Supplemental Materials

Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality

Daniel F. Simola, Lothar Wissler, Greg Donahue, Robert M. Waterhouse, Martin Helmkampf, Julien Roux, Sanne Nygaard, Karl M. Glastad, Darren E. Hagen, Lumi Viljakainen, Justin T. Reese, Brendan G. Hunt, Dan Graur, Eran Elhaik, Evgenia V. Kriventseva, Jiayu Wen, Brian J. Parker, Elizabeth Cash, Eyal Privman, Christopher P. Childers, Monica C. Muñoz-Torres, Jacobus J. Boomsma, Erich Bornberg-Bauer, Cameron Currie, Christine G. Elsik, Garret Suen, Michael A. D. Goodisman, Laurent Keller, Jürgen Liebig, Alan Rawls, Danny Reinberg, Chris D. Smith, Chris R. Smith, Neil Tsutsui, Yannick Wurm, Evgeny M. Zdobnov, Shelley L. Berger, and Jürgen Gadau

Table of contents

List of Contributions

Project Coordination: Jürgen Gadau, **GC Composition Analysis:** Eran Elhaik, Justin T. Reese, Dan Graur, Christine G. Elsik, **Gene Homology (AntOrthoDB), Phylogeny, Gene Set Quality Assessment:** Robert M. Waterhouse, Evgenia V. Krisventseva, Evgeny M. Zdobnov, Lothar Wissler, **Taxonomically Restricted Genes, Genes With Paralogs:** Lothar Wissler, Erich Bornberg-Bauer, **Codon Usage Bias:** Julien Roux, **Gene Family Evolution:** Martin Helmkampf, Lothar Wissler, **Desaturases:** Martin Helmkampf, Elizabeth Cash, **Immune Genes:** Lumi Viljakainen, **Multiple Genome Alignment:** Daniel F. Simola, **Synteny:** Greg Donahue, **Conserved Elements:** Daniel F. Simola, **DNA Methylation:** Karl M. Glastad, Brendan G. Hunt, Michael A. D. Goodisman, **Conserved Structural RNAs:** Sanne Nygaard, Jiayu Wen, Brian J. Parker, **Micro RNA:** Darren E. Hagen, Christine G. Elsik, **Transcription Factor Binding Sites, RNA Expression Analysis:** Daniel F. Simola, Alan Rawls, Jürgen Gadau, **Positive Selection:** Julien Roux, Eyal Privman, **Hymenoptera Genome Database:** Christopher P. Childers, Monica C. Muñoz-Torres, Justin T. Reese, Christine G. Elsik, **Additional Manuscript Preparation:** Chris R. Smith, Christopher D. Smith, Neil Tsutsui, Garret Suen, Cameron Currie, Yannick Wurm, Laurent Keller, Jürgen Liebig, Danny Reinberg, Shelley L. Berger

Materials and Methods

Quality Assessment of Annotated Gene Sets in AntOrthoDB

Orthologs ‘present-in-all-but-one’ species

AntOrthoDB (Supplemental Fig. 2, Supplemental Table 1) orthologous groups were examined to identify groups with genes from 11 out of the 12 species which contain (i) strictly one gene in each species (single-copy) and (ii) at least one species with more than one gene (multi-copy). Orthologous genes that are present in all but one of the 12 insects indicate true gene losses or genes that are missing from the genome annotation or assembly. This analysis revealed that amongst the seven ant species, there are generally few lost or missing genes, apart from *S. invicta* (~500 genes) and *A. echinatior* (~200 genes) (Supplemental Fig. 3).

Potentially missing or missed orthologs

AntOrthoDB orthologous groups delineated across the seven ant species plus *A. mellifera* and *N. vitripennis* were examined to identify those with gene members in honeybee and/or wasp but without gene members in one or two ant species. For each potentially missing or missed ant gene, a seed gene was identified from a closely-related ant species and three TBLASTN searches were performed: the seed protein sequence against the genome of the species where the ortholog appears to be missing, its own genome, and the genome of the bee or wasp (outgroup species). The results were analyzed to distinguish cases where the ortholog is indeed likely to be missing from the assembled genome – either no significant BLAST hits were found, the hits were less significant than those to the outgroup genome, or the ortholog may have been missed by the annotation procedure (or was poorly annotated and hence failed the orthology delineation procedure); otherwise the BLAST hits were more significant than those to the outgroup genome. This analysis identified 3,313 potentially missing or missed ant orthologs from 2,635 orthologous groups (Supplemental Fig. 4): ~200 cases from *LHUMI*, *PBARB*, and *ACEPH*, ~450 cases from *HSALT* and *CFLOR*, ~650 cases from *AECHI*, and ~1,100 cases from *SINVI*. Cases with 'No Hits' or 'Probably Missing' may be true gene losses, which occur relatively frequently in insect evolution, and which may relate to certain specific biological traits of each species. Cases with 'Probably Present' may highlight potential errors with the automatic gene annotation procedures, resulting in incomplete or missed gene annotations. Hence, 'Probably Present' genes should be targeted for manual curation efforts to improve future releases of the official gene sets.

Protein length concordance among orthologs

Employing 4,346 single-copy orthologs defined across the seven ant species, *A. mellifera* (*AMELL*) and *N. vitripennis* (*NVITR*), protein lengths were compared to examine the agreement of predicted ant genes with those from honeybee and wasp (Supplemental Fig. 6 and Supplemental Table 2). This analysis compared some of the most accurately predicted proteins in each species, as conserved single-copy orthologs are often the simplest genes to predict using homology-based approaches. As a baseline, the honey bee – *Nasonia* wasp comparison shows a concordance of 0.91 with more bee proteins that are shorter than wasp their orthologs. Compared to the honey bee, ant protein length concordance values range from 0.91 for *HSALT* to 0.83 for *SINVI*, and *HSALT*, *LHUMI*, *CFLOR*, and *AECHI*, tend towards longer coding-sequence predictions while *PBARB*, *SINVI*, and *ACEPH* tend towards shorter predictions. Compared to the *Nasonia* wasp, ant protein length concordance values range from 0.90 for *HSALT* to 0.81 for *SINVI*, and all the ant species tend towards shorter predictions.

Codon Usage Bias

Complete coding sequences (CDS) corresponding to all annotated genes in the 7 ant genomes were downloaded from http://antlab.sfsu.edu/~antdata/ (Supplemental Table 4). CDS sequences of the 5 outgroups were downloaded from their respective genome project homepages. Sequences quality was controlled as in (Hambuch and Parsch 2005): CDS sequences whose length was not a multiple of three, did not correspond to the length of the predicted protein or contained an internal stop codon were eliminated; the longest CDS of genes showing multiple isoforms was retained; CDS shorter than 100 nt were eliminated as short sequences can affect the measure of codon usage bias.

The analysis was performed both on the full dataset and on the subset of genes having only single-copy orthologs in the 12 species, based on the AntOrthoDB analyses (above). This guarantees that the results are not due to patterns of species-specific genes or species-specific duplicate genes. Only the results from this dataset are described here but the results were virtually unchanged when the complete dataset was used.

Codon usage bias was estimated using the "effective number of codons" measure (ENC or N_C) (Wright 1990). ENC values range from 20 in the case of extreme bias where one codon is exclusively used for each amino-acid, to 61 when the use of alternative synonymous codons is equally likely. ENC is thus a simple measure that can be used to quantify how far the codon usage of genes from different species departs from equal usage of synonymous codons (Drosophila 12 Genomes Consortium 2007, Vicario et al. 2007, Heger and Ponting 2007, Gingold et al. 2011). ENC measures were calculated for all genes of the 12 species using the CodonW program (*http://codonw.sourceforge.net/*) (Supplemental Fig. 9). CodonW reports for %GC and %GC at 3rd positions of synonymous codons (GC3s) in CDS sequences were also used in the analysis.

Cytoplasmic ribosomal proteins for 10 of the 12 genomes were obtained from M. Helmkampf (7 ants + *D. melanogaster*, *A. mellifera* and *N. vitripennis*; see CD Smith et al 2011, CR Smith et al. 2011, Suen et al. 2011). To retrieve ribosomal proteins from the genomes of *P. humanus* and *T. castaneum*, a similar methodology was applied: all ribosomal proteins of *D. melanogaster* were blasted (BLASTP) against the proteomes and all hits with an e-value smaller than 1e–10 were retained (Supplemental Table 4).

The dataset of randomized CDS sequences was created to reproduce the properties of the real dataset regarding CDS lengths and nucleotide compositions: for each protein of the real dataset, a new CDS sequence was created by randomly choosing at each position a new codon among all of the synonymous codons displaying the same %GC content.

Genes With Paralogs

Across the 30 arthropod genomes, the number of genes with paralogs (GWPs), which may serve as a rough approximation for genetic redundancy, was determined. GWP counts were derived from BLASTP-based inference of homology and single-linkage clustered gene families (see section on Gene Family Evolution below). A total of four different sets of criteria were used to define homology, i.e. when two protein sequences are considered homologs based in the BLASTP hit, to prevent misinterpretation due to threshold effects. These sets define when two protein sequences are considered to be homologs which affects the E-value cutoff, the minimum alignment coverage of the local alignment constructed by BLAST, and the minimum percent identity within the local alignment:

	Set 1	*Set 2*	*Set 3*	*Set 4*
Min. E-value	1e–10	1e–20	1e–5	1e–10
Min. alignment coverage	70%	0	30%	50%
Min. sequence identity	30%	0	30%	0

Overall, the number of GWPs is relatively homogeneous among different groups of insects, and the distributions of GWPs seem independent from the exact homolog definition as all sets show comparable trends (Supplemental Fig. 8, Supplemental Table 3). The three mosquitoes (Culicidae) seem to be a slight outlier, showing higher GWP counts than the other tested groups, although it is currently unclear how much this trend might be influenced by the low taxon sampling. GWP counts in ants (Formicidae) are comparable to the other two Hymenoptera and to those in Drosophilidae. Among ants, the highest redundancy was found in the *H. saltator* genome (Supplemental Fig. 8).

Gene Family Evolution

Gene families were identified by a relaxed reciprocal BLAST method (Drosophila 12 Genomes Consortium 2007) and subsequent single-linkage clustering. Clustering of genes into gene families was done using their encoded protein sequences and performing all vs. all BLASTP searches. Gene models were obtained from official gene sets available from 30 arthropod species with fully sequenced genomes. All protein-coding genes were added to a graph as nodes. If a homologous relationship between query and subject proteins was determined with BLAST, a directed edge was added from query to subject. A query gene was considered a homolog to a subject gene if the BLAST E-value was smaller or equal to 1e–10, and the local alignment covered at least 70% of the longer of the two sequences with at least 30% sequence identity. After all BLAST hits were evaluated, non-reciprocal edges were removed from the graph. Finally, gene families were obtained from the graph as subgraphs using single linkage clustering.

For the evolutionary analysis of gene families, only 15 of the 30 arthropod species were retained for the final dataset, including five drosophilids and six ant species. The remaining species were discarded due to large evolutionary distances to the focal ant species (e.g., *Ixodes scapularis*, *Daphnia pulex*), or due to concerns that differences in sequencing depth (e.g., drosophilids sequenced to low coverage) or annotation quality (e.g., *S. invicta*, see below for a detailed explanation) might bias the analyses.

In total, 33,891 gene families were identified, of which 5,681 remained after filtering out families which were inferred to lack members in the most recent common ancestor of the 15 species using the `-filter` option provided by CAFE v2.2 (Hahn et al. 2007), the software used for all subsequent analyses of gene family evolution (Supplemental Fig. 10). This step removed all gene families with members in only hymenopteran or non-hymenopteran insect taxa (which represents the basal split in our species tree), including all lineage-specific gene families. An additional six gene families, predicted to be made up mostly of transposable elements, were discarded due to their exceptionally strong influence on parameter estimates during preliminary analyses, leaving a total of 5,675 families containing 111,420 genes in the final dataset. The largest of these gene families contained 1,822 members across all 15 species. The topology of the species phylogeny required by CAFE were taken from this study and the literature (Drosophila 12 Genomes Consortium 2007, Wiegmann et al. 2009). Divergence times were obtained from timetree.org (Hedges et al. 2006), a public resource reporting consensus estimates of divergence times, resulting in the following ultrametric tree:

```
((tcas:300,((dvir:47,(((dmel:13,dere:13):22,dana:35):2,dpse:37):1
0):228,aaeg:275):25):50,((((((pbar:110,(aech:8,acep:8):102):13,cf
lo:123):3,lhum:126):6,hsal:132):32,amel:164):27,nvit:190):160)
```

To assess the rate and direction of gene family size change along the phylogeny and to identify families which are characterized by significant size changes, we applied five models of gene gain and loss with varying numbers of parameters estimated within CAFE's maximum likelihood framework. Unless noted otherwise, probabilities for gene gain and loss were assumed to be equal, and did not vary between gene families:

– model 1: one parameter for one global rate of gene gain and loss (λ) on all branches of the phylogeny
– model 2: two parameters for two global rates, one for gene gain (λ) and one for gene loss (μ)
– model 3: three parameters, one each for ants, drosophilids, and other taxa
– model 4: three parameters for three rate categories

– model 5: four parameters for four rate categories
– model 6: five parameters for five rate categories

To assign branches to one of the rate categories in model 4–6, a two-parameter analysis was run for each branch, estimating a rate specific to the focal (foreground) branch and a background rate for the remaining branches. The branch-specific rates were then categorized by *k*-means clustering with $k = 3$, $k = 4$ and $k = 5$, respectively. This approach served as an approximation for the fully parameterized model with independent rates for each branch, which did not converge to a single maximum due to its complexity. We employed the likelihood ratio test to find the model which best fit the data.

Gene families with an overall size distribution that differed from the null distribution expected under random birth and death at a significance level of $p \leq 0.01$ were considered as having potentially evolved under the influence of natural selection. By calculating exact *p*-values for all transitions between parent and child nodes of these families (the "Viterbi" method; Hedges et al. 2006), we identified the branches characterized by the most unlikely amount of change. Transitions with a likelihood of $p \leq 0.01$ were considered significant, indicating lineage-specific adaptation. Gene families of interest were functionally annotated using BLAST against the Swiss-Prot (De Bie et al. 2006) and Pfam databases (The UniProt Consortium 2012), and Blast2GO (Punta et al. 2002) and the Gene Ontology database (Conesa et al. 2005). To test whether gene families with significant size changes in ants (i.e., along one or several internal or terminal ant branches) are significantly enriched in certain Gene Ontology terms in comparison to all gene families, we employed the topGO package implementing the elim algorithm which accounts for the tree-like, non-independent structure of GO categories (*p*-value ≤ 0.005) (The Gene Ontology Consortium 2000). All datasets are available from the authors upon request.

Why *S. invicta* was excluded from the gene family evolution analyses

In various analyses on annotated genes, we found hints that the gene annotation of the *S. invicta* genome (version 2.2.3) may not be exhaustive. As reported in the Gene Set Quality Assessment section above, *S. invicta* stands out among the seven ants displaying the highest number of missing genes and the shortest gene models. Similar patterns were found in KEGG pathway annotation (Alexa et al. 2006) obtained from KAAS (Kaneshina et al. 2012) with full proteomes and the BBH method (Supplemental Fig. 14). Despite our efforts to identify missing genes, we therefore excluded *S. invicta* from the gene family analysis to prevent potentially inflated estimates of gene turnover and incorrect ancestral gene counts in the statistical analysis of gene family size variation without a significant loss in phylogenetic resolution.

Desaturases

Desaturase genes were identified by reciprocal blastp using the *D. melanogaster* desat1 gene (CG5887) as query against the official gene sets of all seven ant species and *Acyrthosiphon pisum*, *Anopheles gambiae*, *A. mellifera*, *B. mori*, *D. melanogaster*, *N. vitripennis* and *T. castaneum*. Manual annotation was carried out for the ant species as described elsewhere (CR Smith et al. 2011), and functional gene copies were distinguished from pseudogenes by ORF length and number of premature stop codons.

A total of 179 putatively functional, homologous genes were aligned using the L-INS-i algorithm implemented in MAFFT v6 (Katoh et al. 2002). Ambiguously aligned positions were eliminated by ALISCORE (Moriya et al. 2007). Based on the LG+G substitution model (Misof et al. 2009), a maximum likelihood tree was then constructed using RAxML v.7.2.6 (Le, Gascuel 2008). Nodal confidence values were computed by performing a rapid bootstrap analysis with 500 replicates.

Immune Genes

In the comparison of immune gene contents across insects manually annotated immune genes of *P. barbatus*, *L. humile,* and *A. cephalotes* were used. In addition, immune genes were identified in the genomes of *A. echinatior*, *S. invicta*, *C. floridanus* and *H. saltator* using honeybee immune proteins as a query and the reciprocal best hit approach in the similarity searches as described in (CD Smith et al. 2011, CR Smith et al. 2011, Suen et al. 2011). For gene family characterization hidden Markov model (HMM) profiles (Stamatakis 2006) were made in HMMER3 (Eddy 1998) for the following immune gene families: lysozymes, thioester-containing proteins (TEPs), Gram-negative-bacteria-binding proteins (GNBPs), peptidoglycan-recognition proteins (PGRPs), fibrinogen-related proteins (FREPs), galectins, class B and class C scavenger receptors (SCR-B and SCR-C), clip-domain serine proteases (CLIPs), serine protease inhibitors (serpins) and C-type lectins (CTLs). The profiles were based on alignments of immune gene sequences retrieved from ImmunoDB (*http://cegg.unige.ch/Insecta/immunodb*) and corresponding honeybee sequences. These profiles were used in a HMMER3 search against each ant genome in order to find homologs for each gene family. Detailed immune gene identification based on the same approaches was also made for *N. vitripennis*. For *A. mellifera*, *T. castaneum*, *B. mori* and *Drosophila* data for immune gene family sizes were obtained from published analyses (Durbin et al 1998, Sackton and Clark 2009, Sackton et al. 2007, Zou et al. 2007, Tanaka et al. 2008). The HMM profiles were tested against honeybee, *Drosophila* and *T. castaneum* genomes, and the same number of paralogs for each immune gene family was found as reported in the published analyses except for cSPs, for which a smaller number of paralogs was found in all three species (see Supplemental Table 6).

Multiple Genome Alignment

Whole-genome multiple alignments of ant genome sequences were generated for three taxonomic groups: Formicidae (n=7), Myrmicinae (n=4), and Attini (n=2). These alignments were generated in two stages: (1) identification and (2) alignment of homologous contigs among species. First, homologous DNA sequence contigs from each assembled genome in the taxonomic group of interest were identified using Mercator (Lall et al. 2006), given softmasked versions of each genome sequence and a set of constraints defined as the significant nucleotide alignments between exons of 9,975 single-copy orthologs among the seven ant species. OrthoMCL (Chen et al. 2006) was used to identify these single-copy orthologs, and Blat (Kent 2002) was used to align all pairs of exons for these genes, retaining high scoring pairs (HSPs) with at least 90% sequence identity over at least 22 nt. On average, this yielded 81,811 significant alignments (or 8.2 HSPs per gene) between pairs of species. Mavid (Bray and Pachter 2004) was then used for multiple alignment of the resulting set of homologous contigs, given the phylogenetic tree estimated for the ant species by PAML (Yang 2007) using the single-copy orthologous gene sequences (4-fold degenerate sites codon-based model). Recursive optimization (`--r`) was used during alignment. Gene annotations were ported to the multiple alignments using custom software, removing any annotations with imperfect sequence identity between an individual genome and its alignment (due to spurious homology assessment).

Synteny

AntOrthoDB (above) discovered 244,281 relationships among all possible gene pairs from different ant species. A relationship is supported by synteny if the two genes (i) fall within 20 kb of each other on the seven-species alignment; (ii) share, among their thirty nearest neighbors on the seven-species alignment, at least three one-to-one orthologs; or (iii) share, among their thirty nearest neighbors or all other genes on the same scaffold, at least three confirmed one-to-one orthologs.

Large regions of pairwise conservation (synteny blocks) were assessed in the following way. For every pair of scaffolds from two different species sharing at least one pair of genes related by synteny, a syntenic block was defined as a pair of regions, one on each scaffold, from the most upstream to the most downstream genes involved in syntenic relationships with genes on the opposite scaffold. Blocks are not sub-divided over inversions, rearrangements, or internal syntenic relationships to other scaffolds.

Synteny plots were created in the following way. For a given scaffold in *A. echinatior* (the "pivot species") and a given comparison species, the rank-order of genes on the *A. echinatior* scaffold was used to construct a composite scaffold in the comparison species which minimizes rank disruption (in other words, the composite scaffold is a sequence of scaffolds from the comparison species placed in an order which produces the least departure from the *A. echinatior* rank-order). Orthology relationships are then plotted by their size in *A. echinatior* and their rank in *A. echinatior* (abscissa) and on the composite scaffold (ordinate) and colored by whether there is an inversion.

Conserved Structural RNAs

The EvoFold (Pedersen et al. 2006) RNA structure screen was based on the 7 ant species multiple alignment, which was used for structure prediction by utilizing comparative genomics features of conserved structure, such as compensatory double substitutions and compatible single substitutions. The screen was restricted to a set of conserved alignment segments based on the PhastCons predicted elements, as paired regions of structural RNAs evolve slowly. PhastCons regions were extended by 20 bases and combined when overlapping to also include fast-evolving single-stranded regions. Since EvoFold is sensitive to misaligned sequences, we applied a conservative sequence filter to the extracted alignment segments, which discards sequences with a significant excess number of mismatches given the branch-lengths of the relating phylogenetic tree (Parker et al. 2011). *C. floridanus* was used as reference species and gaps in it removed from the alignments. EvoFold (v.2.0) (Pedersen et al. 2006) was then applied to these filtered alignments in both their forward and reverse directions, in overlapping windows of length 150 bp with an offset of 50 bp. Low-confidence predictions that are short (< 7 base-pairs); with excessive amount of bulges (>50% bulges in stem); based on shallow or low quality alignments (removal of low confidence base pairs with posterior probability <

50%; removal of dangling base pairs; sequences > 25% bp cannot form structure; sequences > 7.5% positions are gapped; sequences > 10% contradictory substitution; entries with sequence counts < 3); or overlapping repeats were eliminated from the prediction set. P-values for double substitutions evidence were computed using a Monte Carlo test as in (Parker et al. 2011), though due to the small number of species in the alignments, an independent test set could not be held out. We defined a high confidence set with P-value < 0.05.

Genomic regions were defined as follows: coding sequences (CDS) annotation from *C. floridanus* was used, 3'UTR and 5'UTR were defined by the 3rd quartile of the UTR sizes in the well-annotated *D. melanogaster* genome (3' UTR length: 600 bp and 5' UTR length: 250 bp). Each prediction was assigned to the genomic region it had the greatest overlap with. GO enrichments for the structures were defined with the overall homologous gene GO annotations (blast2GO results) as a background, to reveal the additional enrichment of GO terms of structures above background. The TopGO package (The Gene Ontology Consortium 2000) in R bioconductor was used for the GO analysis, calculating P values with the "elim" method. Intronic, CDS, and UTR structures were assigned the GO of their enclosing gene (defined by >=1 bp overlap); intergenic regions were excluded. The structures were tested for homology against the structures of RFAM v. 10.1. Hits above the defined RFAM noise cutoff (NC), which are likely homologues, are considered to be significant. All structure predictions are available for viewing and download at *http://people.binf.ku.dk/jeanwen/data/ants*.

Positive Selection

We applied the branch-site test of the program Codeml from the PAML package (Yang 2007, Zhang et al. 2005) to 4,261 gene families which did not experience duplications (single-orthologs families). A model allowing for positive selection at some sites of a protein, on a selected branch of the tree (ratio of number of non-synonymous substitutions per non-synonymous site over number of synonymous substitutions per synonymous site, d_N/d_S or $\omega > 1$) is compared through a Likelihood Ratio Test (LRT) to a model where elevated rates of evolution on this branch are due to relaxed selective constraints ($d_N/d_S \sim 1$) (Zhang et al. 2005, Yang and dos Reis 2011). We successively changed the branch of interest to test for positive selection on 15 branches of the insect phylogeny (Supplemental Fig. 26) and FDR-corrected the ensemble of p-values (Yang and dos Reis 2011, Anisimova and Yang 2007, Kosiol and Anisimova 2012, Benjamini and Hochberg 1995).

Given their impact on the rate of false positives of the positive selection tests (Markova-Raina and Petrov 2011, Fletcher and Yang 2010, Schneider et al. 2009), we took great care at filtering out potential gene predictions and alignments errors. We filtered CDS sequences as described above (see section on codon usage bias). The quality filtering pipeline used for multiple alignments is adapted from the pipeline of the Selectome database (*http://selectome.unil.ch*) (Proux et al. 2008): multiple alignments of protein sequences of gene families were first computed by M-Coffee (Wallace et al. 2006) from the T-Coffee package v8.93 (Notredame et al. 2000), which combines the output of different aligners (mafftgins_msa, muscle_msa, kalign_msa, t_coffee_msa). We kept only amino acid positions where the M-Coffee score was 7 or above, eliminating residues not consistently aligned by different aligners. We then used MaxAlign v1.1

(Gouveia-Oliveira et al. 2007) to remove badly aligned sequences. Finally we used a stringent Gblocks filtering (v0.91b; type = codons; minimum length of a block = 4; no gaps allowed) (Castresana 2000), to remove gap-rich regions from the alignments. The results from this analysis are available for download at the Ant Genomes Portal (*http:hymenopteragenome.org/ant_genomes/*).

To test for functional categories enrichment (Supplemental Table 20) we used the Gene Ontology functional annotation (Ashburner et al. 2000) transferred from the *D. melanogaster* member of each family and extracted from Flybase (*http://flybase.org/static_pages/downloads/FB2011_02/go/gene_association.fb.gz*). We applied a SUMSTAT test (Tintle et al. 2009) and used the LRT value from the positive selection analysis (transformed using the fourth square root to stabilize variance) as score for each gene family. We implemented an algorithm similar to the Elim algorithm of the topGO software (The Gene Ontology Consortium 2000) to decorrelate the graph structure of the Gene Ontology. The false discovery rate was assessed using 100 permutations of scores of gene families.

To check that the dependence of our results to the methodology used, we constructed another dataset including all gene families that could pass CDS quality filters (6,186 families, including families with gene duplications). Sequences were aligned using PRANK (v100701), one of the most realistic aligner currently available (Markova-Raina and Petrov 2011, Fletcher and Yang 2010, Löytynoja and Goldman 2008, Löytynoja and Goldman 2005, Jordan and Goldman 2012). We then filtered alignments based on the confidence score attributed by Guidance (v1.1) (Penn et al. 2010, Privman et al. 2012). Gene family phylogenies were built using RAxML (v7.2.9) (Le and Gascuel 2008). Finally, the site test of Codeml (PAML v4.4e) (Yang et al. 2000) was used to test for positive selection (null model M8a vs. alternative model M8) (Swanson et al. 2003, Wong et al. 2004). The functional categories enriched in positively selected genes in ants identified in this dataset are similar to the ones reported in Supplemental Table 20, supporting that our results are likely not artifactual.

CG compositional analysis

Genomic sequences were partitioned into domains using IsoPlotter (*http://code.google.com/p/isoplotter/*), which employs an algorithm that recursively segments chromosomes by maximizing the difference in CG-content between adjacent subsequences. The process of segmentation terminates when the difference in CG-content between two neighboring domains is no longer statistically significant.

Supplemental Text

Supplemental Text 1. Analysis of 64 TRGs found in all seven ant genomes.

Among the 28,581 TRGs specific to Formicidae, we identified a subset of 64 genes that display no protein sequence similarity to genes outside Formicidae but which have at least significant local similarity among all ants (BLASTP, E < 1e–3). Of these 64 orthologous gene clusters, 62 are strict single-copy gene clusters, i.e., they contain one protein sequence per species. The remaining two clusters are single-copy in six ant species but contain a duplication in one species. We could not detect any Pfam-A domains in these genes, indicating that these genes do not contain any known functional units, despite their broad conservation. For further classification of these 64 ant-specific TRGs, we aligned the 62 strict single-copy gene clusters using MUSCLE. To evaluate alignment conservation, we applied Gblocks (Castresana 2000) with default settings to identify only conserved sequence blocks (minimum length of 10 residues) from the protein multiple sequence alignments. Despite the relaxed homology criterion used to identify the ant-specific TRGs, the vast majority of clusters display substantial sequence conservation. 57 of the 64 gene clusters (89%) contain conserved blocks with a summed length of at least 50 residues, and in 52 of the 64 clusters (81%) at least 50% of all alignment positions are conserved (Fig. 2C). These results suggest that these ant-specific TRGs are present throughout Formicidae and contain highly conserved functional regions.

Supplemental Text 2. Codon usage bias.

The genetic code is redundant with multiple codons encoding the same amino acids. Codon usage bias reflects the fact that not all synonymous codons are used with equal frequencies, often with sharp preferences for some codons compared to others. This phenomenon is present in most organisms ranging from bacteria to animals (Hershberg and Petrov 2008, Duret 2002, Plotkin and Kudla 2011). Codon usage bias is thought to result from a balance between two major forces: selection for translational optimization and mutational biases (Duret 2002, Bulmer 1991, Drummond and Wilke 2008). Analysis of the 12 *Drosophila* species highlighted that selective forces were mainly responsible for codon usage bias in these genomes (Stark et al. 2007). Interestingly, variations in patterns of codon usage bias among these species reflect the variations of strength of translational selection across their phylogeny (Drosophila 12 Genomes Consortium 2007, Vicario et al. 2007, Heger and Ponting 2007). For example translational selection strength was shown to be highest for the species of the *D. melanogaster* group – although a slight genomic reduction in codon bias is observed for *D. melanogaster*. Another interesting example is a striking lineage-specific shift in codon preferences seen *D. willistoni*, which cannot be sufficiently explained by mutation alone, and may have involved directional selection (Vicario et al. 2007, Heger and Ponting 2007). Similarly to the *Drosophila* lineage, it is expected that the study of codon usage bias in the 7 ant species and their outgroups can give us valuable insights into the evolutionary history of these lineages.

