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, et al.

Genome Res. published online May 1, 2013
Access the most recent version at doi:10.1101/gr.155408.113

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

Daniel F. Simola[1,2,20], Lothar Wissler[3,20], Greg Donahue[1,2], Robert M. Waterhouse[4], Martin Helmkampf[5], Julien Roux[6,21], Sanne Nygaard[7], Karl M. Glastad[8], Darren E. Hagen[9,22], Lumi Viljakainen[10], Justin T. Reese[9,22], Brendan G. Hunt[8], Dan Graur[11], Eran Elhaik[12], Evgenia V. Kriventseva[4], Jiayu Wen[13], Brian J. Parker[13], Elizabeth Cash[5], Eyal Privman[6], Christopher P. Childers[9,22], Monica C. Muñoz-Torres[9], Jacobus J. Boomsma[7], Erich Bornberg-Bauer[3], Cameron Currie[14], Christine G. Elsik[9,22], Garret Suen[14], Michael A. D. Goodisman[8], Laurent Keller[6], Jürgen Liebig[5], Alan Rawls[5], Danny Reinberg[15], Chris D. Smith[16], Chris R. Smith[17], Neil Tsutsui[18], Yannick Wurm[6,23], Evgeny M. Zdobnov[4], Shelley L. Berger[1,2,19] and Jürgen Gadau[5,24]

Corresponding Author: Jürgen Gadau

Arizona State University

School of Life Sciences

PO Box 874701

Arizona State University

Tempe, AZ 85287-4701

phone: (480) 965-2349

fax: (480) 965-2519

e-mail: jgadau@asu.edu

Running title: Evolution of social insect genomes

Affiliations

[1]Department of Cell and Developmental Biology, University of Pennsylvania, Philadelphia, PA 19104, USA.

[2]University of Pennsylvania Epigenetics Program, Philadelphia, PA 19104, USA.

[3]Institute for Evolution and Biodiversity, University of Münster, 48149 Münster, Germany.

[4]Department of Genetic Medicine and Development University of Geneva, 1211 Geneva, Switzerland; Swiss Institute of Bioinformatics,1211 Geneva, Switzerland.

[5]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA.

[6]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland.

[7]Centre for Social Evolution, University of Copenhagen, 2100 Copenhagen, Denmark.

[8]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA.

[9]Department of Biology, Georgetown University, Washington, DC 20057, USA.

[10]Department of Biology, University of Oulu, 3000 Oulu, Finland.

[11]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA.

[12]The Johns Hopkins University, Bloomberg School of Public Health, Baltimore, Maryland, 21205, USA; The Johns Hopkins University, School of Medicine, Baltimore, Maryland, 21205, USA.

[13]The Bioinformatics Centre, University of Copenhagen, 2100 Copenhagen, Denmark.

[14]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI 53706, USA.

[15]Department of Biochemistry, New York University, New York, NY 10003, USA; Howard Hughes Medical Institute, New York University, New York, NY 10003, USA.

[16]Center for Computing for Life Science, San Francisco State University, San Francisco, CA 94117, USA.

[17]Department of Biology, Earlham College, Richmond, IN 47374, USA.

[18]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720, USA.

[19]Departments of Biology and Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA.

[20]These authors contributed equally to this work.

[21]Present address: Department of Human Genetics, University of Chicago, Chicago, IL, USA.

[22]Present address: Divisions of Animal and Plant Sciences, University of Missouri, Columbia, MO, USA.

[23]Present address: School of Biology & Chemistry, Queen Mary University of London, London, United Kingdom.

[24]Corresponding author: J.G. (jgadau@asu.edu)

Abstract

Genomes of eusocial insects code for dramatic examples of phenotypic plasticity and social organization. We compared the genomes of seven ants, the honeybee, and various solitary insects to examine whether eusocial lineages share distinct features of genomic organization. Each ant lineage contains ~4,000 novel genes, but only 64 of these genes are conserved among all seven ants. Many gene families have been expanded in ants, notably those involved in chemical communication (e.g., desaturases and odorant receptors). Alignment of the ant genomes revealed reduced purifying selection compared to *Drosophila* without significantly reduced synteny. Correspondingly, ant genomes exhibit dramatic divergence of non-coding regulatory elements, however extant conserved regions are enriched for novel non-coding RNAs and transcription factor binding sites. Comparison of orthologous gene promoters between eusocial and solitary species revealed significant regulatory evolution in both cis (e.g., CREB) and trans (e.g., Forkhead) for nearly 2000 genes, many of which exhibit phenotypic plasticity. Our results emphasize that genomic changes can occur remarkably fast in ants, as two recently diverged leaf-cutter ant species exhibit faster accumulation of species-specific genes and greater divergence in regulatory elements compared to other ants or *Drosophila*. Thus, while the "socio-genomes" of ants and the honeybee are broadly characterized by a pervasive pattern of divergence in gene composition and regulation, they preserve lineage-specific regulatory features linked to eusociality. We propose that changes in gene regulation played a key role in the origins of insect eusociality, whereas changes in gene composition were more relevant for lineage-specific eusocial adaptations.

Introduction

The insect order Hymenoptera encompasses several lineages, including ants, bees, and aculeate wasps, that independently evolved obligate eusociality. Such eusocial lineages are characterized by reproductive division of labor, cooperative brood care, and overlapping generations (Michener 1969). Ants (Formicidae) represent one of the oldest (~130 million years) and most successful exclusively eusocial lineages (Cardinal and Danforth 2011). They have colonized every terrestrial habitat except at the highest latitudes, and they have achieved substantial diversity in both individual and colonial traits. The ecological and evolutionary success of the >15,000 described extant ant species (www.antweb.org) is often attributed to their sociality and

ability to engineer environments, e.g. by building elaborate nests, herding aphids for honeydew, or practicing sustainable agriculture (Crozier and Pamilo 1996, Hölldobler and Wilson 2009).

The genomes of seven ant species, representatives of four major lineages that comprise two-thirds of all ant species, have recently been sequenced and characterized independently (reviewed in Gadau et al. 2012): Jerdon's jumping ant, *Harpegnathos saltator* (Bonasio et al. 2010) (Ponerinae, n=1033 extant species), the globally invasive Argentine ant, *Linepithema humile* (Smith et al. 2011) (Dolichoderinae, n=692), the Florida carpenter ant, *Camponotus floridanus* (Bonasio et al. 2010) (Formicinae, n=2831), and four ants within the hyper-diverse subfamily Myrmicinae (n=6087) the red harvester ant, *Pogonomyrmex barbatus* (Smith et al. 2011), the invasive red imported fire ant, *Solenopsis invicta* (Wurm et al 2011), and two agricultural leaf-cutter species, *Acromyrmex echinatior* and *Atta cephalotes* (Nygaard et al. 2011, Suen et al. 2011). Together with the honeybee *Apis mellifera* (The Honeybee Genome Sequencing Consortium 2006), eight eusocial genomes are now available from two evolutionarily independent lineages. While ants and honeybees are both eusocial Hymenoptera, they differ significantly in many aspects. For instance, ants have wingless and often polymorphic worker castes, enjoy long lifespans for insects, and are descendants of predatory wasps, whereas honeybees only have winged monomorphic workers with limited lifespans and are derived from solitary bees.

Given the remarkable phenotypic diversity among eusocial insects, a key question is to what extent do derived and independent eusocial lineages harbor shared features of genomic organization that enable their eusocial lifestyles (Robinson et al. 2005, Gadau et al. 2012, Fereira et al. 2013). To address this question, we performed a comprehensive characterization of the genomic basis for eusociality, utilizing eight eusocial insect genomes in addition to twenty-two available solitary insect genomes. Our results reveal a variety of lineage-specific changes in both gene composition and gene expression regulation that have facilitated the evolution of eusociality.

Results

Ant genomes harbor thousands of taxonomically restricted genes

Comparison of existing gene annotations from the seven ant genomes showed that while the number of protein-coding genes is only partially explained by genome assembly size (R=0.32; Fig. 1A), the number of orthologous genes shared among these ants varies considerably (3.1-

fold; Supplemental Fig. 1). This suggested that either existing annotations lack many valid genes or ant genomes harbor an abundance of taxonomically restricted genes (TRGs), which have been associated with the evolution of novel functionalities in other systems (reviewed in Khalturin et al. 2009). Previous analyses of individual ant genomes predicted up to 8000 species-specific TRGs (e.g., Smith et al. 2011), in addition to 840 TRGs that are exclusively shared among ants (e.g., Bonasio et al. 2010).

To infer more accurately the origin and abundance of TRGs while minimizing annotation error, we applied two approaches to re-annotate ant and honeybee genomes in terms of gene number and model quality. First, by comparing known protein sequences among species, we identified 3313 genes from 2635 orthologous groups that were missing from existing annotations (Supplemental Figs. 2–4, Supplemental Table 1). Second, we developed a broader approach involving thirty published arthropod genomes and conservative filtering steps and identified 5996 additional, previously missing genes (Supplemental Figs. 5–8, Supplemental Tables 2,3). Thus, our significantly revised genome annotations include 9309 newly annotated genes for the eight eusocial species. These analyses corroborate that the honeybee has an exceptionally low gene number (Fig. 1A): we found only 223 previously missing genes in the honeybee vs. 856 on avg. for ants (Supplemental Fig. 5) and relatively few TRGs compared to other insects. Whether this apparent gene loss is restricted to *Apis* honeybees or is shared with other corbiculate bees will soon be elucidated by ongoing efforts to sequence multiple bee and bumble bee genomes.

Having identified the missing genes from social insect genome annotations, we delineated TRGs for different clades within Hymenoptera and Diptera. Notably, we found 28,581 TRGs that are restricted to Formicidae (ant TRGs), and 42% of these genes appear to be species-specific, given our current taxon sampling (Figs. 1A, B). Thus, we estimate that each ant genome harbors an average of 4,083 TRGs, of which 1,715 appear to be species-specific. Intriguingly, among the remaining, non-species-specific TRGs (i.e., TRGs present in multiple ant genomes), only 64 are present in all seven ant genomes (Supplemental Text 1). These 64 pan-ant TRGs have a median length of 97 aa (25^{th}–75^{th} percentile range of 72–149 aa), are well conserved, and show strong expression support in two ant species (Fig. 2); however, they generally lack caste-specific expression ($n = 3$, FDR < 0.05) and do not encode any known protein domains. These results suggest that a broad “social toolkit” of conserved *de novo* protein-coding genes is not a requirement for eusociality.

Hymenoptera, especially the two leaf-cutter ants, exhibit a faster emergence rate for taxonomically restricted genes than Diptera

Having multiple genomes for two insect orders allowed us to compare rates of TRG emergence between Hymenoptera (n=9) and Diptera (n=15). This comparison revealed that Hymenoptera have both a greater number of TRGs (4658 vs. 3442 average TRGs per species) and a faster TRG emergence rate (31 vs. 14 average TRGs per species per million years) than Diptera (Fig. 1B). Young insect lineages also tend to have a faster TRG emergence rate than older lineages ($R = -0.68$, $P < 0.04$; Fig. 1C). Notably, the two leaf-cutter ant species, which diverged only 8–12 million years (Myr) ago (Schultz and Brady 2008), exhibit the highest number of TRGs (n=6796) and the fastest TRG emergence rate of any sequenced insect lineage, gaining 340 TRGs per species per Myr. In comparison, the *D. melanogaster* subgroup (n=6), also having diverged ~12 Myr ago, gains 115 TRGs per species per Myr. This pattern of rapid but transient expansions of gene content may coincide with dramatic life-history changes associated with early stages of lineage divergence. For example, *A. cephalotes* is distinguished from *A. echinatior* by loss of cuticular actinomycete cultures, physically distinct soldier castes, and claustral colony founding (Fenandez-Marin et al. 2009, Villesen et al 2009). In support of this, most leaf-cutter TRGs (68%) are species-specific.

