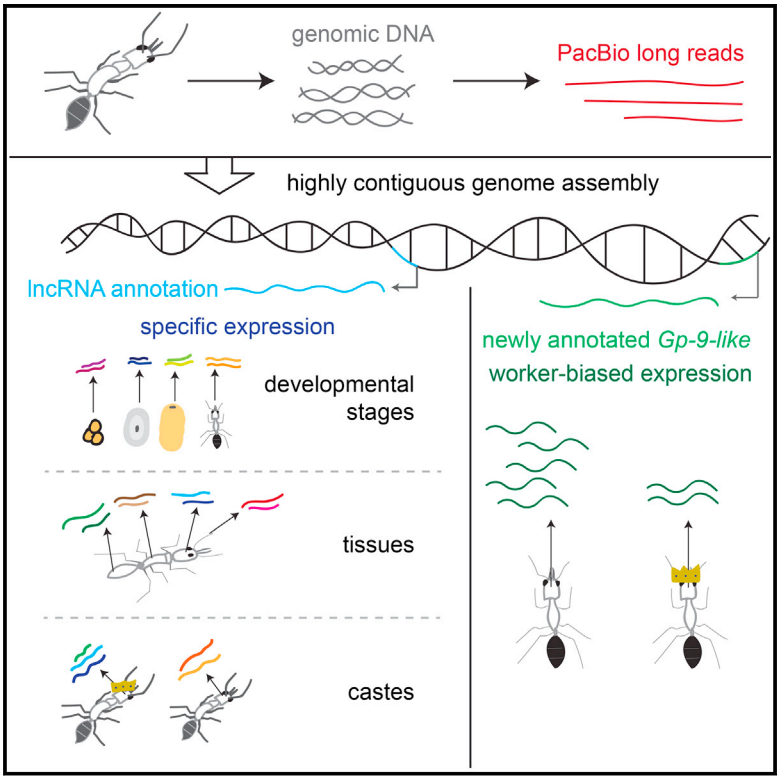
Article

Cell Reports

High-Quality Genome Assemblies Reveal Long Non-coding RNAs Expressed in Ant Brains

Graphical Abstract



Highlights

- Long reads produce highly contiguous genome assemblies for two ant species
- Formerly unannotated gene well studied in other ants has caste-biased expression
- Upgraded genomes allow for annotation of long non-coding RNAs
- Many long non-coding RNAs are expressed in the brain, some in caste-specific manner

Authors

Emily J. Shields, Lihong Sheng, Amber K. Weiner, Benjamin A. Garcia, Roberto Bonasio

Correspondence

rbon@pennmedicine.upenn.edu

In Brief

Using long-read sequencing, Shields et al. upgrade the genome assemblies for two ant species. Their results reveal a protein-coding gene preferentially expressed in worker ants and genes for long non-coding RNAs, several of which were expressed in the brain, in some cases at different levels in workers and reproductives.

Shields et al., 2018, Cell Reports *23*, 3078–3090
June 5, 2018
https://doi.org/10.1016/j.celrep.2018.05.014

CellPress

OPEN ACCESS
CellPress

Cell Reports
Article

High-Quality Genome Assemblies Reveal Long Non-coding RNAs Expressed in Ant Brains

Emily J. Shields,[1,2,3] Lihong Sheng,[1,3] Amber K. Weiner,[1,2,4] Benjamin A. Garcia,[1,4] and Roberto Bonasio[1,3,5,*]
[1]Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[5]Lead Contact
*Correspondence: rbon@pennmedicine.upenn.edu
https://doi.org/10.1016/j.celrep.2018.05.014

SUMMARY

Ants are an emerging model system for neuroepigenetics, as embryos with virtually identical genomes develop into different adult castes that display diverse physiology, morphology, and behavior. Although a number of ant genomes have been sequenced to date, their draft quality is an obstacle to sophisticated analyses of epigenetic gene regulation. We reassembled *de novo* high-quality genomes for two ant species, *Camponotus floridanus* and *Harpegnathos saltator*. Using long reads enabled us to span large repetitive regions and improve genome contiguity, leading to comprehensive and accurate protein-coding annotations that facilitated the identification of a *Gp-9-like* gene as differentially expressed in *Harpegnathos* castes. The new assemblies also enabled us to annotate long non-coding RNAs in ants, revealing caste-, brain-, and developmental-stage-specific long non-coding RNAs (lncRNAs) in *Harpegnathos*. These upgraded genomes, along with the new gene annotations, will aid future efforts to identify epigenetic mechanisms of phenotypic and behavioral plasticity in ants.

INTRODUCTION

The ponerine ant *Harpegnathos saltator* is emerging as a model system to study the epigenetic regulation of brain function and behavior (Bonasio, 2012; Yan et al., 2014). Adult *Harpegnathos* workers can convert to acting queens, called gamergates, that are allowed to mate and lay fertilized eggs. We have previously shown that the worker-gamergate transition is accompanied by changes in brain gene expression (Gospocic et al., 2017), but the epigenetic mechanisms responsible for these changes remain unknown.

Previous work in *Harpegnathos* and in the more conventional Florida carpenter ant *Camponotus floridanus* has suggested that epigenetic pathways, including those that control histone modifications and DNA methylation, might be responsible for differential deployment of caste-specific traits (Bonasio et al., 2010, 2012; Simola et al., 2013a). Pharmacological and molecular manipulation of histone acetylation affects caste-specific behavior in *Camponotus* ants (Simola et al., 2016), suggesting a direct role for epigenetics in their social behavior. Although the molecular mechanisms by which environmental and developmental cues are converted into epigenetic information on chromatin remain subject of intense investigation (Allis and Jenuwein, 2016), it has become clear that non-coding RNAs play an important role in mediating this flow of information (Holoch and Moazed, 2015). In particular, long non-coding RNAs (lncRNAs)—transcripts longer than 200 bp that are not translated into proteins—have been proposed to participate in the epigenetic regulation of gene expression (Bonasio and Shiekhattar, 2014; Rinn and Chang, 2012).

Many proteins that regulate chromatin function bind RNA (Guttman et al., 2011; He et al., 2016; Hendrickson et al., 2016), and it is believed that these interactions might explain the epigenetic function of certain lncRNAs. Among epigenetic factors that bind to and are regulated by lncRNAs are SCML2 and EZH2 (Bonasio et al., 2014; Zhao et al., 2010), subunits of *Polycomb* repressive complex 1 and 2, which maintain lineage specifications and cell identity during development via epigenetic gene repression (Schuettengruber et al., 2017); WDR5 (Yang et al., 2014), a subunit of the MLL complex, which belongs to the *Trithorax* group of epigenetic activators (Schuettengruber et al., 2017); various DNA methyltransferases (Wang et al., 2015); and CTCF (Saldaña-Meyer et al., 2014; Sun et al., 2013), better known as the "master weaver of the genome" because of its role in organizing the genome in 3D loops (Phillips and Corces, 2009). In fact, lncRNAs have been directly implicated in maintaining looping interactions between promoters and enhancers (Lai et al., 2013) and as organizers of 3D genome architecture (Amaral et al., 2018; Engreitz et al., 2016a; Joung et al., 2017).

lncRNAs have been annotated extensively in human (Cabili et al., 2011; Derrien et al., 2012), mouse (Guttman et al., 2009; Pervouchine et al., 2015), model organisms such as zebrafish, *Drosophila melanogaster* and *Caenorhabditis elegans* (Gerstein et al., 2014; Nam and Bartel, 2012; Pauli et al., 2012; Ulitsky et al., 2011; Young et al., 2012), and the bees *Apis mellifera* and *Apis cerana* (Jayakodi et al., 2015); but, to our knowledge, no comprehensive annotation of lncRNAs in ants has been

reported. This may be in part because ant genomes, including those of *Camponotus* and *Harpegnathos* (Bonasio et al., 2010), are still in draft, highly fragmented form due to the prevalent use of whole-genome shotgun sequencing to assemble them. In addition to making lncRNA annotation practically impossible, the fragmented nature of these genome assemblies also hamper the sophisticated genome-wide analyses required for epigenetic research, thus limiting the reach of these species as model organisms.

We upgraded the genomes of *Harpegnathos* and *Camponotus* to megabase level with a combination of *de novo* assembly of Pacific Biosciences (PacBio) long reads, scaffolding with mate pairs and long reads, and polishing with short reads. The contiguity of both genomes was greatly improved while maintaining the high accuracy of the short-read-only assemblies. We used these new assemblies to annotate protein-coding genes and lncRNAs, leading to the discovery of lncRNAs differentially expressed between *Harpegnathos* castes, developmental stages, and tissues. These improvements to the *Harpegnathos* and *Camponotus* genomes will lead to greater understanding of the genetic and epigenetic factors that underlie the behavior of these social insects.

RESULTS

Long-Read Sequencing Improves Contiguity

We sequenced genomic DNA isolated from *Harpegnathos* and *Camponotus* workers using PacBio single-molecule real-time technology, obtaining a sequence coverage of 70× for *Harpegnathos* and 53× for *Camponotus*, compatible with PacBio-only genome assembly (Koren et al., 2017). PacBio reads are much longer than those used for whole-genome shotgun draft assemblies, including the previously reported assemblies for these two ant species (Bonasio et al., 2010), and are thus expected to yield longer contigs and scaffolds with fewer gaps (scheme, Figure 1A).

We used these long reads to assemble the two genomes *de novo* using a multistep process (Figure S1A), starting with the dedicated long read assembler Canu (Koren et al., 2017). Although this initial step produced assemblies that surpassed the contiguity of the current draft genomes (Figure S1A; Table S1), we leveraged long reads and previously generated sequencing data to maximize the quality of the newly assembled genomes.

The new PacBio sequencing-derived assemblies ("2016 assemblies") compared favorably to the short-read assemblies currently available for both ant species ("2010 assemblies"). Despite capturing a larger amount of genomic sequence (Table S1), the number of contigs was dramatically decreased in the 2016 assemblies (Figure 1B) and their average size was more than 30-fold larger than in the 2010 assemblies (Figure 1C), reflecting greatly increased assembly contiguity. Scaffolding was also improved in the 2016 assemblies, which consisted of fewer, larger scaffolds (Figure S1B) and contained fewer gaps than the 2010 assemblies (Table S1). Improvements were also evident in the conventional metrics of assembly quality such as contig and scaffold N50s (Table S1). Overall, the contig N50 size increased by 22-fold (to 885 kb) and 65-fold (to 1.2 Mb) for *Harpegnathos* and *Camponotus*, respectively, and in both assemblies, the scaffold N50 size surpassed 1 Mb (Table S1).

The contig N50 sizes of our improved *Harpegnathos* and *Camponotus* assemblies top almost all other insect genomes available in the NCBI database, with the exception of two genomes also assembled using PacBio long-read sequencing, *Drosophila serrata* (Allen et al., 2017) and *Aedes albopictus* (Miller et al., 2018), as well as the classic model organism *Drosophila melanogaster* (Figure 1D, left). The number and size of scaffolds also compared favorably with other available genomes (Figure 1D, right), and the numbers of gaps in our new assemblies (240 and 326 in *Harpegnathos* and *Camponotus*, respectively) are lower than for any other insect genome in this set, including *Drosophila melanogaster* (Figure 1E).

PacBio reads can span long repetitive sequence that cannot be assembled properly using short reads (Roberts et al., 2013). We found several cases where distinct scaffolds from the 2010 assemblies mapped to a single scaffold (or contig) in the 2016 assemblies, separated by repetitive sequences. For example, scaffolds 921 and 700 from 2010 were joined into a larger scaffold in the improved 2016 assemblies (Figure 1F), separated by ~6.5 kb of repeats spanned by multiple PacBio reads (Figure 1G). Indeed, much of the newly assembled DNA sequence consisted of repeats (Figure S2).

Thus, long PacBio reads allowed us to assemble across longer repeats than previously possible, greatly improving the contiguity of the *Harpegnathos* and *Camponotus* genomes.

The New Assemblies Are Highly Accurate

We countered the high error rate of PacBio sequencing with deep sequence coverage (>50×) and by polishing our assemblies with the short reads from the original draft genomes (Bonasio et al., 2010).

RNA sequencing (RNA-seq) from various developmental stages in both species mapped better to the 2016 assemblies compared to the 2010 draft versions in all cases (Figure 2A), with a lower mismatch rate per base (Figure 2B), demonstrating that our strategy to correct PacBio sequencing errors successfully generated highly accurate genome sequences. The improved mapping rate suggests that the new assemblies capture transcribed but previously unassembled sequence.

Furthermore, alignment of Sanger sequences of 10 (*Harpegnathos*) and 9 (*Camponotus*) ~40-kb fosmid clones (Bonasio et al., 2010) showed similar or higher coverage in the new assemblies compared to the draft 2010 versions (Figure 2C; Table S2).

Neither the RNA-seq nor the fosmids were used in assembly construction, providing an orthogonal method of measuring genome completeness and accuracy.

Improvements in Protein-Coding Annotations

We annotated protein-coding genes using a combination of *ab initio* transcriptome reconstruction, homology-based searches, and *de novo* identification of gene structure (Figure S3A). We used MAKER2 (Holt and Yandell, 2011) to combine these sources of evidence and retained models consistently represented across evidence (Figure S3B) and/or with a protein domain,

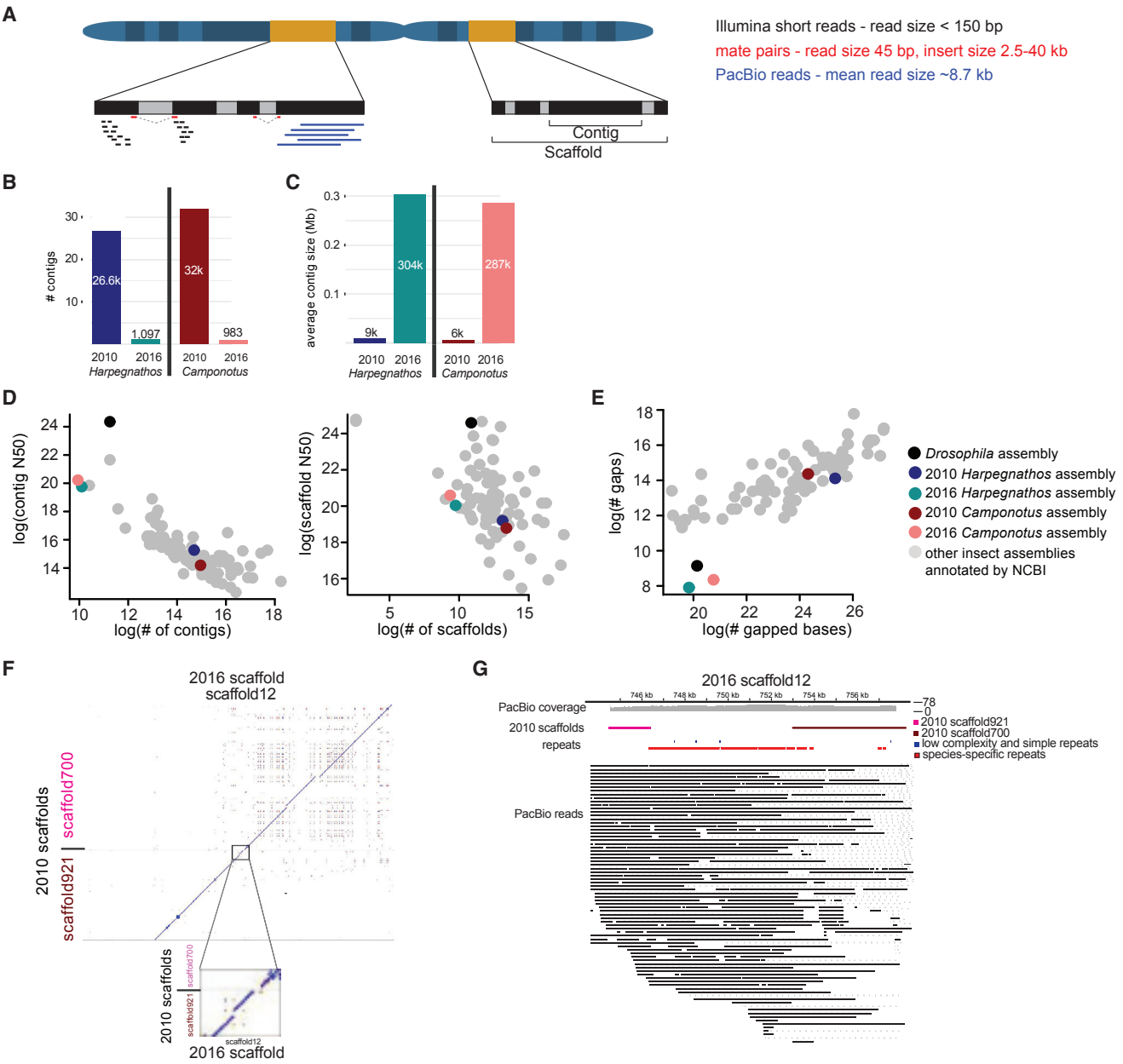


Figure 1. PacBio Sequencing Improves Contiguity of Two Ant Genomes

(A) Scheme showing types of reads used in assembly.

(B and C) Comparison of contig number (B) and average size (C) in 2016 and 2010 assemblies.

