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

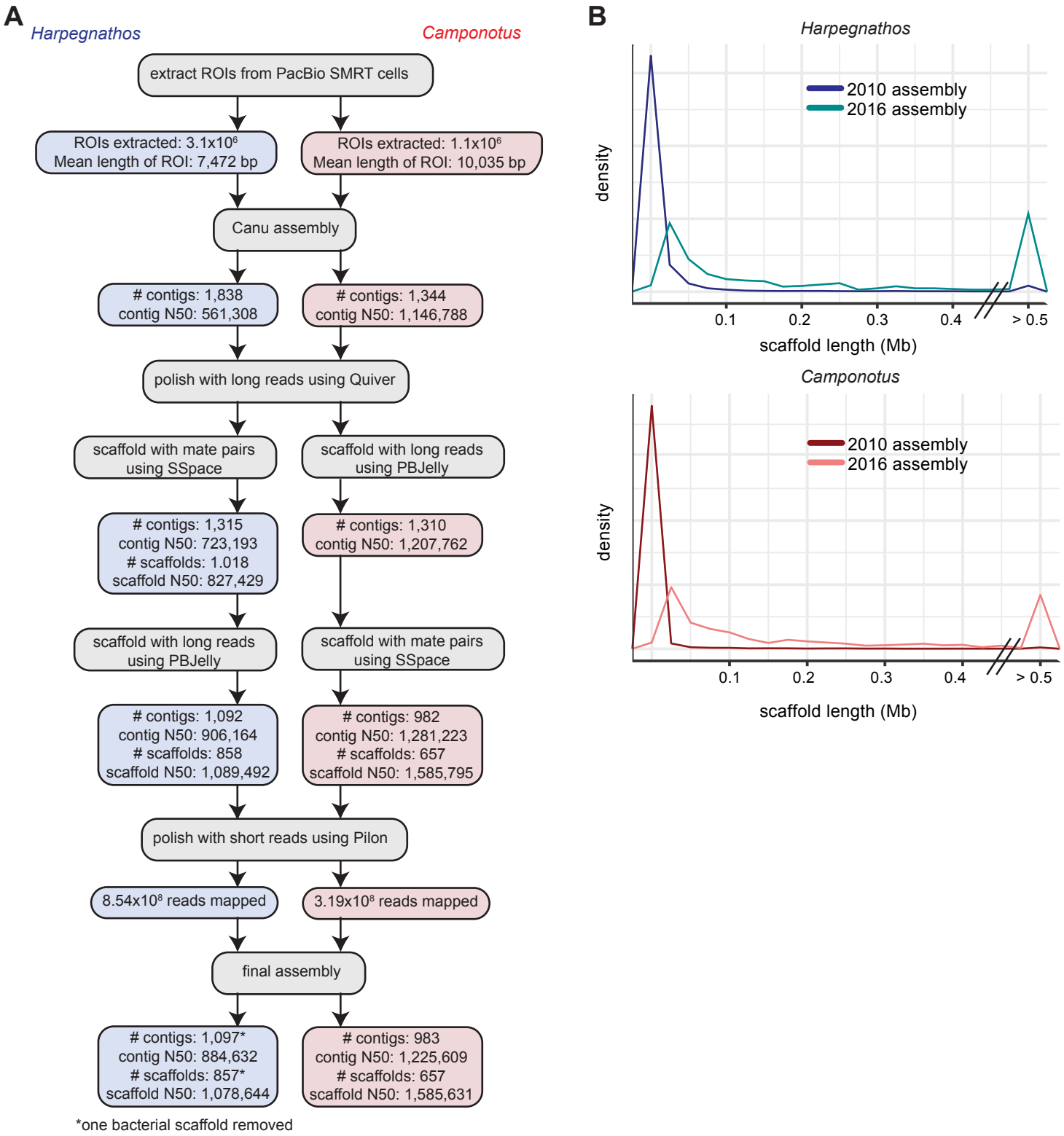


Figure S1. Assembly Pipeline and Associated Metrics, Related to Figure 1

(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.