To compare the levels of codon usage bias among different species, we measured the “effective number of codons” used in CDS sequences (ENC or N_C; Supplemental Fig. 9A) (Drosophila 12 Genomes Consortium 2007, Vicario et al. 2007, Heger and Ponting 2007). Though this is the measure of choice to implement multi-species comparisons

(Gingold et al. 2011), unfortunately it does not differentiate if codon usage bias results from selective forces or mutational bias. So we used three other complementary measures: First, we analyzed the GC content in CDS sequences (%GC; Supplemental Fig. 9C) and the G+C content at 3rd synonymous positions (%GC3s; Supplemental Fig. 9B), which reflect the overall mutational biases experienced by the genomes in their evolutionary history. Second, to better characterize the role of selective forces, we isolated codon usage bias levels of ribosomal genes (Supplemental Fig. 9A; red bars). Ribosomal genes are expected to be under strong selection for optimal codon usage because they are highly and constitutively expressed in most cells of the organism (Heger and Ponting 2007). A reduction of the levels of codon usage bias of these genes is likely to reflect a genome-wide relaxation of selection. Third, because both selective forces and mutational biases may be responsible for codon usage bias in a genome, and to know if nucleotide composition biases are sufficient to explain the observed patterns of codon bias, we created a randomized dataset by randomizing the codon usage in the sequences of the whole dataset, controlling for GC content of codons (see methods below). The ENC levels of genes in this dataset reflect the expectation in the absence of selective forces (Supplemental Fig. 9A; blue bars).

The 12 analyzed species can be gathered in three groups with similar patterns, the first "group" being *D. melanogaster* alone. This species displays a relatively high codon bias and it is well established that this pattern is essentially due to selective pressure acting on synonymous sites (Drosophila 12 Genomes Consortium 2007, Vicario et al. 2007, Heger and Ponting 2007, Duret 2002, Akashi 1994, Powell and Moriyama 1997). The observation of high levels of %GC and %GC3s in CDS sequences confirms this hypothesis, (i) because almost all optimal codons – corresponding to the most abundant tRNAs – are ending by cytosines or guanines (Duret 2002, Shields et al. 1988), (ii) because mutational events in *D. melanogaster* are biased toward A+T (Petrov and Hartl 1999). Selection on codon usage is also reflected by the very strong level of codon usage bias seen in ribosomal genes which are clearly skewed in the distribution of ENC values for protein coding genes. Consistently too, the randomized sequences display a lower codon usage bias.

Second, the seven ant species, *T. castaneum* and *N. vitripennis* display low levels of codon bias, consistent with a relaxation of selective pressure on synonymous sites on their genome. The nucleotide composition of these genomes is relatively balanced: the observed %GC3s is between 0.37 (*C. floridanus*) and 0.53 (*T. castaneum*); %GC is between 0.42 (*C. floridanus*) and 0.47 (*T. castaneum*). These values show that mutational forces were probably insufficient to change the composition of genomes and bias strongly codon usage. The relaxed levels of selection in these species is confirmed by the low level of codon usage bias observed in ribosomal genes sequences, as well as by the relatively low shift between real and randomized datasets.

Finally, strikingly high levels of codon bias are seen in *A. mellifera* and *P. humanus* genes. This is most probably due to strong mutational biases, which are reflected in very low %GC (0.34 and 0.36 respectively) and %GC3s (0.15 and 0.23 respectively) in these genomes. As shown for *A. mellifera* (Jorgensen et al. 2007), synonymous sites of genes tend to adopt the GC content of the region in which they reside and thus reflect the biased nucleotide composition of these genomes. The median levels of codon bias of ribosomal genes are very close to those of protein coding genes, showing that the selective pressure on codon usage is drastically reduced, even on these genes which are usually under strong selection for optimized translation. Finally, the randomized sequences display similar ENC values as the real dataset, indicating that selective forces are not required to explain the codon bias patterns observed in these species.

Overall, this analysis provides evidence for strong selection on codon usage only in *D. melanogaster*. With the exception of *A. mellifera* and *P. humanus* genomes where codon usage is biased by a very extreme G+C content composition of the genome, all other genomes display relatively low levels of codon usage bias. For the 7 ant species in particular, the global reduction of codon usage bias most likely reflects a relaxation of purifying selection acting on these genomes. Interestingly, such a relaxation was previously predicted in relation to the reduction of effective population size (N_e) associated with social life (Bromham and Leys 2005), but no solid evidence of this phenomenon was found so far in eusocial organisms. We should also note that the genomic patterns seen in ants and in another social insect, *A. mellifera*, are drastically different. The reasons for this remain to be examined.

Supplemental Text 3. GC compositional analysis

Animal genomes are not uniform in their long-range sequence composition but are composed of a mosaic of sequence stretches of variable lengths that differ widely in their guanine and cytosine (GC) compositions. These sequences are referred to as compositional domains and are defined here as are genomic DNA segments that have a characteristic GC-content that differs significantly from the GC-content of adjacent compositional domains. Compositional domains can be divided into compositionally homogeneous and compositionally non-homogeneous domains, if their internal homogeneity is lower or higher than that of the chromosome on which they reside, respectively. In classical terminology, compositionally homogeneous domains that are larger than 300 kb are referred to as isochores (Bernardi 2000).

In all animal genomes studied so far, we found that the distribution of compositional-domain lengths showed an abundance of short domains and a paucity of long ones (Weinstock et al. 2006, Bernardi 2000, Richards et al. 2008, Sea Urchin Genome Sequencing Consortium 2006, Bovine Genome Sequencing and Analysis Consortium 2009). The three ant genomes we previously studied are not exceptions in this respect (CD Smith et al. 2011, CR Smith et al. 2011, Suen et al. 2011). Here we performed a comparative analysis of seven ant genomes along with two hymenopteran (*A. mellifera* and *N. vitripennis*) and three non-hymenopteran outgroups (*D. melanogaster*, *A. gambiae*, *T. castaneum*) to provide further insight into the evolution of GC compositional domain architecture.

We performed three analyses, described below. In the first analysis, we calculated the distribution of homogeneous domain lengths. For convenience, domains were divided according to the order of magnitude of their length into: short (10^3–10^4 bp), medium (10^4–10^5 bp), and long (10^5–10^7 bp). Based on the observed goodness-of-fit, we calculated a *p*-value that quantifies the probability that the data were drawn from the hypothesized distribution. In the second analysis, we compared the dispersal of domain GC-contents. In the last analysis, we compared the domain GC-content versus their sizes in a log scale.

Analysis of Compositional-Domain Sizes

The total number of compositional domains per genome varied from about 35,000 in *C. floridanus* to approximately 66,000 in *H. saltator* (Supplemental Table 10). The coefficient of variation for ant genomes is about 28%. We divided the compositional domains into four size classes: 1–10 kb, 10–100 kb, 100 kb to 1 Mb, and 1–10 Mb. Using a G-test goodness-of-fit test, we determined that none of the distributions of domain sizes is similar to any other ($p = 0.03$).

A comparison of the distributions of compositional domain lengths among ants, bee, wasp, beetle, mosquito, and fly showed that bee, wasp, and *H. saltator* have the smallest fraction (0.1–0.3%) of long domains (>100 kb). Long domains are abundant in the ant linage, with the leaf-cutters *A. cephalotes* and *A. echinatior* having the largest domains among all fully sequenced insect genomes (Supplemental Table 10, Supplemental Fig. 17).

Unlike vertebrate genomes, whose GC-content varies from 40% to 45%, ant genomes exhibit variable GC-content, with average GC-content ranging from 32.6% (*A. cephalotes*) to 45.2% (*H. saltator*) and a GC-content standard deviation of 8–10% (Supplemental Fig. 3). The distribution of GC-content within compositional domains varies greatly: in the bee, beetle, and most of the ant genomes, it is right-skewed due to a high frequency of GC-poor domains, in the wasp genome it is bimodal (Supplemental Fig. 18). The *H. saltator* genome is different than the other ant genomes, in that it is also bimodal. We used the Kolmogorov-Smirnov test (Sokal and Rohlf 1995) to determine that none of the compositional domain GC-content distributions is similar to any other ($p<0.01$).

The range of GC-content in hymenopteran domains was the widest among all invertebrates in the analysis, ranging from 1% to 75%, with *C. floridanus* domains setting the lower limit and *A. mellifera* domains setting the upper limit (Supplemental Fig. 18). Interestingly, the decrease in mean genomic GC-content in the ant genome is proportional to the increase in the number of large domains (>100 kb). This is not surprising, as the elimination of GC-rich domains increases the homogeneity of the genome indicated by longer homogeneous domains.

Analysis of Genome Architecture

Comparing the GC-content of compositional domains with their length distributions provides a general view of the invertebrate genomic architecture. Long GC-poor domains are rare among hymenopterans particularly in bee and wasp, compared to the beetle and the two dipterans. Although all genomes in the analysis have similar numbers of long domains (72 to 401) and isochoric domains (44 to 224), their GC-composition varies greatly (Supplemental Fig. 18, Supplemental Table 10). Nearly all long domains in beetle, mosquito, and fly have a GC-content that is within ±5% of their genomic mean GC-content, whereas in bee and wasp about half of the domains have GC-content above the 5% boundary. In ants, there is a trend of GC enrichment for long domains beginning with *H. saltator* and ending with *A. cephalotes*.

Distribution of Genes in Compositional Domains

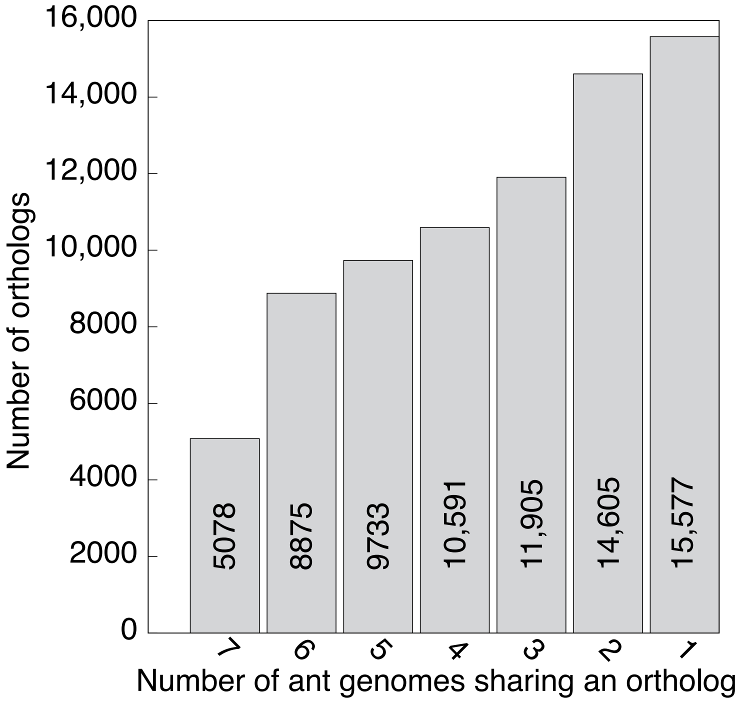
We observed previously that genes in *A. mellifera* have a strong bias toward occurring in the more GC-poor regions of the genome (Weinstock et al. 2006). In contrast, the genomes of all other species studied prior to the availability of an ant genome assembly (including human, fruit fly, worm, mosquito, yeast, body louse and sea urchin) showed either little bias with respect to GC content, or a slight bias toward occurring in more GC-rich regions of the genome (Weinstock et al. 2006, Bernardi 2000, Richards et al. 2008, Sea Urchin Genome Sequencing Consortium 2006, Bovine Genome Sequencing and Analysis Consortium 2009, Werren et al. 2010). We later found that genes in two ant genomes showed no bias (*A. cephalotes*) or a very slight bias (*L. humile* and *P. barbatus*) toward low GC regions (CD Smith et al. 2011, CR Smith et al. 2011, Suen et al. 2011). We therefore were interested in whether all of the currently available ant genomes were similar in this respect.

The relative percent GC of the GC compositional domains containing genes in each species recapitulate the relative percent GC of the overall genome of each species (Supplemental Fig. 17). For example, the cumulative distribution of GC content in compositional domains containing genes for *A. mellifera* lies to the left of that of *L. humile*, which in turn lies to the left of that of *N. vitripennis*. This is consistent with the fact that *A. mellifera* is more GC-poor than *L. humile*, which in turn is more GC-poor than *N. vitripennis*.

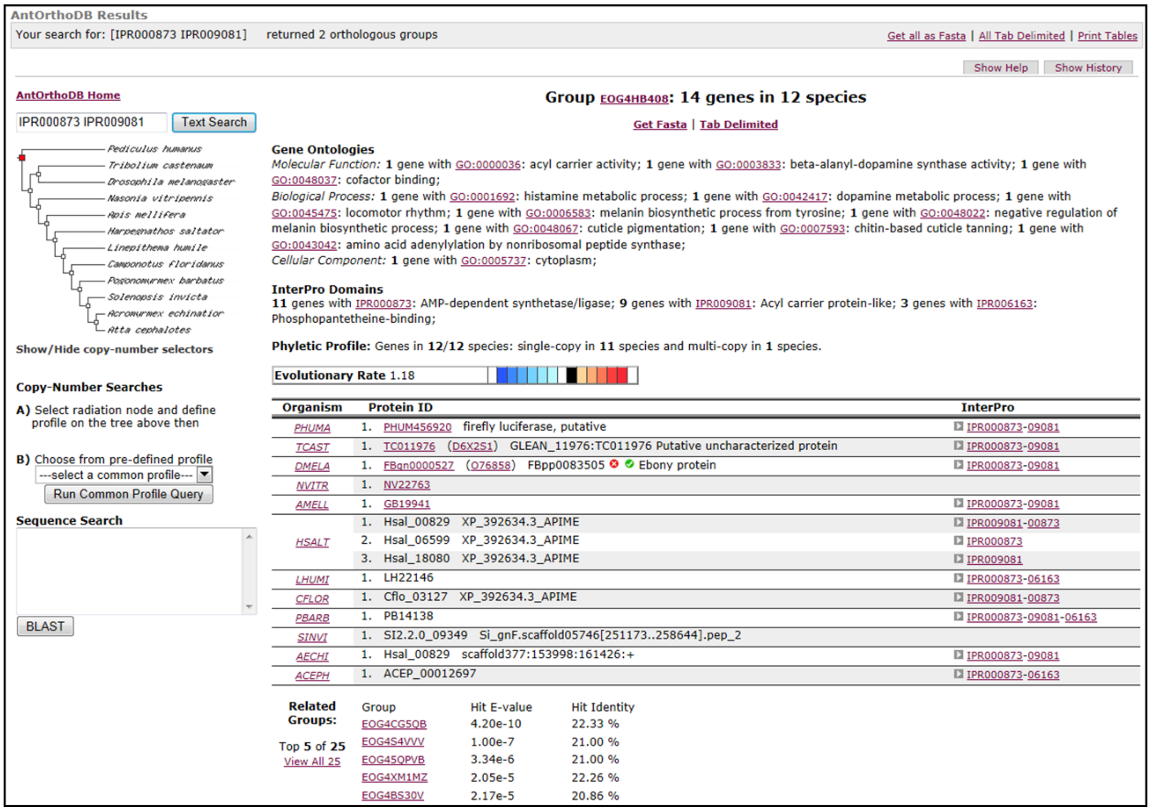
To assess whether genes in these species are biased toward occurring in GC compositional domains of high or low GC content, for each genome we overlaid cumulative distributions of the percent genome that is comprised of compositional domains below a given percent GC (thin lines) onto a similar distribution for only compositional domains that contain genes (thick lines; the same lines shown in Supplemental Fig. 19; Supplemental Fig. 20). Among the genomes studied, genes in the *A. mellifera* and *H. saltator* genomes show the strongest tendency to occur in more GC-poor regions of the genome (Supplemental Fig. 20). For example, about half of genes in *H. saltator* occur in compositional domains whose GC content is less than 40% (thick blue line, x = 40%, y = 0.5), but compositional whose GC content is less than 40% represents only about 25% of the genome (thin blue line, x = 40%, y = 0.20). Further, the cumulative distribution for the GC content of compositional domains containing genes lies to the left of the cumulative distribution for the GC content of all compositional domains (compare thick and thin lines for *A. mellifera* and *H. saltator*). Genes in the *C. floridanus* and *S. invicta* genomes, similar to the previously studied *P. barbatus*, *L. humile* and *N. vitripennis* genomes, showed a slight tendency to occur in GC-poor regions of the genome. (CD Smith et al. 2011, CR Smith et al. 2011, Werren et al. 2010). Genes in *A. echinatior* showed a very slight bias to GC rich regions, while, as previously reported, *A. cephalotes* did not show any bias toward lower percent GC regions (Suen et al. 2011).

Supplemental Text 4. Background on the salivary gland and wing development regulatory networks.

Glands derived from the ectodermal cell layer, including the mandibular, salivary (labial), and metapleural glands are essential for intracolonial communication and the interaction of individuals and their environment (Wurm et al. 2011). Phenotypic plasticity of these glands between specialized worker castes has been reported in some ants, predicting the acquisition of novel regulatory mechanisms (Pavon and Camargo-Mathias 2005, Niculita et al. 2008, Amaral and Machado-Santelli 2008). The underlying genetic regulation of the specification, differentiation, and morphogenesis of these integumentary glands has best been studied in the salivary gland in *Drosophila* (Abrams et al. 2003). The complex interaction of transcriptional activators and repressors specify the pre-ductal cells, activate lineage-specific ductal and secretory morphogenic cassettes and remodeling glands during metamorphosis.



Supplemental Fig. 1. Variation in orthology among ant genomes. The number of orthologous genes shared among different ant genomes is shown as a function of the number of genomes in consideration (i.e., 7 denotes all seven ant genomes considered). Orthology assessed using OrthoDB.



Your search for: [IPR000873 IPR009081]   returned 2 orthologous groups

IPR000873 IPR009081

Copy-Number Searches

A) Select radiation node and define profile on the tree above then

B) Choose from pre-defined profile

Sequence Search

Group EOG4HB408: 14 genes in 12 species

Gene Ontologies
Molecular Function: **1** gene with GO:0000036: acyl carrier activity; **1** gene with GO:0003833: beta-alanyl-dopamine synthase activity; **1** gene with GO:0048037: cofactor binding;
Biological Process: **1** gene with GO:0001692: histamine metabolic process; **1** gene with GO:0042417: dopamine metabolic process; **1** gene with GO:0045475: locomotor rhythm; **1** gene with GO:0006583: melanin biosynthetic process from tyrosine; **1** gene with GO:0048022: negative regulation of melanin biosynthetic process; **1** gene with GO:0048067: cuticle pigmentation; **1** gene with GO:0007593: chitin-based cuticle tanning; **1** gene with GO:0043042: amino acid adenylylation by nonribosomal peptide synthase;
Cellular Component: **1** gene with GO:0005737: cytoplasm;

InterPro Domains
11 genes with IPR000873: AMP-dependent synthetase/ligase; **9** genes with IPR009081: Acyl carrier protein-like; **3** genes with IPR006163: Phosphopantetheine-binding;

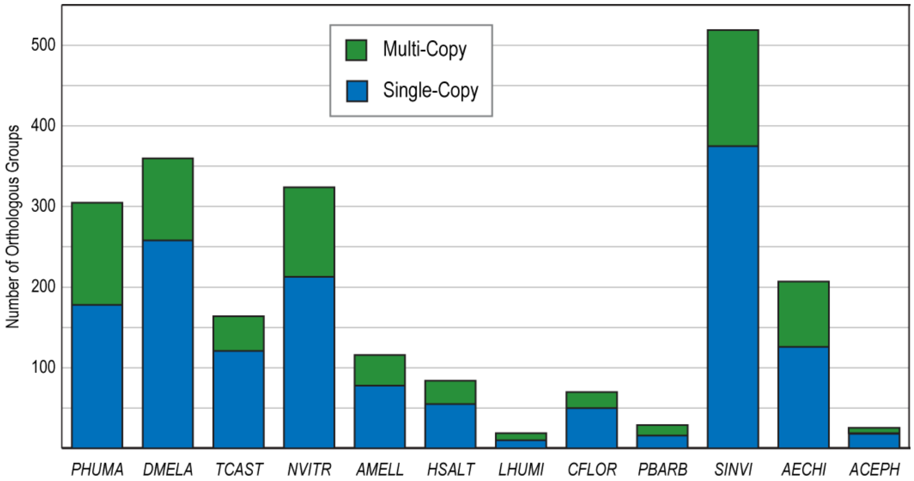
Phyletic Profile: Genes in **12/12** species: single-copy in **11** species and multi-copy in **1** species.

Evolutionary Rate 1.18

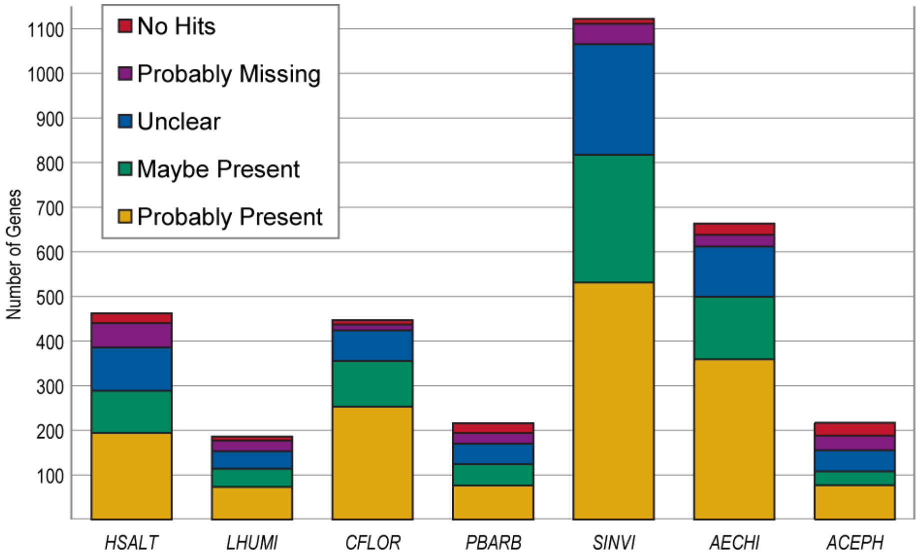
Organism		Protein ID	InterPro
PHUMA	1.	PHUM456920 firefly luciferase, putative	IPR000873-09081
TCAST	1.	TC011976 (D6X2S1) GLEAN_11976:TC011976 Putative uncharacterized protein	IPR000873-09081
DMELA	1.	FBgn0000527 (O76858) FBpp0083505 ✕ ✓ Ebony protein	IPR000873-09081
NVITR	1.	NV22763	
AMELL	1.	GB19941	IPR000873-09081
HSALT	1.	Hsal_00829 XP_392634.3_APIME	IPR009081-00873
	2.	Hsal_06599 XP_392634.3_APIME	IPR000873
	3.	Hsal_18080 XP_392634.3_APIME	IPR009081
LHUMI	1.	LH22146	IPR000873-06163
CFLOR	1.	Cflo_03127 XP_392634.3_APIME	IPR009081-00873
PBARB	1.	PB14138	IPR000873-09081-06163
SINVI	1.	SI2.2.0_09349 Si_gnF.scaffold05746[251173..258644].pep_2	
AECHI	1.	Hsal_00829 scaffold377:153998:161426:+	IPR000873-09081
ACEPH	1.	ACEP_00012697	IPR000873-06163

Related Groups:	Group	Hit E-value	Hit Identity
	EOG4CG5QB	4.20e-10	22.33 %
Top 5 of 25	EOG4S4VVV	1.00e-7	21.00 %
View All 25	EOG45QPVB	3.34e-6	21.00 %
	EOG4XM1MZ	2.05e-5	22.26 %
	EOG4BS30V	2.17e-5	20.86 %

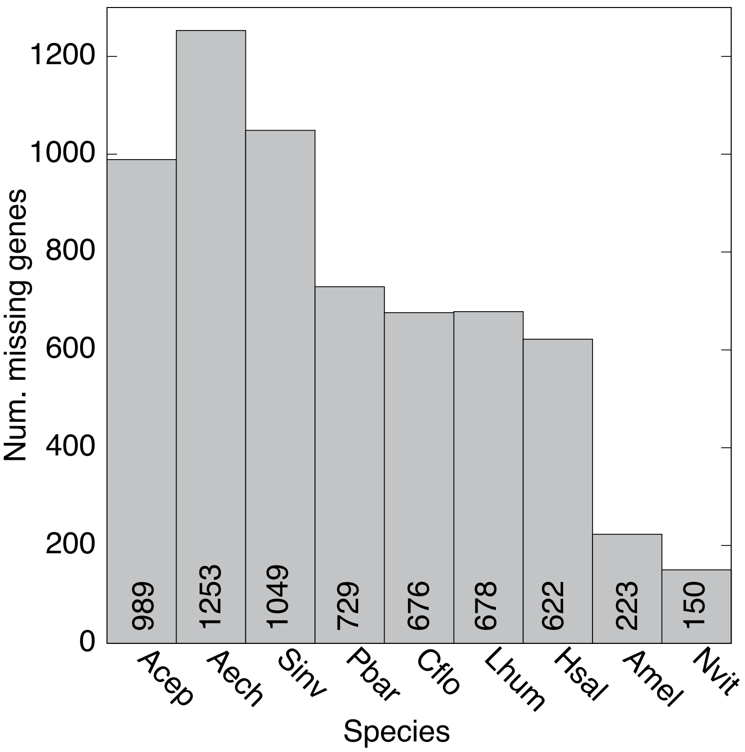
Supplemental Fig. 2. A screenshot from AntOrthoDB (*http://cegg.unige.ch/orthodbants*) shows an example orthologous group with protein descriptors, Gene Ontology and InterPro attributes, phyletic profile, evolutionary rate, and related groups.



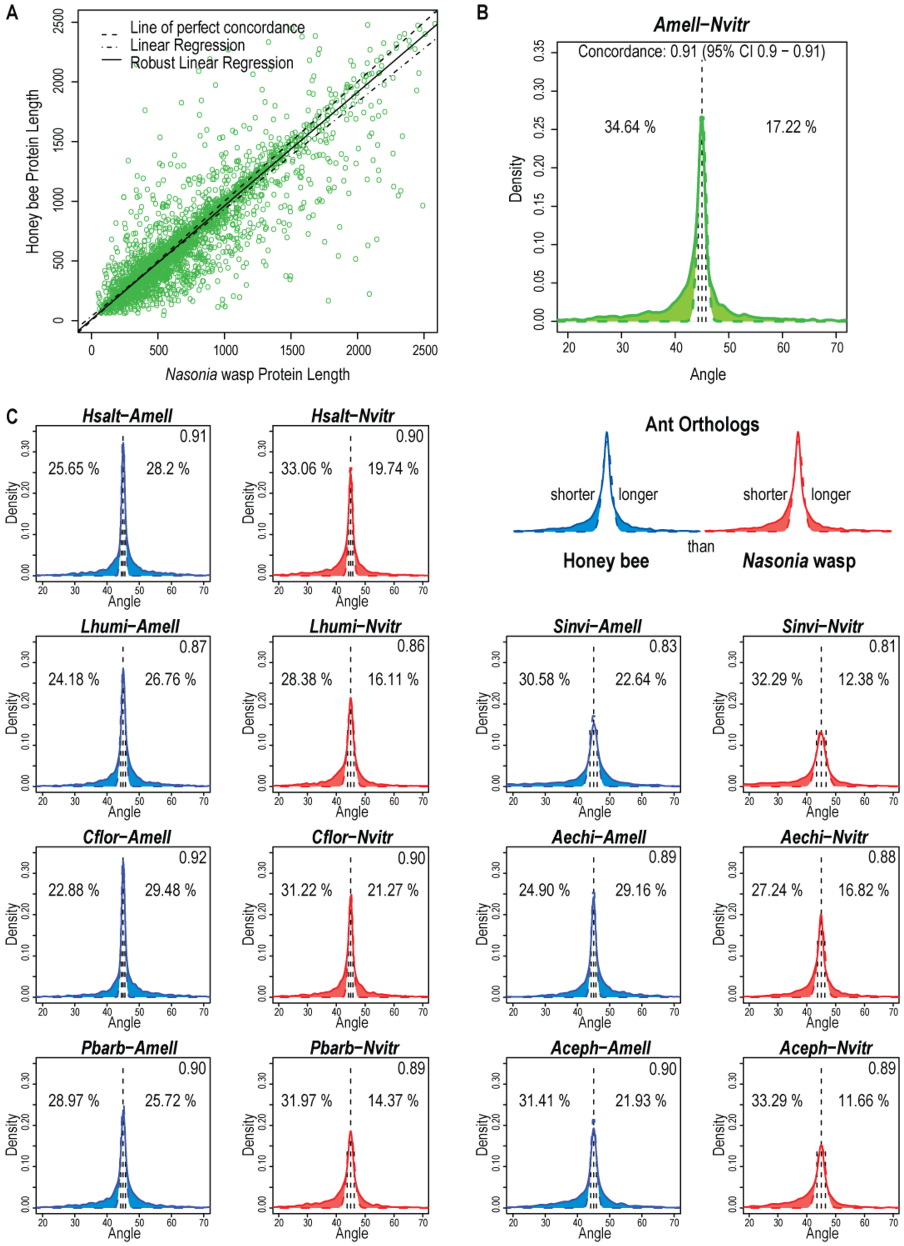
Supplemental Fig. 3. Identification of lost or missing genes by analysis of existing gene annotations. Orthologous groups were identified with genes from 11 out of the 12 species which contain (i) strictly one gene in each species (single-copy) and (ii) at least one species with more than one gene (multi-copy). See Supplemental Table 2 for species abbreviations.