We hypothesize that these rapid TRG expansions observed for Hymenoptera may be due to differences in the rate of gene loss rather than gene gain. Natural selection is expected to be less efficient at removing superfluous genes from populations with small effective population sizes. Haplo-diploid sex determination and reproductive division of labor in eusocial Hymenoptera reduce effective population size relative to solitary and diploid Diptera (Crozier and Pamilo 1996, Gadau et al. 2012). The lack of a significant codon usage bias in ant genomes compared to *Drosophila* further supports the idea of relatively reduced selection efficiency in eusocial Hymenoptera (Supplemental Text 2, Supplemental Fig. 9, Supplemental Table 4).

Extensive gene family evolution in ants targets cytochromes, desaturases, olfactory receptors, and transcription factors

Gene families are sets of paralogs that often display functional similarity. Expansions or contractions in gene families may correspond to adaptive events coupled to life-history transitions (Ranson et al. 2002). To identify gene families in ants that have expanded or contracted due to natural selection, we examined changes in gene family size along branches of

the phylogeny for 15 insects and estimated the rates of change, using a null model of gene family evolution that reflects the expected divergence due to neutral mutation and genetic drift (Supplemental Methods, Supplemental Table 5) (Hahn et al. 2007). We found hundreds of gene family expansions and contractions along each of the terminal branches (Supplemental Fig. 10A) resulting in significant increases in variation for 281 families ($P < 0.01$). Along the branch leading to Formicidae, eleven significant expansions and nine significant contractions have occurred. Functional annotation of these twenty families showed that 55% possess DNA binding capacity, generally characterized by zinc-finger or helix-loop-helix domains (Supplemental Fig. 11); of these 55% of the expanded and 22% of the contracted families may be involved in regulation of transcription. This suggests that changes in the transcription factor (TF) repertoire were important in the initial stages of ant evolution.

In addition, 96 gene families (34% of significant families) show significantly increased variation within Formicidae with several showing repeated expansions and contractions. This includes the P450 cytochrome superfamily, which has been linked to ecdysteroid metabolism and the detoxification of xenobiotics, and odorant receptor and desaturase genes, which are involved in chemical communication, e.g., caste and colony recognition (Nygaard et al. 2011, Smith et al. 2011, Suen et al. 2011). Repeated changes in these families may reflect adaptations to novel ecological niches (e.g., tropics vs. desert or terrestrial vs. arboreal) and/or changes in social organization (e.g., colony size, mode of reproduction, division of labor). For instance, dietary specialization may conceivably demand novel genes to detoxify or metabolize novel compounds, while existing genes that help process undesirable food items could become unnecessary and therefore lost through genetic drift, e.g., P450 cytochrome pseudogenization in *P. barbatus* (Smith et al. 2011) and loss of metabolic pathways in leaf-cutter ants (Nygaard et al. 2011, Suen et al. 2011). Analogously, the efficiency of chemical stimuli varies between environments and communication systems, necessitating tailored changes to groups of desaturases or olfactory receptors.

Desaturase proteins are central to the production of alkenes, a highly variable component of cuticular hydrocarbons reported to transmit complex signals like nest-mate recognition cues in ants (Martin and Drijfout 2009, Zweden van et al. 2010). Manual annotation and phylogenetic analyses of the $\Delta 9$ and $\Delta 11$ desaturase gene families revealed that five ancestral subfamilies are present in all holometabolous insects (Supplemental Fig. 10B). Two of these subfamilies

experienced multiple episodes of gene expansion at various times during hymenopteran evolution, resulting in 10–23 putatively functional genes in ants, compared to 7 genes in *D. melanogaster*. Large numbers of related but non-functional gene fragments scattered throughout ant genomes also suggest that ants frequently altered their desaturase gene repertoire. For example, the invasive species *L. humile* and *S. invicta* possess 25 and 15 pseudogenes, respectively, suggesting that drastic changes in habitat and social organization following novel habitat invasion might have an immediate effect on genes implicated in social communication. Overall, the elevated number of desaturase genes and their variability in sequence and expression might reflect increased demand for chemical signal diversity used in ant social communication. Consistent with this, gene families presumably involved in the perception of these signals (e.g., olfactory receptors; Smith et al. 2011, Zhou et al. 2012) exhibit similar expansions.

Finally, we reanalyzed immune gene families of social and solitary insects, since the honeybee genome is reported to contain only one-third of the immune genes found in *Drosophila* (The Honeybee Genome Sequencing Consortium 2006), yet these families did not emerge from our gene family evolution analysis. The immune gene complements of eusocial insects did not differ from solitary insects as dramatically as previously proposed, as only three of sixteen immune gene families showed significant changes in size between eusocial and solitary groups (FDR < 0.1; Supplemental Table 6).

Ant genomes exhibit a strong degree of synteny

Comparative analysis of large-scale genome structure in the *Drosophila* clade (63 Myr; Tamura et al. 2004) has revealed broad genome-scale similarity with 66% synteny, less than 3-fold change in total genome size, and little change in chromosome number (4–6 chromosomes (Drosophila 12 Genomes Consortium 2007). While genome size among the seven ants shows a similar 3-fold range, ant species cover twice the evolutionary distance, have more variable chromosome numbers (8–22 chromosomes; Gadau et al. 2012), and exhibit extremely high recombination frequencies (e.g., 71 kb/cM for *Pogonomyrmex rugosus*; Sirviö et al. 2011). These differences between Formicidae and *Drosophila* should favor a rapid decline of synteny in ants. To examine this, we first assembled syntenic blocks of homologous sequences using pairwise alignment of single-copy exons among genomes (Supplemental Figs. 12A–C). While only 5.3% of assembled contigs exhibit significant homology among ants, they comprise 3639 syntenic blocks that cover 65% of each genome on average (min. 57%, *H. saltator*; max. 71%, *A.*

cephalotes), largely consistent with estimates from *Drosophila* (66%) (Drosophila 12 Genomes Consortium 2007). Moreover, average synteny increases to 74% among the four Mymicinae, which have comparable divergence time to the 12 sequenced *Drosophila* species, and to 86% between the two leaf-cutter ants (Supplemental Fig. 12A). Thus, despite fragmented genome assemblies and deep evolutionary history, ant genomes show moderate to strong genomic synteny, especially in gene rich, euchromatic regions. Of course, given the variability in repetitive and total genome size in ants (see Fig. 1A), we do suspect that heterochromatic regions may harbor a greater degree of large-scale structural divergence.

Using a subset of 287 syntenic blocks showing strongest synteny in ants (Supplemental Fig. 13), we evaluated the extent of gene inversions and rearrangements in ants and other insects, using the *A. echinatior* genome as a common reference (the assembly of this species has the greatest N50 contig size; Gadau et al. 2012). These highly syntenic blocks average 300 kb, include 10–15 genes per block, and harbor 8749 genes, including 5202 (91%) single-copy genes found in all seven ant genomes (Supplemental Fig. 1). As expected, both inversions and rearrangements increase with evolutionary distance from *A. echinatior*, although inversions appear to be more common overall (Supplemental Fig. 14). Notably, gene order in the *hox* cluster is identical among ants and is consistent with the ordering in *Drosophila* (Supplemental Fig. 15). Interestingly, all ant species display a lower percentage of gene rearrangements (< 4%) compared to *D. melanogaster* or *A. mellifera* (~7%) and much lower compared to the parasitoid wasp *Nasonia vitripennis* (11%) (Supplemental Fig. 14B). In contrast, *D. melanogaster* shows 2.5-fold more gene inversions compared to *A. mellifera,* ants, and *N. vitripennis,* as expected phylogenetically. Thus, some lineages of Hymenoptera may have accumulated specific kinds of structural divergence, including rearrangements, at a faster rate than *Drosophila* and independent of eusociality.

Ant conserved regions harbor an abundance of regulatory elements and are enriched near neuronal genes

Leveraging the high structural homology among ant genomes, we generated global multiple sequence alignments of all 3639 syntenic blocks (Supplemental Fig. 12D) and identified over 1.7 million conserved elements (CEs), including 424 CEs that span at least 1 kb (Supplemental Fig. 16). After conservatively comparing CEs against all annotated exonic sequences, including those from TRGs, and masking likely untranslated regions (UTRs), nearly half (49%) of CEs appear to

be intergenic (Supplemental Fig. 16D). By counting nucleotides delimited by CEs, we estimate that, on average, 18.6% of each ant genome undergoes purifying selection; similar analysis restricted to Myrmicinae (utilizing 74% of the four genome sequences, compared to 65% for seven ants) yielded approximately the same estimate of 20.7%. Purifying selection is greatest for exons (59%), miRNAs (92%), and tRNAs (28%), i.e., explicitly functional DNA sequences (Supplemental Figs. 16D,E). These estimates may be overestimated, as syntenic blocks are generally depleted for repetitive DNA, a major source of evolutionary variation. Thus, ants appear to exhibit 1.8 to 2.8-fold less purifying selection than *D. melanogaster* (37–53%) (Siepel et al. 2005, Sella et al. 2009) and 3-fold more than *Homo sapiens* (5.5%) (Lindblad-Toh et al 2011), consistent with the hypothesis that eusocial insects have reduced selection efficiency (see above).

When grouping genes by the location of proximal CEs, we recovered several significant functional categories. For example, 2883 genes harboring promoter CEs are enriched for 35 categories pertaining to system, organ, and anatomical structure development, signal transduction, and cell differentiation (FDR < 0.05). Also, the 2721 genes associated with the top 5% of CEs (ultraconserved elements; Supplemental Fig. 16F) are enriched for 113 categories that are not represented among all CEs (FDR < 0.05; Supplemental Table 7), 24 (21%) of which identify nervous system regulation as a key process associated with strongest conservation in ants. This is consistent with the significant differences in brain structure seen in many ants between workers and queens, worker sub-castes, and age-dependent worker task groups (Gronenberg et al. 1996).

To examine whether CEs exhibit conservation beyond primary sequence, we predicted their secondary structures and identified 3318 significant structural CEs (Supplemental Methods). Most of these structures are short (91% < 15 nt), likely forming hairpins, and the majority are located near protein-coding genes (61% ≤ 5 kb, 37% ≤ 1 kb) (Supplemental Table 8). While structural CEs are enriched in likely 3' UTRs ($P < 10^{-15}$), similar to vertebrate genomes (Parker et al. 2011), 60% are intergenic, suggestive of functional small noncoding RNAs (see below). Genes near structural CEs are enriched for functional categories related to development (e.g., imaginal disc-derived wing morphogenesis, specification of segmental identity and head) and cellular dynamics (cell motility, cell migration) (Supplemental Table 9).

These results indicate that DNA sequences conserved among ants identify genes and regulatory processes known to be involved in the transition to and elaboration of eusociality.

Given the abundance of conserved regulatory elements in ant genomes, we examined three mechanisms previously implicated in the regulation of social traits or phenotypic plasticity: direct modification of DNA by methylation (Kucharski et al. 2008, Bonasio et al. 2012, Smith et al. 2012), transcriptional and translational regulation by small RNAs (Pauli et al. 2011), and transcriptional regulation by transcription factors (TFs) (Rebeiz et al. 2011).

