi Arabia. [46]Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China. [47]Department of Medicine, University of Hong Kong, Hong Kong. [48]Center for Advanced Study Professor in Entomology and Neuroscience, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [49]Centre for Social Evolution, Department of Biology, Universitetsparken 15, University of Copenhagen, DK-2100 Copenhagen, Denmark.
*These authors contributed equally to this work. †Corresponding author. E-mail: karen.kapheim@usu.edu (K.M.K.); wangj@genomics.org.cn (J.W.); generobi@illinois.edu (G.E.R.); zhanggj@genomics.org.cn (G.Z.)

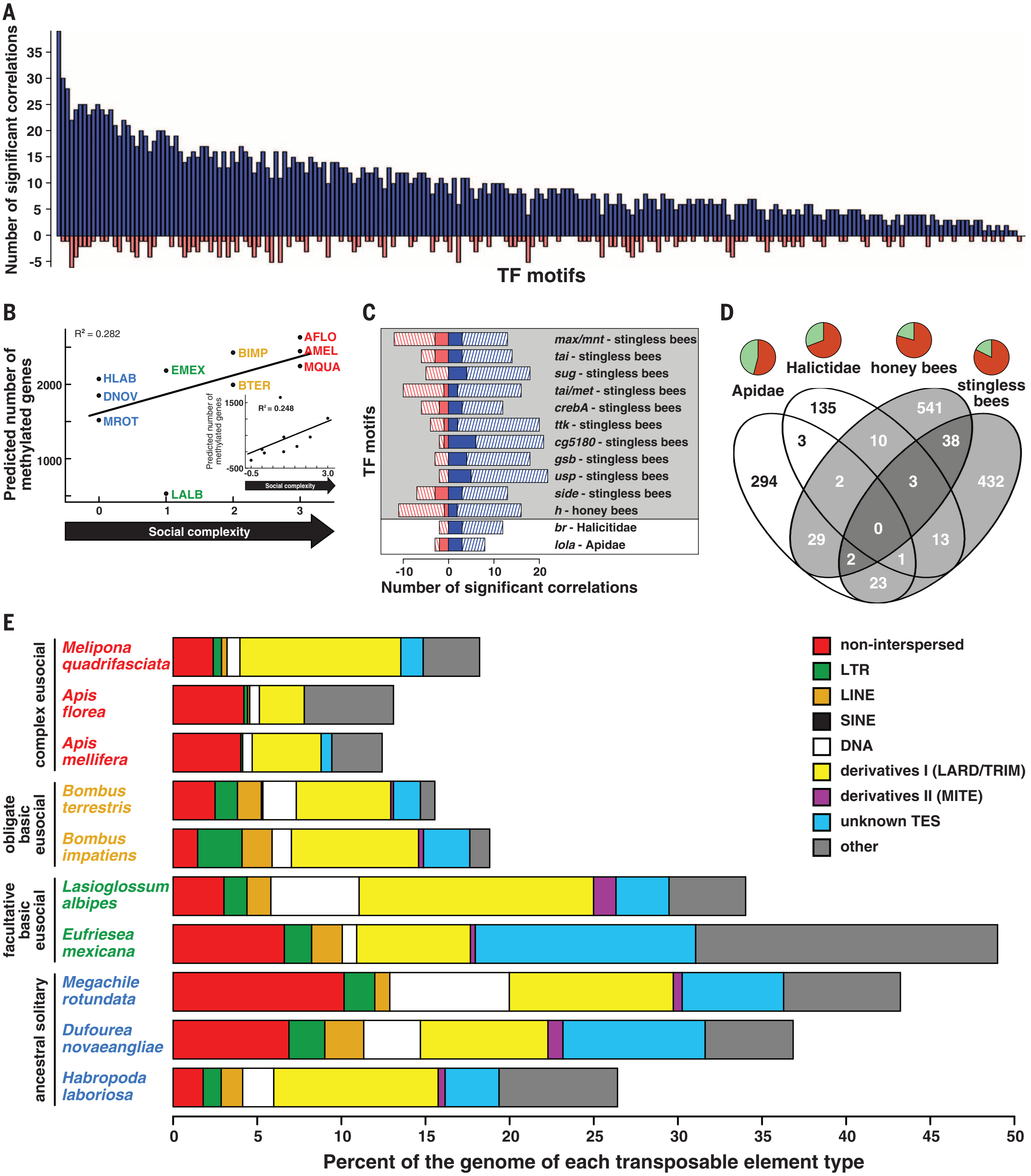

Fig. 2. Genomic signatures of evolutionary transitions from solitary to group life. (**A**) Increasing social complexity is associated with increasing presence of cis-regulatory TFBSs in promoter regions. Each bar represents a TFBS for which presence correlates significantly with social complexity (blue: positive; red: negative). (**B**) Relationship between predicted number of methylated genes and social complexity before and after (inset) phylogenetic correction (see text for statistics). (**C**) TFBS motifs showing a relationship between social complexity and evolutionary rate of coding and noncoding sequences in different lineages. Bar length indicates the number of significant correlations (blue: positive; red: negative) between each motif score and social complexity (from Table 1) among genes evolving faster (solid) or slower (hatched) in lineages with different levels of social complexity [from (D)]. Background shading follows circle shading in Fig. 1. (**D**) Number of genes for which evolutionary rate is faster or slower in lineages with higher compared to lower social complexity. Pie charts represent the proportion of genes evolving slower (light green) or faster (dark orange) with increased social complexity. Venn diagram shading follows circle shading in Fig. 1. (**E**) Complex eusocial species have a reduced proportion of repetitive DNA compared to other bees (see text for statistics). LTR, long terminal repeat; LINE, long interspersed element; SINE, short interspersed element; DNA, DNA transposon; LARD, large retrotransposon derivative; TRIM, terminal repeat retrotransposon in miniature; MITE, miniature inverted-repeat transposable element; TES, transposable elements.

The major transitions in evolution involve a reduction in conflict as the level of natural selection rises from the individual to the group (*1*). Extending this to intragenomic conflict may explain our finding of decreased diversity and abundance of transposable elements (TEs) with increasing social complexity (*7*) (regression after phylogenetic correction, $F = 8.99$, adjusted $R^2 = 0.47$, $P = 0.017$; Fig. 2E, figs. S42 to S44, and supplementary text). This may be a consequence of increased recombination rates among highly eusocial insects (*21*, *22*) or because key features of complex eusociality lead to decreased exposure to parasites and pathogens that horizontally transmit TEs (*4*, *23*). Eusociality in bees may thus provide natural immunity against certain types of intragenomic conflict.

Our results and those in (*10*–*13*) support the prediction that changes in gene regulation are key features of evolutionary transitions in biological organization (*5*). Our results further reveal the convergent adaptive and nonadaptive evolutionary processes common to both the early and advanced stages of multiple independent transitions from solitary to group living. It is now clear that there are lineage-specific genetic changes associated with independent origins of eusociality in bees, and independent elaborations of eusociality in both bees and ants. This includes different sets of genes showing caste-biased expression across species (*24*–*26*) and, as we have shown, evolutionary modifications of TEs, gene methylation, and cis-regulatory patterns associated with the suite of life-history traits that define eusociality. This suggests that if it were possible to "replay life's tape" (*27*), eusociality may arise through different mechanisms each time, but would likely always involve an increase in the complexity of gene networks.

Table 1. Transcription factors (TFs) and corresponding motifs associated with origins and elaborations of eusociality in bees. [Motif names: Fly Factor Survey (6); supplementary text.]

Motif	*D. melanogaster* TFs	Hypergeometric test *P*-value
	Solitary to simple eusociality–Apidae	
lola_PQ_SOLEXA	*Lola*	0.0047
	Solitary to simple eusociality–Halictidae	
br_PL_SOLEXA_5	*Br*	0.0016
	Simple eusociality to complex eusociality–honeybees	
h_SOLEXA_5	*dpn,h*	0.0027
	Simple eusociality to complex eusociality–stingless bees	
Side_SOLEXA_5	*E_spl, HLHm3, HLHm5, HLHm7, HLHmbeta, HLHmdelta, HLHmgamma, Side*	0.0008
usp_SOLEXA	*EcR,svp,usp*	0.0013
CrebA_SOLEXA	*CrebA*	0.0040
CG5180_SOLEXA	CG5180	0.0044
tai_Met_SOLEXA_5	*Mio_bigmax,tai_Met*	0.0045
ttk_PA_SOLEXA_5	*Ttk*	0.0078
gsb_SOLEXA	*gsb,Poxn,prd*	0.0083
tai_SOLEXA_5	*Tai*	0.0100

Table 2. Relative size of select gene families as related to social complexity in bees.

Family	Function	Eusocial bees compared to solitary bees
	Differences among bees	
Major royal jelly	Brood feeding	Expanded only in *Apis*
Sex determination pathway genes	Sex-specific development	Expanded in some eusocial lineages
Odorant receptors	Olfaction	Expanded in complex eusocial lineages
Lipid metabolism genes	Metabolic processing of lipids	Expanded in complex eusocial lineages
	Similarities across bees	
Biogenic amines receptors, neuropeptides, GPCRs*	Neural plasticity	Similar
Insulin-signaling and ecdysone pathway genes	Insect development, caste determination in honeybees, behavioral plasticity as adults	Similar
Immunity	Infectious disease protection	Similar
Cytochrome P450 monooxygenase genes	Detoxification	Similar

*GPCRs, G protein–coupled receptors.

REFERENCES AND NOTES

1. J. Maynard Smith, E. Szathmáry, *The Major Transitions in Evolution* (Oxford Univ. Press, Oxford, UK, 1995).
2. C. D. Michener, *The Social Behavior of the Bees* (Harvard Univ. Press, Cambridge, MA, 1974).
3. A.-M. Klein *et al.*, *Proc. Biol. Sci. B* **274**, 303–313 (2007).
4. H. Hölldobler, E. O. Wilson, *The Superorganism: The Beauty, Elegance and Strangeness of Insect Societies* (Norton, New York, 2009).
5. B. R. Johnson, T. A. Linksvayer, *Q. Rev. Biol.* **85**, 57–79 (2010).
6. L. J. Zhu *et al.*, *Nucleic Acids Res.* **39**, D111–D117 (2011).
7. Materials and methods are available as supplementary materials on *Science* Online.
8. H. Yan *et al.*, *Annu. Rev. Entomol.* **60**, 435–452 (2015).
9. N. Lartillot, R. Poujol, *Mol. Biol. Evol.* **28**, 729–744 (2011).
10. S. H. Woodard *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7472–7477 (2011).
11. B. A. Harpur *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2614–2619 (2014).
12. J. Roux *et al.*, *Mol. Biol. Evol.* **31**, 1661–1685 (2014).
13. D. F. Simola *et al.*, *Genome Res.* **23**, 1235–1247 (2013).
14. J. Romiguier *et al.*, *J. Evol. Biol.* **27**, 593–603 (2014).
15. M. Park, K. Shen, *EMBO J.* **31**, 2697–2704 (2012).
16. J. Gates, G. Lam, J. A. Ortiz, R. Losson, C. S. Thummel, *Development* **131**, 25–36 (2004).
17. D. Brawand *et al.*, *Nature* **478**, 343–348 (2011).
18. D. Molodtsova, B. A. Harpur, C. F. Kent, K. Seevananthan, A. Zayed, *Front. Genet.* **5**, 431 (2014).
19. D. W. Pfaff, A. P. Arnold, A. M. Etgen, R. T. Rubin, S. E. Fahrbach, Eds., *Hormones, Brain and Behavior* (Elsevier, New York, 2009).
20. S. Chandrasekaran *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18020–18025 (2011).
21. L. Wilfert, J. Gadau, P. Schmid-Hempel, *Heredity* **98**, 189–197 (2007).
22. E. S. Dolgin, B. Charlesworth, *Genetics* **178**, 2169–2177 (2008).
23. S. Schaack, C. Gilbert, C. Feschotte, *Trends Ecol. Evol.* **25**, 537–546 (2010).
24. B. Feldmeyer, D. Elsner, S. Foitzik, *Mol. Ecol.* **23**, 151–161 (2014).
25. P. G. Ferreira *et al.*, *Genome Biol.* **14**, R20 (2013).
26. B. G. Hunt *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15936–15941 (2011).
27. S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History* (Norton, New York, 1989).
28. S. Cardinal, B. N. Danforth, *Proc. Biol. Sci.* **280**, 20122686 (2013).
29. S. Cardinal, B. N. Danforth, *PLOS ONE* **6**, e21086 (2011).

ACKNOWLEDGMENTS

Data deposition at National Center for Biotechnology Information: *H. laboriosa*, *D. novaeangliae*, *E. mexicana*, *M. quadrifasciata*, and *M. rotundata* genome assemblies accession numbers PRJNA279436, PRJNA279825, PRJNA279814, PRJNA279820, and PRJNA66515. Funding for genome sequencing and analysis: BGI, a U.S. National Institutes of Health Pioneer Award (DP1 OD006416) to G.E.R., and European Union Marie Curie International Incoming Fellowship (300837) to G.Z. Additional funding: U.S. National Science Foundation grant DEB-0640690 (M.A.D.G.) and DEB-0743154 (G.E.R. and M.E.H); Danish Council for Independent Research grants 10-081390 (C.L.) and 0602-01170B (C.J.P.G.); Lundbeck Foundation (C.J.P.G); Georgia Tech–Elizabeth Smithgall Watts endowment (M.A.D.G.); Marie Curie International Outgoing Fellowship PIOF-GA-2011-303312 (R.M.W.); Swiss National Science Foundation award 31003A-125350, Commission Informatique of the University of Geneva, Schmidheiny Foundation, and Swiss Institute of Bioinformatics (E.M.Z.); and Natural Sciences and Engineering Research Council of Canada Discovery Grant and Early Research Award (A.Z.). This project was conducted under the auspices of the i5K Initiative. We thank the Roy J. Carver Biotechnology Center (sequencing services); N. Lartillot (advice on Coevol); T. Newman (assistance with DNA extractions); and E. Hadley (assistance with figures). Computational support: D. Davidson, N. Band, D. Slater (University of Illinois), J. H. Kidner and H. Scharpenberg (Martin-Luther-University Halle-Wittenberg); Compute Canada; and Center for High Performance Computing at the University of Utah. J. Himes created the illustrations in Fig. 1. We are grateful to T. Pitts-Singer, H. G. Hall, B. N. Danforth, J. Gibbs, and S. Cardinal for providing bee specimens; R. Ayala for identification of *E. mexicana*; J. Vega and the staff at Estación de Biología Chamela, Universidad Nacional Autónoma de México (UNAM), for support during collecting trips; and S. A. Cameron for insightful discussion.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6239/1139/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S44
Tables S1 to S32
Additional Data Tables S1 to S12
References (*30–151*)
Author Contributions

19 December 2014; accepted 6 May 2015
Published online 14 May 2015;
10.1126/science.aaa4788

HUMAN OOCYTES

Error-prone chromosome-mediated spindle assembly favors chromosome segregation defects in human oocytes

Zuzana Holubcová,[1] Martyn Blayney,[2] Kay Elder,[2] Melina Schuh[1]*

Aneuploidy in human eggs is the leading cause of pregnancy loss and several genetic disorders such as Down syndrome. Most aneuploidy results from chromosome segregation errors during the meiotic divisions of an oocyte, the egg's progenitor cell. The basis for particularly error-prone chromosome segregation in human oocytes is not known. We analyzed meiosis in more than 100 live human oocytes and identified an error-prone chromosome-mediated spindle assembly mechanism as a major contributor to chromosome segregation defects. Human oocytes assembled a meiotic spindle independently of either centrosomes or other microtubule organizing centers. Instead, spindle assembly was mediated by chromosomes and the small guanosine triphosphatase Ran in a process requiring ~16 hours. This unusually long spindle assembly period was marked by intrinsic spindle instability and abnormal kinetochore-microtubule attachments, which favor chromosome segregation errors and provide a possible explanation for high rates of aneuploidy in human eggs.

Meiosis in human oocytes is more prone to chromosome segregation errors than mitosis (*1*, *2*), meiosis during spermatogenesis (*3*, *4*), and female meiosis in other organisms (*3*, *5*). Despite its importance for fertility and human development, meiosis in human eggs has hardly been studied. Human oocytes are only available in small numbers, warranting single-cell assays capable of extracting maximal information. Although high-resolution live-cell microscopy is an ideal method, oocyte development in the ovary poses challenges to direct imaging. We therefore established an experimental system (*6*) for ex vivo high-resolution fluorescence microscopy of human oocytes freshly harvested from women undergoing gonadotropin-stimulated in vitro fertilization cycles. To establish the major stages of meiosis in this system, we simultaneously monitored microtubules and chromosomes for ~24 to 48 hours (Fig. 1 and movie S1). Similar to the situation in situ (*7*), human oocytes matured into fertilizable eggs over this time course, as judged by the formation of a polar body. The morphologically identifiable stages (Fig. 1A) at characteristic times after nuclear envelope breakdown [(NEBD), set to 0 hours] provided a time-resolved framework for human oocyte meiosis (Fig. 1B). This reference timeline post-NEBD is used throughout this paper.

Before NEBD, chromosomes were highly condensed and clustered around the nucleolus. Instead of rapidly nucleating microtubules upon NEBD, human oocytes first formed a chromosome aggregate that was largely devoid of microtubules (Fig. 1A; movie S1; and fig. S1, A and B). Microtubules were first observed at ~5 hours, when they started to form a small aster within the chromosome aggregate. As the microtubule aster grew, the chromosomes became individualized and oriented on the surface of the aster with their kinetochores facing inwards. The microtubule aster then extended into an early bipolar spindle that carried the chromosomes on its surface (Fig. 1A; movie S1; and fig. S1, C to E). The chromosomes then entered the spindle but remained distributed throughout the entire spindle volume. Chromosomes first congressed in the spindle center at ~13 hours but continued to oscillate around the spindle equator. Stable chromosome alignment was typically only achieved close to anaphase onset (Fig. 1, A and B, and movie S1). Unexpectedly, the spindle volume increased over the entire course of meiosis, up until anaphase onset (Fig. 1, C and D). The barrel-shaped spindle formed in this process consisted of loosely clustered bundles of microtubules and lacked astral microtubules (movie S2 and fig. S2). At ~17 hours, the oocytes progressed into anaphase and eliminated half of the homologous chromosomes in a polar body. Nearly a day after NEBD, the oocytes had formed a bipolar metaphase II spindle and matured into a fertilizable egg. The stages and timing of meiosis were highly reproducible among oocytes (Fig. 1, A and B) and could also be observed in fixed oocytes (fig. S1, A to I). Importantly, 79.0% of imaged human oocytes extruded a polar body. This indicates that the imaging assays, as well as the methods by which the oocytes were obtained and processed, did not have a prominent effect on meiotic progression.

The surprisingly slow and gradual build-up of the spindle over 16 hours (Fig. 1, C and D) is in stark contrast to mitosis, where spindle assembly takes only ~30 min (*8*), or meiosis in mouse oocytes, where it takes 3 to 5 hours (*9*–*11*). During mitosis, two centrosomes ensure the rapid assembly of a spindle. In oocytes of many species, centrosomes are absent but functionally replaced by microtubule organizing centers (MTOCs) that lack centrioles (*9*, *12*). Human oocytes also lack centrosomes (*13*–*15*), but whether acentriolar MTOCs participate in spindle assembly is unclear (*16*–*19*). We consistently detected pericentrin- and γ-tubulin–positive MTOCs at the spindle poles of mitotic cells and metaphase I and II (MI and MII) mouse oocytes, but never at MI or MII spindles in human oocytes (Fig. 2, A and B, and fig. S3). Thus, our data suggest that meiotic spindles in human oocytes lack detectable MTOCs.

In *Xenopus* egg extracts, chromosomes can serve as sites of microtubule nucleation if centrosomes are absent (*20*). The human oocytes we imaged also initiated microtubule nucleation in the region of the chromosome aggregate (78 of 78 live human oocytes). High-resolution imaging of fixed human oocytes confirmed that microtubules were first nucleated on chromosomes, emanating primarily from kinetochores (Fig. 2C, movie S3, and fig. S4). MTOC-nucleated cytoplasmic asters, such as those seen in chromosomal proximity upon NEBD in mouse oocytes (*9*), could not be detected. Thus, chromosomes, not MTOCs, serve as major sites of microtubule nucleation in human oocytes.

[1]Medical Research Council, Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK. [2]Bourn Hall Clinic, Bourn, Cambridge CB23 2TN, UK.
*Corresponding author. E-mail: mschuh@mrc-lmb.cam.ac.uk

Science AAAS

Genomic signatures of evolutionary transitions from solitary to group living

Karen M. Kapheim *et al.*

Science **348**, 1139 (2015);
DOI: 10.1126/science.aaa4788

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this information is current as of July 9, 2015):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/348/6239/1139.full.html

Supporting Online Material can be found at:
http://www.sciencemag.org/content/suppl/2015/05/13/science.aaa4788.DC1.html

This article **cites 138 articles**, 72 of which can be accessed free:
http://www.sciencemag.org/content/348/6239/1139.full.html#ref-list-1

This article appears in the following **subject collections:**
Evolution
http://www.sciencemag.org/cgi/collection/evolution

Downloaded from www.sciencemag.org on July 9, 2015

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2015 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

www.sciencemag.org/cgi/content/full/science.aaa4788/DC1

Supplementary Materials for

Genomic signatures of evolutionary transitions from solitary to group living

Karen M. Kapheim,* Hailin Pan, Cai Li, Steven L. Salzberg, Daniela Puiu, Tanja Magoc, Hugh M. Robertson, Matthew E. Hudson, Aarti Venkat, Brielle J. Fischman, Alvaro Hernandez, Mark Yandell, Daniel Ence, Carson Holt, George D. Yocum, William P. Kemp, Jordi Bosch, Robert M. Waterhouse, Evgeny M. Zdobnov, Eckart Stolle, F. Bernhard Kraus, Sophie Helbing, Robin F. A. Moritz, Karl M. Glastad, Brendan G. Hunt, Michael A. D. Goodisman, Frank Hauser, Cornelis J. P. Grimmelikhuijzen, Daniel Guariz Pinheiro, Francis Morais Franco Nunes, Michelle Prioli Miranda Soares, Érica Donato Tanaka, Zilá Luz Paulino Simões, Klaus Hartfelder, Jay D. Evans, Seth M. Barribeau, Reed M. Johnson, Jonathan H. Massey, Bruce R. Southey, Martin Hasselmann, Daniel Hamacher, Matthias Biewer, Clement F. Kent, Amro Zayed, Charles Blatti III, Saurabh Sinha, J. Spencer Johnston, Shawn J. Hanrahan, Sarah D. Kocher, Jun Wang,* Gene E. Robinson,* Guojie Zhang*

*Corresponding author. E-mail: karen.kapheim@usu.edu (K.M.K.); wangj@genomics.org.cn (J.W.); generobi@illinois.edu (G.E.R.); zhanggj@genomics.org.cn (G.Z.)

Published 14 May 2015 on *Science* Express
DOI: 10.1126/science.aaa4788

This PDF file includes:

Materials and Methods
Supplementary Text
Author Contributions
Figs. S1 to S44
Tables S1 to S32
Captions for additional data tables S1 to S12
References

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/science.aaa4788/DC1)

Additional Data table S1	Supplementary tables for analysis of TF motifs
Additional Data table S2	Orthogroup IDs and gene IDs of single-copy orthologs in ten bees
Additional Data table S3	Results of Coevol analysis of dN/dS and signatures of selection in ten bees
Additional Data table S4	Single-copy orthologs included in each of the four clade-specific PAML analyses
Additional Data table S5	Genes evolving at a rate consistent with social complexity for each of the four clade-specific analyses
Additional Data table S6	Enriched GO terms for genes evolving rapidly and slowly in association with social complexity from the clade-specific PAML analyses
Additional Data table S7	Manual annotation of major royal jelly protein genes and other developmental genes
Additional Data table S8	Pairwise comparisons of substitutions (per site,+/- SD) in deduced *double sex* (*dsx*) amino acid sequence, calculated within the DNA-binding (OD1) and *dsx*-dimerization domain among insects
Additional Data table S9	Pairwise comparisons of substitutions (per site,+/- SD) in deduced *transformer 2 (tra2)* amino acid sequence, calculated within the RNA-binding (RMM) domain and the remaining part of the protein region among insects
Additional Data table S10	Homology assignments for 190 immune-gene candidate proteins among the sequenced bee genomes. ORTHODB version 7 names given when available, along with official gene ID or,

	where appropriate, manually linked gene ID's (NCBI). Red entries indicate equally likely paralogs in protein set
Additional Data table S11	dN/dS of immunity-related genes across 10 bees
Additional Data table S12	Annotated P450 genes in the ten bee genomes

Materials and Methods

Genome sequencing, assembly, annotation, and orthology

H. laboriosa, E. mexicana, M. quadrifasciata, D. novaeangliae

H. laboriosa males were collected by H. Glenn Hall (University of Florida) in Gainesville, Fl in February 2011. Samples were live caught, stored at -70 °C, and shipped to University of Illinois on dry ice. DNA was extracted from the whole body from each bee independently by Thomas Newman (University of Illinois) and sent to BGI on dry ice. *D. novaeangliae* males were collected by Bryan Danforth, Jason Gibbs, and Sophie Cardinal on the south shore of Lake Ontario in July 2011. Bees were stored at -80 °C and shipped frozen to University of Illinois. Genomic DNA was extracted from whole bees by Thomas Newman, eluted in water, and shipped frozen to BGI. Male *E. mexicana* were collected in Chamela Bioreserve, Jalisco, Mexico by Karen Kapheim in July 2011. Individuals were captured in nets at odor baits scented with eucalyptus essential oil. Males were frozen and then stored in 80% ethanol until transport to University of Illinois. Genomic DNA was extracted by Thomas Newman from whole individuals at University of Illinois and shipped frozen to BGI. Male *Melipona quadrifasciata* were collected by Klaus Hartfelder as they emerged from combs from a colony fertilized with an irradiated male at the University of São Paolo campus. They were snap frozen in liquid nitrogen and stored at -80°C. Samples were carried frozen on person to Penn State and then shipped to University of Illinois. DNA was extracted by Thomas Newman at University of Illinois. DNA was shipped frozen to BGI.

We estimated genome size for *E. mexicana* and *D. novaeangliae* using flow cytometry as described in Hare and Johnston (*30*). In brief, neural tissue from a single adult individual was placed into 1 ml of Galbraith (*31*) buffer in a 2 ml Kontes (Kimble Chase, NJ) Dounce. Neural tissue from an adult female of the sequenced YW strain of *D. melanogaster* (1C = 175 Mb) was added as a standard to the tube with *D. novaeangliae*, while neural tissue from a *Periplaneta americana* male (1 C = 3338 Mb) was added as a standard with *D. novaengliae*. Each sample and standard were ground with 15 strokes of the A pestle, filtered through 40 µm nylon mesh, stained with 25 µg/ml propidium idodide (PI) and refrigerated at 4 ºC for at least 30 minutes in the dark prior to analysis. The amount of PI fluorescence of nuclei from the sample and standard were scored with a Partec (Partec North America, NJ) Cyflow cytometer with a solid-state laser emitting at 532 nm and the relative mean peak fluorescence recorded using Partec software. The amount of DNA in each sample was determined as the ratio of the mean peak position of the sample over the mean peak position of the standard multiplied by the genome size of the standard. For *E. mexicana*, the genome size was an average of replicate samples.

We used 12, 3, 3, and 4 DNA samples to construct libraries of *H. laboriosa*, *E. mexicana*, *M. quadrifasciata*, and *D. novaeangliae*, respectively, with insert sizes varying between 170bp to 20kb (Table S1). Then we applied Illumina platforms to sequence DNA and SOAPdenovo to assemble.

We adopted two different methods to predict genes: homology-based method and a *de novo method*. Results of two methods were integrated by GLEAN. Protein sequences of five species including *Apis mellifera, Bombus impatiens, Acromyrmex echinatior, Drosophila melanogaster* and *Homo sapiens* were used to predict genes by homology-based method. The homology-based follows 4 steps: 1) homology searching across the whole-geneome to get a non-redundant collection of protein sequences using TBLASTN; 2) selection of the most similar proteins for each region with protein homologous matching; 3) connect the short fragments using SOLAR; 4) use Genewise (version 2.0) (*32*) to generate the gene structures based on the homology alignments. Three de novo prediction programs, Augustus (*33*), GlimmerHMM (*34*) and SNAP (*35*), were used to predict genes, with parameters trained with 500-1000 intact genes from homology-based prediction. The genes originating from Augustus that also appeared in GlimmerHMM and SNAP results were picked as the final *de novo* set. The evidence derived from homology-based (5 sets for 5 species) and *de novo* (1 set for 3 programs) analyses were integrated to generate a consensus gene set by GLEAN (*36*). We then filtered genes to exclude transposon related genes and the genes whose cds regions contain more than 30% Ns. We then used transcriptomes assembled from RNA sequencing to improve the gene sets. RNA-seq reads were mapped to the genome by Tophat, and then Cufflinks was used to assemble transcripts. The assembled transcripts were then used to predict ORFs. The transcript-based gene models with intact ORFs that had no overlap with the GLEAN gene set were added to the gene set. In addition, if a transcript-based gene model with intact ORF covered more than one GLEAN gene, we would replace the GLEAN genes with the transcript-based gene model. The transcripts without intact ORFs were used to extend the incomplete GLEAN gene models to find start and stop codons.

Gene functions were assigned to the genes based on the best alignments to Swiss-Prot database (*37*) (release15.10) using Blastp. Then we searched Inter-Pro databases (*38*) (v29.0) including Pfam, PRINTS, PROSITE, ProDom, and SMART databases to find out the motifs and domains of proteins. GO terms for each gene were obtained from the corresponding Inter-Pro entries. All genes were aligned against KEGG proteins (http://www.genome.jp/kaas-bin/kaas_main?mode=partial), and the pathways in which the gene might be involved were derived from the best matched protein in KEGG.

M. rotundata

M. rotundata males were collected by Theresa Pitts-Singer (USDA-ARS) from commercially-supplied bees (JWM). DNA was extracted by Thomas Newman at University of Illinois, using a CsCl preparation.

Sequencing was performed at University of Illinois on an Illumina GAIIx platform, with DNA libraries insert sizes ranging from 475 basepairs to 10 kilobasepairs (Table S1). The reads were error corrected using Quake v0.2 (*39*) with a kmer size=18. In addition to correction, Quake also does adapter trimming. The trimming process reduced the total number of from 672,666,974 to 562,236,800. The error corrected reads were assembled using SOAPdenovo (*40*) v1.05 with a kmer size=47. The intra-scaffold gaps were closed using GapCloser v1.10.

The *M. rotundata* genome was annotated with the automated MAKER annotation pipeline (*41, 42*). The genome was masked for repetitive elements from a custom repeat library made with RepeatModeler (*43*), all the organisms in Repbase (*44*), and a hand-curated list of transposable element proteins supplied by MAKER.

RNA-seq data was generated for assistance with gene model prediction. *M. rotundata* prepupae were staged according to previously described methods (*45*). Total RNA was extracted from diapausing prepupae and post-diapausing prepupae using Trizol (Invitrogen) according to the manufacturer’s protocol. The isolated RNA pellets were stored under absolute ethyl alcohol at -80 °C until needed. The cDNA libraries (3 diapause prepupa and 3 post-diapause prepupae) were constructed by the University of Georgia Genomics facility using TruSeq RNA kit (Illumina). The libraries were pooled and then pair-end sequenced on an Illumina sequencer model HiSeq2000.

The evidence used for annotation consisted of the *Apis mellifera (46),* proteome (available on BeeBase (*47*)), and *Megachile rotundata* sequence available on Genbank, along with RNA-seq data assembled with Trinity (*48*). *ab-initio* gene predictions were generated by GeneMark (*49*), Augustus, and SNAP. This resulted in a gene set of 9,438 genes, 6,249 of which had homology to known proteins as identified with IPRScan (*50*). Following standard MAKER procedures (*51*), we added 1,749 ab-initio gene predictions that contained a known protein domain as identified by IPRScan (*50*).

Re-annotation of 4 other bees

We re-annotated the genome assemblies of *A. florea*, *B. terrestris*, *B. impatiens*, and *M. rotundata* using the same pipeline as described above for the four species sequenced at BGI. We mainly adopted two different methods to predict genes: a homology-based method and *de novo* gene prediction. Results from the two methods were integrated by GLEAN. Protein sequences of five species including *Apis mellifera*, *Bombus impatiens*, *Acromyrmex echinatior*, *Drosophila melanogaster* and *Homo sapiens* were used to predict genes by homology. Three *de novo* prediction programs, Augustus, GlimmerHMM and SNAP, were used to predict genes, with parameters trained using 500-1000 full-length genes identified by homology-based prediction. The Augustus origin genes which also appeared in GlimmerHMM and SNAP results were picked as the final *de novo* set. The gene models derived from homology-based (5 sets for 5 species) and *de novo* (1 sets for 3 programs) were integrated to generate a consensus gene set by GLEAN.

The transcripts assembled from RNA-seq data were used to predicted open reading frames (ORFs). Transcript-based gene models with intact ORFs that had no overlap with the GLEAN gene set were added into the set. In addition, if a transcript-based gene model with an intact ORF covered more than one GLEAN gene, we replaced the GLEAN genes with the transcript-based gene model. Likewise, if an *A. mellifera* homology-based gene model covered more than one GLEAN gene, we replaced the GLEAN gene with the homology-based prediction. The transcripts without intact ORFs were used to extend the incomplete GLEAN gene models to find start and stop codons. After filtering to remove transposon related genes, we obtained a final re-annotated gene set.

We found that *Bombus terrestris* had about 2,000 fewer genes than *B. impatiens* after finishing the above annotation procedures. This may be attributable to incomplete prediction in *B. terrestris,* possibly because of fragmentation in the genome assembly. Therefore after constructing gene families using TreeFam (*52*), we used the extra *B. impatiens* genes to search the *B. terrestris* assembly, and when we found homologous sequences, we used them to build gene models. These good gene models were added to *B. terrestris* gene set, and as a result both *Bombus* species have about 13,000 genes.

Genome assembly versions used for re-annotation

Megachile rotundata

- Assembly described above
- ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/invertebrates/Megachile_rotundata/MROT_1.0/

Apis florea

- sequenced by Baylor College of Medicine Human Genome Sequencing Center with 454 GS FLX Titanium

- assembled with Newbler v. 2.3-PreRelease-10/19/2009 and Phrap
- ftp://ftp.ncbi.nih.gov/genomes/Apis_florea/

Bombus terrestris (*53*)

- sequenced by Baylor College of Medicine Human Genome Sequencing Center with 454 GS FLX Titanium
- assembled with Newbler v. 2.3-PreRelease-10/19/2009 and Phrap
- ftp://ftp.ncbi.nih.gov/genomes/Bombus_terrestris/

Bombus impatiens (*53*)

- sequenced by the Biotechnology Center, University of Illinois (BCUI) with Illumina GAIIx
- assembled at the Center for Bioinformatics and Computational Biology at University of Maryland with SOAPdenovo v. 1.05
- ftp://ftp.ncbi.nih.gov/genomes/Bombus_impatiens/

Orthology

OrthoDB orthology delineation starts with the identification of all best-reciprocal-hits (BRHs) between genes from each pair of species from all-against-all Smith–Waterman protein sequence comparisons (see Table S2 for list of species.) Clusters are then built by progressively merging triangulating BRHs, pair-only BRHs, and finally all in-paralogs. OrthoDB orthology mapping first compares all genes from the species to be mapped to all genes in OrthoDB groups, and then follows the same BRH clustering procedure but only allowing new genes to be added to existing clusters. For further details on OrthoDB methodology and resources please see (*54*-*56*) and www.orthodb.org.

The relative completeness of each gene annotation set was assessed by examining counts of near-universal orthologous groups (with orthologs present in all, but one or two species) at different nodes on the arthropod phylogeny from Arthopoda, to Holometabola, and Hymenoptera. Such apparently rare gene losses may represent real evolutionary events, however, they also highlight potential annotation errors where genes have been completely missed, or only fragments have been predicted, or they have been fused with neighboring gene predictions, or they are completely missing from the assembly. For the same nodes of the phylogeny, the proportions of predicted genes for which orthology could be determined were compared across species to examine fractions of species-specific genes, and sparsely-present, widely-present, and universal orthologs.

Transcription factor characterization

We characterized the genes containing a DNA-binding domain or that were considered basal transcription factors (TFs) in each genome. We first used 'hmmscan' function in Hmmer v. 3.0 (*57*) to query protein sequence file of gene models for each species against the Pfam-A database (*58*) (downloaded Pfam-A.seed from ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/Pfam-A.seed.gz on 30August2012). We then scanned this database with protein sequence files, using the gathering threshold as a cutoff for inclusion and significance (option --cut_ga).

We then filtered the results of this scan for a curated list of TFs. The curated list was created from DNA-binding domain families identified from Transcription factor database v2.0 (*59*) (www.transcriptionfactor.org, downloaded on 30 August 2012). We then added missing ‘basal’ and ‘other’ accession numbers from Huang et al. 2012 (*60*). One exception is: PF10492

(Nrf1_activ_bdg) was found in Amel in Huang et al, but does not come up in the Amel HMMER results against all the Pfam db; PF10491 (Nrf1_DNA-bind) does come up, however, and does seem to also be a TF, based on entry on Pfam website; PF10491 is not in the Huang et al. Figure 1 list, however. We included this as a TF.

Transcription factor binding site analysis

Motif Selection and Scanning

The motifs of ~360 transcription factors were characterized in *Drosophila* with a bacterial one-hybrid system followed by sequencing by FlyFactorSurvey (*6*). For the bee analysis, similarities between the motifs are calculated and a set of 223 representative motifs (one from each cluster) is selected. We searched the orthogroups provided by OrthoDB (*54*) for the *Drosophila* TFs covered by our collection. The orthogroups containing at least one of the fly TFs are examined for copy number variation and the corresponding representative motif is selected for motif scanning.

The cis-regulatory analysis begins by producing a normalized genome-wide scoring profile for each selected TF motif in each of the 10 bee genomes. The first step of this process is to mask out the tandem repeats in each genome with the Tandem Repeat Finder (*61*). Next, each genome is divided into 500 bp windows that overlap by 250 bp. The HMM-based motif scoring program Stubb (*62*) is run on all selected motifs and all genomes to produce a motif score for every window in all genomes. Stubb was run with a fixed transition probability (0.0025) to the motif state and a background state nucleotide distribution learned from 5kbp centers of gene deserts (regions without coding features of length at least 22 kbp) of the corresponding genome. Within each species for each motif, the window scores are rank normalized into score from 0 (best) to 1 (worst). Because a motif composed of mostly C's and G's is expected to find a high Stubb score in a G/C rich window, the normalization considers the window's local G/C content. The procedure separates genomic windows into 20 equal-sized bins based on their G/C content, and performs rank-normalization within each bin separately.

PIC Analysis

To correlate presence of cis-regulatory TF motifs with level of sociality, we use the method of Phylogenetically Independent Contrasts (PIC) (*63*). We focus on the 7204 orthogroups that have at least one representative ortholog from each of the 10 bee species. For each orthogroup, we calculate a p-value for each motif and each species. This p-value is calculated as $P_s^m = 1 - (1 - N_s^m)^{w_s}$ where N_s^m is the best G/C normalized score for motif m among the w_s windows that fall within 5kb upstream and 2 kb downstream of the transcription start site of the gene(s) in the ortholog from species s. To consider the PIC correlation between the motif and sociality for a (motif, orthogroup) pair, at least one species must have evidence of the motif binding with its calculated motif p-value, P_s^m, less than 0.05. The motif p-values are converted to corresponding z-scores from the standard normal distribution. Negative z-scores are thresholded at zero to reflect our belief that the motif score of a window is less sensitive to under-representation of the motif compared to over-representation of the motif. PIC analysis is performed with the ape package (*64*) in R. The analysis relies on the phylogenetic tree, ((Hlab:91, ((Mqua:68, (Bimp:13, Bter:13):55):10, (Emex:62, (Amel:19, Aflo:19):43):16):13):15, Mrot:106):9, (Dnov:85, Lalb:85):30) (based on ref (*28*)). The leaves are assigned sociality values of 0 for solitary species, 1 for facultative basic eusocial, 2 for obligate basic eusocial, and 3 for complex eusocial. The PICs of the sociality labels are then calculated. For each (motif, orthogroup) pair, the motif p-values P_s^m are assigned to the appropriate tree leaves and the motif PICs are calculated. Finally, the coefficient and p-value of the Pearson correlation between the sociality PICs and motif PICs are recorded.

We "estimated" the false discovery rate ("eFDR") of the correlation analysis by creating a randomized control for each tested (motif, orthogroup) pair. In the randomized control, the motif z-scores were shuffled among the leaves of the tree while the sociality labels were held constant. The eFDR for a true correlation p-value is calculated as the quotient of the number of random controls significant at that threshold over the number of significant true tests at that threshold.

Cis-regulatory analysis of molecular evolution orthogroup sets

We next tested for associations between the orthogroups whose coding sequence is evolving at a rate consistent with a sociality hypothesis and the orthogroups whose motif presence in the cis-regulatory regions correlate with sociality. For a particular motif and significance threshold, we find all of the orthogroups in which the motif's presence is correlated with sociality (using the Phylogenetically Independent Contrast method) with a p-value less than the threshold. We then check for enrichment of these motif-based orthogroups within the sets of orthogroups obtained from the molecular evolution (PAML) analysis. We quantify the enrichment with the one- sided p-value of the Fischer exact test. We repeat this process with three different thresholds on the correlation significance threshold (<0.01, <0.05, <0.1, and in the top 100000 p-values across all motifs). To control for multiple hypothesis correction, we rerun the association tests 50 times with randomized motif sets from permutation of the orthogroups labels. The eFDR values (see above) are calculated from looking at the proportion of significant results for each (sociality set, motif) pair separately.

The final analysis more closely examines the relationship between coding and non-coding sequence evolution with respect to sociality. For each origins or elaborations of sociality set defined from the molecular evolution analysis, we examine two characteristics of its orthogroups. The first characteristic is derived from the PAML analysis and quantifies whether the orthogroup's coding sequence is evolving faster or slower with respect to sociality. The quantity is the negative logarithm (base 10) of the ratio of a fitted omega from the branch of one high sociality bee and the fitted omega from the branch of one low sociality bee from the PAML analysis on the H1 tree. The second characteristic is the PIC correlation of the motif's presence with the level of sociality, as described above. It shows whether the TF binding motif was gained or lost in the social bees (Fig. S7). For each (sociality set, motif) pair, we count the number of "significant" orthogroups that have a motif correlation p-value <0.25 and a 3-fold change in omega. We find the signs of these two characteristics for each "significant" orthogroup and report the most frequent pattern.

Characterization of CpG distribution

DNA methylation predominantly targets CpG dinucleotides (5' – 3' cytosine followed by guanine) in insect genomes. Depletion of normalized CpG content (CpG observed/expected [o/e]) represents an evolutionary signature of DNA methylation in animal genomes because methylated cytosines undergo spontaneous deamination to thymine with high frequency (*65*).

CpG o/e was calculated for genomic features as P_{CpG} / (P_C * P_G), where P_{CpG}, P_C, and P_G are the frequencies of CpG, cytosine, and guanine, respectively (*66*). In addition, the corresponding metric, GpC o/e, which controls for GC content, was also calculated using similar methods for genomic features in all taxa. Features with CpG o/e or GpC o/e values of 0 or greater than 6 were excluded from analyses. Short features with a length < 100bp were also not included in analyses.

Sliding-window analysis of CpG o/e across gene positions was conducted using a custom perl script. Window size was set as 200 bp, with a mean CpG o/e value calculated in 20-base increments. Windows were dropped if CpG o/e content was zero for a given window or if > 50% of the window was composed of masked (uninformative) bases. CpG o/e and GpC o/e dendrograms were created using the neighbor joining method on species' gene CDS values for each measure.

In order to predict those genes that are confidently methylated and unmethylated in each of the 10 species, we modeled each species' CpG o/e as a 2 component distribution: the CpG-depleted component associated with DNA methylation, and the non-depleted component associated with unmethylated genes. Using the probability of a gene's membership in a given component as well as the normalized distance (z-score) from the mean, we called genes as either "putatively methylated" or "putatively unmethylated" (>0.7 probability for a given mixture component and >0.75 or < -0.75 CpG o/e SDs from mean respectively), or "undetermined" for genes not satisfying the other classifications' requirements.

Using fractional DNA methylation data from *Apis mellifera (67)* we validated these calls, finding that 96% of the 2456 genes (out of 6138) called as putatively methylated with our method were, in fact, found to be empirically methylated. Similarly, 89.5% of the 1033 genes called as putatively unmethylated in *A. mellifera* were deemed to be empirically unmethylated (Fig. S8). We conducted similar analyses on data from the fire ant *Solenopsis invicta* and the sea squirt *Ciona intestinalis* for which empirically determined methylation data were available (*67, 68*). We found that the balanced accuracy (the mean of the sensitivity and specificity) for predicting methylation status of genes from CpG o/e values was 0.873 and 0.776 for 1:1 orthologs in *A. mellifera* and *S. invicta*, and 0.776 for all genes in *C. intestinalis* (for which an equivalent conservation-limited subset of genes could not be comparably determined). Thus this method does provide a relatively reliable metric of DNA methylation.

In order to evaluate the relationship between CpG depletion and sociality independent of phylogenetic distance, we employed phylogenetic independent contrasts (PICs) (*63*). The Analysis of MicroArrays (APE) R package (*69*) was used to generate PICs for CpG o/e and sociality values for every ortholog group (1-to-1 orthologs), which when correlated, provides an estimate of the association of methylation and sociality for each 1-to-1 ortholog group after controlling for differences in phylogenetic distance between them. The analysis relies on the phylogenetic tree, ((Hlab:91, ((Mqua:68, (Bimp:13, Bter:13):55):10, (Emex:62, (Amel:19, Aflo:19):43):16):13):15, Mrot:106):9, (Dnov:85, Lalb:85):30) (based on ref (*28*)). The leaves are assigned sociality values of 0 for solitary species, 1 for facultative basic eusocial, 2 for obligate basic eusocial, and 3 for complex eusocial. The PICs of the sociality labels are then calculated. We then used these contrasts in a Spearman's rank correlation to assess the relationship between the ranked number of genes predicted to be methylated in each species and social complexity. We used ranked values and non-parametric statistics to account for variation in methylation patterns between species (Fig. S11).

Phylogenetic covariance analysis of GC-corrected dN/dS across ten bees

We used Coevol 1.3c (*9*, *70*, *71*) to model evolutionary processes over the 10-species tree in 5,856 single-copy orthologs identified with OrthoDB (Additional Data table S2). Coevol can infer changes in genic equilibrium GC content, so it can calculate GC*, and dN/dSfrom the tree and sequence alignments. It also can take a matrix of "covariates" and infer their values throughout the tree. In our case the covariates supplied to the program were the sociality and CpG O/E (hereafter abbreviated as CpGOE). Initial runs were done without CpGOE data, so an additional set of 6000 runs were done with CpGOE. Sociality level was encoded as recommended by Lartillot et al, so that after Coevol takes the logarithm we get the values used elsewhere in the project (1=ancestrally solitary, 2=facultative basic eusociality, 3=obligate basic eusociality, 4=obligate complex eusociality). Both the authors of Coevol and others have noted that statistics on ω may differ from statistics on dN and dS separately, as the mean of a quotient is not in general the quotient of the means (*9*, *72*). To address this another 6000 runs were done using a model which estimates dS and dN, rather than dS and ω. Finally, the Coevol software requires a burn-in period for the Markov Chain simulations (currently set at 100 iterations) followed by a long run (set at 1000) to produce reasonable statistics. The p values given are based on the 1000 runs, so p=0.0 really means "no occurences in 1000 runs".

Outputs from a Coevol simulation run on the alignments for one gene include tables of mean and 95% confidence limits for the dS and ω inferred for each of the 10 species, plus an analysis of covariance of the 5 terms dS, ω, GC, Sociality, and CpGOE over the tree. This covariance analysis produces 4 tables: (a) the correlation r between each pair of terms; (b) the posterior probability p for each r; (c) the *partial* correlation between terms when variation due to other terms is controlled, and (d) the probability p for the partial correlations. These tables contain a total of 60 dS, ω, r, and p values for each of the 6000+ genes.