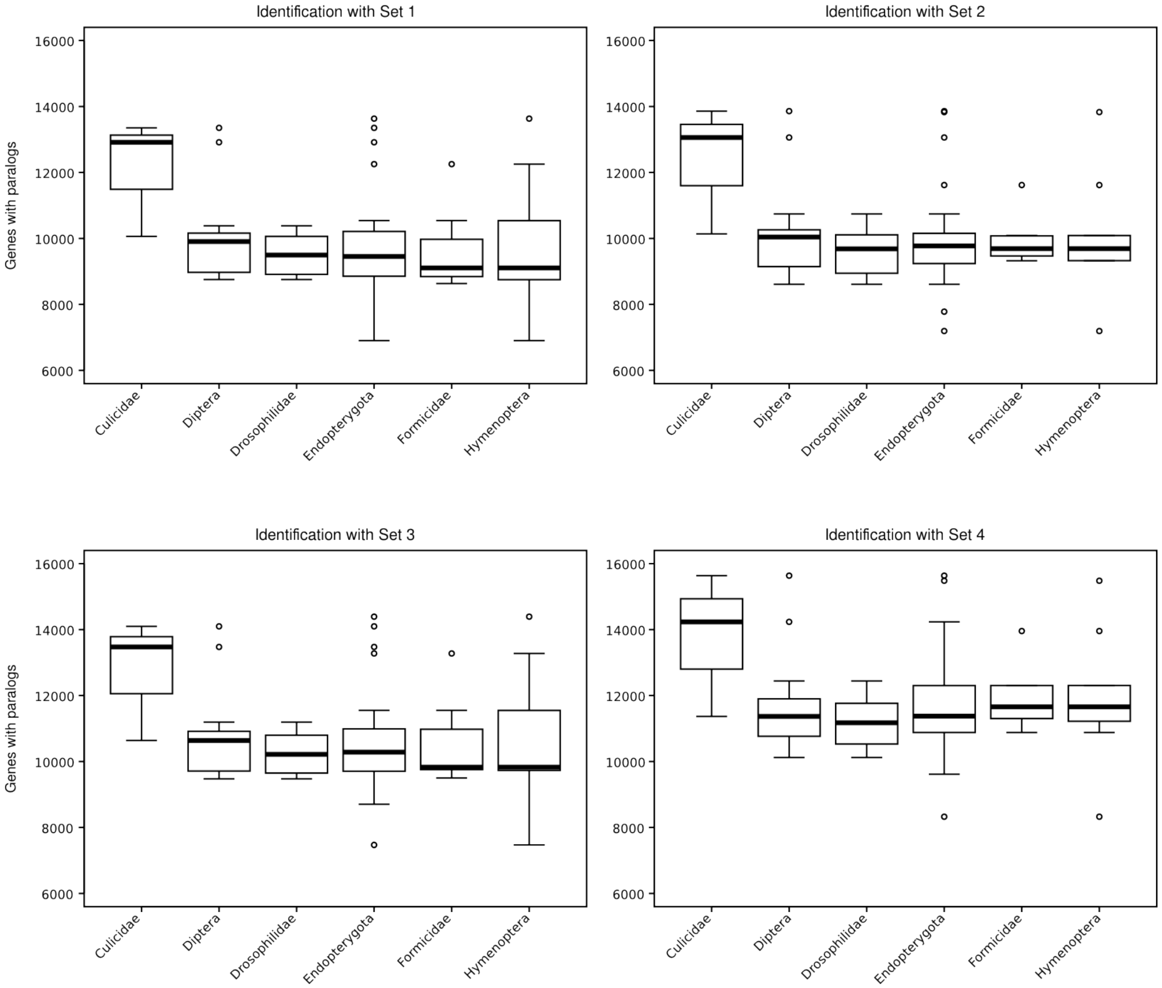
Supplemental Fig. 4. Assembled ant genomes were searched for potentially missing or missed genes (see Supplemental Table 2 for species abbreviations). No Hits: the seed gene had a significant BLAST hit to the ‘outgroup’ genome but none to the ‘missing’ genome, i.e. these orthologs are missing from the genome assemblies. Probably Missing: the seed gene BLAST hit was more significant to the ‘outgroup’ genome than to the ‘missing’ genome, i.e. the ‘missing’ genome hit may correspond to a homolog rather than an ortholog. Unclear: the differences between the seed gene BLAST hits to the ‘missing’ genomes and to the ‘outgroup’ genomes did not allow for clear distinctions to be made. Maybe Present: the seed gene BLAST hit was ‘better’ to the ‘missing’ genome than to the ‘outgroup’ genome, i.e. the ‘missing’ genome hit may correspond to the ortholog and hence these genes may be present. Probably Present: the seed gene BLAST hit was ‘better’ to the ‘missing’ genome than to the ‘outgroup’ genome, i.e. the ‘missing’ genome hit probably corresponds to the ortholog and hence these genes are probably present.



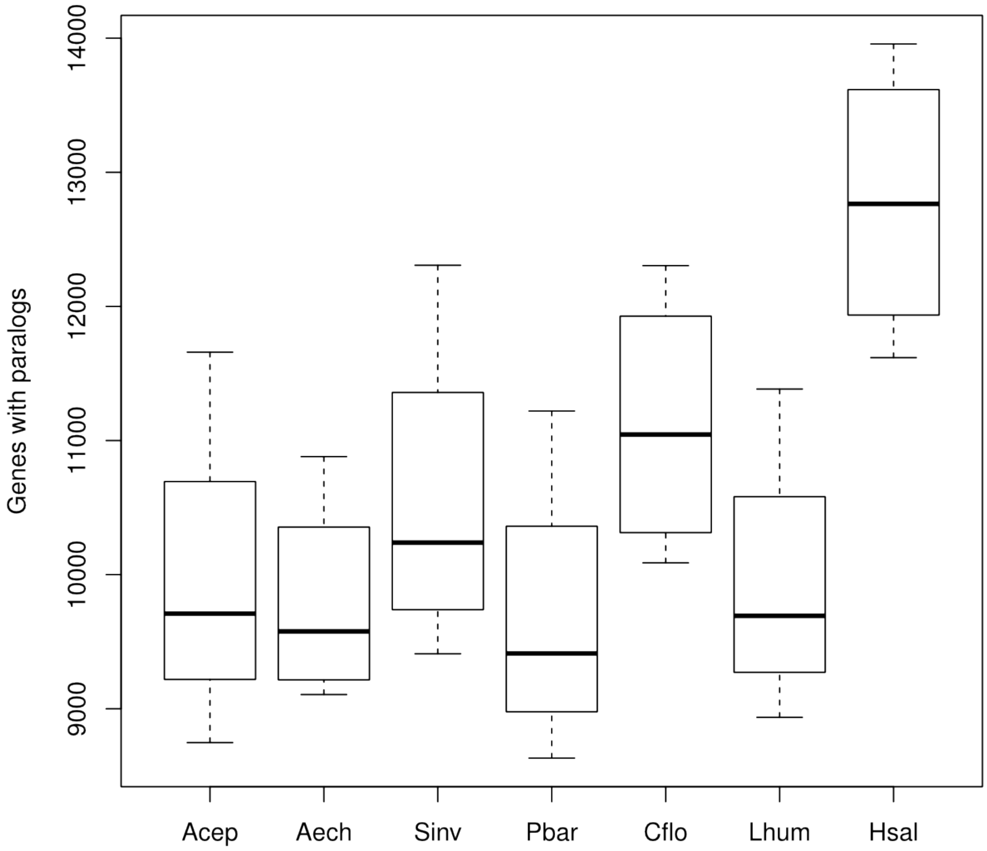
Supplemental Fig. 5. Newly annotated genes that were previously considered species-specific but are present in multiple Hymenoptera genomes. The number of newly annotated genes is shown for each species. See Materials and Methods (Identification of taxonomically restricted genes) for details on the identification procedure that included thirty published arthropod genomes.



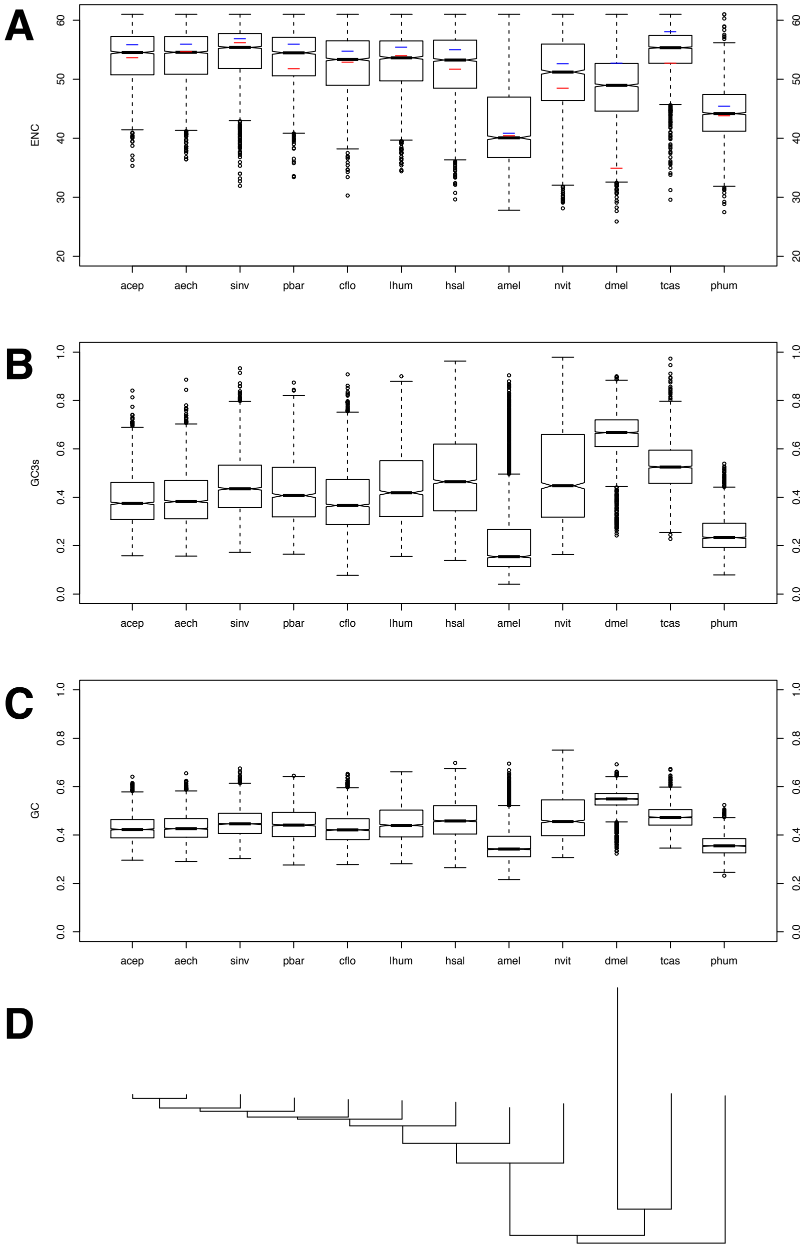
Supplemental Fig. 6. Concordance analysis of protein lengths between *A. mellifera* (blue) or *N. vitripennis* (red) genes and their ant orthologs. The bee-wasp comparison shows the distribution of compared lengths (**A**) with both regressions showing a tendency for bee proteins to be shorter than their wasp orthologs. Plotting the density of data points falling at each degree below and above 45 degrees (**B**) shows the distributions of the deviations from perfect agreement. Comparing to normal fittings of the data (dotted curves), with means fixed at 45 degrees, highlights proportions of significantly shorter bee proteins (**left**) and significantly longer bee proteins (**right**), given the underlying data. Each of the seven ant species is compared to bee and wasp in the same way (**C**). See Supplemental Table 2 for species abbreviations and Supplemental Table 3 for statistics.



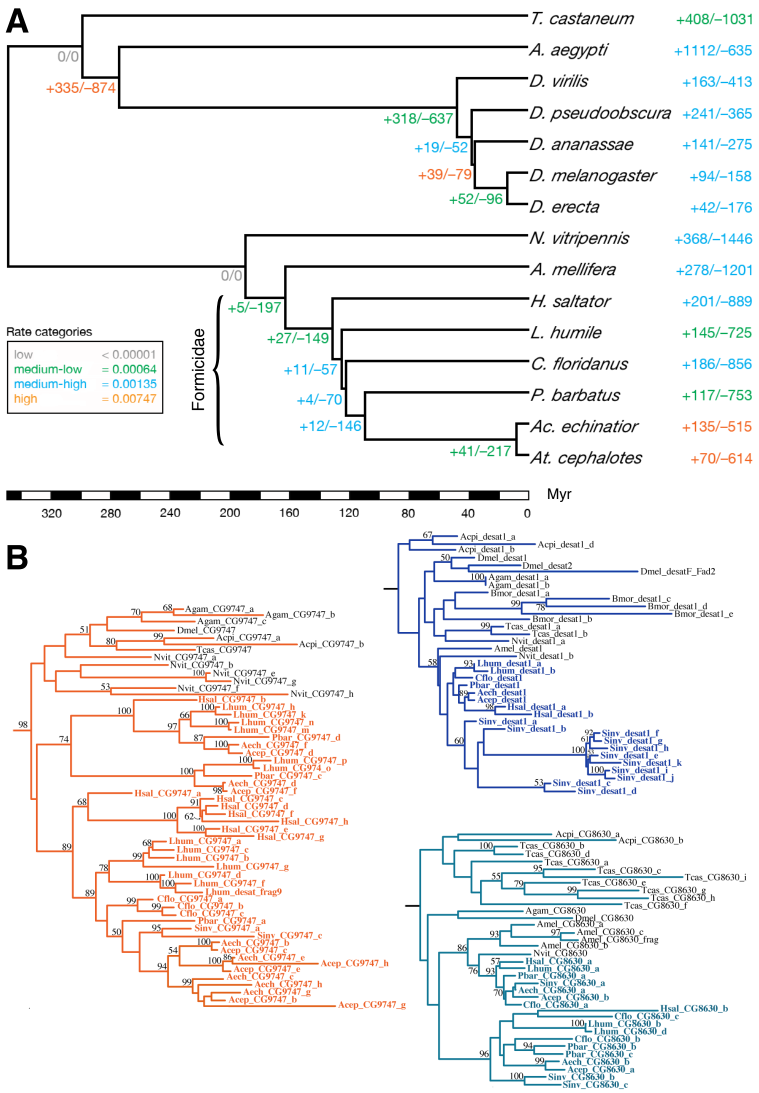
Supplemental Fig. 7. Genes with paralogs (GWP) counts across genomes of several partially overlapping groups of insects.



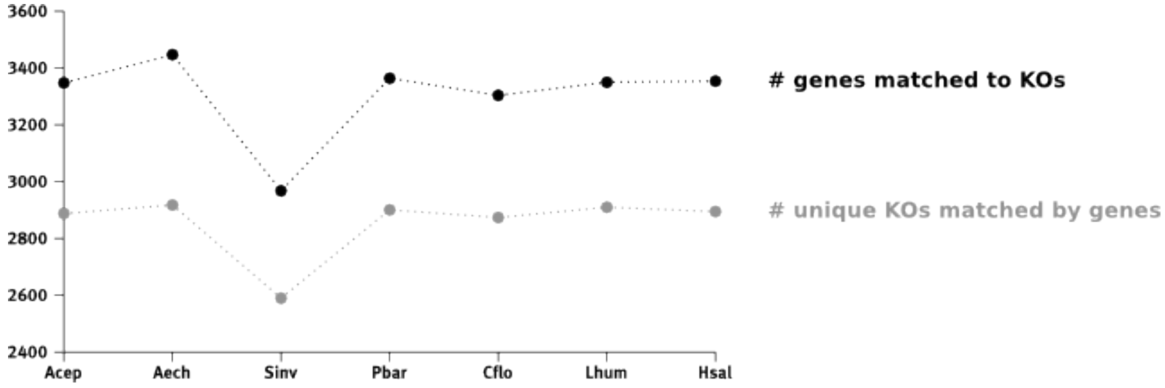
Supplemental Fig. 8. Genes with paralogs (GWP) counts among the seven ant genomes. The four sets of paralog definition were used as replicates per species.



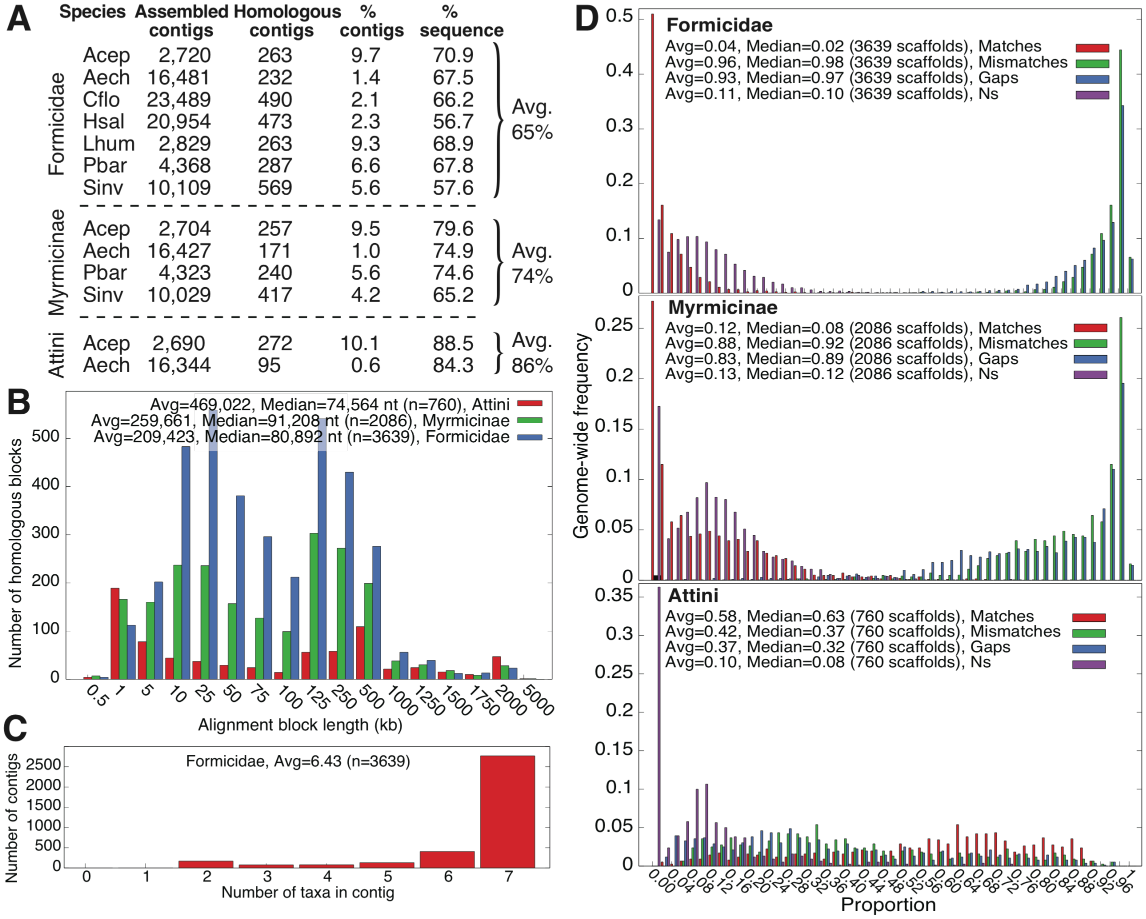
Supplemental Fig. 9. (**A**) Box plot of the distributions of ENC values of CDS sequences of the 12 genomes analyzed. ENC values range between 20 (strong codon usage bias) and 61 (no codon bias). Red bars indicate for each species the median level of ENC observed for CDS sequences of ribosomal genes. Blue bars indicate for each species the median level of ENC for the dataset of randomized CDS sequences. (**B**) Box plot of the distributions of the G+C content at 3rd positions of synonymous codons for CDS sequences of the 12 genomes analyzed. (**C**) Box plot of the distributions of the global G+C content of CDS sequences of the 12 genomes analyzed. (**D**) Phylogeny of the 12 species analyzed.



Supplemental Fig. 10. Gene family evolution along the insect phylogeny. (**A**) Number of gene families that have expanded (+) and contracted (–) along the insect tree of life, as inferred in a maximum likelihood framework (Supplemental Methods). Colors denote one of four estimated rates of average gene gain and loss per gene family per million years (Myr). A model assigning each branch to one of four rate categories fitted the data significantly better than any other model, assessed by likelihood ratio tests (Supplemental Table 6). Note, branches leading to "Formicoida" (six ant species excluding *H. saltator*) or Formicinae and Myrmicinae (excluding *H. saltator* and *L. humile*) proved too short to have accumulated significant changes in gene family size. (**B**) Phylogenetic reconstructions of three desaturase gene subfamilies characterized by both ancestral and recent lineage-specific expansions and contractions in ants. Shown are details of maximum likelihood trees inferred from a dataset encompassing all putatively functional Δ9 and Δ11 desaturase genes that could be identified in 14 holometabolous insect species. Ant genes are highlighted by colored labels; genes in other species are shown in black. Numbers denote nodal confidence values obtained from 500 rapid bootstrap replicates.



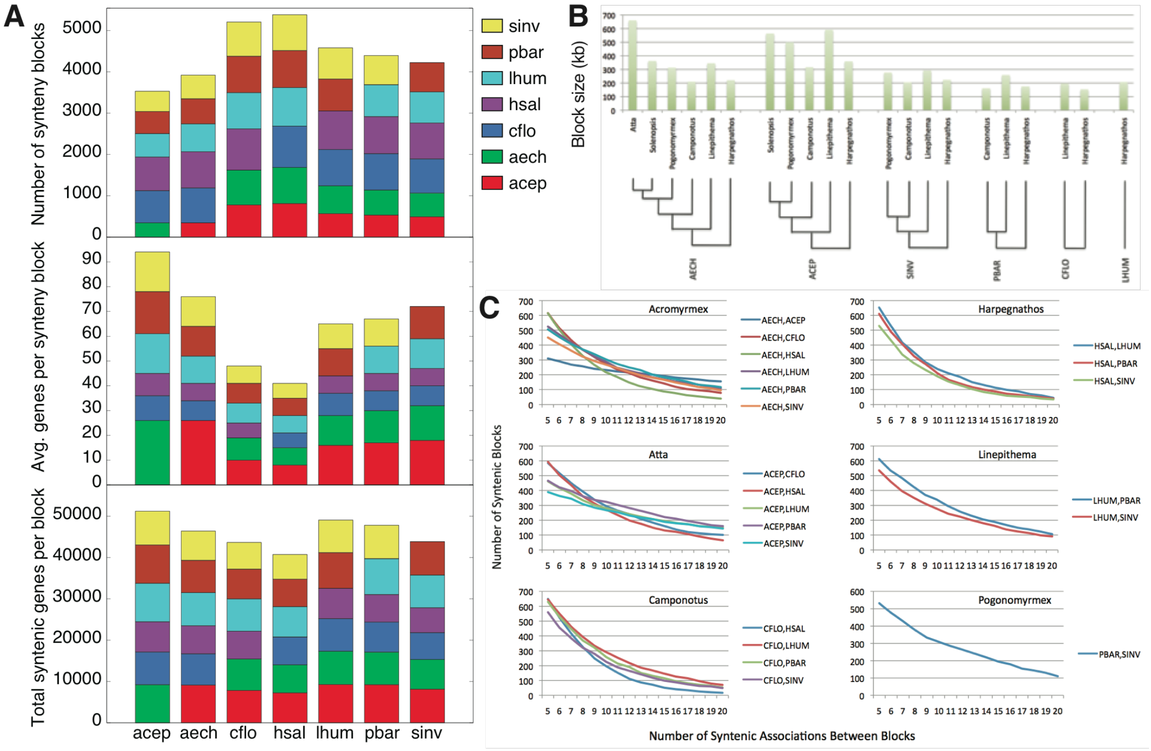
Supplemental Fig. 11. Coverage of KEGG annotation across the seven ant genomes. For each species, the number of annotated genes (black) and KEGG orthology (KO) terms (grey) are given, as multiple genes can map to the same KO term. The coverage is highly similar between all species except *S. invicta* (Sinv), which suggests incomplete genome annotation.



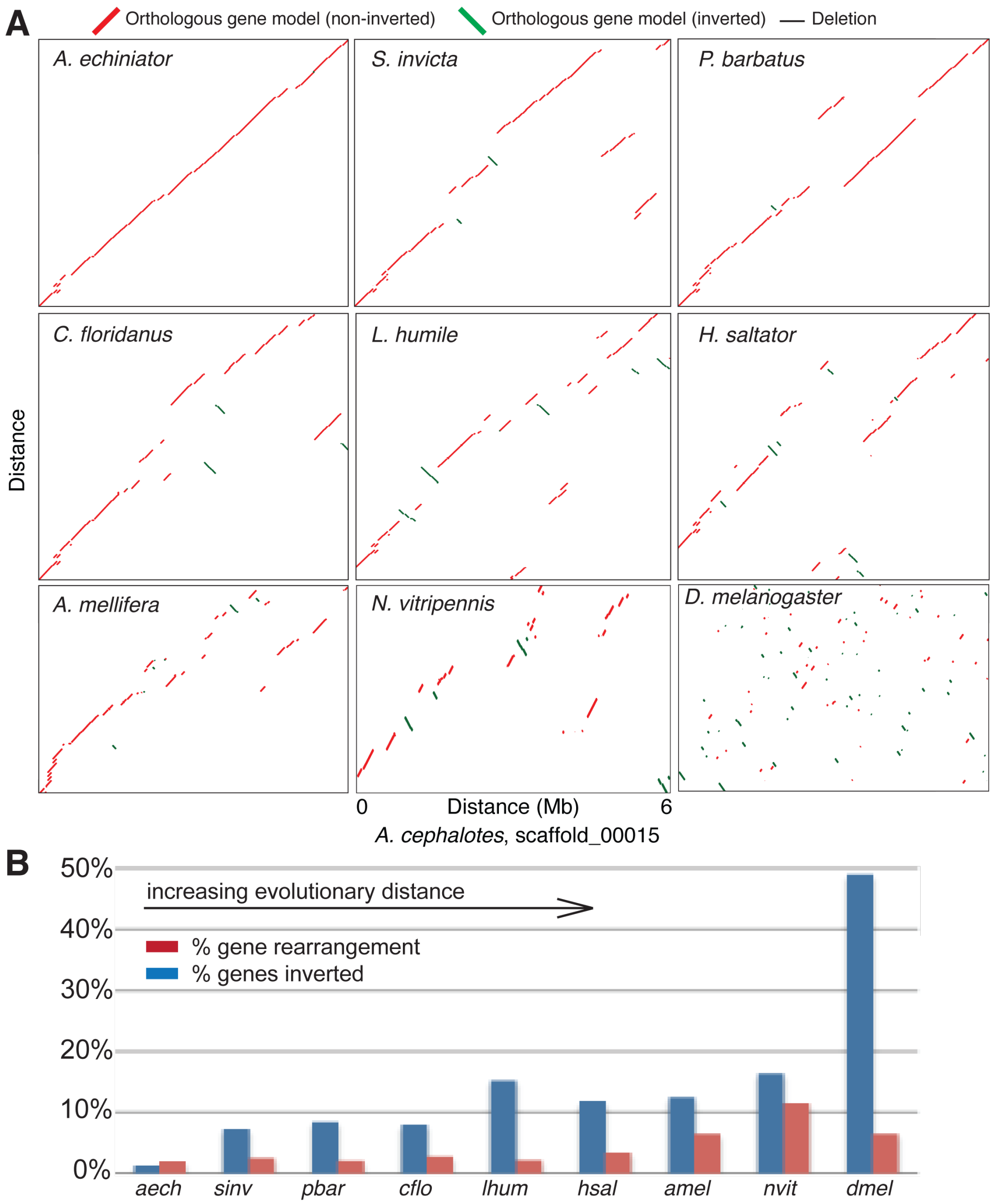
	Species	Assembled contigs	Homologous contigs	% contigs	% sequence	
Formicidae	Acep	2,720	263	9.7	70.9	Avg. 65%
	Aech	16,481	232	1.4	67.5	
	Cflo	23,489	490	2.1	66.2	
	Hsal	20,954	473	2.3	56.7	
	Lhum	2,829	263	9.3	68.9	
	Pbar	4,368	287	6.6	67.8	
	Sinv	10,109	569	5.6	57.6	
Myrmicinae	Acep	2,704	257	9.5	79.6	Avg. 74%
	Aech	16,427	171	1.0	74.9	
	Pbar	4,323	240	5.6	74.6	
	Sinv	10,029	417	4.2	65.2	
Attini	Acep	2,690	272	10.1	88.5	Avg. 86%
	Aech	16,344	95	0.6	84.3	

A
B
Avg=469,022, Median=74,564 nt (n=760), Attini 🟥
Avg=259,661, Median=91,208 nt (n=2086), Myrmicinae 🟩
Avg=209,423, Median=80,892 nt (n=3639), Formicidae 🟦
Number of homologous blocks
0
100
200
300
400
500
0.5
1
5
10
25
50
75
100
125
250
500
1000
1250
1500
1750
2000
5000
Alignment block length (kb)
C
Formicidae, Avg=6.43 (n=3639)
Number of contigs
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6
7
Number of taxa in contig
D
Formicidae
Avg=0.04, Median=0.02 (3639 scaffolds), Matches 🟥
Avg=0.96, Median=0.98 (3639 scaffolds), Mismatches 🟩
Avg=0.93, Median=0.97 (3639 scaffolds), Gaps 🟦
Avg=0.11, Median=0.10 (3639 scaffolds), Ns 🟪
0
0.1
0.2
0.3
0.4
0.5
Myrmicinae
Avg=0.12, Median=0.08 (2086 scaffolds), Matches 🟥
Avg=0.88, Median=0.92 (2086 scaffolds), Mismatches 🟩
Avg=0.83, Median=0.89 (2086 scaffolds), Gaps 🟦
Avg=0.13, Median=0.12 (2086 scaffolds), Ns 🟪
0
0.05
0.1
0.15
0.2
0.25
Genome-wide frequency
Attini
Avg=0.58, Median=0.63 (760 scaffolds), Matches 🟥
Avg=0.42, Median=0.37 (760 scaffolds), Mismatches 🟩
Avg=0.37, Median=0.32 (760 scaffolds), Gaps 🟦
Avg=0.10, Median=0.08 (760 scaffolds), Ns 🟪
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
0.32
0.36
0.40
0.44
0.48
0.52
0.56
0.60
0.64
0.68
0.72
0.76
0.80
0.84
0.88
0.92
0.96
1
Proportion

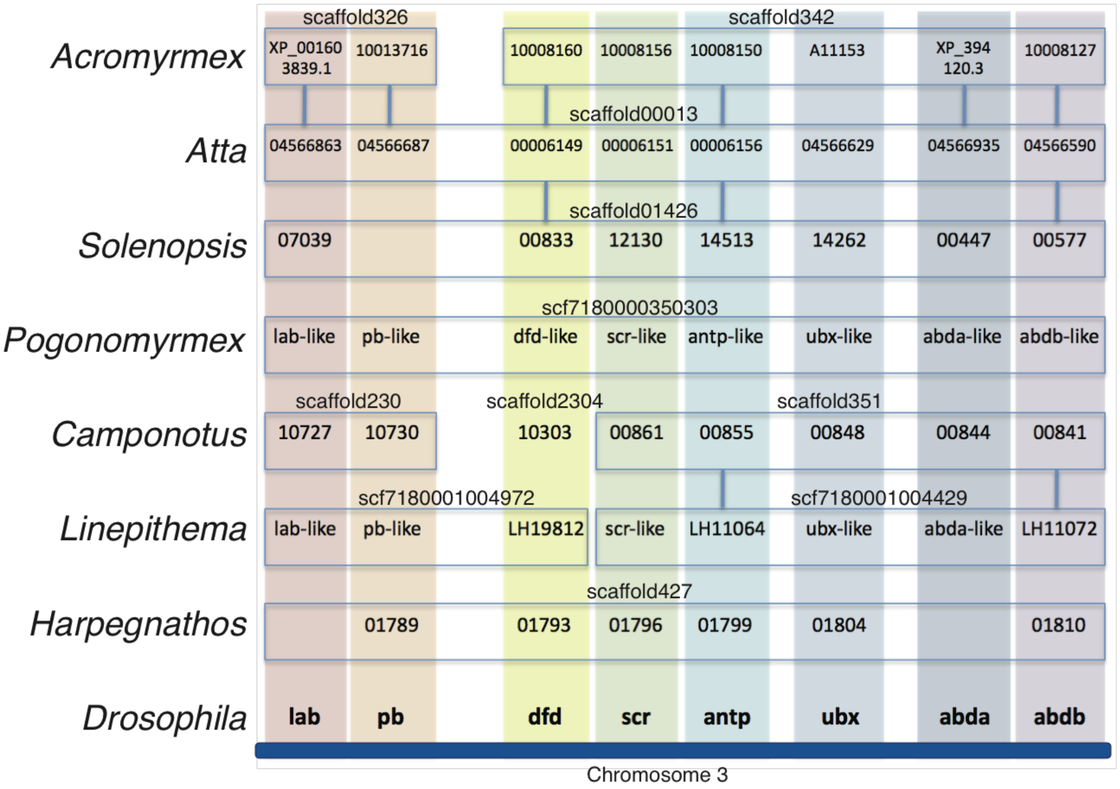
Supplemental Fig. 12. Multiple alignment of seven ant genomes. (**A**) Number of contigs and proportion of each ant genome identified as homologous among Formicidae, Myrmicinae, and Attini. Homologous contigs were determined using Mercator and subsequently aligned using Mavid. The number of homologous contigs identified for each evolutionary grouping are indicated. (**B**) Sequence length distribution of homologous, aligned contigs for each evolutionary grouping. (**C**) Species representation among homologous contigs; average number of species per homologous contig group is indicated. (**D**) Distribution of nucleotide matches, mismatches, gaps, and missing nucleotides (N) across the 7 Formicidae (**top**), 4 Myrmicinae (*Acep*, *Aech*, *Pbar*, *Sinv*) (middle), and 2 Attini (*Acep*, *Aech*) (**bottom**) genomes.