Ant genomes exhibit distinct signatures of DNA methylation

DNA methylation has been implicated in regulating gene function in social insects (Glastad et al. 2011). For example, relative depletion of CpG dinucleotides (CpG O/E, a sequence-based signature of DNA methylation) correlates negatively with DNA methylation and distinguishes classes of genes that are differentially expressed between honeybee queens and workers (Elango et al. 2009). We found that all seven ants are distinct from the honeybee in exhibiting unimodal CpG O/E distributions and significantly less CpG depletion over exons genome-wide (i.e., higher mean CpG O/E) (Fig. 3A). Interestingly however, *H. saltator* (a basal ant in our analysis) exhibits moderately greater exonic CpG depletion than the six other ants (Fig. 3A). In contrast, CpG O/E patterns over introns and promoters are broadly similar across Hymenoptera with little CpG depletion. Moreover, ant CEs do not differ substantially from genomic background in terms of CpG depletion (Fig. 3A). These observations are consistent with a coarser, domain-scale analysis of GC-bias (Supplemental Text 3, Supplemental Figs. 17–21, Supplemental Table 10), in which eusocial insects, especially ants, show a negative relationship between genome-wide GC content and exonic bias towards GC-rich domains ($R = –0.93$, $P < 0.0007$) (Supplemental Fig. 21). These analyses confirm that the honeybee is an outlier among hymenopterans in terms of sequence-based patterns of DNA methylation.

To confirm statistical patterns of DNA methylation experimentally, we generated a complete bisulfite-sequence map for *S. invicta.* We found that high levels of DNA methylation (mCG/CG) correspond well with CpG depletion in exons (Spearman's $R = –0.53$, Fig. 3B; Supplemental Tables 11,12), indicating that some genes in ants are distinguished by DNA methylation, similar to the honeybee (see also Bonasio et al. 2012, Smith et al. 2012). Functional analysis of genes putatively methylated in all seven ant genomes (low CpG O/E) revealed enrichment for housekeeping functions, including transcription, translation, and cellular

metabolic function (FDR < 0.05; Supplemental Tables 13–15), as reported for the honeybee (Elango et al. 2009). Next, we computed the average methylation level of genes, grouped by the number of hymenopteran species with orthologs. Interestingly, levels of DNA methylation (mCG/CG and CpG O/E) increase with evolutionary conservation of protein-coding genes, especially when genes have orthologs in more distant Hymenoptera (Fig. 3C). Thus, highly conserved genes are preferentially targeted by DNA methylation. We also note that CpG O/E is a poor predictor of methylation for genes with paralogs (i.e., multi-copy genes), regardless of orthology (Supplemental Figs. 22–24, Supplemental Tables 11,12). Thus, while ants and the honeybee exhibit significantly distinct statistical patterns predictive of DNA methylation, all insects that possess DNA methyltransferases likely methylate exons of highly conserved genes, indicating a common role for DNA methylation independent of eusociality.

Conserved miRNAs and small noncoding RNAs exhibit caste differential expression

Recent investigations have uncovered novel regulatory roles for various classes of noncoding RNAs (e.g., Loewer et al. 2011). We first evaluated known miRNA genes (Bonasio et al. 2010) and found that 63 are highly conserved among ants (avg. ~80% conservation) (Supplemental Fig. 16D). We then re-annotated the seven ant genomes for miRNAs and uncovered 24 novel loci, 18 of which are specific to ants (Supplemental Table 16). Using RNA-seq gene expression data, we confirmed that a total of 115 miRNAs are expressed in *C. floridanus*, including 20 of the novel miRNAs. Several miRNAs show stage and caste-specific expression (Supplemental Fig. 25), and many, including 12 of the novel ant-specific miRNAs, are predicted to share orthologous gene targets among ant species, typically located in 3’ UTRs (data not shown).

We also utilized several small and polyA+ RNA-seq data sets (Bonasio et al. 2010) to identify over 70,000 CEs that overlap transcribed sequences in *C. floridanus* and *H. saltator*. Most transcribed CEs (~64%) are intergenic, including 23,000 CEs located more than 2 kb upstream of protein-coding genes (Supplemental Table 17); these CEs comprise a class of predominantly small, conserved noncoding RNAs. Many conserved noncoding RNAs show moderate differences in expression level between adult worker castes, notably a group of 2290 RNAs that overlap CpG islands and show the highest median expression difference between *C. floridanus* worker castes (Supplemental Figs. 26A,C). Interestingly, CpG islands are also the only non-exonic regions significantly enriched for CEs (P < 0.01; Supplemental Fig. 16D, top), and CpG island RNA expression levels correlate positively with expression levels of the nearest

downstream protein-coding genes ($0.1 \leq R \leq 0.5$), with stronger correlations generally found for CpG islands closer to genes (Supplemental Fig. 26B). Functional analysis of these downstream genes revealed a striking enrichment for regulatory processes targeting neuron differentiation and neurogenesis, steroid hormone signaling, cell differentiation, and gene expression (FDR < 0.1; Supplemental Table 18). These results support the notion that ant CpG islands, which are broadly hypomethylated in ants (our data and Bonasio et al. 2012), may serve a regulatory role, perhaps by harboring enhancer binding sequences or by being preferentially targeted by regulators of chromatin structure (Ramirez-Carrozzi et al. 2009), where small RNAs may be directly or indirectly involved (Kim et al. 2010, Pauli et al. 2011).

Genome-wide evolution of TF binding sites is more divergent within ants than between eusocial and solitary insects

The genomic organization of sequence-specific TF binding sites (TFBSs) represents a profound source for transcriptional regulatory variation potentially utilized during the evolution of insect sociality (Gadau et al. 2012). We evaluated the extent of TF-mediated regulatory evolution by analyzing the genome-wide occurrence and distribution of 59 TFBSs (Supplemental Fig. 27), corresponding to developmentally expressed TFs that exhibit broad conservation in their genomic copy number (i.e., they do not belong to evolutionarily variable gene families; Supplemental Table 19) and non-adaptive coding sequence evolution (93% of branch tests for positive selection in TF loci were not significant) among insects (Supplemental Fig. 28, Supplemental Table 20). Indeed, 26% of all CEs and 48% of ultraconserved CEs harbor at least one TFBS (Supplemental Fig. 29A), confirming that conserved regions are broadly enriched for regulatory elements in ants. In fact, most of these CEs harbor multiple TFBSs (avg. 4.3 TFBSs/CE; Supplemental Fig. 29B), suggesting that preservation of TF co-regulation is also important.

We proceeded to examine whether ant genes that harbor conserved TFBSs in their promoters (0–2kb upstream of ORFs) exhibit evolutionary changes in TF regulation among insects, including the 8 eusocial species and 20 solitary species. For most TFs, the average number of promoter binding sites remains similar across insects, although a few TFs do show overall gains in Hymenoptera (e.g., *Antennapedia, Giant*) (Fig. 4A). To evaluate the extent of divergence of TF regulation, we compared genome-wide TFBS profiles by computing Euclidean distances between species using the number of TFBSs per gene per TF (Fig. 4A, right;

Supplemental Fig. 30). Eusocial species (especially ants) exhibit striking divergence in promoter TFBSs that is greater than their divergence from solitary species. Moreover, the two leaf-cutter ants differ more from each other than the two most divergent flies, and the honeybee shows greater divergence from ants than from flies. This suggests that while many TFBSs remain conserved in ants, the overall architecture of TF-mediated gene regulation is highly variable across insects, especially between convergently evolved eusocial lineages in Hymenoptera.

Ant genomes exhibit similar patterns of cis-regulatory evolution associated with evolutionary increased gene expression plasticity between worker castes

To evaluate whether gains or losses of TFBSs are specifically maintained in eusocial insects but not solitary insects, we examined whether individual gene promoters exhibit changes in TFBS abundance between eusocial (n=8) and solitary (n=20) species. Indeed, we identified nearly 2000 significant genes (FDR < 0.25; Supplemental Tables 21,22; Supplemental Fig. 29C), which represent potential drivers of the genome-wide divergence pattern (see above). This analysis implicates 30 TFs, 16 of which are associated with over 100 significant genes each (Fig. 4B). Most of the significant TFs show either predominant gains (n=7) or losses (n=8) of TFBSs in the eusocial genomes (e.g., SHN and EMS), although a few TFs show more complex patterns of gains and losses (e.g., CREB) (Fig. 4C). We also identified 292 genes that exhibit significant changes in TFBS abundance for multiple TFs, i.e., apparent targets of concentrated cis-regulatory rewiring (Fig. 4D). These 292 genes are enriched for 41 functional categories involved in hormone regulation and transcription factor activity (FDR < 0.05; Supplemental Table 23) and include *nervous wreck*, which regulates synaptic growth and neurotransmission (Coyle et al. 2004), and choline O-acetyltransferase, a key enzyme for synthesizing the neurotransmitter acetylcholine (Fig. 4D). This suggests that the insect neuro-endocrine system has been targeted for regulatory changes during eusocial evolution. Thus, specific TF regulatory proteins conserved among insects exhibit significant divergence in their targets of regulation between eusocial and solitary lineages.

We next assessed whether any genes exhibit regulatory changes specifically in ants, and we found 141 genes with significant TFBS gains in ants but zero predicted binding sites for the majority (≥ 80%) of other species (Supplemental Fig. 31), including honeybee (Supplemental Fig. 32). Intriguingly, evolution of binding sites for CREB, a TF regulator of long-term memory (Ishimoto et al. 2009) and secretory activity (Abrams and Andrew 2005) in insects, affects the

most genes in this analysis (n=31). This suggests that ants may have preferentially altered the binding distribution of CREB as a possible means to achieve gene expression plasticity between specialized castes (see below). Indeed, evolutionary gains/losses of CREB binding sites perfectly discriminate eusocial from solitary species (and ants from the honeybee) in a principle components analysis (Supplemental Fig. 33). Moreover, one of CREB's cofactors, the transcriptional co-activator and histone acetyltransferase CBP, was recently reported to play a role in maintaining caste-specific gene expression patterns in *C. floridanus* (Simola et al. 2013). These results suggest that while many genes show significant cis-regulatory changes specific to ants, the majority (>90%) of genes with significant eusocial-associated regulatory evolution tend to exhibit similar changes in both ants and the honeybee, broadly suggestive of the importance of cis-regulatory changes in the evolutionary origins of or convergence on eusociality.

Since TF binding events regulate gene expression levels, we proceeded to examine whether changes in TFBS abundance between ant species may be indicative of evolutionary increases in gene expression plasticity between castes within a species. Interestingly, genes with the most significant changes in TFBS abundance between eusocial and solitary insects show elevated levels of plasticity in a socially sophisticated ant, *C. floridanus*, compared to *H. saltator*, whose colonies are smaller and exhibit less reproductive division of labor (avg. 0.39 vs. 0.27, $P < 10^{-8}$; Fig. 4B, bottom). Furthermore, different TFs show significantly different levels of plasticity in *C. floridanus* ($P < 0.05$) but not *H. saltator* ($P < 0.13$) (Fig. 4B, bottom), as well as correlations in plasticity between species when grouped by TF (R=0.65, P=0.002) or for individual genes (R=0.30, P=0.003; Supplemental Fig. 34). Gene targets of nine TFs show greater plasticity compared to all ant orthologs ($P < 0.05$; Fig. 4B, asterisks), notably for *empty spiracles* (EMS), which regulates brain morphogenesis and antennae development in *Drosophila* (Cohen and Jürgens 1990), and CREB (see above), which shows the second largest effect (albeit not significant in this analysis). These results suggest that caste-associated gene expression plasticity is a continuously evolving trait in eusocial insects that is partly determined by TFBS abundance (see also Supplemental Fig. 33).