PAML analysis of dN/dS within clades

We performed PAML analyses of dN/dS within subsets of closely related species (i.e. clades) that vary in social complexity, and included an outgroup to root the tree (Fig. S17). For each set of species, 1:1:1 orthologs were extracted from the OrthoDB analysis (Additional Data table S4). First, we aligned CDSs of each ortholog with Prank. Each site of an alignment had a score to evaluate the aligning quality, in which the largest 1 indicates pretty high quality aligning. Then we masked poor site whose score was smaller than 0.5 to N by Guidance. Last, we performed likelihood ratio tests (LRT) after running PAML to quantify support for each of three hypothesized branch models of dN/dS. The branch models allow the ratio to vary among branches in the phylogeny, and are useful for differing rates of amino acid substitution acting on particular lineages (*73, 74*). The chi-square-derived P-values were calculated and adjusted by FDR method (a FDR cutoff of 0.05). Gene families for which ML scores for H0 and H1 are significantly different, H3 and H1 are significantly different, but H1 and H2 are not significantly different represent the most conservative set of genes evolving in association with eusociality. We disregarded any OrthoGroups with dS > 2, which could indicate saturation. We tested the following models (colors are as in Fig. 1):

Origins of eusociality

1. Apidae - ((Hlab, ((Bter, Bimp), Emex)), Mrot)
   - H0 – ((Hlab #1, ((Bimp #1, Bter #1) #1, Emex #1) #1) #1, Mrot #1)
   - H1 – ((Hlab #1, ((Bimp #2, Bter #2) #2, Emex #2) #2) #3, Mrot #3)
   - H2 – ((Hlab #1, ((Bimp #2, Bter #3) #4, Emex #5) #6) #7, Mrot #7) (free ratio model)
   - H3 – ((Hlab #1, ((Bimp #1, Bter #1) #1, Emex #1) #1) #2, Mrot #2)
2. Halictidae - ((Dnov, Lalb), Mrot)

- H0 – ((Lalb #1, Dnov #1) #1, Mrot #1)
- H1 – ((Lalb #2, Dnov #1) #3, Mrot #3)
- H2 – ((Lalb #2, Dnov #1) #3, Mrot #3) (free ratio model)
- H3 – ((Lalb #1, Dnov #1) #2, Mrot #2)

Elaborations of eusociality

1. Apinae A - (Hlab, ((Aflo, Amel), Emex))
   - H0 – (Hlab #1, (((Amel #1, Aflo #1) #1, Emex #1), #1))
   - H1 – (Hlab #3, (((Amel #2, Aflo #2) #2, Emex #1), #3))
   - H2 – (Hlab #3, (((Amel #5, Aflo #4) #2, Emex #1), #3)) (free ratio model)
   - H3 – (Hlab #1, (((Amel #2, Aflo #2) #2, Emex #2), #1))
2. Apinae B - (Hlab, ((Bter, Bimp), Mqua))
   - H0 – (Hlab #1, (((Bimp #1, Bter #1) #1, Mqua #1) #1))
   - H1 – (Hlab #3, (((Bimp #1, Bter #1) #1, Mqua #2) #3))
   - H2 – (Hlab #3, (((Bimp #5, Bter #4) #1, Mqua #2) #3)) (free ratio model)
   - H3 – (Hlab #1, (((Bimp #2, Bter #2) #2, Mqua #2) #1))

We used in house scripts to calculate enrichment of GO terms among the resulting lists of genes. For every GO term, a p value is calculated representing the probability that the observed numbers of counts could have resulted from randomly distributing this term between the tested gene list and the reference gene list (*75, 76*). The tested gene lists are resulting lists of genes above. The reference gene list is the total genes of one organism annotated, but one can use customized gene list as reference. The p value can be approximated by ChiSquare-Fisher test or hypergeometric test. The p values are corrected by fdr. For GO term with child terms, the observed number of this GO terms will include all child terms.

Convergent evolution analysis

We used a Maximum Likelihood (ML) method based on Parker et al. 2013 (*77*) to detect signatures of molecular convergence in genomic data by identifying genes that fit a "phylogenetic" tree based on phenotypes significantly better than the species phylogeny. We used two hypotheses to identify convergent evolution of sociality. H1 regarded all social species form a monophyletic clade. H2 split social species into three monophyletic clades based on social complexity (Fig. S20). We constructed two versions of these hypothesized trees – one with Emex in the eusocial group and one with Emex in the solitary group, based on the fact that the social biology of Emex is unknown. Alignments of single-copy orthologs (Additional Data table S2) were fitted to the H0, H1 and H2 topologies using RAxML. H0 is the same as species topology and H1, H2 are two hypothesised convergent phylogenies. Since the data and the substitution models were the same, the difference in likelihood between two phylogenies reflects the strength of support for each in the data. In order to clarify unambiguous topologies for H1 and H2, we merged 6,093 inferred gene trees into a majority clade-consensus (MCC) summary phylogeny tree for H1 and H2 using TreeAnnotator v1.7.4. We calculated the log-likelihood of the phylogeny for every site in the alignment using ML. From this, we calculated site-specific likelihood support (ΔSSLS): ΔSSLS = lnLi,H0 - lnLi,Ha. Where ΔSSLS for the ith site is given by the difference in log-likelihood units between the log-likelihood of the ith site under H0 (the species tree) and Ha (the alternative tree; one of H1 or H2) (following ref (*77*)). Positive ΔSSLS thus reflect sites with a better model fit to H0 (the species tree), and negative ΔSSLS reflect sites with a better fit to the alternative (phenotypic) topologies H1 or H2. To test the significance of the ΔSSLS measures, we performed the same comparisons on simulated datasets (phylobayes3.3f, 30 simulated alignments for each ortholog), and calculated their stepwise empirical cumulative density function, based on collated ΔSSLS values with linear interpolation, as in ref (*77*).

We found 16 orthogroups with at least one ΔSSLS < 0 and p < 0.05 (Table S15). We constructed gene trees for these 16 orthogroups using RaxML with cds alignments without any restriction to further to determine whether these genes showed an evolutionary signature of eusociality.

<u>Gene annotations for the Major Royal Jelly genomic region and signaling pathways functionally involved in caste development of bees</u>

The strategy we used for the annotation of these genes was the following:

1) The *Apis mellifera* proteins of the genes obtained from RefSeq database were used as query in BLAST (*78*) searches against all proteins predicted in the ten bee genomes.

2) The same *Apis mellifera* proteins were then used as reference in BLAST searches using the predicted proteins as query.

3) Reciprocal Best Hits (RBH) in the previous two BLAST steps were recovered for each gene.

4) The gene IDs of the RBH procedure were then used to recover the respective scaffolds and gene model from all the predictions for the ten bee species.

5) The respective scaffolds with their gene model predictions were then loaded for manual annotation into an Artemis environment (*79*) implemented in a Linux server.

6) In Artemis, the gene model predictions were checked against the nr GenBank protein database, GenBank nucleotide database (including genomes), and the collection of proteins or genomic sequences of the ten species. Alignments were performed by BLAST or BLAT (*80*). After confirming exon structure, which included the elimination of over-predicted exons or insertion of putatively missing exons, translation start sites, exon/intron borders and stop codons were checked. All genes, with correct or curated gene models were then saved as gff files (Additional Data table S7).

7) Multiple protein sequence alignments for each gene set - the proteins of *mrjp* and flanking *yellow* genes were run together – were generated using T-Coffee (*81*). Phylogenetic trees were generated using PhyML (*82*), a maximum likelihood approach.

<u>Manual annotation of genes in the sex determination pathway</u>

Genome assemblies and annotations of bee species (this study) were used to identify gene copies of interest (*fem, tra2, dsx*, GB47018, GB47045, GB47023) taking Amel v4.5, OGS 3.2 as reference and using various Blast parameters to avoid non-detection errors. In addition, Hidden Markov profile searches (*83*) were performed to screen specifically for *fem* paralogs in bee genomes using HMMer3 on the protein (HMMseach) and nucleotide (nHMMer) levels (*84*). Multiple sequence alignments were generated using MUSCLE (*85*) and optimized manually. To reduce the loss of informative sites due to incomplete or misleading annotations, experimentally validated and publicly available data were used for some species and genes (Tables S19-S21). Genealogies were reconstructed after applying Model Test on the dataset to determine the evolutionary substitution model that best fit the data followed by a maximum likelihood algorithm for tree construction implemented in MEGA vs. 5.2 (*86*). We applied different statistical models implemented in the HyPhy package (*87*) to test for the presence of diversifying selection. We used the genetic algorithm approach (GA-Branch) and the branch-site random effect likelihood (REL) method to identify signs of selection by calculating the distribution of nonsynonymous vs. synonymous per site substitutions (dn/ds) among branches.

Gene family contraction and expansion

Compiling gene clusters

All versus all blast was done for the protein sequences of five species with an E-value cutoff of 1e-5. After conjoining the fragmental alignment for each gene pair by Solar (a program in Treefam), the alignments were passed to calculate the distance between two genes. Then hierarchical clustering algorithm was used to cluster all the genes, with parameters of min_weight=20, min_density=0.34, and max_size=1000. This generated 12,960 gene families.

Filtering

Gene families were filtered such that more than one species within a clade was represented in each gene family. For these purposes, we used the following clades: (HLAB, MQUA, BIMP, BTER), (EMEX, AMEL, AFLO), (MROT), and (DNOV, LALB). This was completed by Gregg Thomas. After filtering, 8,001 gene families, including 91,148 genes remained in the analysis. Gene family sizes ranged from 4-272.

Gene family birth and death rate estimates

Birth and death rate (lambda) was estimated for each gene family using CAFE v2.2(*88*), the filtered gene family size list, and the molecular divergence tree below. Molecular divergence times were estimated from Figure 1 in (*28*).

```
tree
(((Hlab:91,((Mqua:68,(Bimp:13,Bter:13):55):10,(Emex:62,(Amel:19,Aflo:19):43):
16):13):15,Mrot:106):9,(Dnov:85,Lalb:85):30)
```

We first used the command lambda -s to search for the optimized lambda across each branch. We then varied the number of parameters for lambda in 10 additional models of variable birth and death rates across the tree. Models 1-4 tested hypotheses about variation in lambda based on social organization. Model 5 tested whether lambda varied by family. Model 6 tested an independent branch rate hypothesis. Models 5 and 6 failed to converge due to their complexity. Models 7-10 were thus substitute for model 6. To set up models 7-10 we first found maximum likelihood lambdas for each branch independently, by setting 1 rate for the focal branch and another rate for all background branches. These lambda values were then clustered with kmeans clustering with 2, 3, 4, or 5 clusters. Cluster identity was then used to assign lambda parameters in models 7, 8, 9, and 10, respectively. Each model was run at least 5 times to check for convergence. Model 10 failed to converge. Maximum likelihood scores were recorded for each model, and Likelihood Ratio Tests were performed in Stata v. 9.2 to identify the model(s) which best fit the data. Model 9 was the best fit to the data. Gene families for which the size distribution was significantly different from a null model of random birth and death were identified with CAFE v2.2 under model 9. Gene families with family-wide p-vlaue < 0.01 were used for GO enrichment analysis against every gene in each gene family using GOstats (*76*) in R (*69*).

Models tested in CAFE v2.2:

```
# search for optimized lambda
lambda -s

# search for 2 parameter lambda; basic sociality different - model 1
lambda -s -t (((1,((1,(2,2)2)1,(2,(1,1)1)1)1)1,1)1,(1,2)1)
report report.cafe2.2.lambda.model1.13aug2013

# search for 2 parameter lambda; complex sociality different - model 2
lambda -s -t (((1,((2,(1,1)1)1,(1,(2,2)2)1)1)1,1)1,(1,2)1)
report report.cafe2.2.lambda.model2.13aug2013

# search for 2 parameter lambda; all social different from solitary - model 3
lambda -s -t (((1,((2,(2,2)2)2,(2,(2,2)2)2)1)1,1)1,(1,2)1)
report report.cafe2.2.lambda.model3.13aug2013

# search for 3 parameter lambda; solitary, basic, complex all different -
model 4
lambda -s -t (((1,((3,(2,2)2)2,(2,(3,3)3)2)1)1,1)1,(1,2)1)
report report.cafe2.2.lambda.model4.13aug2013

# search for 4 parameter lambda; each family different - model 5
lambda -s -t (((1,((1,(1,1)1)1,(1,(1,1)1)1)1)1,2)4,(3,3)4)
report report.cafe2.2.lambda.model5.13aug2013

# search for 18 parameter lambda; independent lambdas for each branch - model 6
lambda -s -t (((1,((2,(3,4)5)6,(7,(8,9)10)11)12)13,14)15,(16,17)18)
report report.cafe2.2.lambda.model6.13aug2013

# search for 2 parameter lambda; branch rate model, based on kmeans=2 - model 7
lambda -s -t (((1,((1,(2,2)1)1,(1,(2,2)1)1)1)1,1)1,(1,1)1)
report report.cafe2.2.lambda.moremodels.model7.14aug2013

# search for 3 parameter lambda; branch rate model, based on kmeans=3 - model 8
lambda -s -t (((2,((1,(3,3)2)2,(1,(3,3)2)2)2)2,2)2,(2,1)2)
report report.cafe2.2.lambda.moremodels.model8.14aug2013

# search for 4 parameter lambda; branch rate model, based on kmeans=4 - model 9
lambda -s -t (((2,((1,(3,4)2)2,(1,(3,3)2)2)2)2,2)2,(2,1)2)
report report.cafe2.2.lambda.moremodels.model9.14aug2013

# search for 5 parameter lambda; branch rate model, based on kmeans=5 - model
10
lambda -s -t (((3,((1,(5,4)3)3,(2,(4,4)3)3)3)3,3)3,(3,1)3)
report report.cafe2.2.lambda.moremodels.model10.14aug2013
```

Manual annotation of neural plasticity genes (biogenic amine receptors, neuropeptides, and GPCRs)

The assemblies and automatic annotations of the ten bee genomes were blasted (blastp and tblastn) with neuropeptide and protein hormone sequences as well as with neuropeptide and biogenic amine GPCR sequences known from other arthropods. Some of the gene models were manually corrected using the splice site prediction program SPL and the gene prediction program FGENESH+ (http://www.softberry.com/). The translated protein sequences were checked for the presence of a signal peptide (http://www.cbs.dtu.dk/services/SignalP/) (*89*) or seven transmembrane domains (http://www.cbs.dtu.dk/services/TMHMM/).

Manual annotation of genes related to immunity

We identified putative orthologs of known immune proteins by three methods. First, we queried gene sets for each bee species via reciprocal BLASTP from 190 candidate immune genes. When proteins were not found for a species, we searched the appropriate genome assemblies using TBLASTN. Next, we used orthology designations available for 183 of these 190 proteins in OrthoDB 6.0 (*54*) to connect proteins from different bee species. Finally, we constructed whole genome orthology scans using Proteinortho (v.4) (*90*), including all ten bee species and additional model species including *Anopheles gambiae* (Agam_3.7), *Bombyx mori* (Bmor_v100.pep.fa), *Nasonia vitripennis* (Nvit_2.0.pep.fa), *Tribolium casteneum* (Tcas_3.0.pep.fa), and *Drosophila melanogaster* (Dmel-all-translation-r5.54.fa.). We cross-referenced these approaches to determine consensus protein matches for each species. Where two of the three techniques agreed on an orthologous prediction we included that prediction as the best estimate.

For 97 proteins with conserved and nominally 1:1 orthologs across all ten taxa (Additional Data table S10), we leveraged estimates of amino-acid evolution generated by PAML. Specifically, we used Omega (dN/dS) values estimated for each gene to compare across social structure (six social species versus four largely solitary species and species with complex (n = 5) versus simple (N = 1) societies) and across taxonomy (*Apis* species (n = 2) *vs. Bombus* (n = 2) vs. *Habropoda* and *Lasioglossum*. We identified specific classes of immune genes (Additional Data table S10) and determined which classes, and proteins within each class, showed the strongest signs for diversifying and purifying selection.

Manual annotation of cytochrome P450 monooxygenase genes

Genomic regions in the bee genomes likely to encode P450s were located by comparing annotated P450 amino acid sequences from the genomes of *Apis mellifera* (*91*), *Nasonia vitripennis (92)* and *Drosphila melanogaster (93)* with the bee genomes using TBLASTN (E-value < 1E-10). Each genomic region identified as likely to encode a P450 was then aligned with both the set of automated gene predictions for each bee genome using BLASTN (E < 1E-50) and the amino acid sequences of other insect P450s using TBLASTX (E < 1E-10). A gene model was then constructed manually based on these alignments with the aid of a custom BioPerl script (http://bioperl.org). A gene model was designated as putatively functional, but incomplete, if more than 50% of the expected sequence was present and the missing sequence could be attributed to incomplete genome assembly. Pseudogenes were identified as such if more than one error (indel, in-frame stop codon, unmatched intron splice site or truncated sequence) was observed. Many P450 fragments were identified with less than 50% of the expected sequence. If a fragment coded for one of the clearly orthologous P450s in the CYP2 and mitochondrial clans and another fragment of the same gene could be identified these fragments were combined.

Amino acid sequences for all bee P450s were aligned with *N. vitripennis* using Muscle (*85*). P450 clan, family and subfamily membership was assigned based on a phylogenetic tree constructed using PhyML with amino acid substitution model and parameters chosen using jModelTest (*94*) (LG+G+F with 4 categories of substitution and gamma=1.96).

Manual annotations were mapped back to the official gene predictions by comparing GFF files using BedTools (*95*). The list of P450s with "official gene set" names and provisional "P450" names, as well as P450 clan classification, is available.

Manual annotation of inotocin hormone system

We performed an in-depth genome sequence analysis on honey bees and related species, taking advantage of the current inventory of available insect genomes to search for the presence of a bee version of the inotocin system. We used the Hymenoptera Genome Database (*47*) for several of the lineages analyzed here and for the bee, wasp, sawfly, and termite species not included on the Hymenoptera Genome Database, we referenced currently unpublished genome assemblies. We used inotocin receptor and ligand sequences from ants to search for an intact bee version of this hormone system in ten different species. A previous genomics study could not identify an inotocin system in honey bees, but did not address this issue for other bee species (*96*). Using the annotated inotocin receptor and inotocin protein sequences from ants (*97*), we performed tBLASTn searches in the genomes of the following bee species: *Habropoda laboriosa*, *Apis mellifera*, *Apis florea*, *Eufriesea mexicana*, *Melipona quadrifasciata*, *Bombus terrestris*, *Bombus impatiens*, *Megachile rotundata*, *Lasioglossum albipes*, and *Dufourea novaeangliae.* We also searched the genomes of the following insects: *Solenopsis invicta* (ant), *Pogonomyrmex barbatus* (ant), *Acromyrmex echinatior* (ant), *Atta cephalotes* (ant), *Camponotus floridanus* (ant), *Linepithema humile* (ant), *Harpegnathos saltator* (ant), *Polistes dominula* (wasp), *Diachasma alloeum* (wasp), *Nasonia vitripensis* (wasp), *Cephus cinctus* (sawfly), and *Zootermopsis nevadensis* (termite). Based on the location of the inotocin receptor gene along scaffolds in ants, we used neighboring gene regions upstream and downstream of the intact ant receptor and inotocin genes to search for regions of microsynteny and any remains of the inotocin hormone system in bees.

Repetitive elements in the genomes of ten bee species

Abbreviations

GAG	GAG-Protein of retrotransposons
HTT	horizontal transposon transfer
LARD	large retrotransposon derivative
LINE	long interspersed element
LTR	long terminal repeat
MITE	miniatiure inverted-repeat transposable element
ORF	open reading frame
POL	POL-polyprotein of retrotransposons
RT	reverse transcriptase
SSR	simple sequence repeat
SINE	short interspersed element
TASE	Transposase
TE	transposable element
TIR	terminal inverted repeat
TRIM	terminal repeat retrotransposon in miniature

We performed an analysis of repetitive DNA in 10 bee genome assemblies: *Apis mellifera* (Amel4.5, Elsik et al. 2014), *Apis florea* (Aflo1.0), *Apis dorsata* (Ador1.1), *Bombus terrestris* (Bter1.1), *Bombus impatiens* (Bimp2.0), *Eufriesea mexicana* (Emex1.0), *Melipona quadrifasciata* (Mqua1.0), *Habropoda laboriosa* (Hlab1.0), *Megachile rotundata* (Mrot1.0), *Lasioglossum albipes* (Lalb2.0) (*98*) and *Dufourea novaeangliae* (Dnov1.0). For technical reasons the small fraction of the assembly composed of large numbers of very short scaffolds was excluded from the analysis, such as for Emex (<700 bp), Hlab (<300 bp), Lalb (<200bp) and Dnov (<200 bp, only for the *de novo* detection and first round of annotation). Furthermore, in some occasions small parts of sequences were masked by N in order to avoid technical problems related to satellite repeats in the *de novo* analysis pipeline. All analyses were performed with equal parameters and manual classifications with the same software and same persons for each repeat class.

Repetitive elements were detected and annotated with the REPET software package v 2.0 (*99*) consisting of two pipelines integrating a set of bioinformatics programs. First, repeated sequences were detected by similarity (all-by-all blast using BLASTER) and LTR retrotransposons were detected by structural search (LTRharvest). The similarity matches were clustered with GROUPER, RECON and PILER, the structural matches with single-linkage NCBI Blastclust. From each cluster a consensus sequence is generated by multiple alignments with Map. The consensus sequences were analyzed for terminal repeats (TRsearch), tandem repeats (TRF), open reading frames (dbORF.py, REPET) and poly-A tails (polyAtail, REPET). Furthermore the consensuses were screened for matches to nucleotide and amino acid sequences from known transposable elements (RepBase 17.01) (*44*) using BLASTER (tblastx, blastx) as well as searched for HMM profiles (Pfam database 26.0) (*100*) using hmmer3 (*57*). Based on the detected structural features and homologies, the consensuses are classified by PASTEC according to Wicker et al. (*101*). Redundancies (BLASTER, MATCHER) as well as elements classified as SSRs (>0.75 SSR coverage) or unclassified elements built from less than 10 fragments are removed.

This set of *de novo* detected repetitive elements was used to mine the genome in the second pipeline with BLASTER (NCBI BLAST, sensitivity 4, followed by MATCHER), RepeatMasker (NCBI BLAST/ CrossMatch, sensitivity q, cutoff at 200) and CENSOR (NCBI BLAST). False positive matches were removed by an empirical statistical filter. Satellites were detected with TRF, MREPS and RepeatMasker and were then merged. Furthermore the genomic sequences were screened for matching nucleotide and amino acid sequences from known transposable elements using RepBase 17.01 (*44*) via BLASTER (tblastx, blastx) followed by MATCHER. Finally, a removal of redundant TEs, removal of SSR annotations included into TE annotations and "long join procedure" to connect distant fragments was performed. Sequences from the *de novo* repetitive element library which were found to have at least one perfect match in the genome were then used to re-run the whole analysis.

To ensure compatibility and to avoid introducing a bias, we refrained from a manual curation or clustering of the *de novo* detected elements before mining the genome. However, *post hoc* we manually analyzed all elements which were previously classified into class I retrotransposon or class II DNA transposon elements or unclassified elements with detected coding element features (similarity to known transposable elements) due to potential chimeric insertion. From derivative elements (LARD, TRIM, MITE), “potential Hostgene” or unclassified elements (noCat), only elements were analyzed if they were present with more than one copy, occupied more than one Mbp or carried another detected feature, such as potential chimeric TE inserts or had similarity to known protein domains (PFAM). Manual inspection was done with ORF Finder (NCBI), CDD search (NCBI) (*102*), with a search in the most up to date online RepBase database (accessed December 2012-March 2014) via CENSOR (*103*) and phylogenetic analysis for LINE RT domains with RTclass1 (*104*) in order to achieve a detailed classification for each element, determine its potential relation to a family of known elements, to evaluate the completeness, to detect potential active elements or find other similarities to known sequences (NCBI Blast against nucleotide collection). We defined an element to be complete, if it possessed the relevant coding parts with the element-typical domains and the structural features (LTR, TIR). If an intact ORF seemed to cover a complete region including the typical domains (e.g. GAG as well as POL, Tase) then the element is considered to potentially active. If a Tase domain is covered by a truncated ORF or the Tase itself appears to be truncated but is covered by an intact ORF, or if the RT domain is covered by an active ORF but not the remaining element-typical domains, then the element is considered to be maybe potentially active. During the manual classification to at least superfamily level, novel transposable element types not covered by the system of Wicker et al. (*101*) were also considered: *Kolobok*, *Sola*, *Chapaev*, *Ginger*, *Academ*, *Novosib* and *ISL2EU* class II DNA transposons (*105, 106*).

Simple sequence repeats, satellites and other low complexity regions were extracted from the REPET pipeline database and processed with a custom Perl script to calculate the total coverage of these types of repetitive DNA, whereby overlaps with annotated TEs were excluded.

We tested the relationship between social complexity and repetitive elements while accounting for phylogenetic structure, using linearized models following phylogenetic independent constrasting with the R package ape (*69*). The analysis relies on the phylogenetic tree, ((Hlab:91, ((Mqua:68, (Bimp:13, Bter:13):55):10, (Emex:62, (Amel:19, Aflo:19):43):16):13):15, Mrot:106):9, (Dnov:85, Lalb:85):30) (based on ref (*28*)). The leaves are assigned sociality values of 0 for solitary species, 1 for facultative basic eusocial, 2 for obligate basic eusocial, and 3 for complex eusocial. The PICs of the sociality labels are then calculated.

Supplementary Text

Genome sequencing, assembly, annotation, and orthology

H. laboriosa, E. mexicana, M. quadrifasciata, D. novaeangliae

Assembly sizes of these 10 bees vary largely between 230 Mb to 370 Mb except *E. mexicana* whose genome size is 1.03 Gbp (Table S3). These are close to genome size estimates based on flow cytometry (Table S3). *E. mexicana*'s large genome size is due to numerous sections of repeat sequence (64.77%). We used homology-based, *de novo* method and transcriptomes to annotate coding genes. All 10 bees have about 13,000 coding genes and show similar gene features with average gene length of about 10kb; average CDS length of about 1.5kb; about 6 exons per gene; average exon length of about 250bp and average intron length of about 1.6kb. About 70% of genes could be aligned to Inter-Pro, KEGG and Swiss-Prot database in each bee. CEGMA evaluation result indicated these 10 bees' protein-coding gene annotations are of high quality (98.8%~99.6% complete CEGMA genes).

M. rotundata

We recovered a MAKER standard build of 11,187 genes which was used for further manual annotation.

Re-annotation of 4 other bees

The re-annotation results of 4 other bees showed their gene numbers varied from 12,684 to 15,810. The details of gene features are shown in Tables S4 and S5.

Orthology

OrthoDB clustering and mapping was used to evaluate the relative completeness of each species gene sets. The re-annotation of each genome improved the gene sets. More orthologs were identified and the number of potentially missing orthologs was reduced (Fig. S1). The re-annotations predicted more genes for which no orthology could be determined (Fig. S2).

Complete OrthoDB orthology delineation across 26 insect species including 19 hymenopterans identified 21,785 orthologous groups and classified 75.5% of all genes, 78.2% of all hymenopteran genes, and 82.4% of all bee genes (Fig. S3). A conserved core of about 5,500 genes in each bee species (35%-47%) have identifiable orthologs across the representatives of the insect phylogeny and a further ~1,200 to ~2,800 genes in each bee species (10%-21%) show no detectable orthology beyond Hymenoptera.

Transcription factor characterization

All ten bees have a remarkably similar repertoire of genes identified as TFs (Table S6).

Transcription factor binding site analysis

Motif Selection and Scanning

For 34 of the 223 representative motifs, we were unable to find their corresponding fly TFs in any orthogroup from OrthoDB. These motifs are assumed to not be conserved and are eliminated from further analyses (Additional Data table S1). Most of these 34 motifs with missing orthology were from the basic helix-loop-helix and zinc finger protein domain families (Table S7). The collection of motifs from FlyFactorSurvey is focused on these two TF domain families and is more likely to contain *Drosophila* specific TFs with those domains. Most (57%) of the identified TF orthogroups showed perfect copy number conservation across the 10 bees (52% when *Drosophila* is included) (Additional Data table S1). The five orthogroups that showed high copy number variation (variance > 3) had an expansion in either of the facultative basic eusocial bees, *L. albipes* or *E. mexicana*.

PIC Analysis

A significant correlation (p-value less than or equal to 0.01) was found between the sociality PICs and motif PICs for 2,101 (motif, orthogroup) pairs (Additional Data table S1). According to the plot of “eFDR” generated from random controls (Fig. S4), this p-value threshold of 0.01 corresponds to an estimated false discovery rate of 24%. 1,865 (89%) of the significant correlations were positive, meaning that higher social complexity corresponded to higher motif presence. Only 236 (11%) of the significant correlations were negative with higher social complexity corresponding to lower motif presence. We also summarize the significant results by motif by counting the number of orthogroups that were significant for the motif and the sign of the correlation (Additional Data table S1, Fig. S5). 17 of the 189 motifs had significant correlations with sociality for at least 20 orthogroups. 121 (64%) of the representative motifs had at least three times as many orthogroups with positive significant correlations to sociality as orthogroups with negative significant correlations to sociality. This bias of the data for motif presence to positively correlate with sociality is absent in the random control (Fig. S6). 84 orthogroups had at least three motifs that were correlated with sociality in the real data (Additional Data table S1), an observation that was made for only one orthogroup in the negative controls.

Cis-regulatory Analysis of Molecular Evolution Orthogroup Sets

For several motifs, there was an association between the orthogroups where the complexity of sociality correlated with the level of the motif and the orthogroups identified by the PAML-based molecular evolution analysis on both the origins and elaborations of eusociality. The significant associations according to the criteria of an association p-value less than 0.005 or an eFDR less than 0.3 are listed in Table S8. For the motifs for the transcription factors *Longitudinals lacking* and *Broad*, the associations are with the molecular evolution sets that capture the origins of eusociality within bees. 11 motifs for TFs including *Ultraspiracle*, *Taiman*, and *Tramtrack* have a significant association with an elaboration of eusociality gene set. Most of the significant associations were found with the accelerated elaboration genes identified in the stingless bees. These associations identify when the coding and non-coding sequence are evolving in a way related to sociality. We examine these significant associations more closely to investigate whether there is a pattern between the rate of coding sequence evolution and the presence of the motif with respect to sociality. As an example, we find that generally genes that are more conserved in species with increased social complexity are also more likely to contain high strength of the *Ultraspiracle* motif in species with higher social (Fig. S7). When we consider all of our significant association results, we find this pattern of high conservation and high motif presence with respect to sociality is dominant (Table S9).

Characterization of CpG distribution

Analysis of CpG dinucleotide depletion, an evolutionary signature of DNA methylation resulting from the mutability of methylated cytosines (*65, 66*), indicates that DNA methylation is likely to be present in at least nine, and possibly all ten, of the bee genomes (Fig. S8C and S9). In particular, normalized CpG values (CpG o/e) were much lower for coding sequences than for intergenic regions (Fig. S9) – consistent with the targeting of DNA methylation to translated exons, as observed in other investigated insects (*107*). We assigned genes a status of methylated, unmethylated, or undetermined based on the mixture distribution of coding sequence CpG o/e in each species. The mean number of predicted methylated genes per species was 1996.2 (SD: 576.4), and among 10 bee species 1,188 orthologous genes were predicted as methylated in 8 or more species (Table S10). *L. albipes* was the only species that deviated substantially in predicted number of methylated genes, with only 531 (2.6 SDs below the mean; Figs. 2B and Fig. S10), suggesting that the methylation system in *L. albipes* may differ greatly from those in the other bees. This likely reflects a change in DNA methylation status in *L. albipes,* because comparative studies (*8*) suggest the pattern observed in the other 9 bee species is representative of the ancestral state in bees. Alternatively, it is possible that decreased overall methylation is characteristic of the early stages of eusociality. We observed a strong ranked correlation between the level of sociality and predicted number of methylated genes (rho = 0.76, P = 0.0103; rho = 0.64, P = 0.0610 phylogenetically corrected), suggesting that more highly social bee taxa possess a greater number of methylated genes (Fig. S11). The finding of a link between DNA methylation and sociality is consistent with increased regulatory potential in social bees. Indeed, previous research with bees has identified a role for methylation in behavioral and reproductive plasticity (*108-111*). Further investigation is required to determine how life-history factors and population demographics linked with sociality lead to our observed association between number of methylated genes and sociality, and whether it plays a role in the regulation of phenotypic plasticity. Investigating methylation patterns in additional facultatively eusocial species will reveal whether decreased methylation is a general feature of the evolutionary origins of eusociality.

Phylogenetic covariance analysis of GC-corrected dN/dS across ten bees

Within each species, the highest ω genes are generally at low GC12, indicating that GC content could be mimicking the effect of selection by leading to increased ω even without changes in selection (Figs. S12-S14). A regression of log-transformed omega on GC12 and GC3 and their interaction was done separately for each species. In each case, all terms are highly significant. The total amount of variance explained by the regression (R-squared) was significantly correlated with level of sociality (r=0.73, p=0.016, Spearman rank correlation test). Thus structuring of the genome by causes related to sociality increases the tightness of the relationship between ω and gene GC content. The very strong relationship between GC12, GC3 and ω is consistent with an effect of bGC on ω, modulated by GC.

Coevol calculates the equilibrium GC (GC*), at which biased mutation and gBGC balance(*112*), based on the number of GC→AT and AT→GC mutations in each lineage. If biased mutation ratio is a constant over all species, as it is determined largely by DNA biochemistry, then biased gene conversion due to high recombination rates is the primary factor producing different GC* values. It follows that GC* can be used as a surrogate for the rate of gBGC in the vicinity of a gene. Plots of GC* for solitary or facultatively eusocial bees were unimodal, while those for obligately eusocial species were strongly left-skewed or bimodal, indicating that patterns of gBGC are associated with social complexity (Fig. S15). Mean GC* is strongly correlated with both GC3 and GC12 within a species (not shown). To understand the relationship between GC* and ω, we plotted the non-parametrically smoothed (via the loess method) values of ω in each species against (a) GC12 and (b) GC* (Fig. S16). The non-linear relationships between ω and GC12 and GC* ω could be estimated from a joint quadratic fit of these two GC measures. We used a square root transformation to normalize variation in ω. This regression dramatically improved the proportion of variance (R^2) in transformed ω explained, which ranged from 12% for *Dufourea novaeangliae* to 49% for *Lasioglossum albipes* (average: 30%). All regression coefficients were significant, and most (90%) were significant with p<0.000001. We calculated the residual values of transformed ω from the above regression for each gene in each species, then normalized the residuals to a unit variance in each species (producing a z-score). These "corrected ω (dN/dS)" values remove the strong effects of GC12 and GC* within and among species, and are centered around zero (Fig. S16). Values of corrected ω significantly larger than zero should, if most genes experience neutral or slightly deleterious selection, correspond to positive selection, while negative values should correspond to purifying selection. The exceptions to this pattern would occur if most genes were subject to strongly negative or to strongly positive selection. However, Harpur et al. 2014 (*11*) found that the distribution of selection in honey bee *Apis mellifera* meets the assumptions of the nearly-neutral model (Figure 1, op.cit.).

If correcting ω for GC truly removes neutral effects, we should find that corrected ω correlates better with measures of selection that use more information about selection. One such measure is the McDonald-Kreitman test (*113*), which in addition to divergence information also uses polymorphism information to produce what many authors believe to be a more robust estimate of selection effects (*114*). This can be summarized by the estimated population-scaled selection coefficient γ (gamma).

Harpur et. al. (*11*) used divergence and polymorphism information for *Apis mellifera* to estimate γ for most honey bee genes. We correlated γ from their work with uncorrected ω and GC-corrected ω. Although both correlations were positive and highly significant, the correlation of GC-corrected ω and γ explained 76% more of the variance in γ than did uncorrected ω. This strongly suggests that much of the variation in ω removed by GC-correction is poorly related to selection.

We then calculated the Spearman's rank correlation of GC-corrected dN/dS with social complexity. We used this regression coefficient, its significance (p and q values), and the GC-corrected ω values to identify genes under various selection regimes. These values, and their membership in various gene lists are reported in Additional Data table S3.

Positive selection:

1. Regression between GC-corrected ω and social complexity is positive and significant ($p < 0.05$ and $q < 0.1$)
2. The difference in the average GC-corrected ω of complex eusocial species is > 2 standard deviations (1.87) of the average GC-corrected ω of solitary species.
3. The average GC-corrected ω of complex eusocial species is positive.

Relaxed selection:

1. Regression between GC-corrected ω and social complexity is positive and significant ($p < 0.05$ and $q < 0.1$)
2. The difference in the average GC-corrected ω of complex eusocial species is > half of one standard deviation (0.467) of the average GC-corrected ω of solitary species.
3. The average GC-corrected ω of complex eusocial species is negative, indicating constrained evolution.

Purifying selection:

1. Regression between GC-corrected ω and social complexity is negative and significant ($p < 0.05$ and $q < 0.1$)
2. The difference in the average GC-corrected ω of solitary species is > 2 standard deviations (1.87) of the average GC-corrected ω of complex eusocial species.
3. The average GC-corrected ω of complex eusocial species is negative, indicating constrained evolution.

We used *D. melanogaster* orthologs of the honeybee gene in each orthogroup that fit these criteria in a GO enrichment analysis in DAVID (Tables S11-S13).

PAML analysis of dN/dS within clades

Genes with dN/dS significantly consistent with social complexity for each of the four clade-specific analyses are listed in Additional Data table S5 and are summarized in Table S14. There were no genes shared between all four analyses (Fig. 2D). The number of genes overlapping between the two origins of eusociality was not significant (6 shared genes, hypergeometric test p = 0.82). The number of genes overlapping between the two elaborations of eusociality was also not significant (43 shared genes, hypergeometric test p = 0.70). Some of these genes showed overlap with genes identified as undergoing increasing purifying, relaxed, or positive selection in association with increasing social complexity (Fig. S18).

GO terms enriched in each clade-specific analysis are listed in Additional Data table S6. Genes related to signal transduction were common to analyses involving the origins of eusociality in Apidae and Halictidae, but there were no shared GO terms among the analyses involving elaborations of eusociality in honeybees and stingless bees (Fig. S19).

Convergent evolution analysis

Two orthogroups (APO012406 and APO010594) showed signatures of genomic convergence associated with social complexity (Fig. S21). Within these orthogroups, genes of eusocial species cluster separately from solitary bees. *D. melanogaster* orthologs of the *A. mellifera* member of these orthogroups are FBgn0036538 (CG15715) and FBgn0037121 (Rpb8), respectively. These two orthogroups were not found to be rapidly evolving in either the clade-specific PAML analyses or the coevol analysis. This suggests the signature of convergence is not strong for eusocial evolution in bees.

Gene annotations for the Major Royal Jelly genomic region and signaling pathways functionally involved in caste development of bees

mrjp/yellow genes

The *mrjp/mrjp-like* genes and the *yellow* genes flanking these, *y-e3* and *y-h*, respectively (*115*) were first analyzed individually for each species and then as the entire genomic segment which they form. A general conclusion drawn from the multiple alignments and phylogenetic tree analysis for the *mrjp*, *y-e3* and *y-h* genes was that only the genus *Apis* has an expanded *mrjp* gene set, with all other bee species having an *mrjp-like* gene that is most similar to *A. mellifera mrjp9* (Fig. S22). This finding is in line with the recent work by Buttstedt et al. 2013 (*115*). The *y-h* and y-e3 genes form distinct clusters. Only three *yellow* gene models were outliers: one for *Eufriesea mexicana*, where the *y-e3* gene had been split into two gene models and one was now correctly placed, a second *Apis florea yellow* gene was correctly assigned; and an additional *yellow* gene was found for *Dufourea novaeangliae*.

The nomenclature for the *A. melífera/A. florea mrjp* genes will still need some work, as the numbers do not necessarily reflect orthology relationships in the current tree (Fig. S22). Nonetheless, the nine *mrjp* genes for these two species and their order in the genomic region had already been defined (*115*).

As far as the individual gene models go, we made the following corrections: For *Apis mellifera*, all nine *mrjp* CDS regions and their flanking yellow genes were validated, but we noted a clear mistake in the Official Gene set prediction for *mrjp5* (GB55208). The first exons were wrong and the last large exon containing the MRD repeats was missing; actually it was predicted as part of a gene model with a separate GB number (GB55208); for the *mrjp5* gene, the NCBI prediction (NP_001011599.1) appeared to be much more appropriate than the OGS prediction. We also solved minor problems in the *mrjp2* and *y-h* predictions in the GB database (GB55212 and GB55216, respectively), primarily related to predictions of the first exons. These gene models were corrected.

Since the *Apis florea* genome was annotated based on *Apis mellifera*, we solved a similar prediction problem for the *Apis florea mrjp5* gene, which was also corrected.

For *Bombus impatiens* and *B. terrestris* the predictions were all validated. Each species has one *mrjp9-like* gene inserted between a *y-e3* and a *y-h* ortholog.

For *D. novaengliae* the *mrjp* gene prediction was overestimated, in fact it was a fusion of a *y-h* and *mrjp* gene. Consequently, we split this prediction into the two gene models. Within the genomic scaffold that encompasses the y-e3 to y-h region we found an additional gene, provisionally named *y-like*, as we could not establish a clear homology relationship with *y-e3, y-h* or *mrjp9-like.*

For *Eufriesea mexicana* the problems with the gene models in the database required similar corrections as for *Dufourea.*

For *Habropoda laboriosa,* a *y-e3* and an *mrjp9-like* gene were identified, the Hlab predictions were validated.

For *Lasioglossum albipes* the *yellow* gene models and the one for an *mrjp9-like* gene were corrected. A second *mrjp-like* gene was found in the scaffold, indicating a possible duplication event.

For *Melipona quadrifasciata* one of the *yellow* gene models (*y-e3*) was correct, but for the *mrjp* and *y-h* genes these had been fused in a single gene model. For *y-h* we could now define a complete gene model, but the one for *mrjp* remained incomplete. Specifically, the first exon for Mqua *mrjp* could not be completed as it ran into the *y-h* gene model. Apparently, something is missing in the genome assembly for this scaffold. We know that a complete *mrjp* transcript exists for *M. quadrifasciata;* it has recently been found and sequenced by Stefan Albert (unpublished).

For *Megachile rotundata,* one of the *yellow* gene models (*y-e3*) was correct, but for the *mrjp* and *y-h* genes this was also a single, fused gene model. We could define complete gene models for both.

For an overview we compiled the basic data on predicted protein size and exon number (Table S16), as well as on the size of the genomic region comprising the *mrjp* genes and their flanking *yellow* genes (Table S17). For more details, all revised gene models are provided as gff files, with a NOTE-field that informs whether and which corrections were made to the respective original gene model**.**

With respect to the *yellow* genes, *y-h* is generally larger than *y-e3* (~550 aa vs. 400-430 aa), but has less exons, four for *y-h* and five for *y-e3*. Exceptions are *E. mexicana y-h* predicted to encode a 730 aa protein, and y-e3 of *D. novaeangliae* with only 330 aa and four exons instead of five. With respect to the *mrjp* genes, the non-*Apis* bees had generally only one *mrjp-like* gene, and practically all were more similar do *A. mellifer*a *mrjp9* than to the others *mrjp* genes, providing further support for the hypothesis that *mrjp9* is the ancentral *mrjp* gene (*115, 116*). Interestingly, predicted protein size for the MRJPs of the ten bee species were all very close to 410 amino acids, except for *M, quadrifasciata* and *M. rotundata*, with only 288 predicted amino acid residues, but at least for *M. quadrifasciata* we know that the gene prediction is incomplete due to a probably missing region in the genome assembly. It was only for *L. albipes* that we identified a second *mrjp-like* gene in the scaffold containing the *y-e3, mrjp9-like* and *y-h* genes.

When the entire genomic regions were compiled, starting with the translation start site of *y-e3* and ending with the stop codon of *y-h,* the two *Apis* species had of course the largest genomic region, due to the *mrjp* gene duplications. For the two *Bombus* species, the genomic region had a surprisingly conserved size, differing by only 16 bp across over 11 kb. In contrast, for the two halictids, the genomic region in *L. albipes* was two-fold larger than in *D. novaeangliae*, and in the megachilid *M. rotundata* the region was the shortest, less than half the size compared to that of the euglossine *E. mexicana*, which, among the species analyzed herein, is most basal within the clade Apoidea (*117*). Differences in the genomic region covering the gene family may indicate differences in regulatory complexity in the respective *cis*-regulatory regions, or be due to insertions of repetitive sequences. In any case, considering the importance of the yellow genes in color variation of insects, and the specific role of MRJPs in the social life history, particularly of the genus *Apis*, this genomic region holds potential for revealing traces of evolutionary processes in the genomes of bees. It still remains to be shown whether and how this relates to social evolution. For instance, *M. rotundata* and *H. laboriosa,* which are both solitary living and represent the two more basal species based on the bee tree of life (*117*), differ by a factor of four in the size of this genomic region, whereas in the halictids, the genomic region is twice the size in the facultatively social *L. albipes,* compared to the solitary *D. novaeangliae*. The larger size of the genomic region in *L. albipes* coincides with an extra *mrjp-like* gene that we putatively identified in this genomic scaffold. As pseudogenization apparently happened in the MRJPs encoding region of the genus *Apis (115)*, duplication events in this region may also have occurred in other bee species, especially those that exhibit a larger than expected genomic region for an ancestral *mrjp* and its flanking *yellow* genes. The entire region comprising the *yellow* genes and a yellow-derived *mrjp* ancestral gene could be a region of gene family expansion, where paralogs may have undergone specialization or were gradually lost by pseudogenization.

Developmental genes related to caste development

We included in this analysis a set of 12 genes knowingly involved or likely to play a prominent role in caste development in honey bees. We considered that these genes may reveal molecular signatures in relation to the evolution of sociality. For all these genes, the experimentally validated or predicted gene models were retrieved by EggNOG and tblastn searches against the ten bee genomes, and all were manually annotated using the Artemis platform. An overview on CDS size and exon number for each gene is listed in Table S18. For more details, all revised gene models are provided as gff files, with a NOTE-field that informs whether and which corrections were made to the respective original gene model.

The gene with the highest degree of conservation with respect to CDS size (1343-1460) and exon number (11-12) is *egfr*, and interestingly, the Egfr signaling pathway is the one shown to play a prominent role in honey bee caste development, translating the royalactin signal and affecting the JH titer (*118*). At first sight this could mean that the potential for dimorphic development would be deeply embedded in the bees, but for a more comprehensive understanding this would require an analysis of synonymous versus non-synonymous substitutions in this gene.

A second highly conserved gene is *tor*, which also has been shown to play a functionally important role in honey bee caste development (*119, 120*). It is the gene with the largest CDS (2308-2476) among all the developmental genes analyzed herein. The phylogenetic tree for the deduced TOR protein (Fig. S17) is also the one that most closely reflects the bee tree of life (*117*), with the megachilid *M. rotundata* and the antophorid *H. laboriosa* being the most basal species, and the two halictids, *D. novaeangliae* and *L. albipes* clustering together. Among the Apidae, *E. mexicana* TOR is at the most basal position, and TOR of the highly eusocial stingless bee *M. quadrifasciata* is more closely related to TOR of the two primitively eusocial bumble bee species than to honey bee TOR.

Another highly conserved gene is *usp*, with a CDS size varying only slightly in most species (424-438 aa) and six exons, the only exception being *M. quadrifasciata* USP (373 aa and five exons). As the latter figure is based on experimental evidence (*121*), this could have a meaning for regulatory pathways that involve *usp* function at the interface between ecdysone and JH signaling in bees (*122*).

The most variable genes in terms of CDS size and exon number are the HIF-alpha homolog *sima* and *met* (Table S18). HIF-alpha is the critical component in hypoxia signaling and the honey bee *sima* gene was shown to be significantly overexpressed in worker larvae, suggesting an endogenous hypoxia response related to worker caste development in these larvae (*123*). The annotation results even suggest the potential existence of isoforms, at least in *A. florea* and *D. novaeangliae*, this of course requiring experimental confirmation. The second gene, *met*, encodes a Methoprene-tolerant/Germ cell-expressed gene product, which is the functional JH receptor in insects (*124*). Together with the HIF-beta homolog *tango*, which turned out to be less variable than *sima*, these genes belong to the bHLH-PAS protein family, and problems in the prediction of gene models for this complex gene family may in part explain this variability. Nonetheless, the phylogenetic trees for both *sima* and *met* (Fig. S24) were in fairly close agreement with the bee tree of life (*117*).

For the IIS pathway, two receptor isoforms were consistently identified in all bee genomes, this indicating certain conservation within these gene families (Fig. S25). For most bee species, CDS size for the two insulin receptors varied between 1400 and 1700 amino acid residues and six (InR1) and 11 exons (InR2), respectively (Table S18). Exceptions were denoted in the gene models for InR1 of the two halictid species *D. novaeangliae* and *L. albipes*, with gene models smaller than average, and for *A. florea*, which had the shortest prediction amongst all insulin receptors for InR2. For *E. mexicana* there were three predictions for insulin receptors, two of which were identical in terms of CDS size and exon number (Emex01484 and Emex06662) and, not surprisingly, they also clustered very closely in the gene phylogenetic tree. This could indicate a recent duplication for the InR2 gene, or an assembly problem in the current version of the *E. mexicana* genome.