(D) Comparison of *Harpegnathos* and *Camponotus* genome assemblies to other insect genomes using contig number and N50 (left) and scaffold number and N50 (right).

(E) Number of gaps and gapped bases in insect assemblies.

(F) Two 2010 scaffolds, scaffold921 and scaffold700, are depicted along the y axis, with the 2016 scaffold, scaffold12, along the x axis. Dots indicate regions where there is significant sequence similarity. The boundary region between the 2010 scaffolds is shown in the inset.

(G) A genome browser view of region from (F) shows coverage by several PacBio reads that span the stretch of repetitive sequence across the gap between the two 2010 scaffolds.

See also Table S1 and Figures S1 and S2.

annotating 20,317 and 18,620 protein-coding genes for *Harpegnathos* and *Camponotus*, respectively (Figure S3C). The filtered protein-coding annotations recovered slightly higher percentages of a core set of evolutionarily conserved arthropod genes (Simão et al., 2015) compared to the 2010 annotations (Table S3).

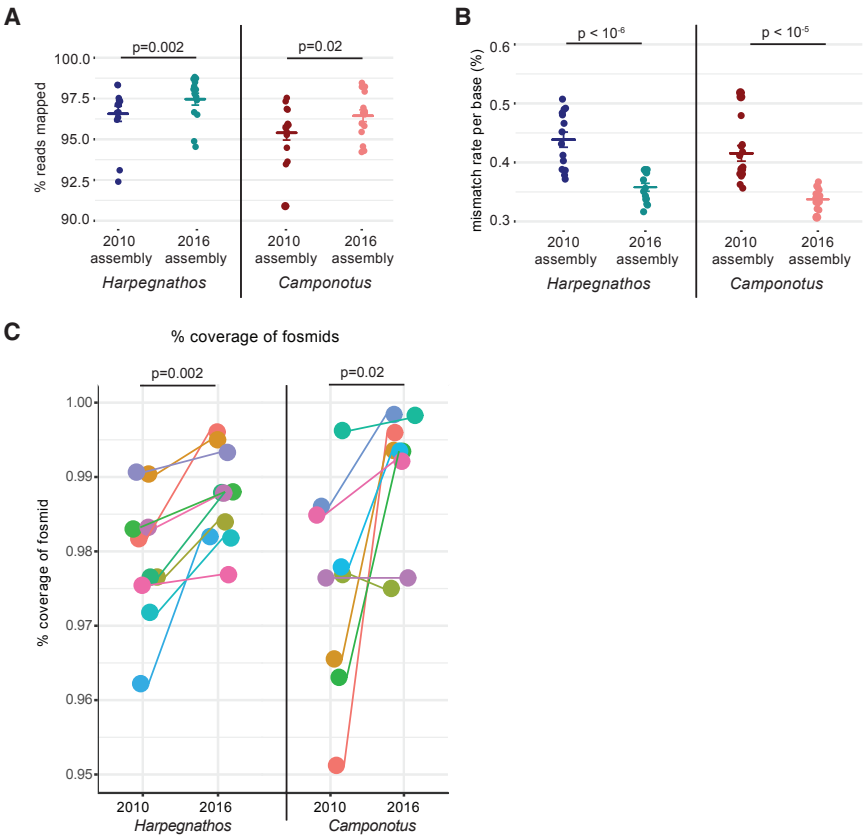


Figure 2. Improved Accuracy of New Assemblies

(A and B) Mapping (A) and sequence mismatch (B) rates for RNA-seq reads from various developmental stages of *Harpegnathos* (n = 14) and *Camponotus* (n = 15) to old and new assemblies. Horizontal bars indicate the means. p values are from two-sided, paired Student's t test. Error bars indicate SEM.

(C) 2010 and 2016 assembly accuracy measured by percentage of fosmid Sanger sequence covered on a single scaffold. Each dot represents a fosmid. p value is from a two-sided Student's t test.

See also Table S2.

The number of gene models encoding proteins conserved throughout evolution was more or less unchanged after the genome update (Figure 3A). Interestingly, a higher percentage of genes in the 2016 assemblies had no homology to known protein-coding genes in human, mouse, and a panel of insects, including several Hymenoptera (Figure 3A, red boxes). A majority of these gene models without homology to known proteins contained at least one recognizable protein families (PFAMs) domain (Figure 3B), suggesting that they might encode true protein-coding genes missed by annotation efforts in related organisms.

We reasoned that the improved assemblies and protein-coding annotations might uncover biologically relevant genes missing in the older versions. *Harpegnathos* workers are characterized by their unique reproductive and brain plasticity that, in absence of a queen, allows some of them to transition to a queen-like phenotypic status called "gamergate" (Bonasio et al., 2010, 2012), which is accompanied by major changes in brain gene expression (Gospocic et al., 2017). Mapping this dataset to the new annotation, we found that a *Gp-9-like* gene had significantly higher expression in worker brains compared to gamergates (Figure 3C). This gene was not previously detected as differentially expressed, likely because its closest homolog in the old annotation contains many sequence disparities (Figure S4A), reducing the RNA-seq coverage mapped to this gene in both castes (Figure S4B). Mass spectrometry analyses identified two peptides mapping exactly to the newly predicted sequence (Figure S4A), confirming the accuracy of the updated gene model.

This *Gp-9-like* gene encodes one of several proteins with homology to a pheromone-binding protein well studied in the fire ant *Solenopsis invicta* because it marks a genomic element associated with the ability of the colony to accept one or more fertile queens (Wang et al., 2013). Other ant species, including *Monomorium pharaonis* (Warner et al., 2017), *Vollenhovia emeryi* (Miyakawa and Mikheyev, 2015), and *Dinoponera quadriceps* (Patalano et al., 2015), have several *Gp-9-like* homologs, some of which display worker-biased expression patterns (Figures S4C–S4E). Many other Hymenoptera also have *Gp-9* or *Gp-9-like* homologs in their genomes (Figure S4F), and much of the *Solenopsis invicta* gene that associates with colony structure is conserved with these *Gp-9-like* gene models, especially within the odorant-binding domain (Figure S4G). Furthermore, this gene is likely under positive selection (significant by chi-square test, degrees of freedom [df] = 1, $\alpha = 0.001$). These observations suggest that the role of this pheromone-binding protein in social organization is more conserved than previously appreciated.

One specific locus with better contiguity and improved protein-coding annotations is the *Hox* cluster, a group of developmental genes conserved throughout metazoa (Finnerty and Martindale, 1998). Homologs for two *Drosophila Hox* cluster genes, *lab* and *abd-A*, were surprisingly missing from the 2010 *Harpegnathos* annotation (Simola et al., 2013b); however, both genes were properly positioned in the *Hox* cluster of the new *Harpegnathos* assembly, in the same order as the corresponding *Drosophila* homologs (Figure 4A). The 2010 annotation did contain gene models overlapping the loci for *lab* and *abd-A*, but they were incomplete (Figures 4B and 4C; data not shown), which had previously prevented their detection by homology searches. The contiguity of the *Hox* cluster is critical to its function, as genes in the cluster are expressed in a collinear fashion during development (Kmita and Duboule, 2003). *Drosophila* and the silkworm *Bombyx mori* have split *Hox* clusters (Negre et al.,

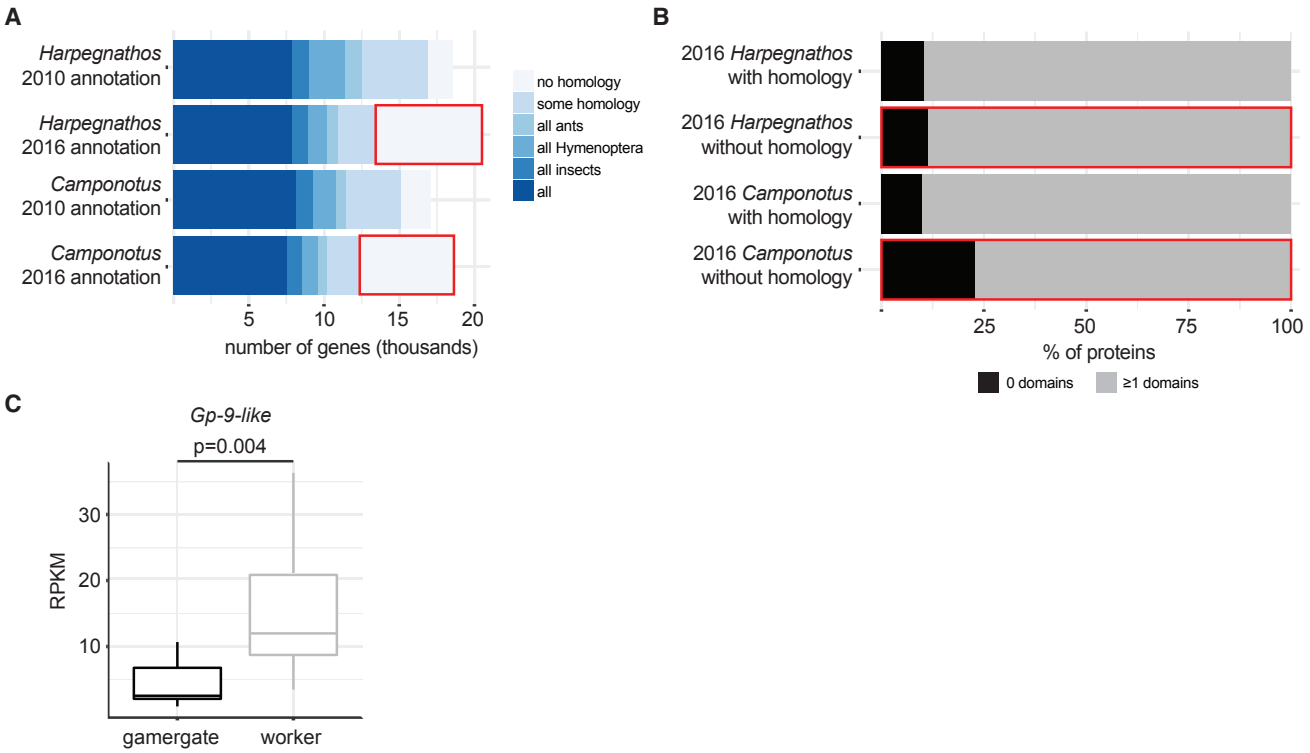


Figure 3. Annotation of Protein-Coding Genes

(A) Number of genes in 2010 and 2016 *Harpegnathos* and *Camponotus* annotations with a homolog in a panel of other ants, Hymenoptera, and animals.

(B) Fraction of genes with no detectable homology, as outlined in red in (A), that contains no (black) or ≥1 (gray) protein family (PFAM) domains.

(C) Expression of the previously unannotated *Gp-9-like* gene in *Harpegnathos* gamergates (n = 12) and workers (n = 11). p value is from a two-sided Student's t test.

See also Table S3 and Figures S3 and S4.

2005; Yasukochi et al., 2004), but many other insects have an intact one (Brown et al., 2002; Devenport et al., 2000; Ferrier and Akam, 1996), including *Apis mellifera* (Dearden et al., 2006). In our previous assemblies, the *Camponotus* cluster was split among three different scaffolds, begging the question of whether this separation was due to the actual relocalization of genes during evolution or simply discontinuous assembly. The improved 2016 assemblies answered this question by showing that the entire *Camponotus Hox* clusters could be assembled into a single, larger scaffold (Figure 4A).

Together, our analyses show that reannotation of the improved assemblies for *Harpegnathos* and *Camponotus* yielded more complete gene sets, better models of already annotated genes, and better contiguity of a tightly regulated gene cluster.

Annotation of lncRNAs

To annotate lncRNAs, we assembled a reference-based transcriptome from RNA-seq of various developmental stages and retained high-confidence transcripts longer than 200 bp not overlapping with existing protein-coding gene models (Figure S5A). Approximately 24% of *de-novo*-assembled *Harpegnathos* and *Camponotus* transcripts met this requirement (Figure S5B).

We filtered our non-coding annotations using PhyloCSF (Lin et al., 2011). Most protein-coding genes in both *Harpegnathos* and *Camponotus* had positive PhyloCSF scores, indicative of a coding model, whereas our newly annotated putative non-coding transcripts were skewed toward negative, non-coding scores (Figure 5A). After filtering by PhyloCSF score, 628 (28.2%) and 683 (30.1%) of the putative non-coding genes in *Harpegnathos* and *Camponotus*, respectively, were retained. We then removed lncRNA gene models with splice junctions to adjacent protein-coding genes, as they might constitute 5′ or 3′ UTRs missed by our protein-coding annotation pipeline (You et al., 2017) (Figure 5B). We also removed lncRNA models containing open reading frames to which we could assign PFAM domains or peptides from mass spectrometry (Figure 5B). We did not consider these models for protein-coding annotations.

At the end of all filtering steps, we annotated 438 and 359 high-confidence lncRNA gene models for *Harpegnathos* and *Camponotus*, respectively (Figures 5B and S5A), which we subdivided according to their spatial relationship to neighboring protein-coding gene models into intervening, promoter-associated, and intronic (Figures S5C and S5D), all of which showed a lack of coding potential, even when considered separately (Figure S6A). We could not detect a substantial number of antisense lncRNAs overlapping exons of protein-coding genes.

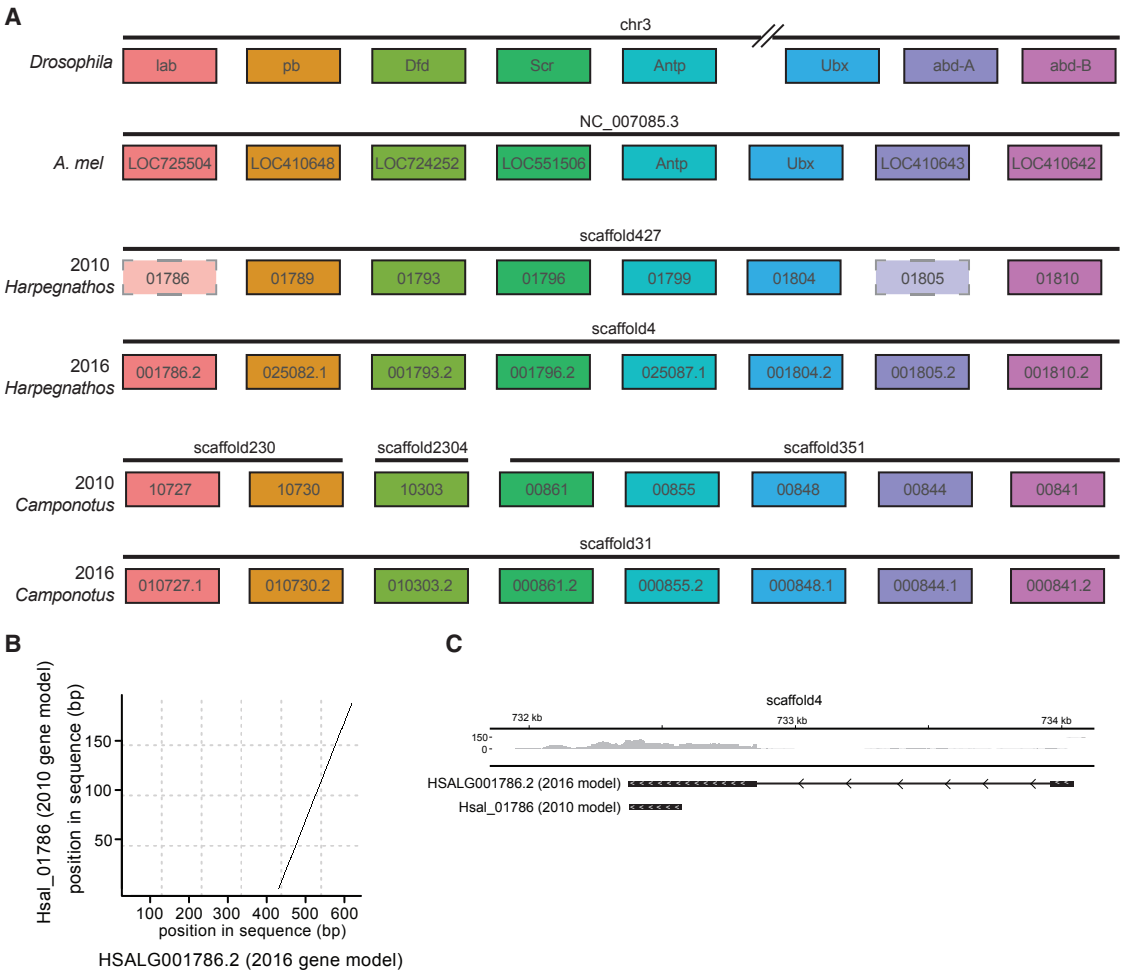


Figure 4. Reassembly of the *Hox* Clusters of *Camponotus* and *Harpegnathos*

(A) Scheme of *Hox* gene organization in (from the top) *Drosophila*, *Apis mellifera*, *Harpegnathos* (old and new assembly), and *Camponotus* (old and new assembly).

(B) Example of a *Hox* gene in *Harpegnathos* updated in 2016 annotation. The 2010 gene model is depicted on the y axis, with the 2016 gene model on the x axis. Dots in the plot indicate regions of significant sequence similarity between 2010 and 2016 models.