Known eusocial pathways exhibit cis and trans regulatory evolution for several TFs

Finally, we analyzed patterns of regulatory evolution in the salivary gland and wing development regulatory networks, which are known to exhibit phenotypic plasticity between workers and queens and between different worker castes and task groups (Abouheif and Wray 2002, Li and

White 2003). We were struck by the over-representation of TFs associated with eusocial regulatory evolution (see above) among key regulators of these networks (Figs. 5A,B; Supplemental Text 4). In particular, *forkhead* (FKH), an essential regulator of insect salivary glands and labial silk gland development in *Bombyx mori* (Mach et al. 1995), has undergone considerable loss of TFBSs in its own promoter in the eusocial genomes (Fig. 5C)—an example of trans-regulatory evolution that may confer pleiotropic effects. Furthermore, the regulatory network for wing development in ants harbors three TFs, *abdominal A* (*abdA*), *snail* (*sna*), and *engrailed* (*en*), for which we found significant changes in TFBS abundance; *engrailed* was previously shown to be down-regulated in wing discs of workers compared to queens during larval development in two ant species (Abouheif and Wray 2002). These observations support the hypothesis that regulatory rewiring both in promoters of loci encoding TF regulators (trans-regulatory evolution) and in promoters of target loci (cis-regulatory evolution) may serve as an evolutionary driving force enabling morphological and behavioral plasticity in eusocial insects.

Discussion

Given the remarkable phenotypic diversity among eusocial insects, a key question is to what extent do derived and independent eusocial lineages harbor shared features of genomic organization that enable a eusocial lifestyle. Our analysis of eight “socio-genomes” (Robinson et al. 2005), together with several solitary insect genomes, suggests that despite a pattern of broad sequence divergence expected by deep evolutionary history, key changes in gene regulation (both in cis and in trans) may have convergently evolved in the early stages of eusocial evolution, whereas changes in gene composition may have been more important for lineage-specific social and ecological adaptations. Ant intergenic sequences are enriched for regulatory elements, including miRNAs, noncoding RNAs, and TF binding sites that are conserved in both DNA sequence and homologous genomic position. Nearly two thousand genes share similar cis-regulatory changes in eusocial compared to solitary insects, and most of these changes are similar in both eusocial lineages of ants and honeybees. These genes are enriched for neuronal and hormonal functions and implicate a few specific TFs in eusocial evolution, notably EMS and CREB. Changes in TFBS abundance are linked to evolutionary increases in phenotypic plasticity from ants with a simpler social organization to those with more complex social organization. The locus encoding one key TF, *forkhead*, also shows significant regulatory evolution in ants, which has potentially facilitated pleiotropic effects in trans for regulatory networks relevant to

eusociality, e.g. exocrine gland and wing development. Finally, six gene families with putative transcription regulatory activity experienced significant gains in family size during early ant evolution.

Our analysis of evolutionary changes in gene composition in ants contextualizes these results. We found an abundance of taxonomically-restricted genes (TRGs) in ants as well as a higher rate in the emergence of TRGs in ants compared to flies, suggesting functional ties to eusocial adaptations. In other systems, TRGs can comprise 10–33% of a species' protein-coding gene complement and have been linked both to morphological adaptations (Kalthurin et al 2009, Tautz and Loso 2011) and to eusocial traits, including caste differentiation (Kamakura 2011) and complex behavioral repertoires (Johnson and Tsutsui 2011). Importantly, while TRGs likely play important roles involved in the elaboration of social adaptations in individual lineages, TRGs that are critical for early eusocial evolution or the maintenance of eusociality should be conserved in multiple ant genomes. We found 64 ant-specific TRGs which are conserved in all seven ant genomes; however, these genes show limited differential expression between adult worker castes in two different species, at least using pooled tissue data. These novel genes may be relevant for the evolution of eusociality in ants, but their specific functional significance remains unclear.

In conclusion, evolutionary changes in gene regulation seem to dominate our view of the shared genomic features associated with the origins of eusociality. However, the broad spectrum of changes observed in eusocial insect genomes suggests that the origin, maintenance and elaboration of insect eusociality was not necessarily restricted to a small set of genes or regulatory elements. Instead, the organization of eusocial insect genomes appears to harbor sufficient degrees of freedom to allow convergence of higher-order complex traits, such as eusociality, from unique, lineage-specific evolutionary trajectories that involve distinct genes and modes of regulation. Such genomic complexity may be especially engendered by ants, where extreme reproductive divisions of labor resulting from a eusocial lifestyle may effectively reduce the strength of natural selection, thereby facilitating rapid sequence divergence among lineages.

Materials and Methods

Complete materials and methods can be found in the Supplemental Materials.

Assessing homology of known genes among insect species (AntOrthoDB)
The OrthoDB orthology delineation procedure (Waterhouse et al. 2011) was employed to delineate orthologous genes at each radiation along the insect species phylogeny, which includes seven sequenced ant and five outgroup insect species (Supplemental Table 1). OrthoDB has been updated to include these results, along with protein descriptors, Gene Ontology, and InterPro attributes (http://cegg.unige.ch/orthodbants) (Supplemental Fig. 2).

Identification of taxonomically restricted genes
Taxonomically restricted genes (TRGs) were identified as protein-coding genes that lack sequence similarity to annotated proteins outside of a focal taxonomic group (e.g., Formicidae; Fig. 1B), using the official gene sets for 30 arthropod species. Ortholog identification was based on all vs. all BLASTP searches among all annotated proteins (E < 1e–3); For Formicidae TRGs, BLAST hits within ant genomes were ignored. Among TRGs, subsets of genes that only occur within an individual species were identified and denoted as lineage-specific genes (LSGs), given our taxon sampling. Among LSGs, genes were conservatively filtered if the gene (a) matched in other proteomes when low complexity filtering was deactivated, (b) matched proteins in SwissProt taxonomic divisions (except invertebrates) as potential contaminations, or (c) appear not to be lineage-specific if a similar sequence with matching gene model was found in the genome. To control for false LSGs, putative homologs of genes that were predicted in only one of the ant species were screened against the genomes of the other eight Hymenoptera using a custom-built pipeline. True missing genes had to (a) yield a significant BLAST hit in another genome using its predicted peptide sequence as query for TBLASTN (E ≤ 1e–5, low complexity filtering activated), and (b) yield a seemingly functional gene model based on the alignment of the protein query against the genomic sequence. GeneWise v2.4.1 was used to align the protein query against the scaffold in a strand and position-specific manner. Strandedness and position (plus/minus 50kb) were derived from the TBLASTN hit. Only GeneWise models with a score > 35, coverage of the query sequence > 75%, and zero indels (ORF-disrupting frameshift mutations) were accepted as valid gene models. Applying this procedure yielded a total of 2,936 (previously lineage-specific) genes, along with 6,369 newly identified genes that produced valid

gene models in multiple Hymenoptera species (Supplemental Fig. 5). This yields a total of 12,054 LSGs within Formicidae (42.2% of all 28,581 Formicidae TRGs).

Insect phylogeny

The phylogeny shown in Fig. 1 was estimated by maximum likelihood from the concatenated alignment of conserved protein sequences of 2,756 single-copy orthologs across 12 insect species, comprising 792,477 well-aligned amino acids. Sequence alignments were performed with MAFFT (Katoh et al. 2002), conserved cores were selected with Gblocks (Castresana 2000), and the phylogeny was built with PhyML (Guindon et al. 2010). Employing 4,346 single-copy orthologs defined across the seven ant species, *A. mellifera* and *N. vitripennis*, protein lengths were compared to examine the agreement of ant genes with those of their bee and wasp orthologs (Supplemental Fig. 6, Supplemental Table 2).

Conserved Elements

Conserved elements were identified from whole-genome sequence alignments using PhastCons (Siepel et al. 2005). Conserved and non-conserved HMMs were estimated with parameters (`--target-coverage 0.25 --expected-length 12 --estimate-trees`), given a phylogenetic tree (described above) for initialization. Resulting conserved elements (CEs) were filtered to remove any regions whose consensus sequence consisted of gaps only. A subset of ultra-conserved elements (UCEs) were identified as those CEs having length $\geq$ 5nt and LOD/length scores in the top 5th percentile of all CEs, where LOD denotes the log odds ratio of the posterior probability of conservation to that of non-conservation across the nucleotides delimited by the CE. Non-coding conserved RNAs are defined as transcribed ant CEs and do not overlap annotated ant exonic sequences. Specifically, for each CE, associated DNA sequences from each of 7 ant genomes as well as their consensus sequences were locally aligned to all annotated exon sequences from each of the genomes using nucleotide BLAST (E-value of 10). A CE was considered to be exonic if any one associated DNA sequence showed significant alignment to any one exon from any ant genome. CEs were annotated as transcribed if at least 25% of the element overlaps an annotated transcript whose expression level is greater than the 5th percentile of genome-wide expression levels in at least 2 independent biological samples, for each species (see RNA Expression Analysis, below).

DNA Methylation

Normalized CpG dinucleotide content (CpG o/e) was calculated using the equation:

$$CpG\,\frac{o}{e} = \frac{length^2}{length} \times \frac{CpG\ count}{C\ count\ \times\ G\ count}$$

Bisulfite-seq data was obtained using genomic DNA from 6 pooled whole-body haploid males of *S. invicta* from a single colony (NCBI GEO accessions GSE39959). Bisulfite conversion and sequencing were performed by Beijing Genomics Institute (Shenzhen, China). Bisulfite treatment was performed using the ZYMO EZ DNA Methylation-Gold kit (Zymo Research Corporation, Irvine, CA, USA). Sequencing was performed using the Illumina platform. Reads passing quality control were mapped using Bowtie and Bismark (Langmead et al. 2009, Krueger and Andrews 2011). Aligned reads were processed using SAMtools to remove PCR duplicated reads (rmdup) and parsed using custom perl scripts to obtain methylated and unmethylated read counts on a per nucleotide basis. Fractional methylation was calculated for each CpG site with ≥ 3 reads as mCG/CG, where mCG is the number of reads with methylated cytosine at a CpG site (according to bisulfite conversion) and CG is the total number of reads with either unmethylated or methylated (converted and unconverted) cytosine at the same CpG site. Fractional methylation values were averaged across each annotated element with data for ≥ 3 sites (otherwise the element was discarded). Functional enrichment was performed using Gene Ontology annotation of single-copy *D. melanogaster* orthologs of *H. saltator* genes belonging to single-copy 7-ant orthologs, as analyzed by the DAVID functional annotation tool.

MicroRNA identification

Hymenoptera small RNA sequences were downloaded from NCBI Sequence Read Archives (SRA) database (Accession numbers: SRX018737, SRX023147–SRX023156) and searched against the ant genome assemblies using BLASTN. Alignments having ≤ 2 nt mismatches were retained and analyzed using MIREAP with a minimum folding energy of –18 kcal/mol (http://sourceforge.net/projects/mireap/). A miRNA was considered ant-specific when the precursor hairpin and subsequent mature miRNA was present in ≥4/7 ant genomes. This conservative strategy ignores other likely ant-specific miRNAs absent in the current assemblies.