With respect to the ecdysone receptor, gene models for one gene with two possible splicing isoforms (EcR-A and EcR-B) could be built for all the ten species. The gene models were fairly conserved in terms of CDS size and exon numbers (Table S18).

Manual annotation of genes in the sex determination pathway

The conserved mechanism of sex determination in holometabolous insects is the *transformer* (*tra*)-*doublesex* (*dsx*) transduction module (*125*). We focused on the molecular evolution of *tra*, its ortholog *feminizer* (*fem*) and paralogous copies such as *complementary sex determiner (csd),* the primary signals of sex determination in *A. mellifera (126, 127)*. Duplications of *tra*/*fem* were observed in several bee lineages (Fig. S26 and Table S22); *tra* duplication has previously been detected in six out of seven ant species (*128*). Orthologous groups identified in OrthoDB7 provide further evidence for more than one *tra/fem* copy in ant and bee species (Sinv, Cflo, Hsa, Labl, see Table S19, S23-S24). Low divergence on synonymous sites among *fem* and its paralogs within non-*Apis* bees (Bter, Bimp, Lalb, Table S25) supports lineage specific, independent gene duplication events (*129*). Selective constraints of *fem* evolution vary when compared among bee lineages (Fig. S26). We found signs of episodic diversifying selection in non-*Apis* species (Fig.

S26). Experimental data obtained for the *fem* paralog in *Bombus* (Bter *fem1*) provide no evidence for a multi-allelic factor as found for *A. mellifera* ((*130*), Biewer and Hasselmann, unpublished).

We focused on the position of genes located in the sex determination locus of *A. mellifera (127)* to evaluate synteny in the other bee genomes. The physical association of the *A. mellifera* genes GB47018 and GB47045 is highly conserved in the orthologs among all bee species and is followed by the *fem*/GB47023 complex (Fig. S27). The *fem* paralog in *Apis* is located within a few kb of *fem*. In species for which a duplicate has been found the exact position is either on another chromosome (*Bombus*) or scaffold (*L .albipes*). RNAi treatment (*131*) showed no effects of GB47018 (synonymous to GB11211), GB47045 (GB13727) and GB47023 (GB30480) on sex determination in *A. mellifera.*

The DNA binding (OD1) and oligomerization (OD2) domains were found in all species (except *E. mexicana*) DSX with amino acid divergence reflecting their phylogenetic relationship (Fig. S28 and Additional Data table S8). The *tra2* gene containing a RNA-binding (RMM) domain (Fig. S29) is on average more diverged at the protein level between *Apis* and non-*Apis* species than outside the bees ($d_{RMM-A-nA}$=0.126+/-0.04 vs. $d_{outRMM-A-nA}$=0.059+/-0.019, Z-test=1.51, P<0.1), with significant higher divergence downstream ($d_{outRMMDown-A-nA}$ = 0.02 +/-0.02, Z-test =2.37, P<0.01) than upstream ($d_{outRMMUp-A-nA}$ = 0.074+/-0.02, Z-test =1.16, n.s.) relative to the RRM domain (Additional Data table S9). Furthermore, *Apis tra2* shows 21 of otherwise fixed amino acid differences compared to non-*Apis* species.

The core of the sex determining pathway in insects *fem(tra)*/*dsx* is conserved over > 250 million years of evolution (Diptera/Hymenoptera). The copy numbers of *fem* paralogs found in bee genomes (Table S19) reflect either lineage specific gene losses (in *Mqua, Mrot, Dnov* and *Hlab*) from a single ancestral duplication event or independent gene duplications (in *Apis, Bter, Bimp, Lalb*). More experimental data are needed to evaluate possible differences in mechanisms of sex determination in bee species with and without *fem* paralogs and the function of the conserved gene complex near the *fem* locus.

Gene family contraction and expansion

Model 9 was a significantly better fit to the data than Model 8, which had the next highest likelihood score (χ^2 = 22.36, p = 2.26E-06). Thus, there are four rates of gene family evolution among the bees. All internal branches evolved under lambda 2. These branches had the second slowest rate of gene family evolution. Terminal branches were divided among lambda 1-4, with lambda 4 being the slowest. Terminal branch lambdas tended to cluster based on social organization. Lambda 1 included Emex, Lalb, and Mqua. Emex and Lalb are the only two facultatively eusocial branches in the dataset, though Mqua has obligate complex eusociality. Lambda 2 included all the internal (i.e. ancestral) branches and the three solitary species in the dataset - Hlab, Mrot, and Dnov. Lambda 3 included Amel, Aflo, and Bimp. Amel and Aflo are obligate complex eusocial, and Bimp is obligate basic eusocial. Bter was the fastest evolving branch, and it is obligate basic eusocial. Branch-specific expansions and contractions are in Fig. S31.

The birth and death rates under model 9 are:

```
lambda 1 = 0.000798 (Emex, Lalb, Mqua, internal branches)
lambda 2 = 0.000239 (Hlab, Mrot, Dnov)
lambda 3 = 0.001818 (Aflo, Amel, Bimp)
lambda 4 = 0.001337 (Bter)
```

Lambda is higher on short terminal branches. This is not likely to be a branch length issue, however, because short internal branches had low lambda.

Gene families with family-wide significant gain/loss were enriched for terms related to olfaction, signal transduction, tachykinan receptor activity, and metabolism – specifically lipid metabolism, (Table S26). Closer inspection of the gene families contributing to the enrichment of odorant binding and lipid metabolism revealed some of these gene families were expanding in one or more eusocial species, as compared to solitary species (Figs. S32-S35).

Manual annotation of neural plasticity genes (biogenic amine receptors, neuropeptides, and GPCRs)

Biogenic amine GPCRs

Insects have 18-22 biogenic amine GPCR genes (*132*, *133*). We have earlier found that *Apis mellifera* contains 20 of them (*132*). When comparing this set from the honey bee with the biogenic amine GPCR genes from other insects, we found one duplication in the honey bee of an octopamine GPCR gene (Am1, see Fig. 3 of ref (*132*)), which was not present in the other insects. Originally, we thought that this finding was interesting and that it perhaps might be related to bee sociality. However, we now find that the same biogenic amine GPCR gene set, including that specific octopamine gene duplication (Am1), is present in all bees, including the non-social bees (Table S27). So Am1 might be bee-specific (it is not present in ants or *Nasonia*), but it is probably not involved in sociality.

Neuropeptides and protein hormones and their GPCRs

Insects have about 30-40 neuropeptide and protein hormone genes and 40-60 neuropeptide/protein hormone GPCR genes (*132, 133*). About 75% of these GPCR genes have been deorphanized in *Drosophila melanogaster* (i.e., the GPCRs have been matched with their ligands). We have previously identified and compared the neuropeptide, protein hormone, and their GPCR genes in a wide variety of insects and other arthropods with a sequenced genome (*134*). Our conclusions from this work are the following: (1) Insects have two sets of neuropeptide, protein hormone, and corresponding GPCR genes: The “core set” of about 20 GPCRs (and their ligands), which occur in each insect with a sequenced genome, and the “variable set” of about 30 GPCRs (and their ligands), which can either be present or absent; (2) The core set is probably related to basic physiological processes common to all insects, while the variable set might be responsible for specific features characteristic for each insect group or species; (3) the combined core and variable gene sets can be represented as a barcode for each sequenced insect species. We hypothesize that this barcode must be related to behavior and, of course, also to evolution.

Table S28 shows the neuropeptide and protein hormone GPCR barcode for the ten bees with a sequenced genome. Whenever a GPCR was present (highlighted in green), we found that its peptide or protein hormone ligands was also present and, vice versa, whenever a GPCR was absent (highlighted in yellow), its peptide or protein hormone ligand was also absent. Therefore, Table S28 represents the barcode for neuropeptide and protein hormone signaling in bees. We can see in Table S28 that the barcodes are identical for all bees independently from their sociality. The same is true for the orphan GPCRs (Table S29). There are, however, very few exceptions. One exception is the absence of kinin signaling in *Eufriesia mexicana*, which represents facultative simple eusocial life history. Another bee known to have facultative simple eusociality, *Lasioglossum albipes*, however, contains kinin signaling (Table S28). Another exception is the absence of sulfakinin in *Bombus terrestris* and *Bombus impatiens*, which both are simple eusocial. However other simple eusocial bees, such as *Lasioglossum albipes* and *Eufriesia mexicana* have sulfakinin signaling. Finally, trissin signaling is absent in *Apis mellifera* and *Apis florea*. However, other complex eusocial bees such as *Melipona quadrifasciata* have trissin signaling (Table S28). In conclusion, therefore, we find no correlation between absence or presence of neuropeptide and protein hormone genes (including GPCR genes) and the degree of sociality in bees.

Manual annotation of genes related to immunity

Our survey methods were largely concordant, although Proteinortho was somewhat more conservative in assigning proteins to orthology groups but more willing to assign multiple proteins to a single group, e.g., list paralogs. While our intent was to identify most likely orthologs for each candidate immune protein, rather than assess gain and loss of proteins, there were not striking differences in protein number for these groups across the ten species. Specifically, the relatively low number of canonical antimicrobial peptides, and classic recognition proteins (beta-glucan receptor proteins [BGRP] or synonymously known as gram-negative binding proteins [GNBP], and peptidoglycan recognition proteins [PGRP]) first observed in *Apis mellifera (135)*, seem to be a general trait of the Apoidea. *Lasioglossum albipes* appears to have one additional BGRP for a total of three (Lalb_11661,Lalb_11660,Lalb_05742). Both *Bombus* species and *Eufriesea mexicana* appear to have relatively scarce PGRPs. Exhaustive surveys of the bumblebees found additional sequences bringing *B. impatiens* to four PGRPs (Bimp13700, Bimp13150, Bimp13701, Bimp0970), *B. terrestris* to 3 (Bter08827,Bter03632,Bter07215), and *E. mexicana* to 2 (Emex03016, Emex11756). The gene predictions here did not predict the presence of STAT in *Megachile rotundata* but we were able to identify the ortholog to this signal transducer and key member of the JAK/STAT pathway in the NCBI RefSeq set for this species. Molecular-evolution comparisons showed differences in the degree of sequence conservation and the drive for amino-acid substitutions, indicating strong purifying selection for Argonaute 1 and three other proteins related to RNA interference (Figs. S36-S37).

Overall, copy numbers for immune proteins are relatively constant across the Apoidea, indicating selection for the maintenance of pathways and functions in this group. Key proteins in the RNAI, Jak/STAT, Toll, and Imd/Relish pathways are conserved across all surveyed species (Additional Data table S10). There were no striking differences in gene family size for immune proteins, suggesting that the Apoidea as a whole carry fewer immune-related proteins than do other Holometabola for which genomic data exist. A broad view of sequence evolution for immune-related proteins and processes can provide novel insights into which proteins in a family are indeed involved with immunity (Figs. S36-S37). For example, C-type lectins show a remarkable range from proteins apparently under positive selection to those under strongly purifying selection. Knowing the evolutionary rate of substitution in family members can help predict which are involved with tracking fast-evolving parasites and which are involved in long-stable processes. Similarly, four proteins in the RNAi pathway showed the highest levels of purifying selection, while others in this group did not show this pattern.

There were several missing short proteins (e.g., for the antimicrobial peptide Apidaecin) that might reflect assembly issues (especially for *Bombus terrestris*, where this protein almost certainly exists) but might also reflect novelty, e.g., in the case of Apidaecin an origin in the

Apini.

Manual annotation of cytochrome P450 monooxygenase genes

Annotated P450 genes for each species are listed in Additional Data table S12. All bees appear to share a reduction in the complement of cytochrome P450 monooxygenase genes encoded in the genome relative to other insect genomes, including other Hymenoptera. Enzymes in the P450 superfamily are important for synthesis and breakdown of pheromones, endogenous signaling compounds and for metabolizing xenobiotics, including plant allelochemicals and pesticides (*136*). Bee genomes encode just 41 to 58 putatively functional P450s (Table S30) while most other insect genomes encode between 59 (*Acyrthosiphon pisum*) (*137*) and 196 (*Culex quinquefasciatus*) (*138*) P450s. Only the genome of the louse, *Pediculus humanus*, encodes fewer P450s, just 37 (*139*). Much of the reduction in P450 gene diversity in bees occurs in the CYP4 clan. Bee genomes encode just 4 to 6 CYP4 genes while other insect genomes encode 23-57 CYP4 genes (*140*). The function of these missing P450s may be related to pheromone processing (*141*), wax production (*142*), or fatty acid metabolism (*143*) rather than dietary xenobiotics.

Accounting for P450s encoded in the genome does not necessarily indicate the number of functional P450s that are transcribed and translated into functional proteins in living bees. Genes that are identified as “putatively functional” based on conceptual translation may not have a genuine function in living bees – this may be particularly true of the highly dynamic CYP3 clan P450s associated with xenobiotic metabolism where a “birth and death” model of evolution is expected (*136*).

The reduction in P450s in bee genomes does not appear to be related to the haplodiploid system of sex determination as the genomes of other haplodiploid Hymenoptera encode between 72 (*Pogonomyrmex barbatus*) (*144*) and 111 (*Linepithema humile*) (*145*) P450s. As the current analysis makes clear, the bees’ reduction in P450s appears to be unrelated to an protection afforded social insects as the genomes of the social bees and ants do not encode fewer P450s than solitary bees or a solitary wasp (*Nasonia vitripennis*), with 92 P450s (*92*). The reduced P450 gene complement in bees is likely related to the relatively innocuous diet of nectar and pollen consumed by larval and adult bees. However, these foods may contain flavonoids and flavonoid derivatives, which honeybees are capable of metabolizing through CYP6AS subfamily P450s classified in the detoxicative CYP3 clan (*146*). Bee genomes encode between 6 and 17 CYP6AS P450s that may contribute to detoxication of dietary toxins.

Manual annotation of inotocin hormone system

Genome analyses reveal a pseudogene remnant in certain bee species but intact inotocin hormone systems in ants, wasps, sawflies, and termites. We found intact inotocin receptor (ITR) and inotocin genes in sawflies, wasps, ants, and termites (Fig. S38). By contrast, we were unable to locate any intact inotocin receptor genes or inotocin genes in the bee species we analyzed. We did, however, locate exons 2-5 in *Dufourea novaeangliae* and exons 2-4 in *Bombus terrestris* and *Bombus impatiens* (Fig. S41). To confirm that these sequences were indeed remnants of the pseudogenized inotocin receptor in bees, we mapped regions of microsynteny conserved between ants and bees near the locus with the remaining inotocin receptor exons (Fig. S39-S40). The *RGP1*, *loc 100643488*, and *DNA repair XRCC* genes upstream of the inotocin receptor (*ITR*) pseudogene remnants in the bee species mentioned were also microsyntenic with *ITR* in *Pogonomyrmex barbatus*, *Camponotus floridanus*, and *Harpegnathos saltator* (Fig. S39 and

S40). In *Solenopsis invicta, Acromyrmex echinatior,* and *Linepithema humile*, *RGP1* and *loc 100643488* were the only genes with any shared microsynteny near the *ITR* remnants in bees (Fig. S39 and S40). The gene *loc 100643488* was also microsyntenic in the wasp species *Polistes dominula* (Fig. S39 and S40). There were no microsyntenic genes found in the wasp species *Diachasma alloeum* and *Nasonia vitripensis*.

Repetitive elements in the genomes of ten bee species

The analyses of repetitive sequences in 10 bee genomes (*Apis mellifera, A. florea*, *Bombus terrestris*, *B. impatiens*, *Eufriesea mexicana*, *Melipona quadrifasciata*, *Habropoda laboriosa*, *Megachile rotundata*, *Lasioglossum albipes* and *Dufourea novaeangliae*) showed the presence of elements across the known diversity of transposable and other repetitive elements, yet with large differences between the species.

The number of elements of *de novo* detected interspersed repeats ranges from 190 to 3315 covering between 3.61 to 26.63 % of the analyzed genome assemblies (Tables S31-S32 and Fig. 2E). These elements are comprised of retrotransposons (class I, 0.1 – 4.61 %), DNA transposons (class II, 0.57 – 7.13 %), novel/unknown elements (0 – 13.13 %) as well as their derivatives (2.7 – 20.18 %) (Table S32).

Classified Retrotransposons of the LTR and LINE type are the most frequent retrotransposable elements in the bee genomes. *Copia* elements were almost absent (0.02 – 0.05 %) in the highly eusocial species Amel, Aflo and Mqua, more common (0.1 – 0.3 %) in the primitively or facultatively eusocial Bter, Bimp and Lalb, and most frequent (0.45 – 1.4 %) in the solitary Emex, Mrot and Dnov. *Gypsy* elements show a similar pattern with absence in Amel or infrequent (0.01 – 0.15 %) occurrence in Aflo, Mqua, but also Emex, and higher in the other species, especially the bumblebees which have been mainly invaded by two different Gypsy elements. *BelPao* elements were absent in Amel, infrequent (<0.2%) in Emex, Aflo, and Bter. Retroviruses of *Gypsy*-like Errantiviridae were scarce (0.01 – 0.06 %) and only found in Bter, Bimp in which also Gypsy elements proliferated (Fig. S42).

Of the nonLTR retroelements (LINE) *Jockey* and *I* are most common (up to 1.06 and 1.72 %), followed by *R2* and *RTE*. Except *R2*, LINEs are absent in Amel and Aflo. Similarly Mqua showed very low amounts of LINEs. Other retroid elements (SINE, DIRS; PLE) were scarce in all species (Fig. S43 and Tables S31-S32). A very large fraction of the genome is represented by retroid elements classified as LARD or TRIM, retrotransposon derivatives without a coding region, accounting for 2.7 – 13.95 % (Fig. 2E and Table S32). Some of the TRIM elements reached high copy numbers, particularly in Emex.

Of Class II DNA transposons, almost all superfamilies could be detected in the bee genomes, whereby Helitron and Polinton elements were scarce. The majority of the DNA transposons belong to the TIR types. Commonly most frequent are *Tc1*/*Mariner*, followed by *PiggyBac* and *hAT* in some of the genomes (Table S32). The diversity is lowest in Amel and Aflo which only have elements of 2 superfamilies, followed by Emex, Mqua, Bter, Bimp and Hlab with 4-6, and Mrot, Lalb and Dnov with 10-12 (Table S31). Derivatives of TIR-DNA-Transposon (MITE) could be detected in most genomes, whereby they were absent or almost absent in Aflo, Amel and Mqua (Fig. S44).

Besides the well classified sequences, numerous elements could not be assigned to a superfamily or even class. The latter contains a larger number of elements (0 – 13.12 % of the genome) which could represent novel types. Not categorized elements or detected elements which contained no typical transposable element feature, but profiles from protein coding genes were separately annotated and don't belong to the interspersed DNA as they are likely to represent host genome sequences detected due to their repetitive characteristics, such as common protein domains, similar members of gene families, and duplicated genes. In fact genes of Ankyrin repeat-containing proteins, csd, p450, 28S rRNA, odorant receptors and major royal jelly proteins were frequently found as repetitive sequences. Both groups together comprise between 0.82 to 17.88 % of the genome (Table S32). Emex has by far the highest amounts of these elements due to unusually high copy numbers of a few long not categorized elements. These do neither show any typical features of transposable elements, nor any sequence similarity to known sequences.

Detailed analyses of the detected elements proved numerous Retro- and DNA transposons to be complete in their structure, to be potentially active or partly active, or to contain a RT/Tase domain. Typically, those elements which were found to be potentially active elements were higher in copy number and appeared to be present as insert in other repetitive elements more frequently (data not shown).

The precise family relationships of the well classified or other elements in the analyzed bee genomes are not fully resolved yet, but several elements appear to be shared between certain species as indicated by high similarity to identical elements in RepBase (RepBase 17.01) (*44*) (data not shown). Among different bee species, several DNA transposons of the *Mariner*, *PiggyBac* and *hAT* superfamilies, but also their derivatives (MITE) were found to be present in at least two bee species. The majority of the *Mariner* elements belong to the large group of *Mariner-1_Tbel* families which are present in many organisms (*147*, *148*). Shared retroid elements belong to the *R2* LINEs and *Gypsy* LTR, but also *Copia* LTR and *RTE* (LINE). In general they were found to be more infrequent. In addition to elements shared among bee species genomes, numerous elements of the above mentioned types as well as 5S RNA related repeats, Kolobok (TIR) and CACTA (TIR) appeared to have high sequence similarity to known elements (RepBase) from distant organisms, including ants (*S. invicta*, *L. humile*), jewel wasp (*N. vitripennis*). flies (*Drosophila* spp., *Ceratitis rosa*, *C. amoena*, *Trirhithrum coffeae*), flour beetle (*Tribolium castaneum*), silkmoth (*Bombyx mori*), pea aphid (*Acyrthosiphon pisum)*, earwing (*Forficula auricularia*), but also the flatworm (*Schmidtea mediterranea*), Cnidaria (*Hydra magnipapillata*), endoparasitic nematode (*Heterohabditis bacteriophora*, bat (*Myotis*), frog (*Xenopus tropicalis*), the anole (*Anolis carolinensis*), hyrax (*Procavia capensis*), cow (*Bos taurus*), the primates *Microcebus murinus* and human, and a coelacanth (*Latimeria chalumnae*).

We did not perform a particular scan for known viruses, but while inspecting the transposable element sequences, some conserved protein domains like viral helicase, ANK and PRANC or sequences similar to Baculoviridae, Herpesviridae, Bracovirus (*Cotesia*), Megavirus, Caulimoviridae, Chordopoxviridae and Poxviridae were found (data not shown). Furthermore we detected repetitive sequences which show very high sequence similarity to *Wolbachia* endosymbionts, particularly in the genome of Dnov and in low amounts in Mrot.

The proportion of the genome containing repetitive elements decreases with increasing social complexity. Social complexity is a significant predictor of total repeat content of the genome, independent of phylogeny (glm with phylogenetic independent contrasts, F = 8.99, adjusted R^2 = 0.47, p = 0.017).

Author Contributions
K.M.K. and H.P. are joint first authors. S.L.S., D.P., H.M.R., M.E.H., B.J.F., and A.H. sequenced and assembled the *M. rotundata* genome. M.Y., D.E., C.H., and B.J.F. annotated the

M. rotundata genome. G.D.Y., W.P.K., and J.B. sequenced the *M. rotundata* transcriptome. H.P., C.L., and G.Z. sequenced and assembled genomes for *H. laboriosa*, *D. novaeangliae*, *M. quadrifasciata*, and *E. mexicana*. H.P., C.L., and G.Z. annotated the ten genomes. R.M.W. and E.M.Z. provided OrthoDB support, assessed the quality of the annotations, and identified orthologous genes across the ten species. K.M.K. characterized TFs within each genome. E.S., F.B.K., S.H., and R.F.A.M. analysed transposable elements. K.M.G., B.G.H., and M.A.D.G. analysed CpG distributions. F.H. and C.J.P.G. performed manual annotation of genes related to neural plasticity. D.G.P., F.M.F.N., M.P.M.S., E.D.T., Z.L.P.S., and K.H. performed manual annotation of genes related to development. J.D.E. and S.M.B. performed manual annotation of immunity genes. R.M.J. and B.L. performed manual annotation of cytochrome P450 genes. J.H.M., H.M.R., and B.R.S. annotated genes in the inotocin hormone system. M.H., D.H., and M.B. analysed sex-determination genes. K.M.K. and H.P. performed gene family contraction and expansion analysis. C.F.K., H.P., C.L., K.M.K., A.Z., and G.Z. performed molecular evolution analyses of protein coding genes. C.B., K.M.K., and S.S. performed the TFBS analysis. J.S.J. and S.J.H. analysed genome sizes. S.D.K., C.L., and G.Z. sequenced and provided access to the *L. albipes* genome. J.W., G.E.R., and G.Z. initiated the project. K.M.K, G.E.R., and G.Z. designed the study. K.M.K. and G.E.R. wrote the paper with contributions from all the authors.

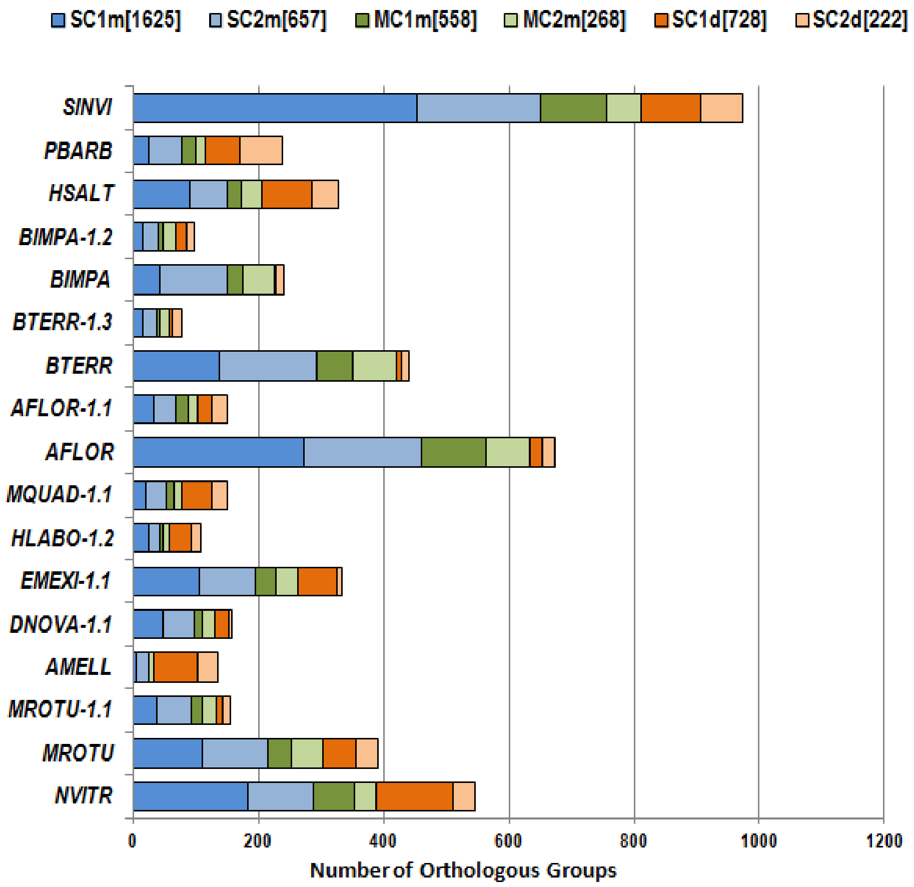

Fig. S1.

Counts of near-universal Hymenoptera OrthoDB6 orthologous groups and mapped gene sets with one or two species exhibiting apparent gene losses or duplications (see methods for descriptions of key labels). The OrthoDB6 species with new annotations in each case show that the re-annotation process has correctly identified more orthologs and reduced the numbers of potentially missing orthologs (blue and green). Amongst the final sets of gene annotations for the bee species, only *EMEXI* shows an elevated count of potentially missing orthologs. SC1m – single-copy orthologs in all, but missing from 1 of the selected species; SC2m – single-copy orthologs in all, but missing from 2 of the selected species; MC1m – orthologs present, with some multi-copy orthologs in all, but missing from 1 of the selected species; MC2m – orthologs present with some multi-copy orthologs in all, but missing from 2 of the selected species; SC1d – single-copy orthologs in all, but duplications in 1 of the selected species; SC2d – single-copy orthologs in all, but duplications in 2 of the selected species; [n] indicates the total number of orthologous groups for the category.

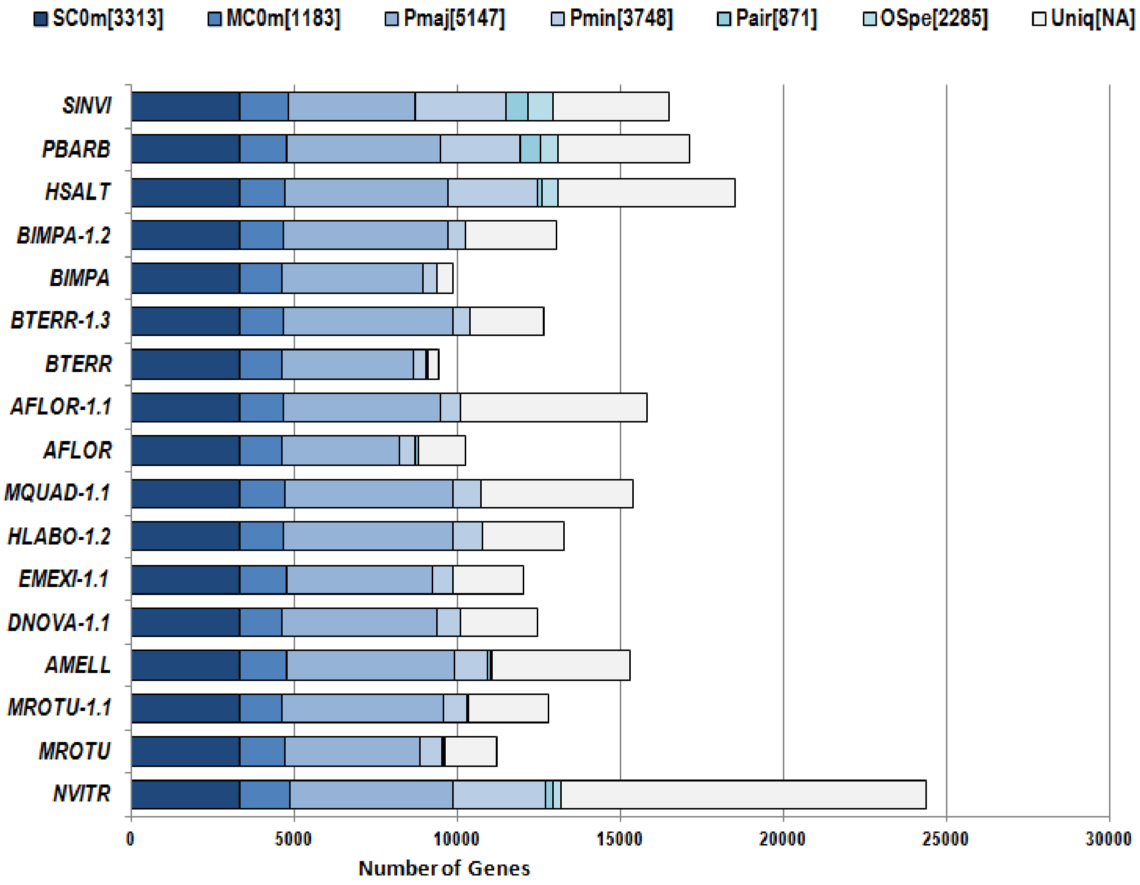

Fig. S2.

Phyletic distributions of Hymenoptera OrthoDB6 orthologous groups and mapped gene sets (see methods for descriptions of key labels). The OrthoDB6 species with new annotations in each case show that the re-annotation process has correctly identified more orthologs, and at the same time it has predicted more genes for which no orthology could be determined. SC0m – single-copy orthologs in all of the sleected species (zero missing; M0m – orthologs present, with some multi-copy orthologs, in all of the selected species (zero missing); Pmaj – orthologs present in the majority (>50%), but not all, of the selected species; Pmin – orthologs present in the minority (<=50%), but not just in pairs, of the selected species; Pair – orthologs present only in pairs of the selected species; OSpe – orthologs present in at least one other species from OrthoDB but not shown on the chart; Uniq – the fraction of genes for which no orthology could be determined; [n] represents the total number of orthologous groups for the category

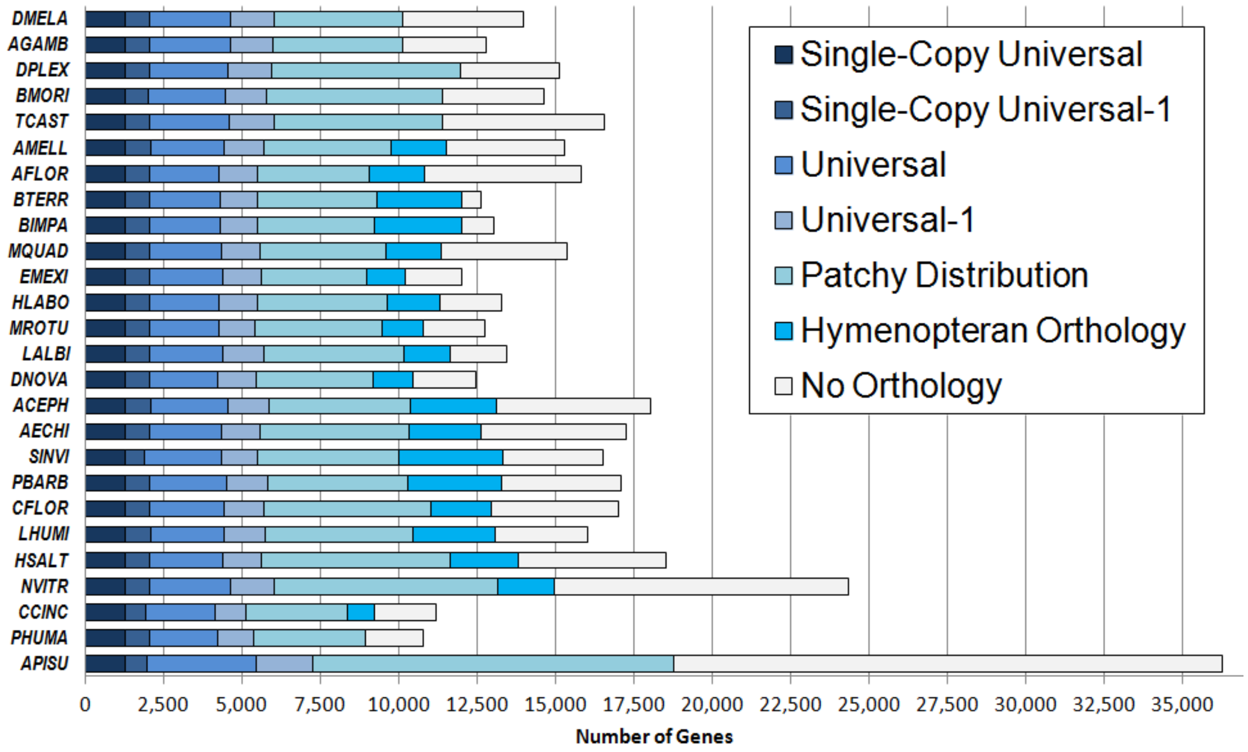

Fig. S3.

Complete OrthoDB orthology delineation across 26 insect species including 19 hymenopterans identified 21,785 orthologous groups and classified 75.5% of all genes, 78.2% of all hymenopteran genes, and 82.4% of all bee genes. A conserved core of about 5,500 genes in each bee species (35%-47%) have identifiable orthologs across the representatives of the insect phylogeny and a further ~1,200 to ~2,800 genes in each bee species (10%-21%) show no detectable orthology beyond Hymenoptera.

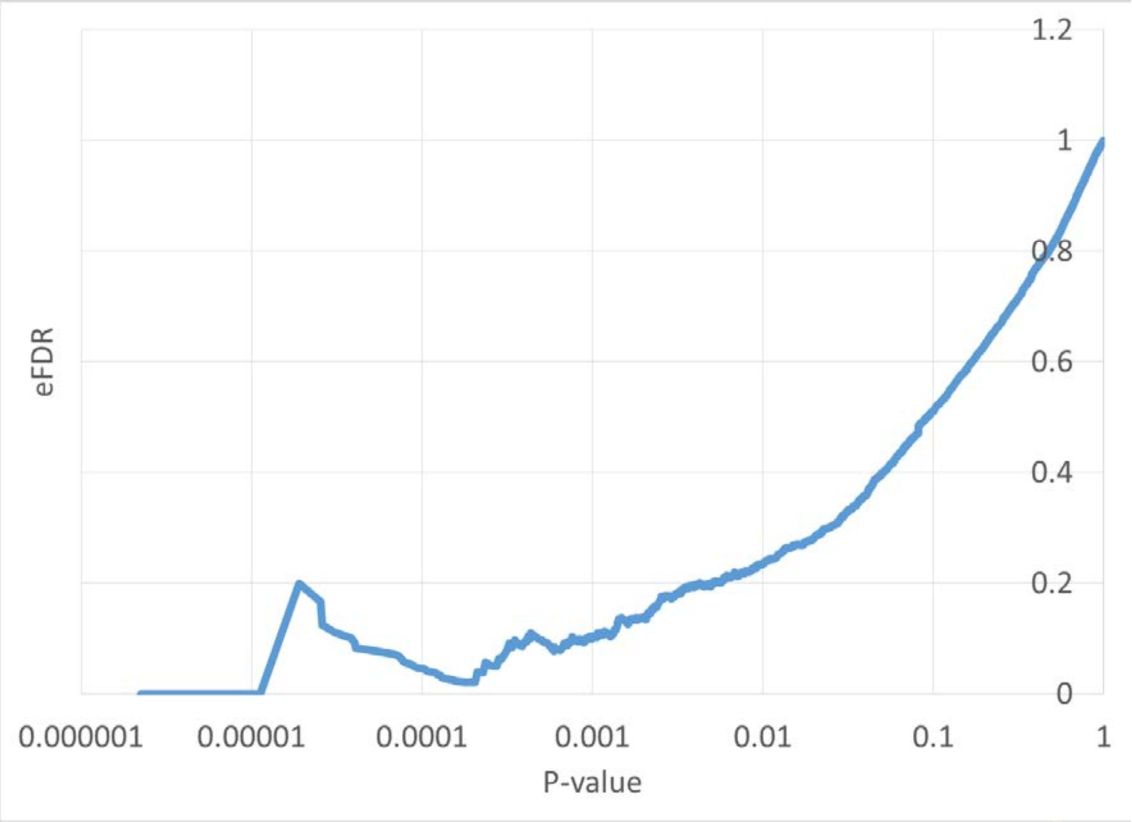

Fig. S4.
eFDR for different p-value thresholds.

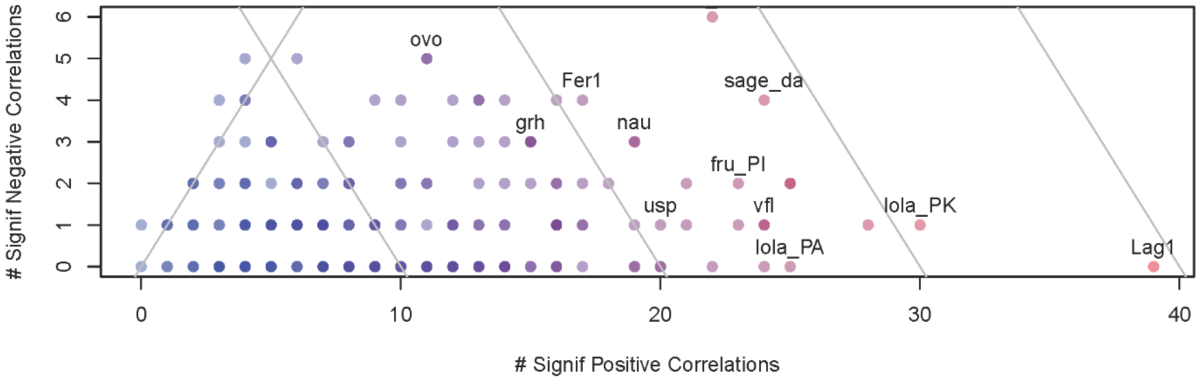

Fig. S5.

Significant correlations per TF motif. Plot of data from Additional Data table S1. Each point shows the number of significant positive/negative correlations from the PIC analysis for a particular motif. Select motifs with many significantly correlated orthogroups are labeled.

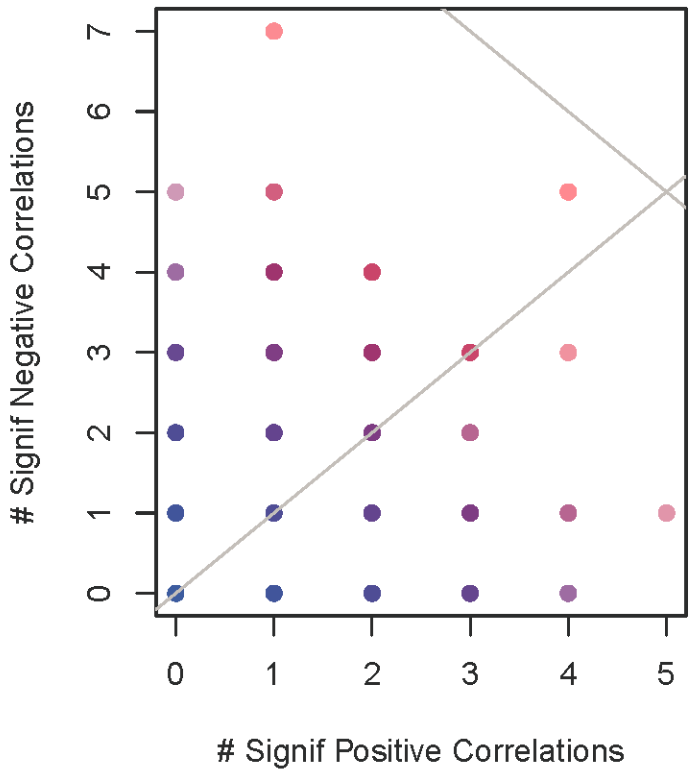

Fig. S6.

Significant correlations per TF motif in random controls. Equivalent to Fig. S5 except with data from random controls rather than real data. Figure shows much fewer significant results per motif with no bias towards positive correlations.

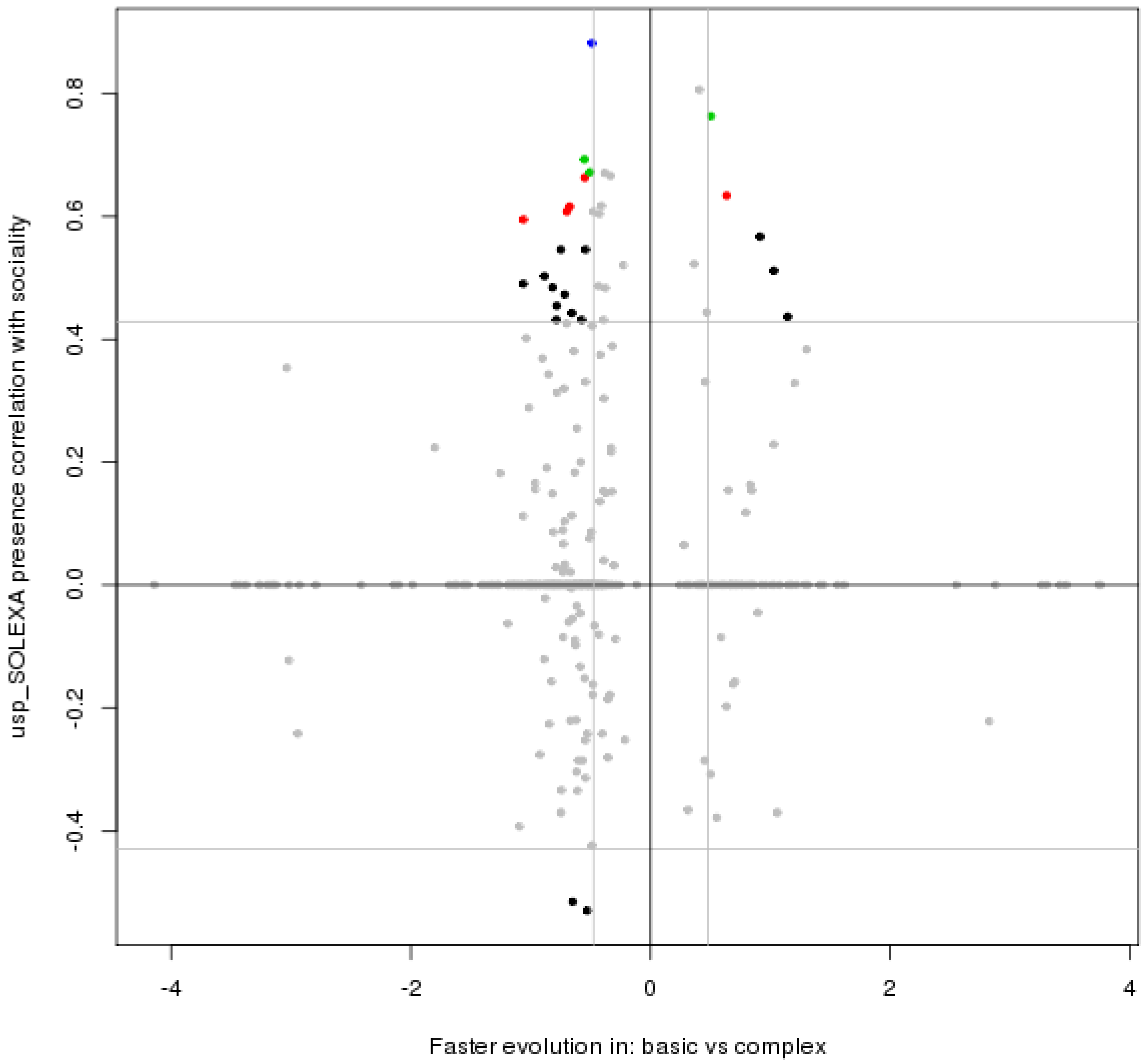

Fig. S7.

Relationship between coding sequence evolution and USP motif presence evolution in genes identified in the elaborations of sociality in stingless bees. Each point in the plot is an orthogroup in the results from the PAML analysis studying molecular evolution in stingless bees. The x-axis is the log10 ratio of the evolutionary rate between the complex and basic eusociality species. The vertical gray lines indicate a 3-fold change in the rate. The y-axis is the correlation values from the PIC analysis for the orthogroups with the USP motif. The horizontal gray lines indicate a correlation significance of 0.25. The colored points are orthogroups with significant correlations between motif and sociality and evolutionary rate and sociality. The locations of these colored points are examined for patterns of evolution.

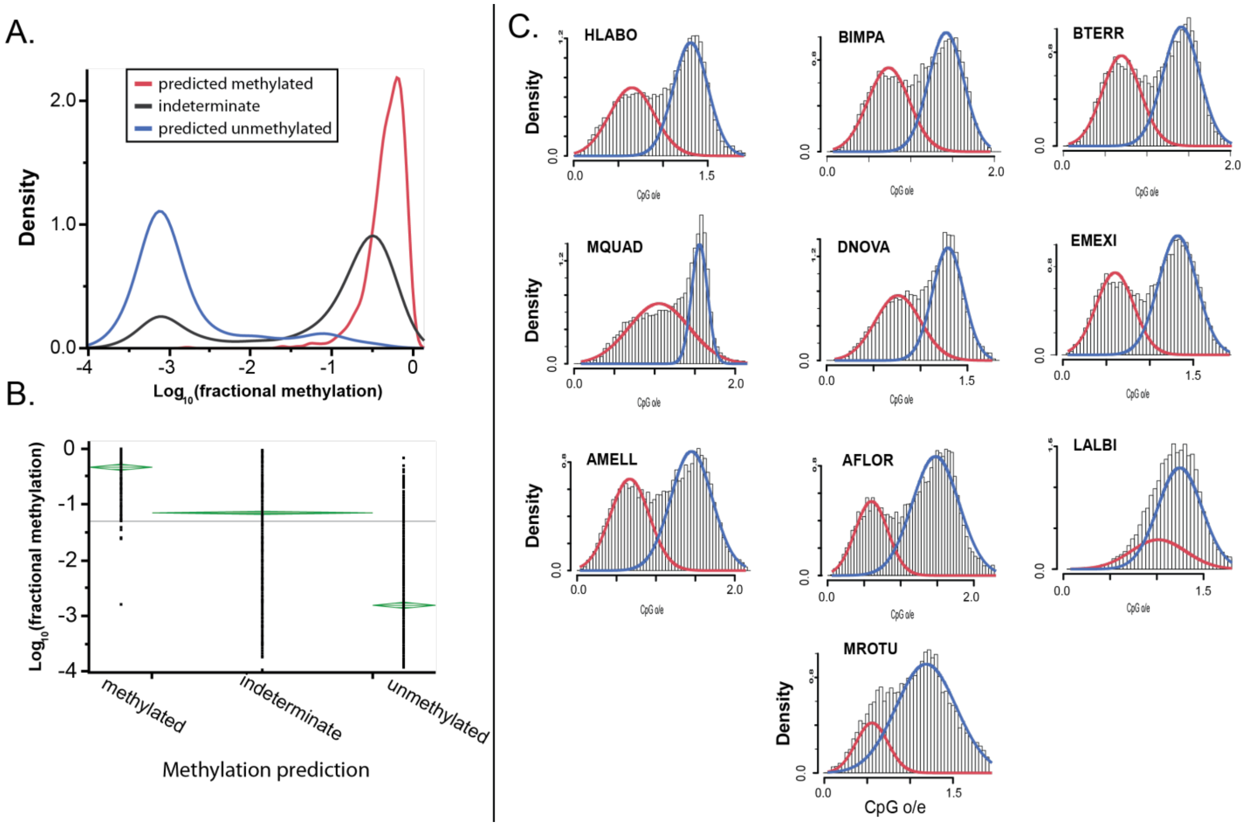

Fig. S8.