(C) RNA-seq from various developmental stages in *Harpegnathos* shows extension of the gene model past the 2010 boundaries. The 2010 and 2016 gene models are shown under the RNA-seq coverage track. Scale on RNA-seq track indicates reads per million.

lncRNAs in other organism are less conserved, are shorter, have fewer exons, and, overall, are expressed at lower levels than protein-coding genes (Quinn and Chang, 2016). We detected most of these features in our ant lncRNAs; they were less conserved than protein-coding genes (Figure 5C), regardless of their genomic localization (Figure S6B); they had a smaller number of exons (Figure S6C); and they were expressed at lower levels than protein-coding genes (Figure S6D). However, the length distribution of the ant lncRNAs was similar to that of protein-coding genes (Figure S6E), which was a departure from what was observed in mammals, *Drosophila*, and *C. elegans* (Cabili et al., 2011; Nam and Bartel, 2012; Young et al., 2012). lncRNAs in other genomes tend to overlap with transposable elements at a higher rate than protein-coding genes, suggesting a role for these sequences in their function and diversification (Kapusta et al., 2013; Kelley and Rinn, 2012). We observe this in ant lncRNAs as well (Figure S6F).

Expression Patterns of lncRNAs

If our lncRNA gene models comprise functional loci with potential for epigenetic regulation we should be able to observe their differential expression in a number of relevant comparisons, such as through developmental stages, in different tissues, and perhaps even the same tissue from different castes.

We determined whether lncRNA transcription was differentially regulated during life transitions in *Harpegnathos*. We analyzed whole-body RNA-seq datasets from embryos, larvae, pupae, and adult workers. We clustered relative changes in the expression levels of lncRNAs across these samples into groups with distinct kinetics (Figure 6A), which allowed us to identify

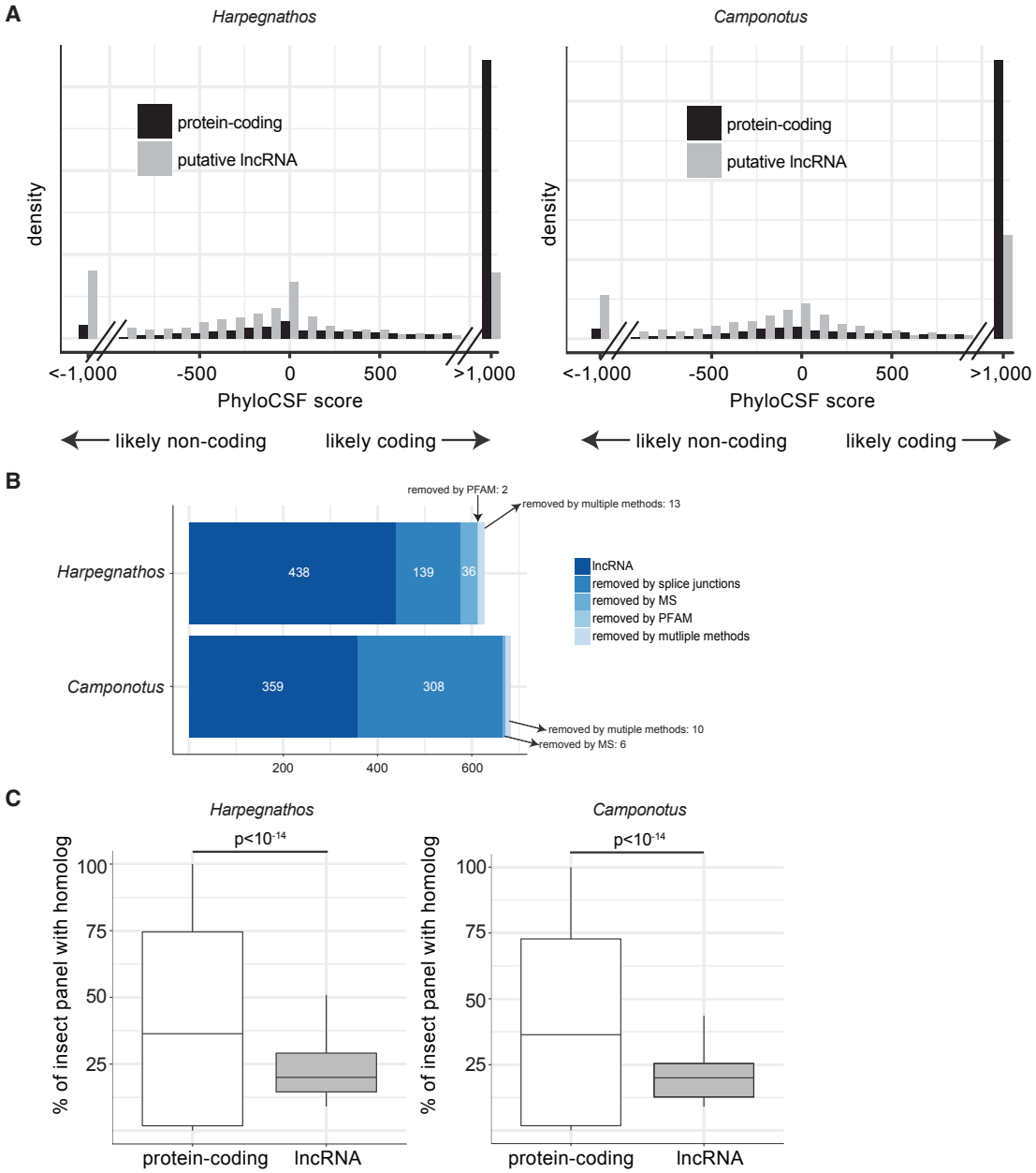


Figure 5. Annotation of Long Non-Coding RNAs

(A) PhyloCSF scores for transcripts with no overlap to coding sequences (gray) and known protein-coding genes (black). The x axis indicates the PhyloCSF scores in decibans, which represents the likelihood ratio of a coding model versus a non-coding model. Negative values indicate that a gene model is more likely to be non-coding than coding.

(B) Filtering of lncRNA using stranded, spliced RNA-seq reads and mass spectrometry.

(C) Boxplot for the number of homologs (BLASTN e-value $<10^{-3}$) found in other insect genomes for lncRNAs compared to protein-coding gene models.

See also Figures S5 and S6.

early development lncRNAs (Figure 6A, clusters 1–4), adult lncRNAs (clusters 8–10), and a set of lncRNAs predominantly expressed in the pupal stage (clusters 6 and 7), a critical phase in the life of holometabolous insects characterized by pronounced cell proliferation, morphogenesis, and neuronal remodeling.

We also identified lncRNAs with tissue-specific expression in *Harpegnathos* adults by comparing the transcriptomes of antenna, non-visual brain (the central part of the brain after removal of the optic lobes; Gospocic et al., 2017), fat body, and ovary. Many lncRNAs were expressed specifically in one tissue, especially the brain (Figure 6B), perhaps indicating a dedicated

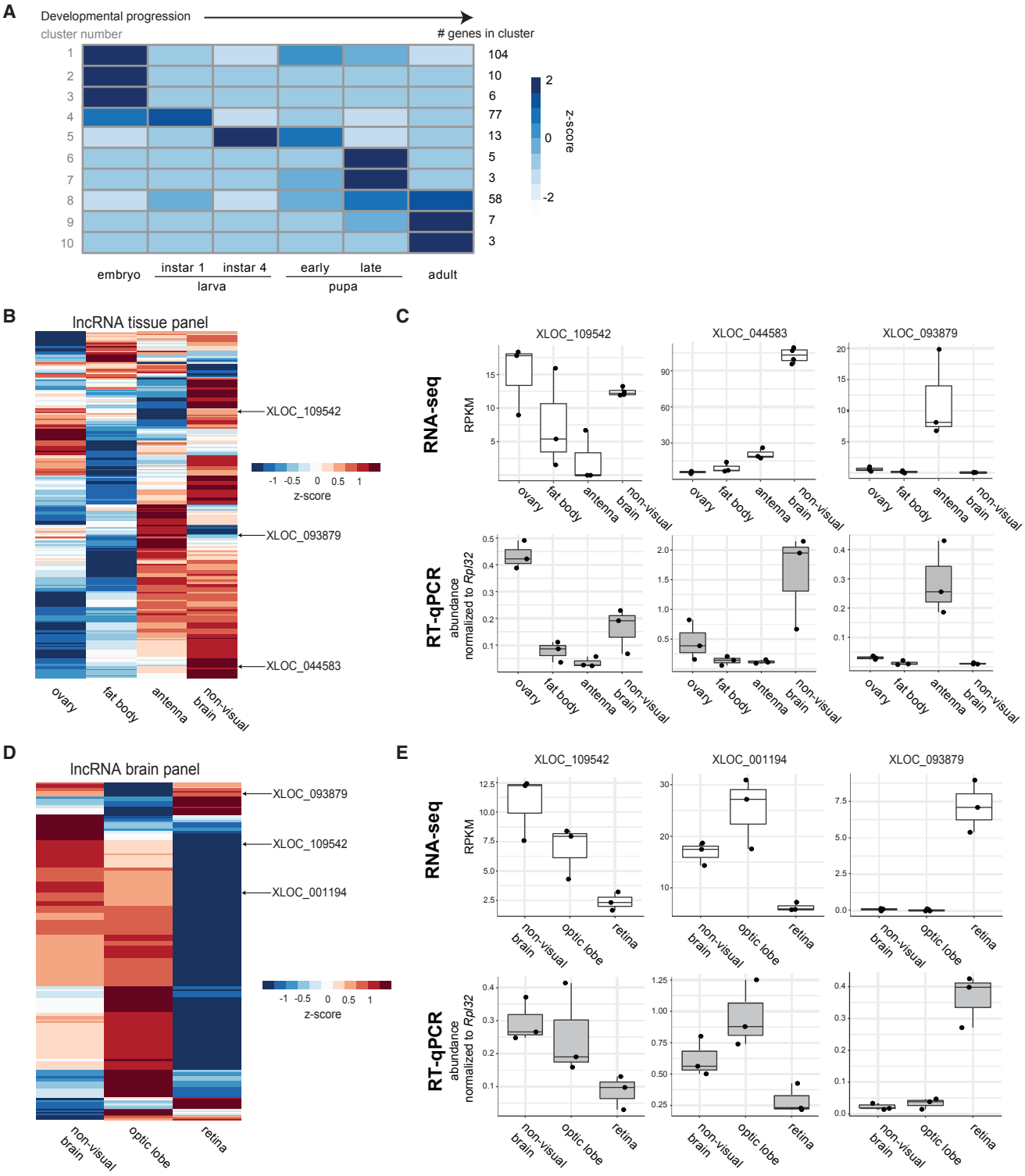


Figure 6. Differential Expression of LncRNAs in *Harpegnathos* Developmental Stages, Tissues, and Brain Regions

(A) K-means clustering of changes in lncRNA expression across the indicated developmental stages (all n = 2). The cluster number is displayed to the left of the heatmap, while the number of lncRNAs in each cluster is shown to the right.

(B) Heatmap of lncRNA expression patterns from RNA-seq in ovary, fat body, antenna, and non-visual brain (all n = 3). Heatmap shows *Z* scores of log(RPKMs) (read per kilobase per million) by row. Arrows point to lncRNAs that have expression specific to one or more tissues.

(C) RNA-seq and qRT-PCR for the three lncRNAs highlighted in (B).

(legend continued on next page)

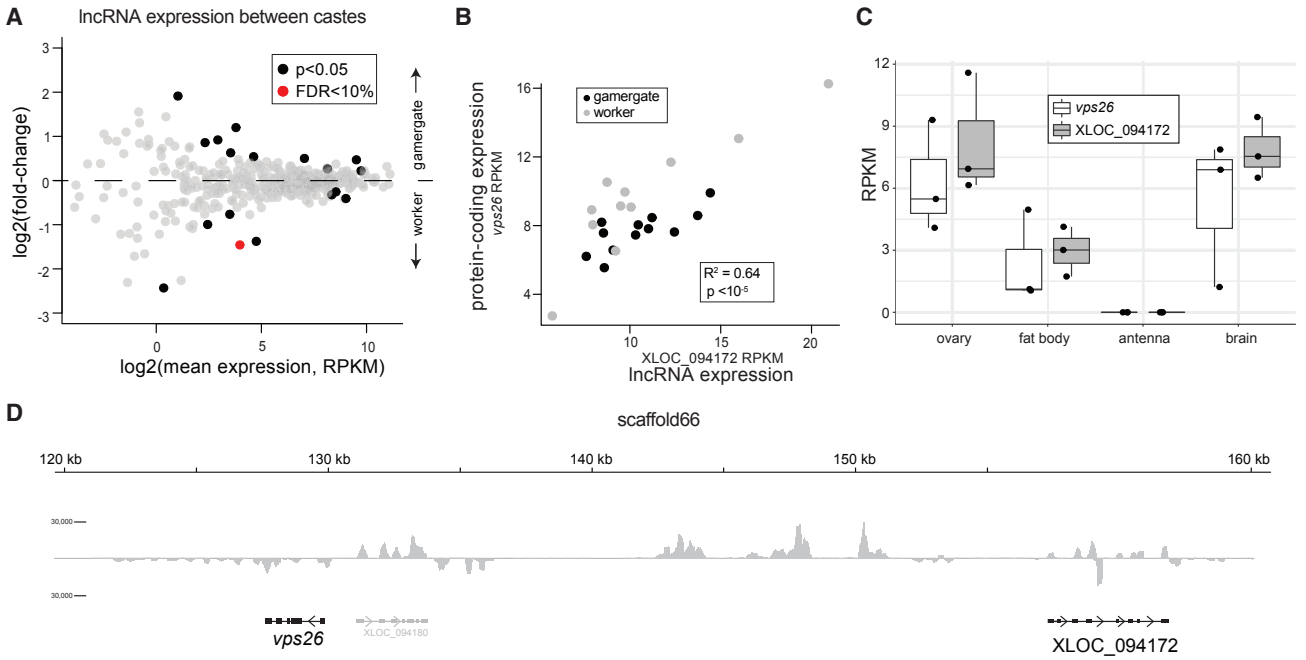


Figure 7. Differential LncRNA Expression and LncRNA/Protein-Coding Co-regulation in *Harpegnathos* Castes

(A) MA plot of lncRNAs in RNA-seq data comparing worker and gamergates (Gospocic et al., 2017). Genes with unadjusted p < 0.05 are highlighted in black, genes with <10% false discovery rate (FDR) in red. Data are from ≥10 biological replicates per condition (individual ants; worker, n = 11; gamergate, n = 12).

(B) The expression levels of XLOC_094172 lncRNA (x axis) and the protein-coding gene *vps26* (y axis) correlate in both gamergate and worker. Each dot represents one biological sample (worker, n = 11; gamergate, n = 12). p value from Pearson correlation is indicated.

(C) Expression patterns of XLOC_094172 and *vps26* in worker brains by RNA-seq in non-visual brain, optic lobe, and retina (all n = 3).

(D) Positions of XLOC_094172 and *vps26* on scaffold66, with RNA-seq coverage from combined workers (n = 11) and gamergates (n = 12). Scale on RNA-seq track indicates reads per million.

See also Figure S7.

function. We validated the tissue-specific expression of three lncRNAs by RT-qPCR (Figure 6C): XLOC_109542, which showed higher expression in ovary and non-visual brain; XLOC_044583, with highest expression in the brain; and XLOC_093879, which was restricted to the antenna (and the retina; see below).

As lncRNAs have been previously shown to be expressed in different regions of the mouse brain (Mercer et al., 2008), we compared lncRNA expression levels in RNA-seq data from non-visual brain, optic lobe, and retina. We used region-specific controls *corazonin* (non-visual brain), *Gabbr2* (optic lobe), and *Arr2* (retina) to ensure that our dissections had been performed accurately (Figure S7A). We detected many lncRNAs with higher expression in one region of the brain (Figure 6D), and validated three by RT-qPCR (Figure 6E): XLOC_109542, which was expressed at higher levels in the non-visual brain; XLOC_001194, expressed at higher levels in the optic lobe; and XLOC_093879, restricted to the retina (and the antenna, see above).

We previously showed that the adult caste transition between worker and gamergates in *Harpegnathos* is accompanied by major changes in protein-coding gene expression (Gospocic et al., 2017). We reanalyzed that dataset in the context of our new lncRNA annotation and found 17 lncRNAs that were differentially expressed in worker and gamergate brains with a p value cutoff of 0.05 (Figure 7A). We also looked for lncRNAs that might be responsible for co-regulating a protein-coding gene. XLOC_094172 caught our attention because its expression strongly correlated with that of the neighboring protein-coding gene *vps26* (Figures 7B, 7C, and S7B), a subunit of the retromer complex implicated in neurological disorders (Linhart et al., 2014; McMillan et al., 2017). This lncRNA and its co-regulated protein-coding gene are ~22 kb apart and on opposite strands (Figure 7D), excluding the possibility that they are spanned by the same primary transcript. Instead, we propose that this lncRNA controls expression of the protein-coding gene, as is the case for several *cis*-acting lncRNAs in other organisms (Engreitz et al., 2016b).

(D) Heatmap of lncRNA expression patterns from RNA-seq in non-visual brain, optic lobe, and retina (all n = 3). Heatmap shows *Z* scores of log(RPKMs) by row. Arrows point to lncRNAs that have expression specific to one or more regions.

(E) RNA-seq and qRT-PCR for the three lncRNAs highlighted in (D).

See also Figure S7.