Target analysis was performed after aligning novel, ant-specific miRNAs to the annotated ant genes. As a proxy for 3' UTRs, 750 bp of sequence downstream from each stop codon was extracted. miRNA:mRNA target analysis was performed using miRanda with parameters: `-sc`

`140 -en 20`. Target predictions were considered conserved if at least four ant orthologous ant genes were targeted by the respective miRNA using OrthoDB.

Identification and Analysis of Transcription Factor Binding Sites

Identification of TFBS in insects. Position weight matrices (PWMs) for 56 transcription factors (TFs) were taken from DMMPMM (Bigfoot) and iDMMPMM (Kulakovskiy et al. 2009). PWMs for CREB and the promoter elements CPE and DPE were obtained from Transfac (http://www.gene-regulation.com). Sequence motifs showing significant similarity to each PWM were predicted using pwm_scan (Levy and Hannenhalli 2002). For each of 28 insect species (as described in text), each PWM was scanned across the 2 kb promoters of protein-coding genes whose orthology among species was established by OrthoDB. To obtain stringent TFBS predictions, an empirical score distribution was estimated for each PWM for each species' genome as the set of all nominally significant ($P < 0.05$) motif scores identified within the target set of promoter sequences (`pwm_scan -s 1 -p ln(0.05)`). Candidate binding sites were then selected for each genome as those scoring in the top 0.02% among this set of nominally significant sites (`pwm_scan -s 2 -p ln(1/5000)`), where 1/5000 is the recommended p-value roughly estimating expected frequency (Levy and Hannenhalli 2002).

Between species comparison of number of TFBSs. Each gene of interest was tested for showing a difference in the number of proximal promoter binding sites for each TF in 8 eusocial species compared to 20 solitary species using a two-tailed Mann-Whitney (rank sum) U-test. To control for potential bias due to variation in nucleotide frequency, GC-bias was estimated from the promoter sequences of each genome and used to scale the number of binding sites for each motif x: GC(x)/avg(GC). Only genes having TFBS estimates for at least three eusocial and five solitary (or vice versa) species were used for evaluation. Significant genes were identified using a Benjamini-Hochberg false discovery rate (FDR) of 25%.

Genes with TFBS changes for multiple TFs. Each gene showing a significant change in TFBS abundance between eusocial and solitary genomes was also tested for significant change for multiple TFs by summing the number of 2 kb promoter TFBSs for all TFs and testing for significant difference in total TFBS abundance between eusocial and solitary genomes using a T-test. Genes with $|T|$ values in the top 15% overall were retained.

RNA Expression Analysis

Raw RNA-seq expression data for *C. floridanus* and *H. saltator* were downloaded from NCBI GEO using accession number GSE22680 (Bonasio et al. 2010) or GSE37523 (Simola et al. 2013). Raw sequence reads were mapped using Bowtie+Tophat (Langmead et al. 2009) allowing 1 mismatch and up to 50 alignments per read (`-v 1 -k 50 --best`) and default parameter values otherwise. Expression levels for previously annotated gene models were quantified with these maps using Cufflinks (Trapnell et al. 2010), correcting for fragment bias (`--frag-bias-correct`) and uncertain alignment location (`--multi-read-correct`) and default parameter values otherwise. Expression levels are reported as $log_2(FPKM+1)$ unless otherwise stated.

Data access

Sequencing data for DNA methylation in *S. invicta* have been deposited in NCBI Gene Expression Omnibus (GSE39959). Additional Supplemental files are freely available for download from the Hymenoptera Genome Database (Munoz-Torres et al. 2011): http://hymenopteragenome.org/ant_genomes/?q=consortium_datasets.

Acknowledgements

This work was supported by grants from NSF to JG and CRS (IOS-0920732) and a Howard Hughes Medical Institute Collaborative Innovation Award #2009005 to DR, SLB, and JL. DFS was supported in part by a NRSA post-doctoral fellowship from the University of Pennsylvania Department of Cell and Developmental Biology. LV was supported in part by Academy of Finland grant #130290. BH, KG, MG were supported by the U. S. National Science Foundation (grant numbers DEB-1011349, DEB-0640690, and IOS-0821130) and the Georgia Tech-Elizabeth Smithgall Watts endowment. YW was supported in part by the BBSRC (grant BB/K004204/1). SN and JJB were supported by the Danish National Research Foundation. RMW was supported by SNSF 125350 & 143936 to EMZ. JR, EP, LK, and YW were supported by grants from the Swiss NSF and En ERC advanced grant. We thank Mira Han for help with running and interpreting the gene family evolution analyses.

Figure Legends

Figure 1. Overview of protein-coding gene composition and genome size in Hymenoptera. (**A**) Gene and genome content in seven ant species and honeybee (red) and representative solitary insects (blue) as outgroups. Orthology delineation among protein-coding genes from 12 insects identified orthologs present in all (Universal, n=12) or almost all (Broad, $10 \leq n \leq 11$) species,

conserved as single-copy genes or with paralogs (with duplications). Differential gene losses leave orthologs shared among fewer species across the phylogeny (Patchy, n<10). Remaining ant genes exhibit orthology with honeybee (*AMELL*, *Apis mellifera*) and/or jewel wasp (*NVITR*, *Nasonia vitripennis*) (Hymenoptera), among ants (Formicidae), or lack orthology (Undetectable). Total estimated genome sizes vary among Hymenoptera, largely due to repetitive regions (orange bars), however hymenopterans share a non-repetitive core of ~200 megabases (Mb) (green bars). A maximum-likelihood species tree computed from the concatenated alignment of all universal single-copy orthologs confirms the established ant phylogeny (Moreau et al. 2006). Rates of molecular evolution are comparable to the other hymenopterans, flour beetle (*TCAST*, *Tribolium castaneum*; genome size ~200 Mb), and body louse (*PHUMA*, *Pediculus humanus*; genome size ~108 Mb), but are much slower than the dipteran representative (*DMELA*, *Drosophila melanogaster*; genome size 175 Mb). Ant species: *HSALT*, *Harpegnathos saltator*; *LHUMI*, *Linepithema humile*; *CFLOR*, *Camponotus floridanus*; *PBARB*, *Pogonomyrmex barbatus*; *SINVI*, *Solenopsis invicta*; *AECHI*, *Acromyrmex echinatior*; *ACEPH*, *Atta cephalotes*. (**B**) Occurrence (blue) and emergence rate (red) of taxonomically restricted genes (TRGs) in different taxonomic clades of Hymenoptera (colors) and Diptera (gray). The youngest clades of both Hymenoptera and Diptera exhibit the highest rates of TRG accumulation. Age is measured as the time between the most distant members of each group and hence does not reflect a clade's absolute age. (**C**) Rate of change of TRGs versus divergence time, for eight species groupings. Pearson correlation coefficient shown. P-value computed using a two-tailed T-test.

Figure 2. Analysis of 64 pan-ant taxonomically restricted genes (TRGs). (**A**) RNA expression support for 64 TRGs that are orthologous among all seven ant species but not found in other genomes. (**A**) RNA expression levels, estimated as $\log_2$(FPKM+1), are shown for various developmental stages and adult castes of *C. floridanus* and *H. saltator*. (**B**) Expression correlation between adult worker castes in *C. floridanus* (major vs. minor, green) and *H. saltator* (gamergate vs. worker, blue) for the 64 novel ant TRGs; Pearson correlation coefficients are shown. Inset shows histogram of differences in gene expression levels between castes (major – minor in green, gamergate – worker in blue) per gene. (**C**) Length distribution (in amino acids) of the 64 novel ant TRGs. Inset shows distribution of the number (left) and percentage (right) of conserved alignment positions (see Supplemental Text 1).

Figure 3. DNA methylation profiles in ant genomes. (**A**) Normalized CpG content (CpG O/E) of different genomic elements, including exons, introns, and promoter regions (1.5 kb upstream of coding sequence start sites) for protein-coding genes, non-genic conserved elements, and genome-wide background (1 kb fragments). Exons show the strongest evidence of CpG depletion in ants, indicating they are the most highly methylated regions of the genome in all taxa (confirmed by Bisulfite-seq; below). Introns also show slight depletion of CpGs in ants suggesting some intron methylation. (**B**) Scatterplot of $\log_{10}$(mCG/CG) methylation levels estimated by Bisulfite-seq versus CpG O/E for coding sequences in *S. invicta* reveals a bimodal distribution of gene body methylation. (**C**) Average methylation levels (mCG/CG) for protein-coding genes in *S. invicta* males, grouped according to the number of taxa in Hymenoptera with orthologs for each gene; indicating conserved genes tend to be highly methylated. Error bars indicate 95 percent confidence intervals for the mean.

Figure 4. Evolution of transcription factor binding sites (TFBS) in insects. (**A**) Heatmap showing number of promoter TFBSs per gene for 59 TFs in 28 insect species (n=4,189 genes associated with 2kb promoter CEs in ants). Species (rows), ordered by phylogenetic grouping, are denoted as solitary (blue) or eusocial (red). TFs were clustered hierarchically using average linkage by computing Euclidean distance between TFBS profiles over all queried genes. On right, boxplots show distributions of Euclidean distance values for pairs of species, computed using genome-wide TFBS abundance profiles over genes and TFs (see Supplemental Fig. 30). Each boxplot reflects a group of paired comparisons. P-values estimated by 2-tailed Mann-Whitney U-test. $^{**}P < 10^{-5}$, $^{***}P < 10^{-10}$. (**B**) Genes and TFs exhibiting significant TFBS evolution between solitary and eusocial groups. 3,231 of 4,189 genes had sufficient data for significance testing. *denotes TF with significant promoter TFBS evolution (2-tail Mann-Whitney U-test; FDR < 0.25). Top two rows indicate numbers of genes showing significant gain or loss of binding sites for the specified TF. Bottom row indicates proportion of significant genes showing more TFBSs in eusocial compared to solitary insects. Over 93% of tested genes are single-copy in the ant genomes. Bottom panels show mean and standard error of the standard deviation in RNA expression levels (Y-axis) for 96 genes with greatest significance in multiple TFs (top 5%), grouped by TF. Expression levels estimated by $\log_2$(FPKM+1). FPKM, fragments

per kilobase per million reads. *denotes significantly increased caste variation in RNA expression (compared to all ant orthologs, Background, $P < 0.05$). (**C**) TFBS abundance profiles for significant genes, shown for three TFs. Species order (X-axis) as in (A). (**D**) TFBS abundance profiles for two neuronal genes with significance in multiple TFs. Cell colors are row-normalized. Periods (.) denote missing data. P-values were computed by a Mann-Whitney U-test. (**E**) mRNA expression level estimates for the genes in (**D**), shown for different worker castes in *H. saltator* (reproductive/non-reproductive) and *C. floridanus* (major/minor). Error bars indicate standard error over three biological replicates. **FDR < 0.01, *FDR < 0.25.