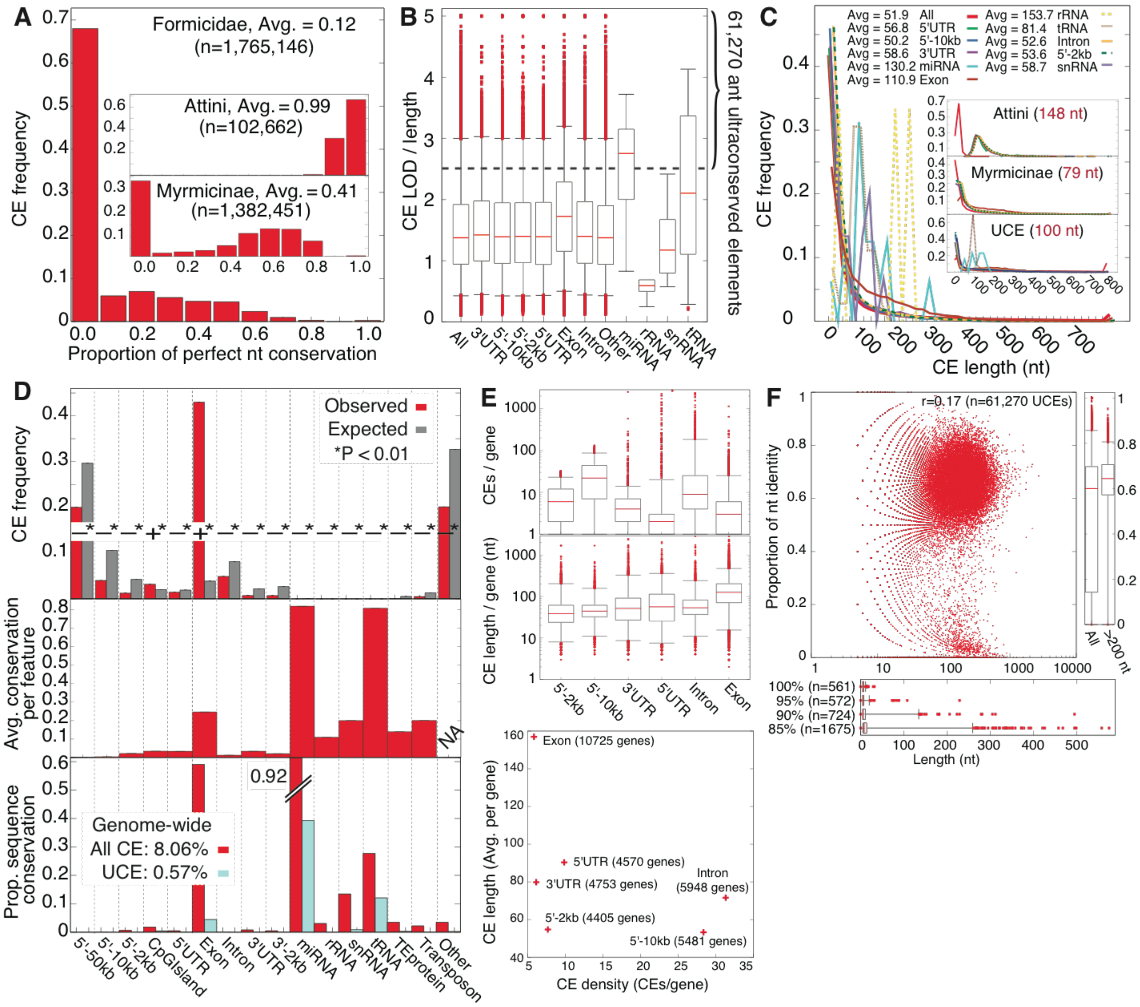
Supplemental Fig. 13. Characteristics of ant synteny blocks. (**A**) Number of synteny blocks for each ant species relative to one other species (top); Average number of genes per synteny block for each ant species relative to another species (middle); Total syntenic genes per synteny block for each ant species relative to another species (bottom). (**B**) Average length of synteny blocks (kilobases, kb). (**C**) Relationship between number of synteny blocks for each species versus the number of syntenic associations between blocks, used as a cutoff for synteny block identification.



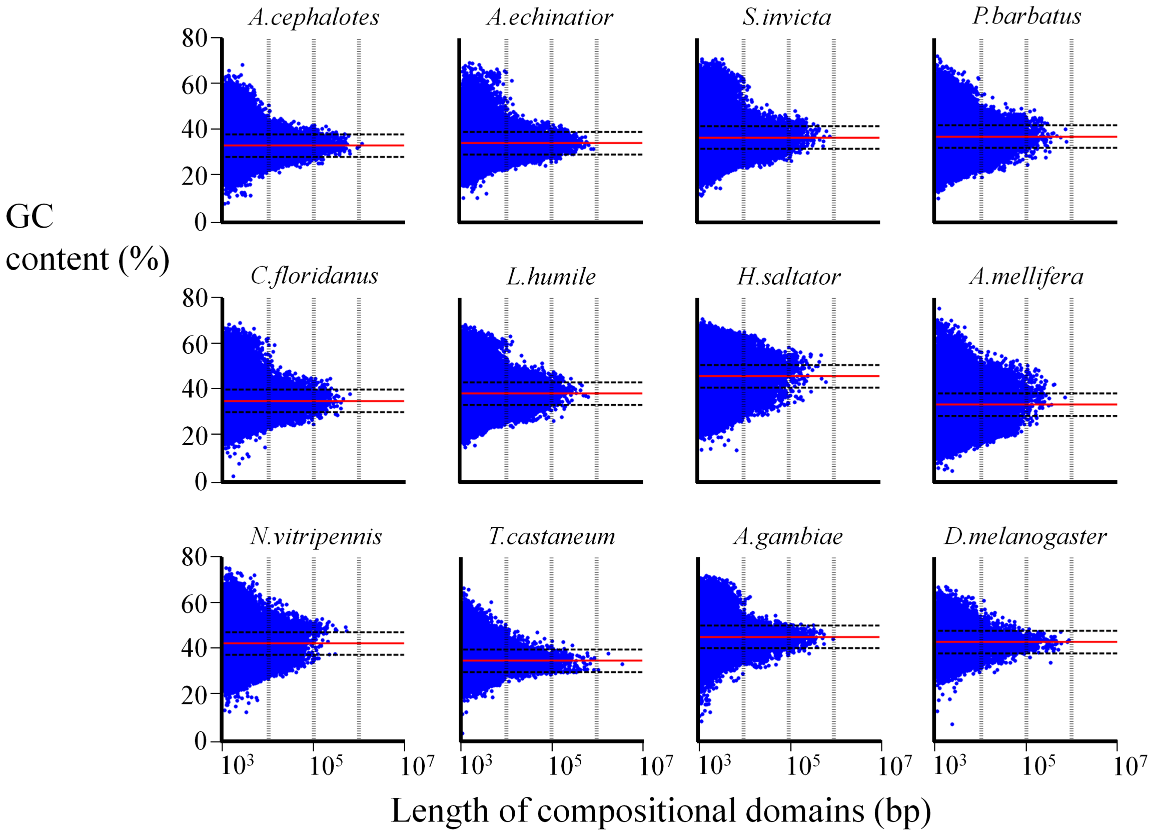
Supplemental Fig. 14. Syntenic fragmentation with increasing evolutionary distance. (**A**) Synteny plots for scaffold 00015 in *A. cephalotes* compared to six other ant and three other insect genomes. Horizontal axes show *A. cephalotes* gene order and vertical axes map orthologous gene models in the respective species. Red and green lines represent gene models in the same or reverse orientation to *A. cephalotes*, respectively. Horizontal gaps represent deletions in the other species. (**B**) Genome-wide proportion of genes inverted (blue) or rearranged (red) between *A. cephalotes* and other species, for all scaffolds greater than one megabase in size. Gene rearrangement is quantified for a target species by ranking genes by order along *A. cephalotes* scaffolds, counting the shift in each gene's rank order relative to *A. cephalotes*, and normalizing this rank shift to the maximum possible number of shifts over all genes. Species are arranged in order of increasing evolutionary distance to *A. cephalotes*.



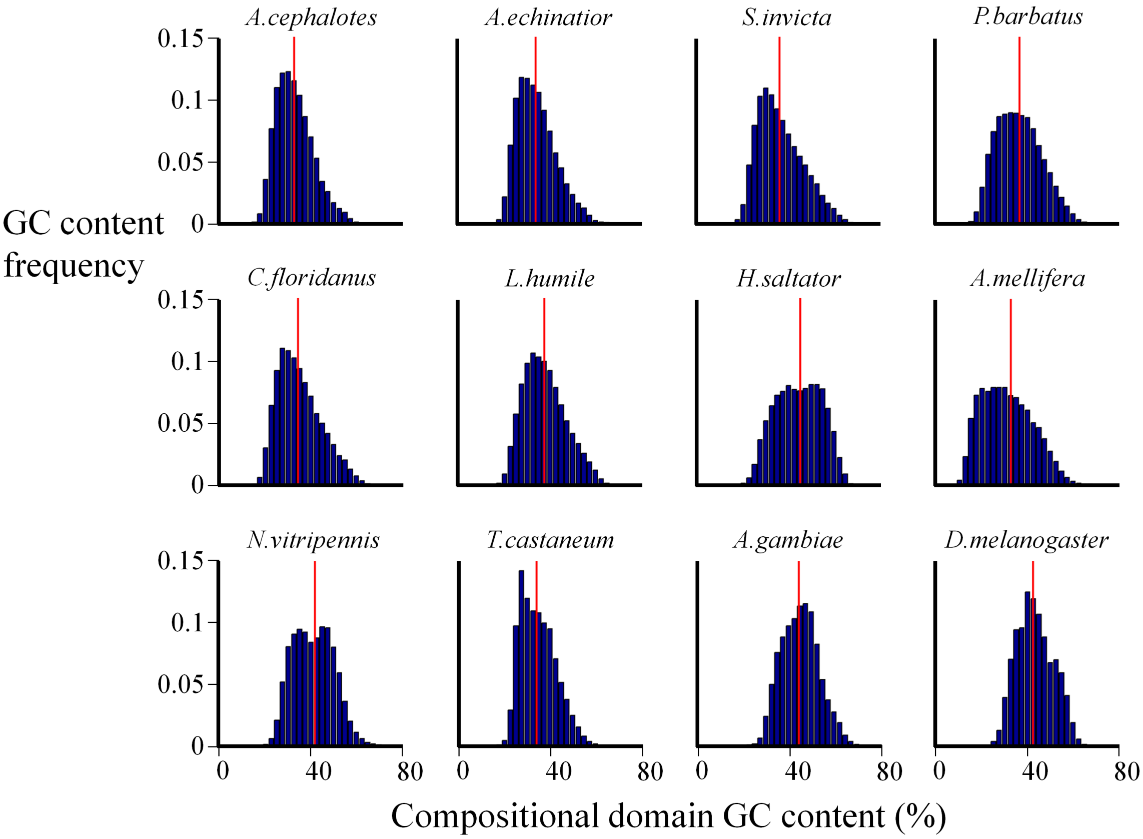
Supplemental Fig. 15. Hox cluster gene order is conserved among ants. Bottom row shows Hox cluster gene order along chromosome 3 in *D. melanogaster.* Columns indicate gene orthology among species. Boxes indicate individual scaffolds in the current genome assembly for each ant species. Vertical lines denote syntenic links between species (see Methods).



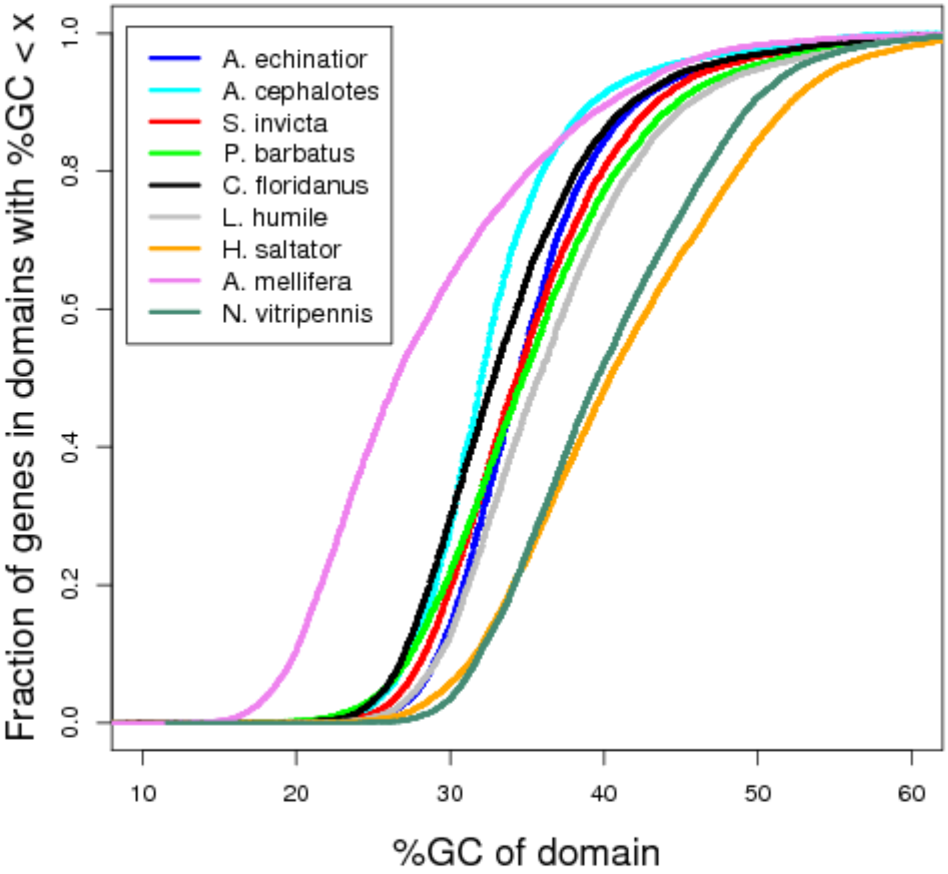
Supplemental Fig. 16. Distribution of conserved elements in ants. (**A**) Number of conserved elements (CEs) and proportion of perfect nucleotide conservation identified from aligned ant genomes (posterior probability of conservation > 0.95) for each of three evolutionary groupings: Formicidae (n=7), Myrmicinae (n=4), Attini (n=2). (**B**) CE conservation scores for Formicidae, estimated by LOD score normalized by CE length. Whiskers indicate outer 5% of distributions. Dashed line indicates 95% cutoff for isolation of ultra-conserved elements (UCEs; n=61,270). (**C**) CE length distributions, grouped by genic feature: Exon, Intron, 5' UTR, 3' UTR, proximal (2 kb) and distal (10 kb) promoters. Noncoding gene classes (miRNA, rRNA, tRNA, snRNA) are also included for CEs in Formicidae. Inside panels show length distributions for Attine and Myrmicine CEs and for auCEs (**bottom**). (**D, top**) Distribution of CEs over annotated genic regions (red), compared to random expectation (gray). Expected distributions were generated by randomly sampling sequences (with the same length distribution as observed CEs) from the 7 genome alignment and assessing genic feature distribution over 100 replicates. Genic regions significantly enriched (+) or depleted (–) for CEs are indicated ($P > 0.99$). (**Middle**) Conservation of each annotated genic feature, assessed as the proportion of DNA sequence for each genic feature that is covered by a single CE, averaged over all features for a given region. (**Bottom**) Estimates of the proportion of each genic region with sequence conservation (thus under purifying selection). Overall estimates of proportion of genome-wide conservation using CEs or UCEs are indicated; note these estimates are based on the total aligned nucleotide sequence among the seven ant genomes. Error bars indicate 1 SE. (**E**) Distributions of number (**left**) and length (**right**) of CEs per protein-coding gene. Whiskers indicate outer 5% of distributions. Right, scatterplot of average CE density vs. length per protein-coding gene; sample sizes are indicated for each genic region. (**F**) Scatterplot of CE length vs. proportion of perfect (7-way) nucleotide identity for 61,270 ultraconserved elements (UCEs). Right, length distribution for all UCEs and the subset of 11,574 UCEs greater than 200 nt in length. Bottom, length distributions for UCEs grouped by minimum percent nucleotide identity.



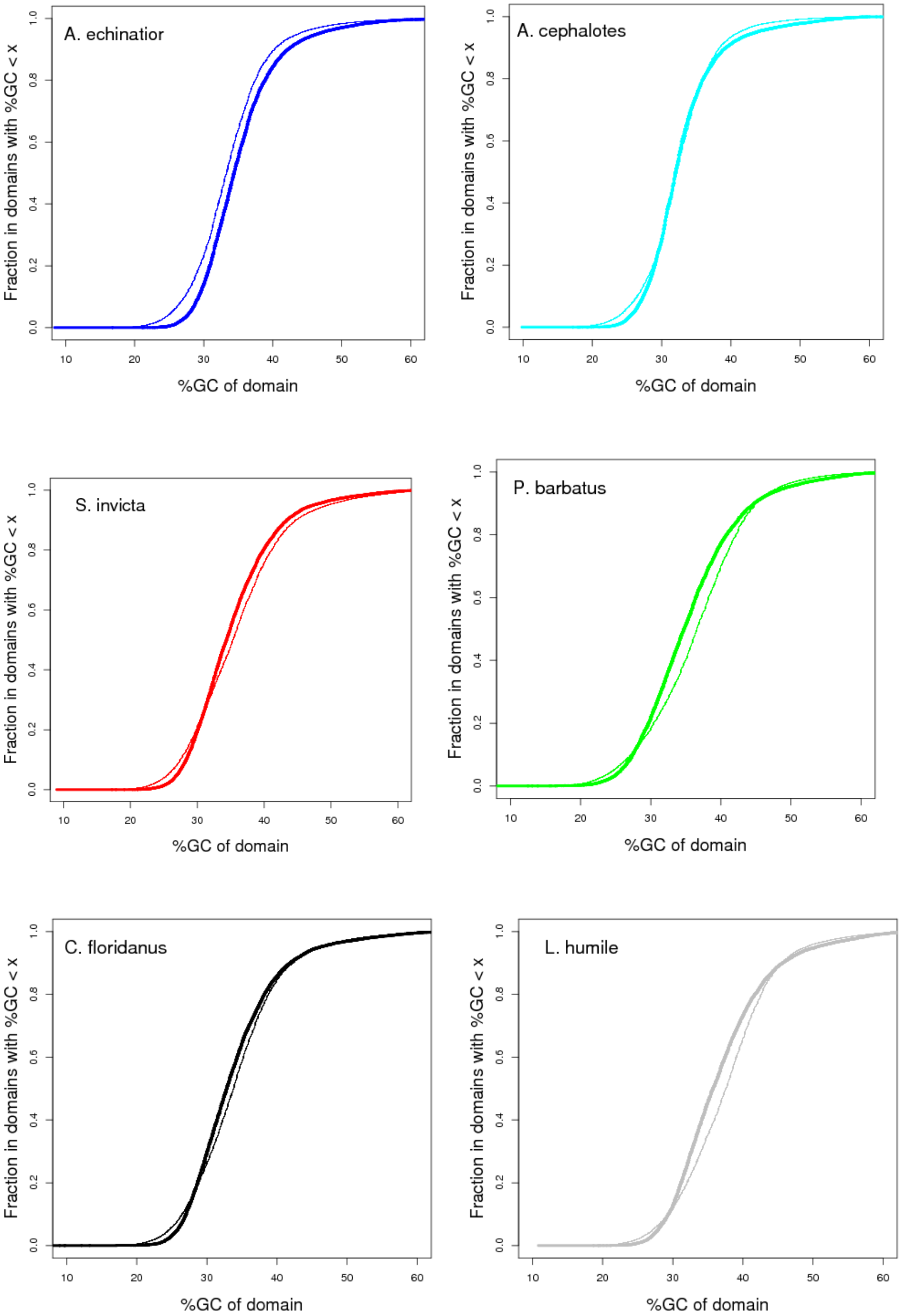
Supplemental Fig. 17. Compositional domain GC-content versus domain lengths on a log scale. The middle horizontal line (solid red) represents the mean genome GC-content within margins of ±5% (dashed black).

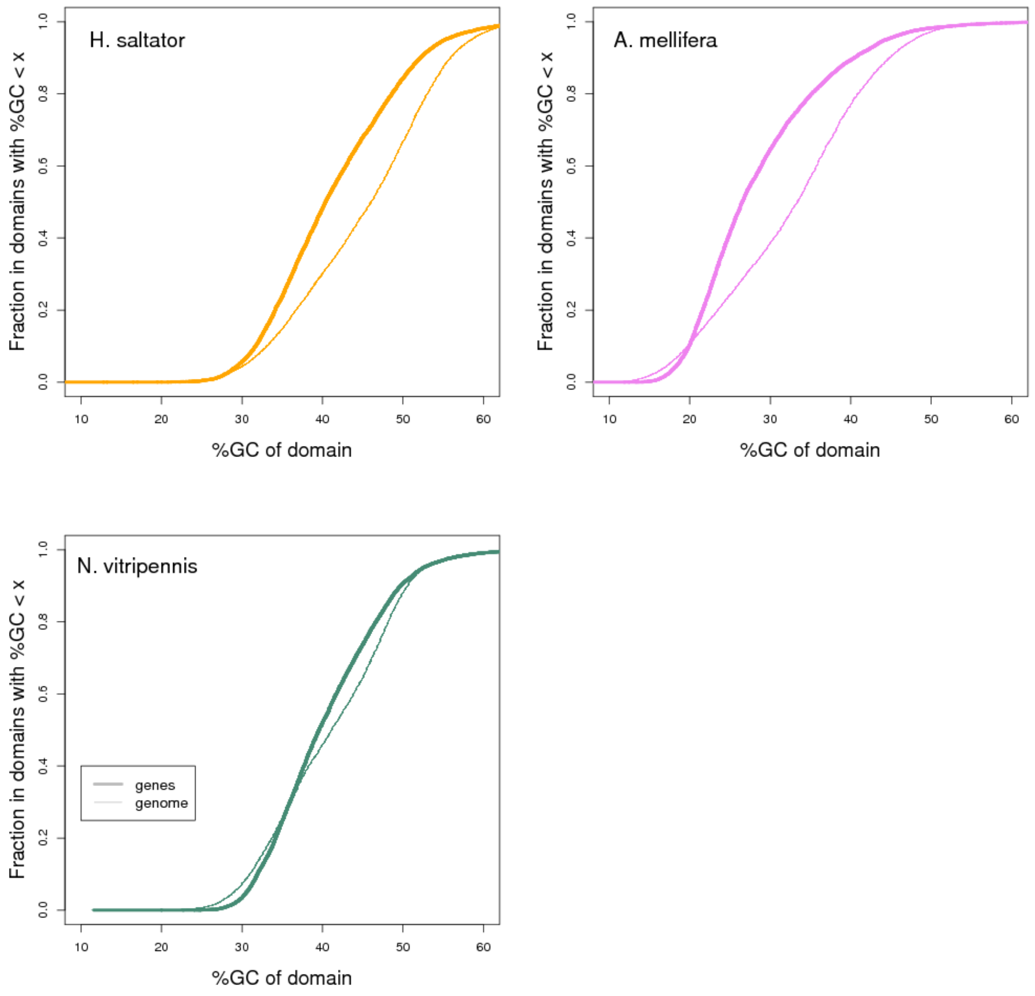


Supplemental Fig. 18. Compositional domain GC-content frequency distribution. The middle horizontal line (solid red) represents the mean genome GC-content.

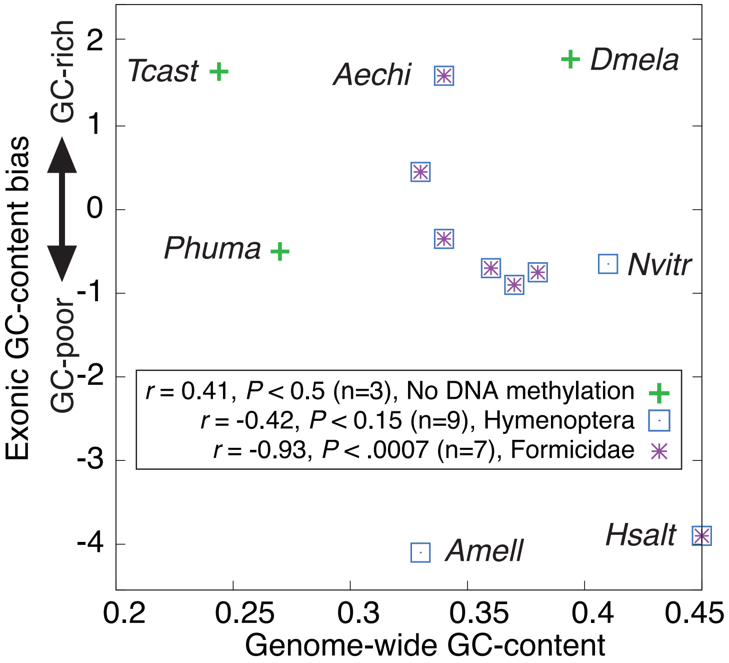


Supplemental Fig. 19. Cumulative distribution of the percent GC of all the genes in the nine species studied. Any point on this curve as the fraction of genes that exists in compositional domains less than a given percent GC. For example, a point on the *L. humile* curve at x = 33 and y = 0.4 indicates that 0.4 (40%) of genes are in poor compositional domains (GC_u<33%).