We also confirmed the expression in the brain of lncRNAs with homologs in other insects (Jayakodi et al., 2015; Li et al., 2012) (Figure S7C). Most notably, the *Harpegnathos* homolog of CASK regulatory gene (CRG), a lncRNA involved in locomotor behavior in *Drosophila* (Li et al., 2012), was expressed in neurons throughout the brain, as demonstrated by RNA-seq analyses (Figure S7D; XLOC_081169) and by its co-localization with the pan-neuronal marker *elav* by fluorescence *in situ* hybridization (FISH) (Figures S7D and S7E).

To confirm that our example lncRNAs are bona fide lncRNAs, we utilized an orthogonal method of measuring coding potential used in other lncRNA annotations (Jayakodi et al., 2015; Nam and Bartel, 2012; Ulitsky et al., 2011; Young et al., 2012), the Coding Potential Calculator (CPC) (Kong et al., 2007). CPC scores correlated strongly with PhyloCSF scores ($p < 10^{-15}$ for both *Harpegnathos* and *Camponotus*) and scored as non-coding all lncRNAs shown in Figures 6, 7, and S7. The accuracy of our lncRNA gene models was further confirmed by the fact that all qRT-PCR reactions yielded products of the expected size (Figure S7F).

Thus, our improved genome assemblies allowed us to annotate lncRNAs, several of which displayed developmental-, brain-, or caste-specific expression patterns, which suggests that they might have important roles in development and brain function.

DISCUSSION

Social insects offer a unique perspective from which to study epigenetics (Bonasio, 2012; Yan et al., 2014). Striking morphological and behavioral differences between castes include phenotypes relevant to translational research, such as social behavior, aging, and development. These traits can be studied on an organism level within a natural social context, as full colonies can be maintained in the laboratory. However, to analyze these complex traits at a molecular level, proper genomic tools must be developed. We previously assembled ant genomes generating a workable draft using the best technology at the time, whole-genome shotgun using short Illumina reads (Bonasio et al., 2010); however, the fragmented nature of these draft genomes presented an obstacle to epigenomic studies.

Here, we used PacBio long reads to reassemble *de novo* the genomes of the two ant species currently in use as models in our laboratory, *Camponotus floridanus* and *Harpegnathos saltator*, and produced accurate assemblies with scaffold N50 sizes larger than 1 Mb and a number of gaps smaller than in all other insect genomes available on NCBI at the time of writing (Figure 1E). Although other insect assemblies have larger scaffold N50s than our new ant assemblies, which might be helpful for evaluating structural variations and interactions at great length scales, many *cis* regulatory and epigenomic mechanisms take place at short-to-medium range and their study is facilitated by longer gap-free regions of sequence (i.e., longer contigs). Thus, we prioritized contig length in our new assemblies, and chose to pursue greater PacBio sequencing coverage rather than techniques used to improve scaffold N50, such as optical mapping and proximity ligation.

Our greatly improved *Harpegnathos* and *Camponotus* assemblies deliver several critical benefits to further development of these ant species into molecular model organisms: (1) more comprehensive protein-coding annotations and more complete gene models (Figure 3; Table S3), (2) more continuity of co-regulated gene clusters (Figure 4), (3) high-quality lncRNA annotations (Figure 5), and (4) the ability to detect regulatory mechanisms functioning in *cis* at distances of 10–100 kb (Figure 7).

Although the annotation of protein-coding genes did not suffer excessively from the draft status of the 2010 assemblies, the new annotations contain potentially relevant genes that were previously missing. Most notably, a *Gp-9-like* gene previously unannotated in the *Harpegnathos* genome was found to be differentially expressed in worker brains compared to gamergates (Figure 3C). The importance of *Gp-9* in ant biology is well established, as it was one of the first genetic markers discovered in ants for a colony-level phenotype. In the fire ant *Solenopsis invicta*, *Gp-9* maps to a cluster of genes involved in a large genomic rearrangement that governs the choice between a polygyne (multiple queens) or monogyne (one queen) colony (Ross, 1997; Wang et al., 2013). We found that a *Gp-9-like* homolog is expressed at different levels in *Harpegnathos* castes as well as in three other ant species with different social structures (Figures S4C–S4E), suggesting a conserved role for this gene in colony organization and opening an avenue for future investigation on its molecular function.

Another advance granted by our improved genome assemblies was the ability to annotate lncRNAs. We developed a custom pipeline and discovered over 300 high-confidence lncRNAs in both *Harpegnathos* and *Camponotus*. The mechanism of action and biological impact of lncRNAs is the subject of intense investigation in various model systems and in several cases a dedicated role in brain function has been advocated, based in part on their expression patterns (Bonasio, 2012; Bonasio and Shiekhattar, 2014).

Most lncRNAs are believed to act in *cis* to regulate expression of neighboring genes (Bonasio and Shiekhattar, 2014; Engreitz et al., 2016b; Lee, 2012); therefore, an extended view of protein-coding genes in the vicinity of lncRNAs is critical to understand their regulatory role, and this information is provided by our updated genomes. Thanks to the increased continuity of the new assemblies, we were able to identify a lncRNAs-mRNA pair whose brain expression patterns were correlated, suggesting a potential regulatory relationship (Figures 7B–7D). Similar cases have been described in mammals; one lncRNA, *HAR1F*, is co-expressed in human Cajal-Retzius neurons with *reelin*, a protein-coding gene that regulates cortical development (Pollard et al., 2006). A murine lncRNA, *Dali*, regulates the expression of the nearby transcription factor *Pou3f3*, which is involved in nerve and growth development (Chalei et al., 2014). Our findings on brain- and caste-specific lncRNAs as well as lncRNAs co-expressed with protein-coding gene will allow us to prioritize candidates for future studies on the neuroepigenetic functions of lncRNAs in ants. This prospect is particularly intriguing in *Harpegnathos*, where we discovered major transcriptional changes that accompany the rewiring of the brain during adult caste transitions (Gospocic et al., 2017) and where we recently showed the feasibility of genetic manipulation of the germline (Yan et al., 2017).

EXPERIMENTAL PROCEDURES

Genome Assembly Strategy

The reads of insert extracted from raw PacBio data were error corrected, trimmed, and assembled by Canu v1.3 (Koren et al., 2017). Quiver was used to polish the assemblies, which were then scaffolded with extracted subreads from the PacBio data using PBJelly (English et al., 2012), and with mate pairs using SSpace-Standard (Boetzer et al., 2011). The assemblies were polished with paired-end Illumina short reads using Pilon (Walker et al., 2014).

Annotation of Protein-Coding Genes

Protein-coding genes were annotated on the *Harpegnathos* and *Camponotus* assemblies using iterations of the MAKER2 pipeline. The MAKER2 pipeline was run four times, each step updated with hidden Markov models trained on the previous step. On the fourth run, gene models produced directly from RNA-seq and homology were reported, and all gene models were filtered using the annotation edit distance and the presence of a PFAM domain, as detailed in Supplemental Experimental Procedures.

Annotation of lncRNA Genes

RNA-seq reads from various developmental stages of *Harpegnathos* and *Camponotus* were assembled using two reference-based transcriptome assemblers, Trinity and Stringtie. Transcripts common among the two methods that did not overlap with protein-coding genes were designated as putative lncRNAs. LncRNAs were further filtered using the PhyloCSF Omega Test (Lin et al., 2011), mass spectrometry, and presence of splice junctions to protein-coding genes.

qPCR

For qRT-PCR, 1 ng RNA was assayed per 10 μL reaction using the RNA-to-Ct single-step kit (Thermo Fisher). The RNA for *Rpl32*, encoding a ribosomal protein, was used as a normalization control.

Heatmaps and Clustering of lncRNA Expression Levels

Expression patterns of differentially expressed lncRNAs in the developmental stages of *Harpegnathos* were clustered based on the quantile-normalized log-fold expression changes between each pair of samples. K-means clustering with a preset number of clusters (10) and maximum number of iterations (50) was performed on this quantile-normalized matrix. Heatmaps were plotted using the pheatmap package for R, with color scaling by row.

In Situ Hybridization

500-bp DNA probes were designed against XLOC_081169 and included T7 (sense) and SP6 (anti-sense) RNA polymerase promoters. RNA *in situ* hybridization were performed according to published protocols (Morris et al., 2009; Søe et al., 2011), with minor modifications. Chromogenic ISH sections were imaged with a DS-Ri1 Digital Microscope Camera from Nikon. Fluorescent ISH sections were imaged with a Leica SPE laser scanning confocal microscope.

Sequencing Data

The accession number for the RNA-sequencing data generated for this study is GEO: SuperSeries GSE102605. The accession number for the raw PacBio reads of the insert, as well as assembled genomes, is BioProject: PRJNA445978.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and three tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.05.014.

ACKNOWLEDGMENTS

The authors thank Janko Gospocic for providing ant samples; Karl Glastad for helpful comments on genome assembly; Katy Munson and the UW PacBio Sequencing service for performing PacBio sequencing; Shawn Little for helping with FISH; and Yoseph Barash, Ben Voight, and Paul Babb for comments on the manuscript. R.B. thanks Danny Reinberg (NYU) as well as Guojie Zhang, Cai Li, Zhensheng Chen, and Luohao Xu (BGI) for their intellectual support and efforts during an initial attempt at annotating lncRNAs. R.B. acknowledges financial support from the NIH (DP2MH107055), the Searle Scholars Program (15-SSP-102), the Linda Pechenik Montague Investigator Award, and the Charles E. Kaufman Foundation (KA2016-85223). E.J.S. acknowledges financial support from the NIH (T32HG000046). B.A.G. was supported by the NIH (GM110174 and HL122993).

AUTHOR CONTRIBUTIONS

A.K.W. performed mass spectrometry experiments and analysis in B.A.G.'s lab. L.S. performed brain and tissue dissections and carried out *in situ* hybridization experiments. All remaining experiments and analyses were performed by E.J.S. The manuscript was written by E.J.S. and R.B., with input from L.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 28, 2017
Revised: April 4, 2018
Accepted: May 3, 2018
Published: June 5, 2018

REFERENCES

Allen, S.L., Delaney, E.K., Kopp, A., and Chenoweth, S.F. (2017). Single-molecule sequencing of the *Drosophila serrata* genome. G3 (Bethesda) *7*, 781–788.

Allis, C.D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nat. Rev. Genet. *17*, 487–500.

Amaral, P.P., Leonardi, T., Han, N., Viré, E., Gascoigne, D.K., Arias-Carrasco, R., Büscher, M., Pandolfini, L., Zhang, A., Pluchino, S., et al. (2018). Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. Genome Biol. *19*, 32.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.

Bonasio, R. (2012). Emerging topics in epigenetics: ants, brains, and noncoding RNAs. Ann. N Y Acad. Sci. *1260*, 14–23.

Bonasio, R., and Shiekhattar, R. (2014). Regulation of transcription by long noncoding RNAs. Annu. Rev. Genet. *48*, 433–455.

Bonasio, R., Zhang, G., Ye, C., Mutti, N.S., 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, N.S., 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.

Bonasio, R., Lecona, E., Narendra, V., Voigt, P., Parisi, F., Kluger, Y., and Reinberg, D. (2014). Interactions with RNA direct the Polycomb group protein SCML2 to chromatin where it represses target genes. eLife *3*, e02637.

Brown, S.J., Fellers, J.P., Shippy, T.D., Richardson, E.A., Maxwell, M., Stuart, J.J., and Denell, R.E. (2002). Sequence of the Tribolium castaneum homeotic complex: the region corresponding to the Drosophila melanogaster antennapedia complex. Genetics *160*, 1067–1074.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Chalei, V., Sansom, S.N., Kong, L., Lee, S., Montiel, J.F., Vance, K.W., and Ponting, C.P. (2014). The long non-coding RNA Dali is an epigenetic regulator of neural differentiation. eLife *3*, e04530.

Dearden, P.K., Wilson, M.J., Sablan, L., Osborne, P.W., Havler, M., McNaughton, E., Kimura, K., Milshina, N.V., Hasselmann, M., Gempe, T., et al. (2006). Patterns of conservation and change in honey bee developmental genes. Genome Res. *16*, 1376–1384.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Devenport, M.P., Blass, C., and Eggleston, P. (2000). Characterization of the Hox gene cluster in the malaria vector mosquito, Anopheles gambiae. Evol. Dev. *2*, 326–339.

English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., and Gibbs, R.A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE *7*, e47768.

Engreitz, J.M., Ollikainen, N., and Guttman, M. (2016a). Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. Nat. Rev. Mol. Cell Biol. *17*, 756–770.

Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016b). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.

Ferrier, D.E., and Akam, M. (1996). Organization of the Hox gene cluster in the grasshopper, Schistocerca gregaria. Proc. Natl. Acad. Sci. USA *93*, 13024–13029.

Finnerty, J.R., and Martindale, M.Q. (1998). The evolution of the Hox cluster: insights from outgroups. Curr. Opin. Genet. Dev. *8*, 681–687.

Gerstein, M.B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J.B., Davis, C.A., Hillier, L., Sisu, C., Li, J.J., et al. (2014). Comparative analysis of the transcriptome across distant species. Nature *512*, 445–448.

Gospocic, J., Shields, E.J., Glastad, K.M., Lin, Y., Penick, C.A., Yan, H., Mikheyev, A.S., Linksvayer, T.A., Garcia, B.A., Berger, S.L., et al. (2017). The neuropeptide Corazonin controls social behavior and caste identity in ants. Cell *170*, 748–759.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300.

He, C., Sidoli, S., Warneford-Thomson, R., Tatomer, D.C., Wilusz, J.E., Garcia, B.A., and Bonasio, R. (2016). High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells. Mol. Cell *64*, 416–430.

Hendrickson, D., Kelley, D.R., Tenen, D., Bernstein, B., and Rinn, J.L. (2016). Widespread RNA binding by chromatin-associated proteins. Genome Biol. *17*, 1–18.

Holoch, D., and Moazed, D. (2015). RNA-mediated epigenetic regulation of gene expression. Nat. Rev. Genet. *16*, 71–84.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics *12*, 491.

Jayakodi, M., Jung, J.W., Park, D., Ahn, Y.-J., Lee, S.-C., Shin, S.-Y., Shin, C., Yang, T.-J., and Kwon, H.W. (2015). Genome-wide characterization of long intergenic non-coding RNAs (lincRNAs) provides new insight into viral diseases in honey bees Apis cerana and Apis mellifera. BMC Genomics *16*, 680.

Joung, J., Engreitz, J.M., Konermann, S., Abudayyeh, O.O., Verdine, V.K., Aguet, F., Gootenberg, J.S., Sanjana, N.E., Wright, J.B., Fulco, C.P., et al. (2017). Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. Nature *548*, 343–346.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. *9*, e1003470.

Kelley, D., and Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. *13*, R107.

Kmita, M., and Duboule, D. (2003). Organizing axes in time and space; 25 years of colinear tinkering. Science *301*, 331–333.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. *35*, W345-9.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Lai, F., Orom, U.A., Cesaroni, M., Beringer, M., Taatjes, D.J., Blobel, G.A., and Shiekhattar, R. (2013). Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. Nature *494*, 497–501.

Lee, J.T. (2012). Epigenetic regulation by long noncoding RNAs. Science *338*, 1435–1439.

Li, M., Wen, S., Guo, X., Bai, B., Gong, Z., Liu, X., Wang, Y., Zhou, Y., Chen, X., Liu, L., and Chen, R. (2012). The novel long non-coding RNA CRG regulates Drosophila locomotor behavior. Nucleic Acids Res. *40*, 11714–11727.

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275–i282.

Linhart, R., Wong, S.A., Cao, J., Tran, M., Huynh, A., Ardrey, C., Park, J.M., Hsu, C., Taha, S., Peterson, R., et al. (2014). Vacuolar protein sorting 35 (Vps35) rescues locomotor deficits and shortened lifespan in Drosophila expressing a Parkinson's disease mutant of leucine-rich repeat kinase 2 (LRRK2). Mol. Neurodegener. *9*, 23.

McMillan, K.J., Korswagen, H.C., and Cullen, P.J. (2017). The emerging role of retromer in neuroprotection. Curr. Opin. Cell Biol. *47*, 72–82.

Mercer, T.R., Dinger, M.E., Sunkin, S.M., Mehler, M.F., and Mattick, J.S. (2008). Specific expression of long noncoding RNAs in the mouse brain. Proc. Natl. Acad. Sci. USA *105*, 716–721.

Miller, J.R., Koren, S., Dilley, K.A., Puri, V., Brown, D.M., Harkins, D.M., Thibaud-Nissen, F., Rosen, B., Chen, X.-G., Tu, Z., et al. (2018). Analysis of the Aedes albopictus C6/36 genome provides insight into cell line utility for viral propagation. Gigascience *7*, 1–13.

Miyakawa, M.O., and Mikheyev, A.S. (2015). QTL mapping of sex determination loci supports an ancient pathway in ants and honey bees. PLoS Genet. *11*, e1005656.

Morris, C.A., Benson, E., and White-Cooper, H. (2009). Determination of gene expression patterns using in situ hybridization to Drosophila testes. Nat. Protoc. *4*, 1807–1819.

Nam, J.-W., and Bartel, D.P. (2012). Long noncoding RNAs in C. elegans. Genome Res. *22*, 2529–2540.