Figure 5. Transcription factor mediated signaling pathways controlling salivary gland development. (**A**) *Sex combs reduced* (SCR) in combination with *extradenticle* (EXD) and homothorax (HTH) direct the specification of cells to the salivary gland fate in PS2 of the *Drosophila* embryo; these TFs are essential for the downstream regulation of genes required for gland cell differentiation and morphogenesis. Boundaries of salivary gland development are restricted along the anterior/posterior axis by *abdominal B* (Abd-B) and *teashirt* (TSH), along the dorsal axis by *decapentaplegic* (DPP) signaling, and along underlying mesoderm by *twist* (TWI) and snail (SNA). *Epidermal growth factor* (EGF) signaling determines the decision to differentiate into duct or secretory cells. (**B**) Regulation of programmed cell death of embryonic salivary gland by 20-hydroxyecdysone (20E). Cell death is inhibited by *forkhead* (FKH) expression in embryonic salivary gland cells. A pulse of 20E at the late larval stage triggers *broad* (BR-C) mediated FKH inhibition. A second pulse of 20E at the prepupal stage leads to BR-C and *ecdysone-induced protein 74EF* (EIP74EF) directed transcription of apoptotic genes, including *wrinkled* (HID). TFs associated with TFBS evolution in over 100 target genes (cis-regulatory evolution; see Fig. 4) are shown in red. The promoter of the *forkhead* locus (yellow), which encodes a TF involved in the regulation of secretory cells, shows significant changes in the gain or loss of promoter TFBSs in ants (trans-regulatory evolution), as detailed in (**C**). On right, mRNA expression level estimates for *forkhead*, shown for different worker castes in *Harpegnathos saltator* (reproductive/non-reproductive) and *Camponotus floridanus* (major/minor). Error bars indicate standard error over three biological replicates.

References

Abouheif E, Wray GA. 2002. Evolution of the gene network underlying the wing polyphenism in ants. *Science* **297:**249–252.

Abrams EW, Andrew DJ. 2005. CrebA regulates secretory activity in the *Drosophila* salivary gland and epidermis. *Development* **132:**2743–2758.

Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin, N, Donahue G, Yang P, Li Q, Li C, et al. 2010. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329:**1068–1071.

Bonasio R, Li Q, Lian J, Mutti NS, Jin L, Zhao H, Zhang P, Wen P, Xiang H, Ding Y, et al. 2012. Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Curr Biol* **22:**1755–1764.

Cardinal S, Danforth BN. 2011. The antiquity and evolutionary history of social behavior in bees. *PLoS ONE* **66:**e21086.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17:**540–552.

Cohen SM, Jürgens G. 1990. Mediation of *Drosophila* head development by gap-like segmentation genes. *Nature* **346:**482–485.

Coyle IP, Koh YH, Lee, WCM, Slind J, Fergestad T, Littleton JT, Ganetzky B. 2004. Nervous wreck, an SH3 adaptor protein that interacts with Wsp, regulates synaptic growth in *Drosophila*. *Neuron* **41:**521–534.

Crozier RH, Pamilo P. 1996. Evolution of social insect colonies: Sex allocation and kin selection. Oxford University Press, Oxford and New York.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:**203–218.

Elango B, Hunt BG, Goodisman MAD, Yi SV. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, *Apis mellifera*. *Proc Natl Acad Sci USA* **106:**11206–11211.

Ferreira PG, Patalano S, Chauhan R, Ffrench-Constant R, Gabaldon T, Guigo R, Sumner S. 2013. Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. *Genome Biol* **14:**R20.

Fernández-Marín H, Zimmerman JK, Nash DR, Boomsma JJ, Wcislo WT. 2009. Reduced biological control and enhanced chemical pest management in the evolution of fungus farming in ants. *Proc R Soc B* **276:**2263–2269.

Gadau J, Helmkampf M, Nygaard S, Roux J, Simola DF, Smith CR, Suen G, Wurm Y, Smith CD. 2012. The genomic impact of 100 million years of social evolution in seven ant species. *Trends Genet* **28:**14–21.

Glastad KM, Hunt BG, Yi SV, Goodisman MAD. 2011. DNA methylation in insects: on the brink of the epigenomic era. *Insect Mol Biol*. **20:**553-565.

Gronenberg W, Heeren S, Hölldobler B. 1996. Age-dependent and task-related morphological changes in the brain and the mushroom bodies of the ant *Camponotus floridanus*. *J Exp Biol* **199:**2011–2019.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59:**307–321.

Hahn MW, Han MV, Han SG. 2007. Gene family evolution across 12 *Drosophila* genomes. *PLoS Genet* **3:**e197.

Hölldobler B, Wilson EO. 2009. The Superorganism. W.W. Norton & Company, New York and London.

Ishimoto H, Sakai T, Kitamoto T. 2009. Ecdysone signaling regulates the formation of long-term courtship memory in adult *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **106:**6381–6386.

Johnson BR, Tsutsui ND. 2011. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC Genomics* **12:**164.

Kamakura M. 2011. Royalactin induces queen differentiation in honeybees. *Nature* **473:**478–483.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30:**3059–3066.

Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet* **25:**404–413.

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465:**182–187.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27:**1571–1572.

Kucharski R, Maleszka j, Foret S, Maleszka R. 2008. Nutritional control of reproductive status in honeybees via DNA methylation. *Science* **319:**1827–1830.

Kulakovskiy IV, Favorov AF, Makeev VJ. 2009. Motif discovery and motif finding from genome-mapped DNase footprint data. *Bioinformatics* **25:**2318–2325.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:**R25.

Levy S, Hannenhalli S. 2002. Identification of transcription factor binding sites in the human genome. *Mammalian Genome* **13:**510–514.

Li T-R, White KP. 2003. Tissue-specific gene expression and ecdysone-regulated genomic networks in *Drosophila*. *Developmental Cell* **5:**59–72.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478:**476–82.

Loewer S, Cabili MN, Guttman M, Loh Y-H, Thomas K, Park IH, Garber M, Curran M, Onder T, Agarwal S, et. al. 2011. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42:**1113–1117.

Mach V, Takiya S, Ohno K, Handa H, Imai T, Suzuki Y. 1995. Silk Gland Factor-1 involved in the regulation of Bombyx Sericin-1 gene contains Fork Head motif. *J Biol Chem* **270:**9340–9346.

Martin S, Drijfhout F. 2009. A review of ant cuticular hydrocarbons. J Chem Ecol **35:**1151–116.

Michener CD 1969. Comparative social behavior of bees. *Annu Rev Entomol* **14:**299–342.

Moreau CS, Bell CD, Vila R, Archibald SB, Pierce NE. 2006. Phylogeny of the ants: diversification in the age of angiosperms. *Science* **312:**101–104.

Munoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, Childs KL, Anzola JM, Milshina N, Elsik CG. 2011. Hymenoptera genome database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res* **39:**D658–D662.

Nygaard S, Zhang G, Schiott M, Li C, Wurm Y, Hu H, Zhou J, Ji L, Qiu F, Rasmussen M, et al. 2011. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res* **21:**1339–1348.

Parker BJ, Moltke I, Roth A, Washietl S, Wen J, Kellis M, Breaker R, Pedersen JS. 2011. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res* **21:**1929–1943.

Pauli A, Rinn JL, Schier AF. 2011. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet* **12:**136–149.

Ramirez-Carrozzi VR, Braas D, Bhatt DM, Cheng CS, Hong C, Doty KR, Black JC, Hoffmann A, Carey M, Smale ST. 2009. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* **138:**114–128.

Ranson H, Claudianos C, Ortelli F, Abrall C, Hemingway J, Sharakhova MV, Unger MF, Collins FH, Feyereisen, R. 2002. Evolution of supergene families associated with insecticide resistance. *Science* **298:**179–181.

Rebeiz M, Jikomes N, Kassner VA, Carroll SB. 2011. Evolutionary origin of novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. *Proc Natl Acad Sci USA* **108:**10036–10043.

Robinson GE, Grozinger CM, Whitfield CW. 2005. Sociogenomics: social life in molecular terms. *Nat Rev Genet* **6:**257–270.

Schultz TR, Brady SG. 2008. Major evolutionary transitions in ant agriculture. *Proc Natl Acad Sci USA* **105:**5435–5440.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? *PLoS Genet* **56:**e1000495.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:**1034–1050.

Simola DF, Ye C, Mutti NS, Dolezal K, Bonasio R, Liebig J, Reinberg D, Berger S. 2013. A chromatin link to caste identity in the carpenter ant *Camponotus floridanus*. *Genome Res* **23:**486–496.

Sirviö A, Pamilo P, Johnson RA, Page RE, Jr, Gadau J. 2011. Origin and evolution of the dependent lineages in the genetic caste determination system of *Pogonomyrmex* spp. *Evolution* **65:**869–884.

Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, et al. 2011. Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile). Proc Natl Acad Sci USA* **108:**5673–5678.

Smith CR, Mutti NS, Jasper WC, Naidu A, Smith CD, Gadau J. 2012. Patterns of DNA methylation in development, division of labor and hybridization in an ant with genetic caste determination. *PLoS ONE* **78:**e42433.

Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R, et al. 2011. Draft genome of the red harvester ant *Pogonomyrmex barbatus*. *Proc Natl Acad Sci USA* **108:**5667–5672.

Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, et al. 2011. The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *PLoS Genet* **7:**e1002007.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly *Drosophila* evolution revealed by mutation clocks. *Mol Biol Evol* **21:**36–44.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet* **12:**692–702.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol* **28:**511–515.

Villesen P, Murakami T, Schultz TR, Boomsma JJ. 2009. Identifying the transition between single and multiple mating of queens in fungus-growing ants. *Proc R Soc Lond B* **269:**1541–1548.

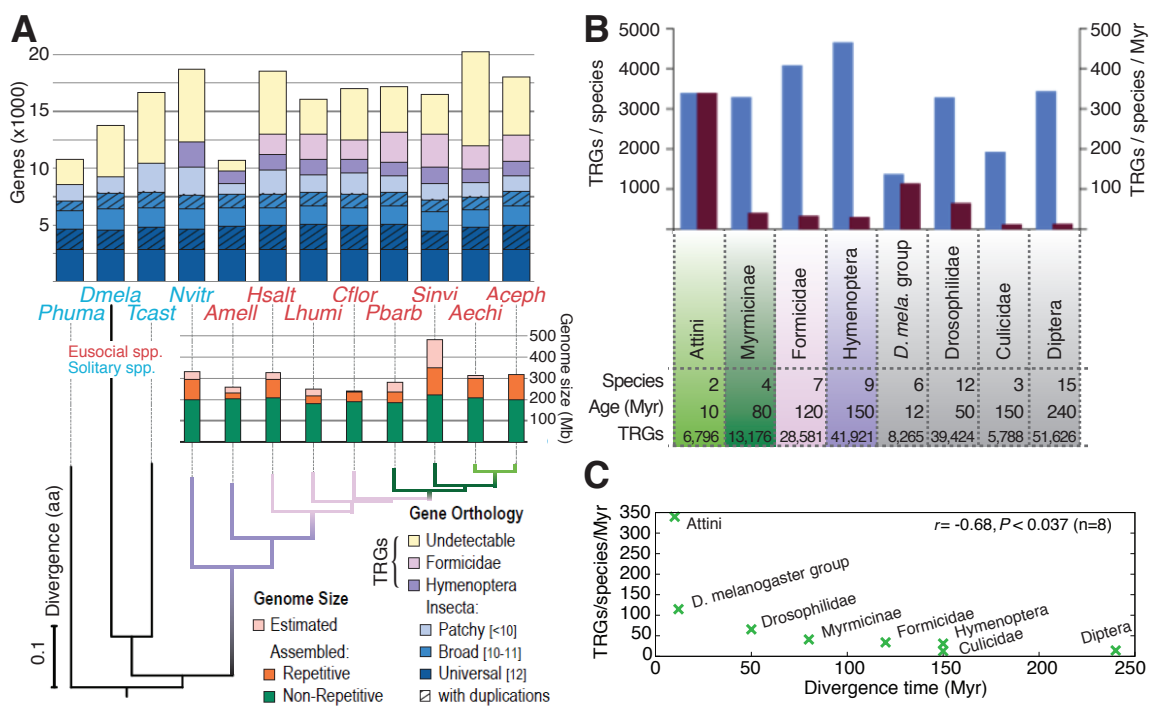
Waterhouse RM, Zdobnov EM, Tegenfeldt F, Li J, Kriventseva EV. 2011. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res* **39:**D283–288.