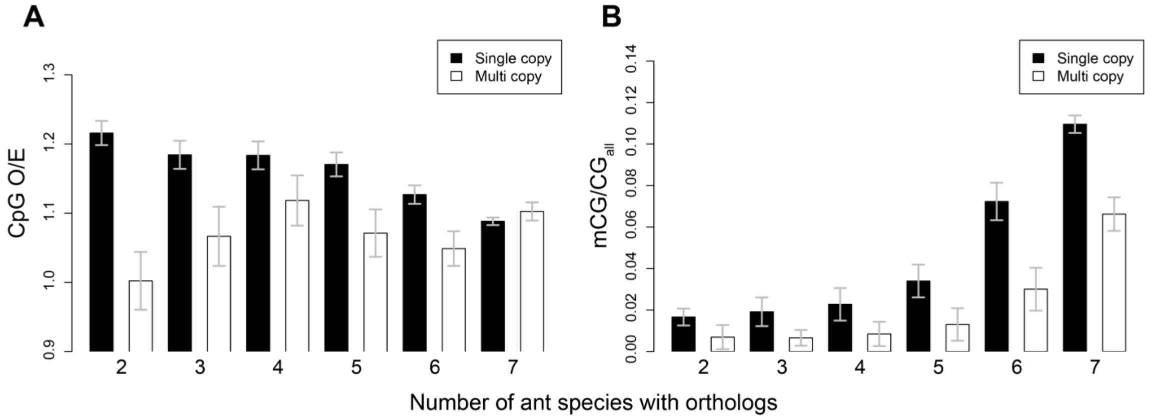

A. echinatior
A. cephalotes
S. invicta
P. barbatus
C. floridanus
L. humile
Fraction in domains with %GC < x
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
%GC of domain



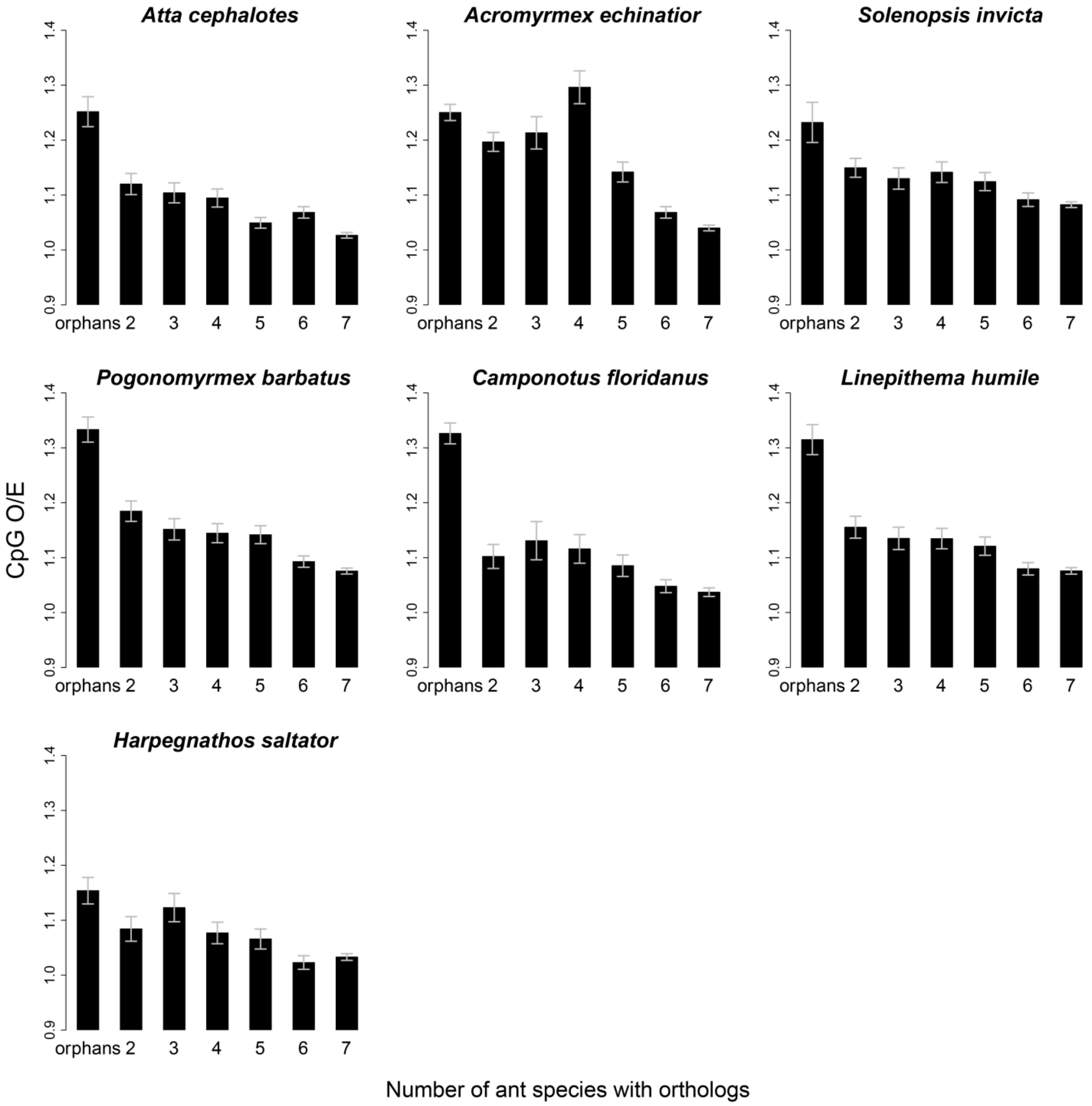
Supplemental Fig. 20. Cumulative distribution of the percent of each genome that is comprised of compositional domains below a given percent GC (thin lines) and the similar distribution for only compositional domains that contain genes (thick lines). If there is no tendency for genes to occur in compositional domains of a particular GC content, these two curves will be essentially the same.



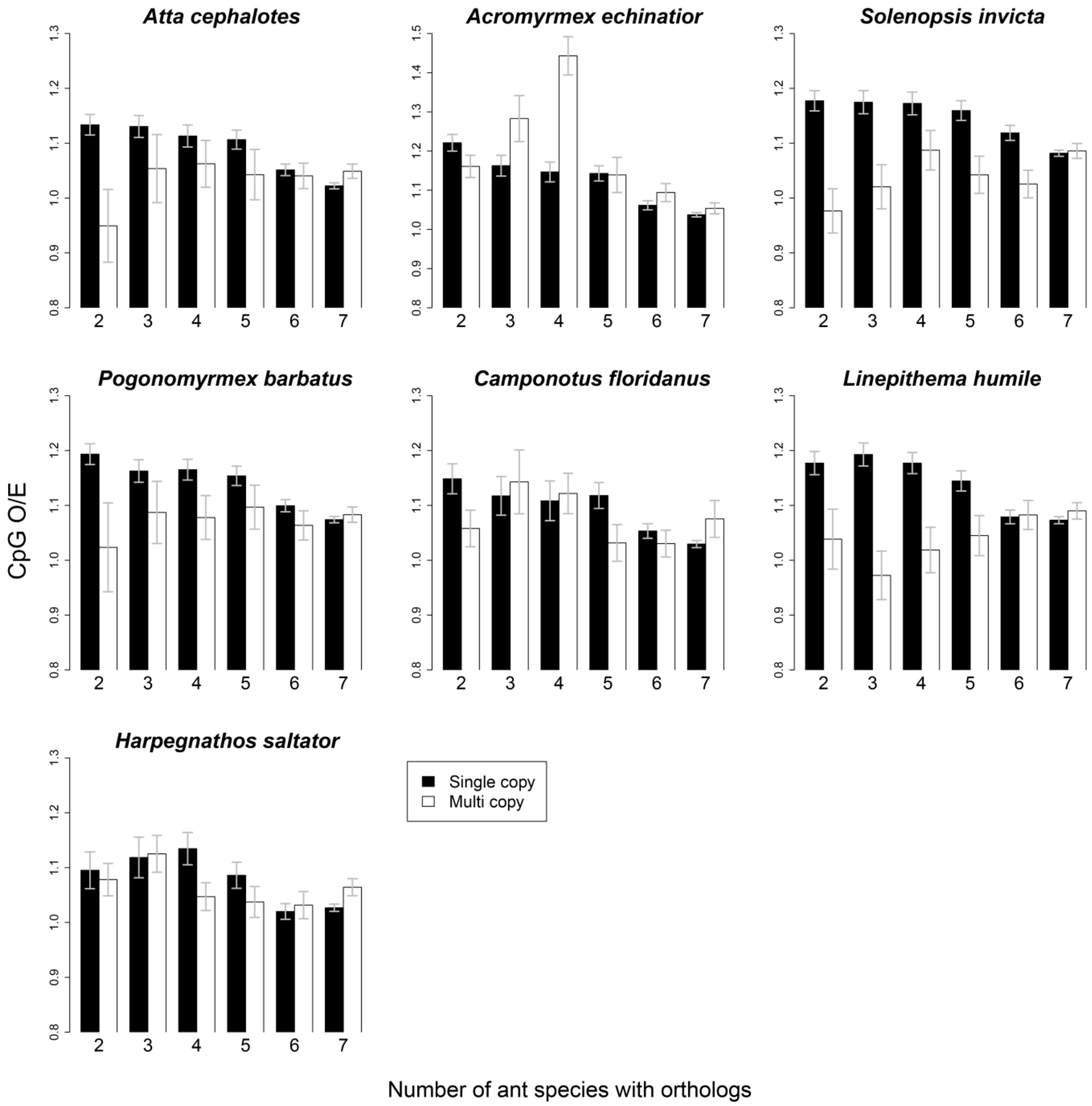
Supplemental Fig. 21. Scatterplot of genome-wide GC-content versus average GC-content bias for all protein-coding genes in 12 insect genomes. GC-content bias was computed as the difference in GC dinucleotide frequency of individual genes compared to the genome-wide background (Supplemental Fig. 20). Unexpectedly, ants cover the whole range of GC-content bias observed in animals. Pearson correlations were computed for three groups: insects lacking DNA methylation (Dmela, Phum, Tcast), Hymenoptera (n=9), and Formicidae (n=7). P-values computed using one-sample T-test.



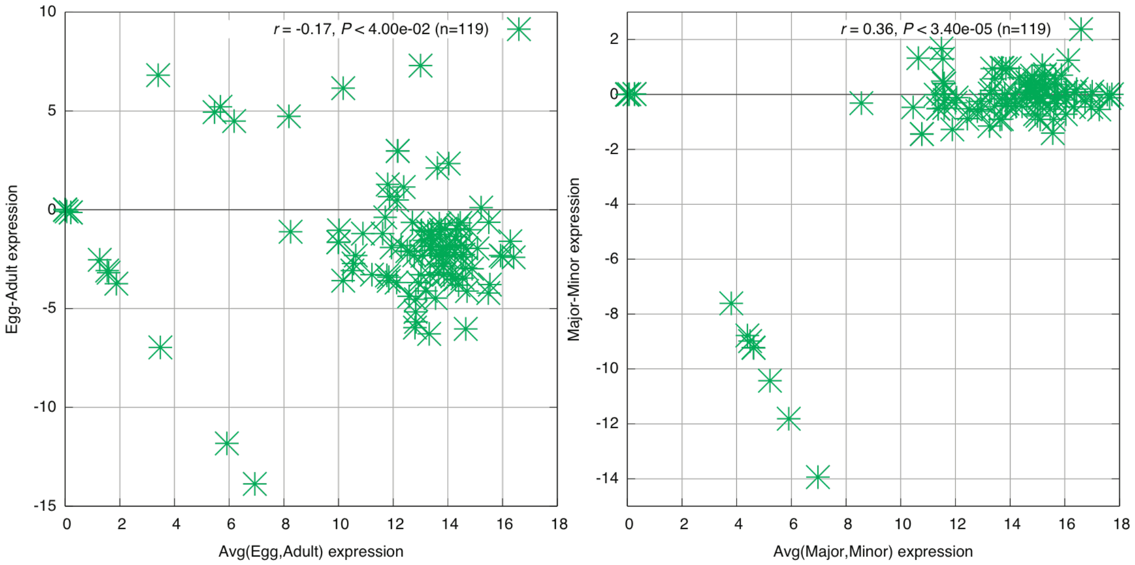
Supplemental Fig. 22. Normalized CpG content (CpG O/E) faithfully reflects DNA methylation of single copy, but not multi copy, genes. (**A**) *Solenopsis invicta* CpG O/E values of coding sequences suggest orthologs that are single copy in all lineages (single copy) exhibit increasing methylation with increasing prevalence of orthology among seven ant taxa, but this pattern does not hold for orthologs that are multi-copy in some lineages (multi copy). (**B**) In contrast, fractional methylation data (mCG/CG$_{all}$) demonstrate that methylation is correlated with the taxonomic prevalence of orthologs for single copy and multi copy genes. Notably, genes with orthologs in all seven ant genomes exhibit the highest methylation levels among single copy and multi copy orthologs. These results suggest that CpG O/E is not a good indicator of DNA methylation for multi copy genes (see Supplemental Table 8). Means and 95% confidence intervals are plotted.



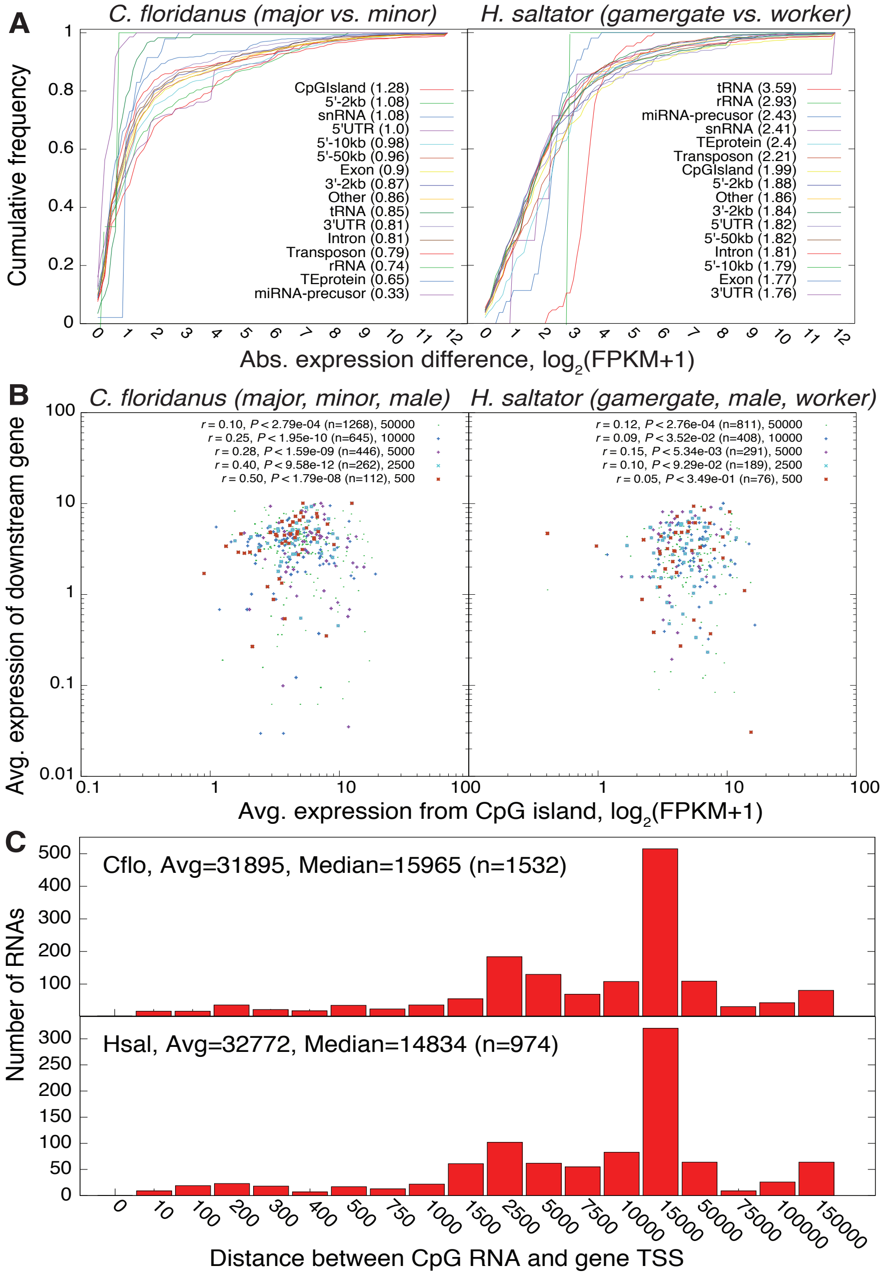
Supplemental Fig. 23. DNA methylation levels differ according to gene conservation. Normalized CpG content of coding sequences within genes (CpG O/E) are grouped according to the number of species with orthology. The relatively high CpG O/E values in orphan genes suggest they are largely unmethylated, whereas the relatively low CpG O/E values suggest seven-species orthologs are the primary targets of DNA methylation. Differences are highly significant in each species (Kruskal-Wallis *P* < 0.0001). Means and 95% confidence intervals are plotted.



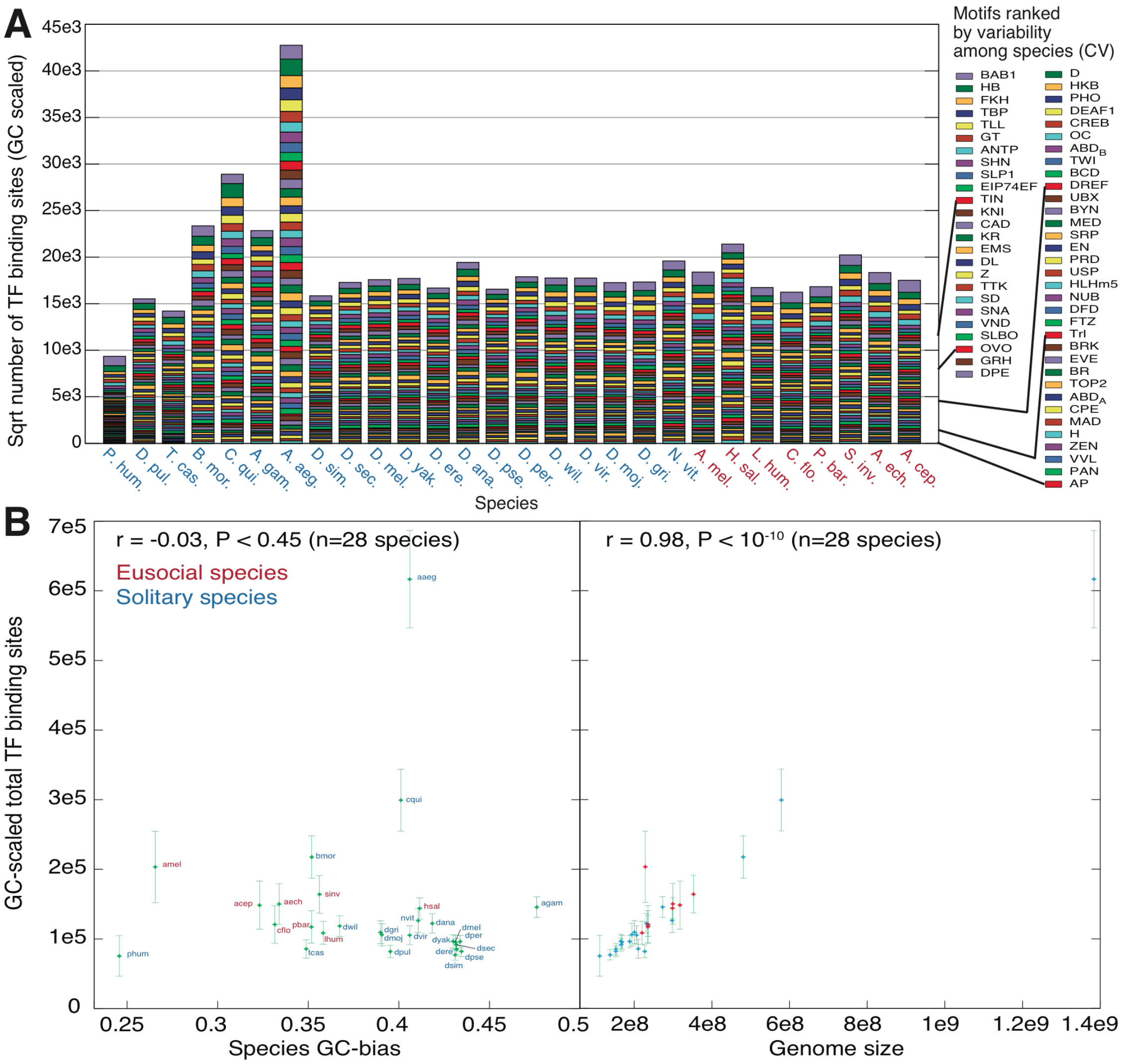
Supplemental Fig. 24. Normalized CpG content (CpG O/E) of coding sequences grouped according to the number of species with orthologs that are either multi copy in some lineages (multi copy) or single copy in all lineages (single copy). Estimates of methylation from single copy orthologs (see Supplemental Fig. 21) consistently suggest that seven-way orthologs are the primary targets of DNA methylation, and that DNA methylation infrequently targets taxonomically restricted or fast evolving genes. In contrast, CpG O/E of multi copy orthologs does not follow this trend and may not reflect methylation status (see Supplemental Fig. 21). Differences are highly significant in all species among single-copy orthologs (Kruskal-Wallis *P* < 0.0001) and are significant in each species except *P. barbatus* among multi-copy orthologs (Kruskal-Wallis *P* < 0.05). Means and 95% confidence intervals are plotted.



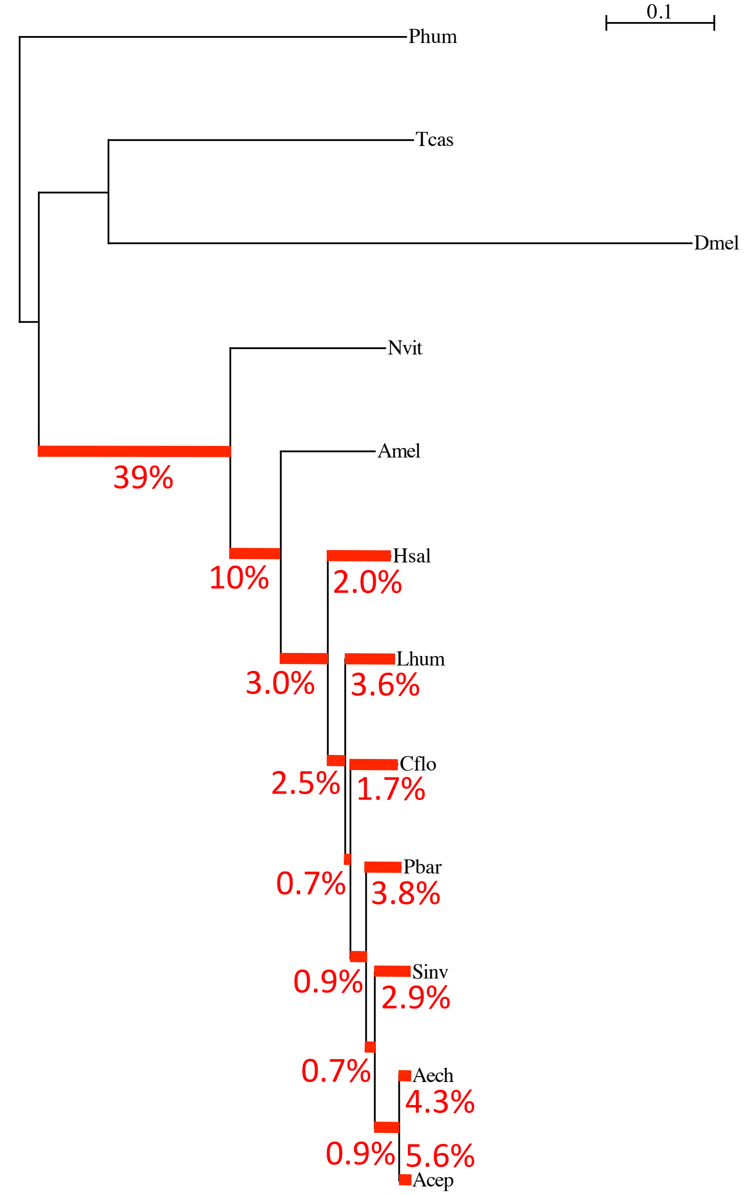
Supplemental Fig. 25. Expression of miRNA genes in *C. floridanus*. Left panel compares miRNA expression averaged over egg and adult stages (major, minor, male) with differential expression between egg and adult stages. Right panel compares miRNA expression averaged over female worker castes (major, minor) with differential expression between these castes. Expression estimates derived from small RNA-Seq data and quantified as $\log_2$(FPKM+1).



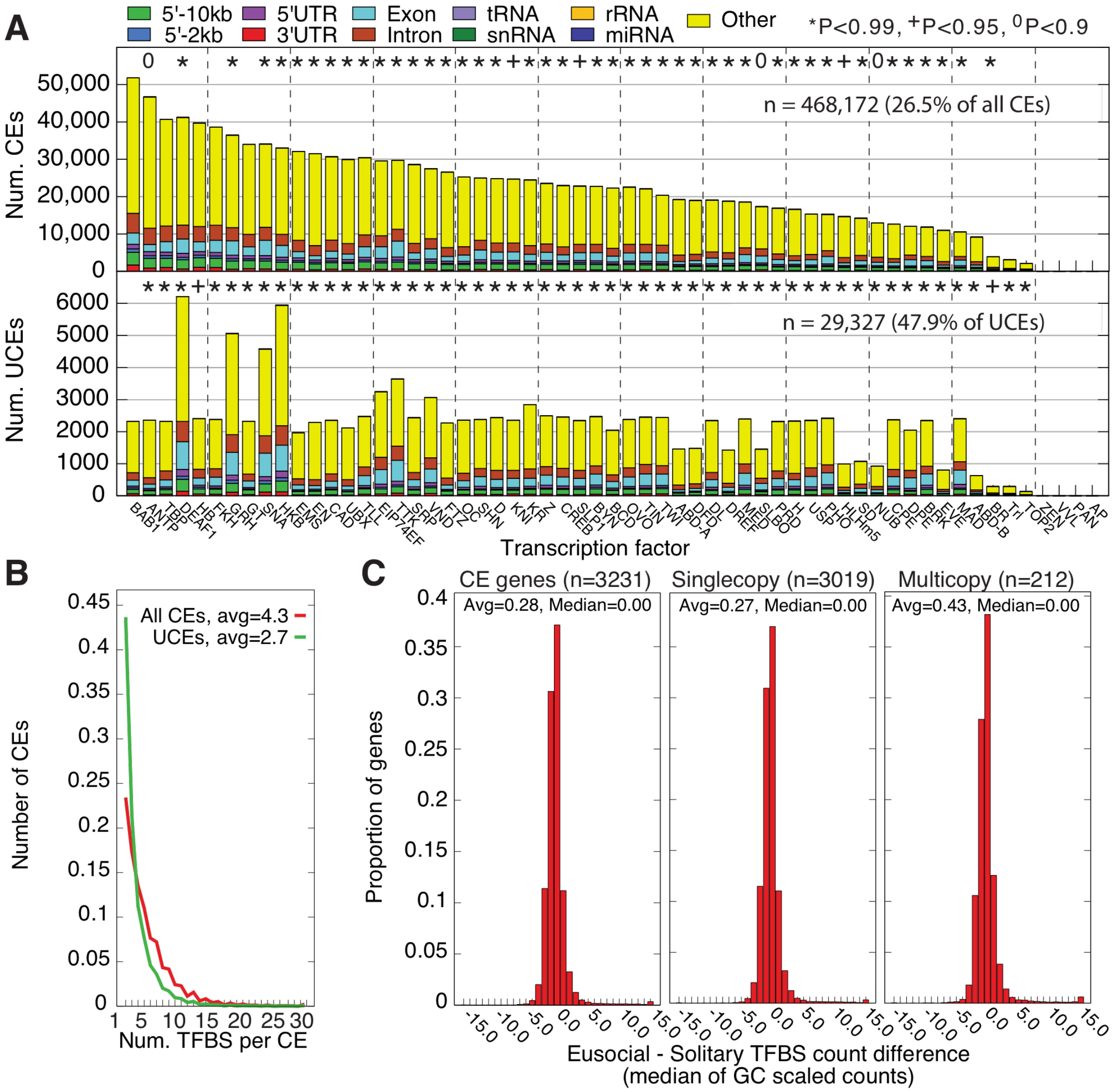
Supplemental Fig. 26. Expression of noncoding RNAs that overlap CEs. (**A**) Cumulative distributions of worker caste variation in gene expression of noncoding RNAs overlapping different genic regions, for *C. floridanus* on left and *H. saltator* on right. Values in parentheses indicate absolute difference in caste expression for the 50th percentile of features for a given region; regions are ranked by this statistic. (**B**) Relationship between expression level of small RNAs overlapping CpG islands and expression level of nearest downstream protein-coding gene, grouped by distance to nearest gene. Expression levels reflect average $\log_2$(FPKM+1) over three castes, as indicated. Pearson correlation coefficients are reported. (**C**) Distribution of distances between conserved CpG RNAs and nearest downstream protein-coding genes.



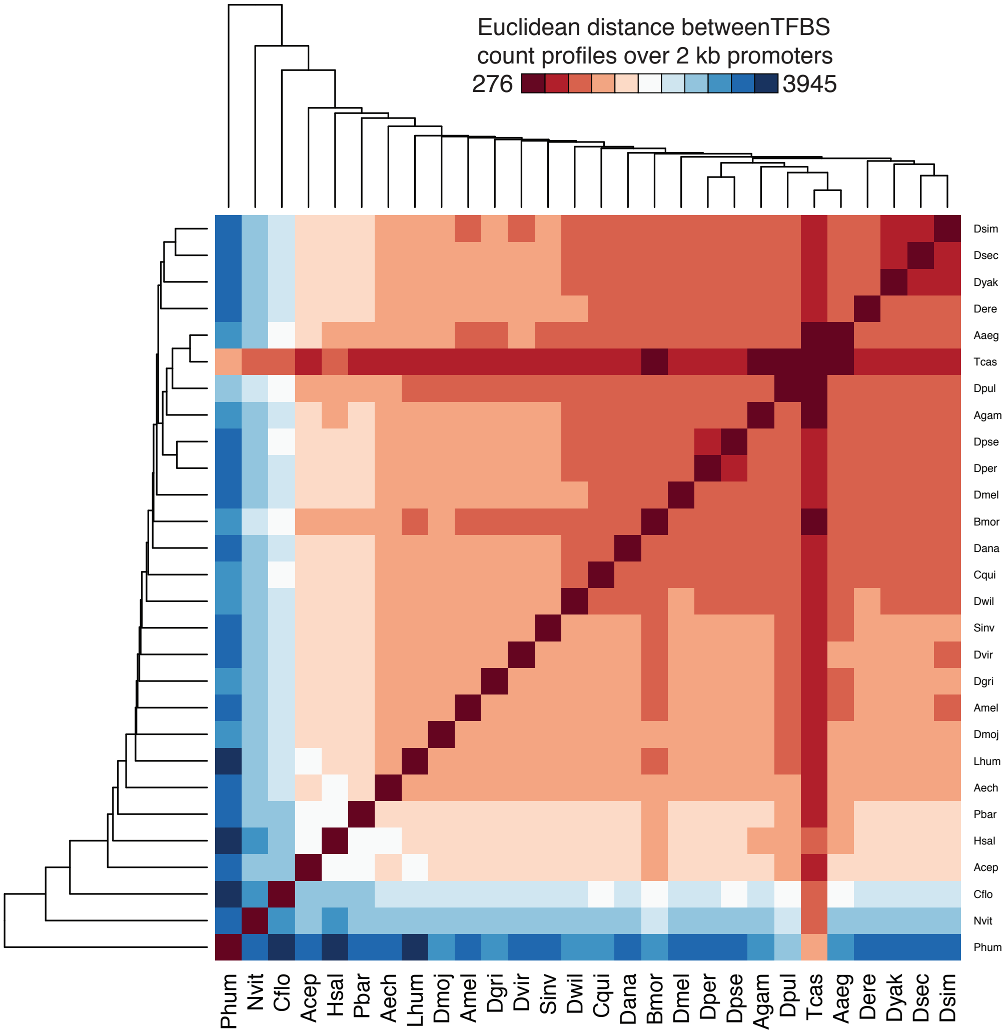
Supplemental Fig. 27. Genome-wide distribution of TF binding sites in insects. (**A**) Genome-wide distributions of total binding sites predicted for each species, separated by TF. TFs are ranked by species variability, Variance(|TFBS_i|)/Mean(|TFBS_i|), for each TF i. Mean and variance are computed across species. (**B**) Comparison of GC-bias (left) and Genome size (right) versus number of binding sites predicted for 28 insect species, averaged over 59 motifs. Binding site number for each motif is scaled by the ratio GC(x)/avg(GC), which corrects for variation in GC-bias among species. Error bars indicate 1 SEM over TFs. Red and blue text indicates eusocial and solitary species, respectively. Most genomes show very similar TFBS distributions (A), and we found no significant relationship to GC-content (P<0.45), though GC content is variable (B, left). However, 96% of observed variation in the number of TFBSs among species is explained by overall genome-size ($P<10^{-10}$) (B, right), as expected by our assumption of constant TFBS occurrence probability among species (1 TFBS per 5000nt).