Negre, B., Casillas, S., Suzanne, M., Sánchez-Herrero, E., Akam, M., Nefedov, M., Barbadilla, A., de Jong, P., and Ruiz, A. (2005). Conservation of regulatory sequences and gene expression patterns in the disintegrating Drosophila Hox gene complex. Genome Res. *15*, 692–700.

Patalano, S., Vlasova, A., Wyatt, C., Ewels, P., Camara, F., Ferreira, P.G., Asher, C.L., Jurkowski, T.P., Segonds-Pichon, A., Bachman, M., et al. (2015). Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. Proc. Natl. Acad. Sci. USA *112*, 13970–13975.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Pervouchine, D.D., Djebali, S., Breschi, A., Davis, C.A., Barja, P.P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L.H., et al. (2015). Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat. Commun. *6*, 5903.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211.

Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Quinn, J.J., and Chang, H.Y. (2016). Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. *17*, 47–62.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Roberts, R.J., Carneiro, M.O., and Schatz, M.C. (2013). The advantages of SMRT sequencing. Genome Biol. *14*, 405.

Ross, K.G. (1997). Multilocus evolution in fire ants: effects of selection, gene flow and recombination. Genetics *145*, 961–974.

Saldaña-Meyer, R., González-Buendía, E., Guerrero, G., Narendra, V., Bonasio, R., Recillas-Targa, F., and Reinberg, D. (2014). CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. Genes Dev. *28*, 723–734.

Schuettengruber, B., Bourbon, H.M., Di Croce, L., and Cavalli, G. (2017). Genome regulation by Polycomb and Trithorax: 70 years and counting. Cell *171*, 34–57.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

Simola, D.F., Ye, C., Mutti, N.S., Dolezal, K., Bonasio, R., Liebig, J., Reinberg, D., and Berger, S.L. (2013a). A chromatin link to caste identity in the carpenter ant Camponotus floridanus. Genome Res. *23*, 486–496.

Simola, D.F., Wissler, L., Donahue, G., Waterhouse, R.M., Helmkampf, M., Roux, J., Nygaard, S., Glastad, K.M., Hagen, D.E., Viljakainen, L., et al. (2013b). Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. Genome Res. *23*, 1235–1247.

Simola, D.F., Graham, R.J., Brady, C.M., Enzmann, B.L., Desplan, C., Ray, A., Zwiebel, L.J., Bonasio, R., Reinberg, D., Liebig, J., and Berger, S.L. (2016). Epigenetic (re)programming of caste-specific behavior in the ant Camponotus floridanus. Science *351*, aac6633.

Søe, M.J., Møller, T., Dufva, M., and Holmstrøm, K. (2011). A sensitive alternative for microRNA in situ hybridizations using probes of 2′-O-methyl RNA + LNA. J. Histochem. Cytochem. *59*, 661–672.

Sun, S., Del Rosario, B.C., Szanto, A., Ogawa, Y., Jeon, Y., and Lee, J.T. (2013). Jpx RNA activates Xist by evicting CTCF. Cell *153*, 1537–1551.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537–1550.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE *9*, e112963.

Wang, J., Wurm, Y., Nipitwattanaphon, M., Riba-Grognuz, O., Huang, Y.-C., Shoemaker, D., and Keller, L. (2013). A Y-like social chromosome causes alternative colony organization in fire ants. Nature *493*, 664–668.

Wang, L., Zhao, Y., Bao, X., Zhu, X., Kwok, Y.K.Y., Sun, K., Chen, X., Huang, Y., Jauch, R., Esteban, M.A., et al. (2015). LncRNA Dum interacts with Dnmts to regulate Dppa2 expression during myogenic differentiation and muscle regeneration. Cell Res. *25*, 335–350.

Warner, M.R., Mikheyev, A.S., and Linksvayer, T.A. (2017). Genomic signature of kin selection in an ant with obligately sterile workers. Mol. Biol. Evol. *34*, 1780–1787.

Yan, H., Simola, D.F., Bonasio, R., Liebig, J., Berger, S.L., and Reinberg, D. (2014). Eusocial insects as emerging models for behavioural epigenetics. Nat. Rev. Genet. *15*, 677–688.

Yan, H., Opachaloemphan, C., Mancini, G., Yang, H., Gallitto, M., Mlejnek, J., Leibholz, A., Haight, K., Ghaninia, M., Huo, L., et al. (2017). An engineered orco mutation produces aberrant social behavior and defective neural development in ants. Cell *170*, 736–747.

Yang, Y.W., Flynn, R.A., Chen, Y., Qu, K., Wan, B., Wang, K.C., Lei, M., and Chang, H.Y. (2014). Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. eLife *3*, e02046.

Yasukochi, Y., Ashakumary, L.A., Wu, C., Yoshido, A., Nohata, J., Mita, K., and Sahara, K. (2004). Organization of the Hox gene cluster of the silkworm, Bombyx mori: a split of the Hox cluster in a non-Drosophila insect. Dev. Genes Evol. *214*, 606–614.

You, B.H., Yoon, S.H., and Nam, J.W. (2017). High-confidence coding and noncoding transcriptome maps. Genome Res. *27*, 1050–1062.

Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.L., and Ponting, C.P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the Drosophila melanogaster genome. Genome Biol. Evol. *4*, 427–442.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

Cell Reports, Volume *23*

Supplemental Information

High-Quality Genome Assemblies Reveal Long Non-coding RNAs Expressed in Ant Brains

Emily J. Shields, Lihong Sheng, Amber K. Weiner, Benjamin A. Garcia, and Roberto Bonasio

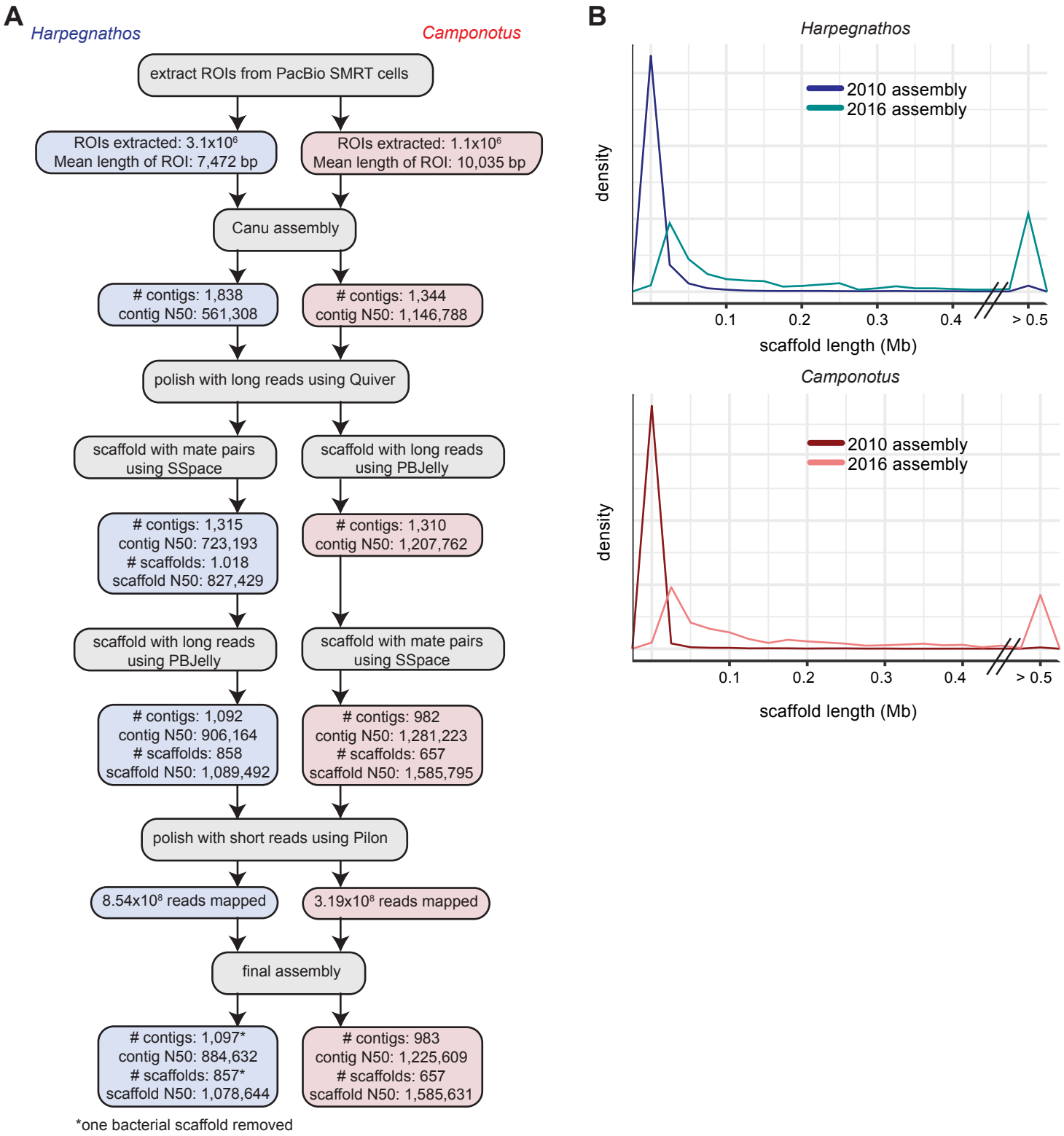


Figure S1. Assembly Pipeline and Associated Metrics, Related to Figure 1

(A) Steps performed at each point of the assembly process are listed along with relevant metrics.

(B) Density plots of scaffold lengths for *Harpegnathos* (top) and *Camponotus* (bottom) assemblies. 2010 assemblies have many short scaffolds, as shown by the large peaks below 0.1 Mb. In contrast, the 2016 assemblies have a greater number of longer scaffolds, with many scaffolds larger than 0.5 Mb.

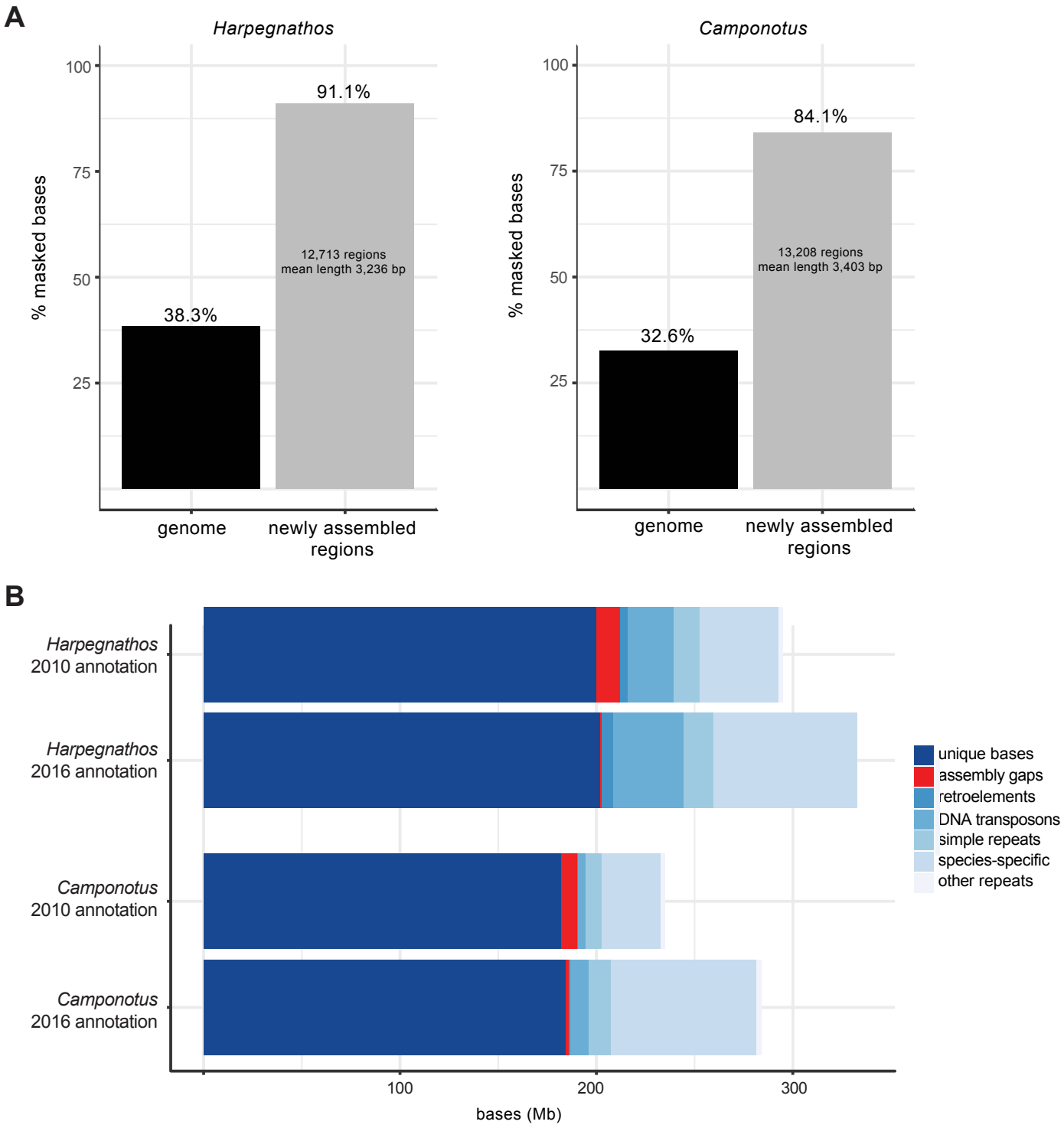


Figure S2. Repeat Content in 2010 and 2016 Assemblies, Related to Figure 1

(A) The percentage of masked bases is given for the whole genome and "newly assembled regions," which is defined as any stretch of the 2016 genome assembly with a >1 kb gap in matched 2010 assembly sequence. The new sequence content of the 2016 *Harpegnathos* (left) and *Camponotus* (right) assemblies contains a greater percentage of bases masked compared to background genome levels.

(B) 2016 *Harpegnathos* and *Camponotus* assemblies capture more repeat content than 2010 assemblies and have a comparable number of unique bases. Number of bases of the genome assigned to various repeat categories by RepeatMasker using "harpegnathos saltator" as the species, with additional species-specific repeat libraries constructed using RepeatScout, in *Harpegnathos* and *Camponotus* 2010 and 2016 assemblies. Species-specific repeats were detected using 2016 assemblies.

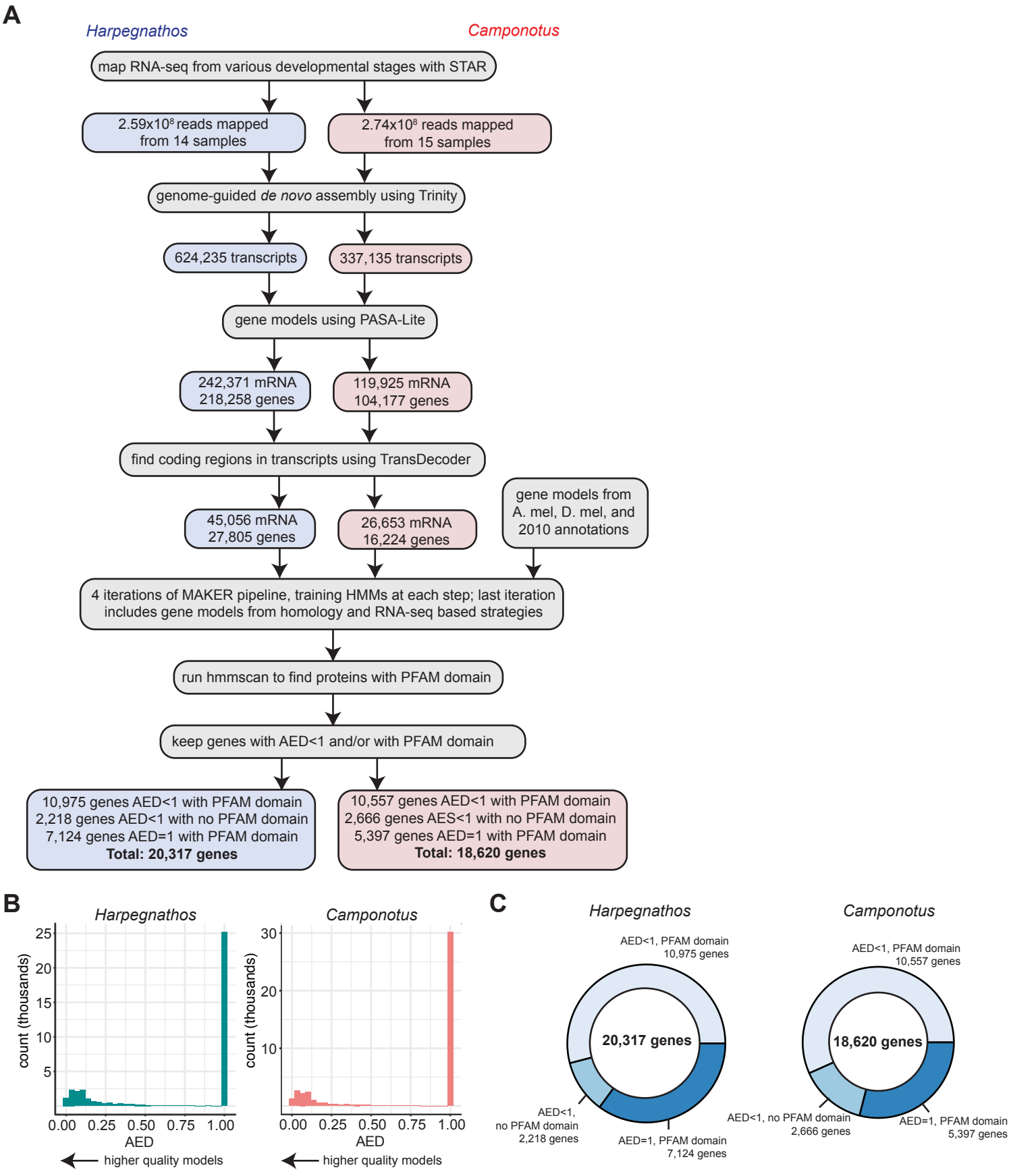


Figure S3. Protein-Coding Annotation Pipeline and Associated Metrics, Related to Figure 3

(A) Steps performed at each point in annotation are listed along with relevant metrics.