Validation of CpG o/e with empirical methylation data from *A. mellifera*. A) Densities and B) means comparing CpG o/e-based DNA methylation predictions to empirically-determined DNA methylation level in the honey bee (log10 transformed fractional methylation levels). C) CpG o/e of gene frames for each species as split by mixtools, illustrating the probability distribution of each of two components: putatively methylated genes (red) and putatively unmethylated genes (blue).

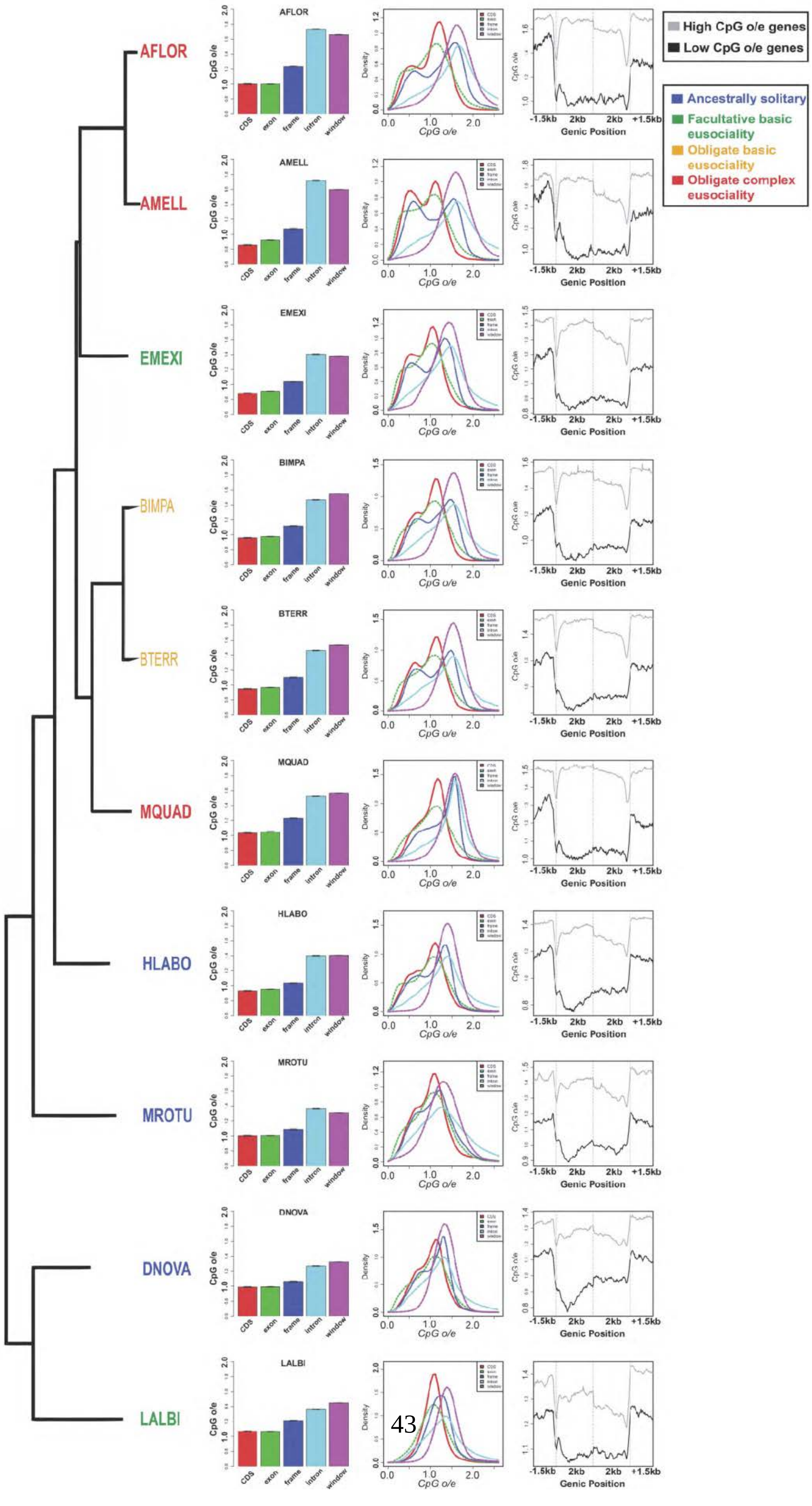
AFLOR
AMELL
EMEXI
BIMPA
BTERR
MQUAD
HLABO
MROTU
DNOVA
LALBI
CpG o/e
CDS
exon
frame
intron
window
Density
Genic Position
-1.5kb
2kb
2kb
+1.5kb
High CpG o/e genes
Low CpG o/e genes
Ancestrally solitary
Facultative basic eusociality
Obligate basic eusociality
Obligate complex eusociality

Fig. S9.

Patterns of DNA methylation in genomic features as determined by analysis of CpG depletion. The dendrogram depicts species relationships based on four-fold degenerate sites. Bar plots provide mean CpG o/e values, with 95% confidence intervals, for CpG o/e of coding sequences (CDS), exons, gene frames (exons and introns combined), introns, and intergenic windows (1kb fragments; genome window); density plots are shown for the same genomic features. The far right column contains spatial plots of mean CpG o/e values generated from sliding window analysis of genes and proximal regions; vertical dashed lines indicate start codon and stop codon. High and low CpG o/e genes are determined as those with CpG o/e values above and below the mean, respectively. Data are shown for (A) the bees *A. florea* (AFLOR), *A. mellifera* (AMELL), *E. mexicana* (EMEXI), *B. impatiens* (BIMP), *B. terrestris* (BTERR), *Mel. quadrifasciata* (MQUAD), *H. laboriosa* (HLABO), *Meg. rotundata* (MROTU), *D. novaeangliae* (DNOVA), and *L. albipes* (LALBI).

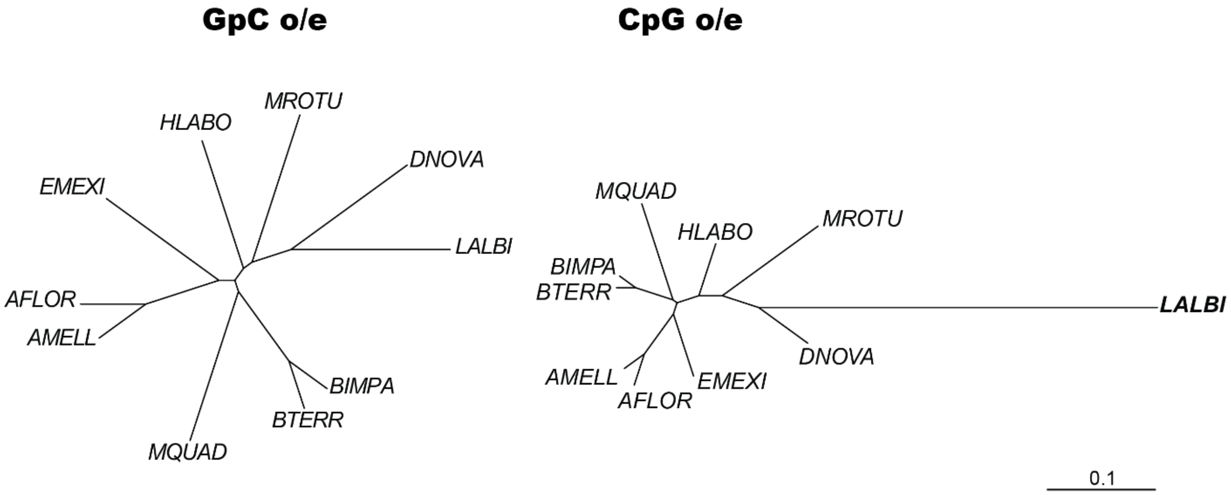

Fig. S10.

Neighbor joining trees constructed from Spearman's rank correlation distance matrices (using 1 – rho as pairwise distances) of coding sequence CpG o/e and GpC o/e strongly suggest that *L. albipes* is distinct in terms of CpG depletion, but not GpC depletion. Notably, the GpC o/e tree recapitulates the species relationships observed based on 4-fold degenerate sites, whereas the CpG o/e tree reveals *L. albipes* as a strong outlier.

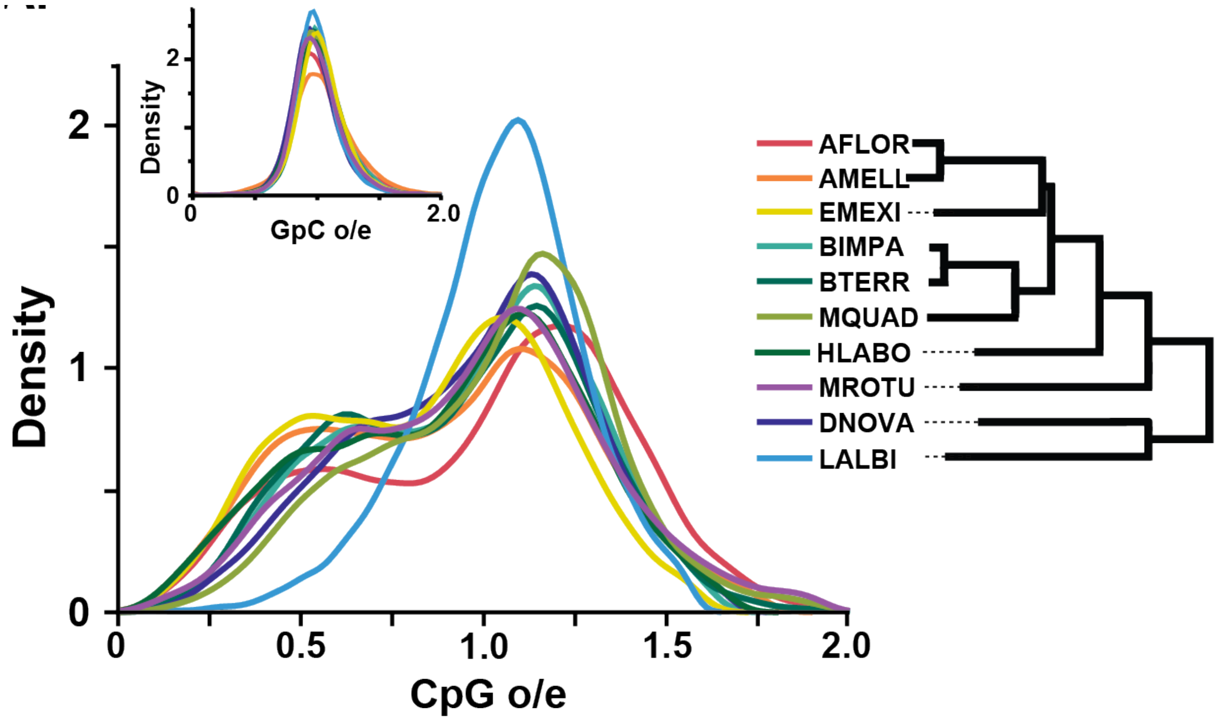

Fig. S11.

Mutational signature of DNA methylation in coding sequences of bees. CpG o/e smoothed distributions of Kernal probability density for coding sequences from each bee species, as well as for GpC o/e (control dinucleotide; inset); tree in legend generated from 4-fold degenerate sites of all aligned species. Almost all bees show bimodal distribution of CpG o/e, suggesting substantial levels of DNA methylation.

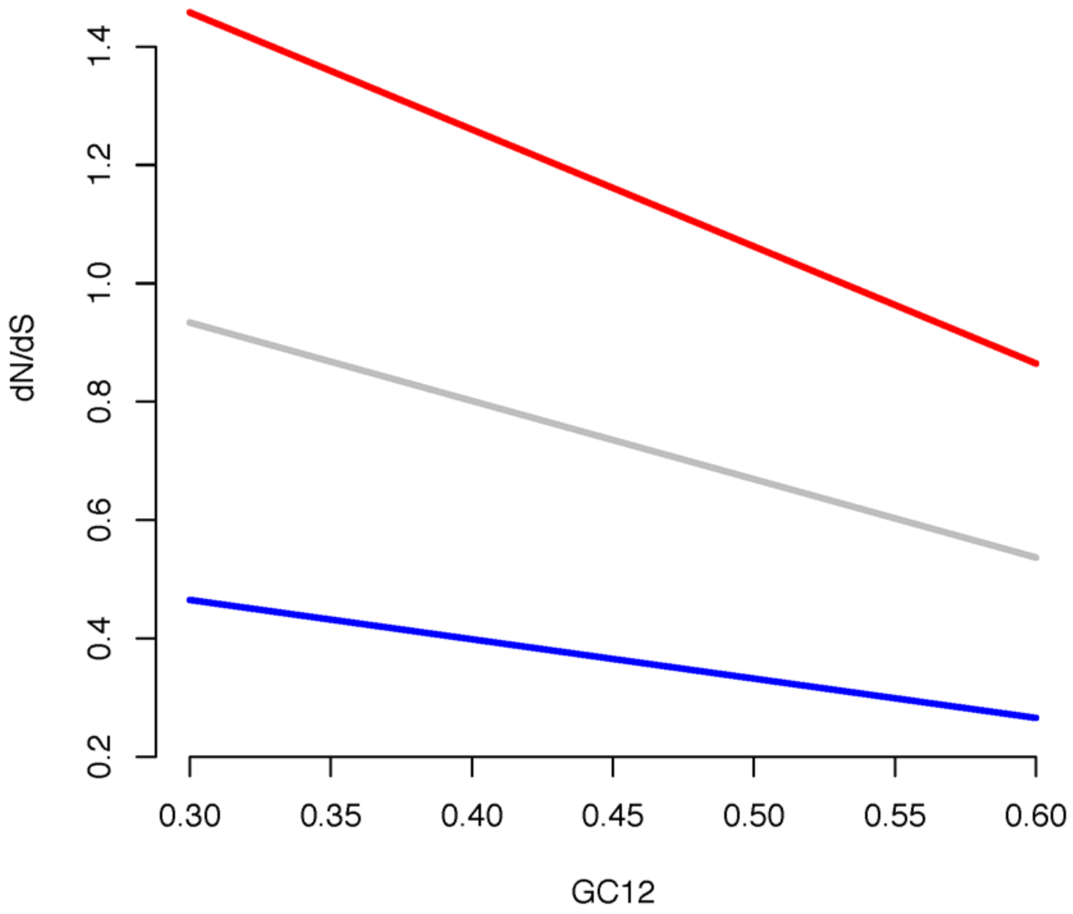

Fig. S12.

The effect of biased Gene Conversion on dN/dS depends on GC content. In this figure the GC3 content of a gene is held fixed, as is the selection coefficient s and the strength b of biased gene conversion. GC12 varies on the horizontal axis. Calculated dN/dS is higher at lower GC12 whether the underlying selection coefficient s is negative (blue line), neutral (gray line) or positive (red line). Thus the effect of GC12 on dN/dS via bGC is largely independent of the distribution of fitness effects. The value of dN/dS is calculated from the Malécot-Kimura formula (*149*) separately for probability of fixation of new alleles representing GC-increasing mutation (AT→GC mutations, probability increased by bGC) and for probability of fixation of GC-decreasing mutations (GC→AT, probability decreased by bGC) and then using the average of the probabilities weighted by the proportion of GC parent alleles at non-synonymous sites (GC12) for dN and the proportion of GC parent alleles at synonymous sites (GC3) for dS. This relationship was originally noted by Bulmer (*150*) and Li (*151*).

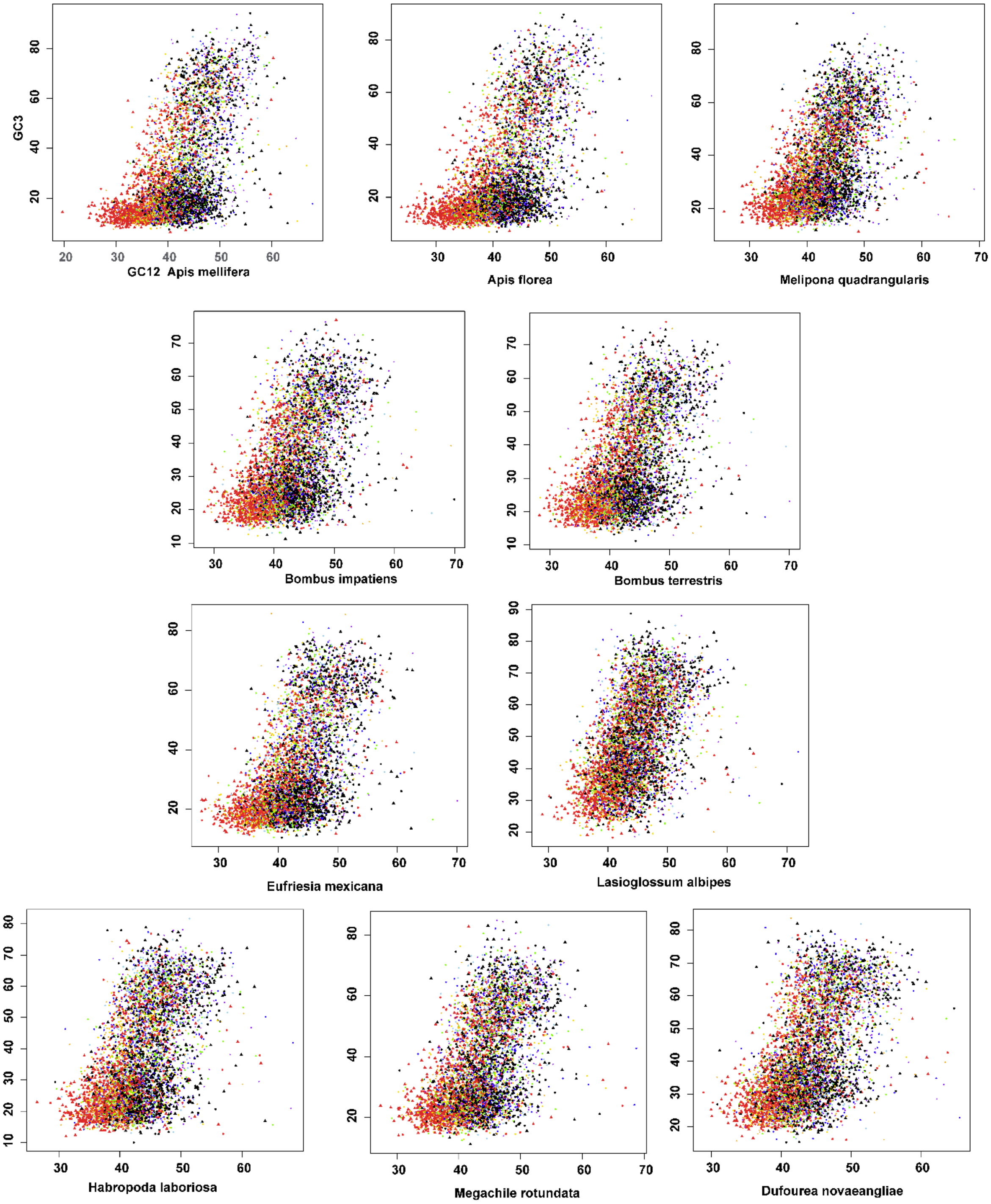

Fig. S13.

Distribution of GC12, GC3, and measured ω for each species. This figure shows the distributions of genes by GC12 (x-axis) and GC-3 (y-axis). Each gene is a point colored by its coevol-determined omega. Red-orange mark the highest omega genes in a species, black-blue are the lowest. Within each species, the highest omega genes are generally at low GC12. Note that GC12 has a stronger effect on measured ω than GC3.

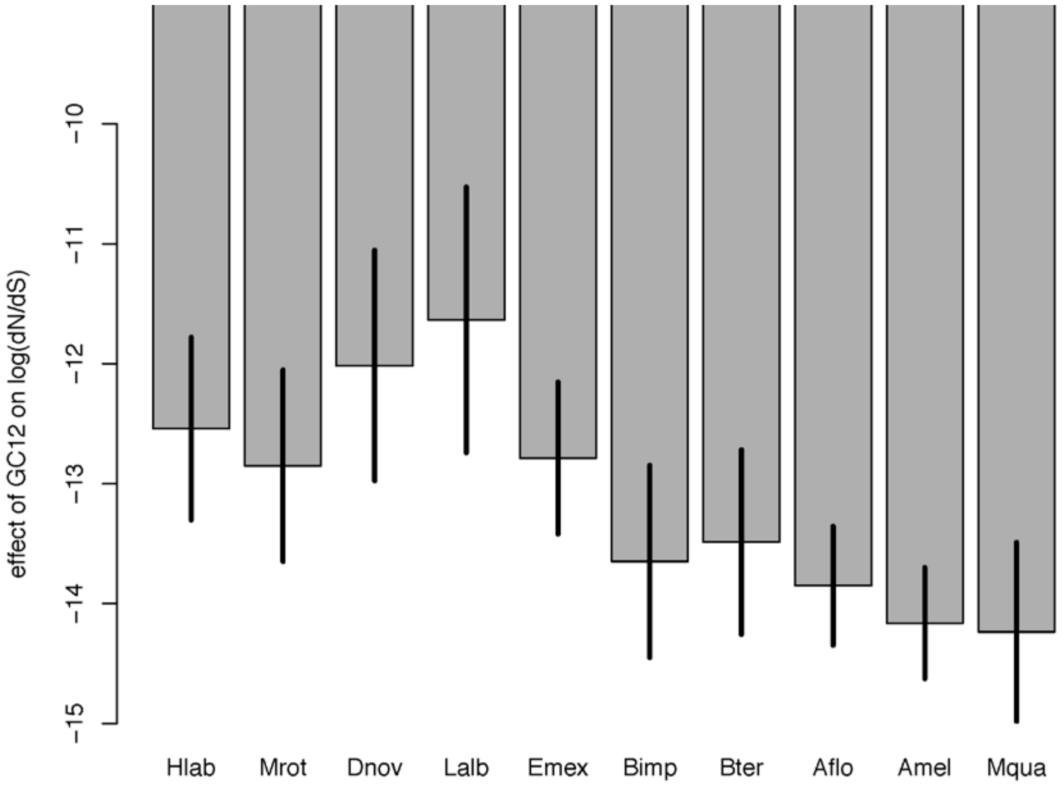

Fig. S14.

The slope of the regression of log(ω) on GC12 for 10 species. Lower GC12 increases ω, so the regression coefficients plotted are negative. An increase of GC12 from 0.45 to 0.40 would on average decrease ω by 50% in the 3 highly eusocial species. The effect of GC12 on ω is not as strong in the 5 non-social or facultatively eusocial species as in the 5 obligately eusocial species (Welch t-test t=5.49, p=0.001).

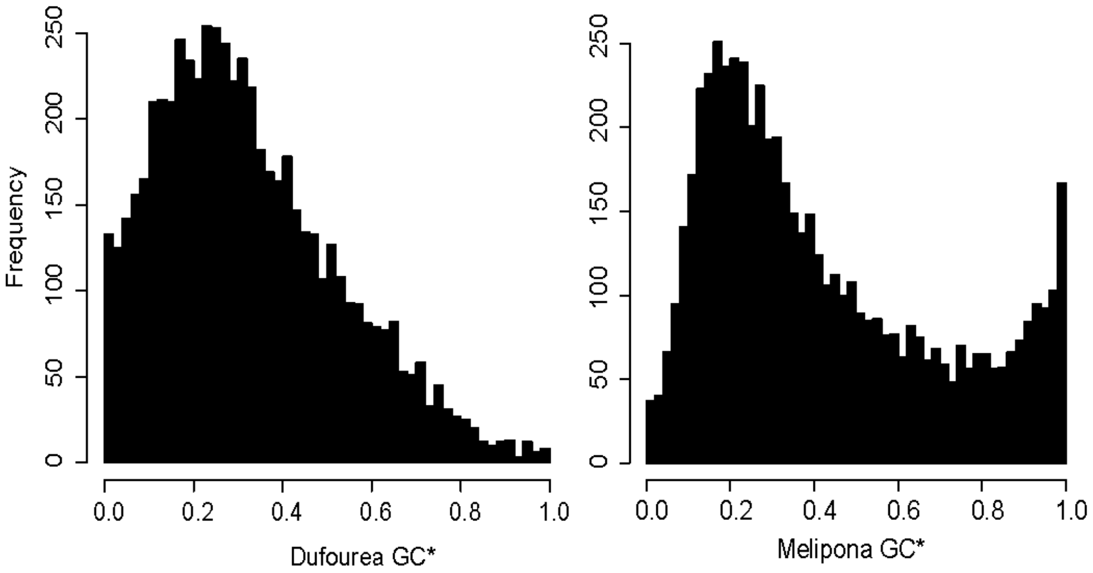

Fig. S15.

CoEvol-calculated GC* (equilibrium GC) for all genes in a facultatively eusocial (*Dufourea novaeangliae*) and complex eusocial (*Melipona quadrifasciata*) species. Note the skew towards low GC* but the significant minority of genes with high GC* in *Melipona*.

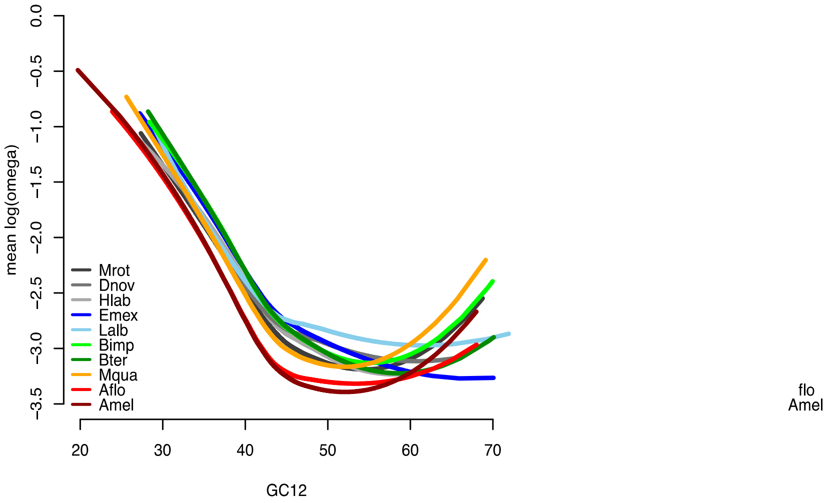

Fig. S16.

ω versus GC12 and GC* in ten bee species. For each species the smoothed ω (via the non-parametric loess method) is plotted against (a) GC12, and (b) GC*. The parabolic or U shapes are in fact best fit by quadratic regressions (highly significantly), rather than linear regressions. This is the basis of the GC-corrected ω. For each of the 10 species, a best-fit equation quadratic in GC12 and GC* was fit. Interaction terms were non-significant and were not included. Residuals from the equation are zero-centered and were scaled to have unit variance, thus producing z-scores. This is GC-corrected ω. If the bulk of genes are neutral or very mildly deleterious, points near the fitted curves (GC-corrected ω near zero) are under neutral or nearly neutral conditions. Points well above the fitted curves (GC-corrected ω greater than zero) are experiencing adaptive selection in addition to neutral selection. Finally, points well below the fitted curves (GC-corrected ω less than zero) are experiencing adaptive selection in addition to neutral selection.

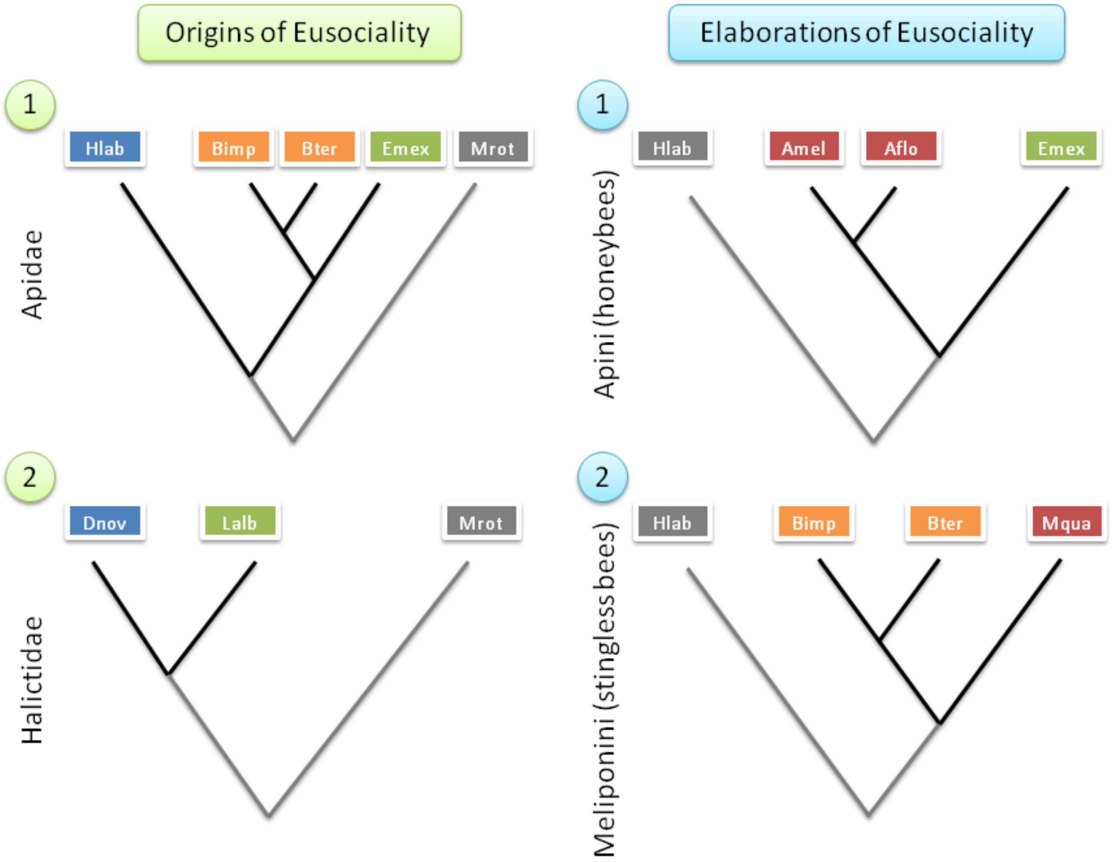

Fig. S17.

Species and trees used for clade-specific PAML analyses. Species used in each clade-specific PAML analysis included independent evolutionary transitions from solitary (blue) to basic eusociality (green – facultative, orange – obligate) and two independent evolutionary transitions from basic eusociality to complex eusociality (red). Each analysis included an outgroup (grey) to root the tree.

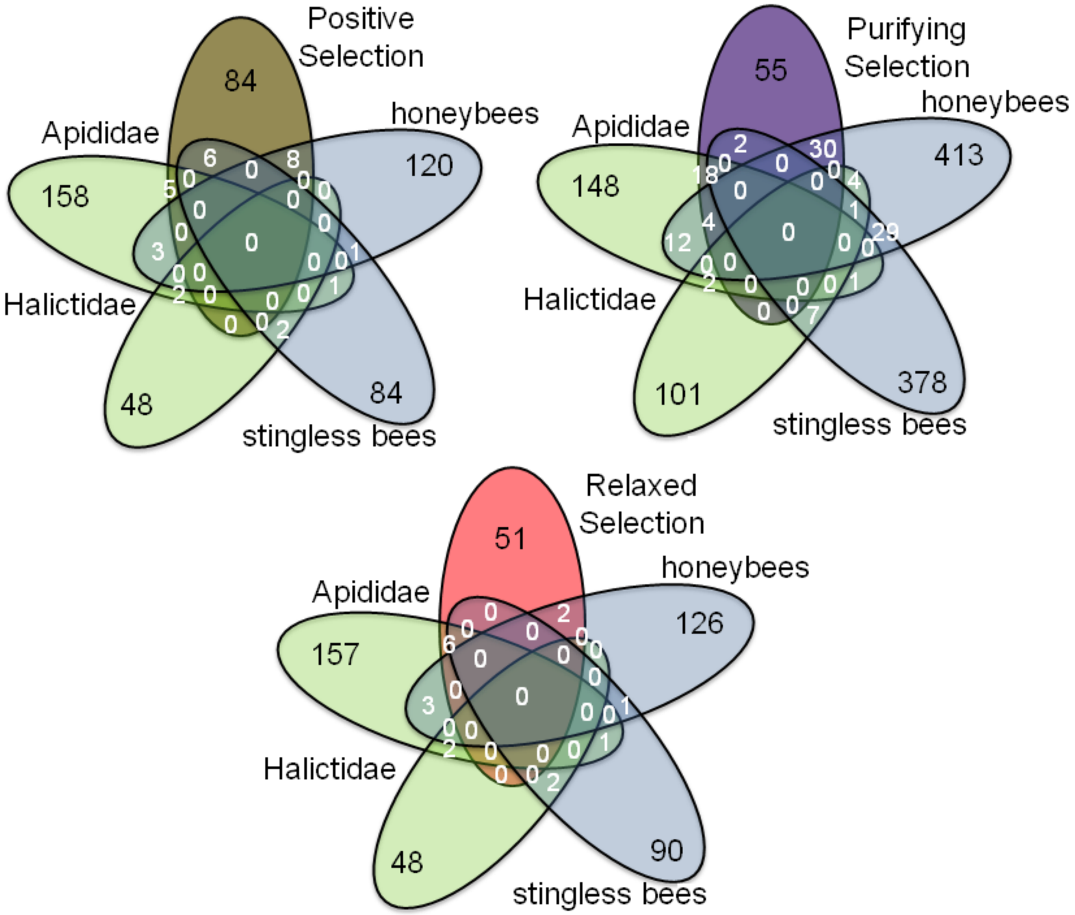

Fig. S18.

Number of genes overlapping between clade-specific PAML analyses of dN/dS consistent with origins (green) and elaborations (blue) of eusociality and analyses of positive (brown), relaxed (orange), and purifying (purple) selection associated with increasing social complexity in all ten bee species.

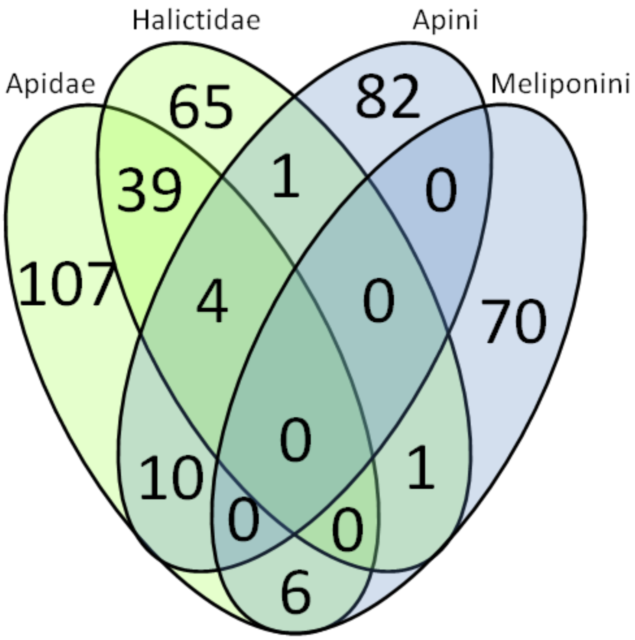

Fig. S19.

Number of GO terms significantly (FDR p < 0.05) enriched among genes significant in the clade-specific PAML analyses. The green shaded sections represent two independent origins of eusociality – one in Apidae and one in Halictidae. The blue shaded sections represent two independent elaborations of eusociality – one in honey bees (Apini) and one in stingless bees (Meliponini).

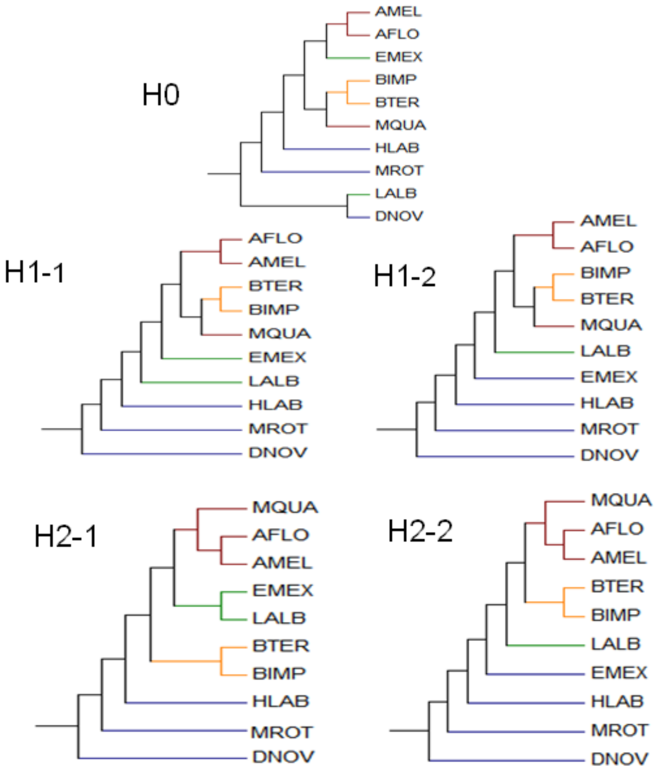

Fig. S20.

Alternative tree topologies used in convergent evolution analysis. H0 is the accepted species phylogeney (based on (*28*)). H1 places all eusocial species in a monophyletic clade. H2 places species in monophyletic clades based on social complexity. Subset 1 of each alternative hypothesis treats *E. mexicana* as facultatively eusocial. Subset 2 of each alternative hypothesis treats *E. mexicana* as solitary. Branch colors follow Fig. 1.

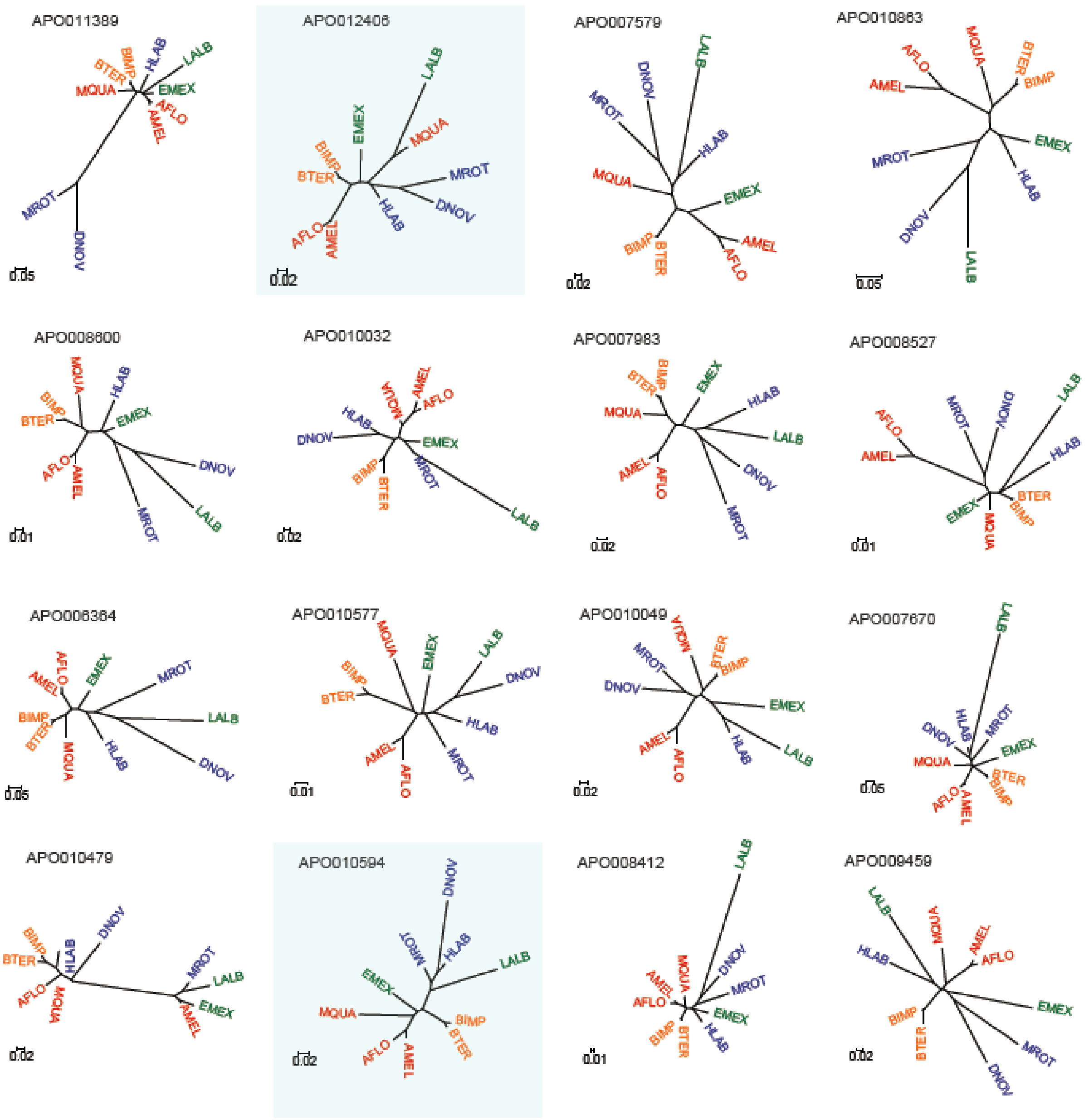

Fig. S21.
Gene trees for the 16 candidates of convergent evolution. These 16 orthogroups are a significantly better fit to an alternative topology based on social phenotype than to the species tree. RAxML trees of the gene members of these orthogroups indicate only two (shaded blue background) are evolving differently in eusocial vs noneusocial species.

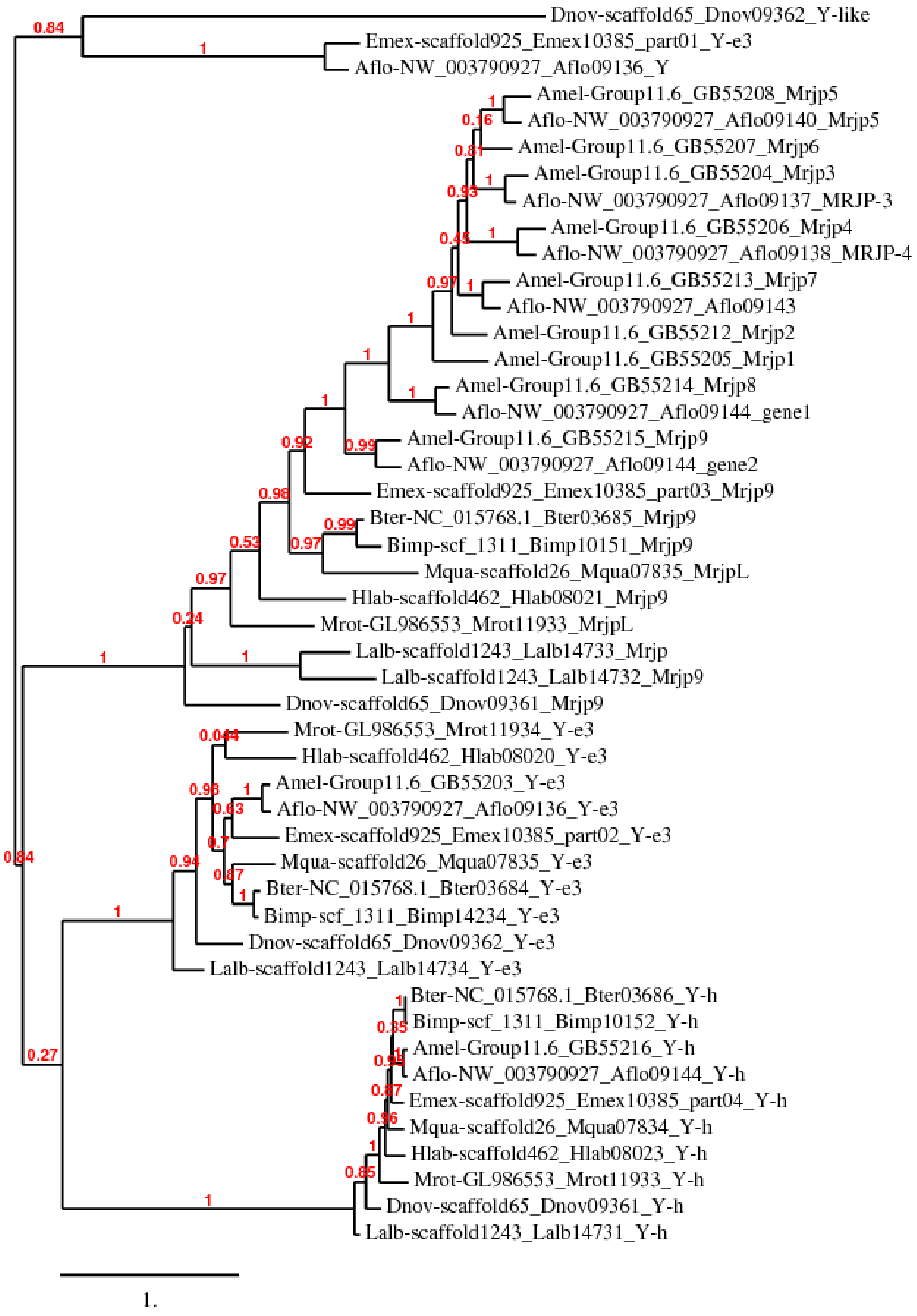
0.84
Dnov-scaffold65_Dnov09362_Y-like
1
Emex-scaffold925_Emex10385_part01_Y-e3
Aflo-NW_003790927_Aflo09136_Y
1
Amel-Group11.6_GB55208_Mrjp5
0.16
Aflo-NW_003790927_Aflo09140_Mrjp5
0.81
Amel-Group11.6_GB55207_Mrjp6
1
Amel-Group11.6_GB55204_Mrjp3
0.93
Aflo-NW_003790927_Aflo09137_MRJP-3
1
Amel-Group11.6_GB55206_Mrjp4
0.45
Aflo-NW_003790927_Aflo09138_MRJP-4
0.97
1
Amel-Group11.6_GB55213_Mrjp7
Aflo-NW_003790927_Aflo09143
1
Amel-Group11.6_GB55212_Mrjp2
1
Amel-Group11.6_GB55205_Mrjp1
1
Amel-Group11.6_GB55214_Mrjp8
1
Aflo-NW_003790927_Aflo09144_gene1
0.92
0.99
Amel-Group11.6_GB55215_Mrjp9
Aflo-NW_003790927_Aflo09144_gene2
0.98
Emex-scaffold925_Emex10385_part03_Mrjp9
0.99
Bter-NC_015768.1_Bter03685_Mrjp9
0.53
0.97
Bimp-scf_1311_Bimp10151_Mrjp9
0.97
Mqua-scaffold26_Mqua07835_MrjpL
Hlab-scaffold462_Hlab08021_Mrjp9
0.24
Mrot-GL986553_Mrot11933_MrjpL
1
1
Lalb-scaffold1243_Lalb14733_Mrjp
Lalb-scaffold1243_Lalb14732_Mrjp9
Dnov-scaffold65_Dnov09361_Mrjp9
0.044
Mrot-GL986553_Mrot11934_Y-e3
Hlab-scaffold462_Hlab08020_Y-e3
0.98
1
Amel-Group11.6_GB55203_Y-e3
0.63
Aflo-NW_003790927_Aflo09136_Y-e3
0.7
Emex-scaffold925_Emex10385_part02_Y-e3
0.84
0.94
Mqua-scaffold26_Mqua07835_Y-e3
0.87
Bter-NC_015768.1_Bter03684_Y-e3
1
Bimp-scf_1311_Bimp14234_Y-e3
1
Dnov-scaffold65_Dnov09362_Y-e3
Lalb-scaffold1243_Lalb14734_Y-e3
1
Bter-NC_015768.1_Bter03686_Y-h
0.35
Bimp-scf_1311_Bimp10152_Y-h
0.95
1
Amel-Group11.6_GB55216_Y-h
Aflo-NW_003790927_Aflo09144_Y-h
0.27
0.87
Emex-scaffold925_Emex10385_part04_Y-h
0.96
Mqua-scaffold26_Mqua07834_Y-h
1
Hlab-scaffold462_Hlab08023_Y-h
0.65
Mrot-GL986553_Mrot11933_Y-h
1
Dnov-scaffold65_Dnov09361_Y-h
Lalb-scaffold1243_Lalb14731_Y-h
1.

Fig. S22.

Phylogenetic tree for the *mrjp* and their flanking yellow genes (*y-e3* and *y-h*) in the ten bee species. The tree was generated using T-Coffee with PhyML settings, bootstrap values are in red. Species names are represented by their three-letter acronyms, followed by genome scaffold number, gene set number and abbreviated gene name.