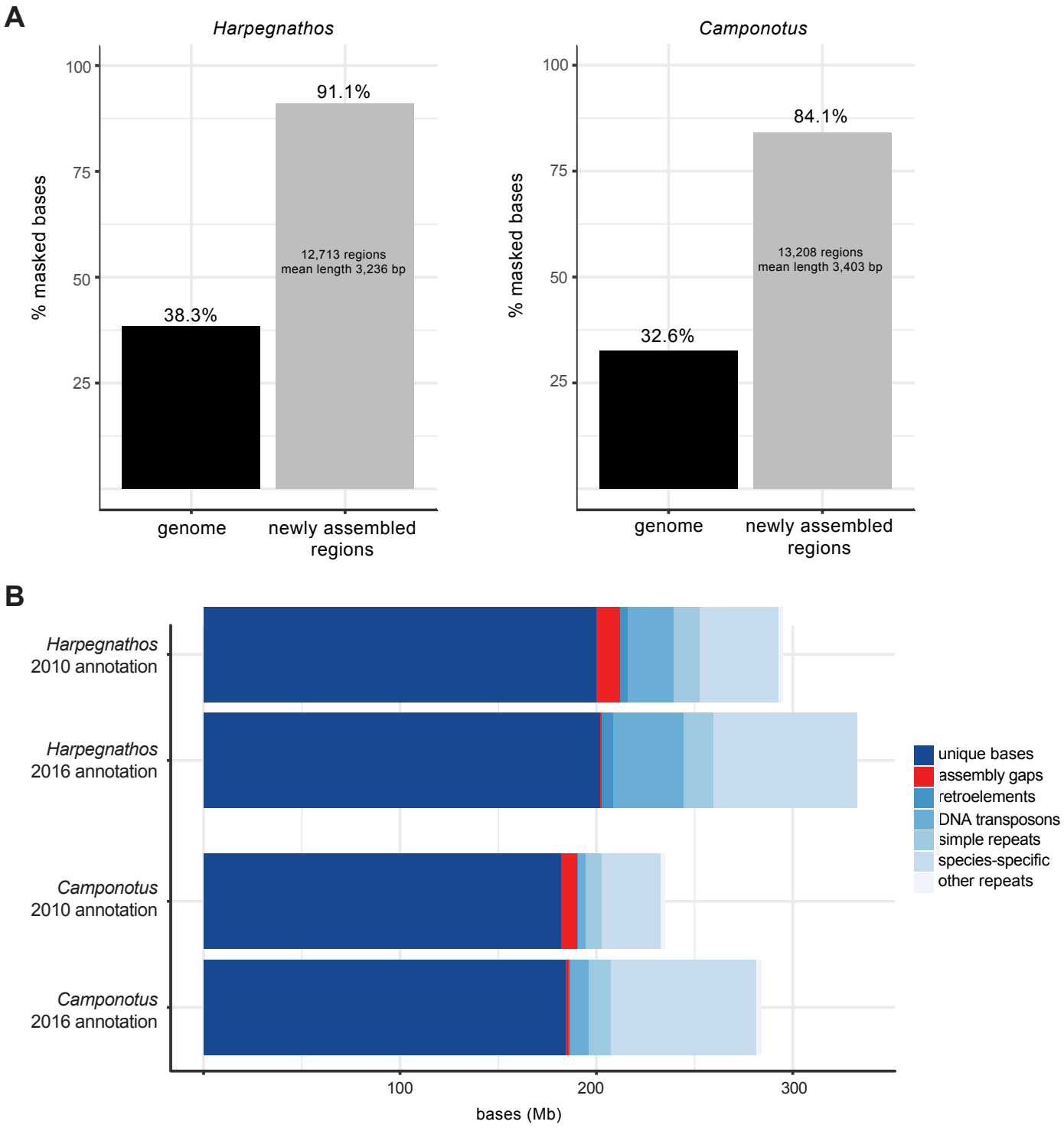


Figure S2. Repeat Content in 2010 and 2016 Assemblies, Related to Figure 1

(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.