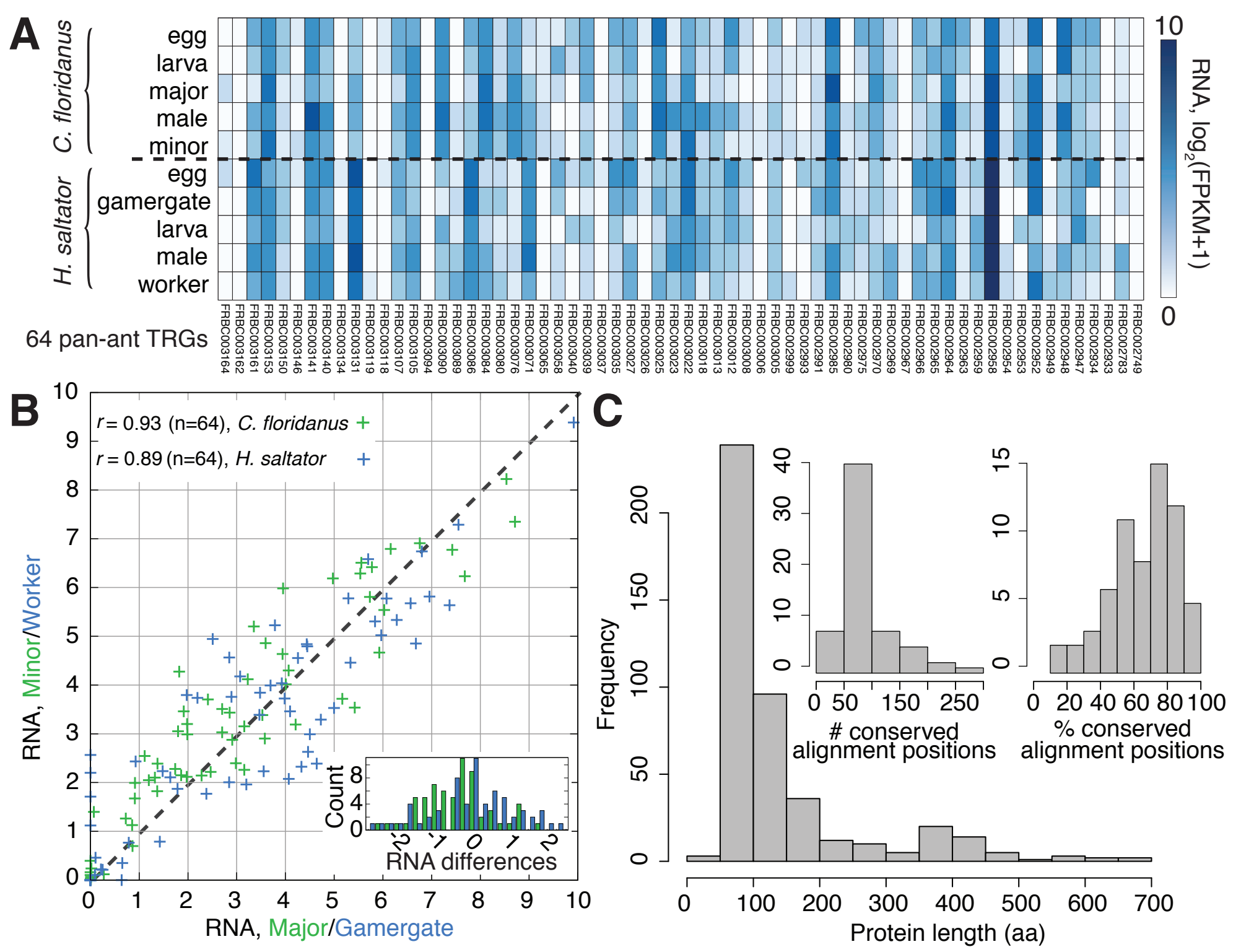
The Honeybee Genome Sequencing Consortium. 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443:**931–949.

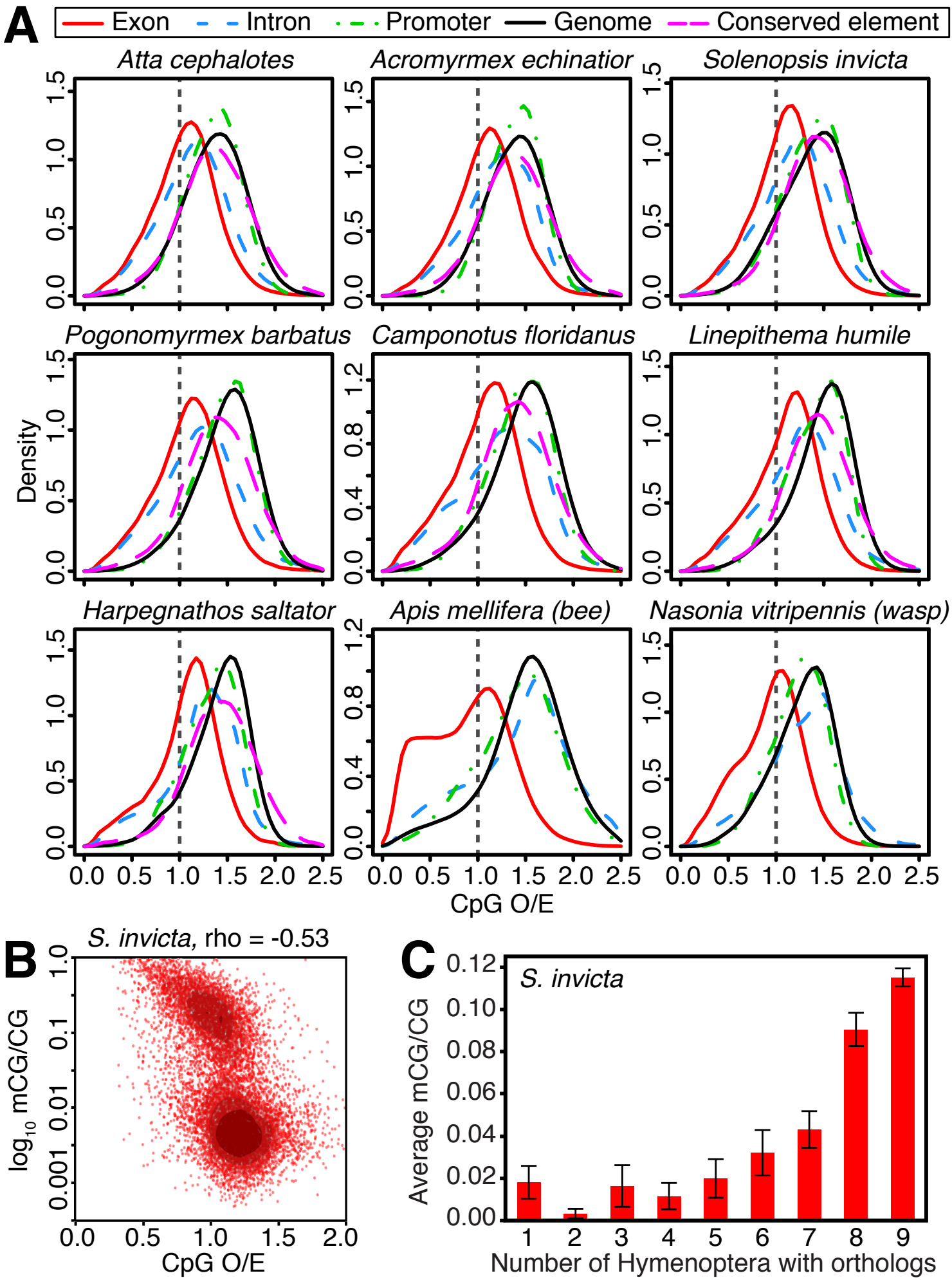
Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, et al. 2011. The genome of the fire ant *Solenopsis invicta*. *Proc Natl Acad Sci USA* **108:**5679–5684.

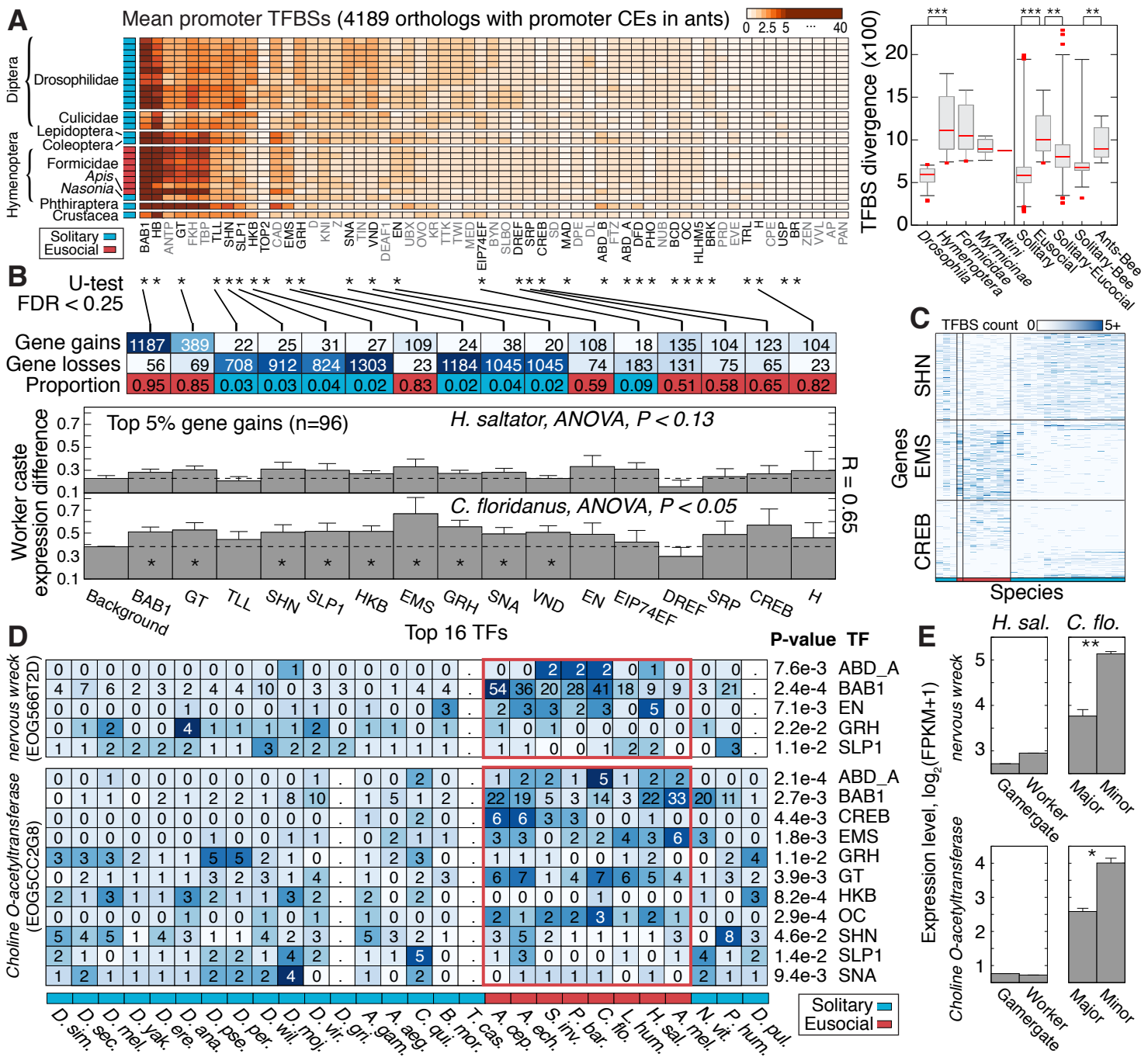
Zhou X, Slone JD, Rokas A, Berger SL, Liebig J, Ray A, Reinberg D, Zwiebel L. 2012. Phylogenetic and transcriptomic analysis of chemosensory receptors in a pair of divergent ant species reveals sex-specific signatures of odor coding. *PLoS Genet* **8:**e1002930.