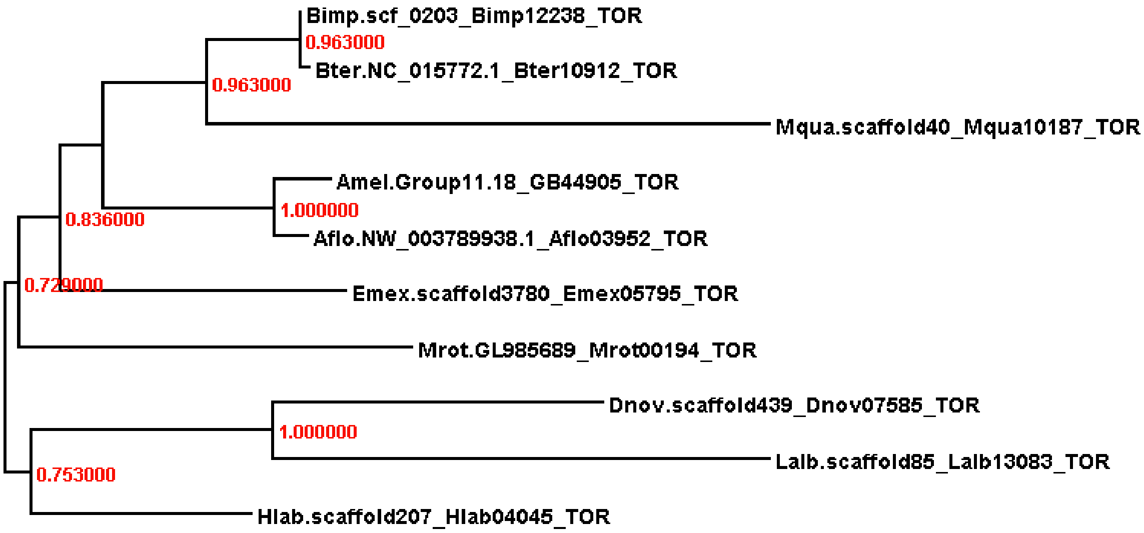
Bimp.scf_0203_Bimp12238_TOR
0.963000
Bter.NC_015772.1_Bter10912_TOR
0.963000
Mqua.scaffold40_Mqua10187_TOR
Amel.Group11.18_GB44905_TOR
1.000000
0.836000
Aflo.NW_003789938.1_Aflo03952_TOR
0.729000
Emex.scaffold3780_Emex05795_TOR
Mrot.GL985689_Mrot00194_TOR
Dnov.scaffold439_Dnov07585_TOR
1.000000
Lalb.scaffold85_Lalb13083_TOR
0.753000
Hlab.scaffold207_Hlab04045_TOR

Fig. S23.

Phylogenetic tree for the *tor* gene CDS in the ten bee species. The tree was generated using T-Coffee with PhyML settings, bootstrap values are in red. Species names are represented by their three-letter acronyms, followed by gene set number.

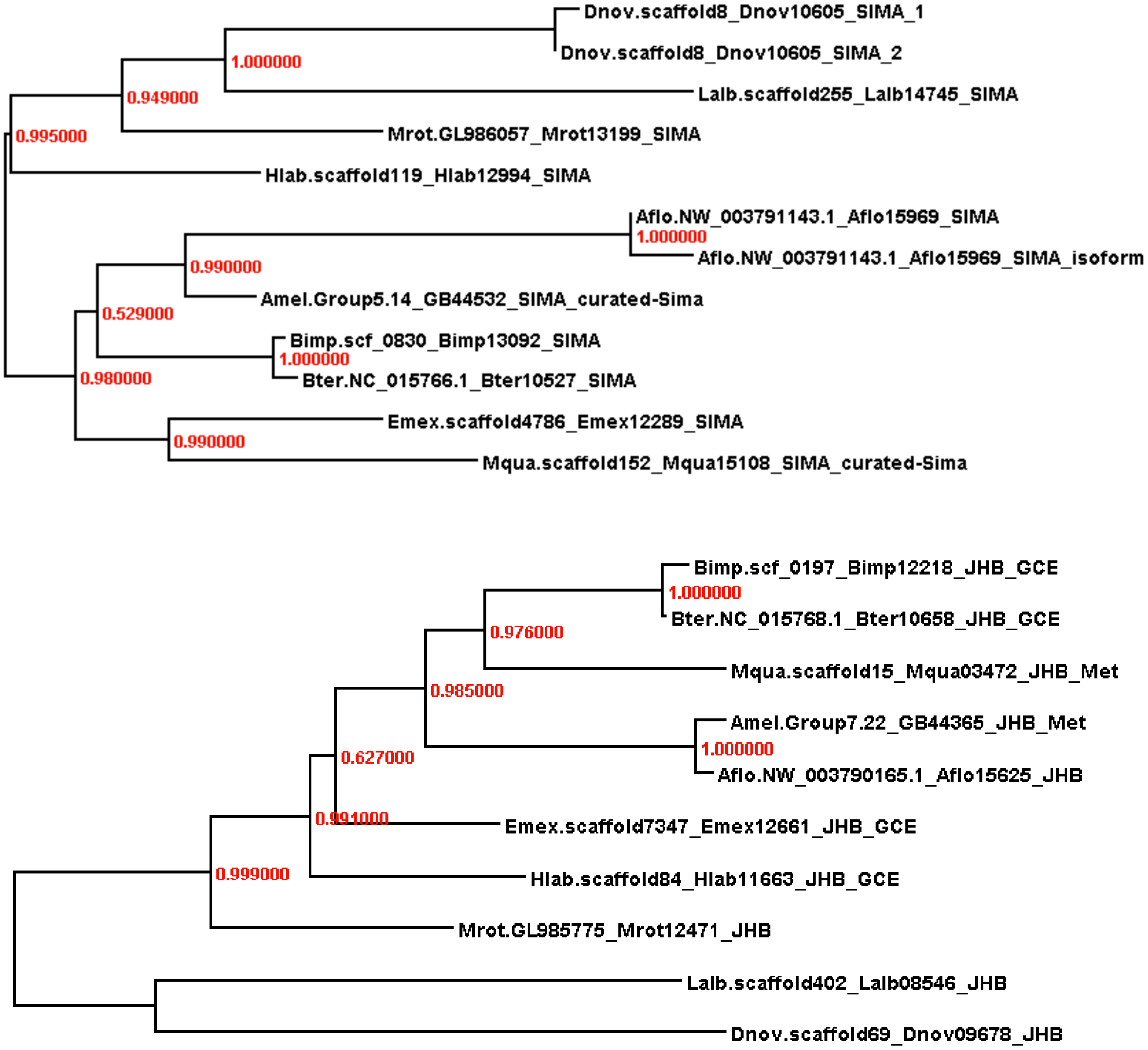
Dnov.scaffold8_Dnov10605_SIMA_1
Dnov.scaffold8_Dnov10605_SIMA_2
1.000000
0.949000
Lalb.scaffold255_Lalb14745_SIMA
0.995000
Mrot.GL986057_Mrot13199_SIMA
Hlab.scaffold119_Hlab12994_SIMA
Aflo.NW_003791143.1_Aflo15969_SIMA
1.000000
Aflo.NW_003791143.1_Aflo15969_SIMA_isoform
0.990000
Amel.Group5.14_GB44532_SIMA_curated-Sima
0.529000
Bimp.scf_0830_Bimp13092_SIMA
1.000000
Bter.NC_015766.1_Bter10527_SIMA
0.980000
Emex.scaffold4786_Emex12289_SIMA
0.990000
Mqua.scaffold152_Mqua15108_SIMA_curated-Sima
Bimp.scf_0197_Bimp12218_JHB_GCE
1.000000
Bter.NC_015768.1_Bter10658_JHB_GCE
0.976000
Mqua.scaffold15_Mqua03472_JHB_Met
0.985000
Amel.Group7.22_GB44365_JHB_Met
1.000000
Aflo.NW_003790165.1_Aflo15625_JHB
0.627000
0.991000
Emex.scaffold7347_Emex12661_JHB_GCE
Hlab.scaffold84_Hlab11663_JHB_GCE
0.999000
Mrot.GL985775_Mrot12471_JHB
Lalb.scaffold402_Lalb08546_JHB
Dnov.scaffold69_Dnov09678_JHB

Fig. S24.

Phylogenetic trees for the HIF-alpha gene (sima, upper panel) and the *methoprene-tolerant/germ cell-expressed* gene (Met, GCE, JHB, lower panel). Despite their variability in predicted CDS size and exon number, these genes belonging to the HLH-PAS family strikingly well reflect the bee tree of life (*117*). Species names are represented by their three-letter acronyms, followed by gene set number. Bootstrap values are in red.

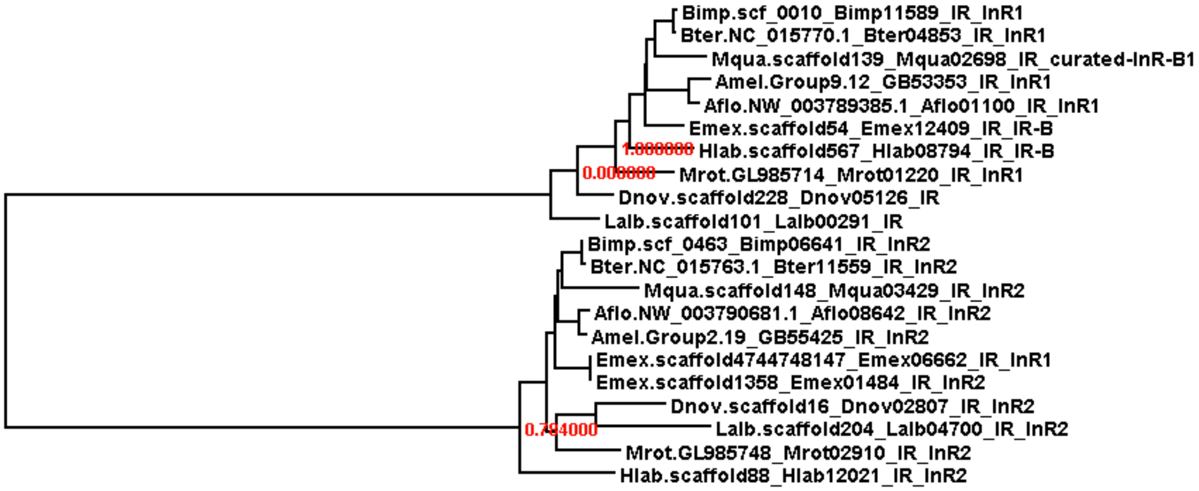
Bimp.scf_0010_Bimp11589_IR_InR1
Bter.NC_015770.1_Bter04853_IR_InR1
Mqua.scaffold139_Mqua02698_IR_curated-InR-B1
Amel.Group9.12_GB53353_IR_InR1
Aflo.NW_003789385.1_Aflo01100_IR_InR1
Emex.scaffold54_Emex12409_IR_IR-B
1.000000
Hlab.scaffold567_Hlab08794_IR_IR-B
0.000000
Mrot.GL985714_Mrot01220_IR_InR1
Dnov.scaffold228_Dnov05126_IR
Lalb.scaffold101_Lalb00291_IR
Bimp.scf_0463_Bimp06641_IR_InR2
Bter.NC_015763.1_Bter11559_IR_InR2
Mqua.scaffold148_Mqua03429_IR_InR2
Aflo.NW_003790681.1_Aflo08642_IR_InR2
Amel.Group2.19_GB55425_IR_InR2
Emex.scaffold4744748147_Emex06662_IR_InR1
Emex.scaffold1358_Emex01484_IR_InR2
Dnov.scaffold16_Dnov02807_IR_InR2
0.784000
Lalb.scaffold204_Lalb04700_IR_InR2
Mrot.GL985748_Mrot02910_IR_InR2
Hlab.scaffold88_Hlab12021_IR_InR2

Fig. S25.

Phylogenetic trees for the insulin receptors of the ten bee species. The two insulin receptors showed very distinct clustering in the gene trees, and within each cluster the bee tree of life structure was fairly well represented. For *Eufriesea mexicana*, a third prediction for an insulin receptor (Emex06662) was found in the genome assembly, with the same predicted CDS size and exon number as Emex 01484. Species names are represented by their three-letter acronyms, followed by gene set number. Bootstrap values are in red.

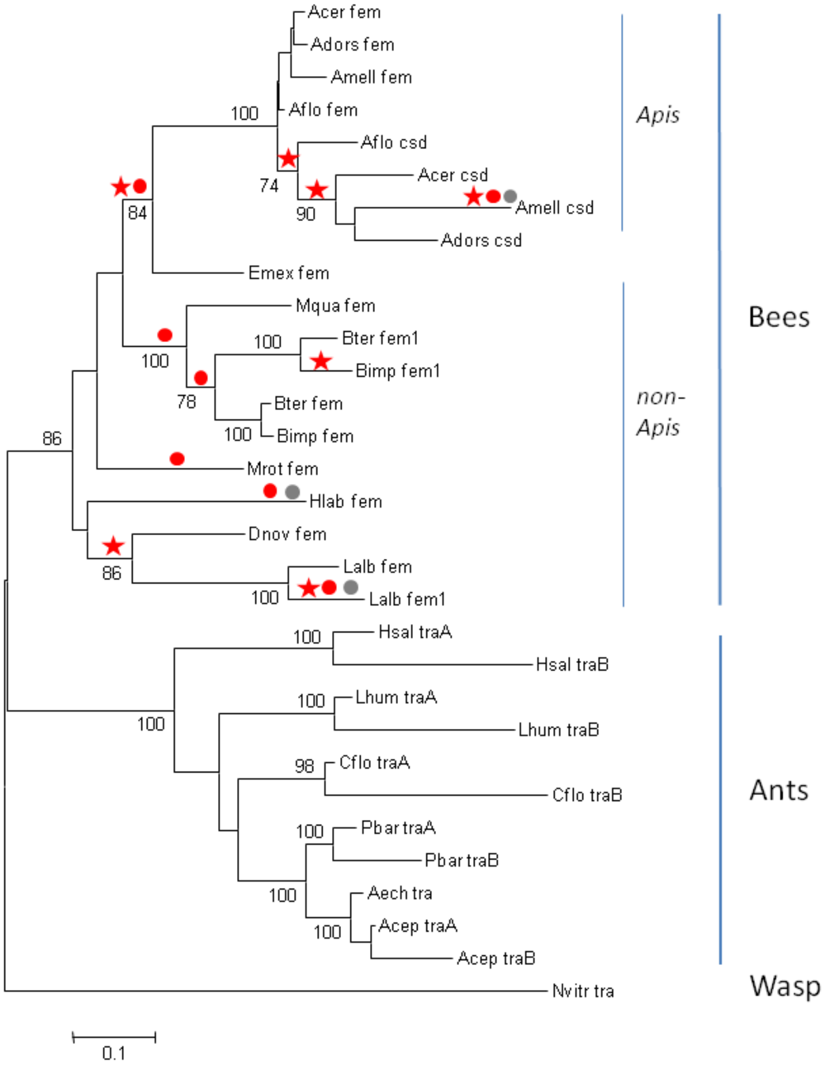

Fig. S26.

Molecular phylogenetic analysis of the sex-determining genes *fem/tra* in 19 hymenopteran species (ten bees from this study + *Apis cerana*, *A. dorsata*, *Harpegnathos saltator*, *Linepithema humile*, *Camponotus floridanus*, *Pogonomyrmex barbatus*, *Acromyrmex echinatior*, *Atta cephalotes*, *Nasonia vitripennis*) and signs of diversifying selection in bee lineages. The evolutionary history of 31 deduced amino acid sequences was inferred by using the Maximum Likelihood method based on the JTT matrix-based model. The tree with the highest log likelihood (-6819.26) and bootstrap values above 70 is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+*G*, parameter = 1.54)). The tree scale represents number of substitutions per site. All positions with less than 90% site coverage were eliminated. Analyses for detecting signs of diversifying selection among bee lineages were performed using *HyPhy (87).* Branches with dn/ds > 1.3 detected by the GA-model are indicated by red asterisk. Dots indicate signs of episodic diversifying selection (grey for ω = 1, red for ω > 5), detected by the branch-site REL model.

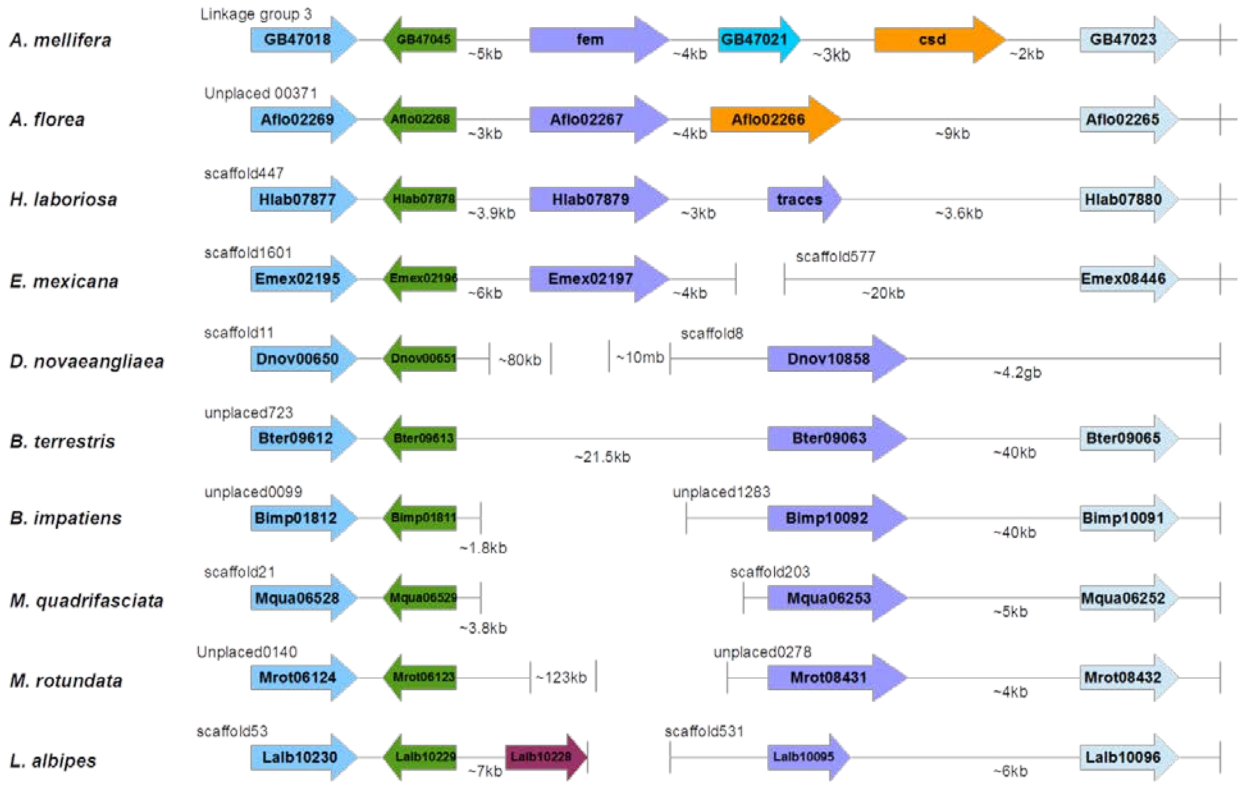

Fig. S27.

Overview of the relative gene position in the *A. mellifera* sex determination locus and corresponding orthologs in nine bee genomes. Orthologous genes among bee species are coded by the same color. Relative gene position on corresponding linkage group (for Amel v4.5 with OGS v3.2) and scaffold number are shown. Gene orientations in the genome are represented by arrows. The *fem* paralogs for *B. terrestris*, *B. impatiens* and *L. albipes* are not shown as they locate outside of this region.

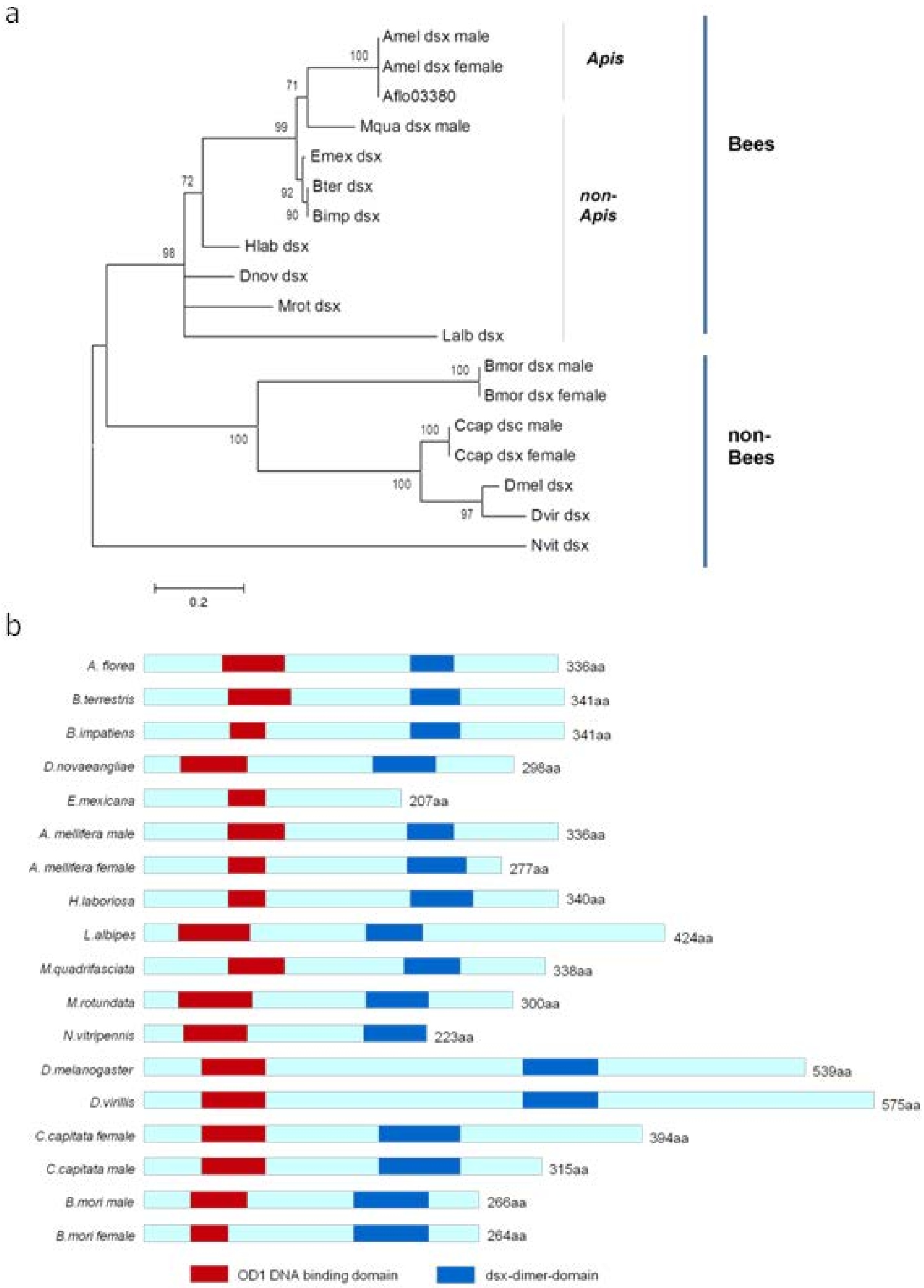

Fig. S28.

Evolutionary relationship based on amino distance using maximum likelihood algorithm (a) and protein scheme of doublesex (b) Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*, Nvitr-*Nasonia vitripennis*, Ccap-*Ceratitis capitata*, Dmel-*Drosophila melanogaster*, Dvir-*Drosophila virilis*, Bmor-*Bombyx mori.*

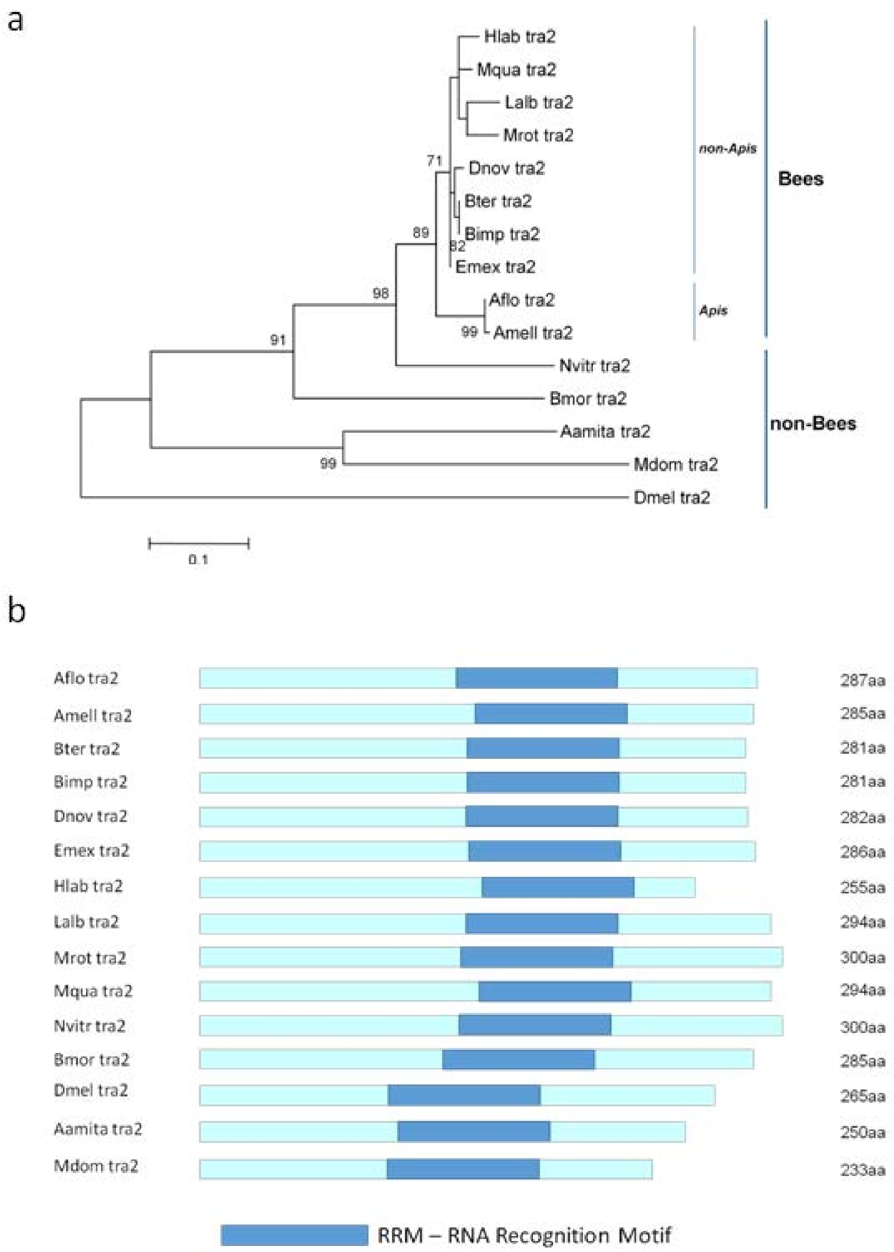

Fig. S29.

Evolutionary relationship based on amino distance using maximum likelihood algorithm (a) and protein scheme of transformer 2 (b). Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*, Nvitr-*Nasonia vitripennis*, Dmel-*Drosophila melanogaster*, Aamita-*Anastrepha amita*, Mdom-*Musca domestica*, Bmor-*Bombyx mori.*

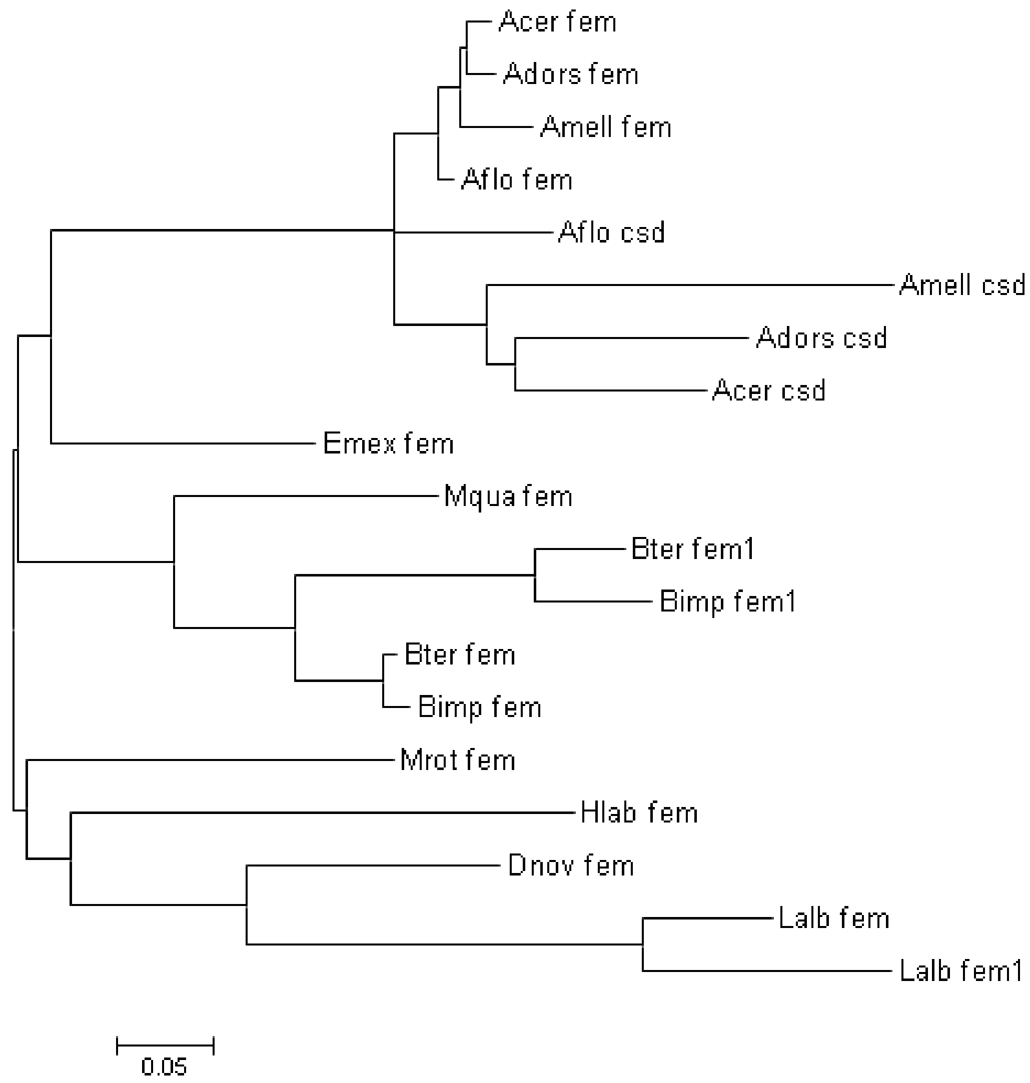

Fig. S30.

Genealogy used for the detection of selection within the *HyPhy* program package. The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model. The tree with the highest log likelihood (-4405.2467) is shown. A discrete Gamma distribution was used to model evolutionary rate differences among sites (2 categories (+*G*, parameter = 1.8)). The analysis involved 19 amino acid sequences. All positions with less than 90% site coverage were eliminated.

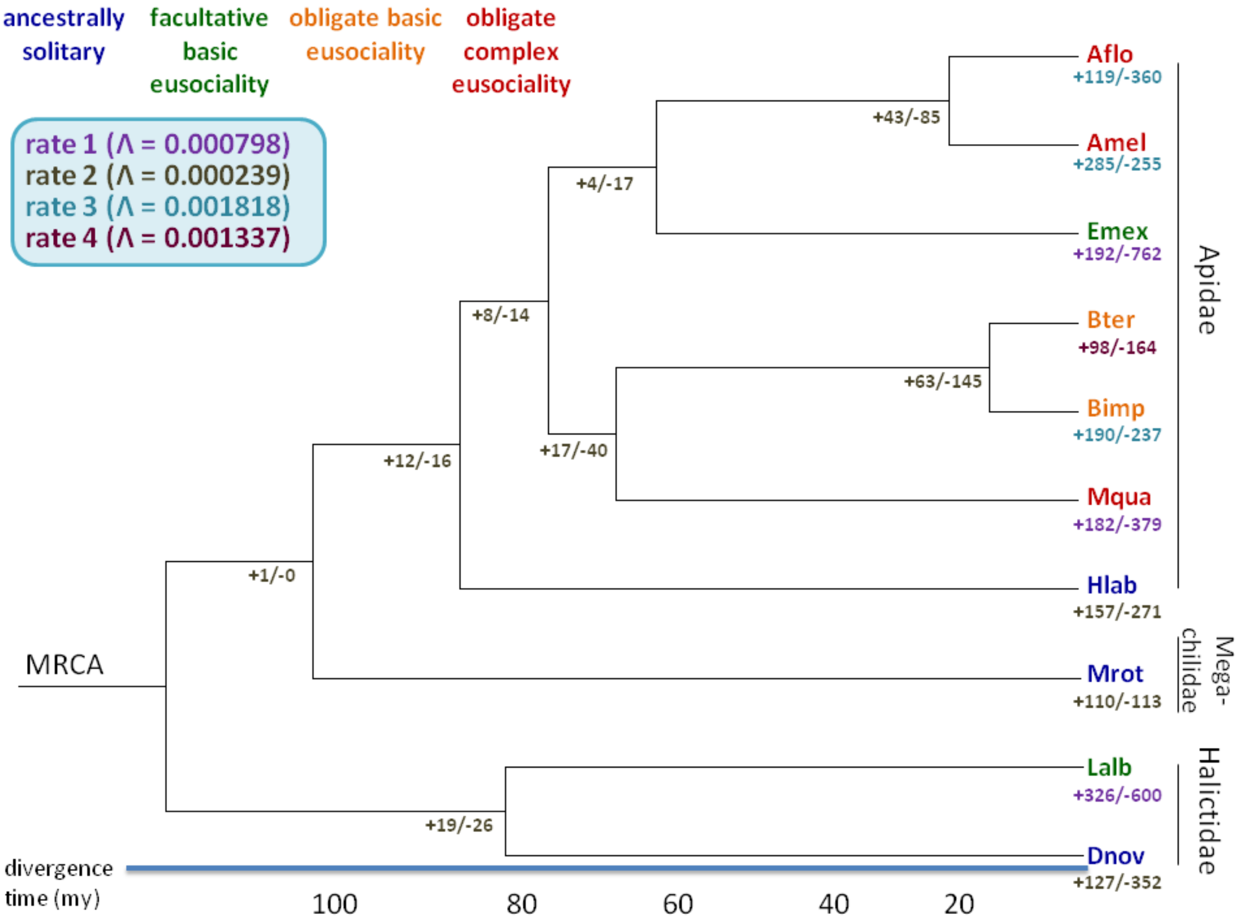

Fig. S31.

Gene family expansion and contraction among ten bees. Lambda values are rates of gene family birth/death in each lineage based on the best fitting model. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

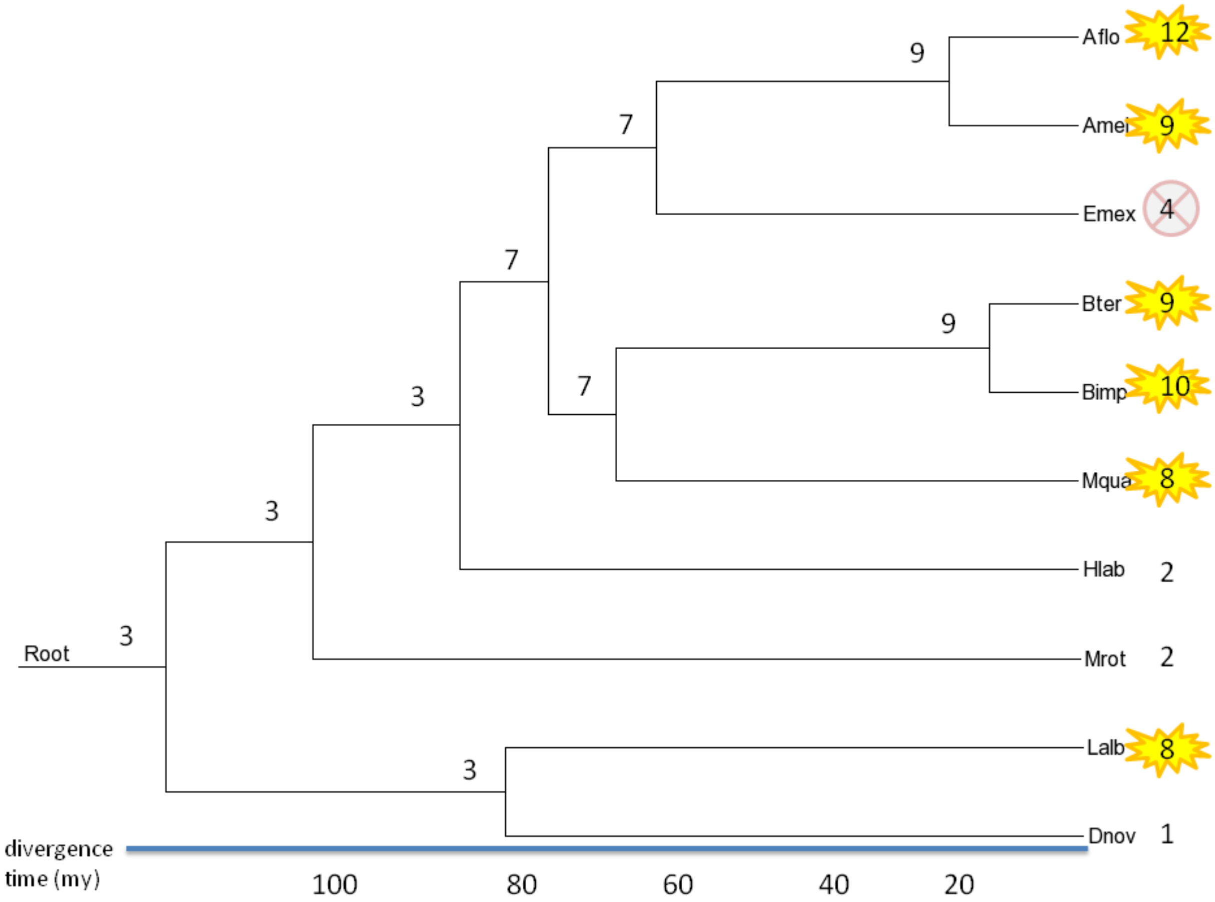

Fig. S32.

Expansion of gene family 395 (related to odorant binding) in all eusocial species except *E. mexicana*, which had a loss of these genes. Numbers indicate the number of genes in this family in each lineage. Yellow stars highlight expansion in terminal branches, circled-X highlights contraction in terminal branches. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*.

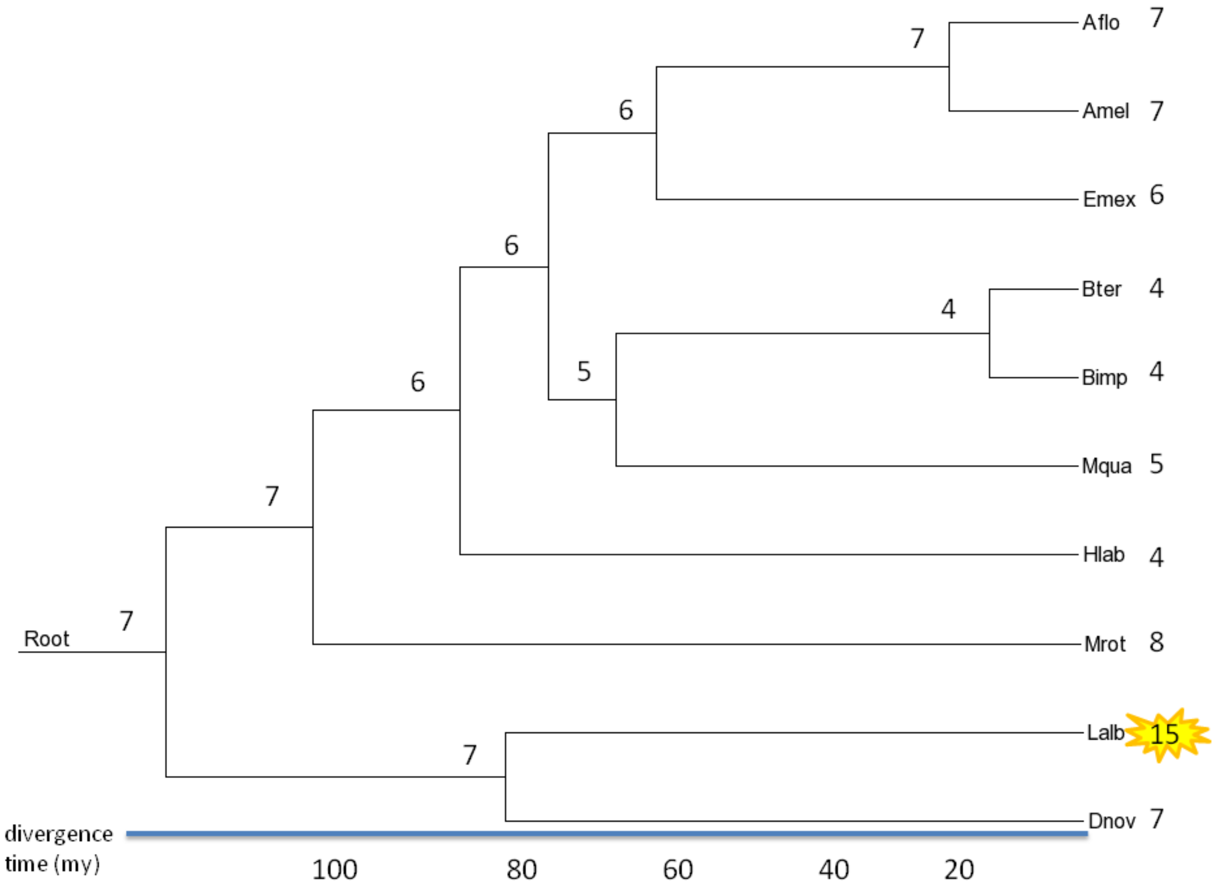

Fig. S33.

Expansion of gene family 5886 (related to lipid metabolism) in the facultatively eusocial *L. albipes*. Numbers indicate the number of genes in this family in each lineage. Yellow star highlights expansion in terminal branch. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens,* Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

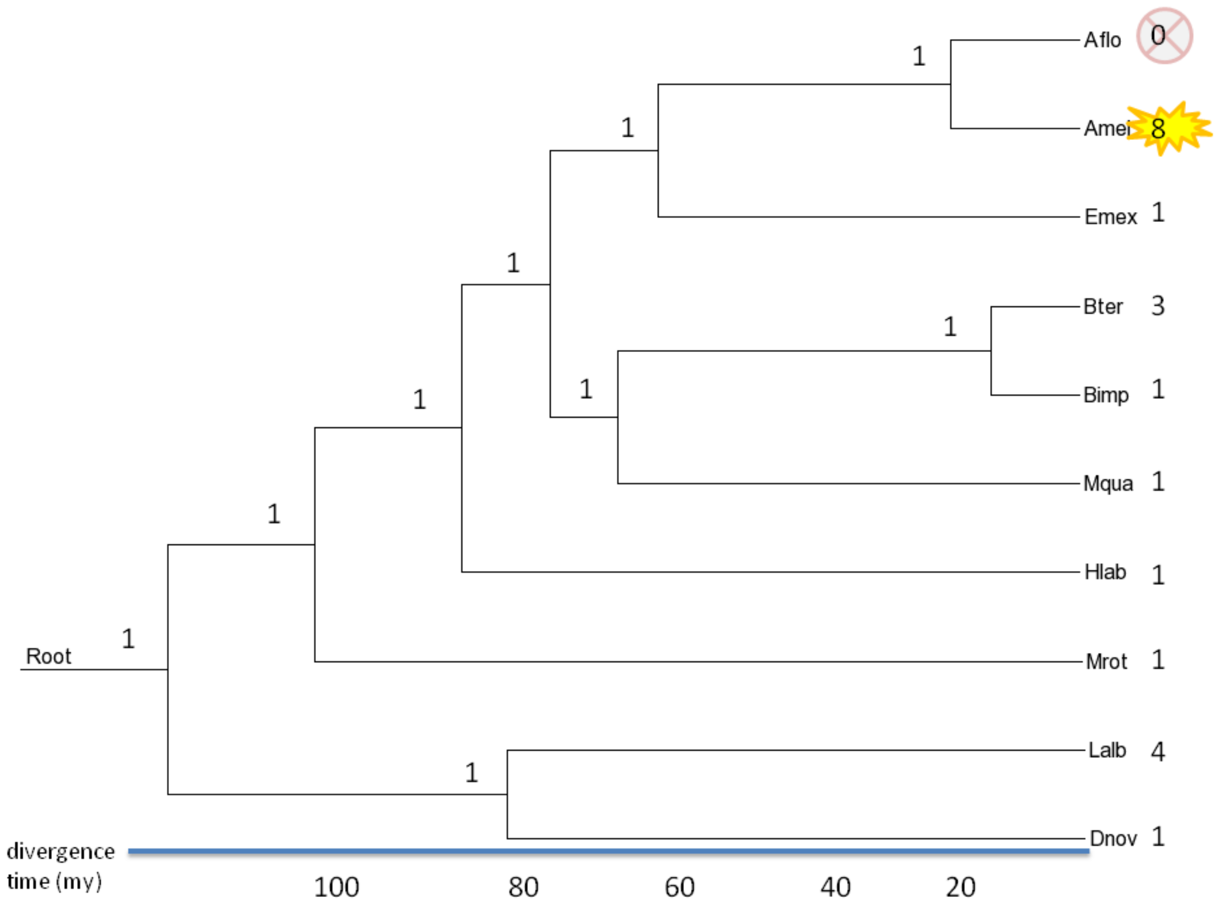

Fig. S34.

Expansion of gene family 221 (related to lipid metabolism) in honeybees with subsequent loss in the dwarf honeybee (*A. florea*). Numbers indicate the number of genes in this family in each lineage. Yellow star highlights expansion in terminal branch; circled-cross highlights contraction on terminal branch. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*.

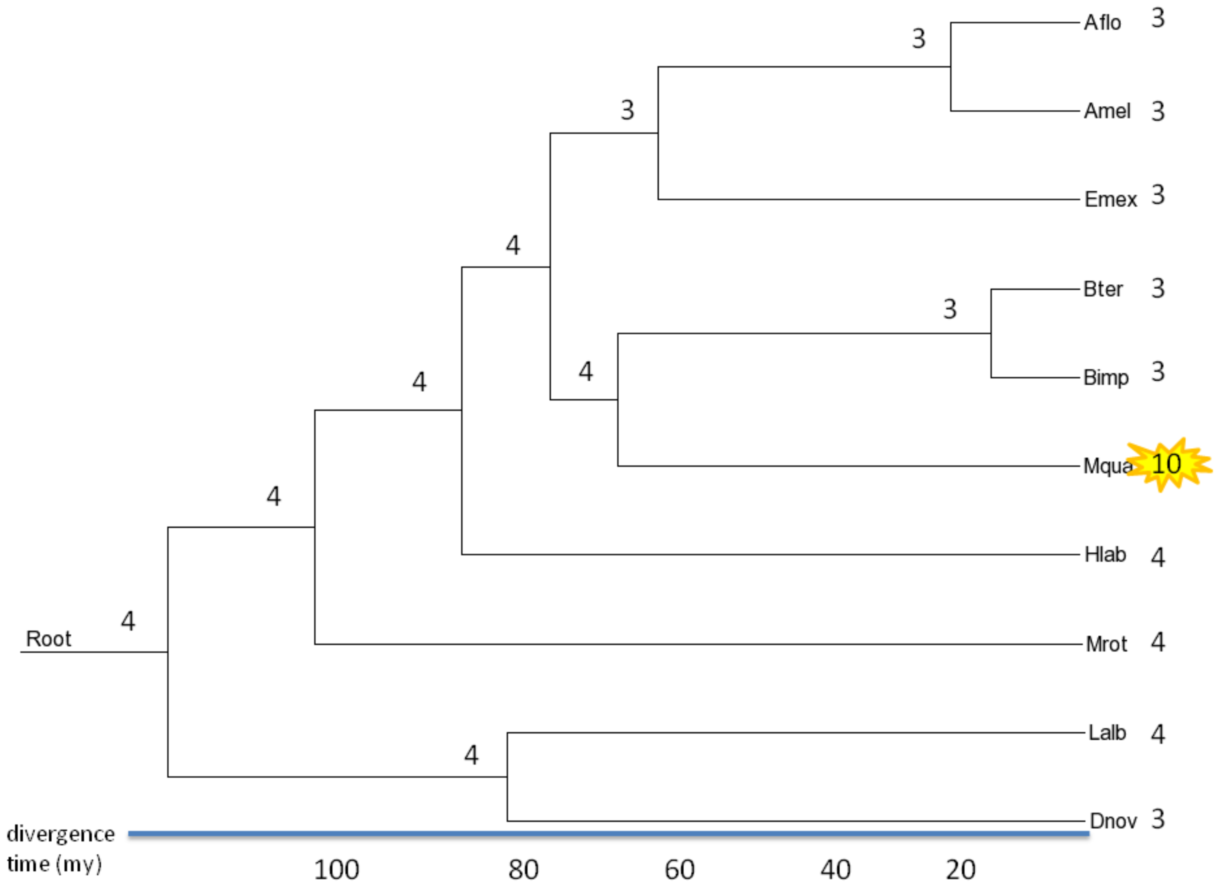

Fig. S35.