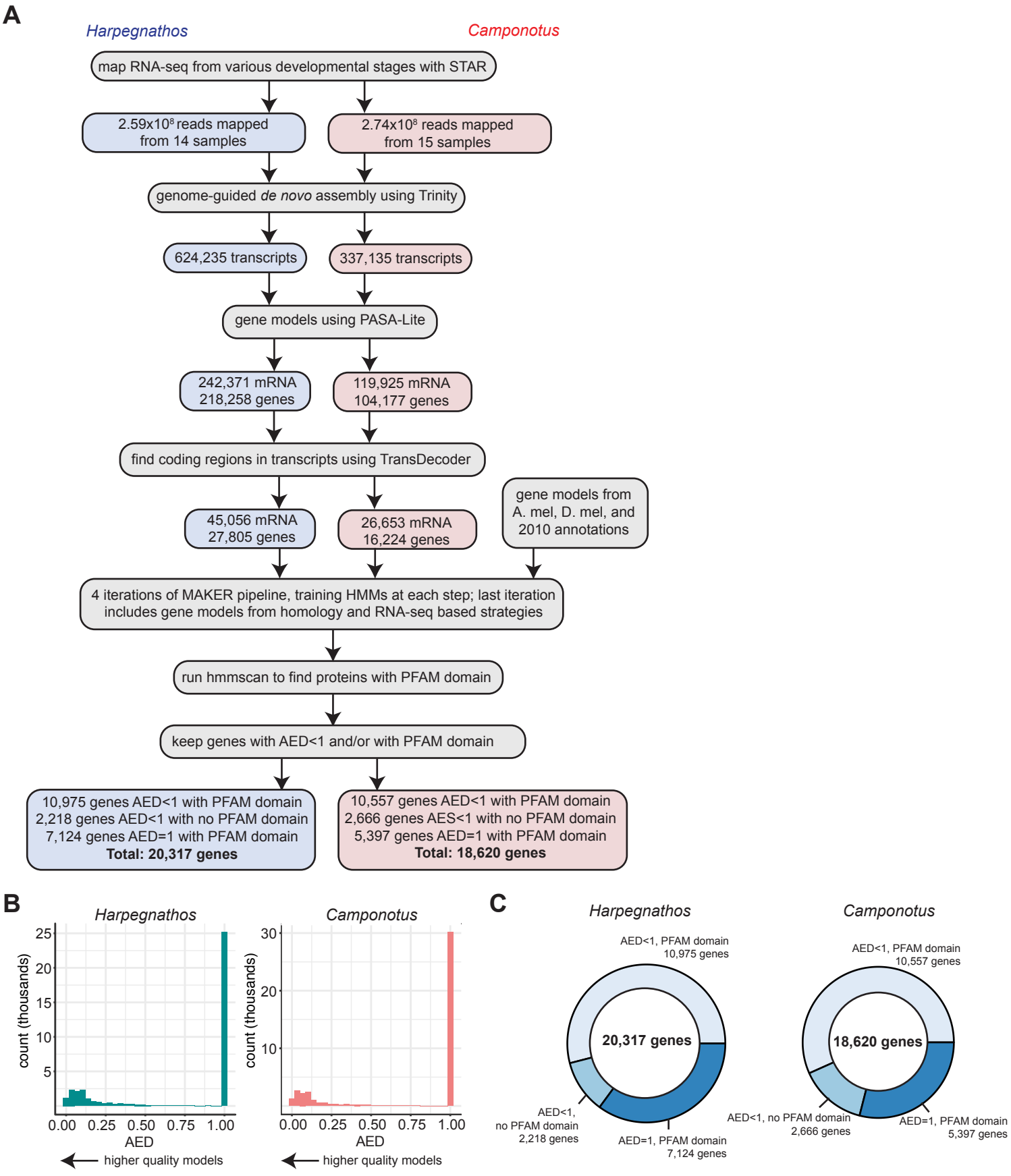


Figure S3. Protein-Coding Annotation Pipeline and Associated Metrics, Related to Figure 3

(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.