(B) Annotation edit distance (AED) reported by MAKER for all gene models for *Harpegnathos* (left) and *Camponotus* (right). AED represents the agreement between the different sources of evidence (homology, sequence-based, RNA-seq based). A lower AED corresponds to a gene model with more agreement between evidence types.

(C) Number of genes with AED<1 and/or PFAM domain in final *Harpegnathos* (left) and *Camponotus* (right) annotations.

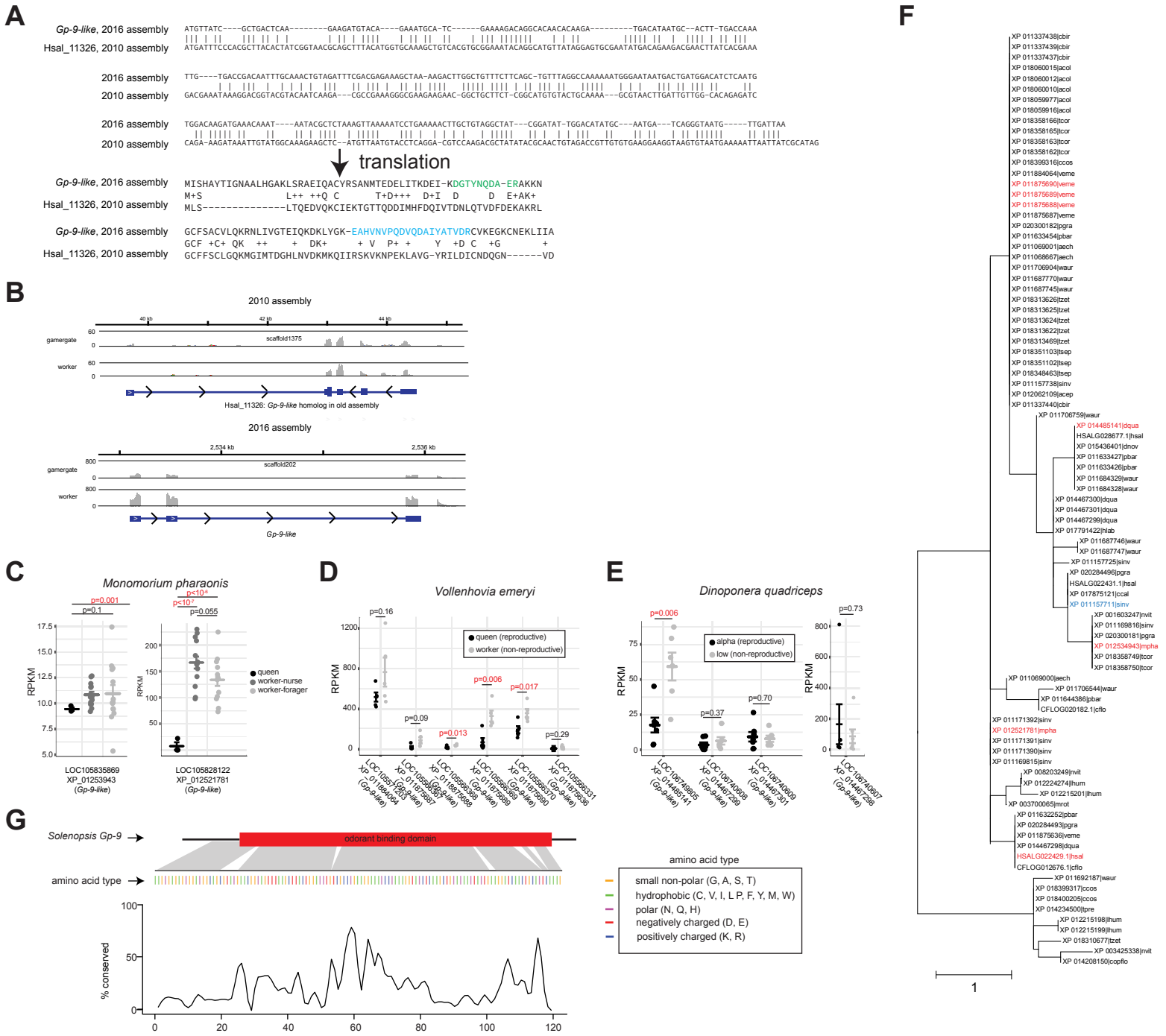


Figure S4. *Gp-9* Expression in Other Ants and Comparison to Old Gene Model, Related to Figure 3

(A) Comparison of *Gp-9-like* gene models in the 2016 and its closest homolog (by similarity of the associated protein) in the 2010 annotation by nucleotide (top) and protein (bottom) sequence. Color highlights on the protein alignment for the new model indicate the 2 peptides detected by mass spectrometry.

(B) Genome browser snapshots of RNA-seq coverage of the 2016 *Gp-9-like* gene and its 2010 homolog (left) and quantification. RNA-seq from workers (n=11) or gamergates (n=12) was mapped to the 2010 or 2016 genome using the same settings.

(C) Expression in heads of queens (n=3) and workers, either foragers (n=13) or nurses (n=14), in the Myrmicine ant *Monomorium pharaonis* of all genes annotated as *Gp-9-like*. P-values are from a Student's t-test.

(D) Expression in full bodies of queens (reproductive, n=5) and workers (non-reproductive, n=5) in the Myrmicine ant *Vollenhovia emeryi* of all genes annotated as *Gp-9-like*. P-values are from a Student's t-test.

(E) Expression in brains of alpha (reproductive, n=7) or low (non-reproductive, n=6) in the Ponerine ant *Dinoponera quadriceps* of all genes annotated as *Gp-9-like*. P-values are from a Student's t-test.

(F) Maximum likelihood phylogenetic tree constructed from multiple species alignment of *Gp-9* and *Gp-9-like* protein sequences in insects. Differentially-expressed genes from Figures 3, and S4C–E are in red. *Solenopsis Gp-9* is in blue.

(G) Conservation by position of *Harpegnathos Gp-9-like* gene. % conservation refers to number of *Gp-9* and *Gp-9-like* models from (F) with same residue as *Harpegnathos* HSALG022429.1 in a multi-species alignment, and was smoothed using smooth.spline, spar=0.2.

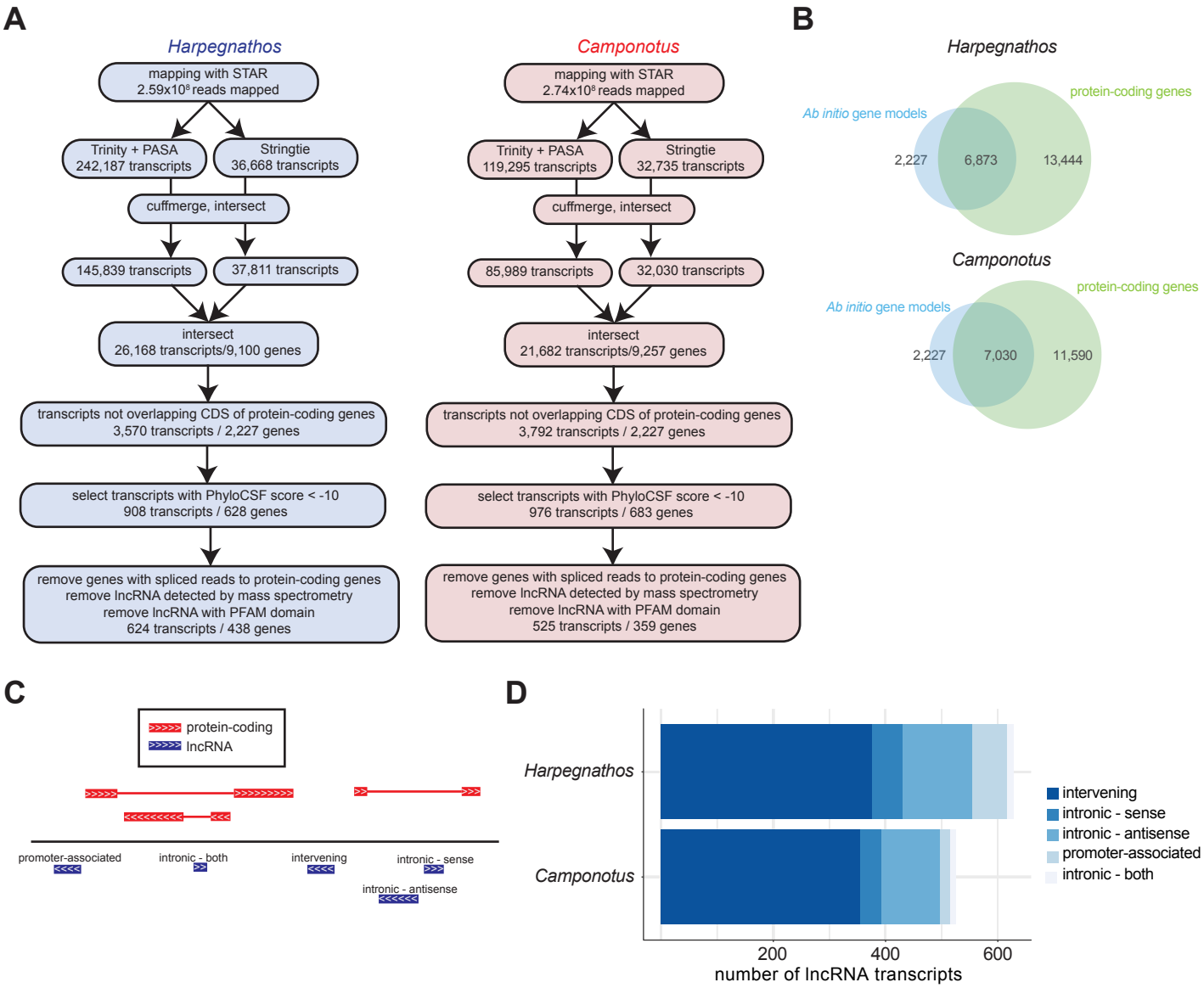


Figure S5. LncRNA Annotation, Related to Figure 5

(A) Steps performed during annotation process are listed along with relevant metrics. 75% reciprocal overlap threshold was required for cuffmerge overlap with PASA or Stringtie, and for overlap between cuffmerge/PASA and cuffmerge/Stringtie. For protein-coding overlap, a transcript was considered intergenic if no base pairs overlapped (strand-specific) between the transcript and a protein-coding gene. The PhyloCSF Omega Test mode was used to detect transcripts with low coding potential (PhyloCSF score < -10).

(B) Venn diagram for the overlap between *ab initio* transcript assembled by Trinity and Stringtie with protein-coding gene models in *Harpegnathos* (top) and *Camponotus* (bottom).

(C) Schematic for the classification of lncRNAs based on their position relative to protein-coding genes. The lncRNA models were divided into promoter-associated (lncRNAs within 1 kb of promoter of gene and transcribed in the opposite direction), intronic - both (lncRNAs contained in introns of two genes in opposite directions), intervening (no overlap with protein coding genes, exluding promoter-associated), intronic - antisense (lncRNAs contained in intron of antisense gene), intronic - sense (lncRNAs contained in intron of sense gene).

(D) Number of lncRNAs in each category listed in (C).

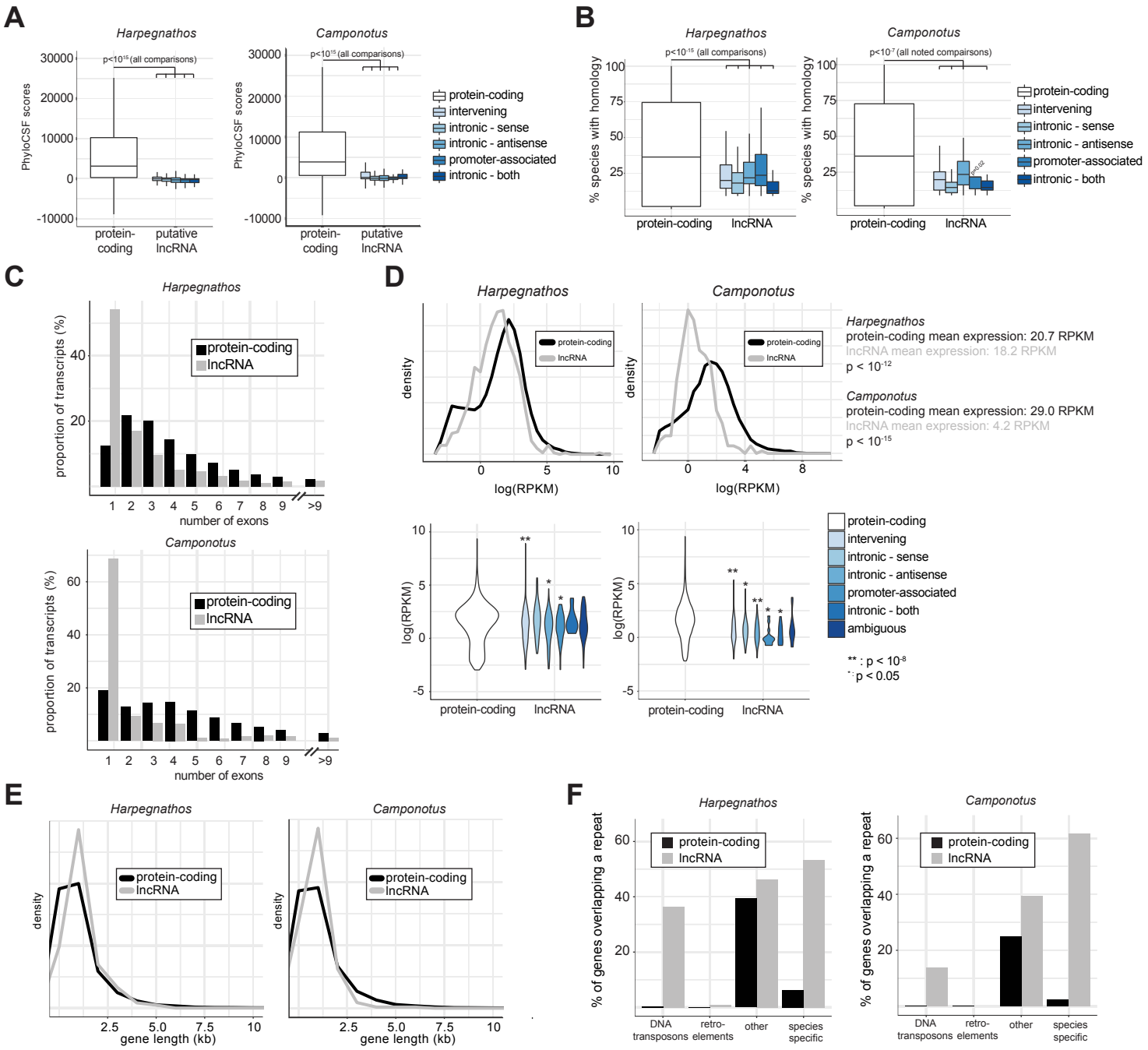


Figure S6. Characteristics of Ant LncRNAs, Related to Figure 5

(A) PhyloCSF scores for putative lncRNAs separated in classes based on their relationship with neighboring protein-coding genes (as in Fig. S5C–D). P-values are from two-sided Student's t-tests.

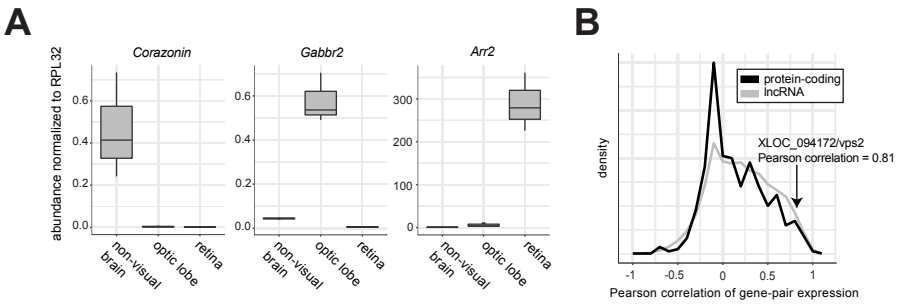
(B) The transcriptomes of 54 insects and 1 outgroup (*Homo sapiens*) were searched for transcripts with significant similarity to protein-coding and lncRNA transcripts. BLASTN hits with an evalue < 10^{-3} were kept as homologs, as in Figure 5C. P-values are from two-sided Student's t-tests.