Expansion of gene family 2565 (related to lipid metabolism) in stingless bees (*M. quadrifasciata*). Numbers indicate the number of genes in this family in each lineage. Yellow star highlights expansion in terminal branch. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

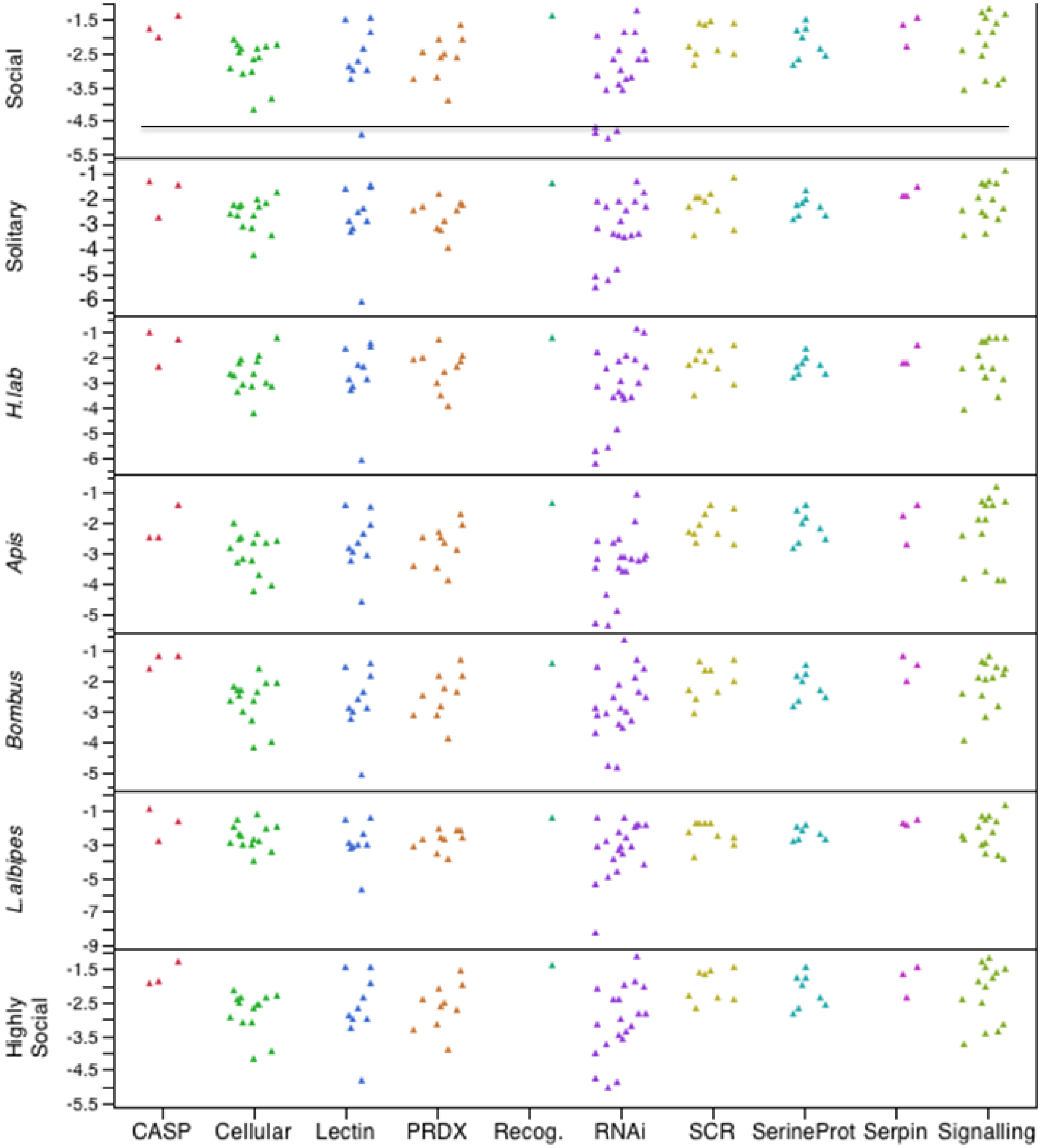

Fig. S36.

dN/dS generated for 97 orthologous proteins related to immunity divided by taxon and social structure.

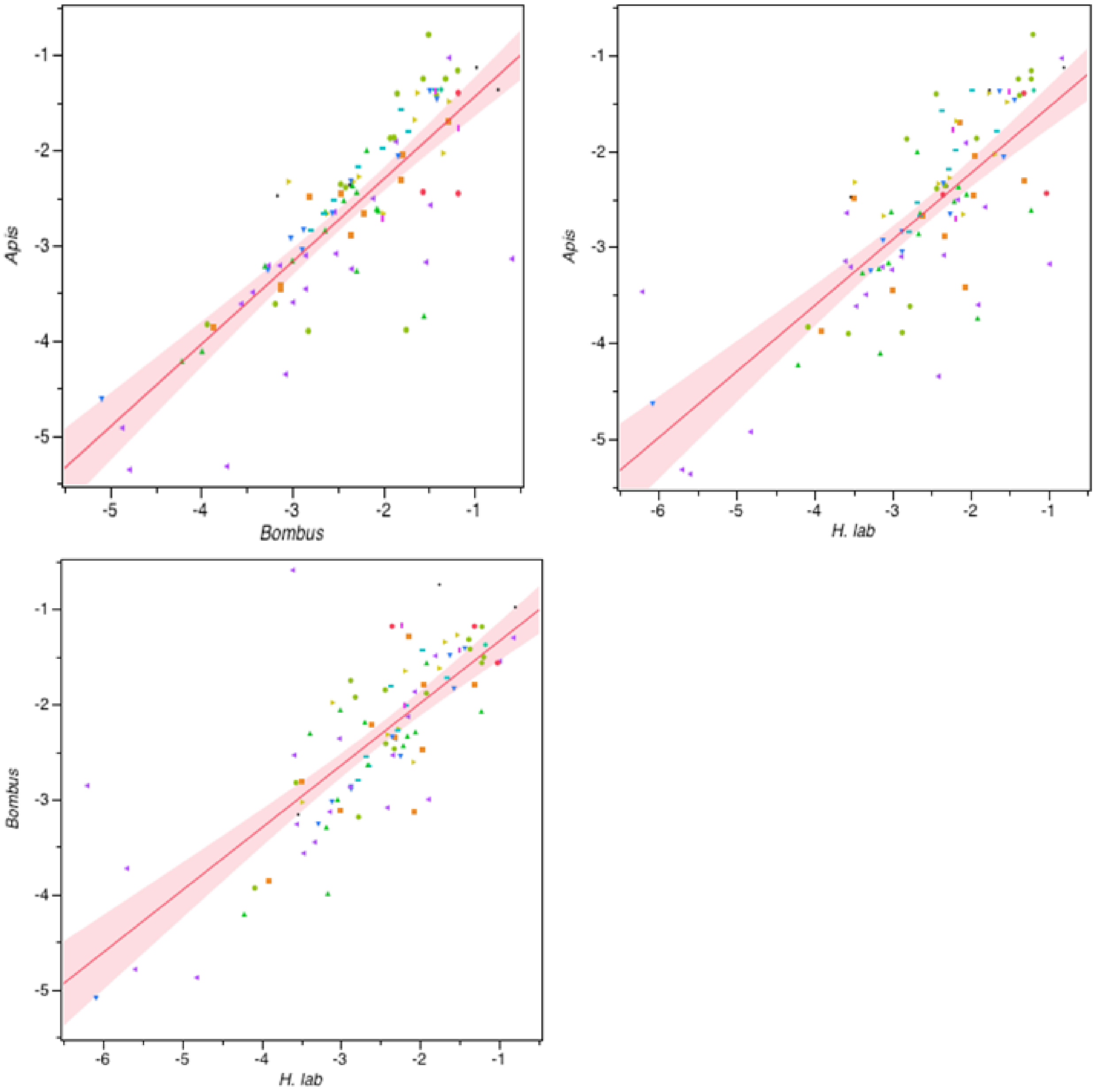

Fig. S37.
dN/dS generated for 97 orthologous proteins related to immunity contrasted between *Apis*, *Bombus*, and *Habropoda* as pairwise charts.

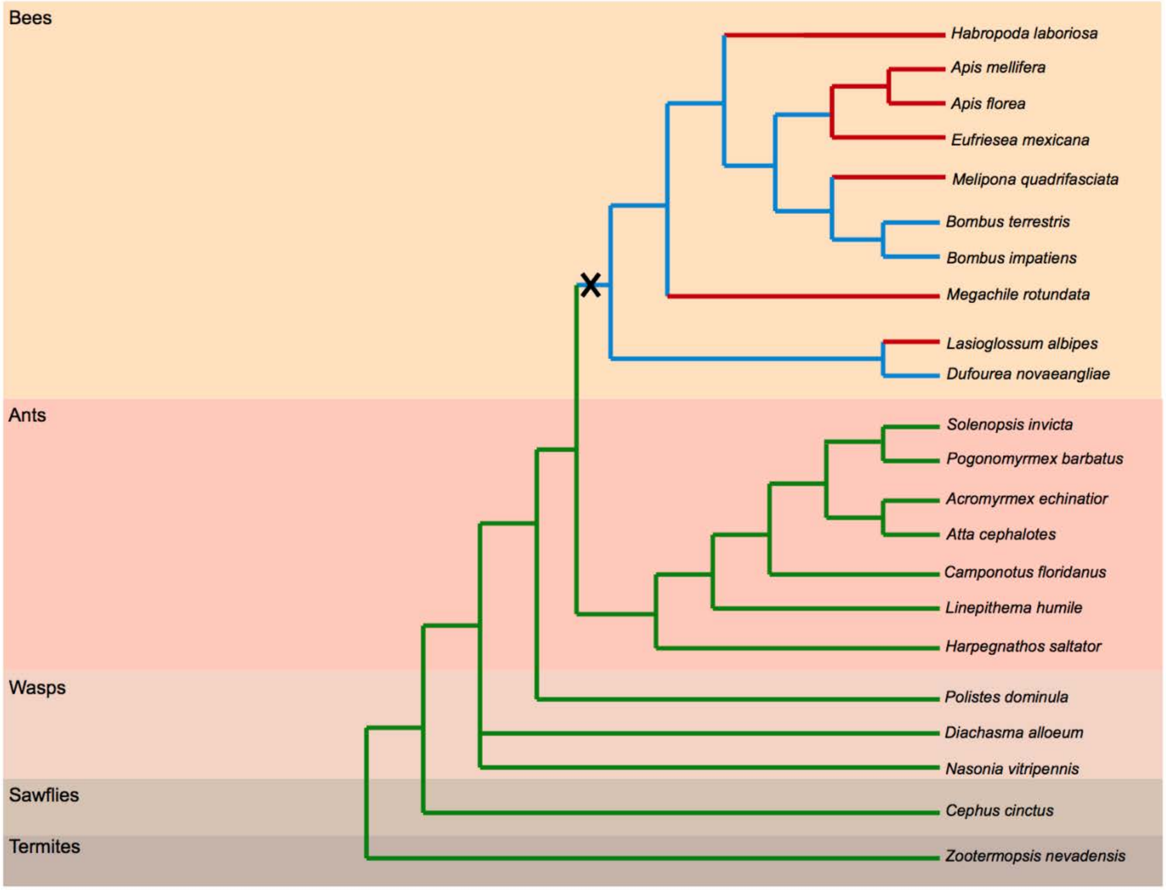

Fig. S38.

Cladogram illustrating the evolutionary loss of the inotocin receptor in the Hymenoptera. A termite (included as a social insect outgroup), sawfly, three wasps, and seven ants all possess intact genes for the inotocin hormone receptor and ligand (green). The “X” at the base of bees denotes a pseudogenization event that disrupted the inotocin receptor during the evolution of bees. The bee species *Bombus impatiens*, *Bombus terrestris*, and *Dufourea novaeangliae* all possess pseudogene remnants of the inotocin receptor (blue). The rest of the bee species analyzed, *Megachile rotundata*, *Melipona quadrifasciata*, *Apis florea*, *Apis mellifera*, *Eufriesea mexicana* and *Habropoda laboriosa* do not have any remains of the inotocin receptor pseudogene (red). None of the bee species in this cladogram possess any remains of the gene encoding the inotocin precursor ligand.

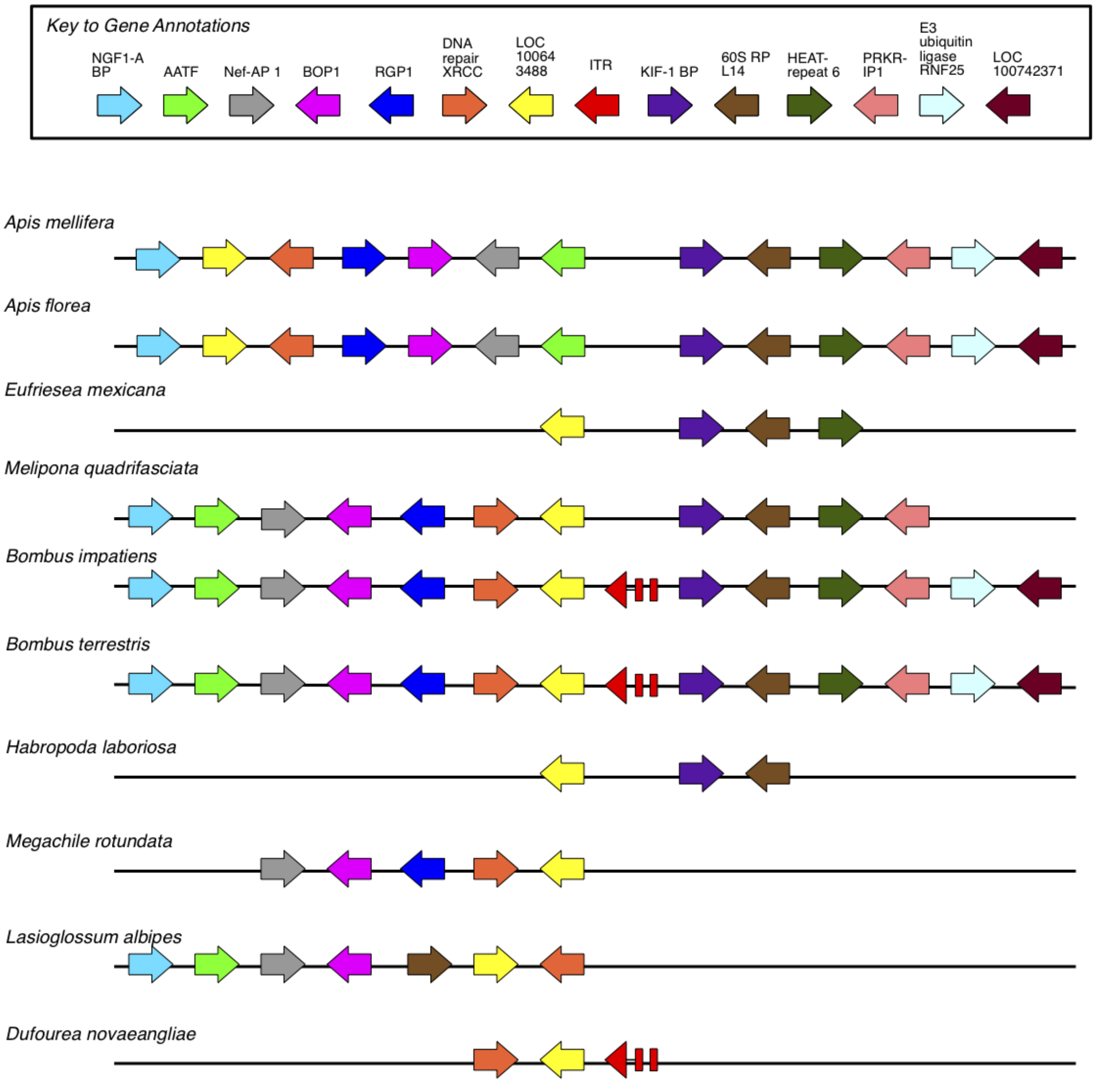

Fig. S39.

Microsynteny analysis of the DNA region containing the inotocin receptor (*ITR*) pseudogene in bees. The color of each arrow denotes specific genes in this region with predicted annotations. The direction of each arrow denotes the orientation of each gene at this locus. The broken red arrow illustrates that the inotocin receptor (*ITR*) is a pseudogene remnant in these species.

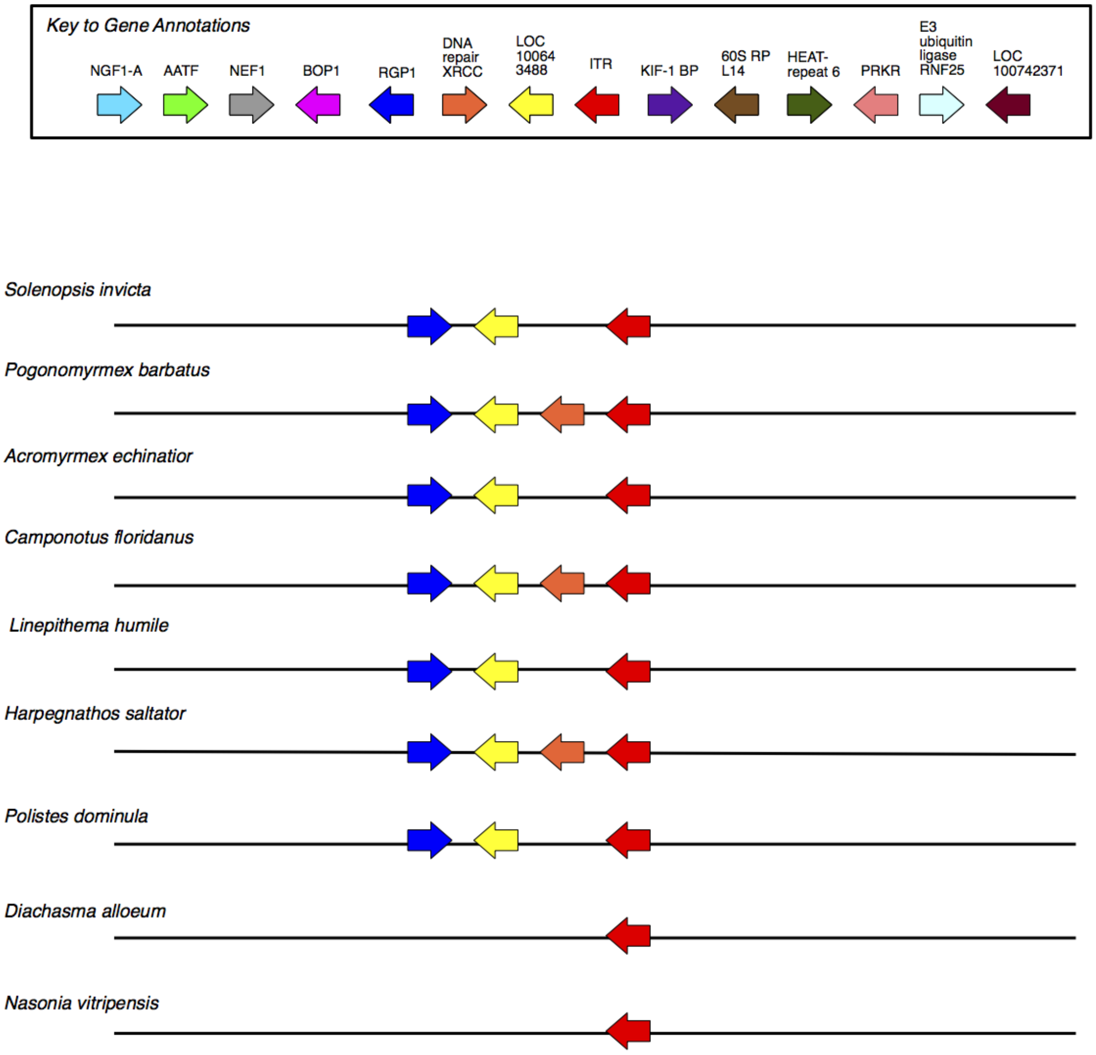

Fig. S40.

Microsynteny analysis of the DNA region containing the intact inotocin receptor gene (*ITR*) in ants, wasps, sawflies, and termites. A few genes in ants (*RGP1*, *loc100643488*, and *DNA repair XRCC*) are microsyntenic with *ITR*, and *loc100643488* is microsyntenic with *ITR* in the wasp *Polistes dominula*.

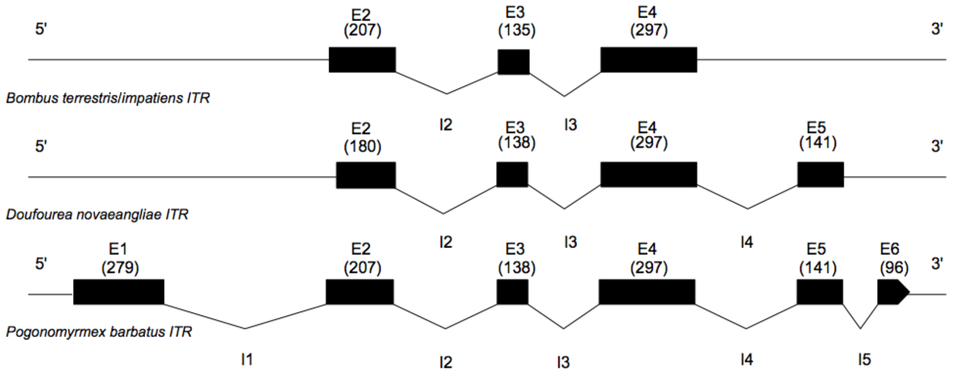

Fig. S41.

Gene and pseudogene structures of ant (*Pogonomyrmex barbatus*) *ITR*, *Dufourea novagensis ITR*, and *Bombus terrestris*/*impatiens ITR*. Exons are shown in black with size in bp of each in parentheses. Introns are indicated between each exon by indented lines.

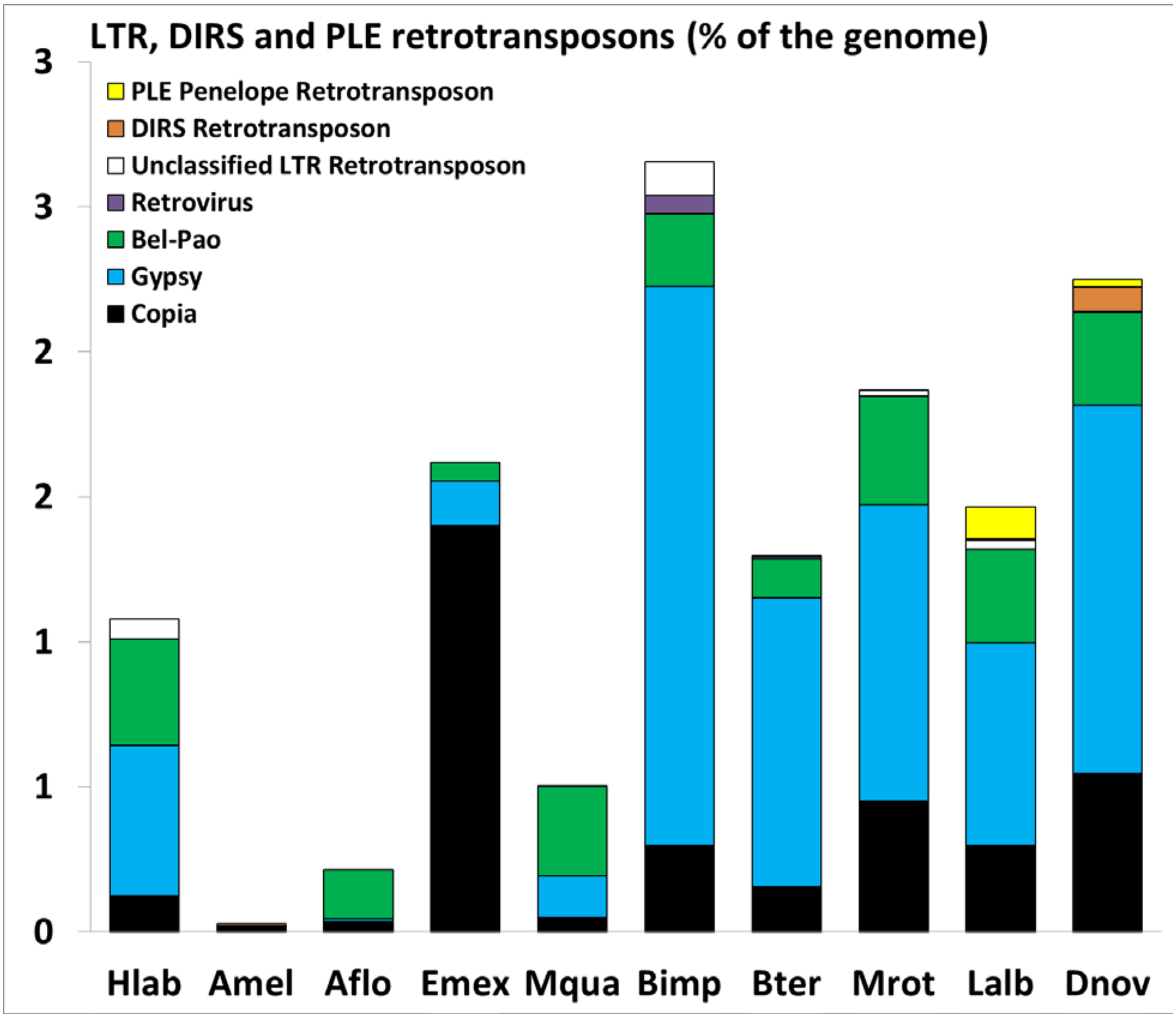

Fig. S42.

LTR, DIRs, and PLE retrotransposons (% of the genome). Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

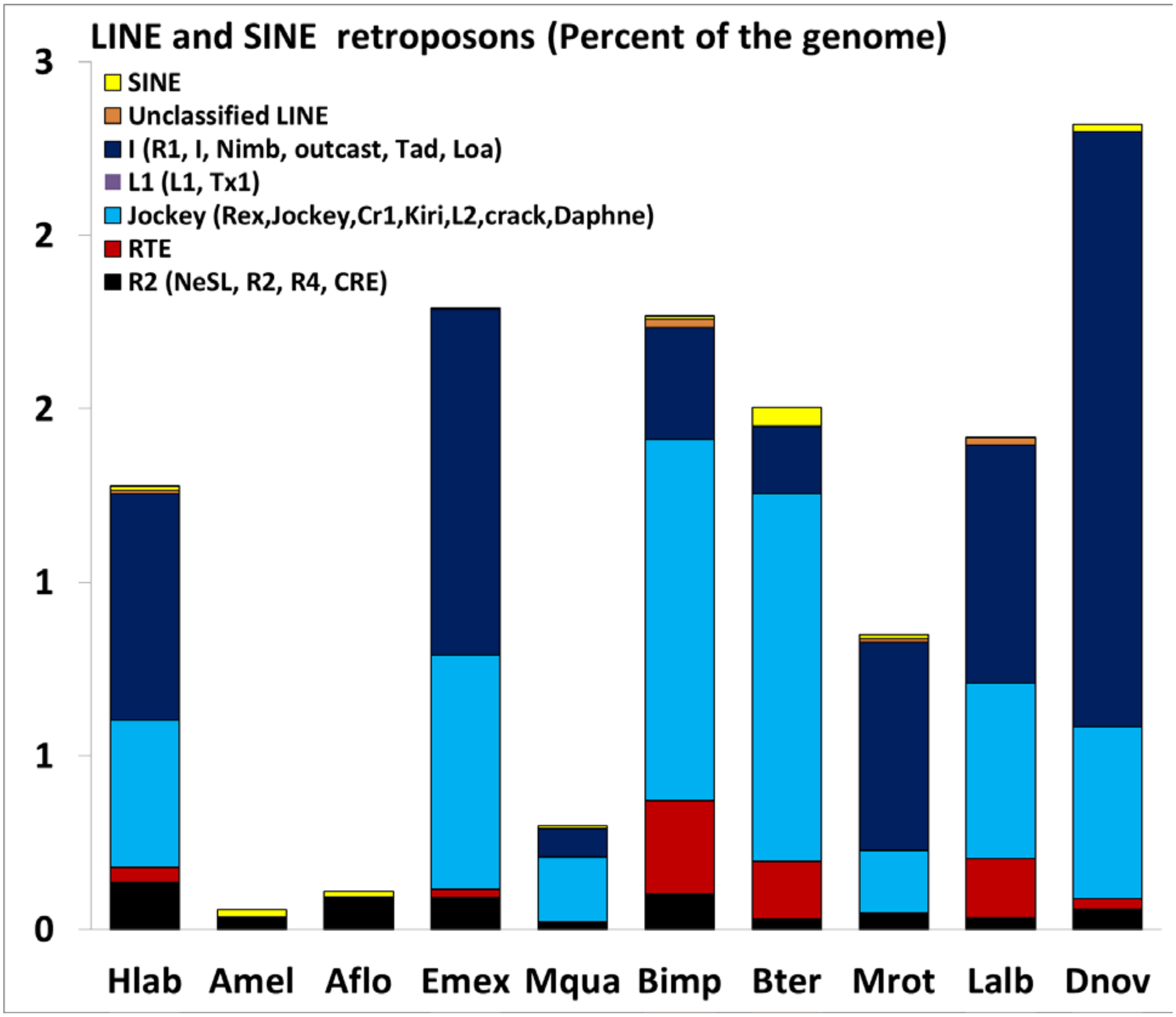

Fig. S43.

LINE and SINE retrotransposons (% of the genome) in each of the ten bees. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

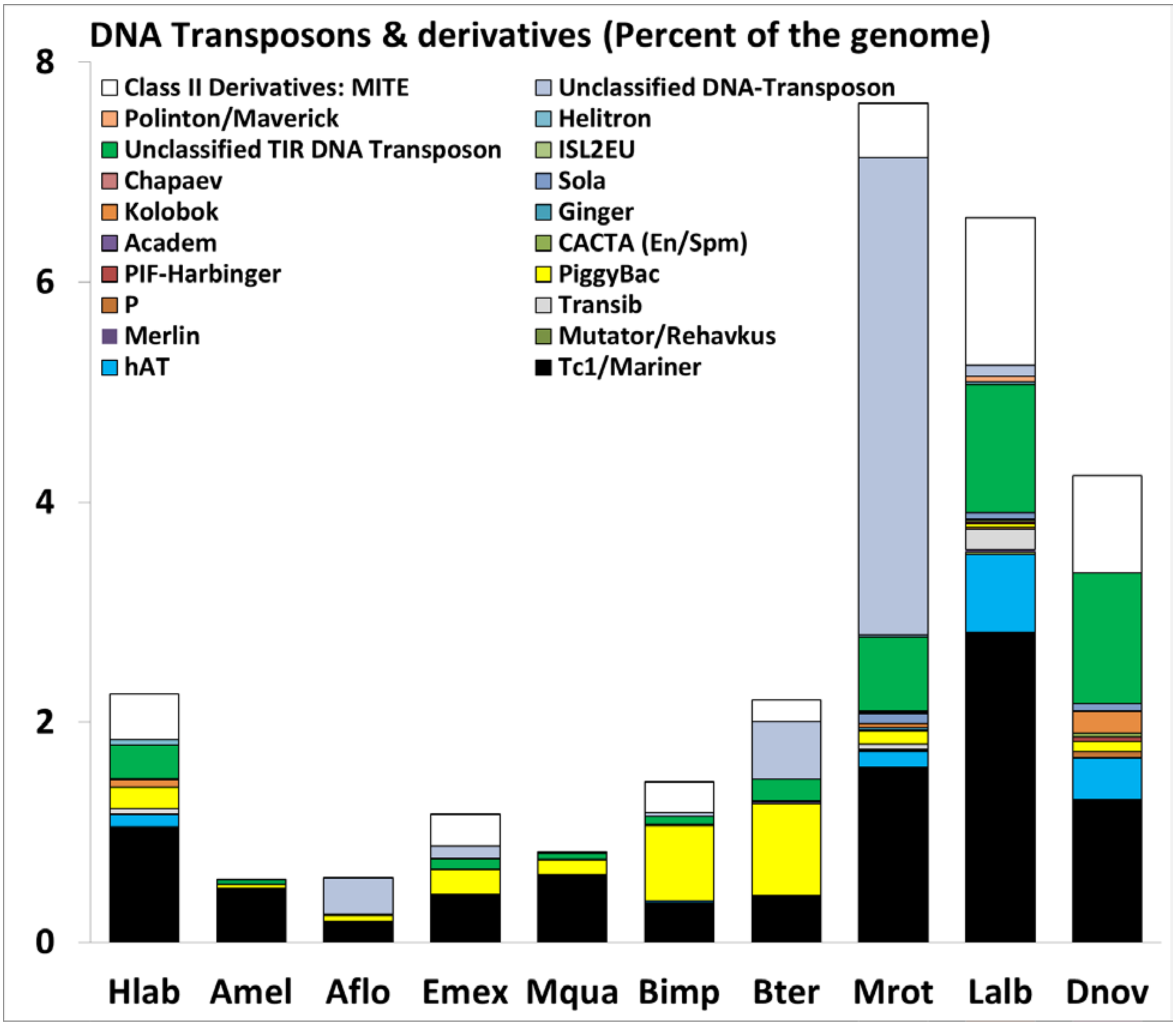

Fig. S44.

DNA transposons & derivatives (% of the genome). Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*.

Table S1.
De novo genome sequencing, assembly, and annotation methods.

Species	Sequencing	Insert sizes	Amount of sequencing	Assembly	Annotation
Megachile rotundata	Illumina GAIIx (U Illinois)	475 bp, 1.1 kb, 1.5 kb, 3 kb, 5.3 kb, 8-10 kb	74,330 Mbp	SOAPdenovo + GapCloser	Augustus, GlimmerHMM and SNAP + homology + transcriptome – integrated with GLEAN
Eufriesea mexicana	Illumina HiSeq (BGI)	170 bp, 500 bp, 2 kb, 5 kb, 10 kb	75,844 Mbp	SOAPdenovo	
Melipona quadrifasciata		170 bp, 500 bp, 2 kb, 5 kb	123,860 Mbp	SOAPdenovo + ALLPATHS LG	
Dufourea novaeangliae		170 bp, 500 bp, 2 kb, 5 kb	32,567 Mbp	SOAPdenovo	
Habropoda laboriosa		170 bp, 500 bp, 2 kb, 5 kb, 10 kb, 20 kb	38,786 Mbp	SOAPdenovo	

Table S2.

The 26 insect species and annotation sets used for complete OrthoDB orthology delineation.

Species Gene Set	Source	Gene Count
ACEPH_v1.2	Hymenoptera Genome Database, http://hymenopteragenome.org/ant_genomes	18,062
AECHI_v3.8	Hymenoptera Genome Database, http://hymenopteragenome.org/ant_genomes	17,277
AFLOR_v1.1	This project's re-annotation	15,810
AGAMB_v3.7	VectorBase, https://www.vectorbase.org/downloads	12,810
AMELL_v3.2	http://www.hgsc.bcm.tmc.edu/collaborations/insects/bees/Amel_4.5_OGSv3.2	15,314
APISU_v2.1	AphidBase, http://www.aphidbase.com/aphidbase/downloads	36,275
BIMPA_v1.2	This project's re-annotation	13,049
BMORI_GLEAN	SilkDB, http://www.silkdb.org/silkdb/doc/download.html	14,623
BTERR_v1.3	This project's re-annotation	12,648
CCINC_MAKER	Hugh Robertson: http://weatherby.genetics.utah.edu/daniel_downloads/cephus_cinctus	11,206
CFLOR_v3.3	Hymenoptera Genome Database, http://hymenopteragenome.org/ant_genomes	17,015
DMELA_v5.50	FlyBase, ftp://ftp.flybase.net/releases/FB2013_02/dmel_r5.50	13,967
DNOVA_v1.1	This project's re-annotation	12,453
DPLEX_v2.0	MonarchBase, http://monarchbase.umassmed.edu/resource.html	15,130
EMEXI_v1.1	This project's re-annotation	12,022
HLABO_v1.2	This project's re-annotation	13,279
HSALT_v3.3	Hymenoptera Genome Database, http://hymenopteragenome.org/ant_genomes	18,518
LALBI_v5.4	This project's re-annotation	13,448
LHUMI_v1.2	Hymenoptera Genome Database, http://hymenopteragenome.org/ant_genomes	16,048
MQUAD_v1.1	This project's re-annotation	15,368
MROTU_v1.1	This project's re-annotation	12,770
NVITR_v2.0	http://arthropods.eugenes.org/EvidentialGene/nasonia/genes	24,369
PBARB_v1.2	Hymenoptera Genome Database, http://hymenopteragenome.org/ant_genomes	17,100
PHUMA_v1.2	VectorBase, https://www.vectorbase.org/downloads	10,772
SINVI_v2.2.3	Hymenoptera Genome Database, http://hymenopteragenome.org/ant_genomes	16,513
TCAST_v3.0	BeetleBase, ftp://ftp.bioinformatics.ksu.edu/pub/BeetleBase/3.0	16,565

Table S3.

De novo genome sequencing, assembly, and annotation results.

			Assembly				
Species	Genome size (Mbp)[ǂ]	Sequencing depth (X)	Contig N50 (bp)	Scaffold N50 (bp)	GC content	# Genes	CEGMA[§] genes (>80% overlap)
Megachile rotundata	273	272.3	64,153	1,699,680	37%	12,770	247 (196)
Eufriesea mexicana	1,032[¥]	120.0	883	2,427	41%	12,022	247 (193)
Melipona quadrifasciata	257	126.7	11,556	1,864,352	39%	15,368	245 (191)
Dufourea novaeangliae	291[Ψ]	133.3	23,525	2,397,596	40%	12,453	247 (193)
Habropoda laboriosa	377	201.4	14,958	1,338,707	39%	13,279	247 (208)

[ǂ] Based on k-mer analysis of sequences.

[§] Number of genes found to correspond to CEGMA genes and, in (), the number of genes with an 80% overlap with CEGMA genes.

[¥] Flow cytometry estimate is 1939.7 ± 41.6 Mbp.

[Ψ] Flow cytometry estimate is 342.0 Mbp.

Table S4.

Protein coding gene properties among ten bees. Aflo, Bimp, and Bter were re-annotated with the pipeline used for the five de novo species. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*.

Species	Genome size (Mb)	Number	Average gene length (kb)	Average CDS length (kb)	Average exons per gene	Average exon length (bp)	Average intron length (kb)
Amel	234.07	15,314	6.94	1.27	5.32	237.82	1.31
Aflo	230.47	15,810	7.67	1.27	5.20	244.93	1.52
Emex	1,031.84	12,022	8.00	1.56	6.22	250.76	1.23
Mqua	256.95	15,368	12.08	1.39	6.73	206.79	1.87
Bimp	249.19	13,050	9.58	1.54	5.96	259.05	1.62
Bter	248.66	12,648	9.92	1.51	6.00	251.70	1.68
Hlab	376.63	13,279	7.52	1.42	5.46	260.43	1.37
Mrot	272.66	12,770	10.26	1.55	6.10	254.28	1.71
Lalb	340.69	13,448	7.75	1.47	5.72	257.29	1.33
Dnov	290.96	12,453	8.81	1.50	5.74	261.44	1.54

Table S5.

Functional annotation of the re-annotated ten bee genomes. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

Species	number(% of total gene)				
	InterPro	**GO**	**KEGG**	**Swissprot**	**Annotated**
Amel	9150 (59.75)	7314(47.76)	5484(35.81)	9421(61.52)	10034(65.52)
Aflo	8613 (54.48)	6910(43.71)	5265(33.30)	8782(55.55)	9343(59.10)
Emex	8347 (69.43)	6757(56.21)	5202(43.27)	8656(72.00)	9115(75.82)
Mqua	8592 (55.91)	6853(44.59)	5318(34.60)	8929(58.10)	9521(61.95)
Bimp	8465 (64.87)	6815(52.22)	5505(42.18)	9099(69.72)	9579(73.40)
Bter	8488 (67.11)	6777(53.58)	5069(40.08)	8697(68.76)	9211(72.83)
Hlab	8563 (64.49)	6802(51.22)	5281(39.77)	9081(68.39)	9685(72.93)
Mrot	8205 (64.25)	6573(51.47)	4994(39.11)	8683(68.00)	9193(71.99)
Lalb	8686 (64.59)	6933(51.55)	5343(39.73)	9203(68.43)	10001(74.37)
Dnov	8287 (66.55)	6680(53.64)	5082(40.81)	8565(68.78)	9033(72.54)

Table S6.

Transcription factors (TF) in ten bees. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*.

Species	# genes with TF domains	# genes with basal TF domain	# genes with other TF domain	# TF domain types detected	# other TF domain types detected	# basal TF domain types detected	top TF	2nd top TF	3rd top TF
Aflo	667	45	629	160	125	35	zf-C2H2	Homeobox	HLH
Amel	695	45	734	160	125	35	zf-C2H2	Homeobox	HLH
Bimp	662	44	623	151	117	34	zf-C2H2	Homeobox	HLH
Bter	657	43	621	148	114	34	zf-C2H2	Homeobox	HLH
Dnov	670	46	630	150	115	35	zf-C2H2	Homeobox	HLH
Emex	623	42	659	146	113	33	zf-C2H2	Homeobox	HLH
Hlab	690	45	651	146	112	34	zf-C2H2	Homeobox	HLH
Lalb	745	49	702	147	114	33	zf-C2H2	Homeobox	HLH
Mqua	652	47	611	145	111	34	zf-C2H2	Homeobox	HLH
Mrot	652	45	613	142	108	34	zf-C2H2	Homeobox	HLH

Table S7.
Summary of TFs absent in orthogroups by DNA binding domain. For each family of DNA binding domain, lists of the 223 TF motifs, the total number and the number and percentage of motifs which had none of the represented TFs in any OrthoDB orthogroup.

Domain	**Total**	**Missing**	**%Missing**
ZF	78	19	24.4%
bHLH	31	4	12.9%
HD	25	1	4.0%
MADF	20	1	5.0%
bZIP	11	2	18.2%
other	58	7	12.1%
total	**223**	**34**	**15.2%**

Table S8.

Association results between molecular evolution and motif analysis. Each row lists the name of the gene set from the PAML analysis of molecular evolution and the name of the motif that defined a gene set from the PIC analysis (as well as the family of the DNA binding domain of its TF). For each threshold on defining the motif gene set (0.1, 0.05, 0.1, top 100000 results), significant association p-values are listed as well as the best across all thresholds.

Set Name	Motif	Domain	0.01	0.05	0.1	100000	Best Pval
F1: solitary to basic eusociality - Apidae	lola_PQ_SOLEXA	ZF		0.0047			0.0047
F2: solitary to basic eusociality - Halictidae	br_PL_SOLEXA_5	ZF		0.0016			0.0016
F3: basic to complex eusociality - Apinae A	h_SOLEXA_5	bHLH				0.0027	0.0027
F4: basic to complex eusociality - Apinae B	Side_SOLEXA_5	bHLH		0.0008			0.0008
F4: basic to complex eusociality - Apinae B	usp_SOLEXA	NHR				0.0013	0.0013
F4: basic to complex eusociality - Apinae B	gsb_SOLEXA	PAX	0.0020		0.0080		0.0020
F4: basic to complex eusociality - Apinae B	ttk_PA_SOLEXA_5	ZF		0.0045	0.0060	0.0033	0.0033
F4: basic to complex eusociality - Apinae B	CrebA_SOLEXA	bZIP			0.0040		0.0040
F4: basic to complex eusociality - Apinae B	CG5180_SOLEXA	MADF				0.0044	0.0044
F4: basic to complex eusociality - Apinae B	tai_Met_SOLEXA_5	bHLH	0.0045				0.0045
F4: basic to complex eusociality - Apinae B	sug_SOLEXA_5	ZF			0.0065		0.0065
F4: basic to complex eusociality - Apinae B	tai_SOLEXA_5	bHLH	0.0100				0.0100
F4: basic to complex eusociality - Apinae B	Max_Mnt_SOLEXA_5	bHLH			0.0100		0.0100

Table S9.

Patterns of association results between molecular evolution and motif analysis. For each association, results identified in Table S8, lists the name of the PAML gene set and the name of the motif defining the gene sets from the PIC analysis. The number of orthogroups with a 3-fold change in molecular evolution rate and significant (p-value < 0.25) correlation between motif and sociality level are reported ("Sig") as well as the percentage these orthogroups form of all orthogroups from the PAML gene set ("Sig%"). The number of orthogroup that specifically follow the pattern where higher social complexity is correlated with stronger motif presences and a slower evolutionary rate are listed ("Pattern") as well as their percentage of the "Sig" orthogroups.

Set Name	Motif	Sig	Sig%	Pattern	Pattern%
F1: solitary to basic eusociality - Apidae	lola_PQ_SOLEXA	11	3.10%	5	45.50%
F2: solitary to basic eusociality - Halictidae	br_PL_SOLEXA_5	14	8.40%	9	64.30%
F3: basic to complex eusociality - Apinae A	h_SOLEXA_5	27	4.30%	14	51.90%
F4: basic to complex eusociality - Apinae B	Side_SOLEXA_5	20	3.90%	10	50.00%
F4: basic to complex eusociality - Apinae B	usp_SOLEXA	24	4.70%	17	70.80%
F4: basic to complex eusociality - Apinae B	gsb_SOLEXA	21	4.10%	14	66.70%
F4: basic to complex eusociality - Apinae B	ttk_PA_SOLEXA_5	23	4.50%	15	65.20%
F4: basic to complex eusociality - Apinae B	CrebA_SOLEXA	24	4.70%	18	75.00%
F4: basic to complex eusociality - Apinae B	CG5180_SOLEXA	18	3.50%	9	50.00%
F4: basic to complex eusociality - Apinae B	tai_Met_SOLEXA_5	26	5.10%	14	53.80%
F4: basic to complex eusociality - Apinae B	sug_SOLEXA_5	23	4.50%	14	60.90%
F4: basic to complex eusociality - Apinae B	tai_SOLEXA_5	20	3.90%	11	55.00%
F4: basic to complex eusociality - Apinae B	Max_Mnt_SOLEXA_5	25	4.90%	10	40.00%

Table S10.

Numbers of genes that are putatively methylated, unmethylated, or of undetermined methylation for each of 10 bee species. Aflor-*Apis florea*, Amel-*Apis mellifera*, Bterr-*Bombus terrestris*, Bimpa-*Bombus impatiens*, Dnova-*Dufourea novaengliae*, Emexi-*Eufriesea mexicana*, Lalbi-*Lasioglossum albipes*, Mrotu-*Megachile rotundata*, Mquad-*Melipona quadrifasciata*, Hlabo-*Habropoda laboriosa*.

	methylated	Undetermined	unmethylated
AMELL	2456	2649	1033
AFLOR	2642	3013	563
BIMPA	2436	2542	1160
BTERR	1997	3578	663
DNOVA	1857	3174	1107
EMEXI	2194	2898	1046
HLABO	2078	2969	1091
LALBI	531	5049	558
MQUAD	2249	3769	320
MROTU	1522	3847	669
Mean ± Stdev	1996.2 ± 576.4	3348.8 ± 707.1	821.0 ± 282.8

Table S11.

GO terms enriched ($p < 0.05$) among genes under positive selection in association with social complexity in ten bees.