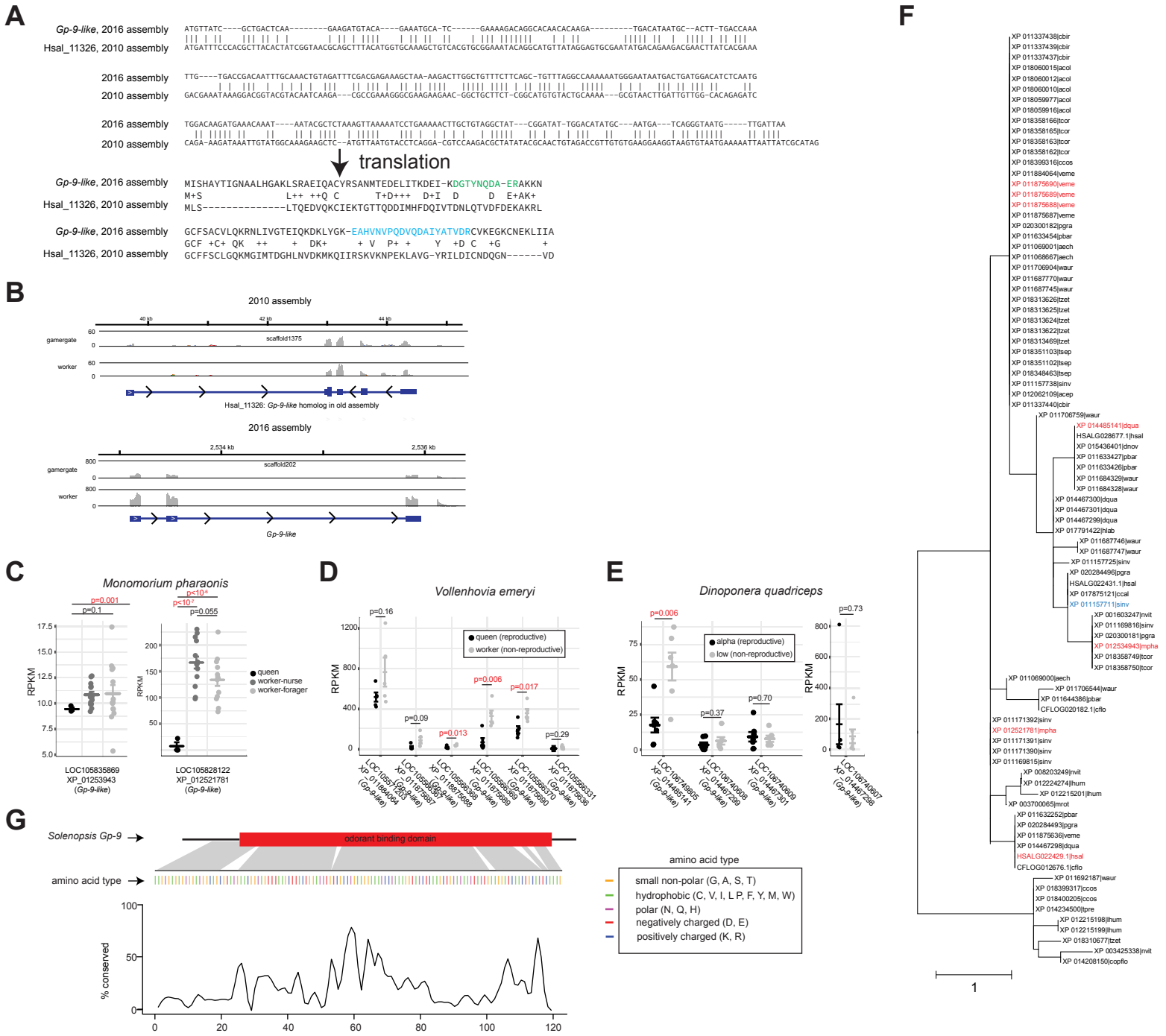


Figure S4. *Gp-9* Expression in Other Ants and Comparison to Old Gene Model, Related to Figure 3

(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.