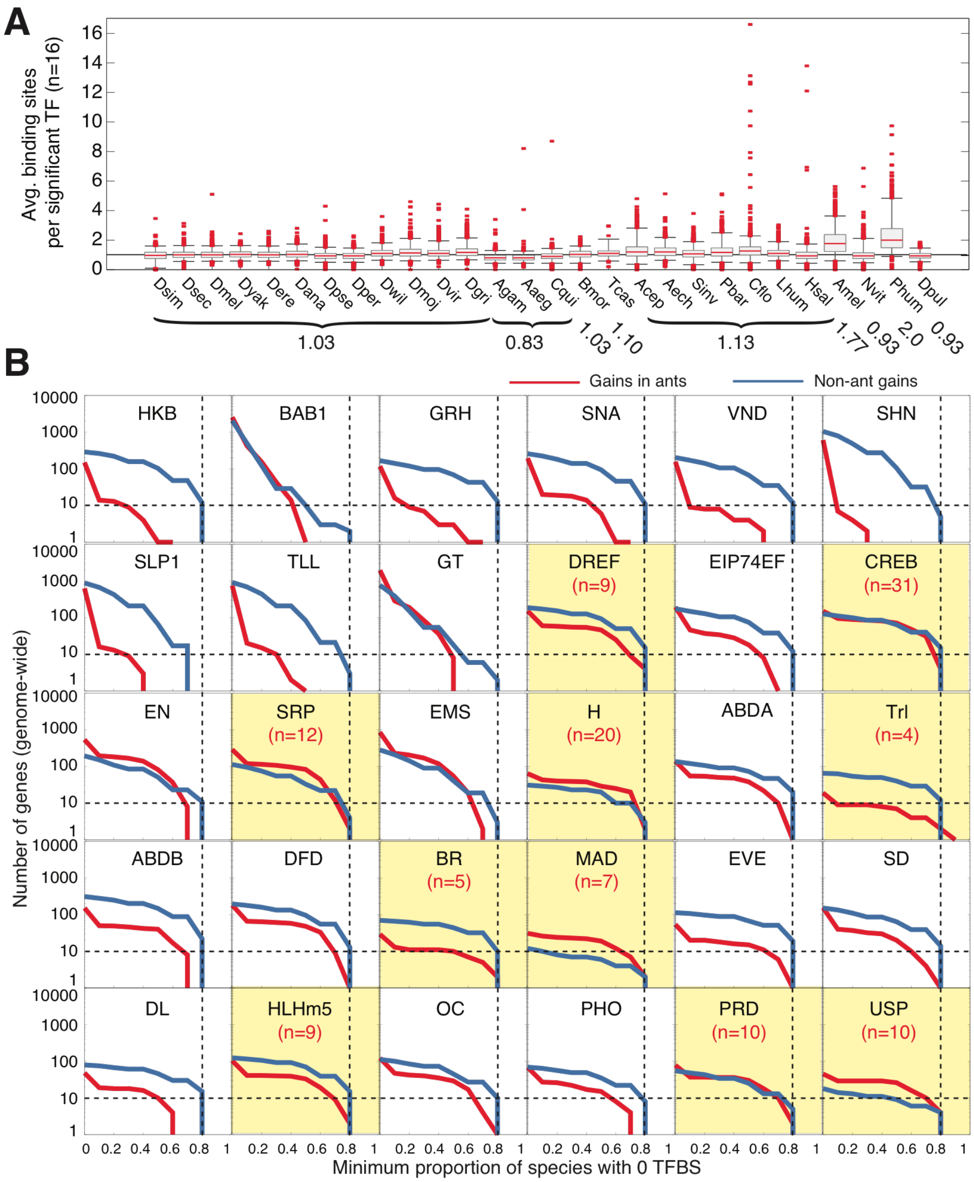
Supplemental Fig. 28. Phylogeny of the 12 species used in the positive selection analysis (see section on phylogeny above). All families tested did not include duplications, so their topology is following the species tree. In red are the 15 branches which were used as foreground branches in successive runs of the branch-site test for each families. The percentages indicated on each of these branches represent the proportion of gene families that display a significant signal for positive selection at a FDR threshold of 10%.



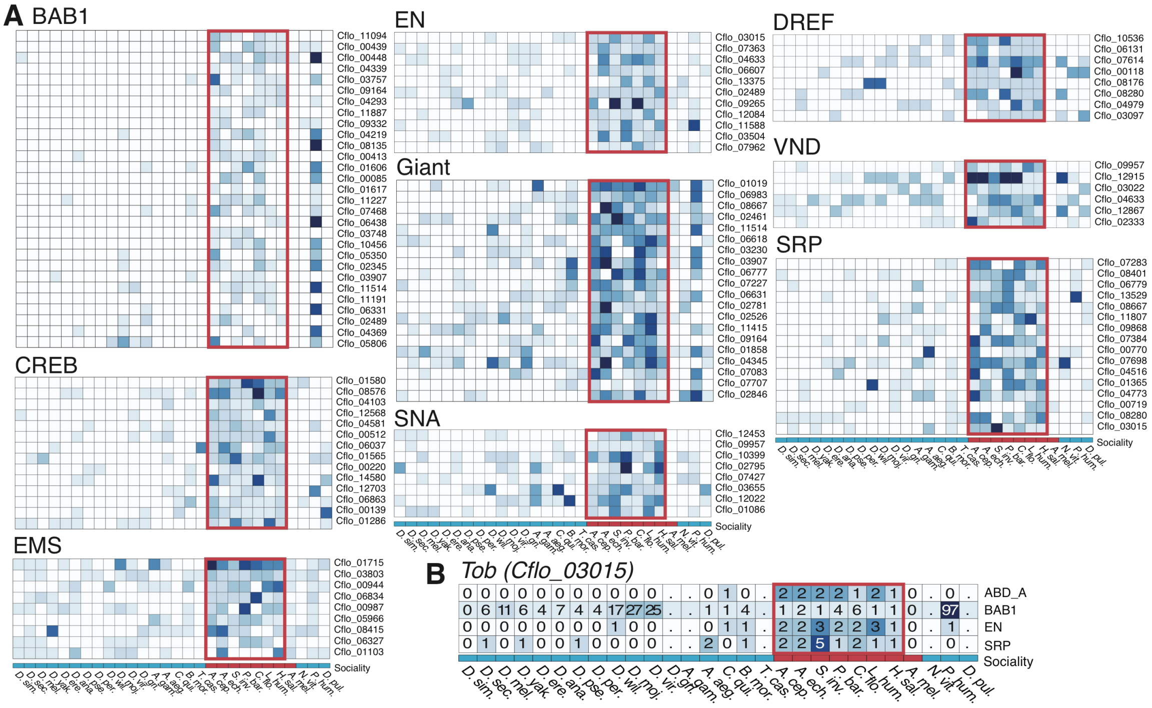
Supplemental Fig. 29. Genome-wide distribution of TF binding sites over genic regions. (**A**) Distribution of predicted TF binding sites (TFBSs) among ant conserved elements (CEs) for 59 TF sequence elements, sorted by total binding sites. Each stacked bar shows the binding site proportions among genic regions. Results for all CEs (**top**) and ultra-conserved elements (UCEs; **bottom**). P-values indicate whether more binding sites for a TF were found among CEs than among random sequences, computed by randomly sampling homologous sequences from the whole-genome alignment and counting predicted TFBSs (n=100; random sequences match the CE length distribution); $^{*}P < 0.99$; $^{+}P < 0.95$; $^{o}P < 0.9$. (**B**) Distributions of number of binding sites per individual CE, for all CEs (red) and the subset of ultraconserved elements (auCEs). (**C**) Distributions of the difference in GC-scaled number of TF binding sites between eusocial and solitary species per gene, pooling differences for all 59 TFs. Only binding sites occurring within 2 kb of the predicted transcription start site of homologous protein-coding genes were included. Genome-wide distributions for all (**left**), single-copy (**middle**), and multi-copy (**right**) genes are shown.



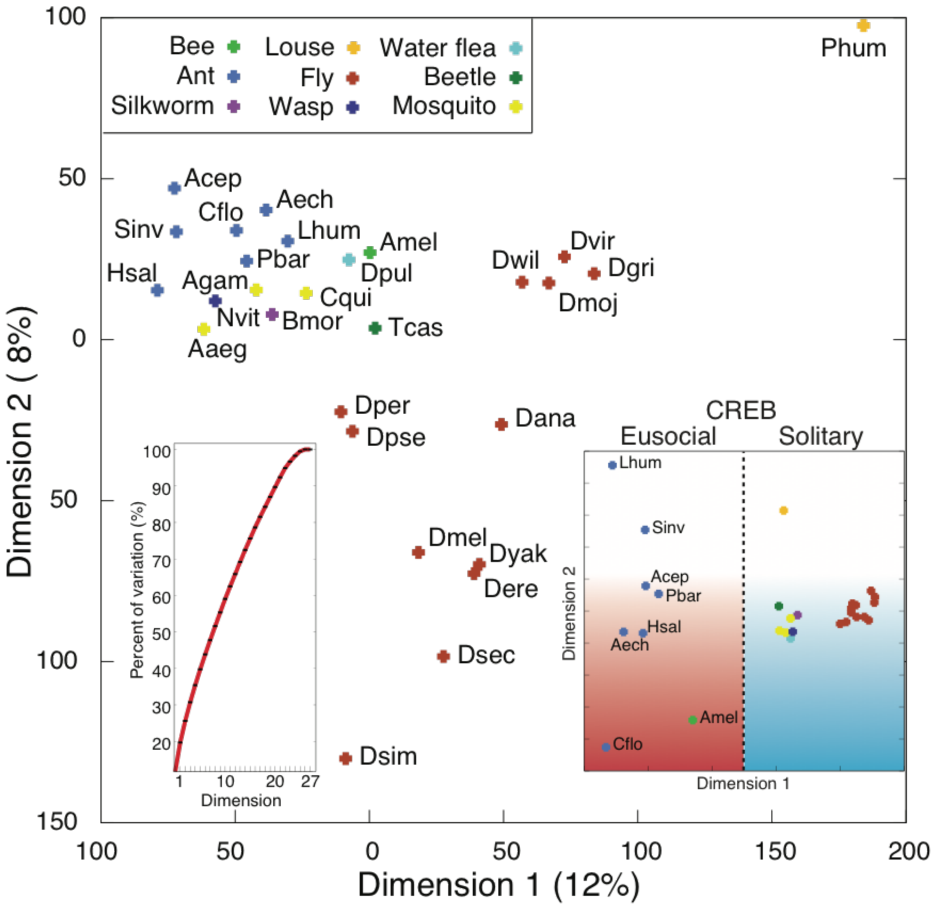
Supplemental Fig. 30. Hierarchical clustering of insect species using TFBS count profiles in 2kb promoters of protein-coding genes for 57 TFs. Clustering was performed using average linkage and Euclidean distance.



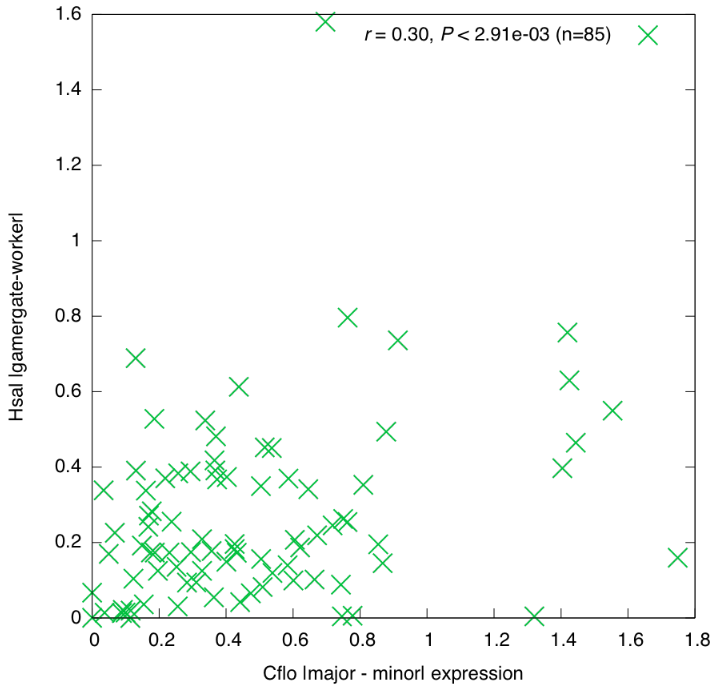
Supplemental Fig. 31. Conservation and divergence of TFBSs for significant TFs. (**A**) Distributions of the average number of TFBSs per gene among target genes for the top 16 TFs significantly associated with eusocial regulatory evolution (see Fig. 5B). Boxes denote 25–75% percentiles; whiskers denote inner 95% of data; outliers shown as red dots. The overall mean number of binding sites per TF is reported for major taxonomic groups and is computed as the median binding sites for each species, averaging over species per taxonomic group as indicated. (**B**) Genome-wide distributions of the number of genes associated with gains (red) or losses (blue) in TF binding sites in the ant lineages (n=7), compared to all other insect species (n=21). X-axis denotes the proportion of species, either ants (blue) or non-ants (red) for which the specified TF shows 0 predicted binding sites. Distributions for the 30 TFs associated with significant regulatory evolution are shown. 10 TFs highlighted in yellow have a positive number of genes with at least 80% of species showing 0 binding sites (blue, ants; red, non-ants).



Supplemental Fig. 32. Heatmaps illustrating number of TF binding sites per gene across insect species (n=28) for genes that show significantly increased binding sites in eusocial lineages without concordant increases in *A. mellifera*. A total of 141 genes met this criteria. (**A**) Heatmaps for nine TFs associated with at least 5 genes are shown. (**B**) Heatmap for a single gene *Tob,* which shows 0 predicted binding sites in *A. mellifera* for four TFs (ABD_A, BAB1, EN, SRP), despite significant changes in ants compared to solitary species. *Tob* encodes a cell antiproliferative protein that interacts with multiple signaling proteins to regulate cell proliferation (Jia and Meng 2007).



Supplemental Fig. 33. Principle components visualization of TF binding site evolution among insects. Singular value decomposition was applied to the 1955 gene x 28 species matrix whose values represent the total number of TF binding sites for the 16 TFs with significant binding site evolution (see Fig. 5B). Resulting species vectors were projected onto the top 2 eigenvectors (dimensions of covariation), as shown. Proportion of variation in TF binding sites among species explained by each dimension is shown on axes and in left inset plot. Right inset plot shows similar analysis of TF binding sites for a single TF, CREB. Vertical dashed line separates eusocial from solitary insects.



Supplemental Fig. 34. Correlation of variation in worker caste expression between species. Plasticity in worker caste gene expression was computed as the absolute value of the standard deviation in $\log_2$(FPKM+1) expression levels between major and minor samples (*C. floridanus*) and gamergate and worker samples (*H. saltator*). 85 of the 96 genes with concentrated regulatory evolution in TFBSs for multiple TFs per gene are shown (those with data in both species). Correlation computed using Pearson metric.

Supplemental Table 1. Organism and gene set information for the twelve insects included in AntOrthoDB (*http://cegg.unige.ch/orthodbants*).

Organism	**Common Name**	**Code**	**Gene Set**	**Gene Count**
Pediculus humanus	Body Louse	*PHUMA*	PhumU1.2	10,773
Drosophila melanogaster	Fruit Fly	*DMELA*	FB5.29	13,752
Tribolium castaneum	Flour Beetle	*TCAST*	Tcas_3.0	16,645
Nasonia vitripennis	Jewel Wasp	*NVITR*	OGS_v1.2	18,731
Apis mellifera	Honey Bee	*AMELL*	Amel_pre_release2	10,699
Harpegnathos saltator	Jumping Ant	*HSALT*	OGS_3.3	18,564
Linepithema humile	Argentine Ant	*LHUMI*	OGS_1.2	16,116
Camponotus floridanus	Carpenter Ant	*CFLOR*	OGS_3.3	17,064
Pogonomyrmex barbatus	Harvester Ant	*PBARB*	OGS_1.2	17,189
Solenopsis invicta	Fire Ant	*SINVI*	OGS_2.2.3	16,522
Acromyrmex echinatior	Leaf-cutter Ant	*AECHI*	OGS_1.0	20,243
Atta cephalotes	Leaf-cutter Ant	*ACEPH*	OGS_1.2	18,093

Supplemental Table 2. Orthologous protein length agreement between each of the seven ant species and the honeybee *A. mellifera* (*AMELL*) and the wasp *N. vitripennis* (*NVITR*). Concordances with 95% confidence limits (Conf. Lim.) are shown, as well as proportions of longer or shorter ant proteins compared to their bee or wasp orthologs.

	AMELL			
Species	Concordance	Conf. Lim.	Longer	Shorter
HSALT	0.91	0.90–0.91	28.20%	25.65%
LHUMI	0.87	0.86–0.87	26.76%	24.18%
CFLOR	0.92	0.91–0.92	29.48%	22.88%
PBARB	0.90	0.89–0.91	25.72%	28.97%
SINVI	0.83	0.82–0.84	22.64%	30.58%
AECHI	0.89	0.88–0.89	29.16%	24.90%
ACEPH	0.90	0.90–0.91	21.93%	31.41%

	NVITR			
Species	Concordance	Conf. Lim.	Longer	Shorter
HSALT	0.90	0.89–0.90	19.74%	33.06%
LHUMI	0.86	0.85–0.87	16.11%	28.38%
CFLOR	0.90	0.89–0.90	21.27%	31.22%
PBARB	0.89	0.89–0.90	14.37%	31.97%
SINVI	0.81	0.80–0.82	12.38%	32.29%
AECHI	0.88	0.87–0.88	16.82%	27.24%
ACEPH	0.89	0.88–0.89	11.66%	33.29%

Supplemental Table 3. Genes with paralog (GWP) counts for different definitions of paralogs (Set 1–4) across 30 arthropod species.

Species	**GWPs for Set 1**	**GWPs for Set 2**	**GWPs for Set 3**	**GWPs for Set 4**
Aaeg	12915	13060	13474	14234
Acep	8748	9690	9729	11659
Acyp	22313	21498	24038	25322
Aech	9106	9325	9829	10880
Agam	10061	10137	10637	11368
Amel	6900	7192	7471	8327
Bmor	7872	7780	8705	9615
Cflo	10539	10088	11551	12304
Cqui	13352	13860	14098	15636
Dana	9287	9237	10047	10884
Dere	8978	9054	9716	10645
Dgri	9705	9856	10388	11313
Dmel	8754	8610	9475	10123
Dmoj	8773	8826	9504	10364
Dper	10383	10743	11194	12442
Dpse	10213	10154	10992	11862
Dpul	19803	19648	21036	22434
Dsec	10015	10370	10757	11945
Dsim	8964	9512	9704	11037
Dvir	8854	8832	9594	10415
Dwil	10111	10065	10842	11669
Dyak	9905	10042	10643	11616
Hsal	12253	11618	13276	13957
Isca	9671	9007	10868	11442
Lhum	8936	9607	9779	11384
Nvit	13632	13830	14392	15481
Pbar	8632	9323	9502	11220
Phum	5731	5785	6363	7135
Sinv	9410	10068	10410	12307
Tcas	9496	9526	10182	11005

Supplemental Table 4. Summary of the dataset used for codon usage bias analysis (asterisks indicate genes that passed all quality filters).

Acro-nym	***Species name***	***Gene set version***	***No. of genes****	***Proportion of genes****	***Proportion of 1-copy orthologs****	***No. of ribosomal genes****
Acep	*Atta cephalotes*	OGS_1.2	15401	85.1	82.4	19
Aech	*Acromyrmex echinatior*	OGS_1.0	19926	98.4	94.6	68
Amel	*Apis mellifera*	Amel_pre_release2	9930	92.8	93.6	70
Cflo	*Camponotus floridanus*	OGS_3.3	16355	95.8	99.9	85
Dmel	*Drosophila melanogaster*	FB5.29	13687	99.5	99.9	89
Hsal	*Harpegnathos saltator*	OGS_3.3	17191	92.6	99.9	114
Lhum	*Linepithema humile*	OGS_1.2	11917	74.0	69	5
Nvit	*Nasonia vitripennis*	OGS_v1.2	14086	75.2	64.5	7
Pbar	*Pogonomyrmex barbatus*	OGS_1.2	12252	71.3	58.3	14
Phum	*Pediculus humanus*	PhumU1.2	10725	99.6	99.7	77
Sinv	*Solenopsis invicta*	OGS_2.2.3	15817	95.7	97.9	66
Tcas	*Tribolium castaneum*	T cas_3.0	16609	99.8	97.4	98

Supplemental Table 5. Rates of gene gain and loss estimated by considering each branch at a time independently from all others. Branches were assigned to rate categories based on *k*-means clustering with *k* = 4. The subsequent model with four rate categories (Fig. 2B) fitted the data significantly better than all other tested models.

Branch leading to	**Branch-specific rate**	**Rate category**
T. castaneum	0.00048	2
Non-hymenopteran Holometabola	0.00000	1
D. virilis	0.00126	3
Drosophilidae	0.00077	2
D. melanogaster	0.00216	3
D. melanogaster + *D. erecta*	0.00091	2
D. erecta	0.00177	3
D. melanogaster group	0.00666	4
D. ananassae	0.00122	3
Drosophila subgenus *Sophophora*	0.00105	3
D. pseudoobscura	0.00231	3
Diptera	0.00380	4
A. aegypti	0.00126	3
P. barbatus	0.00092	2
Myrmicinae	0.00152	3
A. echinatior	0.01192	4
Attini	0.00026	2
A. cephalotes	0.01224	4
Myrmicinae + Formicinae	0.00228	3
C. floridanus	0.00130	3
Formicoida	0.00106	3
L. humile	0.00086	2
Formicidae	0.00061	2
H. saltator	0.00134	3
Aculeata	0.00077	2
A. mellifera	0.00113	3
Hymenoptera	0.00000	1
N. vitripennis	0.00159	3

Supplemental Table 6. Immune gene families and their sizes in selected insects (taxon abbreviations as in Supplemental Table 4). For explanations of gene family acronyms see Material and Methods, Immune Genes. Values for *A. mellifera*, *T. castaneum* and *D. melanogaster* were obtained from published analyses (CR Smith et al. 2012, Pauli et al. 2011, Elango et al. 2009).

	Aech	*Acep*	*Cflo*	*Hsal*	*Lhum*	*Pbar*	*Sinv*	*Amel*	*Nvit*	*Tcas*	*Bmor*	*Dmel*
	Social Hymenoptera											
Recognition												
GNBP	2	2	2	2	4	2	4	4	3	3	4	5
PGRP	4	4	4	4	6	5	3	4	13	8	12	13
FREP	1	1	1	1	1	1	1	2	1	7	3	14
Galectins	2	3	3	3	3	3	2	3	3	3	4	6
SCR-B	8	9	12	8	9	9	10	10	12	16	13	13
SCR-C	1	1	1	1	1	1	1	1	1	1	1	4
CTL*	11	11	12	12	12	13	10	10	28	16	21	33
TEP	3	3	3	3	3	3	3	3	3	4	3	13
Modulation												
cSP†	8	6	9	14	8	7	4	18 (10)	13	48 (20)	15	46 (22)
Serpin	8	7	11	8	7	7	9	7	12	31	26	30
Effectors												
Abaecin	1	1	0	1	0	2	1	1	3	0	0	0
defensin	1	1	2	1	1	5	2	2	6	4	1	1
hymenoptaecin	1	1	1	1	1	1	1	1	2	0	0	0
other AMPs*	0	0	0	0	0	0	0	0	33	8	29	20
lysozyme	4	5	2	1	2	2	4	3	2	4	4	13
PPO*	1	1	1	1	1	1	1	1	2	3	2	3

† Numbers in parenthesis are gene family counts obtained using the same HMMER profile as used for the ants and *N. vitripennis*.
* Denotes significant difference in family size between eusocial (n=8) and solitary (n=4) species using a Mann-Whitney U-test (FDR < 0.1).

Supplemental Table 7. GO terms enriched among ant ultra-conserved elements.

Name	GOID	P	Q(GW)	Q(FG)	Count
regulation of multicellular organismal process	GO:0051239	3.86E–11	0.326	0.076	283
cell development	GO:0048468	1.07E–09	0.310	0.086	323
organ morphogenesis	GO:0009887	5.81E–09	0.314	0.073	274
cellular component movement	GO:0006928	6.55E–09	0.329	0.056	209
localization of cell	GO:0051674	6.55E–09	0.329	0.056	209
cell surface receptor linked signaling pathway	GO:0007166	8.11E–09	0.305	0.085	318
generation of neurons	GO:0048699	1.10E–08	0.321	0.063	234
calcium ion binding	GO:0005509	1.28E–08	0.326	0.056	211
regulation of transcription, DNA-dependent	GO:0006355	1.88E–08	0.294	0.105	393
regulation of RNA metabolic process	GO:0051252	1.94E–08	0.293	0.108	403
neurogenesis	GO:0022008	2.06E–08	0.317	0.064	240
cell adhesion	GO:0007155	3.04E–08	0.330	0.050	187
biological adhesion	GO:0022610	3.51E–08	0.330	0.050	187
neuron differentiation	GO:0030182	3.99E–08	0.320	0.058	218
integral to plasma membrane	GO:0005887	6.27E–08	0.310	0.068	255
intrinsic to plasma membrane	GO:0031226	6.66E–08	0.310	0.069	257
embryonic development	GO:0009790	7.46E–08	0.309	0.070	261
cell morphogenesis	GO:0000902	9.97E–08	0.325	0.050	188
transcription, DNA-dependent	GO:0006351	1.11E–07	0.288	0.111	417
molecular transducer activity	GO:0060089	1.49E–07	0.297	0.086	323
signal transducer activity	GO:0004871	1.49E–07	0.297	0.086	323
RNA biosynthetic process	GO:0032774	1.59E–07	0.287	0.112	419
synapse	GO:0045202	2.21E–07	0.335	0.041	153
cellular component morphogenesis	GO:0032989	3.98E–07	0.316	0.054	202
cell motility	GO:0048870	7.09E–07	0.327	0.042	158
cell-cell signaling	GO:0007267	8.39E–07	0.315	0.051	192
system process	GO:0003008	8.99E–07	0.292	0.087	324
neuron development	GO:0048666	9.62E–07	0.320	0.046	173
locomotion	GO:0040011	1.13E–06	0.317	0.048	180
transmembrane receptor activity	GO:0004888	1.13E–06	0.324	0.043	159
cell projection organization	GO:0030030	1.60E–06	0.317	0.047	176
regulation of cell differentiation	GO:0045595	1.90E–06	0.325	0.040	150
receptor activity	GO:0004872	2.23E–06	0.300	0.065	245
embryonic morphogenesis	GO:0048598	2.91E–06	0.326	0.038	142
regulation of neuron differentiation	GO:0045664	2.91E–06	0.374	0.020	76
cell migration	GO:0016477	3.30E–06	0.327	0.037	139
cell morphogenesis involved in differentiation	GO:0000904	3.30E–06	0.327	0.037	139
cell fate commitment	GO:0045165	3.32E–06	0.361	0.023	87
transcription regulator activity	GO:0030528	3.84E–06	0.285	0.091	342
transcription factor activity	GO:0003700	5.08E–06	0.301	0.059	220
regulation of neurogenesis	GO:0050767	5.17E–06	0.361	0.022	84
central nervous system development	GO:0007417	6.69E–06	0.313	0.045	168

cell projection	GO:0042995	7.12E–06	0.297	0.064	238
muscle system process	GO:0003012	7.19E–06	0.362	0.021	80
cell proliferation	GO:0008283	7.21E–06	0.295	0.067	251
regulation of biological quality	GO:0065008	8.56E–06	0.278	0.109	408
heart development	GO:0007507	9.06E–06	0.341	0.027	102
pattern specification process	GO:0007389	9.12E–06	0.330	0.032	121
positive regulation of developmental process	GO:0051094	1.10E–05	0.311	0.044	165
metal ion binding	GO:0046872	1.18E–05	0.260	0.226	844
sequence-specific DNA binding	GO:0043565	1.31E–05	0.313	0.042	156
neuron projection	GO:0043005	1.51E–05	0.318	0.037	140
tissue development	GO:0009888	1.54E–05	0.301	0.053	199
muscle contraction	GO:0006936	1.60E–05	0.361	0.020	75
regulation of cell communication	GO:0010646	1.61E–05	0.295	0.061	229
regulation of cell development	GO:0060284	1.71E–05	0.345	0.024	91
receptor binding	GO:0005102	1.84E–05	0.313	0.041	152
chordate embryonic development	GO:0043009	2.14E–05	0.318	0.036	136
embryonic development ending in birth or egg hatching	GO:0009792	2.31E–05	0.313	0.040	148
cell morphogenesis involved in neuron differentiation	GO:0048667	2.34E–05	0.323	0.033	123
regulation of nervous system development	GO:0051960	2.42E–05	0.345	0.023	87
neuron projection development	GO:0031175	2.45E–05	0.318	0.036	133
regulation of system process	GO:0044057	2.55E–05	0.328	0.030	111
cell junction	GO:0030054	2.65E–05	0.317	0.036	135
regulation of localization	GO:0032879	2.82E–05	0.307	0.043	162
regulation of cellular component organization	GO:0051128	3.12E–05	0.311	0.040	150
ion binding	GO:0043167	3.14E–05	0.258	0.230	861
locomotory behavior	GO:0007626	3.17E–05	0.337	0.025	95
transmission of nerve impulse	GO:0019226	3.24E–05	0.309	0.041	155
cation binding	GO:0043169	3.35E–05	0.258	0.228	852
sensory organ development	GO:0007423	3.44E–05	0.317	0.035	132
cell part morphogenesis	GO:0032990	3.48E–05	0.319	0.034	126
transcription factor binding	GO:0008134	3.63E–05	0.308	0.041	155
neuron projection morphogenesis	GO:0048812	3.80E–05	0.322	0.032	119
cell projection morphogenesis	GO:0048858	3.80E–05	0.320	0.033	122
ion channel activity	GO:0005216	4.21E–05	0.333	0.026	97
muscle organ development	GO:0007517	4.32E–05	0.328	0.028	105
behavior	GO:0007610	4.77E–05	0.305	0.043	161
synapse part	GO:0044456	4.98E–05	0.327	0.028	105
protein amino acid phosphorylation	GO:0006468	5.11E–05	0.300	0.048	178
cytoskeletal part	GO:0044430	5.28E–05	0.291	0.059	222
regulation of transcription from RNA polymerase II promoter	GO:0006357	5.29E–05	0.296	0.052	196
channel activity	GO:0015267	5.50E–05	0.330	0.026	99
passive transmembrane transporter activity	GO:0022803	5.50E–05	0.330	0.026	99
substrate-specific channel activity	GO:0022838	6.01E–05	0.330	0.026	98

neurological system process	GO:0050877	6.38E–05	0.287	0.067	249
cation channel activity	GO:0005261	6.85E–05	0.351	0.019	71
brain development	GO:0007420	7.21E–05	0.316	0.033	122
protein kinase activity	GO:0004672	7.53E–05	0.307	0.039	145
phosphate metabolic process	GO:0006796	8.25E–05	0.281	0.078	291
phosphorus metabolic process	GO:0006793	8.88E–05	0.280	0.078	291
cell fate determination	GO:0001709	9.24E–05	0.404	0.011	40
synaptic transmission	GO:0007268	9.35E–05	0.310	0.036	133
negative regulation of developmental process	GO:0051093	9.77E–05	0.301	0.043	160
gated channel activity	GO:0022836	1.02E–04	0.339	0.021	80
blood vessel morphogenesis	GO:0048514	1.03E–04	0.341	0.021	78
vasculature development	GO:0001944	1.05E–04	0.331	0.024	90
axonogenesis	GO:0007409	1.08E–04	0.322	0.028	104
phosphorylation	GO:0016310	1.09E–04	0.283	0.068	256
death	GO:0016265	1.11E–04	0.283	0.070	260
regulation of cell proliferation	GO:0042127	1.13E–04	0.299	0.045	167
cell death	GO:0008219	1.19E–04	0.283	0.069	258
regulation of anatomical structure morphogenesis	GO:0022603	1.19E–04	0.338	0.021	80
regulation of cellular component movement	GO:0051270	1.23E–04	0.356	0.017	62
negative regulation of gene expression	GO:0010629	1.25E–04	0.306	0.037	139
negative regulation of cell differentiation	GO:0045596	1.26E–04	0.344	0.019	72
response to external stimulus	GO:0009605	1.29E–04	0.290	0.056	208
regulation of cell projection organization	GO:0031344	1.40E–04	0.385	0.012	45
cytoskeletal protein binding	GO:0008092	1.43E–04	0.308	0.035	132
regulation of neuron projection development	GO:0010975	1.45E–04	0.394	0.011	41
blood vessel development	GO:0001568	1.45E–04	0.330	0.024	88
regulation of locomotion	GO:0040012	1.49E–04	0.354	0.017	62
extracellular region	GO:0005576	1.57E–04	0.281	0.072	268

Supplemental Table 8. Structure, length, and location of EvoFold predicted conserved RNA structures within the conserved elements (CEs) from the seven-way genome alignments. A total of 3318 putative RNA structures were identified, 1223 of which were considered high-confidence. A database of specific structures may be viewed with the following URL: *http://people.binf.ku.dk/jeanwen/data/ants/*.