Category	Term	Count	PValue	Fold Enrichment	Benjamini
GOTERM_BP_FAT	GO:0006096~glycolysis	3	0.010337	18.42308	0.992272
GOTERM_BP_FAT	GO:0019320~hexose catabolic process	3	0.027846	11.05385	0.998651
GOTERM_BP_FAT	GO:0006007~glucose catabolic process	3	0.027846	11.05385	0.998651
GOTERM_BP_FAT	GO:0006412~translation	7	0.028634	2.865812	0.989241
GOTERM_BP_FAT	GO:0046365~monosaccharide catabolic process	3	0.030524	10.52747	0.973403
GOTERM_BP_FAT	GO:0046164~alcohol catabolic process	3	0.0333	10.04895	0.957996
GOTERM_BP_FAT	GO:0044275~cellular carbohydrate catabolic process	3	0.036172	9.61204	0.943511
GOTERM_CC_FAT	GO:0033279~ribosomal subunit	5	0.015286	4.822531	0.665015
GOTERM_CC_FAT	GO:0005761~mitochondrial ribosome	4	0.018395	6.648936	0.482685
GOTERM_CC_FAT	GO:0000313~organellar ribosome	4	0.018395	6.648936	0.482685
GOTERM_CC_FAT	GO:0005840~ribosome	5	0.026078	4.111842	0.464937
GOTERM_CC_FAT	GO:0015935~small ribosomal subunit	3	0.044849	8.370536	0.557129
GOTERM_MF_FAT	GO:0003735~structural constituent of ribosome	5	0.033392	3.949973	0.989795
GOTERM_MF_FAT	GO:0004045~aminoacyl-tRNA hydrolase activity	2	0.041091	46.87302	0.941118

Table S12.

GO terms enriched ($p < 0.05$) among genes under relaxed selection in association with social complexity in ten bees.

Category	Term	Count	PValue	Fold Enrichment	Benjamini
GOTERM_BP_FAT	GO:0006397~mRNA processing	7	0.001413	5.330684	0.477439
GOTERM_BP_FAT	GO:0006367~transcription initiation from RNA polymerase II promoter	5	0.001808	9.032055	0.339869
GOTERM_BP_FAT	GO:0006352~transcription initiation	5	0.002328	8.443008	0.299905
GOTERM_BP_FAT	GO:0016071~mRNA metabolic process	7	0.00312	4.569157	0.301319
GOTERM_BP_FAT	GO:0006366~transcription from RNA polymerase II promoter	5	0.00614	6.472973	0.431846
GOTERM_BP_FAT	GO:0008380~RNA splicing	5	0.012208	5.320252	0.609243
GOTERM_BP_FAT	GO:0006351~transcription, DNA-dependent	5	0.015978	4.916182	0.652218
GOTERM_BP_FAT	GO:0032774~RNA biosynthetic process	5	0.017382	4.794795	0.634339
GOTERM_BP_FAT	GO:0006396~RNA processing	7	0.021852	3.037596	0.675933
GOTERM_BP_FAT	GO:0006461~protein complex assembly	5	0.028394	4.131685	0.733431
GOTERM_BP_FAT	GO:0070271~protein complex biogenesis	5	0.028394	4.131685	0.733431
GOTERM_BP_FAT	GO:0006357~regulation of transcription from RNA polymerase II promoter	5	0.034623	3.883784	0.770154
GOTERM_CC_FAT	GO:0044451~nucleoplasm part	7	0.01242	3.348214	0.572516
GOTERM_CC_FAT	GO:0031981~nuclear lumen	8	0.020051	2.665245	0.497749
GOTERM_CC_FAT	GO:0005654~nucleoplasm	7	0.020493	3.004808	0.37458
GOTERM_MF_FAT	GO:0016251~general RNA polymerase II transcription factor activity	6	5.35E-04	8.401138	0.069248
GOTERM_MF_FAT	GO:0003702~RNA polymerase II transcription factor activity	8	9.37E-04	4.764825	0.060864
GOTERM_MF_FAT	GO:0016563~transcription activator activity	4	0.020882	6.515168	0.610396
GOTERM_MF_FAT	GO:0003713~transcription coactivator activity	3	0.021697	12.60171	0.520416
GOTERM_MF_FAT	GO:0000166~nucleotide binding	14	0.026872	1.793501	0.518096
GOTERM_MF_FAT	GO:0030528~transcription regulator activity	9	0.033859	2.258797	0.53666
GOTERM_MF_FAT	GO:0016455~RNA polymerase II transcription mediator activity	3	0.039101	9.20894	0.53398

Table S13.
GO terms enriched ($p < 0.05$) among genes under purifying selection in association with social complexity in ten bees.

Category	Term	Count	PValue	Fold Enrichment	Benjamini
GOTERM_BP_FAT	GO:0007041~lysosomal transport	3	0.006885	22.45313	0.981304
GOTERM_BP_FAT	GO:0015031~protein transport	8	0.012581	3.050955	0.973914
GOTERM_BP_FAT	GO:0007034~vacuolar transport	3	0.013108	16.32955	0.920606
GOTERM_BP_FAT	GO:0045184~establishment of protein localization	8	0.013433	3.012579	0.857373
GOTERM_BP_FAT	GO:0070727~cellular macromolecule localization	7	0.013838	3.40752	0.79916
GOTERM_BP_FAT	GO:0033060~ocellus pigmentation	3	0.018207	13.81731	0.828636
GOTERM_BP_FAT	GO:0008055~ocellus pigment biosynthetic process	3	0.018207	13.81731	0.828636
GOTERM_BP_FAT	GO:0046158~ocellus pigment metabolic process	3	0.018207	13.81731	0.828636
GOTERM_BP_FAT	GO:0046152~ommochrome metabolic process	3	0.018207	13.81731	0.828636
GOTERM_BP_FAT	GO:0006727~ommochrome biosynthetic process	3	0.018207	13.81731	0.828636
GOTERM_BP_FAT	GO:0008104~protein localization	9	0.019687	2.541863	0.805264
GOTERM_BP_FAT	GO:0006886~intracellular protein transport	6	0.020533	3.665816	0.775475
GOTERM_BP_FAT	GO:0034613~cellular protein localization	6	0.022217	3.5925	0.762587
GOTERM_BP_FAT	GO:0018130~heterocycle biosynthetic process	4	0.023177	6.302632	0.740948
GOTERM_BP_FAT	GO:0006726~eye pigment biosynthetic process	3	0.030457	10.56618	0.802027
GOTERM_BP_FAT	GO:0042441~eye pigment metabolic process	3	0.030457	10.56618	0.802027
GOTERM_BP_FAT	GO:0034622~cellular macromolecular complex assembly	5	0.035581	3.887987	0.824306
GOTERM_BP_FAT	GO:0048069~eye pigmentation	3	0.037513	9.453947	0.816231
GOTERM_CC_FAT	GO:0044459~plasma membrane part	8	0.033594	2.44898	0.979668
GOTERM_MF_FAT	GO:0008565~protein transporter activity	4	0.018686	6.835648	0.952916
GOTERM_MF_FAT	GO:0043566~structure-specific DNA binding	3	0.02582	11.53516	0.879835

Table S14.

Summary of clade-specific PAML results.

	# orthologs	# of significant genes	# significant after filtering dS > 2	dN/dS faster with increasing social complexity	dN/dS slower with increasing social complexity
Origins - Apidae	7477	356	354	169	185
Origins - Halictidae	7881	171	167	52	115
Elaborations – honeybees	7602	630	625	132	493
Elaborations – stingless bees	8288	514	512	94	418

Table S15.

OrthoGroups that were a significantly better fit to a phenotypic tree based on social complexity than a species phylogeny.

#family_ID	**H0-H1-1, ∆SSLS**	**p value**	**H0-H1-2,∆SSLS**	**p value**	**H0-H2-1,∆SSLS**	**p value**	**H0-H2-2,∆SSLS**	**p value**
APO011389	-0.088821385	0	-0.086768426	0	-0.068930805	0	-0.063323036	0
APO012406	-0.005493333	0	-0.002248949	0	-0.002242203	0	-0.002236342	0
APO007579	-0.005058937	0	-0.00162623	0.033	0.012009438	NA	0.004601045	NA
APO010863	-0.004913059	0	-0.009169975	0	0.013247869	NA	-0.006943956	0
APO008600	-0.005782891	0	-0.005783452	0	0.007590083	NA	-0.00325316	0
APO010032	-0.00043809	0	-0.000264701	0	-0.009123757	0	-0.008042697	0
APO007983	-0.001067432	0	0.006883438	NA	0.021403635	NA	0.013479183	NA
APO008527	-0.0016562	0	-0.000636956	0.067	-0.000804122	0.1	-0.000636987	0.067
APO006364	-0.004777289	0	0.003363384	NA	0.032659178	NA	0.016960827	NA
APO010577	0.002641128	NA	-0.000159429	0	0.008039735	NA	0.000355495	NA
APO010049	-0.009340548	0	-0.009340821	0	-0.006173867	0	-0.012804684	0
APO007670	-0.003396187	0	-0.002393883	0	-0.000750039	0.133	-0.002584817	0
APO010479	-0.001677179	0	0.085774216	NA	0.013046837	NA	0.095810477	NA
APO010594	-0.003499089	0.033	0.008718356	NA	0.011352644	NA	0.0085047	NA
APO008412	-0.01812622	0	-0.017207145	0	0.003116395	NA	-0.017764921	0
APO009459	-0.000454998	0.1	-0.000454723	0.1	-0.001272658	0	-0.001272316	0

Table S16.

Predicted protein size (amino acid residues) and exon numbers of mrjp/mrjp-like and the flanking yellow genes y-e3 and y-h of the ten bee species.

Species	Gene name	Protein size (aa)	Exon number
Apis florea	yellow-e3	423	5
	yellow-h	522	4
	mrjp 3	553	6
	mrjp 4	494	6
	mrjp 5	598	7
	mrjp	438	6
	mrjp	423	6
	mrjp	354	5
	yellow	413	5
Apis mellifera	yellow-e3	424	5
	yellow-h	552	4
	mrjp 1	432	6
	mrjp 2	452	6
	mrjp 3	544	6
	mrjp 4	464	6
	mrjp 5	589	7
	mrjp 6	437	6
	mrjp 7	445	6
	mrjp 8	416	6
	mrjp 9	423	6
Bombus impatiens	yellow-e3	424	5
	yellow-h	550	4
	mrjp (mrjp9-like)	411	6
Bombus terrestris	yellow-e3	424	5
	yellow-h	551	4
	mrjp (mrjp9-like)	411	6
Dufourea novaeangliae	yellow-e3	330	4
	yellow-h	563	4
	mrjp (mrjp9-like)	399	6
	yellow-like	323	3
Eufriesea mexicana	yellow-e3 part 1	540	6
	yellow-e3 part 2	424	5
	yellow-h	729	7
	mrjp (mrjp9-like)	404	6

Habropoda laboriosa	yellow-e3	424	5
	yellow-h	552	4
	mrjp	416	6
Lasioglossum albipes	yellow-e3	423	5
	yellow-h	558	4
	mrjp (mrjp9-like)	407	6
	mrjp	417	6
Melipona quadrifasciata	yellow-e3	398	5
	yellow-h	542	4
	mrjp (incomplete)	287	4
Megachile rotundata	yellow-e3	398	5
	yellow-h	542	4
	mrjp (mrjp-like)	287	4

Table S17.

Genomic region comprising end of yellow-e3 CDS until start of yellow-h CDS in bp.

Species	Size of genomic region (bp)
Apis florea	72,141
Apis mellifera	80,877
Bombus impatiens	11,842
Bombus terrestris	11,858
Dufourea novaeangliae	15,591
Eufriesea Mexicana	16,633
Habropoda laboriosa	26,298
Lasioglossum albipes	29,769
Melipona quadrifasciata	12,473
Megachile rotundata	6,479

Table S18.

Signaling pathways and nuclear receptor gene models for the ten bee species. Informed are protein size (amino acid residues) and exon number. Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*, Bimp-*Bombus impatiens*, Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

gene	Aflo	Amel	Bimp	Bter	Dnov	Emex	Hlab	Lalb	Mqua	Mrot
InR1	1406	1439	1416	1415	1280	1349	1414	1120	1571	1412
	6	6	6	6	5	7	6	2	7	6
InR2	646	1498	1726	1726	1734	1546	1491	1628	1543	1526
	4	11	11	11	11	11	11	13	12	11
TOR	2451	2442	2438	2450	2438	2438	2456	2438	2307	2475
	5	6	5	5	5	7	5	6	8	5
Egfr	1427	1444	1428	1453	1414	1342	1428	1431	1464	1459
	11	11	11	11	11	11	11	11	12	11
Sima	905	845	1171	1127	1324	912	1106	856	1003	1068
	12	12	13	15	14	13	14	12	13	14
	1315				1176					
	14				14					
Tango	671	640	612	663	611	655	683	615	634	672
	12	10	10	12	10	11	13	10	10	12
Fatiga	477	374	486	487	481	478	476	503	262	492
	4	2	5	5	4	4	4	4	3	4
EcR-A	606	630	583	589	587	583	592	482		586
	5	5	6	5	5	6	5	5		6
EcR-B	545	560	537	538	534		537		605	537
	5	5	5	5	5		5		6	5
USP	427	427	437	427	422	427	427	423	372	427
	6	6	6	6	6	6	6	6	5	6
Ftz-F1	506	716	607	680	645	675	638	651	743	895
	6	6	7	7	7	7	6	8	7	10
Met	907	839	464	972	682	445	864	518	886	868
	13	12	8	13	11	8	12	8	13	13

Table S19.

Data set used for the molecular evolutionary analysis of *fem/tra* copies.

Gene	Species (ID)	this study	GenBank
fem	Aflo02267	x	
	Amell		EU100941
	Acer		EU100936
	Adors		EU100938
csd	Aflo02266	x	
	Amell		EU100892
	Adors		EU100932
	Acer		EU100907
fem1	Bter		XM_003394645
fem	Bter		NM_001280924
fem	Bimp		XM_003493748
fem1	Bimp14017	x	
fem	Dnov10858	x	
fem	Emex02197	x	
fem	Hlab07879	x	
fem	Lalb_10095	x	
fem1	Lalb_10228	x	
fem	Mrot08431	x	
fem	Mqua06253	x	
tra	Nvitr		NM_001134827
		Privman et al. (2013)	
tra	Aech_tra	x	
tra	Acep_traA	x	
tra	Acep_traB	x	
tra	Pbar_traA	x	
tra	Pbar_traB	x	
tra	Cflo_traA	x	
tra	Cflo_traB	x	
tra	Lhum_traA	x	
tra	Lhum_traB	x	
tra	Hsal_traA	x	
tra	Hsal_traB	x	

Table S20.

Data set used for the molecular evolutionary analysis of *tra2* copies.

species	this study	GenBank
Apis florea	Aflo04333	
Apis mellifera		GB47305
Bombus terrestris		XM_003398958.1
Bombus impatiens		XM_003485456.1
Dufourea novaengliae	Dnov04996	
Eufriesea mexicana	Emex10063	
Lasioglossum albipes	Labl_13632	
Megachile rotundata	Mrot01874	
Melipona quadrifasciata	Mqua09093	
Habropoda laboriosa	Hlab07817	
Nasonia vitripennis		XM_001601056
Drosophila melanogaster		FBgn23633
Anastrepha amita		FN658617.1
Musca domestica		AY847518.1
Bombyx mori		NM_001126233

Table S21.

Data set used for the molecular evolutionary analysis of *dsx* copies.

species	this study	GenBank
Apis florea	Aflo03380	
Apis mellifera male		EU236954
Apis mellifera female		EU236957
Bombus terrestris	Bter02495	
Bombus impatiens	Bimp05107	
Dufourea novaeangliae	Dnov10920	
Eufriesea mexicana	Emex00223	
Habropoda laboriosa	Hlab01811	
Lasioglossum albipes	Lalb_01555	
Megachile rotundata	Mrot01361	
Melipona quadrifasciata	Mqua08015	
Ceratitis capitata male		AF434935
Ceratitis capitata female		AF435087
Drosophila melanogaster		NM_169202.1
Nasonia vitripennis		NM_001162517.1
Drosophila virillis		XM_002056562.1
Bombyx mori male		NM_001111345.1
Bombyx mori female		NM_001043406.1

Table S22.

Copy number of *fem/tra* and paralogs in Hymenoptera. Copy numbers are obtained from annotated bee genomes (this study), manual analysis, and orthologous group entries (OrthoDB7). Numbers in brackets indicate questionable copy numbers resulting from potential pseudogenes.

species	fem/tra	fem/tra paralogs
Apis mellifera	1	1 (2)
Apis cerana	1	1
Apis dorsata	1	1
Apis florea	1	1
Bombus terrestris	1	1
Bombus impatiens	1	1
Melipona quadrifasciata	1	-
Eufriesea mexicana	1	(1)
Habropoda laboriosa	1	(1)
Megachile rotundata	1	-
Duforea novaeanglia	1	-
Lasioglossum albipes	1	1 (3)
Atta cephalotes	1	1
Acromyrmex echinatior	1	-
Solenopsis invicta	1	1 (2)
Pogonomyrmex barbatus	1	1
Camponotus floridanus	1	1 (2)
Linepithema humile	1	1
Harpegnathos saltator	1	1 (2)
Nasonia vitripennis	1	-

Table S23.

Groups of fem/tra orthologs and paralogs in Hymenoptera obtained from OrthoDB7.

				AA
Gene	**Protein**	**Organism**	**Code**	**Length**
GB47020	GB47020-PA	Apis mellifera	AMELL	343
GB47021	GB47021-PA	Apis mellifera	AMELL	256
GB47022	GB47022-PA	Apis mellifera	AMELL	599
Aflo02266	Aflo02266	Apis florea	AFLOR	442
Aflo02267	Aflo02267	Apis florea	AFLOR	402
Bimp10092	Bimp10092	Bombus impatiens	BIMPA	549
Bimp14017	Bimp14017	Bombus impatiens	BIMPA	425
Bter09063	Bter09063	Bombus terrestris	BTERR	369
Bter10437	Bter10437	Bombus terrestris	BTERR	355
Lalb_10095	Lalb_10095	Lasioglossum albipes	LALBI	296
Lalb_10228	Lalb_10228	Lasioglossum albipes	LALBI	243
Lalb_10388	Lalb_10388	Lasioglossum albipes	LALBI	189
Lalb_12702	Lalb_12702	Lasioglossum albipes	LALBI	167
Mrot08431	Mrot08431	Megachile rotundata	MROTU	463
Hlab07879	Hlab07879	Habropoda laboriosa	HLABO	606
Emex02197	Emex02197	Eufriesea mexicana	EMEXI	468
Emex02510	Emex02510	Eufriesea mexicana	EMEXI	190
Mqua06253	Mqua06253	Melipona quadrifasciata	MQUAD	430
HSAL15631	HSAL15631-PA	Harpegnathos saltator	HSALT	104
HSAL15632	HSAL15632-PA	Harpegnathos saltator	HSALT	140
HSAL23152	HSAL23152-PA	Harpegnathos saltator	HSALT	188
LH15834	LH15834-PA	Linepithema humile	LHUMI	246
LH15839	LH15839-PA	Linepithema humile	LHUMI	227
CFLO18904	CFLO18904-PA	Camponotus floridanus	CFLOR	188
CFLO18905	CFLO18905-PA	Camponotus floridanus	CFLOR	267
CFLO27011	CFLO27011-PA	Camponotus floridanus	CFLOR	504
PB18766	PB18766-PB	Pogonomyrmex barbatus	PBARB	452
PB18777	PB18777-PA	Pogonomyrmex barbatus	PBARB	77
SINV24204	SINV24204-PA	Solenopsis invicta	SINVI	64
SINV24207	SINV24207-PA	Solenopsis invicta	SINVI	73
SINV24215	SINV24215-PA	Solenopsis invicta	SINVI	192
AECH27269	AECH27269-PA	Acromyrmex echinatior	AECHI	435
ACEP25429	ACEP25429-PA	Atta cephalotes	ACEPH	705
Dnov10858	Dnov10858	Dufourea novaeangliae	DNOVA	417
Nasvi2EG005321	Nasvi2EG005321t1	Nasonia vitripennis	NVITR	405

Table S24.
Fem/tra orthologs, Gene ID and scaffold number.

species	**Fem-orthologs**	**scaffold**
A.mellifera	fem	LG3
	csd	LG3
A.florea	Aflo02267 (fem)	Unplaced 00371
	Aflo02266(csd)	Unplaced 00371
B.terrestris	Bter09063	unplaced723
	Bter10437 (fem1)	LK3.6
B.impatiens	Bimp10092	unplaced1283
	Bimp14017 (fem1)	Scaffold0565
H.laboriosa	Hlab07879	scaffold447
E.mexicana	Emex02197 (fem)	scaffold1601
	Emex02510	scaffold16922
D.novaeangliae	Dnov10858	scaffold8
L.albipes	Lalb10228 (fem1)	scaffold53
	Lalb10095 (fem)	scaffold531
	Lalb12702	scaffold7928
	Lalb10388	scaffold519
M.rotundata	Mrot08431	unplaced0278
M.quadrifasciata	Mqua06253	scaffold203

Table S25.

Pairwise synonymous (ds) and nonsynonymous (dn) divergence per site and its ratio (dn/ds) between fem and fem1 copies of bees.

gene ID	gene ID	ds	dn	dn/ds
Bter fem	Bimp fem	0.036	0.010	0.281
Bter fem1	Bimp14017 fem1	0.062	0.046	0.752
Bter fem	Bimp14017 fem1	0.156	0.134	0.856
Bimp fem	Bimp14017 fem1	0.164	0.134	0.816
Bter fem1	Bter fem	0.171	0.113	0.663
Bter fem1	Bimp fem	0.164	0.113	0.691
Lalb 10095fem	Lalb 10228fem1	0.156	0.094	0.604
Lalb 10095fem	Lalb 10388	0.066	0.086	1.306
Lalb 10095fem	Lalb 12702	0.210	0.155	0.736
Lalb 10388	Lalb 10228fem1	0.157	0.099	0.630
Lalb 10388	Lalb 12702	0.220	0.146	0.665
Lalb 10228fem1	Lalb 12702	0.090	0.088	0.978
Amell fem	Acer fem	0.066	0.021	0.32
Amell fem	Ad fem	0.101	0.028	0.28
Amell fem	Emex02197 fem	0.268	0.171	0.64
Amell fem	Bimp fem	0.328	0.193	0.59
Amell fem	Bter fem	0.335	0.197	0.59
Amell fem	Mqua06253 fem	0.369	0.215	0.58
Amell fem	Hlab07879 fem	0.372	0.269	0.72
Amell fem	Mrot08431 fem	0.390	0.199	0.51
Amell fem	Dnov10858 fem	0.525	0.262	0.50
Amell fem	Lalb 10095 fem	0.707	0.325	0.46

Note: *Bombus* and *Lasioglossum* represent *fem* ortholog and paralog comparison, indicating recent divergence of lineage specific gene copies. As reference, *Apis mellifera fem* comparisons to other bee *fem* orthologs are given.

Table S26.

GO term enrichment among gene families with a significant ($p < 0.05$) family-wide birth/death rate.

GO category	GOID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
MF	GO:0005549	0.000182	4.552371	2.50535	10	101	odorant binding
MF	GO:0004984	0.000828	4.490883	2.009241	8	81	olfactory receptor activity
MF	GO:0017171	0.001231	4.199469	2.133268	8	86	serine hydrolase activity
MF	GO:0004872	0.001894	2.653041	5.804475	14	234	receptor activity
MF	GO:0004252	0.001923	4.38883	1.785992	7	72	serine-type endopeptidase activity
MF	GO:0070011	0.002714	2.645063	5.382782	13	217	peptidase activity, acting on L-amino acid peptides
MF	GO:0004871	0.003485	2.466346	6.201362	14	250	signal transducer activity
MF	GO:0016717	0.00355	39.82119	0.099222	2	4	oxidoreductase activity, acting on paired donors, with oxidation of a pair of donors resulting in the reduction of molecular oxygen to two molecules of water
MF	GO:0016747	0.007496	3.795281	1.736381	6	70	transferase activity, transferring acyl groups other than amino-acyl groups
MF	GO:0008080	0.007623	5.740173	0.793774	4	32	N-acetyltransferase activity
MF	GO:0016491	0.011681	1.963542	9.351654	17	377	oxidoreductase activity
MF	GO:0016712	0.024805	Inf	0.024805	1	1	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen
MF	GO:0008934	0.024805	Inf	0.024805	1	1	inositol monophosphate 1-phosphatase activity
MF	GO:0004995	0.024805	Inf	0.024805	1	1	tachykinin receptor activity
MF	GO:0016149	0.024805	Inf	0.024805	1	1	translation release factor activity, codon specific
MF	GO:0052745	0.024805	Inf	0.024805	1	1	inositol phosphate phosphatase activity

MF	GO:0050660	0.028316	4.9925	0.669747	3	27	flavin adenine dinucleotide binding
MF	GO:0016620	0.039862	7.22938	0.322471	2	13	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor
MF	GO:0004888	0.041812	2.496332	2.548875	6	107	transmembrane signaling receptor activity
MF	GO:0004348	0.049	39.56579	0.049611	1	2	glucosylceramidase activity
MF	GO:0004022	0.049	39.56579	0.049611	1	2	alcohol dehydrogenase (NAD) activity
MF	GO:0004019	0.049	39.56579	0.049611	1	2	adenylosuccinate synthase activity
MF	GO:0004655	0.049	39.56579	0.049611	1	2	porphobilinogen synthase activity
MF	GO:0003997	0.049	39.56579	0.049611	1	2	acyl-CoA oxidase activity
BP	GO:0007600	0.000305	4.322344	2.700675	10	103	sensory perception
BP	GO:0003008	0.000448	4.09621	2.831776	10	108	system process
BP	GO:0007608	0.001119	4.33407	2.123832	8	81	sensory perception of smell
BP	GO:0032501	0.001129	3.350926	3.749481	11	143	multicellular organismal process
BP	GO:0006629	0.012589	2.796992	3.145313	8	122	lipid metabolic process
BP	GO:0006508	0.021161	2.093154	6.266615	12	239	proteolysis
BP	GO:0055114	0.023931	1.942248	7.892264	14	301	oxidation-reduction process
BP	GO:0006022	0.030031	3.098958	1.75675	5	67	aminoglycan metabolic process
BP	GO:0016042	0.037861	7.557576	0.314642	2	12	lipid catabolic process
CC	GO:0016020	0.007422	1.888536	28.70018	39	1163	membrane

Table S27.

Biogenic amine receptors. The numbers refer to the gene ID numbers for each species. The first column gives the names of the honey bee receptors (see ref. (*132*)). When these names have a question mark, they are still orphans (no ligands known). Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*,Bimp-*Bombus impatiens*,Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*.

receptor/ligand	Hlab	Emex	Mqua	Mrot	Dnov	Amel	Aflo	Bter	Bimp	Lalb
Am1/octopamine	04677	07481	12742	03571	08228	52879	NW_003790810.1	on GL898805	05943	05278
Am2/octopamine	04678	07479	12741	03572	08232	52878	08798	09763	05942	03476
Am3/octopamine	04675	10416	12746	03570	09344	49696	07791	09094	04133	05276
Am4/octopamine	04680	07474+07475	10430	09484	08196	43263	01024	03955	05941	12035
Am5/octopamine	07232	06234	08792	02805	02559	52910	10257	08391	10762	11934
Am6/dopamine	00739	08028	01455	07061	04066	50154	10538	08001	06523	10591
Am7/octop./tyr.	03456	01810	09005	05781	04436	47385	13429	00084	05834	04055
Am8/dopamine	12286+84	02041+42	13358+13348+13359	01169	07288+07291	42577	01197	04775	03116	11697+06069
Am9/dopamine	11370	03073	06267	12087	02463	50192	12746+04095	07938	02769	13751
Am10/?	04936+04937	05426+05427	08402+08398	08104	02510+02512	50583	03445+10398	08191	03257	13804+13800
Am11/serotonin	09004	01067+01068	04366	06478	01873	48005	05049	06984	17521	00417
Am12/serotonin	03866+03867	02541+02542	11810+11814	08384+08386	scaf.17+scaf451	45788	00341+00340	09501+09502	18142/10763+08588	12369+12371
Am13/tyramine	00637	07990+07991	12068	04424	00308	53912	09627	00401	08669	08136
Am14/ACh	02409	04403	02616	02603	11443	41397	12921	04590	01960	11461
Am15/ACh	07657	00878	08290	02405	06011	51689	NW_003790299+NW003791902	08571	09104	01483
Am16/serotonin	01635	10689	00656	01194	05297	42606	01302	04727	03116	06947
Am17/?	10098	01801+09990+09993	10742+10743+10745	03216	09020+09023	51374	13858+13856	05159	04190	08937+00102

Am18/?	08642	03989	00839+00842	07790+07792	07263+07264+11105	42323	09537+00022	00446	05541	01429
Am19/?	05214	08670	01892	04110	00383	54467	14174	02920	06436	11841
A2a/adenosine	02187	12569	04032	08547	00860	51506	14673	01419	05424	06898

Table S28.

Neuropeptide and protein hormone genes (indicated by np) and their GPCR genes (indicated by –R, LGR, or DLGR) present in the ten bees with a sequenced genome. The numbers refer to the gene IDs for each species. Highlighted in green means present genes; highlighted in yellow means genes that were absent (nd = not determined). Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*,Bimp-*Bombus impatiens*,Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

np/np-R	Hlab	Emex	Mqua	Mrot	Dnov	Amel	Aflo	Bter	Bimp	Lalb
ACP	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
ACP-R	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
AKH	scaffold801	03193	scaffold 6	08765	02031	50220	NW_003789956.1	11980	13636	02451
AKH-R	10338	01257	03794	08227	06650	53230	01758	04483	25350	10934
Allatotropin	08711	Scaffold 2087	07358	00733	08938	48348	03531	05123	04225	08192
AT-R	09695+09696	06996	13402	11781	11795	44046	00858	09157	10582	10639
Ast-A	10800	02174	04662	05637	09230	47928	03246	03999	05900	09996
Ast-A-R	04320	06675	15049	08051	04546	43574	08948	07794	12768	08599
Ast-B	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
Ast-B-R	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
Ast-C	04226	02525	scaf34	00221	07601	43201	05848 (no stop)	06194	04108	02287
Ast-CC	04225	02524	09240	00220	07600	43120	05849	06193	04109	02286
Ast-C-R	02113	06675	03939	09437	05273	55818	14770	01350	02162	10319
bursicon-a	07932	12246	15255	13064	11928	45446	15571	11366	11829	13906
bursicon-b	07933	12247	15258	13065	11927	45445	15572	11367	11828	13907
LGR2	05689	08404	04614	06178	01374	45694	10803	06683	10035	04224
Capa	09952? scaf65	10142	12803	05622	09324	42539	15002	04174	03987	07713
capa-R1	11638	00504	03363	03993	02766	55630	11751	00966	06698	06083+06084
capa-R2	00100	00505	03364	03992	02765	55629	11752	00964	12769	00913
CCHa-1	04614	08564	10948	05634	scaffold 168	40377	04770	04058	03873	scaffold 207
CCHa1-R	09006	01072	04363	12761	01872	40169	05042	06982	03448	00416
CCHa-2	04613	08565	10950	05635	08281 (part.)	40536	03185	04057	03872	04767
CCHa2-R	13369	11036	15134	12746	12071	40053	15595	10210	03447	00414+00415

np/np-R	Hlab	Emex	Mqua	Mrot	Dnov	Amel	Aflo	Bter	Bimp	Lalb
CCAP	05306	02948	08431	11982	scaffold 139	50604	10438	08339	11665	04990
CCAP-R	03786	05682	07533	01739	11097	54316	02631	08837	10482	05597
corazonin	06294	11614	08814	08830	02158	53951	09679	00330	10356 (not process.)	02717
corazonin-R	01799	12632	08123+25	08374+08373	05476	44824	15667	05562	20585	06198+00772
DH/Calc	04807	00438	08596	10444	10634	45734	10747	02636	07483	08757
DH/Calc-R	07809	10071	07649	01865	05000	47217	04326	10918	01145	13227
DH/CRF	01251	01665	07414	06784	00167	48796	07798	10111	09858	06497
DH/CRF-R	06044	04324	01705	06044	00807	49166	02175	01831	25051	10406
EH	02327	04994	14985	09124	00542	49648	07536	02024	03054	00344
Elevenin	11517	scaffold 1358	04997	02941	03629	XP_003251276.1	11866	09370	20439 (v1.1)	12272
ETH	08806	01138	03107	XP_003704398.1	03310	40094	XP_003692656.1	06941	11121	05452
ETH-Ra	11768	02845+02846	08045	09556	09419	48241	06373	03469	NT_176882	01380
ETH-Rb	scaffold86	scaffold 18034	08042	09557	09418	AADG06000427.1	scaffold 1351	AELG01000878.1	NT_177001	01379
FMRFa	05699	scaffold 5767	05491	07917	02284	41296	10864	08886	05492	04233
FMRFa-R	12855	04568	10036	11550	05682	51916	14442	01242	25879	12654
GPA2	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
GPB5	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
DLGR1	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
ILP-A	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
ILP-B	00122	11949	00161	10255	03731	43560	00880	07647	10206	03027
ILP-C	10806	02180	00075	05643	08237	54524	03252	04004	05894	06717+06718
inotocin	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
inotocin-R	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
ITP	06888	02451	06924	10102	12670	47095	06078	05033	07596	04969

np/np-R	Hlab	Emex	Mqua	Mrot	Dnov	Amel	Aflo	Bter	Bimp	Lalb
kinin	scaffold18	nd	04906	05535	scaffold 196	49225	09338	04019	06912	13028
kin-R	06920	nd	03244	01967	01446	52101	09179	06529	06121	05580
myosuppressin	03972	05746	13651	08469	06832	44942	03898	05627	06424	05079
MS-R1	00601	04205	00392	02323	02962	51670	09472	08678	18814	00446
MS-R2	nd									
natalisin	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
natalisin-R	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
neuroparsin	04825	scaffold 994	14448	04349	scaffold 8	45741	10730	02480	05125	scaffold 685
NPF	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
NPF-R	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
NPF-like ("NPY")	08971	01049	04395	06507	08654	50693	02113	07003	11019	08815
NPLP-1	04414	10888	02802	03558	05475	43772	12655	07371	13064	00796
orcokinin	04276	05407	09192+09193	AFJA01000102.1	08422	43091	05808	05654	NT_176780	scaf. 333 (part.)
PDF	04170	07290	01568	01857	07567	43156	04035	05866	03701	13724
PDF-R	01527	03923	14612	02489	05392	40478	04825	09913	05294	13665
proctolin	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
proc-R	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
PTTH	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
pyrokinin	07456	09688	10669	05170	06921	46057	06752	03068	00484	05899
PK1-R	10790	10608	04674	05464	09238	42135	07504	02054	08599	04370
PK2-R1	08235	02235	01214+01215	05985	00496	40337	03232	09255	05907	09984
PK2-R2	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
RYa	06366	04654	02910	04695	04789	52705	08862	07469	NT_176446	08080
RYaR	05336	05119	12623	04123	05826	43519	02512	05728 (84)	05584	14588

np/np-R	Hlab	Emex	Mqua	Mrot	Dnov	Amel	Aflo	Bter	Bimp	Lalb
sex peptide	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
Sex pep.-R/Ast-B-R	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
SIFa	08807	01138	03106	XP_003249680.1	03309	40093	04910	06940	11122	05452
SIFa-R	10712	03827	06813	06255	06240	47119	10684	08356	00455	08315+09472
sNPF	00975	scaffold 7107	12853	00161	07789	50444	08252	12330	01294	07116
sNPF-R	01500	01982	13267	10770	05273	42678	01427	04447	01442	10995
sulfakinin	02593	10199	10077	04020	10357	48701	12114	nd	nd	scaffold 158
SK-R1	02359	07632	13498	01999	04241	45613	00001	nd	nd	scaf.475+2406
SK-R2	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
tachykinin	02907	05941	04905	05536	04172	49248	09337	04018	06911	13031
TK-R	10251	02828+002829	00577+00580	06278	06228	49971	04199	07798	14914	03103+03102
trissin	09683	06988	scaffold 81	AFJA01011719.1	11897	nd	nd	09167	02443	00142
trissin-R	11231	scaffold 14206	00528	02527	11715	nd	nd	08639	06209	00600

Table S29.

Orphan GPCRs annotated in the ten bees. The numbers refer to the gene IDs for each species. The first column gives the names of the honey bee receptors (see ref.(*132*)). Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*,Bimp-*Bombus impatiens*,Dnov-*Dufourea novaeangliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa*

receptor	Hlab	Emex	Mqua	Mrot	Dnov	Amel	Aflo	Bter	Bimp	Lalb
Am20/ CG13229	00333	03797	nd	10524	10715	46789	011477	02560	NT_176967	08797
Am23	05138	08292	07891	09814	09485	55046	06192	03347	NT_177487	03895
Am24	07403+07405	01962+01963	11878+11879	05125+05127	00057	42369	04592	09645	NT_177000	05888
Am38	06420	scaffold 1614	02846	nd	04753	52747	08582	07518	10293	00782
Am39	06362	05643	02979	04520	11537	54178	07726	11902	00556	00816
Am48/ LGR3	06667	06866	10255	09657	11161	51938	08788	00553	09441	00157
Am49	01189	01284	01662	06081	00412	55751	02129	01779	03128	scaffold 431
Am50	07005	scaffold 3966	14390	10618	03938	50824	03127	00689	NT_176463	02190
Am51	08703	03496	07349	00725	08947	48344	03539	05132	04217	08201
Am52	06003	01207	05423	01435	scaffold 02007	47749	01632	01887	01254	02007
Am54	01626	12365	14648	02466	11678	52820	07238	09833	05212	00633
Am56	09240	10978	01349	01268	05089	40478	01136	04817	08476	12386

Table S30.

Functional (F) and Pseudogenized (P) P450 genes in the ten bee species.

Species	Sociality	Total P450s		Mitochondrial		CYP2		CYP3		CYP4	
		F	P	F	P	F	P	F	P	F	P
Habropoda laboriosa	ancestrally solitary	41	1	6	0	8	2	22	0	5	0
Apis mellifera	obligate complex eusociality	47	5	6	0	8	0	29	5	4	0
Apis florea	obligate complex eusociality	46	3	7	0	8	0	27	2	4	1
Eufriesea mexicana	facultative simple eusociality	50	4	6	0	8	0	31	4	5	0
Melipona quadrifasciata	obligate complex eusociality	57	7	6	0	7	1	38	6	6	0
Bombus impatiens	obligate simple eusociality	44	7	6	0	7	0	27	6	4	1
Bombus terrestris	obligate simple eusociality	44	7	6	0	7	0	27	6	4	1
Megachile rotundata	ancestrally solitary	50	1	6	0	8	0	32	1	4	0
Lasioglossum albipes	facultative simple eusociality	45	3	6	0	8	0	26	3	5	0
Dufourea novaeangliae	ancestrally solitary	58	2	6	0	8	0	39	2	5	0
Nasonia vitripennis		92	9	6	0	8	0	48	8	30	1

Table S31.

All repetitive DNA in ten bee species (fraction of genome in %).

elements type, repeat type, class, order, superfamily	Hlab	Amel	Aflo	Emex	Mqua	Bimp	Bter	Mrot	Lalb	Dnov
all repetitive DNA (intersp./non-intersp./other)	**26.41**	**12.42**	**13.10**	**49.00**	**18.21**	**16.23**	**14.01**	**43.23**	**34.18**	**36.97**
Repetitive & transposable elements (intersp./non-inters	**19.39**	**9.46**	**7.84**	**31.12**	**14.90**	**15.10**	**13.19**	**36.35**	**29.67**	**31.80**
Non-interspersed repeats (sum)	**1.79**	**4.05**	**4.22**	**6.62**	**2.39**	**1.45**	**2.51**	**10.15**	**3.04**	**6.89**
SSR	0.78	0.62	1.56	5.90	1.13	0.49	1.59	7.43	2.72	5.98
Low complexity	0.54	3.42	2.67	0.72	1.02	0.80	0.84	0.79	0.32	0.70
Satellite	0.48	0.02	0.00	0.00	0.25	0.16	0.09	1.94	0.00	0.21
Interspersed repeats (all) (sum)	**17.60**	**5.40**	**3.61**	**24.50**	**12.51**	**13.65**	**10.68**	**26.20**	**26.63**	**24.91**
Class I, Class II (sum)	**4.22**	**0.67**	**0.91**	**4.32**	**1.61**	**5.83**	**4.85**	**9.89**	**8.15**	**7.97**
Derivatives of Class I and II, unclassified/novel (sun	**13.38**	**4.74**	**2.70**	**20.18**	**10.90**	**7.82**	**5.83**	**16.31**	**18.48**	**16.94**
Class I - Retrotransposons (sum)	**2.38**	**0.10**	**0.32**	**3.45**	**0.80**	**4.65**	**2.85**	**2.76**	**2.91**	**4.61**
LTR Retrotransposon (sum)	**1.08**	**0.02**	**0.21**	**1.62**	**0.50**	**2.65**	**1.30**	**1.87**	**1.35**	**2.14**
Copia	0.12	0.02	0.04	1.40	0.05	0.30	0.15	0.45	0.30	0.55
Gypsy	0.52	0.00	0.01	0.15	0.14	1.93	1.00	1.02	0.70	1.27
Bel-Pao	0.37	0.00	0.17	0.06	0.31	0.25	0.13	0.38	0.32	0.32
Retrovirus	0.00	0.00	0.00	0.00	0.00	0.06	0.01	0.00	0.00	0.00
Unclassified LTR Retrotransposon	0.07	0.00	0.00	0.00	0.00	0.12	0.01	0.02	0.03	0.00
DIRS Retrotransposon	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.09**
PLE Penelope Retrotransposon	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.11**	**0.02**
LINE (non-LTR) Retrotransposon (sum)	**1.26**	**0.04**	**0.09**	**1.79**	**0.29**	**1.76**	**1.45**	**0.84**	**1.42**	**2.30**
R2 (NeSL, R2, R4, CRE)	0.14	0.03	0.09	0.09	0.02	0.10	0.03	0.05	0.03	0.06
RTE	0.04	0.00	0.00	0.02	0.00	0.27	0.17	0.00	0.17	0.03
Jockey (Rex,Jockey,Cr1,Kiri,L2,crack,Da	0.42	0.00	0.00	0.67	0.19	1.04	1.06	0.18	0.50	0.50
L1 (L1, Tx1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I (R1, I, Nimb, outcast, Tad, Loa)	0.65	0.00	0.00	1.00	0.08	0.32	0.19	0.60	0.69	1.72
Unclassified LINE	0.01	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.02	0.00
SINE (sum)	**0.01**	**0.02**	**0.02**	**0.00**	**0.01**	**0.01**	**0.05**	**0.01**	**0.00**	**0.02**
5S-SINE	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
Unclassified SINE	0.01	0.02	0.02	0.00	0.01	0.01	0.05	0.01	0.00	0.02
Unclassified Retrotransposon	**0.00**	**0.00**	**0.00**	**0.04**	**0.00**	**0.23**	**0.05**	**0.02**	**0.03**	**0.04**
Class I Derivatives (sum)	**9.76**	**4.09**	**2.70**	**6.77**	**9.54**	**7.54**	**5.62**	**9.76**	**13.95**	**7.58**
LARD	7.30	3.10	2.36	3.62	9.35	6.87	4.85	7.75	11.85	7.28
TRIM	2.46	0.99	0.35	3.15	0.19	0.67	0.78	2.01	2.10	0.30
Class II - DNA Transposons (sum)	**1.84**	**0.57**	**0.59**	**0.87**	**0.81**	**1.18**	**2.00**	**7.13**	**5.24**	**3.36**
TIR (sum)	**1.79**	**0.57**	**0.26**	**0.76**	**0.81**	**1.15**	**1.48**	**2.77**	**5.07**	**3.35**
Tc1/Mariner	1.05	0.49	0.19	0.43	0.62	0.36	0.43	1.59	2.82	1.29
hAT	0.11	0.00	0.00	0.00	0.00	0.02	0.00	0.15	0.71	0.38
Mutator/Rehavkus	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.00
Merlin	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.01
Transib	0.06	0.00	0.00	0.01	0.00	0.00	0.00	0.05	0.18	0.00
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.05
PiggyBac	0.19	0.04	0.05	0.22	0.13	0.69	0.83	0.12	0.03	0.09
PIF-Harbinger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.04
CACTA (En/Spm)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.03
Academ	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
Ginger	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00
Kolobok	0.07	0.00	0.00	0.00	0.00	0.01	0.00	0.03	0.00	0.20
Sola	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.06	0.07
Chapaev	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
ISL2EU	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Unclassified TIR DNA Transposon	0.30	0.04	0.01	0.10	0.05	0.07	0.20	0.68	1.16	1.19
Helitron	**0.05**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.02**	**0.03**	**0.00**
Polinton/Maverick	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.05**	**0.00**
Unclassified DNA-Transposon	**0.00**	**0.00**	**0.33**	**0.11**	**0.00**	**0.03**	**0.52**	**4.34**	**0.10**	**0.00**
Class II Derivatives: MITE	**0.41**	**0.00**	**0.00**	**0.29**	**0.01**	**0.28**	**0.20**	**0.50**	**1.34**	**0.88**
Unclassified/novel/putative element (sum)	**3.20**	**0.65**	**0.00**	**13.12**	**1.34**	**2.80**	**1.58**	**6.05**	**3.19**	**8.48**
Other repetitive DNA (sum)	**7.02**	**2.96**	**5.26**	**17.88**	**3.30**	**1.13**	**0.82**	**6.88**	**4.52**	**5.17**
Not categorized	2.59	0.53	0.00	17.27	0.06	0.35	0.30	1.33	1.44	1.37
Potential Hostgene	4.43	2.44	5.26	0.61	3.25	0.78	0.52	5.53	3.07	3.68
Wolbachia	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.12
genomesize (bp, as used for REPET)	294.12	234.09	230.47	557.42	256.55	249.19	245.20	272.66	337.96	290.96

Table S32.