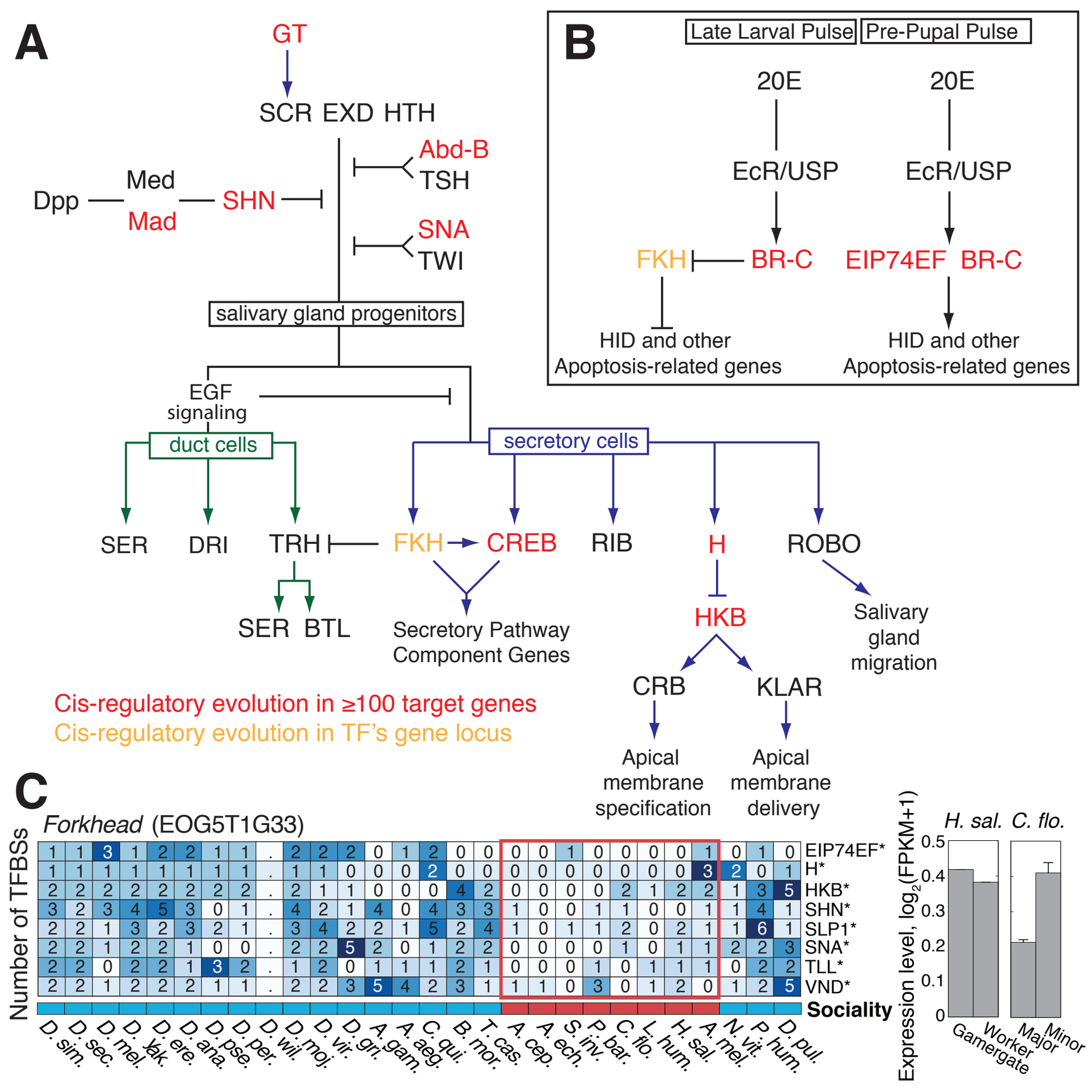
Zweden van JS, Brask JB, Christensen JH, Boomsma JJ, Linksvayer TA, D'Ettore P. 2010. Blending of heritable recognition cues among ant nestmates creates distinct colony gestalt odours but prevents within-colony nepotism. *J Evol Biol* 23:1498–1508.


A
Genes (x1000)
20
15
10
5
Dmela
Nvitr
Hsalt
Cflor
Sinvi
Aceph
Phuma
Tcast
Amell
Lhumi
Pbarb
Aechi
Eusocial spp.
Solitary spp.
Genome size (Mb)
500
400
300
200
100
Divergence (aa)
0.1
Gene Orthology
TRGs
Undetectable
Formicidae
Hymenoptera
Genome Size
Estimated
Assembled:
Repetitive
Non-Repetitive
Insecta:
Patchy [<10]
Broad [10-11]
Universal [12]
with duplications
B
TRGs / species
5000
4000
3000
2000
1000
TRGs / species / Myr
500
400
300
200
100
Attini
Myrmicinae
Formicidae
Hymenoptera
D. mela. group
Drosophilidae
Culicidae
Diptera
Species
2
4
7
9
6
12
3
15
Age (Myr)
10
80
120
150
12
50
150
240
TRGs
6,796
13,176
28,581
41,921
8,265
39,424
5,788
51,626
C
TRGs/species/Myr
350
300
250
200
150
100
50
0
Attini
r= -0.68, P< 0.037 (n=8)
D. melanogaster group
Drosophilidae
Myrmicinae
Formicidae
Hymenoptera
Culicidae
Diptera
0
50
100
150
200
250
Divergence time (Myr)


A
C. floridanus
egg
larva
major
male
minor
H. saltator
egg
gamergate
larva
male
worker
64 pan-ant TRGs
RNA, log2(FPKM+1)
10
0
B
r = 0.93 (n=64), C. floridanus
r = 0.89 (n=64), H. saltator
RNA, Minor/Worker
RNA, Major/Gamergate
Count
RNA differences
C
Frequency
Protein length (aa)
conserved alignment positions
% conserved alignment positions


A
Exon
Intron
Promoter
Genome
Conserved element
Atta cephalotes
Acromyrmex echinatior
Solenopsis invicta
Pogonomyrmex barbatus
Camponotus floridanus
Linepithema humile
Harpegnathos saltator
Apis mellifera (bee)
Nasonia vitripennis (wasp)
Density
CpG O/E
B
S. invicta, rho = -0.53
log10 mCG/CG
CpG O/E
C
S. invicta
Average mCG/CG
Number of Hymenoptera with orthologs


A
Mean promoter TFBSs (4189 orthologs with promoter CEs in ants)
0 2.5 5 ··· 40
Diptera
Drosophilidae
Culicidae
Lepidoptera
Coleoptera
Hymenoptera
Formicidae
Apis
Nasonia
Phthiraptera
Crustacea
Solitary
Eusocial
BAB1 HB ANTP GT FKH TBP TLL SHN SLP1 HKB TOP2 CAD EMS GRH D KNI Z SNA TIN VND DEAF1 EN UBX OVO KR TTK TWI MED EIP74EF BYN SLBO DREF SRP CREB SD MAD DPE DL ABD_B FTZ ABD_A DFD PHO NUB BCD OC HLHM5 BRK PRD EVE TRL H CPE USP BR ZEN VVL AP PAN
TFBS divergence (x100)
0 5 10 15 20
*** *** ** **
Drosophila
Hymenoptera
Formicidae
Myrmicinae
Attini
Solitary
Eusocial
Solitary
Solitary-Bee
Eusocial
Ants-Bee
B
U-test
FDR < 0.25
Gene gains 1187 389 22 25 31 27 109 24 38 20 108 18 135 104 123 104
Gene losses 56 69 708 912 824 1303 23 1184 1045 1045 74 183 131 75 65 23
Proportion 0.95 0.85 0.03 0.03 0.04 0.02 0.83 0.02 0.04 0.02 0.59 0.09 0.51 0.58 0.65 0.82
Worker caste expression difference
Top 5% gene gains (n=96)
H. saltator, ANOVA, P < 0.13
C. floridanus, ANOVA, P < 0.05
R = 0.65
Background BAB1 GT TLL SHN SLP1 HKB EMS GRH SNA VND EN EIP74EF DREF SRP CREB H
Top 16 TFs
C
TFBS count 0 5+
Genes
SHN
EMS
CREB
Species
D
P-value TF
nervous wreck (EOG566T2D)
7.6e-3 ABD_A
2.4e-4 BAB1
7.1e-3 EN
2.2e-2 GRH
1.1e-2 SLP1
Choline O-acetyltransferase (EOG5CC2G8)
2.1e-4 ABD_A
2.7e-3 BAB1
4.4e-3 CREB
1.8e-3 EMS
1.1e-2 GRH
3.9e-3 GT
8.2e-4 HKB
2.9e-4 OC
4.6e-2 SHN
1.4e-2 SLP1
9.4e-3 SNA
D. sim. D. sec. D. mel. D. yak. D. ere. D. ana. D. pse. D. per. D. wil. D. moj. D. vir. D. gri. A. gam. A. aeg. C. qui. B. mor. T. cas. A. cep. A. ech. S. inv. P. bar. C. flo. L. hum. H. sal. A. mel. N. vit. P. hum. D. pul.
Solitary
Eusocial
E
H. sal. C. flo.
nervous wreck
Choline O-acetyltransferase
Expression level, log2(FPKM+1)
** *
Worker Gamergate Major Minor


A
GT
SCR EXD HTH
Abd-B
TSH
Dpp
Med
Mad
SHN
SNA
TWI
salivary gland progenitors
EGF
signaling
duct cells
secretory cells
SER
DRI
TRH
FKH
CREB
RIB
H
ROBO
SER BTL
Secretory Pathway
Component Genes
HKB
Salivary
gland
migration
CRB
KLAR
Apical
membrane
specification
Apical
membrane
delivery
B
Late Larval Pulse
Pre-Pupal Pulse
20E
20E
EcR/USP
EcR/USP
FKH
BR-C
EIP74EF
BR-C
HID and other
Apoptosis-related genes
HID and other
Apoptosis-related genes
Cis-regulatory evolution in ≥100 target genes
Cis-regulatory evolution in TF's gene locus
C
Forkhead (EOG5T1G33)
Number of TFBSs
EIP74EF*
H*
HKB*
SHN*
SLP1*
SNA*
TLL*
VND*
Sociality
D. sim.
D. sec.
D. mel.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
A. gam.
A. aeg.
C. qui.
B. mor.
T. cas.
A. cep.
A. ech.
S. inv.
P. bar.
C. flo.
L. hum.
H. sal.
A. mel.
N. vit.
P. hum.
D. pul.
H. sal.
C. flo.
Expression level, log2(FPKM+1)
0.5
0.4
0.3
0.2
0.1
0
Worker
Gamergate
Major
Minor