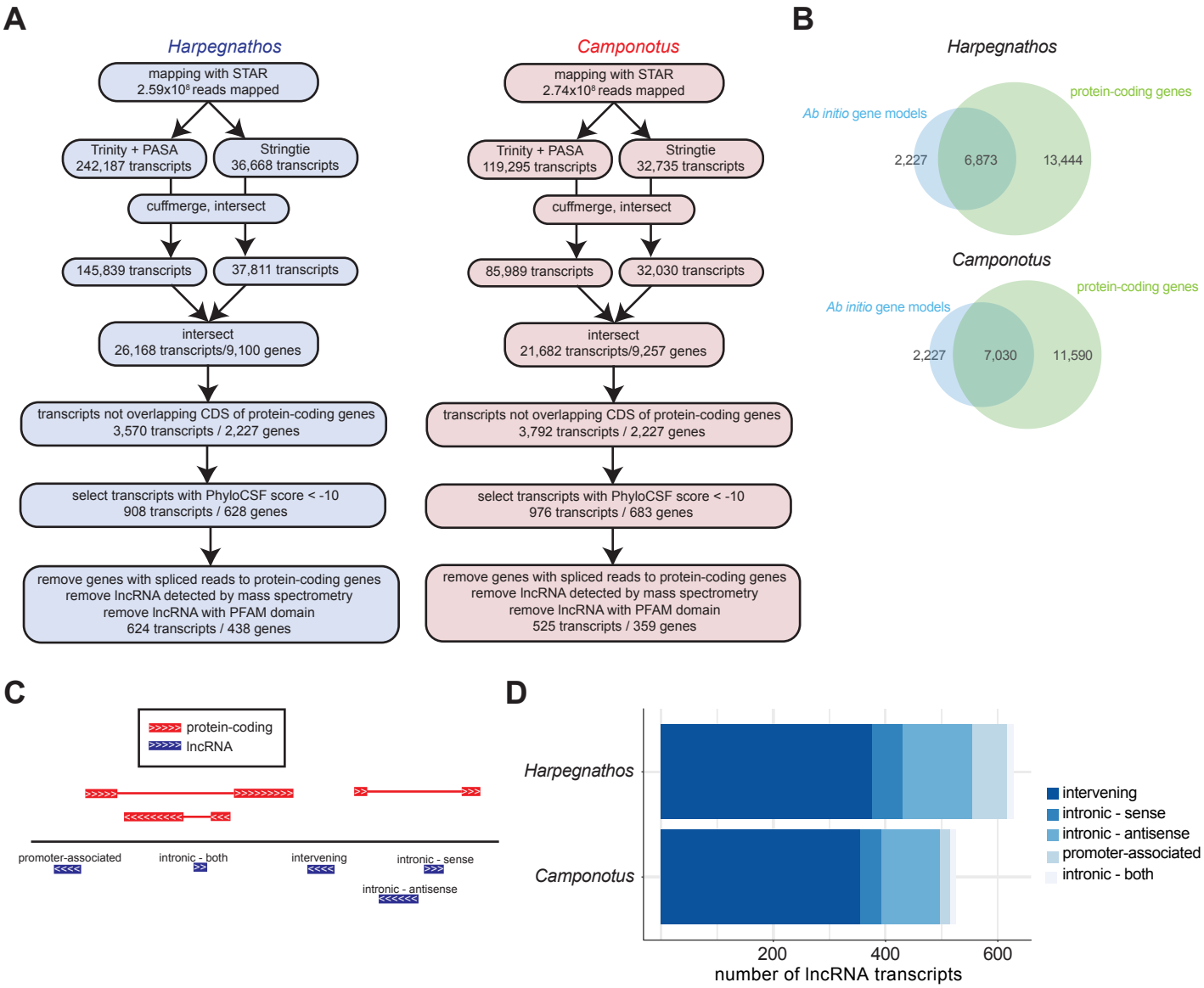


Figure S5. LncRNA Annotation, Related to Figure 5

(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).