All repetitive DNA in ten bee species (total length). Aflo-*Apis florea*, Amel-*Apis mellifera*, Bter-*Bombus terrestris*,Bimp-*Bombus impatiens*,Dnov-*Dufourea novaengliae*, Emex-*Eufriesea mexicana*, Lalb-*Lasioglossum albipes*, Mrot-*Megachile rotundata*, Mqua-*Melipona quadrifasciata*, Hlab-*Habropoda laboriosa.*

elements type, repeat type, class, order, superfamily	**Hlab**	**Amel**	**Aflo**	**Emex**	**Mqua**	**Bimp**	**Bter**	**Mrot**	**Lalb**	**Dnov**
all repetitive DNA (intersp./non-intersp./other)	**77684912**	**29068292**	**30190085**	**273129914**	**46705349**	**47416859**	**38237370**	**117866611**	**115516976**	**107564526**
Repetitive & transposable elements (intersp./non-intersp.) (sum)	**57035257**	**22134229**	**18058147**	**173470912**	**38226793**	**44598254**	**36228680**	**99114283**	**100257143**	**92517085**
Non-interspersed repeats (sum)	**5272523**	**9486745**	**9732547**	**36897204**	**6142062**	**3607445**	**6157797**	**27681003**	**10262039**	**20051241**
SSR	2291001	1441651	3585001	32902049	2893194	1221891	3886911	20256535	9185906	17386320
Low complexity	1584390	8001104	6147546	3995155	2613417	1983832	2048410	2147860	1072939	2050788
Satellite	1397132	43990	0	0	635451	401722	222476	5276608	3194	614133
Interspersed repeats (all) (sum)	**51762734**	**12647484**	**8325600**	**136573708**	**32084731**	**40990809**	**30070883**	**71433280**	**89995104**	**72465844**
Class I, Class II (sum)	**12417350**	**1558968**	**2100558**	**24089796**	**4117772**	**14528804**	**11897247**	**26960158**	**27554440**	**23176668**
Derivatives of Class I and II, unclassified/novel (sum)	**39345384**	**11088516**	**6225042**	**112483912**	**27966959**	**26462005**	**18173636**	**44473122**	**62440664**	**49289176**
Class I - Retrotransposons (sum)	**6998335**	**223588**	**745307**	**19219318**	**2052237**	**11595235**	**6988747**	**7512622**	**9832797**	**13406009**
LTR Retrotransposon (sum)	**3169516**	**49549**	**494607**	**9019559**	**1287914**	**6614301**	**3182801**	**5089695**	**4557561**	**6220102**
Copia	362567	43892	82310	7804423	125148	744572	377616	1229521	1005441	1591485
Gypsy	1527156	0	18317	852961	370326	4800968	2444045	2780246	2357986	3694684
Bel-Pao	1083239	0	393980	362175	791129	624191	330778	1024292	1092103	929207
Retrovirus	0	0	0	0	0	154261	17964	0	0	0
Unclassified LTR Retrotransposon	196554	5657	0	0	1311	290309	12398	55636	102031	4726
DIRS Retrotransposon	**0**	**12472**	**0**	**0**	**0**	**0**	**0**	**0**	**16219**	**251602**
PLE Penelope Retrotransposon	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**379612**	**69400**
LINE (non-LTR) Retrotransposon (sum)	**3714743**	**83103**	**215540**	**9961032**	**745208**	**4378358**	**3557856**	**2282770**	**4790664**	**6688957**
R2 (NeSL, R2, R4, CRE)	397472	72107	212201	511443	54793	252959	74479	126840	112953	167751
RTE	128271	0	0	127470	0	672481	406544	6123	578662	87166
Jockey (Rex,Jockey,Cr1,Kiri,L2,crack,Daphne)	1249609	0	0	3760479	479529	2589707	2596829	486450	1702091	1441488
L1 (L1, Tx1)	0	0	0	0	0	0	0	0	0	0
I (R1, I, Nimb, outcast, Tad, Loa)	1919996	10996	0	5561640	210886	804031	474349	1635717	2321104	4992552
Unclassified LINE	19395	0	3339	0	0	59180	5655	27640	75854	0
SINE (sum)	**43960**	**47278**	**35160**	**26200**	**19115**	**23929**	**130569**	**29815**	**0**	**58141**
5S-SINE	67853	22660	0	7383	0	0	0	51308	0	0
Unclassified SINE	43960	47278	35160	26200	19115	23929	130569	29815	0	58141
Unclassified Retrotransposon	**2263**	**8526**	**0**	**205144**	**0**	**578647**	**117521**	**59034**	**88741**	**117807**
Class I Derivatives (sum)	**28715918**	**9566616**	**6225042**	**37749713**	**24481541**	**18796216**	**13787390**	**26621719**	**47132064**	**22054762**
LARD	21482238	7256932	5427933	20206065	23998741	17118557	11886397	21142421	40038343	21195132
TRIM	7233680	2309684	797109	17543648	482800	1677659	1900993	5479298	7093721	859630
Class II - DNA Transposons (sum)	**5419015**	**1335380**	**1355251**	**4870478**	**2065535**	**2933569**	**4908500**	**19447536**	**17721643**	**9770659**
TIR (sum)	**5272586**	**1335380**	**589980**	**4222911**	**2065535**	**2855991**	**3626522**	**7566105**	**17130658**	**9760942**
Tc1/Mariner	3084357	1147521	429808	2377923	1578340	887660	1044373	4337353	9515240	3760356
hAT	327473	0	0	24723	4100	39932	0	396318	2392928	1107440
Mutator/Rehavkus	0	0	0	0	0	0	0	28013	90144	0
Merlin	0	0	0	0	0	0	0	7174	67709	14734
Transib	172584	0	0	40249	0	0	0	139065	617915	8160
P	0	0	0	0	0	0	0	0	71703	153493
PiggyBac	554070	87963	125924	1236239	341051	1719974	2036512	328460	99193	265315
PIF-Harbinger	0	0	0	0	0	0	0	0	64868	119129
CACTA (En/Spm)	0	0	0	0	0	0	0	18663	54812	98696
Academ	0	0	0	0	0	0	55467	0	0	0
Ginger	0	0	0	0	6720	0	0	68702	0	0
Kolobok	192531	0	0	0	0	20644	4969	92521	12311	584958
Sola	47830	0	0	4943	0	2453	0	238329	210051	191012
Chapaev	0	0	0	0	0	0	0	43790	0	0
ISL2EU	0	0	0	0	0	0	0	23760	0	0
Unclassified TIR DNA Transposon	893741	99896	34248	538834	135324	185328	485201	1843957	3933784	3457649
Helitron	**146429**	**0**	**0**	**20938**	**0**	**0**	**0**	**53023**	**89048**	**3718**
Polinton/Maverick	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**166716**	**5999**
Unclassified DNA-Transposon	**0**	**0**	**765271**	**626629**	**0**	**77578**	**1281978**	**11828408**	**335221**	**0**
Class II Derivatives: MITE	**1213883**	**3751**	**0**	**1622234**	**37623**	**691797**	**499865**	**1351492**	**4543216**	**2569950**
Unclassified/novel/putative element (sum)	**9415583**	**1518149**	**0**	**73111965**	**3447795**	**6973992**	**3886381**	**16499911**	**10765384**	**24664464**
Other repetitive DNA (sum)	**20649655**	**6934063**	**12131938**	**99659002**	**8478556**	**2818605**	**2008690**	**18752328**	**15259833**	**15047441**
Not categorized	7620204	1233884	0	96274260	148046	874291	737560	3633614	4867909	3974052
Potential Hostgene	13029451	5700179	12131938	3384742	8330510	1944314	1271130	15080643	10391924	10720172
Wolbachia	0	0	0	0	0	0	0	38071	0	353217
genomesize (bp, as used for REPET)	294117654	234087000	230467781	557421532	256546924	249185056	245204244	272660569	337961498	290964114

Additional Data table S1 (separate file)

Supplementary tables for analysis of TF motifs

Additional Data table S2 (separate file)

Orthogroup IDs and gene IDs of single-copy orthologs in ten bees

Additional Data table S3 (separate file)

Results of Coevol analysis of dN/dS and signatures of selection in ten bees

Additional Data table S4 (separate file)

Single-copy orthologs included in each of the four clade-specific PAML analyses

Additional Data table S5 (separate file)

Genes significantly consistent with social complexity for each of the four clade-specific analyses

Additional Data table S6 (separate file)

Enriched GO terms for genes evolving rapidly and slowly in association with social complexity from the clade-specific PAML analyses

Additional Data table S7 (separate file)

Manual annotation of major royal jelly protein genes and other developmental genes

Additional Data table S8 (separate file)

Pairwise comparisons of substitutions (per site,+/- SD) in deduced *double sex* (*dsx*) amino acid sequence, calculated within the DNA-binding (OD1) and *dsx*-dimerization domain among insects

Additional Data table S9 (separate file)

Pairwise comparisons of substitutions (per site,+/- SD) in deduced *transformer 2 (tra2)* amino acid sequence, calculated within the RNA-binding (RMM) domain and the remaining part of the protein region among insects

Additional Data table S10 (separate file)

Homology assignments for 190 immune-gene candidate proteins among the sequenced bee genomes. ORTHODB version 7 names given when available, along with official gene ID or, where appropriate, manually linked gene ID's (NCBI). Red entries indicate equally likely paralogs in protein set

Additional Data table S11 (separate file)

dN/dS of immunity-related genes across 10 bees

Additional Data table S12 (separate file)

Annotated P450 genes in the ten bee genomes

References and Notes

1. J. Maynard Smith, E. Szathmáry, *The major transitions in evolution*. (Oxford University Press, Oxford, England, 1995).

2. C. D. Michener, *The social behavior of the bees*. (Harvard University Press, Cambridge, MA, 1974).

3. A.-M. Klein, B. E. Vaissière, J. H. Cane, I. Steffan-Dewenter, S. A. Cunningham, C. Kremen, T. Tscharntke, Importance of pollinators in changing landscapes for world crops. *Proc. Biol. Sci. B* **274**, 303–313 (2007). Medline doi:10.1098/rspb.2006.3721

4. H. Hölldobler, E. O. Wilson, *The superorganism: the beauty, elegance and strangeness of insect societies*. (Norton Press, New York, NY, 2009).

5. B. R. Johnson, T. A. Linksvayer, Deconstructing the superorganism: Social physiology, groundplans, and sociogenomics. *Q. Rev. Biol.* **85**, 57–79 (2010). Medline doi:10.1086/650290

6. L. J. Zhu, R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. A. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe, S. Sinha, S. A. Wolfe, M. H. Brodsky, FlyFactorSurvey: A database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.* **39** (Database), D111–D117 (2011). Medline doi:10.1093/nar/gkq858

7. Materials and methods are available as supplementary material online.

8. H. Yan, R. Bonasio, D. F. Simola, J. Liebig, S. L. Berger, D. Reinberg, DNA methylation in social insects: How epigenetics can control behavior and longevity. *Annu. Rev. Entomol.* **60**, 435–452 (2015). Medline doi:10.1146/annurev-ento-010814-020803

9. N. Lartillot, R. Poujol, A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.* **28**, 729–744 (2011). Medline doi:10.1093/molbev/msq244

10. S. H. Woodard, B. J. Fischman, A. Venkat, M. E. Hudson, K. Varala, S. A. Cameron, A. G. Clark, G. E. Robinson, Genes involved in convergent evolution of eusociality in bees. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7472–7477 (2011). Medline doi:10.1073/pnas.1103457108

11. B. A. Harpur, C. F. Kent, D. Molodtsova, J. M. Lebon, A. S. Alqarni, A. A. Owayss, A. Zayed, Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2614–2619 (2014). Medline doi:10.1073/pnas.1315506111

12. J. Roux, E. Privman, S. Moretti, J. T. Daub, M. Robinson-Rechavi, L. Keller, Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.* **31**, 1661–1685 (2014). Medline doi:10.1093/molbev/msu141

13. D. F. Simola, L. Wissler, G. Donahue, R. M. Waterhouse, M. Helmkampf, J. Roux, S. Nygaard, K. M. Glastad, D. E. Hagen, L. Viljakainen, J. T. Reese, B. G. Hunt, D. Graur, E. Elhaik, E. V. Kriventseva, J. Wen, B. J. Parker, E. Cash, E. Privman, C. P. Childers, M. C. Muñoz-Torres, J. J. Boomsma, E. Bornberg-Bauer, C. R. Currie, C. G. Elsik, G.

Suen, M. A. Goodisman, L. Keller, J. Liebig, A. Rawls, D. Reinberg, C. D. Smith, C. R. Smith, N. Tsutsui, Y. Wurm, E. M. Zdobnov, S. L. Berger, J. Gadau, Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res.* **23**, 1235–1247 (2013). Medline doi:10.1101/gr.155408.113

14. J. Romiguier, J. Lourenco, P. Gayral, N. Faivre, L. A. Weinert, S. Ravel, M. Ballenghien, V. Cahais, A. Bernard, E. Loire, L. Keller, N. Galtier, Population genomics of eusocial insects: The costs of a vertebrate-like effective population size. *J. Evol. Biol.* **27**, 593–603 (2014). doi:10.1111/jeb.12331

15. M. Park, K. Shen, WNTs in synapse formation and neuronal circuitry. *EMBO J.* **31**, 2697–2704 (2012). Medline doi:10.1038/emboj.2012.145

16. J. Gates, G. Lam, J. A. Ortiz, R. Losson, C. S. Thummel, *rigor mortis* encodes a novel nuclear receptor interacting protein required for ecdysone signaling during *Drosophila* larval development. *Development* **131**, 25–36 (2004). Medline doi:10.1242/dev.00920

17. D. Brawand, M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grützner, S. Bergmann, R. Nielsen, S. Pääbo, H. Kaessmann, The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011). Medline doi:10.1038/nature10532

18. D. Molodtsova, B. A. Harpur, C. F. Kent, K. Seevananthan, A. Zayed, Pleiotropy constrains the evolution of protein but not regulatory sequences in a transcription regulatory network influencing complex social behaviors. *Front. Genet.* **5**, 431 (2014). Medline doi:10.3389/fgene.2014.00431

19. D. W. Pfaff, A. P. Arnold, A. M. Etgen, R. T. Rubin, S. E. Fahrbach, Eds., *Hormones, Brain and Behavior* (Elsevier, New York, 2009).

20. S. Chandrasekaran, S. A. Ament, J. A. Eddy, S. L. Rodriguez-Zas, B. R. Schatz, N. D. Price, G. E. Robinson, Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18020–18025 (2011). doi:10.1073/pnas.1114093108

21. L. Wilfert, J. Gadau, P. Schmid-Hempel, Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* **98**, 189–197 (2007). Medline doi:10.1038/sj.hdy.6800950

22. E. S. Dolgin, B. Charlesworth, The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics* **178**, 2169–2177 (2008). Medline doi:10.1534/genetics.107.082743

23. S. Schaack, C. Gilbert, C. Feschotte, Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.* **25**, 537–546 (2010). Medline doi:10.1016/j.tree.2010.06.001

24. B. Feldmeyer, D. Elsner, S. Foitzik, Gene expression patterns associated with caste and reproductive status in ants: Worker-specific genes are more derived than queen-specific ones. *Mol. Ecol.* **23**, 151–161 (2014). Medline doi:10.1111/mec.12490

25. P. G. Ferreira, S. Patalano, R. Chauhan, R. Ffrench-Constant, T. Gabaldón, R. Guigó, S. Sumner, Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. *Genome Biol.* **14**, R20 (2013). Medline doi:10.1186/gb-2013-14-2-r20

26. B. G. Hunt, L. Ometto, Y. Wurm, D. Shoemaker, S. V. Yi, L. Keller, M. A. D. Goodisman, Relaxed selection is a precursor to the evolution of phenotypic plasticity. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15936–15941 (2011). Medline doi:10.1073/pnas.1104825108

27. S. J. Gould, *Wonderful Life: The Burgess Shale and the Nature of History* (Norton, New York, 1989).

28. S. Cardinal, B. N. Danforth, Bees diversified in the age of eudicots. *Proc. Biol. Sci.* **280**, 20122686 (2013). Medline doi:10.1098/rspb.2012.2686

29. S. Cardinal, B. N. Danforth, The antiquity and evolutionary history of social behavior in bees. *PLOS ONE* **6**, e21086 (2011). Medline doi:10.1371/journal.pone.0021086

30. E. W. Hare, J. S. Johnston, in *Molecular methods for evolutionary genetics,* V. Orgogozo, M. Rockman, Eds. (Humana, New York, 2011).

31. D. W. Galbraith, K. R. Harkins, J. M. Maddox, N. M. Ayres, D. P. Sharma, E. Firoozabady, Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* **220**, 1049–1051 (1983). Medline doi:10.1126/science.220.4601.1049

32. E. Birney, M. Clamp, R. Durbin, GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004). Medline doi:10.1101/gr.1865504

33. M. Stanke, O. Schöffmann, B. Morgenstern, S. Waack, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006). Medline doi:10.1186/1471-2105-7-62

34. W. H. Majoros, M. Pertea, S. L. Salzberg, TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004). Medline doi:10.1093/bioinformatics/bth315

35. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004). Medline doi:10.1186/1471-2105-5-59

36. C. G. Elsik, A. J. Mackey, J. T. Reese, N. V. Milshina, D. S. Roos, G. M. Weinstock, Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007). Medline doi:10.1186/gb-2007-8-1-r13

37. A. Bairoch, B. Boeckmann, S. Ferro, E. Gasteiger, Swiss-Prot: Juggling between evolution and stability. *Brief. Bioinform.* **5**, 39–55 (2004). Medline doi:10.1093/bib/5.1.39

38. E. M. Zdobnov, R. Apweiler, InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001). Medline doi:10.1093/bioinformatics/17.9.847

39. D. R. Kelley, M. C. Schatz, S. L. Salzberg, Quake: Quality-aware detection and correction of sequencing errors. *Genome Biol.* **11**, R116 (2010). Medline doi:10.1186/gb-2010-11-11-r116

40. R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam, J. Wang, SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012). Medline doi:10.1186/2047-217X-1-18

41. C. Holt, M. Yandell, MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011). Medline doi:10.1186/1471-2105-12-491

42. B. L. Cantarel, I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado, M. Yandell, MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008). Medline doi:10.1101/gr.6743907

43. A. F. A. Smit, R. Hubley, *RepeatModeler* (2011).

44. J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005). Medline doi:10.1159/000084979

45. W. P. Kemp, J. Bosch, B. Dennis, B. J., B. Dennis, Oxygen consumption during the life cycles of the prepupae-wintering bee *Megachile rotundata* and the adult-wintering bee *Osmia lignaria* (Hymenoptera: Megachilidae). *Ann. Entomol. Soc. Am.* **97**, 161–170 (2004). doi:10.1603/0013-8746(2004)097[0161:OCDTLC]2.0.CO;2

46. G. M. Weinstock, G. E. Robinson, R. A. Gibbs, G. M. Weinstock, G. M. Weinstock, G. E. Robinson, K. C. Worley, J. D. Evans, R. Maleszka, H. M. Robertson, D. B. Weaver, M. Beye, P. Bork, C. G. Elsik, J. D. Evans, K. Hartfelder, G. J. Hunt, H. M. Robertson, G. E. Robinson, R. Maleszka, G. M. Weinstock, K. C. Worley, E. M. Zdobnov, K. Hartfelder, G. V. Amdam, M. M. G. Bitondi, A. M. Collins, A. S. Cristino, J. D. Evans, H. Michael, G. Lattorff, C. H. Lobo, R. F. A. Moritz, F. M. F. Nunes, R. E. Page, Z. L. P. Simões, D. Wheeler, P. Carninci, S. Fukuda, Y. Hayashizaki, C. Kai, J. Kawai, N. Sakazume, D. Sasaki, M. Tagami, R. Maleszka, G. V. Amdam, S. Albert, G. Baggerman, K. T. Beggs, G. Bloch, G. Cazzamali, M. Cohen, M. D. Drapeau, D. Eisenhardt, C. Emore, M. A. Ewing, S. E. Fahrbach, S. Forêt, C. J. P. Grimmelikhuijzen, F. Hauser, A. B. Hummon, G. J. Hunt, J. Huybrechts, A. K. Jones, T. Kadowaki, N. Kaplan, R. Kucharski, G. Leboulle, M. Linial, J. T. Littleton, A. R. Mercer, R. E. Page, H. M. Robertson, G. E. Robinson, T. A. Richmond, S. L. RodriguezZas, E. B. Rubin, D. B. Sattelle, D. Schlipalius, L. Schoofs, Y. Shemesh, J. V. Sweedler, R. Velarde, P. Verleyen, E. Vierstraete, M. R. Williamson, M. Beye, S. A. Ament, S. J. Brown, M. Corona, P. K. Dearden, W. A. Dunn, M. M. Elekonich, C. G. Elsik, S. Forêt, T. Fujiyuki, I. Gattermeier, T. Gempe, M. Hasselmann, T. Kadowaki, E. Kage, A. Kamikouchi, T. Kubo, R. Kucharski, T. Kunieda, M. Lorenzen, R. Maleszka, N. V. Milshina, M. Morioka, K. Ohashi, R. Overbeek, R. E. Page, H. M. Robertson, G. E. Robinson, C. A. Ross, M. Schioett, T. Shippy, H. Takeuchi, A. L. Toth, J. H. Willis, M. J. Wilson, H. M. Robertson, E. M. Zdobnov, P. Bork, C. G. Elsik, K. H. J. Gordon, I. Letunic, K. Hackett, J. Peterson, A. Felsenfeld, M. Guyer, M. Solignac, R. Agarwala, J. M. Cornuet, C. G. Elsik, C. Emore, G. J. Hunt, M. Monnerot, F. Mougel, J. T. Reese, D. Schlipalius, D. Vautrin, D. B. Weaver, J. J. Gillespie, J. J. Cannone, R. R. Gutell, J. S. Johnston, C. G.

Elsik, G. Cazzamali, M. B. Eisen, C. J. P. Grimmelikhuijzen, F. Hauser, A. B. Hummon, V. N. Iyer, V. Iyer, P. Kosarev, A. J. Mackey, R. Maleszka, J. T. Reese, T. A. Richmond, H. M. Robertson, V. Solovyev, A. Souvorov, J. V. Sweedler, G. M. Weinstock, M. R. Williamson, E. M. Zdobnov, J. D. Evans, K. A. Aronstein, K. Bilikova, Y. P. Chen, A. G. Clark, L. I. Decanini, W. M. Gelbart, C. Hetru, D. Hultmark, J.-L. Imler, H. Jiang, M. Kanost, K. Kimura, B. P. Lazzaro, D. L. Lopez, J. Simuth, G. J. Thompson, Z. Zou, P. De Jong, E. Sodergren, M. Csűrös, A. Milosavljevic, J. S. Johnston, K. Osoegawa, S. Richards, C.-L. Shu, G. M. Weinstock, C. G. Elsik, L. Duret, E. Elhaik, D. Graur, J. T. Reese, H. M. Robertson, H. M. Robertson, C. G. Elsik, R. Maleszka, D. B. Weaver, G. V. Amdam, J. M. Anzola, K. S. Campbell, K. L. Childs, D. Collinge, M. A. Crosby, C. M. Dickens, C. G. Elsik, K. H. J. Gordon, L. S. Gramates, C. M. Grozinger, P. L. Jones, M. Jorda, X. Ling, B. B. Matthews, J. Miller, N. V. Milshina, C. Mizzen, M. A. Peinado, J. T. Reese, J. G. Reid, H. M. Robertson, G. E. Robinson, S. M. Russo, A. J. Schroeder, S. E. St Pierre, Y. Wang, P. Zhou, H. M. Robertson, R. Agarwala, C. G. Elsik, N. V. Milshina, J. T. Reese, D. B. Weaver, K. C. Worley, K. L. Childs, C. M. Dickens, C. G. Elsik, W. M. Gelbart, H. Jiang, P. Kitts, N. V. Milshina, J. T. Reese, B. Ruef, S. M. Russo, A. Venkatraman, G. M. Weinstock, L. Zhang, P. Zhou, J. S. Johnston, G. Aquino-Perez, J. M. Cornuet, M. Monnerot, M. Solignac, D. Vautrin, C. W. Whitfield, S. K. Behura, S. H. Berlocher, A. G. Clark, R. A. Gibbs, J. S. Johnston, W. S. Sheppard, D. R. Smith, A. V. Suarez, N. D. Tsutsui, D. B. Weaver, X. Wei, D. Wheeler, G. M. Weinstock, K. C. Worley, P. Havlak, B. Li, Y. Liu, E. Sodergren, L. Zhang, M. Beye, M. Hasselmann, A. Jolivet, S. Lee, L. V. Nazareth, L.-L. Pu, R. Thorn, G. M. Weinstock, V. Stolc, G. E. Robinson, R. Maleszka, T. Newman, M. Samanta, W. A. Tongprasit, K. A. Aronstein, C. Claudianos, M. R. Berenbaum, S. Biswas, D. C. de Graaf, R. Feyereisen, R. M. Johnson, J. G. Oakeshott, H. Ranson, M. A. Schuler, D. Muzny, R. A. Gibbs, G. M. Weinstock, J. Chacko, C. Davis, H. Dinh, R. Gill, J. Hernandez, S. Hines, J. Hume, L. R. Jackson, C. Kovar, L. Lewis, G. Miner, M. Morgan, L. V. Nazareth, N. Nguyen, G. Okwuonu, H. Paul, S. Richards, J. Santibanez, G. Savery, E. Sodergren, A. Svatek, D. Villasana, R. Wright; Honeybee Genome Sequencing Consortium, Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**, 931–949 (2006). Medline doi:10.1038/nature05260

47. M. C. Munoz-Torres, J. T. Reese, C. P. Childers, A. K. Bennett, J. P. Sundaram, K. L. Childs, J. M. Anzola, N. Milshina, C. G. Elsik, Hymenoptera Genome Database: Integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res.* **39** (Database), D658–D662 (2011). Medline doi:10.1093/nar/gkq1145

48. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011). Medline doi:10.1038/nbt.1883

49. V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, M. Borodovsky, Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990 (2008). Medline doi:10.1101/gr.081612.108

50. E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, R. Lopez, InterProScan: Protein domains identifier. *Nucleic Acids Res.* **33** (Web Server), W116-20 (2005). Medline doi:10.1093/nar/gki442

51. M. S. Campbell, M. Law, C. Holt, J. C. Stein, G. D. Moghe, D. E. Hufnagel, J. Lei, R. Achawanantakun, D. Jiao, C. J. Lawrence, D. Ware, S. H. Shiu, K. L. Childs, Y. Sun, N. Jiang, M. Yandell, MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014). Medline doi:10.1104/pp.113.230144

52. H. Li, A. Coghlan, J. Ruan, L. J. Coin, J. K. Hériché, L. Osmotherly, R. Li, T. Liu, Z. Zhang, L. Bolund, G. K. Wong, W. Zheng, P. Dehal, J. Wang, R. Durbin, TreeFam: A curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580 (2006). Medline doi:10.1093/nar/gkj118

53. B. M. Sadd, S. M. Barribeau, G. Bloch, D. C. de Graaf, P. Dearden, C. G. Elsik, J. Gadau, C. J. Grimmelikhuijzen, M. Hasselmann, J. D. Lozier, H. M. Robertson, G. Smagghe, E. Stolle, M. Van Vaerenbergh, R. M. Waterhouse, E. Bornberg-Bauer, S. Klasberg, A. K. Bennett, F. Câmara, R. Guigó, K. Hoff, M. Mariotti, M. Munoz-Torres, T. Murphy, D. Santesmasses, G. V. Amdam, M. Beckers, M. Beye, M. Biewer, M. M. Bitondi, M. L. Blaxter, A. F. Bourke, M. J. Brown, S. D. Buechel, R. Cameron, K. Cappelle, J. C. Carolan, O. Christiaens, K. L. Ciborowski, D. F. Clarke, T. J. Colgan, D. H. Collins, A. G. Cridge, T. Dalmay, S. Dreier, L. du Plessis, E. Duncan, S. Erler, J. Evans, T. Falcon, K. Flores, F. C. Freitas, T. Fuchikawa, T. Gempe, K. Hartfelder, F. Hauser, S. Helbing, F. C. Humann, F. Irvine, L. S. Jermiin, C. E. Johnson, R. M. Johnson, A. K. Jones, T. Kadowaki, J. H. Kidner, V. Koch, A. Köhler, F. B. Kraus, H. M. Lattorff, M. Leask, G. A. Lockett, E. B. Mallon, D. S. Antonio, M. Marxer, I. Meeus, R. F. Moritz, A. Nair, K. Näpflin, I. Nissen, J. Niu, F. M. Nunes, J. G. Oakeshott, A. Osborne, M. Otte, D. G. Pinheiro, N. Rossié, O. Rueppell, C. G. Santos, R. Schmid-Hempel, B. D. Schmitt, C. Schulte, Z. L. Simões, M. P. Soares, L. Swevers, E. C. Winnebeck, F. Wolschin, N. Yu, E. M. Zdobnov, P. K. Aqrawi, K. P. Blankenburg, M. Coyle, L. Francisco, A. G. Hernandez, M. Holder, M. E. Hudson, L. Jackson, J. Jayaseelan, V. Joshi, C. Kovar, S. L. Lee, R. Mata, T. Mathew, I. F. Newsham, R. Ngo, G. Okwuonu, C. Pham, L. L. Pu, N. Saada, J. Santibanez, D. Simmons, R. Thornton, A. Venkat, K. K. Walden, Y. Q. Wu, G. Debyser, B. Devreese, C. Asher, J. Blommaert, A. D. Chipman, L. Chittka, B. Fouks, J. Liu, M. P. O'Neill, S. Sumner, D. Puiu, J. Qu, S. L. Salzberg, S. E. Scherer, D. M. Muzny, S. Richards, G. E. Robinson, R. A. Gibbs, P. Schmid-Hempel, K. C. Worley, The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol.* **16**, 76 (2015). Medline

54. R. M. Waterhouse, F. Tegenfeldt, J. Li, E. M. Zdobnov, E. V. Kriventseva, OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* **41**, D358–D365 (2013). Medline doi:10.1093/nar/gks1116

55. R. M. Waterhouse, E. M. Zdobnov, F. Tegenfeldt, J. Li, E. V. Kriventseva, OrthoDB: The hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res.* **39** (Database), D283–D288 (2011). Medline doi:10.1093/nar/gkq930

56. E. V. Kriventseva, N. Rahman, O. Espinosa, E. M. Zdobnov, OrthoDB: The hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res.* **36** (Database), D271–D275 (2008). Medline doi:10.1093/nar/gkm845

57. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39** (suppl), W29-37 (2011). Medline doi:10.1093/nar/gkr367

58. R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, M. Punta, Pfam: The protein families database. *Nucleic Acids Res.* **42** (D1), D222–D230 (2014). Medline doi:10.1093/nar/gkt1223

59. D. Wilson, V. Charoensawan, S. K. Kummerfeld, S. A. Teichmann, DBD—taxonomically broad transcription factor predictions: New content and functionality. *Nucleic Acids Res.* **36** (Database), D88–D92 (2008). Medline doi:10.1093/nar/gkm964

60. L. Huang, T. Cheng, P. Xu, T. Fang, Q. Xia, *Bombyx mori t*ranscription factors: Genome-wide identification, expression profiles and response to pathogens by microarray analysis. *J. Insect Sci.* **12**, 40 (2012). Medline doi:10.1673/031.012.4001

61. G. Benson, Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999). Medline doi:10.1093/nar/27.2.573

62. S. Sinha, E. van Nimwegen, E. D. Siggia, A probabilistic method to detect regulatory modules. *Bioinformatics* **19** (Suppl 1), i292–i301 (2003). Medline doi:10.1093/bioinformatics/btg1040

63. J. Felsenstein, Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985). doi:10.1086/284325

64. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004). Medline doi:10.1093/bioinformatics/btg412

65. A. P. Bird, DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980). Medline doi:10.1093/nar/8.7.1499

66. N. Elango, B. G. Hunt, M. A. Goodisman, S. V. Yi, DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11206–11211 (2009). Medline doi:10.1073/pnas.0900301106

67. A. Zemach, I. E. McDaniel, P. Silva, D. Zilberman, Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010). Medline doi:10.1126/science.1186366

68. B. G. Hunt, K. M. Glastad, S. V. Yi, M. A. Goodisman, Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. *Genome Biol. Evol.* **5**, 591–598 (2013). Medline doi:10.1093/gbe/evt030

69. R Development Core Team, R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2011).

70. N. Lartillot, Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Mol. Biol. Evol.* **30**, 356–368 (2013). Medline doi:10.1093/molbev/mss231

71. N. Lartillot, Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol. Biol. Evol.* **30**, 489–502 (2013). Medline doi:10.1093/molbev/mss239

72. J. B. Wolf, A. Künstner, K. Nam, M. Jakobsson, H. Ellegren, Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. *Genome Biol. Evol.* **1**, 308–319 (2009). Medline doi:10.1093/gbe/evp030

73. Z. Yang, Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568–573 (1998). Medline doi:10.1093/oxfordjournals.molbev.a025957

74. R. Nielsen, Z. Yang, Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936 (1998). Medline

75. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009). Medline doi:10.1093/nar/gkn923

76. R. Gentleman, S. Falcon, *Package 'GOstats'*. (ed. 2.26.0, 2013).

77. J. Parker, G. Tsagkogeorga, J. A. Cotton, Y. Liu, P. Provero, E. Stupka, S. J. Rossiter, Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013). Medline doi:10.1038/nature12511

78. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997). Medline doi:10.1093/nar/25.17.3389

79. K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, B. Barrell, Artemis: Sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000). Medline doi:10.1093/bioinformatics/16.10.944

80. W. J. Kent, BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002). Medline doi:10.1101/gr.229202. Article published online before March 2002

81. C. Notredame, D. G. Higgins, J. Heringa, T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000). Medline doi:10.1006/jmbi.2000.4042

82. S. Guindon, F. Delsuc, J.-F. Dufayard, O. Gascuel, in *Bioinformatics for DNA Sequence Analysis,* D. Posada, Ed. (Humana Press, New York, NY, 2009), vol. 537, chap. 6, pp. 113-137.

83. S. R. Eddy, Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998). Medline doi:10.1093/bioinformatics/14.9.755

84. S. R. Eddy, A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on Genome Informatics* **23**, 205-211 (2009).

85. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004). Medline doi:10.1093/nar/gkh340

86. K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011). Medline doi:10.1093/molbev/msr121

87. S. L. Pond, S. D. Frost, S. V. Muse, HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* **21**, 676–679 (2005). Medline doi:10.1093/bioinformatics/bti079

88. T. De Bie, N. Cristianini, J. P. Demuth, M. W. Hahn, CAFE: A computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006). Medline doi:10.1093/bioinformatics/btl097

89. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011). Medline doi:10.1038/nmeth.1701

90. M. Lechner, S. Findeiss, L. Steiner, M. Marz, P. F. Stadler, S. J. Prohaska, Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011). Medline doi:10.1186/1471-2105-12-124

91. C. Claudianos, H. Ranson, R. M. Johnson, S. Biswas, M. A. Schuler, M. R. Berenbaum, R. Feyereisen, J. G. Oakeshott, A deficit of detoxification enzymes: Pesticide sensitivity and environmental response in the honeybee. *Insect Mol. Biol.* **15**, 615–636 (2006). Medline doi:10.1111/j.1365-2583.2006.00672.x

92. J. G. Oakeshott, R. M. Johnson, M. R. Berenbaum, H. Ranson, A. S. Cristino, C. Claudianos, Metabolic enzymes associated with xenobiotic and chemosensory responses in *Nasonia vitripennis*. *Insect Mol. Biol.* **19** (Suppl 1), 147–163 (2010). Medline doi:10.1111/j.1365-2583.2009.00961.x

93. N. Tijet, C. Helvig, R. Feyereisen, The cytochrome P450 gene superfamily in *Drosophila melanogaster*: Annotation, intron-exon organization and phylogeny. *Gene* **262**, 189–198 (2001). Medline doi:10.1016/S0378-1119(00)00533-3

94. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012). Medline doi:10.1038/nmeth.2109

95. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). Medline doi:10.1093/bioinformatics/btq033

96. E. Stafflinger, K. K. Hansen, F. Hauser, M. Schneider, G. Cazzamali, M. Williamson, C. J. Grimmelikhuijzen, Cloning and identification of an oxytocin/vasopressin-like receptor and its ligand from insects. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3262–3267 (2008). Medline doi:10.1073/pnas.0710897105

97. C. W. Gruber, M. Muttenthaler, Discovery of defense- and neuropeptides in social ants by genome-mining. *PLOS ONE* **7**, e32559 (2012). Medline doi:10.1371/journal.pone.0032559

98. S. D. Kocher, C. Li, W. Yang, H. Tan, S. V. Yi, X. Yang, H. E. Hoekstra, G. Zhang, N. E. Pierce, D. W. Yu, The draft genome of a socially polymorphic halictid bee, *Lasioglossum albipes*. *Genome Biol.* **14**, R142 (2013). Medline doi:10.1186/gb-2013-14-12-r142

99. T. Flutre, E. Duprat, C. Feuillet, H. Quesneville, Considering transposable element diversification in *de novo* annotation approaches. *PLOS ONE* **6**, e16526 (2011). Medline doi:10.1371/journal.pone.0016526

100. M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, R. D. Finn, The Pfam protein families database. *Nucleic Acids Res.* **40** (D1), D290–D301 (2012). Medline doi:10.1093/nar/gkr1065

101. T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, A. H. Schulman, A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007). Medline doi:10.1038/nrg2165

102. A. Marchler-Bauer, J. B. Anderson, M. K. Derbyshire, C. DeWeese-Scott, N. R. Gonzales, M. Gwadz, L. Hao, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, D. Krylov, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, N. Thanki, R. A. Yamashita, J. J. Yin, D. Zhang, S. H. Bryant, CDD: A conserved domain database for interactive domain family analysis. *Nucleic Acids Res.* **35** (Database), D237–D240 (2007). Medline doi:10.1093/nar/gkl951

103. O. Kohany, A. J. Gentles, L. Hankus, J. Jurka, Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**, 474 (2006). Medline doi:10.1186/1471-2105-7-474

104. V. V. Kapitonov, S. Tempel, J. Jurka, Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**, 207–213 (2009). Medline doi:10.1016/j.gene.2009.07.019

105. V. V. Kapitonov, J. Jurka, A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.* **9**, 411–412, author reply 414 (2008). Medline doi:10.1038/nrg2165-c1

106. Y. W. Yuan, S. R. Wessler, The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7884–7889 (2011). Medline doi:10.1073/pnas.1104208108

107. B. G. Hunt, K. M. Glastad, S. V. Yi, M. A. Goodisman, The function of intragenic DNA methylation: Insights from insect epigenomes. *Integr. Comp. Biol.* **53**, 319–328 (2013). Medline doi:10.1093/icb/ict003

108. H. E. Amarasinghe, C. I. Clayton, E. B. Mallon, Methylation and worker reproduction in the bumble-bee (*Bombus terrestris*). *Proc. Biol.Sci.* **281**, 20132502 (2014). Medline doi:10.1098/rspb.2013.2502

109. R. Kucharski, J. Maleszka, S. Foret, R. Maleszka, Nutritional control of reproductive status in honeybees via DNA methylation. *Science* **319**, 1827–1830 (2008). Medline doi:10.1126/science.1153069

110. G. A. Lockett, R. Kucharski, R. Maleszka, DNA methylation changes elicited by social stimuli in the brains of worker honey bees. *Genes Brain Behav.* **11**, 235–242 (2012). Medline doi:10.1111/j.1601-183X.2011.00751.x

111. F. Lyko, S. Foret, R. Kucharski, S. Wolf, C. Falckenhayn, R. Maleszka, The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLOS Biol.* **8**, e1000506 (2010). Medline doi:10.1371/journal.pbio.1000506

112. A. Eyre-Walker, L. D. Hurst, The evolution of isochores. *Nat. Rev. Genet.* **2**, 549–555 (2001). Medline doi:10.1038/35080577

113. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652–654 (1991). Medline doi:10.1038/351652a0

114. A. Eyre-Walker, P. D. Keightley, Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol. Biol. Evol.* **26**, 2097–2108 (2009). Medline doi:10.1093/molbev/msp119

115. A. Buttstedt, R. F. A. Moritz, S. Erler, Origin and function of the major royal jelly proteins of the honeybee (*Apis mellifera*) as members of the yellow gene family. *Biol. Rev. Camb. Philos. Soc.* **89**, 255–269 (2014). Medline doi:10.1111/brv.12052

116. M. D. Drapeau, S. Albert, R. Kucharski, C. Prusko, R. Maleszka, Evolution of the Yellow/Major Royal Jelly Protein family and the emergence of social behavior in honey bees. *Genome Res.* **16**, 1385–1394 (2006). Medline doi:10.1101/gr.5012006

117. S. M. Hedtke, S. Patiny, B. N. Danforth, The bee tree of life: A supermatrix approach to apoid phylogeny and biogeography. *BMC Evol. Biol.* **13**, 138 (2013). Medline doi:10.1186/1471-2148-13-138

118. M. Kamakura, Royalactin induces queen differentiation in honeybees. *Nature* **473**, 478–483 (2011). Medline doi:10.1038/nature10093

119. A. Patel, M. K. Fondrk, O. Kaftanoglu, C. Emore, G. Hunt, K. Frederick, G. V. Amdam, The making of a queen: TOR pathway is a key player in diphenic caste development. *PLOS ONE* **2**, e509 (2007). Medline doi:10.1371/journal.pone.0000509

120. N. S. Mutti, A. G. Dolezal, F. Wolschin, J. S. Mutti, K. S. Gill, G. V. Amdam, IRS and TOR nutrient-signaling pathways act via juvenile hormone to influence honey bee caste fate. *J. Exp. Biol.* **214**, 3977–3984 (2011). Medline doi:10.1242/jeb.061499

121. A. Teles, T. Mello, A. Barchuk, Z. Simões, *Ultraspiracle* of the stingless bees *Melipona scutellaris* and *Scaptotrigona depilis*: cDNA sequence and expression profiles during pupal development. *Apidol* **38**, 462–471 (2007). doi:10.1051/apido:2007035

122. A. R. Barchuk, V. L. Figueiredo, Z. L. Simões, Downregulation of ultraspiracle gene expression delays pupal development in honeybees. *J. Insect Physiol.* **54**, 1035–1040 (2008). Medline doi:10.1016/j.jinsphys.2008.04.006

123. S. V. Azevedo, O. A. Caranton, T. L. de Oliveira, K. Hartfelder, Differential expression of hypoxia pathway genes in honey bee (*Apis mellifera* L.) caste development. *J. Insect Physiol.* **57**, 38–45 (2011). Medline doi:10.1016/j.jinsphys.2010.09.004

124. M. Jindra, S. R. Palli, L. M. Riddiford, The juvenile hormone signaling pathway in insect development. *Annu. Rev. Entomol.* **58**, 181–204 (2013). Medline doi:10.1146/annurev-ento-120811-153700

125. D. Bopp, G. Saccone, M. Beye, Sex determination in insects: Variations on a common theme. *Sex Dev.* **8**, 20–28 (2014). Medline doi:10.1159/000356458

126. M. Beye, M. Hasselmann, M. K. Fondrk, R. E. Page Jr., S. W. Omholt, The gene *csd* is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell* **114**, 419–429 (2003). Medline doi:10.1016/S0092-8674(03)00606-8

127. M. Hasselmann, X. Vekemans, J. Pflugfelder, N. Koeniger, G. Koeniger, S. Tingek, M. Beye, Evidence for convergent nucleotide evolution and high allelic turnover rates at the complementary sex determiner gene of Western and Asian honeybees. *Mol. Biol. Evol.* **25**, 696–708 (2008). Medline doi:10.1093/molbev/msn011

128. E. Privman, Y. Wurm, L. Keller, Duplication and concerted evolution in a master sex determiner under balancing selection. *Proc. Biol. Sci.* **280**, 20122968 (2013). Medline doi:10.1098/rspb.2012.2968

129. V. Koch, I. Nissen, B. D. Schmitt, M. Beye, Independent evolutionary origin of *fem* paralogous genes and complementary sex determination in hymenopteran insects. *PLOS ONE* **9**, e91883 (2014). Medline doi:10.1371/journal.pone.0091883

130. S. Lechner, L. Ferretti, C. Schöning, W. Kinuthia, D. Willemsen, M. Hasselmann, Nucleotide variability at its limit? Insights into the number and evolutionary dynamics of the sex-determining specificities of the honey bee *Apis mellifera*. *Mol. Biol. Evol.* **31**, 272–287 (2014). Medline doi:10.1093/molbev/mst207

131. T. Gempe, M. Hasselmann, M. Schiøtt, G. Hause, M. Otte, M. Beye, Sex determination in honeybees: Two separate mechanisms induce and maintain the female pathway. *PLOS Biol.* **7**, e1000222 (2009). Medline doi:10.1371/journal.pbio.1000222

132. F. Hauser, G. Cazzamali, M. Williamson, W. Blenau, C. J. Grimmelikhuijzen, A review of neurohormone GPCRs present in the fruitfly *Drosophila melanogaster* and the honey bee *Apis mellifera*. *Prog. Neurobiol.* **80**, 1–19 (2006). Medline doi:10.1016/j.pneurobio.2006.07.005

133. F. Hauser, G. Cazzamali, M. Williamson, Y. Park, B. Li, Y. Tanaka, R. Predel, S. Neupert, J. Schachtner, P. Verleyen, C. J. Grimmelikhuijzen, A genome-wide inventory of neurohormone GPCRs in the red flour beetle *Tribolium castaneum*. *Front. Neuroendocrinol.* **29**, 142–165 (2008). Medline doi:10.1016/j.yfrne.2007.10.003

134. S. Nygaard, G. Zhang, M. Schiøtt, C. Li, Y. Wurm, H. Hu, J. Zhou, L. Ji, F. Qiu, M. Rasmussen, H. Pan, F. Hauser, A. Krogh, C. J. P. Grimmelikhuijzen, J. Wang, J. J. Boomsma, The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res.* **21**, 1339–1348 (2011). Medline doi:10.1101/gr.121392.111

135. J. D. Evans, K. Aronstein, Y. P. Chen, C. Hetru, J. L. Imler, H. Jiang, M. Kanost, G. J. Thompson, Z. Zou, D. Hultmark, Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol. Biol.* **15**, 645–656 (2006). Medline doi:10.1111/j.1365-2583.2006.00682.x

136. R. Feyereisen, Arthropod CYPomes illustrate the tempo and mode in P450 evolution. *Biochim. Biophys. Acta* **1814**, 19–28 (2011). Medline doi:10.1016/j.bbapap.2010.06.012

137. T. I. A. G. Consortium; International Aphid Genomics Consortium, Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLOS Biol.* **8**, e1000313 (2010). Medline doi:10.1371/journal.pbio.1000313

138. T. Yang, N. Liu, Genome analysis of cytochrome P450s and their expression profiles in insecticide resistant mosquitoes, *Culex quinquefasciatus*. *PLOS ONE* **6**, e29418 (2011). Medline doi:10.1371/journal.pone.0029418

139. S. H. Lee, J. S. Kang, J. S. Min, K. S. Yoon, J. P. Strycharz, R. Johnson, O. Mittapalli, V. M. Margam, W. Sun, H. M. Li, J. Xie, J. Wu, E. F. Kirkness, M. R. Berenbaum, B. R. Pittendrigh, J. M. Clark, Decreased detoxification genes and genome size make the human body louse an efficient model to study xenobiotic metabolism. *Insect Mol. Biol.* **19**, 599–615 (2010). Medline doi:10.1111/j.1365-2583.2010.01024.x

140. R. Feyereisen, in *Insect Molecular Biology and Biochemistry,* L. I. Gilbert, Ed. (Academic Press, London, 2012).

141. M. Maïbèche-Coisne, A. A. Nikonov, Y. Ishida, E. Jacquin-Joly, W. S. Leal, Pheromone anosmia in a scarab beetle induced by in vivo inhibition of a pheromone-degrading enzyme. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11459–11464 (2004). Medline doi:10.1073/pnas.0403537101

142. Y. Qiu, C. Tittiger, C. Wicker-Thomas, G. Le Goff, S. Young, E. Wajnberg, T. Fricaux, N. Taquet, G. J. Blomquist, R. Feyereisen, An insect-specific P450 oxidative decarbonylase for cuticular hydrocarbon biosynthesis. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14858–14863 (2012). Medline doi:10.1073/pnas.1208650109

143. N. L. Kirischian, J. Y. Wilson, Phylogenetic and functional analyses of the cytochrome P450 family 4. *Mol. Phylogenet. Evol.* **62**, 458–471 (2012). Medline doi:10.1016/j.ympev.2011.10.016

144. C. R. Smith, C. D. Smith, H. M. Robertson, M. Helmkampf, A. Zimin, M. Yandell, C. Holt, H. Hu, E. Abouheif, R. Benton, E. Cash, V. Croset, C. R. Currie, E. Elhaik, C. G. Elsik, M. J. Favé, V. Fernandes, J. D. Gibson, D. Graur, W. Gronenberg, K. J. Grubbs, D. E. Hagen, A. S. Viniegra, B. R. Johnson, R. M. Johnson, A. Khila, J. W. Kim, K. A. Mathis, M. C. Munoz-Torres, M. C. Murphy, J. A. Mustard, R. Nakamura, O. Niehuis, S. Nigam, R. P. Overson, J. E. Placek, R. Rajakumar, J. T. Reese, G. Suen, S. Tao, C. W. Torres, N. D. Tsutsui, L. Viljakainen, F. Wolschin, J. Gadau, Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5667–5672 (2011). Medline doi:10.1073/pnas.1007901108

145. C. D. Smith, A. Zimin, C. Holt, E. Abouheif, R. Benton, E. Cash, V. Croset, C. R. Currie, E. Elhaik, C. G. Elsik, M. J. Fave, V. Fernandes, J. Gadau, J. D. Gibson, D. Graur, K. J. Grubbs, D. E. Hagen, M. Helmkampf, J. A. Holley, H. Hu, A. S. Viniegra, B. R. Johnson, R. M. Johnson, A. Khila, J. W. Kim, J. Laird, K. A. Mathis, J. A. Moeller, M. C. Muñoz-Torres, M. C. Murphy, R. Nakamura, S. Nigam, R. P. Overson, J. E. Placek, R. Rajakumar, J. T. Reese, H. M. Robertson, C. R. Smith, A. V. Suarez, G. Suen, E. L. Suhr, S. Tao, C. W. Torres, E. van Wilgenburg, L. Viljakainen, K. K. Walden, A. L. Wild, M. Yandell, J. A. Yorke, N. D. Tsutsui, Draft genome of the globally widespread and

invasive Argentine ant (*Linepithema humile*). *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5673–5678 (2011). Medline doi:10.1073/pnas.1008617108

146. W. Mao, S. G. Rupasinghe, R. M. Johnson, A. R. Zangerl, M. A. Schuler, M. R. Berenbaum, Quercetin-metabolizing CYP6AS enzymes of the pollinator *Apis mellifera* (Hymenoptera: Apidae). *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **154**, 427–434 (2009). Medline doi:10.1016/j.cbpb.2009.08.008

147. S. G. Oliveira, W. Bao, C. Martins, J. Jurka, Horizontal transfers of Mariner transposons between mammals and insects. *Mob. DNA* **3**, 14 (2012). Medline doi:10.1186/1759-8753-3-14

148. H. M. Robertson, Multiple Mariner transposons in flatworms and hydras are related to those of insects. *J. Hered.* **88**, 195–201 (1997). Medline doi:10.1093/oxfordjournals.jhered.a023088

149. M. Kimura, *The Neutral Theory of Molecular Evolution*. (Cambridge Univ. Press, Cambridge, 1983).

150. M. Bulmer, The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907 (1991). Medline

151. W.-H. Li, Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345 (1987). Medline doi:10.1007/BF02134132