	All predictions (n=3318)	**High confidence (n=1223)**
Structure shapes		
Hairpin	2980	1049
Clover shaped	15	7
Complex shaped	154	73
Y shaped	169	94
Structure length		
Short (≤ 15 bp)	3007	1118
Long (> 15 bp)	311	105
Structure location		
3' UTR	132	42
5' UTR	47	13
Intron	745	283
Intergenic	2003	694
CDS	391	191

Supplemental Table 9. Top 25 enriched GO categories of the EvoFold predicted conserved structural RNAs. While the P-values reported are not corrected for multiple testing, they were estimated using the TopGO "elim" method, which reduces redundancy in the GO analysis. The full table (less significant hits and the Molecular Function and Cell Component hierarchies), as well as tables of enriched categories for both the high-confidence structure set and for only intronic or UTR structures, can be browsed at the following URL: *http://people.binf.ku.dk/jeanwen/data/ants/evofold/*.

Accession	**GO Biological Process**	**P-value**
GO:0048870	cell motility	0.000042
GO:0007476	imaginal disc-derived wing morphogenesis	0.000093
GO:0009792	embryo development ending in birth or egg hatching	0.001300
GO:0007380	specification of segmental identity, head	0.001500
GO:0035295	tube development	0.001600
GO:0048598	embryonic morphogenesis	0.001700
GO:0007166	cell surface receptor linked signaling pathway	0.001700
GO:0016477	cell migration	0.002000
GO:0030182	neuron differentiation	0.002300
GO:0048675	axon extension	0.003400
GO:0007631	feeding behavior	0.004200
GO:0002251	organ or tissue specific immune response	0.004300
GO:0048468	cell development	0.004300
GO:0001745	compound eye morphogenesis	0.004400
GO:0010927	cellular component assembly involved in morphogenesis	0.005400
GO:0035220	wing disc development	0.005400
GO:2000026	regulation of multicellular organismal development	0.006700
GO:0007431	salivary gland development	0.006800
GO:0048569	post-embryonic organ development	0.007300
GO:0006928	cellular component movement	0.008100
GO:0009628	response to abiotic stimulus	0.009200
GO:0010556	regulation of macromolecule biosynthetic process	0.009500
GO:0006935	chemotaxis	0.009600
GO:0035071	salivary gland cell autophagic cell death	0.009900
GO:0030902	hindbrain development	0.009900

Supplemental Table 10. Distribution of GC/AT compositional-domain lengths.

Order	**Species**	**Number of compositional domains**				**Total number**	**Assembly size (Mb)***
		1—10 kb (%)	10–100 kb (%)	100 kb–1 Mb (%)	1–10 Mb (%)		
Hymenoptera	*A. cephalotes*	32,887 (88)	4,042 (11)	399 (1.1)	2 (0.01)	37,330	281
	A. echinatior	36,282 (88)	4,411 (11)	372 (0.9)	0 (0)	41,065	289
	S. invicta	54,878 (92)	4,376 (7)	294 (0.5)	0 (0)	59,548	311
	P. barbatus	35,604 (90)	3,637 (9)	192 (0.5)	0 (0)	39,433	220
	C. floridanus	30,714 (88)	3,804 (11)	202 (0.6)	0 (0)	34,720	221
	L. humile	31,978 (89)	3,755 (10)	188 (0.5)	0 (0)	35,921	213
	H. saltator	61,849 (94)	3,985 (6)	144 (0.2)	0 (0)	65,978	281
	A. mellifera	42,006 (91)	3,944 (9)	150 (0.3)	0 (0)	46,100	230
	N. vitripennis	51,064 (93)	3,870 (7)	72 (0.1)	0 (0)	55,006	238
Coleoptera	*T. castaneum*	15,432 (90)	1,535 (9)	183 (1)	3 (0.02)	17,153	131
Diptera	*A. gambiae*	36,941 (92)	3,185 (8)	231 (0.6)	0 (0)	40,357	223
	D. melanogaster	12,297 (85)	1,973 (14)	154 (1.1)	0 (0)	14,424	120

* Number of non-ambiguous nucleotides in the assembly

Supplemental Table 11. Spearman's rank correlations between bisulfite-seq fractional methylation levels and CpG O/E of genes in *Solenopsis invicta* males according to orthology data. CpG O/E is a strong predictor of empirically obtained levels of methylation for conserved single copy genes, but not multi copy genes.

Number of taxa with orthology	**Transcription unit (exons and introns)**		**Coding sequence**	
	Single copy	**Multi copy**	**Single copy**	**Multi copy**
1 [a]	0.1029 *		0.0791 ^{NS}	
2	0.056 ^{NS}	0.153 ^{NS}	0.011 ^{NS}	0.1555 *
3	0.061 ^{NS}	0.172 *	0.0257 ^{NS}	0.0917 ^{NS}
4	–0.023 ^{NS}	0.118 ^{NS}	–0.0658 ^{NS}	0.1004 ^{NS} 4
5	–0.175 ****	0.147 *	–0.1624 ****	0.036 ^{NS}
6	–0.456 ****	0.016 ^{NS}	–0.3678 ****	–0.0042 ^{NS}
7	–0.669 ****	–0.291 ****	–0.5545 ****	–0.2644 ****

^{NS} $P > 0.05$, * $P < 0.05$, **** $P < 0.0001$

[a] Orphan genes have not been annotated as single copy or multi copy

Supplemental Table 12. Bisulfite-seq fractional methylation levels (mCG/CG_{all}) of coding sequences in *Solenopsis invicta* males according to different CpG O/E cutoffs (among genes with single copy orthologs present in seven species). CpG O/E values differ among genes according to empirically obtained levels of methylation.

Cutoff	**CpG o/e cutoff value**	**Direction**	**Number of genes**	**Mean fractional methylation (± SEM)**
Mean CpG o/e	1.088	above	2525	0.041 (±0.001)
Mean CpG o/e + 0.5 SD	1.193	above	1822	0.029 (±0.001)
Mean CpG o/e + 1 SD	1.298	above	792	0.020 (±0.002)
Mean CpG o/e + 2 SD	1.507	above	70	0.018 (±0.005)
Mean CpG o/e	1.088	below	2525	0.193 (±0.004)
Mean CpG o/e – 0.5 SD	0.983	below	1546	0.249 (±0.005)
Mean CpG o/e – 1 SD	0.878	below	866	0.311 (±0.008)
Mean CpG o/e – 2 SD	0.669	below	229	0.436 (±0.018)

Supplemental Table 13. Gene Ontology functional enrichment for putatively methylated genes according to the presence of lower than mean coding sequence CpG O/E values in all seven ant taxa (among single copy orthologs present in seven species). P-value calculated by the Benjamini and Hochberg FDR method.

Accession	GO Biological Process	Fold enrichment in class	P-value
GO:0010467	gene expression	1.54	2.32E –05
GO:0016070	RNA metabolic process	1.75	3.93E –05
GO:0006461	protein complex assembly	2.48	1.78E –04
GO:0070271	protein complex biogenesis	2.48	1.78E –04
GO:0009987	cellular process	1.12	2.69E –04
GO:0044260	cellular macromolecule metabolic process	1.28	2.85E –04
GO:0044237	cellular metabolic process	1.22	3.18E –04
GO:0006367	transcription initiation from RNA polymerase II promoter	3.16	3.18E –04
GO:0006366	transcription from RNA polymerase II promoter	2.67	3.59E –04
GO:0006352	transcription initiation	3.08	4.77E –04
GO:0043933	macromolecular complex subunit organization	1.95	9.42E –04
GO:0044249	cellular biosynthetic process	1.38	0.001
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.39	0.001
GO:0032774	RNA biosynthetic process	2.33	0.001
GO:0034645	cellular macromolecule biosynthetic process	1.46	0.001
GO:0065003	macromolecular complex assembly	2.02	0.001
GO:0006351	transcription, DNA-dependent	2.34	0.001
GO:0009059	macromolecule biosynthetic process	1.45	0.001
GO:0009058	biosynthetic process	1.35	0.002
GO:0034641	cellular nitrogen compound metabolic process	1.31	0.007
GO:0043170	macromolecule metabolic process	1.19	0.012
GO:0006412	Translation	1.76	0.013
GO:0044085	cellular component biogenesis	1.46	0.037
GO:0006974	response to DNA damage stimulus	2.19	0.038
GO:0006281	DNA repair	2.25	0.039
GO:0006807	nitrogen compound metabolic process	1.26	0.040

Supplemental Table 14. Gene Ontology functional enrichment for putatively unmethylated genes according to the presence of higher than mean coding sequence CpG O/E values in all seven ant taxa (among single copy orthologs present in seven species). P-value calculated by the Benjamini and Hochberg FDR method.

Accession	**GO Biological Process**	**Fold enrichment in class**	**P-value**
GO:0048856	anatomical structure development	1.35	5.61E –08
GO:0048731	system development	1.38	5.74E –08
GO:0032501	multicellular organismal process	1.27	6.82E –08
GO:0030154	cell differentiation	1.41	8.70E –07
GO:0007166	cell surface receptor linked signal transduction	1.64	1.05E –06
GO:0048869	cellular developmental process	1.38	1.67E –06
GO:0048468	cell development	1.48	1.72E –06
GO:0007165	signal transduction	1.48	4.60E –06
GO:0007275	multicellular organismal development	1.27	5.03E –06
GO:0007399	nervous system development	1.47	5.11E –06
GO:0009653	anatomical structure morphogenesis	1.35	6.07E –06
GO:0032502	developmental process	1.24	1.09E –05
GO:0022008	Neurogenesis	1.52	2.32E –05
GO:0048699	generation of neurons	1.51	7.89E –05
GO:0048513	organ development	1.33	1.28E –04
GO:0007186	G-protein coupled receptor protein signaling pathway	1.93	1.80E –04
GO:0009887	organ morphogenesis	1.44	1.83E –04
GO:0007610	Behavior	1.64	1.90E –04
GO:0007411	axon guidance	1.95	4.43E –04
GO:0007409	Axonogenesis	1.73	9.19E –04
GO:0030182	neuron differentiation	1.48	0.001
GO:0048666	neuron development	1.50	0.003
GO:0009888	tissue development	1.46	0.004
GO:0007155	cell adhesion	1.76	0.005
GO:0022610	biological adhesion	1.76	0.005
GO:0030030	cell projection organization	1.45	0.005
GO:0065007	biological regulation	1.14	0.005
GO:0006928	cell motion	1.56	0.006
GO:0007626	locomotory behavior	1.78	0.007
GO:0051239	regulation of multicellular organismal process	1.52	0.007
GO:0000902	cell morphogenesis	1.38	0.009
GO:0030534	adult behavior	2.00	0.010
GO:0042221	response to chemical stimulus	1.56	0.011
GO:0003008	system process	1.41	0.014
GO:0050794	regulation of cellular process	1.14	0.015

GO:0007167	enzyme linked receptor protein signaling pathway	1.74	0.018
GO:0050789	regulation of biological process	1.13	0.021
GO:0050877	neurological system process	1.39	0.024
GO:0006468	protein amino acid phosphorylation	1.49	0.029
GO:0008407	bristle morphogenesis	2.33	0.032
GO:0048812	neuron projection morphogenesis	1.46	0.032
GO:0032989	cellular component morphogenesis	1.31	0.040
GO:0031175	neuron projection development	1.45	0.040
GO:0048858	cell projection morphogenesis	1.41	0.043
GO:0048667	cell morphogenesis involved in neuron differentiation	1.44	0.044
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	1.80	0.045

Supplemental Table 15. Species-level Gene Ontology functional enrichment (Benjamini and Hochberg FDR P-value < 0.05) for putatively methylated genes with low coding sequence CpG O/E according to a cutoff of one SD below the mean (among single copy orthologs present in seven species). Functional enrichment associated with lineage-specific methylation does not appear to deviate qualitatively from patterns of functional enrichment observed for genes that are methylated in all ants (Supplemental Table 13).

Accession	**GO Biological Process**	***Acep***	***Aech***	***Sinv***	***Pbar***	***Cflo***	***Lhum***	***Hsal***
GO:0009987	cellular process	×	×	×	×	×	×	×
GO:0010467	gene expression	×	×	×	×	×	×	×
GO:0016070	RNA metabolic process	×	×	×	×	×	×	×
GO:0044237	cellular metabolic process	×		×	×	×	×	×
GO:0044260	cellular macromolecule metabolic process	×		×	×	×	×	×
GO:0006461	protein complex assembly	×	×		×	×		×
GO:0065003	macromolecular complex assembly	×	×		×	×		×
GO:0070271	protein complex biogenesis	×	×		×	×		×
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	×		×	×		×	
GO:0006351	transcription, DNA-dependent				×	×	×	×
GO:0006366	transcription from RNA polymerase II promoter				×	×	×	×
GO:0006396	RNA processing	×			×	×	×	
GO:0006412	Translation			×		×	×	×
GO:0009058	biosynthetic process				×	×	×	×
GO:0032774	RNA biosynthetic process				×	×	×	×
GO:0034660	ncRNA metabolic process	×			×	×	×	
GO:0043933	macromolecular complex subunit organization		×		×	×		×
GO:0044249	cellular biosynthetic process				×	×	×	×
GO:0006352	transcription initiation				×	×		×
GO:0006367	transcription initiation from RNA polymerase II promoter				×	×		×
GO:0006399	tRNA metabolic process				×	×	×	
GO:0006807	nitrogen compound metabolic process	×			×		×	
GO:0008152	metabolic process				×		×	×
GO:0009059	macromolecule biosynthetic process				×		×	×
GO:0022613	ribonucleoprotein complex biogenesis			×		×	×	
GO:0034641	cellular nitrogen compound metabolic process	×			×		×	
GO:0034645	cellular macromolecule biosynthetic process				×		×	×
GO:0043170	macromolecule metabolic process				×		×	×

GO:0034470	ncRNA processing		×	×	
GO:0042254	ribosome biogenesis		×	×	
GO:0044238	primary metabolic process	×		×	
GO:0000022	mitotic spindle elongation				×
GO:0000279	M phase				×
GO:0007051	spindle organization				×
GO:0007052	mitotic spindle organization				×
GO:0022403	cell cycle phase				×
GO:0051231	spindle elongation				×

Supplemental Table 16. Genomic coordinates of novel miRNA conserved across ant species. (*) denotes miRNA conserved in all Hymenoptera. (#) denotes Aculeata-specific miRNA. (Table continues on next page.)

ID	*A. cephalotes*	*A. echinatior*	*S. invicta*	*P. barbatus*	*C. floridanus*
1*	Scaffold00009 298195:298279:–	scaffold12 255692:255776:–	Si_gnF.scaffold00206 1720505:1720587:–	scf7180000350292 641659:641741:–	scaffold1446 156946:157025:–
2*	Scaffold00011 1845047:1845144:+	scaffold327 41375:41465:+	Si_gnF.scaffold03557 488207:488300:–	scf7180000350374 732174:732267:–	scaffold493 518629:518713:–
3*	Scaffold00015 526022:526097:–	scaffold99 3513473:3513548:+	Si_gnF.scaffold03776 230425:230499:+	scf7180000350301 242704:242777:–	scaffold1141 50615:50690:–
4*	Scaffold00004 4279845:4279928:+	scaffold527 608040:608121:–	Si_gnF.scaffold06788 607865:607946:+	scf7180000350270 1181197:1181276:–	scaffold710 227009:227091:+
5*	Scaffold00076 524670:524762:–	scaffold39 883344:883434:–	Si_gnF.scaffold03949 26730:26821:+	scf7180000349939 285288:285373:+	scaffold620 8460:8550:–
6#	Scaffold00074 740977:741075:–	scaffold293 677141:677238:–	Si_gnF.scaffold06735 264755:264844:–	scf7180000350310 231490:231569:–	scaffold56 38359:38422:+
7#	Scaffold00018 2590478:2590566:+	scaffold140 2023033:2023122:–	Si_gnF.scaffold06792 466767:466850:+	scf7180000350289 171709:171792:–	scaffold487 1168024:1168090:+
8#	Scaffold00022 176789:176876:–	scaffold88 511352:511439:–	Si_gnF.scaffold05788 503497:503574:–	scf7180000349970 430080:430158:–	scaffold316 1103878:1103956:–
9#	Scaffold00022 176741:176822:–	scaffold88 511304:511394:–	Si_gnF.scaffold05788 503438:503531:–	scf7180000349970 430030:430123:–	scaffold316 1103820:1103905:–
10#	Scaffold00053 1140357:1140443:–	scaffold182 221834:221922:+	Si_gnF.scaffold03294 1453491:1453581:–	scf7180000350222 916084:916171:–	scaffold409 110435:110520:+
11	Scaffold00015 521926:522004:–	scaffold99 3517560:3517640:+	Si_gnF.scaffold03776 234724:234803:+	scf7180000350301 241823:241901:–	scaffold1141 45851:45928:–
12	Scaffold00019 2075903:2075977:+	scaffold485 554961:555033:–	Si_gnF.scaffold01122 1466372:1466445:+	scf7180000350360 189174:189248:+	scaffold1001 129602:129677:–
13	Scaffold00019 476958:477050:–	scaffold294 521785:521877:+	Si_gnF.scaffold06340 384039:384131:+	scf7180000349958 1228972:1229064:+	scaffold1221 174724:174815:–
14	Scaffold00005 5055741:5055775:+	scaffold288 1120031:1120094:+	Si_gnF.scaffold04519 2131:2203:+		scaffold486 352289:352360:–
15	Scaffold00001 822537:822628:–	scaffold220 1135597:1135687:–	Si_gnF.scaffold02694 1023049:1023140:–	scf7180000350285 165772:165862:+	scaffold437 196606:196694:+
16	Scaffold00076 525036:525113:–	scaffold39 883701:883779:–	Si_gnF.scaffold03949 26440:26515:+	scf7180000349939 284934:285012:+	scaffold620 9595:9671:–
17	Scaffold00013 3558919:3558992:+	scaffold342 709632:709710:+	Si_gnF.scaffold01426 531424:531497:+	scf7180000350303 1673055:1673128:+	scaffold351 990202:990279:+
18	Scaffold00034 3145482:3145580:–	scaffold50 2574310:2574385:–	Si_gnF.scaffold07124 468044:468141:–	scf7180000350378 2077483:2077580:+	scaffold372 172336:172416:+
19	Scaffold00008 22919:23005:–	scaffold309 1033552:1033638:–	Si_gnF.scaffold06738 1922595:1922674:+	scf7180000350035 171683:171768:–	scaffold263 930276:930361:–
20	Scaffold00026 1206522:1206599:–	scaffold310 219997:220074:+	Si_gnF.scaffold06207 3421852:3421935:–	scf7180000350381 2220671:2220743:–	scaffold1357 1047:1118:–
21	Scaffold00049 228114:228198:+	scaffold758 168792:168876:+	Si_gnF.scaffold06735 1026055:1026141:–	scf7180000349994 401180:401265:–	scaffold407 1874944:1875027:+
22	Scaffold00015 4711497:4711578:+	scaffold283 1722440:1722522:–	Si_gnF.scaffold06899 487914:487990:–	scf7180000350119 179725:179806:–	scaffold1533 86168:86242:+
23	Scaffold00011 1922519:1922590:–	scaffold327 115890:115961:–	Si_gnF.scaffold03557 396658:396715:+	scf7180000350374 676678:676739:+	
24	Scaffold00088 497795:497871:–	scaffold39 385467:385545:–	Si_gnF.scaffold05266 353110:353186:+	scf7180000350371 729092:729168:–	scaffold770 87970:88047:– scaffold770 93493:93570:–
25	Scaffold00002 498215:498307:–		Si_gnF.scaffold00514 2519367:2519456:–	scf7180000349954 628272:628364:+	scaffold1826 717945:718036:–
26	Scaffold00021 84547:84629:–	scaffold574 412982:413064:–	Si_gnF.scaffold06735 1597341:1597423:–	scf7180000349994 928090:928172:–	scaffold407 1304940:1305021:+
27	Scaffold00055 319865:319943:+	scaffold18 939140:939218:–	Si_gnF.scaffold05266 161013:161078:–	scf7180000349880 101508:101590:+	scaffold1702 155864:155929:+
28	Scaffold00015 4711359:4711446:+	scaffold283 1722572:1722660:–	Si_gnF.scaffold06899 488038:488123:–	scf7180000350119 179857:179941:–	scaffold1533 86038:86116:+
29	Scaffold00032 2852256:2852333:+	scaffold501 525447:525521:+	Si_gnF.scaffold01506 313600:313680:+	scf7180000350194 706179:706249:–	scaffold638 184201:184276:–

Supplemental Table 16. Continued.

ID	*L. humile*	*H. saltator*
1[*]	scf7180001004419 138805:138885:+	scaffold186 184082:184161:+
2[*]	scf7180001004917 825128:825213:–	scaffold155 695798:695882:–
3[*]	scf7180001004993 1158427:1158503:+	scaffold2362 45806:45882:–
4[*]	scf7180001004868 673945:674024:–	scaffold2261 400323:400405:–
5[*]	scf7180001004958 44294:44380:+	scaffold355 180638:180696:+
6[#]	scf7180001005010 75718:75789:+	scaffold846 496341:496429:+
7[#]	scf7180001004973 699038:699104:–	scaffold896 110866:110932:–
8[#]	scf7180001005077 16454:16531:–	scaffold1632 34670:34756:+
9[#]	scf7180001005077 16395:16484:–	scaffold1632 34714:34806:+
10[#]	scf7180001004914 983486:983570:–	scaffold829 1445687:1445772:–
11	scf7180001004993 1161141:1161219:+	scaffold2362 41385:41464:–
12	scf7180001004913 318801:318875:–	scaffold125 1186156:1186236:–
13	scf7180001004905 218388:218479:+	scaffold31 451061:451152:–
14	scf7180001004659 1548615:1548687:+	scaffold120 133411:133482:–
15	scf7180001004947 294745:294831:–	scaffold1545 18020:18101:–
16	scf7180001004958 43895:43973:+	scaffold355 180456:180530:+
17	scf7180001004429 567555:567632:–	scaffold427 997677:997754:–
18	scf7180001004715 877351:877431:+	scaffold284 273801:273881:+
19	scf7180001005010 600479:600566:+	scaffold710 107334:107414:+
20		scaffold220 172582:172655:+
21	scf7180001004456 483993:484075:+	
22	scf7180001004952 137094:137167:–	
23	scf7180001004917 740996:741065:+	
24	scf7180001004940 2630045:2630122:–	
25	scf7180001004986 193126:193198:–	
26		
27		
28		
29		

Supplemental Table 17. Transcription of ant conserved elements. See Materials and Methods for data sets and analysis parameters.

		C. floridanus			***H. saltator***		
Sample	**CEs**	**Transcripts**	**Expr. CE %**	**CE overlap %**	**Transcripts**	**Expr. CE %**	**CE overlap %**
5' 50kb	620,248	16,042	2.6%	91.0%	17,231	2.8%	92.0%
5' 10kb	190,210	5,540	2.9%	89.7%	5,392	2.8%	90.4%
5' 2kb	52,268	1,655	3.2%	87.8%	1,366	2.6%	88.4%
5' UTR	34,375	932	2.7%	87.6%	718	2.1%	88.6%
CpG island	64,016	2,284	3.6%	90.3%	1,570	2.5%	88.2%
Exon	763,028	20,430	2.7%	86.6%	20,288	2.7%	88.3%
Intron	96,431	1,877	1.9%	90.6%	1,831	1.9%	91.7%
3' UTR	33,878	1,636	4.8%	91.5%	1,636	4.8%	91.2%
3' 2kb	39,988	2,820	7.1%	91.9%	2,923	7.3%	91.8%
miRNA	63	63	100.0%	73.6%	53	84.1%	65.1%
rRNA	4	3	75.0%	100.0%	1	25.0%	100.0%
snRNA	47	47	100.0%	97.6%	7	14.9%	82.3%
tRNA	134	127	94.8%	94.5%	84	62.7%	90.6%
TEprotein	4,260	1,499	35.2%	91.8%	802	18.8%	93.2%
Transposon	15,506	1,239	8.0%	90.8%	919	5.9%	90.6%
Other	596,098	17,169	2.9%	90.7%	16,571	2.8%	92.1%
Total		**73,363**			**71,392**		
Non-exonic		**52,933**			**51,104**		
Intergenic		**45,510**			**45,053**		

Supplemental Table 18. GO terms enriched among 118 genes nearest to conserved transcribed CpG islands.