(C) Number of exons per protein-coding and lncRNA transcript.

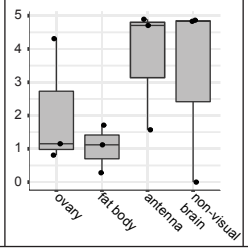
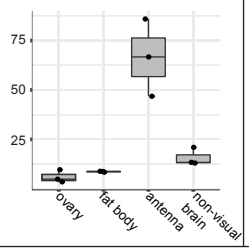
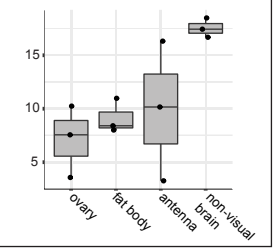
(D) Expression levels of protein coding and lncRNA genes together (top) and split by location (bottom) in *Harpegnathos* and *Camponotus* developmental stages (same samples as in Figure 6A). N=2 for each condition with the exception of *Camponotus* male (n=1). "Ambiguous" indicates that the gene has isoforms that fall into different location categories. P-values are from a two-sample KS test. *, p<0.05, **, p<10^{-10}

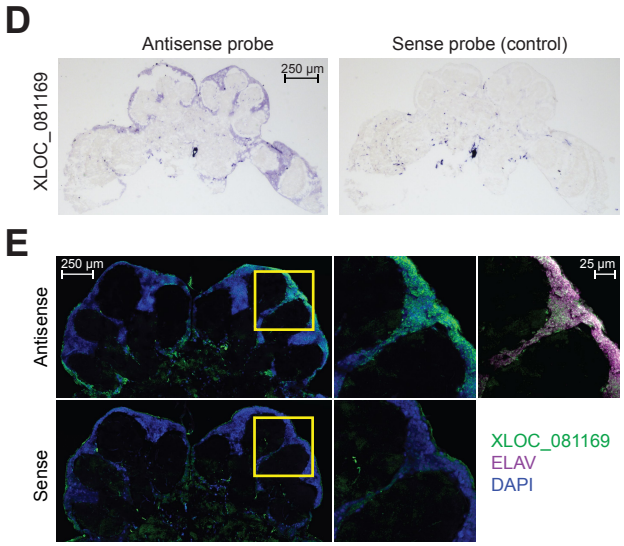
(E) Length distribution of protein-coding and lncRNAs.

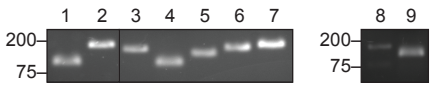
(F) Percent of protein-coding and lncRNA genes in *Harpegnathos* (left) and *Camponotus* (right) that overlap annotated repeats. DNA transposons and retroelements consist of all repeats annotated as "DNA transposons" or "retroelements," respectively, in the *harpegnathos* RepeatMasker library, while "other" consists of all other repeats in the *harpegnathos* RepeatMasker library (small RNA, satellites, simple repeats, low complexity repeats). "Species-specific" consists of repeats from libraries constructed from the 2016 *Harpegnathos* or *Camponotus* assembly.



C

	XLOC_000066	XLOC_093437	XLOC_081169
homolog	*Apis cerana*, AC.lincRNA.17570.1	*Apis cerana*, AC.lincRNA.17570.1	*Drosophila melanogaster* CRG
citation	Jayakodi *et al., BMC Genomics*, 2015	Jayakodi *et al., BMC Genomics*, 2015	Li *et al., NAR*, 2012
localization	brain, hypopharyngeal gland	brain, antenna, hypopharyngeal gland, fat body	high in head, low in thorax, absent in abdomen
notes			regulates CASK, required for locomotor behavior
Harpegnathos RNA-seq (RPKM)			





lane	gene	exp. size of PCR product
1	XLOC_001194	110 bp
2	XLOC_109542	180 bp
3	XLOC_044583	158 bp
4	XLOC_093879	105 bp
5	XLOC_094172	137 bp
6	XLOC_000066	158 bp
7	XLOC_093437	169 bp
8	XLOC_081169	158 bp
9	XLOC_067142	136 bp

Figure S7. Additional LncRNA Validations, Related to Figures 6 and 7

(A) Controls for brain region panel RT-qPCR (n=3 for all brain regions). *Corazonin* is expected to be expressed in non-visual brain, *Gabbr2* in optic lobe, and *Arr2* in retina.

(B) Density plot of Pearson correlations between expression (RPKM) of each protein-coding gene (black) and lncRNA gene (gray) to the nearest protein-coding gene in 12 gamergate and 11 worker brain samples.

(C) Expression levels by RNA-seq in a *Harpegnathos* tissue panel (same data as Figure 6B) for three lncRNAs with homology to other insects.

(D and E) *In situ* hybridization with indicated antisense (*elav*, XLOC_081169) and sense probes on serial frozen sections from *Harpegnathos* worker brains using DIG-coupled probes followed by chromogenic detection (D) or directly conjugated fluorescent probes and counterstaining with DAPI (E). A magnified view of neurons in the mushroom bodies is shown in (E) to demonstrate the colocalization of XLOC_081169 with a pan-neuronal marker, *elav*.

(F) Agarose gel for RT-qPCR products for lncRNAs tested in Figures 6, 7, and S7C.

Table S1. Genome Quality Metrics, Related to Figure 1

	Harpegnathos		*Camponotus*	
	2010 assembly	**2016 assembly**	**2010 assembly**	**2016 assembly**
number of contigs	26,592	1,097	31,883	983
contig N50	39,378	884,632	18,762	1,225,609
number of scaffolds	8,893	857	10,791	657
scaffold N50 (bp)	601,965	1,078,644	451,320	1,585,631
longest scaffold (bp)	2,276,656	3,353,128	2,671,896	10,163,455
number of gaps	17,699	240	21,092	326
number of Ns	11,466,753	933,241	8,173,001	1,771,909
total size (bp)	294,465,601	335,266,283	232,685,334	284,009,204

Table S2. Alignment Metrics for Fosmid Sequences, Related to Figure 2

	Harpegnathos saltator			
	2010 annotation		**2016 annotation**	
fosmid	**coverage**	**length of containing scaffold (bp)**	**coverage**	**length of containing scaffold (bp)**
danthaxa	98.2%	290,101	99.6%	1,117,838
danthcxa	99.0%	1,978,266	99.5%	1,753,804
danthdxa	97.7%	573,047	98.4%	589,170
danthexa	98.3%	1,163,245	98.8%	1,313,330
danthfxa	97.7%	699,624	98.8%	706,479
danthgxa	97.2%	761,569	98.2%	789,757
danthhxa	96.2%	771,335	98.2%	715,156
danthjxa	99.1%	472,718	99.3%	2,893,175
danthkxa	98.3%	984,739	98.8%	1,372,191
danthlxa	97.5%	2,276,656	97.7%	2,621,353
average	97.9%	997,130	98.7%	1,387,225

	Camponotus floridanus			
	2010 annotation		**2016 annotation**	
fosmid	**coverage**	**length of containing scaffold (bp)**	**coverage**	**length of containing scaffold (bp)**
dantcaxa	95.1%	422,032	99.6%	1,595,274
dantcbxa	96.6%	794,750	99.4%	10,163,455
dantccxa	97.7%	544,812	97.5%	2,199,574
dantcdxa	96.3%	588,856	99.3%	7,565,888
dantcexa	99.6%	903,130	99.8%	4,458,663
dantcfxa	97.8%	903,130	99.3%	4,458,663
dantchxa	98.6%	677,527	99.8%	3,484,605
dantcjxa	97.6%	404,019	97.6%	4,397,941
dantckxa	98.5%	468,586	99.2%	4,581,408
average	97.5%	634,093	99.1%	4,767,163

Table S3. Quality Metrics for Protein-Coding Annotation, Related to Figure 3

	Harpegnathos		***Camponotus***	
	2010 assembly	**2016 assembly**	**2010 assembly**	**2016 assembly**
# genes in annotation	18,564	20,317	17,064	18,620
BUSCO results				
complete	98.4%	98.6%	97.2%	98.1%
incomplete or missing	1.6%	1.4%	2.8%	1.9%

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Ant colonies and husbandry

Ants were housed in plaster nests in a clean, temperature- (25°C) and humidity- (50%) controlled ant facility on a 12-hour light/dark cycle. *Harpegnathos* ants were fed three times per week with live crickets. *Camponotus* ants were fed twice weekly with excess supplies of water, 20% sugar water (sucrose cane sugar), and Bhatkar-Whitcomb diet (Bhatkar and Whitcomb, 2016). The *Harpegnathos* colony was descended from the colony sequenced for the original 2010 genome assembly, which was originally collected as a gamergate colony in Karnataka, India in 1999 and bred in various laboratories since (Bonasio et al., 2010; Gospocic et al., 2017). The *Camponotus* colony was collected in Long Key, Florida in November 2011.

Long read DNA library preparation and sequencing

High molecular weight genomic DNA was extracted from 36 *Harpegnathos* and 42 *Camponotus* recently eclosed workers. Gasters were removed before sample homogenization to reduce contamination from commensal bacteria. Size selection and sequencing was performed by the University of Washington PacBio Sequencing service using BluePippin size selection and P6-C4 chemistry, RSII platform. Reads of insert (ROIs) were extracted using SMRT analysis software. The RS_ReadsOfInsert.1 protocol was used, with the parameters 0 minimum full passes and 75% minimum predicted accuracy. 34 SMRT cells were processed for *Harpegnathos*, producing $3.1x10^6$ ROIs containing $2.3x10^{10}$ total bases, for a mean ROI length of 7,471 bp. 17 SMRT cells were processed for *Camponotus*, producing $1.1x10^6$ ROIs containing $1.0x10^{10}$ total bases, for a mean ROI length of 9,934 bp.

Genome assembly strategy

The extracted ROIs were error corrected, trimmed, and assembled by Canu v1.3 (Koren et al., 2017). Error correction and assembly were performed with default parameters with the following changes: corMhapSensitivity = high, corMinCoverage = 0, errorRate = 0.03, minOverlapLength = 499. Quiver was used to polish the assemblies, using the SMRT Analysis protocol RS_Resequencing with default parameters. Scaffolding using both long reads and mate pairs was performed for both *Harpegnathos* and *Camponotus* assemblies, but mate pair scaffolding was done first in *Harpegnathos* and long read scaffolding was done first in *Camponotus*. SSpace-Standard (Boetzer et al., 2011) was used to scaffold the assemblies using mate pair sequencing data with inserts of 2.2 kb (*Harpegnathos*: 5 libraries, *Camponotus*: 1 library), 2.3 kb (*Camponotus*: 1 library), 2.4 kb (*Camponotus*: 1 library), 2.5kb (*Harpegnathos*: 1 library), 5kb (*Harpegnathos:* 4 libraries, *Camponotus*: 2 libraries), 9kb (*Harpegnathos*: 1 library), 10kb (*Harpegnathos*: 1 library, *Camponotus*: 1 library), 20kb (*Harpegnathos*: 1 library, *Camponotus*: 1 library), or 40k (*Harpegnathos*: 1 library, *Camponotus*: 1 library). Standard parameters were used. For scaffolding with long reads, subreads were extracted from PacBio sequencing data using bash5tools with the following parameters: minLength=500, minReadScore=0.8. PBJelly (English et al., 2012) was then used to perform the scaffolding, following the normal protocol. After scaffolding with mate pairs and PacBio subreads, the assemblies were polished using paired-end Illumina short reads and the tool Pilon to produce the final assemblies. One *Harpegnathos* scaffold showed high similarity to a bacterial genome and was removed.

Repeat masking and evaluation of repeats in new sequence content

Although repeat masking was performed by the MAKER2 pipeline internally during the protein-coding gene annotation step, RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org) was also run independently to compare repeats in the 2010 genome assemblies to the 2016 assemblies and to produce a masked genome FASTA. First, the genomes were masked with RepeatMasker and the "*Harpegnathos saltator*" library. Custom repeat libraries were then constructed using RepeatScout on the 2016 genomes with default parameters. These libraries were used in RepeatMasker to find species-specific repeats. Next, we detected non-interspersed repeat sequences with RepeatMasker run with the "-no int" option. Finally, we used Tandem Repeat Finder (Benson, 1999) with the following parameters: match=2, mismatch=7, delta=7, PM=80, PI=10, minscore=50, MaxPeriod=12.

To detect new sequence content, the 2010 genomes were broken into 500 bp non-overlapping windows, then aligned to the 2016 assemblies using Bowtie2 (Langmead and Salzberg, 2012).

Comparison of 2016 *Harpegnathos* and *Camponotus* assemblies to other insects

Other insects used for comparison included all insects with scaffold-level genomes annotated by NCBI as of 5/8/17 (n=81). Scaffold number, contig number, scaffold N50, contig N50, number of gaps, and number of gapped bases were obtained from the genome FASTA available for download on the NCBI website.

BLAST was used to find homologs to *Harpegnathos* and *Camponotus* genes in the 2010 and 2016 annotations. We searched an ant panel consisting of 16 ants (*Wasmannia auropunctata, Pogonomyrmex barbatus, Ooceraea biroi, Atta cephalotes, Atta colombica, Trachymyrmex cornetzi, Cyphomyrmex costatus*, *Acromyrmex echinatior, Vollenhovia emeryi, Linepithema humile, Solenopsis invicta, Monomorium pharaonis, Dinoponera quadriceps, Trachymyrmex septentrionalis, Trachymyrmex zeteki*) and a Hymenoptera panel consisting of 16 non-ant Hymenopterans (*Orussus abietinus, Diachasma alloeum, Ceratina calcarata, Polistes canadensis, Apis cerana, Microplitis demolitor, Polistes dominula, Apis dorsata, Apis florea, Copidosoma floridanum, Bombus impatiens, Trichogramma pretiosum, Megachile rotunda, Bombus terrestris, Nasonia vitripennis*). To qualify for "all insects" in **Figure 3A**, the gene had to have a homolog in at least 90% of ants, Hymenoptera, and in *Drosophila melanogaster*. To qualify for "mammals and insects," the gene had to meet the same requirements for "all insects" and have a homolog in both *Mus musculus* and *Homo sapiens*.

Fosmid analysis

Ten Sanger sequenced fosmids (Bonasio et al., 2010) with an average length of 36,755 bp were analyzed for *Harpegnathos,* and 11 fosmids with a mean length of 37,610 bp were analyzed in *Camponotus*. The scaffold with the most hits for each fosmid in both 2010 and 2016 genome assemblies was found using BLAST. Next, the fosmid and the scaffold with the closest matches were globally aligned. The coverage (how many of the fosmid bases matched with the genome) and the length of the scaffold containing the fosmid were reported.

Annotation of protein-coding genes

Protein-coding genes were annotated on the *Harpegnathos* and *Camponotus* assemblies using iterations of the MAKER2 pipeline (Holt and Yandell, 2011). Inputs to the protein homology evidence section of MAKER2 were FASTA files of proteins in *Apis mellifera, Drosophila melanogaster,* and the previous *Harpegnathos* or *Camponotus* annotation. RNA-seq was provided as EST evidence. RNA-seq was processed using PASA_Lite, a version of PASA (Haas et al., 2003) that does not require MySQL. First, a genome-guided transcriptome reassembly was produced using Trinity (Grabherr et al., 2011). The transcriptome was aligned against the genome using BLAT with the following parameters: -f 3 –B 5 –t 4. The alignments were used as input to PASA_Lite, which produces spliced gene models. The PASA_Lite output was further processed with TransDecoder (Haas B. and Papanicolaou A.), a tool that searches for coding regions within transcripts.

The first iteration of MAKER2 was run with the settings est2genome=1 and protein2genome=1, indicating that both models directly from RNA-seq and homology mapping were output. No SNAP (Korf, 2004) hidden Markov model (HMM) was provided in the first iteration. Augustus (Keller et al., 2011) HMMs were provided; in the first run of maker, the *Camponotus_floridanus* parameters provided with Augustus were used for *Camponotus*, and parameters trained on an earlier version of the *Harpegnathos* genome were used for *Harpegnathos*. After the first MAKER2 run, SNAP and Augustus HMMs were trained using the output of the previous step. High confidence gene models were extracted using BUSCO v2 (Simão et al., 2015), a tool that measures the completeness of a transcriptome set. BUSCO searches for the presence of conserved orthologs in the transcriptome, and also can produce a list of which genes are complete gene models. Only these complete models were used to train Augustus and SNAP.

The second iteration of MAKER2 was run with the same homology and RNA-seq inputs, but with the new HMMs and the GFF from the previous step included as an option in the re-annotation parameters section, and with est2genome=0 and protein2genome=0. After the second MAKER2 iteration, HMMs were trained using the same steps as above, and the process was repeated two more times. On the fourth MAKER2 run, est2genome and protein2genome were turned on, producing gene models directly from RNA-seq and homology. The gene models from the last iteration of MAKER2 were filtered using the reported annotation edit distance (AED), which measures the level of agreement between different sources of evidence (Eilbeck et al., 2009) and the presence of a PFAM domain. PFAM domains

were detected using HMMer v3.1b2 (http://hmmer.org) with the PFAM-A database. Genes were retained if they had either an AED < 1 or a PFAM domain, or both.