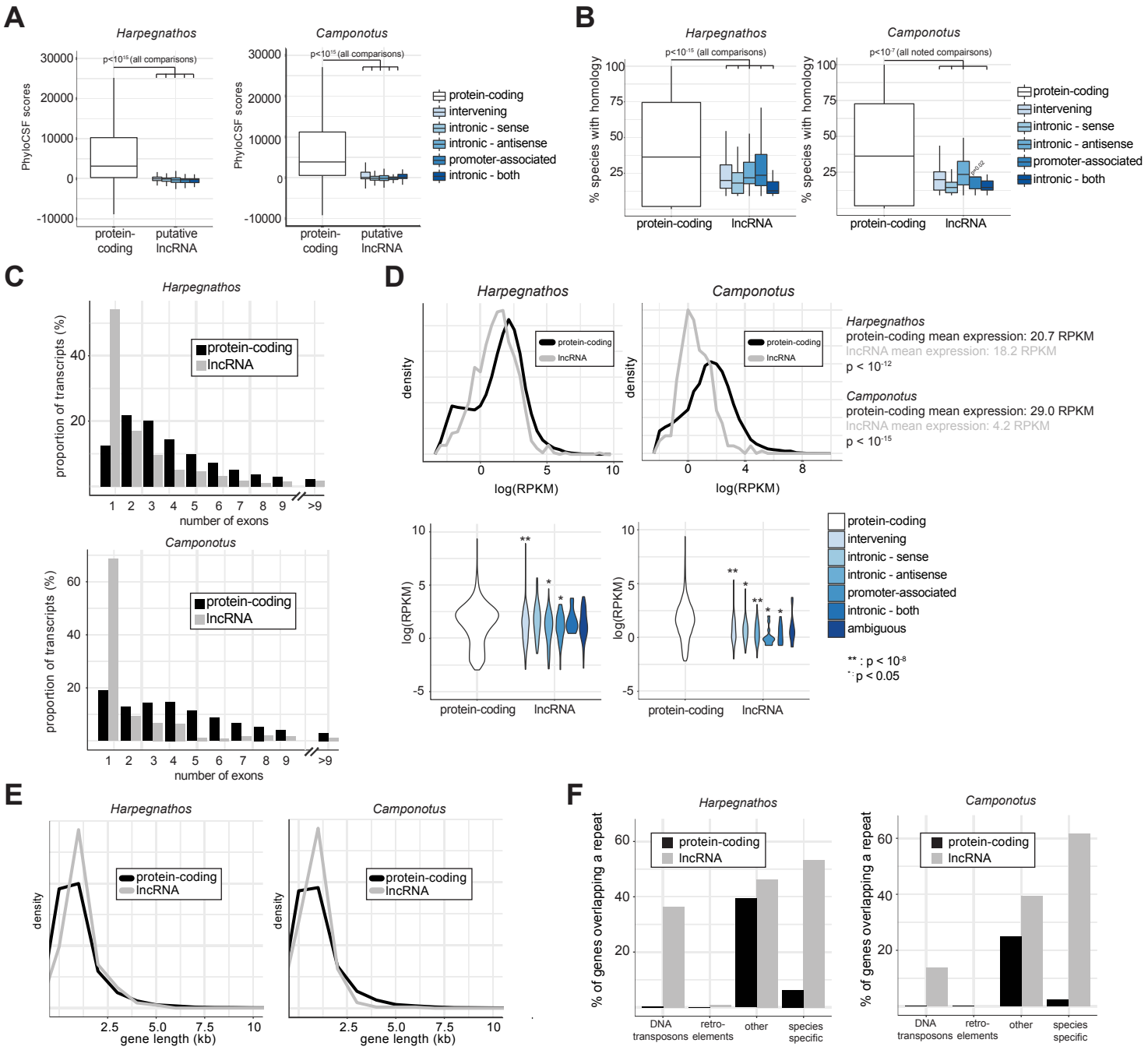


Figure S6. Characteristics of Ant LncRNAs, Related to Figure 5

(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.