Name	GOID	P	Q(GO)	Q(FG)	Count
regulation of primary metabolic process	GO:0080090	0.0	0.02	0.314	37
regulation of neuron differentiation	GO:0045664	0.0	0.067	0.076	9
regulation of nervous system development	GO:0051960	0.0	0.057	0.085	10
regulation of macromolecule metabolic process	GO:0060255	0.0	0.02	0.305	36
negative regulation of cellular process	GO:0048523	0.0	0.025	0.203	24
regulation of metabolic process	GO:0019222	0.0	0.019	0.322	38
regulation of neurogenesis	GO:0050767	0.0	0.057	0.076	9
regulation of biological process	GO:0050789	0.0	0.016	0.466	55
regulation of cellular biosynthetic process	GO:0031326	0.0	0.02	0.271	32
regulation of biosynthetic process	GO:0009889	0.0	0.02	0.271	32
negative regulation of metabolic process	GO:0009892	0.0	0.032	0.136	16
regulation of cellular metabolic process	GO:0031323	0.0	0.019	0.305	36
regulation of macromolecule biosynthetic process	GO:0010556	0.0	0.021	0.263	31
regulation of cellular component organization	GO:0051128	0.0	0.038	0.11	13
negative regulation of biological process	GO:0048519	0.0	0.023	0.212	25
negative regulation of steroid hormone receptor signaling pathway	GO:0033144	0.0	0.375	0.025	3
biological regulation	GO:0065007	0.0	0.015	0.483	57
multicellular organismal development	GO:0007275	0.0	0.018	0.305	36
developmental process	GO:0032502	0.0	0.018	0.339	40
regulation of protein metabolic process	GO:0051246	0.0	0.033	0.119	14
cellular developmental process	GO:0048869	0.0	0.022	0.22	26
regulation of multicellular organismal process	GO:0051239	0.0	0.028	0.144	17
regulation of cell development	GO:0060284	0.0	0.05	0.076	9
regulation of cellular process	GO:0050794	0.0	0.016	0.441	52
cell differentiation	GO:0030154	0.0	0.022	0.212	25
regulation of cell differentiation	GO:0045595	0.0	0.037	0.102	12
negative regulation of macromolecule metabolic process	GO:0010605	0.0	0.031	0.127	15
cellular macromolecule biosynthesis	GO:0034961	0.0	0.018	0.305	36
negative regulation of cell differentiation	GO:0045596	0.0	0.055	0.068	8
anatomical structure development	GO:0048856	0.0	0.019	0.28	33
macromolecule biosynthesis	GO:0043284	0.0	0.018	0.305	36
multicellular organismal process	GO:0032501	0.0	0.017	0.347	41
regulation of steroid hormone receptor signaling pathway	GO:0033143	0.0	0.3	0.025	3
regulation of gene expression	GO:0010468	0.0	0.019	0.254	30
negative regulation of cellular metabolic process	GO:0031324	0.0	0.031	0.119	14
translation elongation factor activity	GO:0003746	0.0	0.133	0.034	4
negative regulation of signal transduction	GO:0009968	0.0	0.056	0.059	7
regulation of cellular protein metabolic process	GO:0032268	0.0	0.033	0.102	12
system development	GO:0048731	0.0	0.019	0.254	30
response to reactive oxygen species	GO:0000302	0.0	0.067	0.051	6
protein complex binding	GO:0032403	0.0	0.055	0.059	7
regulation of transcription from RNA polymerase II promoter	GO:0006357	0.0	0.029	0.119	14
cellular macromolecule biosynthetic process	GO:0034645	0.0	0.017	0.305	36
regulation of translation	GO:0006417	0.0	0.053	0.059	7
macromolecule biosynthetic process	GO:0009059	0.0	0.017	0.305	36
negative regulation of developmental process	GO:0051093	0.0	0.031	0.102	12

Supplemental Table 19. TF gene loci are broadly conserved among insects.

			Gene copy number																											
	Name	**TF**	**Dsim**	**Dsec**	**Dmel**	**Dyak**	**Dere**	**Dana**	**Dpse**	**Dper**	**Dwil**	**Dmoj**	**Dvir**	**Dgri**	**Agam**	**Aaeg**	**Cqui**	**Bmor**	**Tcas**	**Acep**	**Aech**	**Sinv**	**Pbar**	**Cflo**	**Lhum**	**Hsal**	**Amel**	**Nvit**	**Phum**	**Dpul**
1	abdominal A	ABD_A	1	1	1	1	1	1	1	1	1	1	1	1	1?	?	?	1	1	1	?	?	1	?	1	1	1	1	1	?
2	abdominal B	ABD_B	1	1	1	1	1	1	1	1	1	1	1	1	1	1?	1	1	1	1	1	1	?	1	1	1	?	1	1	?
3	bric-a-brac1	BAB1	1	1	1	1	1	1	1	1	1	1	1	1	?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	broad	BR	?	?	1	?	1	1	1	1	1	1	?	1	1	1	1	1	1	?	1	?	?	1	?	1	2?	1	1	1
5	cyclic-AMP response element binding protein	CREB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	deformed	DFD	1	1	1	1	1	1	1	1	?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	dorsal	DL	1	1	1	1	1	1	1	1	1	?	1	?	1	2?	3?	1	1	1?	1	1	1	1	1?	1	2?	1	1	1
8	DNA replication-related element factor	DREF	1	1	1	1	1	1	1	1	1	1	1	1	1	?	?	?	?	1	1	1	1	1?	1	1?	1	1?	1	1
9	Ecdysone-induced protein 74EF	EIP74EF	1	1	1	1	1	1	1	1	1	1	1	1	1	?	1	1	1	1	?	1	1	1	1	?	1	1	1	1
10	empty spiracles	EMS	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	engrailed	EN	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	even skipped	EVE	?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	?
13	grainy head	GRH	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	?	1	1	1	1	1	1	1	1	?	1
14	giant	GT	?	1	1	1	1	?	?	?	?	?	?	?	?	?	?	?	?	1	1	1	?	1	1	1	1	?	?	?
15	hairy	H	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	?
16	huckebein	HKB	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	?	2?	1	1	1
17	E(spl) region transcript m5	HLHm5	1	1	1	1	1	1	1	1	1	1?	1	1	?	?	?	?	?	1	1	1	1	1	1	1	1	?	?	?
18	mothers against decapentaplegic	MAD	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1?	1	1	1	?	1	1	1	1	2?	1	1
19	ocelliless	OC	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	?	1?	1	1	1	1	1	1	1	1	1	1?	1
20	pleiohomeotic	PHO	1	1	1	1	1	1	1	1	1	1	1	1	?	?	1	1	1	1	1	1	1	2?	1	?	1	1	1	1
21	paired	PRD	1	1	1	1	1	1	1	1	1	1	1	1	1	?	1?	1	1	?	1	1?	?	1	?	1?	1?	1	1	1
22	scalloped	SD		1	1		1	1	1	1	1	1	1	1	1	2?	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	schnurri	SHN	1	1	1	1	1	1	?	1?	?	1	?	1	?	?	1	1	1	1	1	1	1	1	1		1	1	1	1?
24	sloppy paired 1	SLP1	1	1	1	1	1	1	?	?	1	1	1	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
25	snail	SNA	1	1	1	1	1	1	1	1	1	1	1	1	1?	1?	1?	1?	1?	1?	1?	1?	1?	1?	1?	1?	1?	1?	1?	1?
26	serpent	SRP	1?	1?	1	1	1	1	1	1	1	1	1?	1	?	?	?	?	1	?	?	?	?	1	?	?	1	?	?	1
27	tailless	TLL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
28	trithorax-like	TRL	1	1	1	1	1	1	1	?	1	1	1	1	1	1	1	1	1	1	1	?	1	1	1	1	1	1	?	?
29	ultraspiracle	USP	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	?	1	1	1?	1	1	1	1	1
30	ventral nervous system defective	VND	1?	1	1	1	1	1	1	1	1	1	?	1	1?	?	?	?	1	1?	1?	1	1?	1	1?	1	1	?	?	1

Supplemental Table 20. Gene Ontology category enrichment for positively selected genes in ants. False discovery rates are calculated based on randomizations (100 tests with permutation of the scores attributed to genes). Categories with FDR < 20% are reported.

GO ID	**Onto-logy**	**GO name**	**p**	**FDR**
GO:0000313	CC	organellar ribosome	1.43E–10	0
GO:0006120	BP	mitochondrial electron transport, NADH to ubiquinone	1.05E–09	0
GO:0005759	CC	mitochondrial matrix	1.63E–09	0
GO:0005762	CC	mitochondrial large ribosomal subunit	1.09E–07	0.0025
GO:0005746	CC	mitochondrial respiratory chain	4.53E–07	0.003333333
GO:0005747	CC	mitochondrial respiratory chain complex I	1.26E–06	0.003333333
GO:0008137	MF	NADH dehydrogenase (ubiquinone) activity	3.21E–05	0.012857143
GO:0005763	CC	mitochondrial small ribosomal subunit	0.000178056	0.047272727
GO:0008038	BP	neuron recognition	0.000228635	0.047272727
GO:0008344	BP	adult locomotory behavior	0.000819594	0.086153846
GO:0042254	BP	ribosome biogenesis	0.001112847	0.099333333
GO:0003735	MF	structural constituent of ribosome	0.001159211	0.099333333
GO:0044459	CC	plasma membrane part	0.001581373	0.115625
GO:0006508	BP	proteolysis	0.002209659	0.143529412
GO:0006412	BP	translation	0.002519768	0.1455
GO:0016491	MF	oxidoreductase activity	0.002753425	0.1455
GO:0004872	MF	receptor activity	0.002823959	0.1455
GO:0055114	BP	oxidation-reduction process	0.003831834	0.164583333
GO:0008237	MF	metallopeptidase activity	0.003874749	0.164583333
GO:0061134	MF	peptidase regulator activity	0.004628046	0.178461538
GO:0002520	BP	immune system development	0.005283503	0.185666667
GO:0048534	BP	hemopoietic or lymphoid organ development	0.005283503	0.185666667
GO:0016616	MF	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	0.00531916	0.185666667
GO:0016836	MF	hydro-lyase activity	0.005464343	0.185666667

Supplemental Table 21. TFs and genes associated with TFBS evolution in eusocial genomes. Highlighted rows indicate significant TFs. Significance assessed using Mann-Whitney U-test (FDR < 0.25). NS, not significant.

		Genome-wide (n=6673)					Genes with promoter CEs (n=1966)				
	Overall	Total			Prop.	Prop.	Total			Prop.	Prop.
TF	change, P	sig.	Gains	Losses	gain	1 copy	sig.	Gains	Losses	gain	1 copy
ABD_A	+, 6.6e–2	111	81	30	0.73	0.883	97	60	37	0.62	0.928
ABD_B	+, 1.4e–2	74	65	9	0.88	0.838	24	21	3	0.88	0.958
ANTP	+, 1e–10	0	0	0	NA	NA	0	0	0	NA	NA
AP	NS	0	0	0	NA	NA	0	0	0	NA	NA
BAB1	+, 1e–10	0	0	0	NA	NA	1243	1187	56	0.95	0.893
BCD	NS	0	0	0	NA	NA	0	0	0	NA	NA
BR	NS	0	0	0	NA	NA	7	7	0	1.00	0.857
BRK	–, 1.2e–5	0	0	0	NA	NA	0	0	0	NA	NA
BYN	–, 7e–3	0	0	0	NA	NA	0	0	0	NA	NA
CAD	NS	0	0	0	NA	NA	0	0	0	NA	NA
CPE	–, 5.1e–2	0	0	0	NA	NA	0	0	0	NA	NA
CREB	+, 6.9e–2	436	295	141	0.68	0.876	188	123	65	0.65	0.920
D	–, 5.9e–6	0	0	0	NA	NA	0	0	0	NA	NA
DEAF1	–, 1.1e–3	0	0	0	NA	NA	0	0	0	NA	NA
DFD	+, 4.6e–3	20	18	2	0.90	0.9	12	9	3	0.75	0.833
DL	–, 2.4e–6	0	0	0	NA	NA	3	1	2	0.33	1.000
DPE	NS	0	0	0	NA	NA	0	0	0	NA	NA
DREF	NS	0	0	0	NA	NA	266	135	131	0.51	0.887
EIP74EF	–, 5.1e–2	0	0	0	NA	NA	201	18	183	0.09	0.910
EMS	+, 1.6e–11	513	424	89	0.83	0.858	132	109	23	0.83	0.932
EN	+, 8.5e–4	0	0	0	NA	NA	182	108	74	0.59	0.901
EVE	NS	18	16	2	0.89	0.833	5	3	2	0.60	1.000
FKH	+, 2e–12	0	0	0	NA	NA	0	0	0	NA	NA
FTZ	NS	0	0	0	NA	NA	0	0	0	NA	NA
GRH	–, 1e–10	0	0	0	NA	NA	1208	24	1184	0.02	0.897
GT	+, 1e–10	0	0	0	NA	NA	458	389	69	0.85	0.910
H	NS	305	254	51	0.83	0.862	127	104	23	0.82	0.906
HB	NS	0	0	0	NA	NA	0	0	0	NA	NA
HKB	–, 1e–10	0	0	0	NA	NA	1330	27	1303	0.02	0.898
HLHm5	+, 1.1e–2	15	13	2	0.87	0.933	4	4	0	1.00	1.000
KNI	–, 1.6e–3	0	0	0	NA	NA	0	0	0	NA	NA
KR	–, 4.8e–6	0	0	0	NA	NA	0	0	0	NA	NA
MAD	NS	55	48	7	0.87	0.855	7	7	0	1.00	1.000
MED	–, 1.5e–9	0	0	0	NA	NA	0	0	0	NA	NA
NUB	+, 1.7e–2	174	129	45	0.74	0.879	0	0	0	NA	NA
OC	NS	0	0	0	NA	NA	4	3	1	0.75	0.750
OVO	NS	0	0	0	NA	NA	0	0	0	NA	NA
PAN	NS	0	0	0	NA	NA	0	0	0	NA	NA
PHO	–, 5.8e–8	0	0	0	NA	NA	4	1	3	0.25	0.750
PRD	NS	0	0	0	NA	NA	4	4	0	1.00	0.750
SD	NS	3	3	0	1.00	1	5	4	1	0.80	1.000
SHN	–, 1e–10	0	0	0	NA	NA	937	25	912	0.03	0.904
SLBO	–, 3.8e–2	0	0	0	NA	NA	0	0	0	NA	NA
SLP1	–, 1.1e–12	0	0	0	NA	NA	855	31	824	0.04	0.904
SNA	–, 1e–10	0	0	0	NA	NA	1083	38	1045	0.04	0.887
SRP	+, 4.8e–2	0	0	0	NA	NA	179	104	75	0.58	0.872
TBP	+, 1e–10	0	0	0	NA	NA	0	0	0	NA	NA
TIN	–, 3.8e–14	0	0	0	NA	NA	0	0	0	NA	NA
TLL	–, 5.9e–10	0	0	0	NA	NA	730	22	708	0.03	0.908

TOP2	NS	319	279	40	0.87	0.828	0	0	0	NA	NA
TRL	NS	77	74	3	0.96	0.909	33	31	2	0.94	0.939
TTK	–, 2.3e–3	0	0	0	NA	NA	0	0	0	NA	NA
TWI	–, 6.8e–7	0	0	0	NA	NA	0	0	0	NA	NA
UBX	NS	0	0	0	NA	NA	0	0	0	NA	NA
USP	–, 4.8e–2	6	6	0	1.00	1	3	2	1		1.000
VND	–, 7.8e–16	0	0	0	NA	NA	1065	20	1045	0.02	0.908
VVL	NS	0	0	0	NA	NA	0	0	0	NA	NA
Z	–, 3.6e–3	0	0	0	NA	NA	0	0	0	NA	NA
ZEN	NS	0	0	0	NA	NA	0	0	0	NA	NA

Supplemental Table 22. GO analysis of genes exhibiting TFBS evolution in eusocial genomes, by comparing 1793 OrthoDB groups (genes) showing significant promoter-associated social evolution and conservation against 9236 genes with conservation (non-significant eusocial changes). GO terms pass FDR < 0.01.

Name	GOID	P	Q(GO)	Q(FG)	Count
cellular_component	GO:0005575	7.31E–20	0.200	0.791	1609
molecular_function	GO:0003674	1.36E–19	0.200	0.794	1614
Cell	GO:0005623	5.55E–17	0.201	0.742	1508
cell part	GO:0044464	5.55E–17	0.201	0.742	1508
binding	GO:0005488	1.55E–16	0.203	0.692	1406
biological_process	GO:0008150	2.78E–16	0.200	0.759	1544
cellular process	GO:0009987	7.33E–14	0.202	0.667	1355
protein binding	GO:0005515	1.01E–11	0.208	0.494	1005
intracellular	GO:0005622	2.83E–09	0.198	0.628	1276
intracellular part	GO:0044424	7.78E–09	0.198	0.612	1245
regulation of biological process	GO:0050789	9.75E–09	0.209	0.387	787
biological regulation	GO:0065007	1.48E–08	0.207	0.412	838
cellular metabolic process	GO:0044237	2.57E–08	0.202	0.496	1008
metabolic process	GO:0008152	6.45E–08	0.200	0.542	1101
primary metabolic process	GO:0044238	6.54E–08	0.201	0.503	1022
signal transduction	GO:0007165	1.50E–07	0.223	0.202	410
membrane	GO:0016020	1.68E–07	0.207	0.367	746
regulation of cellular process	GO:0050794	2.65E–07	0.207	0.365	743
cell communication	GO:0007154	2.75E–07	0.219	0.228	463
estrogen biosynthetic process	GO:0006703	3.22E–07	0.909	0.005	10
testosterone 17-beta-dehydrogenase activity	GO:0050327	3.22E–07	0.909	0.005	10
estrogen metabolic process	GO:0008210	3.42E–07	0.846	0.005	11
organelle	GO:0043226	1.08E–06	0.198	0.533	1084
cytoplasm	GO:0005737	1.76E–06	0.199	0.476	967
intracellular organelle	GO:0043229	2.12E–06	0.197	0.531	1079
anatomical structure development	GO:0048856	2.61E–06	0.217	0.206	418
macromolecule metabolic process	GO:0043170	3.65E–06	0.201	0.416	845
multicellular organismal development	GO:0007275	3.67E–06	0.214	0.225	458
macromolecule metabolism	GO:0043283	4.19E–06	0.201	0.410	833
cellular macromolecule metabolism	GO:0034960	4.37E–06	0.203	0.380	773
cellular macromolecule metabolic process	GO:0044260	4.49E–06	0.203	0.383	779
multicellular organismal process	GO:0032501	4.86E–06	0.210	0.267	543
membrane-bounded organelle	GO:0043227	5.34E–06	0.198	0.493	1002
retinoic acid receptor activity	GO:0003708	5.88E–06	0.769	0.005	10
retinoid-X receptor activity	GO:0004886	5.88E–06	0.769	0.005	10
thyroid hormone receptor activator activity	GO:0010861	5.88E–06	0.769	0.005	10
thyroid hormone receptor coactivator activity	GO:0030375	5.88E–06	0.769	0.005	10
protein complex	GO:0043234	6.47E–06	0.216	0.200	406
intracellular membrane-bounded organelle	GO:0043231	7.12E–06	0.198	0.491	998
macromolecular complex	GO:0032991	7.75E–06	0.211	0.243	494
system development	GO:0048731	7.81E–06	0.217	0.189	384
cellular biosynthetic process	GO:0044249	9.90E–06	0.207	0.295	599

developmental process	GO:0032502	1.03E–05	0.210	0.255	519
biosynthetic process	GO:0009058	1.26E–05	0.206	0.301	611
molecular transducer activity	GO:0060089	1.43E–05	0.235	0.100	203
signal transducer activity	GO:0004871	1.43E–05	0.235	0.100	203
estradiol 17-beta-dehydrogenase activity	GO:0004303	1.72E–05	0.714	0.005	10
receptor activator activity	GO:0030546	1.72E–05	0.714	0.005	10
steroid dehydrogenase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	GO:0033764	1.72E–05	0.714	0.005	10
nitrogen compound metabolic process	GO:0006807	1.75E–05	0.205	0.310	631
cell differentiation	GO:0030154	2.69E–05	0.223	0.137	278
cellular component organization	GO:0016043	3.03E–05	0.211	0.221	450
organ development	GO:0048513	4.20E–05	0.219	0.152	310
cellular developmental process	GO:0048869	4.22E–05	0.221	0.142	288
transferase activity	GO:0016740	4.56E–05	0.219	0.151	306

Supplemental Table 23. GO analysis of 292 genes exhibiting concentrated regulatory rewiring of multiple TFs.

Name	**GOID**	**P**	**Q(GW)**	**Q(FG)**	**Count**
estrogen biosynthetic process	GO:0006703	1.21E–11	0.500	0.032	10
testosterone 17-beta-dehydrogenase activity	GO:0050327	1.21E–11	0.500	0.032	10
estrogen metabolic process	GO:0008210	2.25E–11	0.476	0.032	10
vitamin D receptor binding	GO:0042809	3.47E–11	0.333	0.039	12
retinoic acid receptor activity	GO:0003708	4.04E–11	0.455	0.032	10
retinoid-X receptor activity	GO:0004886	4.04E–11	0.455	0.032	10
thyroid hormone receptor activator activity	GO:0010861	4.04E–11	0.455	0.032	10
thyroid hormone receptor coactivator activity	GO:0030375	4.04E–11	0.455	0.032	10
estradiol 17-beta-dehydrogenase activity	GO:0004303	6.98E–11	0.435	0.032	10
receptor activator activity	GO:0030546	6.98E–11	0.435	0.032	10
steroid dehydrogenase activity, acting on the CH-OH group of donors	GO:0033764	6.98E–11	0.435	0.032	10
thyroid hormone receptor binding	GO:0046966	2.68E–10	0.286	0.039	12
retinoic acid receptor binding	GO:0042974	4.71E–10	0.370	0.032	10
steroid dehydrogenase activity	GO:0016229	4.71E–10	0.370	0.032	10
androgen metabolic process	GO:0008209	2.27E–09	0.323	0.032	10
ligand-dependent nuclear receptor transcription coactivator activity	GO:0030374	5.55E–09	0.256	0.035	11
nuclear hormone receptor binding	GO:0035257	2.15E–08	0.181	0.042	13
receptor regulator activity	GO:0030545	2.72E–08	0.256	0.032	10
hormone receptor binding	GO:0051427	4.98E–08	0.169	0.042	13
ligand-dependent nuclear receptor activity	GO:0004879	5.89E–08	0.208	0.035	11
cellular_component	GO:0005575	1.16E–07	0.030	0.771	239
hormone biosynthetic process	GO:0042446	1.20E–07	0.222	0.032	10
molecular_function	GO:0003674	2.18E–07	0.030	0.774	240
transcription coactivator activity	GO:0003713	5.99E–07	0.105	0.055	17
steroid biosynthetic process	GO:0006694	3.23E–06	0.129	0.039	12
cellular hormone metabolic process	GO:0034754	5.72E–06	0.149	0.032	10
binding	GO:0005488	9.49E–06	0.030	0.668	207
protein heterodimerization activity	GO:0046982	1.08E–05	0.085	0.055	17
cell	GO:0005623	1.14E–05	0.030	0.710	220
cell part	GO:0044464	1.14E–05	0.030	0.710	220
receptor signaling protein activity	GO:0005057	1.16E–05	0.114	0.039	12
oxidoreductase activity, acting on the CH-OH group of donors	GO:0016616	2.68E–05	0.092	0.045	14
hormone metabolic process	GO:0042445	3.51E–05	0.111	0.035	11
protein binding	GO:0005515	3.77E–05	0.032	0.477	148
transcription cofactor activity	GO:0003712	3.87E–05	0.074	0.058	18
transcription activator activity	GO:0016563	5.63E–05	0.067	0.065	20
succinate-CoA ligase activity	GO:0004774	7.10E–05	0.400	0.013	4
regulation of hormone levels	GO:0010817	7.36E–05	0.089	0.042	13
biological_process	GO:0008150	1.19E–04	0.029	0.713	221
ammonia ligase activity	GO:0016211	1.61E–04	0.333	0.013	4
oxidoreductase activity, acting on CH-OH group of donors	GO:0016614	1.68E–04	0.078	0.045	14

Supplemental References

E. W. Abrams, M. S. Vining, D. J. Andrew. Trends Cell Biol. 13, 247-254 (2003).

M. Adams et al., Science 287, 2185 (2000).

H. Akashi, Genetics 136, 927 (1994).

A. Alexa, A. Rahnenfhrer, T. Lengauer, Bioinformatics 22, 1600 (2006).

J. B. Amaral, G. M Machado-Santelli. Micron 39, 1222-1227 (2008). M. Anisimova, Z. Yang, Mol. Biol. Evol. 24, 1219 (2007).

M. Ashburner et al., Nat. Genet. 25, 25 (2000).

Y. Benjamini, Y. Hochberg, J. R. Stat. Soc. Ser. B Stat. Methodol. 57, 289 (1995).

G. Bernardi, Gene 241, 3 (2000).

L. Bromham, R. Leys, Mol. Biol. Evol. 22, 1393 (2005).

Bovine Genome Sequencing and Analysis Consortium. Science 324, 522 (2009).

N. Bray, L. Pachter, Genome Res. 14, 693 (2004).

M. Bulmer, Genetics 129, 897 (1991).

J. Castresana, Mol. Biol. Evol. 17, 540 (2000).

F. Chen, A. J. Mackey, C. J. Stoeckert, Jr., D. S. Roos, Nucleic Acids Res. D363 (2006).

A. Conesa et al., Bioinformatics 21, 3674 (2005).

T. De Bie, J. P. Demuth, N. Cristianini, M. W. Hahn, Bioinformatics 22, 1269 (2006).

D. A. Drummond, C. O. Wilke, Cell 134, 341 (2008).

R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis. Probabilistic Models of Proteins and Nucleic Acids (Cambridge University Press, 1998).

L. Duret, Curr. Opin. Genet. Dev. 12, 640 (2002).

S. R. Eddy, Bioinformatics 14, 755 (1998).

E. Elhaik, D. Graur, K. Josic, G. Landan. Nucleic Acids Res. 38, e158 (2010).

W. Fletcher, Z. Yang, Mol. Biol. Evol. 27, 2257 (2010).

H. Gingold, Y. Pilpel, Mol. Syst. Biol. 7, 481 (2011).

R. Gouveia-Oliveira, P. Sackett, A. Pedersen, BMC Bioinformatics 8, 312 (2007).

M. Kaneshina, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, Nucleic Acids Res. 40, D109 (2012).

E. F. Kirkness et al., Proc. Natl. Acad. Sci. U.S.A. 107, 12168 (2010).

C. Kosiol, M. Anisimova, Methods Mol. Biol. 856, 113 (2012).

T. M. Hambuch, J. Parsch, Genetics 170, 1691 (2005).

S. B. Hedges, J. Dudley, S. Kumar, Bioinformatics 22, 2971 (2006).

A. Heger, C. P. Ponting, Genetics 177, 1337 (2007).

R. Hershberg, D. A. Petrov, Annu. Rev. Genet. 42, 287 (2008).

D. W. Huang, B. T. Sherman, R. A. Lempicki, Nat. Protocols 4, 44 (2008).

S. Jia., A. Meng, Developmental Dynamics 236, 913 (2007).

G. Jordan, N. Goldman, Mol. Biol. Evol. 29, 1125 (2012).

F. G. Jorgensen, M. H. Schierup, A. G. Clark, Mol. Biol. Evol. 24, 611 (2007).

K. Katoh, K. Misawa, K. Kuma, T. Miyata, Nucleic Acids Res. 30, 3059 (2002).

W. J. Kent, Genome Res. 12, 656 (2002).

F. Krueger, S. R. Andrews, Bioinformatics 27, 1571 (2011).

S. Lall et al., Curr. Biol. 16, 460–471 (2006).

H. Li et al., Bioinformatics 25, 2078 (2009).

S. Q. Le, O. Gascuel, Mol. Biol. Evol. 25, 1307 (2008).

A. Löytynoja, N. Goldman, Science 320, 1632 (2008).

A. Löytynoja, N. Goldman, Proc. Natl. Acad. Sci. U.S.A. 102, 10557 (2005).

P. Markova-Raina, D. Petrov, Genome Res. 21, 863 (2011).

V. Matys et al., Nucleic Acids Res. 31, 374 (2003).

B. Misof, K. Misof, Syst. Biol. 58, 21 (2009).

Y. Moriya, M. Itoh, S. Okuda, A. Yoshizawa, M. Kanehisa, Nucleic Acids Res. 35, W182 (2007).

H. Niculita, J. Billen, J, L. Keller. Arthropod Structure & Development 36, 135– 141 (2007).

C. Notredame, D. G. Higgins, J. Heringa, J. Mol. Biol. 302, 205 (2000).

L. F. Pavon, M. I. Camargo-Mathias. Micron 36, 449–460 (2005).

J. S. Pedersen, G. Bejerano, A.Siepel, K. Rosenbloom, K. Lindblad-Toh et al., PLoS Comput. Biol. 2, e33 (2006).

O. Penn, E. Privman, G. Landan, D. Graur, T. Pupko, Mol. Biol. Evol. 27, 1759 (2010).

D. A. Petrov, D. L. Hartl, Proc. Natl. Acad. Sci. U.S.A. 96, 1475 (1999).

J. B. Plotkin, G. Kudla, Nat. Rev. Genet. 12, 32 (2011).

E. Privman, O. Penn, T. Pupko, Mol. Biol. Evol. 29, 1 (2012).

E. Proux, R. A. Studer, S. Moretti, M. Robinson-Rechavi, Nucleic Acids Res., Database Issue 37, D404 (2008).

J. R. Powell, E. N. Moriyama, Proc. Natl. Acad. Sci. U.S.A. 94, 7784 (1997).

M. Punta et al., Nucleic Acids Res. Database Issue 40, D290 (2002).

S. Richards et al., Nature 452, 949 (2008).

T. B. Sackton, A. G. Clark, BMC Genomics 10, 259 (2009).

T. B. Sackton et al., Nat. Genet. 39, 1461 (2007).

A. Schneider et al., Gen. Biol. Evol. 1, 114 (2009).

Sea Urchin Genome Sequencing Consortium. Science 314, 941 (2006).

A. Stamatakis, Bioinformatics 22, 2688 (2006). D. C. Shields, P. M. Sharp, D. G. Higgins, F. Wright, Mol. Biol. Evol. 5, 704 (1988).

R. R. Sokal, F. J. Rohlf. Biometry, 3rd ed. W. H. Freeman and Company, G12 (1995).

A. Stark et al., Nature 450, 219–232 (2007).

W. J. Swanson, R. Nielsen, Q. Yang, Mol. Biol. Evol. 20, 18 (2003).

H. Tanaka et al., Insect Biochem. Mol. Biol. 38, 1087 (2008).

The Gene Ontology Consortium, Nat. Genet. 25, 25 (2000).

The UniProt Consortium, Nucleic Acids Res. 40, D71 (2012).

N. Tintle, B. Borchers, M. Brown, A. Bekmetjev, BMC Proceedings 3, S96 (2009).

S. Vicario, E. N. Moriyama, J. R. Powell, BMC Evol. Biol. 7, 226 (2007).

I. M. Wallace, O. O'Sullivan, D. G. Higgins, C. Notredame, Nucleic Acids Res. 34, 1692 (2006).

J. H. Werren et al., Science 327, 343 (2010).

W. S. Wong, Z. Yang, N. Goldman, R. Nielsen, Genetics 168, 1041 (2004).

F. Wright, Gene 87, 23 (1990).

Z. Yang, R. Nielsen, N. Goldman, A.-M. K. Pedersen, Genetics 155, 431 (2000).

Z. Yang, Mol. Biol. Evol. 24, 1586 (2007).

Z. Yang, M. dos Reis, Mol. Biol. Evol. 28, 1217 (2011).

J. Zhang, R. Nielsen, Z. Yang, Mol. Biol. Evol. 22, 2472 (2005).

Z. Zou et al., Genome Biol. 8, R177 (2007).