Gene identifiers (IDs, e.g. HSALG000001) were assigned to genes based on the presence of homolog in the 2010 annotation. If the 2016 had a perfect match at the nucleotide level in the 2010 assembly, it retained the old ID with the version 1 (e.g. HSALG000001.1). If the 2016 model significantly matched at the protein level, but not at the nucleotide level, it retained the old ID with the version 2 (e.g. HSALG000001.2). If multiple 2010 genes were significant matches, multiple 2016 genes matched to the same 2010 gene, or no homolog was present in the old assembly, a new ID was issued.

The *Harpegnathos* annotation contains 2,912 gene models with 100% identity to old gene models, 7,308 updated gene models, and 10,097 gene models that are reported as "new" by homology searches. The *Camponotus* annotation contains 2,483 gene models with 100% identity to old gene models, 8,335 updated gene models, and 7,802 "new" gene models. Many of these "new" genes have homology to multiple genes in the old annotations. Using an e-value of $1e^{-5}$, 84% of the 2010 *Harpegnathos* gene models and 88% of the 2010 *Camponotus* models have homology to a gene in the new annotation, suggesting that many gene models in the old annotation were incomplete or fragmented.

Assessment of annotation quality

The transcriptome completeness was measured using BUSCO v2, which searches for the presence of well conserved orthologs in a transcriptome. The *arthropoda* set was used as the test lineage.

RNA sequencing and analysis

RNA for developmental stage analysis was extracted from ants at various developmental stages for both *Camponotus* and *Harpegnathos*. Tissue panel RNA samples were collected only from *Harpegnathos*.

For library preparation, polyA+ RNA was isolated from 500 ng total RNA using Dynabeads Oligo$(dT)_{25}$ (Thermo Fisher) beads and constructed into strand-specific libraries using the dUTP method (Parkhomchuk et al., 2009). UTP-marked cDNA was end-repaired (Enzymatics, MA), tailed with deoxyadenine using Klenow exo^- (Enzymatics), and ligated to custom dual-indexed adapters with T4 DNA ligase (Enzymatics). Libraries were size-selected with SPRIselect beads (Beckman Coulter, CA) and quantified by qPCR before and after amplification. The developmental stage libraries, used for annotation, were sequenced as 75 nts single-end reads; all other libraries were sequenced as 38/38 paired-end reads.

RNA-seq reads were aligned to the genome using STAR (Dobin et al., 2013) with default parameters. The mapping rate and mismatch rate per base (Figure 2A–B) were reported by STAR. Read counts were assigned to genes using DEGseq (Wang et al., 2009). Differential expression analysis was performed used DESeq2 (Love et al., 2014). LncRNA selected for developmental stages lncRNA expression clustering were lncRNAs with a p-adjusted < 0.05 in differential expression analysis, indicating an FDR of <5%.

***Hox* cluster analysis**

To detect whether the genome annotation captured the genes in the *Hox* cluster, we searched for *Drosophila melanogaster Hox* genes in the *Apis mellifera* genome, as well as the 2010 and 2016 *Harpegnathos* and *Camponotus* annotations. The gene was denoted as present if there was a significant (e-value < 1e-5) hit using standard megablast parameters.

Differential expression of *Gp-9* homologs

RNA-seq from full bodies of *Vollenhovia emeryi* (PRJDB3517, RNA-seq from 5 queens and 5 workers) (Miyakawa and Mikheyev, 2015) and brains of *Dinoponera quadriceps* (GSE59525, RNA-seq from 7 alpha and 6 low ants) (Patalano et al., 2015) was aligned to the respective genome and mapped to NCBI annotated features. The RPKM table provided on the Linksvayer lab website (https://web.sas.upenn.edu/linksvayer-lab/data/) as supplemental data from PRJDB3164 (Warner et al., 2017) was used to compare RNA-seq data from heads of *Monomorium pharaonis* queens (n=3), foragers (n=13), and nurses (n=14). All genes annotated as "*Gp9*" or a "*Gp9-like*" were evaluated for differences in expression between reproductive (queen or alpha) and non-reproductive (worker, low, forager, or nurse) ants. RPKMs between castes were compared using Student's t-tests.

Phylogenetic tree construction and selection analysis of *Gp-9/Gp-9-like*

To find homologs of *Gp-9* and *Gp-9-like*, we searched for any gene annotated in NCBI databases as "*pheromone-binding protein Gp-9"* or "*pheromone-binding protein Gp9-like*," returning 74 gene models among Hymenoptera (not including *Harpegnathos* and *Camponotus*, for which we used any gene model in our updated annotations with homology to *Gp-9-like* or *Gp-9*). The species that have a homolog in this analysis are *Wasmannia auropunctata, Solenopsis invicta, Vollenhovia emeryi, Trachymyrex cornetzi, Atta colombica, Trachymyrmex zeteki, Pogonomyrmex barbatus, Dinoponera quadriceps, Pseudomyrmex gracilis, Acromyrmex echinatior, Trachymyrmex septentrionalis, Cyphomyrmex costatus, Linepithema humile, Ooceraea biroi, Nasonia vitripennis, Monomorium pharaonis, Megachile rotunda, Dufourea novaeangliae, Trichogramma pretiosum, Atta cephalotes, Ceratina calcarata, Habropoda laboriosa, and Copidosoma floridanum*.

Analysis of the selection pattern of this gene family was performed by contrasting the likelihood of the null model (beta, dN/dS =1) and the alternative model (beta, dN/dS ≥1). We aligned the protein sequences using MEGA7 (Kumar et al., 2016), and then used this to align the codons of the coding sequences from these gene models using PAL2NAL (Suyama et al., 2006). We then used the site test of Codeml from the program PAML (Yang, 2007), similar to a previously used strategy to infer positive selection in ant genomes (Roux et al., 2014). We compared the likelihoods of the null model M8a (beta and ω, ω=1) and the alternative model M8 (beta and ω with ω ≥ 1). We compared the likelihood ratios with a chi-square distribution with 1 degree of freedom (Roux et al., 2014) and as suggested in the PAML user guide.

A phylogenetic tree for the protein sequences was constructed using MEGA7 (Kumar et al., 2016) using the default Maximum Likelihood settings: Jones-Taylor-Thornton substitution model, uniform rates among sites, and Nearest-Neighbor-Interchange as the ML Heuristic Method.

Annotation of lncRNAs

RNA-seq reads from various developmental stages of *Harpegnathos* (embryo, instar 1 larva, instar 4 larva, early pupa, late pupa, adult worker, male) and *Camponotus* (embryo, instar 1 larva, instar 4 larva, late pupa minor, late pupa major, minor, male) were assembled using two reference-based transcriptome assemblers, Trinity (Haas et al., 2003) and Stringtie (Pertea et al., 2015). The transcripts produced from these two methods were merged using cuffmerge (Trapnell et al., 2012), then each reassembled transcriptome was intersected (reciprocal 75% overlap required) with the merged transcripts to produce a file for each method with transcripts from the same set. Transcripts from both methods were then intersected (required 75% reciprocal overlap). Finally, this high-confidence transcriptome was intersected with the coding sequences of protein-coding genes, and only transcripts with no overlap to protein-coding genes were designated as intervening. Transcripts were further split by location for some analyses: "intervening" denotes no overlap with protein-coding genes, "intronic-sense" indicates the transcript is an intron of a gene in the same orientation, "intronic-antisense" indicates the transcript is in an intron of a gene in the opposite orientation, "intronic-both" indicates the gene is intronic to a gene in the sense and antisense direction, and "promoter-associated" indicates that the overlap is within 1,000 bp of a promoter of a protein-coding gene transcribed from the opposite strand. The intervening transcripts were collapsed into loci based on cuffmerge results for some analyses.

BLAST was used to find homologs for intervening transcripts in a panel of 54 insects and an outgroup (human). Only hits with an e-value of 10^{-3} were kept. A multispecies alignment was performed for each transcript using MAFFT. TimeTree (Kumar et al., 2017) was used to create a phylogeny complete with branch lengths of the insect panel and either *Harpegnathos* or *Camponotus*. The phylogeny was rooted using the R package *ape*, with *Homo sapiens* as the outgroup. Using this phylogeny and the multispecies alignment, the PhyloCSF Omega Test mode was run, with all reading frames in the sense direction tested, to assess the coding potential of each transcript. PhyloCSF scores are given in the form of a likelihood ratio, in the units of decibans. A score of x means the coding model is x times more likely than the non-coding model (for example, if x=10, the coding model is 10 times more likely; if x=-10, the non-coding model is 10 times more likely). Transcripts with a score < -10 were considered lncRNAs.

We also removed lncRNAs that are likely to be fragments of protein-coding genes. Using stranded RNA-seq, we removed any lncRNA gene with either more than 5 reads, or >1% of the total reads mapping to the gene, connecting it to a protein-coding gene. We also removed lncRNAs that contained peptides detected using mass spectrometry (see below).

Coding Potential Calculator (CPC) (Kong et al., 2007) was used to confirm the non-coding status of lncRNA chosen as examples in the differential expression analyses. The UniRef90 database was used as a BLAST database.

Mass spectrometry analysis

Sample preparation

Ant brains without optic lobes were dissected in ice cold HBSS with proteinase inhibitors and immediately snap-frozen in liquid nitrogen. Individual brains were homogenized in 100 µL of extraction buffer (8 M urea, 50 mM ammonium bicarbonate pH 8) with proteinase inhibitors. Protein concentration was determined by BCA assay. Five µg of total protein extract were reduced for 1 h at 56ºC by adding 1M DTT to final concentration of 5 mM, followed by 45 min alkylation in 10 mM IAA. Proteins were first digested with Lys-C (1:100 ratio of enzyme:protein) for 4 h at 37ºC; followed by trypsin digestion (1:100 ratio of enzyme:protein) overnight. Proteins samples were prepared for MS by subjecting them to solid phase extraction. The bottom of a 200 mL pipette tip was sealed with a 0.4 mm-diameter-disk of C18 material (Millipore) to make a stage-tip. The stage-tip was activated with 100 mL of acetonitrile, equilibrated with 100 mL of 0.1% acetic acid, and loaded with samples, each followed by a brief centrifugation. After washing with 0.1% acetic acid, peptides were eluted into 100 mL of 50% acetonitrile, 0.1% TFA in water. The elution was lyophilized in a SpeedVac concentrator and resuspended in 20 mL of 0.1% formic acid.

Mass Spectrometry Analysis

LC-MS analysis was carried out using an EASY-nLC nano HPLC (Thermo Scientific) coupled to a Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific), equipped with a nano-electrospray source. Ionization source parameters were set to: positive mode; capillary temperature, 275 C; spray voltage and 2.5 kV. Samples were separated on an in-house analytical column (75 µM inner diameter) packed with ReproSil-Pur 120 C18-AQ resin 3 mm. The gradient length was 195 minutes at 2%-28% (100% ACN, 0.1% formic acid) at a flow rate of 300 nL per minute. Data was acquired using data-dependent acquisition. More specifically, the mass spectrometer was set to perform a full MS scan (350 – 1200 m/z) in the Orbitrap with a resolution of 120,000 FWHM (at 200 m/z), an AGC Target of 5.0e5 and maximum injection time of 50 ms. Peptides were subjected to HCD fragmentation (collision energy = 30%) and detected in the ion trap with an AGC target of 1e4 and maximum injection time of 120 ms.

Data Analysis

Mass spectrometry raw files were searched using MaxQuant version 1.6.0.1. 2016 *Harpegnathos* and *Camponotus* using the protein-coding annotation and by translating putative lncRNA transcript models in all three possible forward frames and considering open reading frames ≥ 10 amino acids. MS/MS were searched using Andromeda (Cox et al., 2011). During the search, variable modifications were specified as methionine oxidation and N-terminal acetylation while fixed modification included carbamidomethyl cysteine. Trypsin, which cleaves after Lysine (R) and Arginine (K) was indicated as the digestive enzyme, with two permitted miscleavages. The main search tolerance was set to 4.5 ppm with the first search tolerance of 20 ppm. One or more razor or unique peptides were needed for protein identification and intensity based absolute quantification (iBAQ) was utilized for label-free quantification (Krey et al., 2014). False discovery rate (FDR < 0.01) was set at the peptide level and all other settings were standard.

***In situ* hybridization**

Probe synthesis

For lncRNA XLOC_081169 probes, 500 bp DNA sequence of lncRNA XLOC_081169 with T7 (sense) and SP6 (anti-sense) promoter were synthesized (IDT). For *Elav* probes, we generated cDNA from total ant brain RNA by reverse transcription using SuperScript III kit (Invitrogen); T7 (sense) and SP6 (anti-sense) promoter sequence were added by PCR. Probe were synthesized following published protocols (Morris et al., 2009) with minor modifications. For fluorescent probes, 35% aminoallyl-UTP (10 mM ATP, CTP, GTP (each), 6.5 mM UTP, 3.5 mM aminoallyl-UTP) was added into the *in vitro* transcription reaction. After ethanol precipitation, we incubated the amino-modified RNA solution (14 µg RNA in 20 µl 0.2 M pH 9 carbonate buffer) with Atto 565/633 NHS ester solution (12 µL 5 mg/mL Atto 565/633 NHS ester in anhydrous DMF) at room temperature for 2 h. We purified probes twice with RNeasy Plus Mini Kit (Qiagen).

RNA in situ *hybridization*

RNA *in situ* hybridization (ISH) were performed according to published protocols (Morris et al., 2009; Søe et al., 2011) with modifications. Formalin-fixed OCT-embedded (4% paraformaldehyde [PFA]; Alfa Aesar, LOT:Z22C046) sections of ant brains were prepared as follows. Sections were serially cut to 8 µm thickness with a Cryostat (Thermo Scientific Microm HM550), mounted on Fisherbrand Superfrost Plus Microscope slides, and stored at 70% ethanol at 4ºC. Upon use, sections were washed two times in PBST (15 min) and once in 5X SSC (15 min). For optimal ISH

performance, tissue sections were incubated in prehybridization buffer (5X SSC, 4M urea, 50 µg/mL heparin, 1% SDS and 0.1% Tween 20, 50 µg/mL yeast tRNA, pH to 4.5 with citric acid) in a hybridization oven at 55ºC at least 1 h. Hybridization mixtures were prepared by adding probe to hybridization buffer (5X SSC, 4M urea, 50 µg/mL heparin, 1% SDS and 0.1% Tween 20, pH to 4.5 with citric acid) to a final concentration of 1 ng/µL and heated to 80ºC(10 min) prior to be applied to the tissue section. Hybridizations were performed at 55ºC overnight and subsequently washed in 0.1X SSC for 30 min at the hybridization temperature. Sections were washed in PBST three times. For fluorescent ISH (FISH), sections were stained with DAPI for 10 min in the dark, then washed in PBST three times and mounted with Fluoroshield histology mounting medium. For DIG-labeled probes, sections were incubated in blocking buffer (20% sheep serum in TBST) at room temperature at least 1 h and subsequently anti-DIG-AP (Roche Applied Science) diluted to 0.375 U/ml in blocking buffer at 4ºC overnight, then washed in PBS three times and then washed in freshly made high pH buffer (100mM NaCl, 100 mM Tris pH 9.5, 50 mM $MgCl_2$, 0.1% Tween20). Sections were stained with staining solution in high pH buffer in the dark by adding 4.5 µL NBT and 3.5 µL X-Phosphate per mL. The reaction was stopped by washing three times in PBST and slides were mounted with Fluoroshield histology mounting medium.

Imaging

For chromogenic ISH, sections were imaged with a DS-Ri1 Digital Microscope Camera from Nikon. For FISH, sections were imaged in a single confocal slice with a Leica SPE laser scanning confocal microscope with a 63x HCX PL APO CS 1.4 NA objective using pixel dimensions of 150 nm x 150 nm. Overlapping tiles, each representing an area of 77 x 77 µm, were assembled into a single image using TileScan in the Leica analysis software.

SUPPLEMENTAL REFERENCES

Benson, G. (1999). Tandem Repeats Finder: a program to analyse DNA sequences. Nucleic Acids Res. *27*, 573–578.

Bhatkar, A., and Whitcomb, W.H. (2016). Artificial Diet for Rearing Various Species of Ants. Florida Entomol. *53*, 229–232.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R., Olsen, J., and Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. *10*, 1794–1805.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Eilbeck, K., Moore, B., Holt, C., and Yandell, M. (2009). Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics *10*, 1–15.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652.

Haas, B.J., Delcher, A.L., Mount S.M., S.M., Wortman, J.R., Smith, R.K., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. *31*, 5654–5666.

Haas B., and Papanicolaou A. Transdecoder. <http://transdecoder.github.io/>.

Keller, O., Kollmar, M., Stanke, M., and Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics *27*, 757–763.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics *5*, 59.

Krey, J.F., Wilmarth, P.A., Shin, J.B., Klimek, J., Sherman, N.E., Jeffery, E.D., Choi, D., David, L.L., and Barr-Gillespie, P.G. (2014). Accurate label-free protein quantitation with high- and low-resolution mass spectrometers. J. Proteome Res. *13*, 1034–1044.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. *33*, 1870–1874.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. *34*, 1812–1819.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 1–21.

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. *37*, e123.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Roux, J., Privman, E., Moretti, S., Daub, J.T., Robinson-Rechavi, M., and Keller, L. (2014). Patterns of positive selection in seven ant genomes. Mol. Biol. Evol. *31*, 1661–1685.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34*, 609–612.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2009). DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics *26*, 136–138.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.