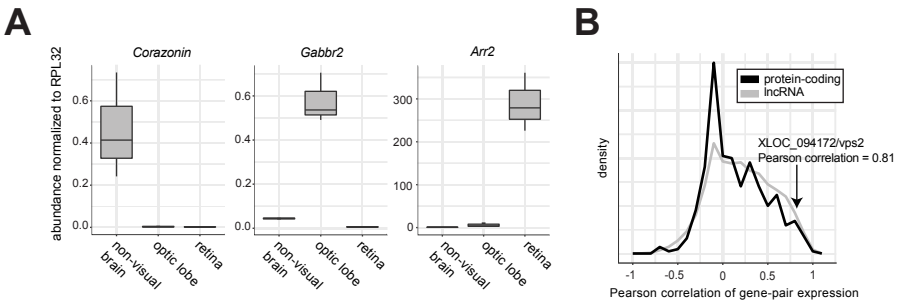
(B) The transcriptomes of 54 insects and 1 outgroup (*Homo sapiens*) were searched for transcripts with significant similarity to protein-coding and lncRNA transcripts. BLASTN hits with an evalue < 10^{-3} were kept as homologs, as in Figure 5C. P-values are from two-sided Student's t-tests.

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

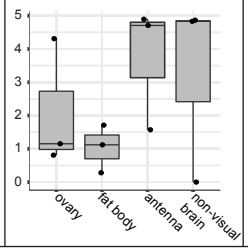
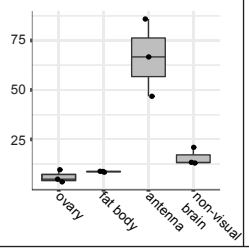
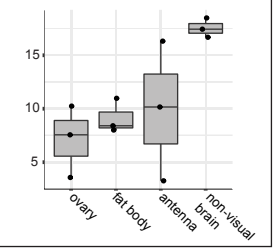
(D) Expression levels of protein coding and lncRNA genes together (top) and split by location (bottom) in *Harpegnathos* and *Camponotus* developmental stages (same samples as in Figure 6A). N=2 for each condition with the exception of *Camponotus* male (n=1). "Ambiguous" indicates that the gene has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

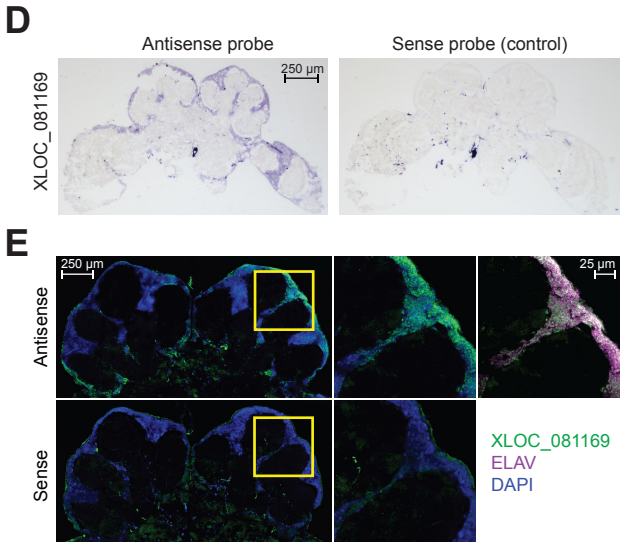
(E) Length distribution of protein-coding and lncRNAs.

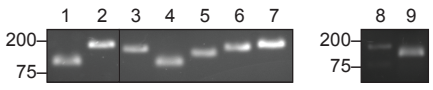
(F) Percent of protein-coding and lncRNA genes in *Harpegnathos* (left) and *Camponotus* (right) that overlap annotated repeats. DNA transposons and retroelements consist of all repeats annotated as "DNA transposons" or "retroelements," respectively, in the *harpegnathos* RepeatMasker library, while "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.