SUPPLEMENTARY INFORMATION

Towards reconstructing the ancestral brain gene-network regulating caste differentiation in ants

Bitao Qiu[1], Rasmus Stenbak Larsen[1], Ni-Chen Chang[2], John Wang[2], Jacobus J. Boomsma[1,*], Guojie Zhang[1,3,4,*]

Contents

1. Technical and statistical limitations of using whole body transcriptomes and transcriptome-based assemblies for cross-species comparisons

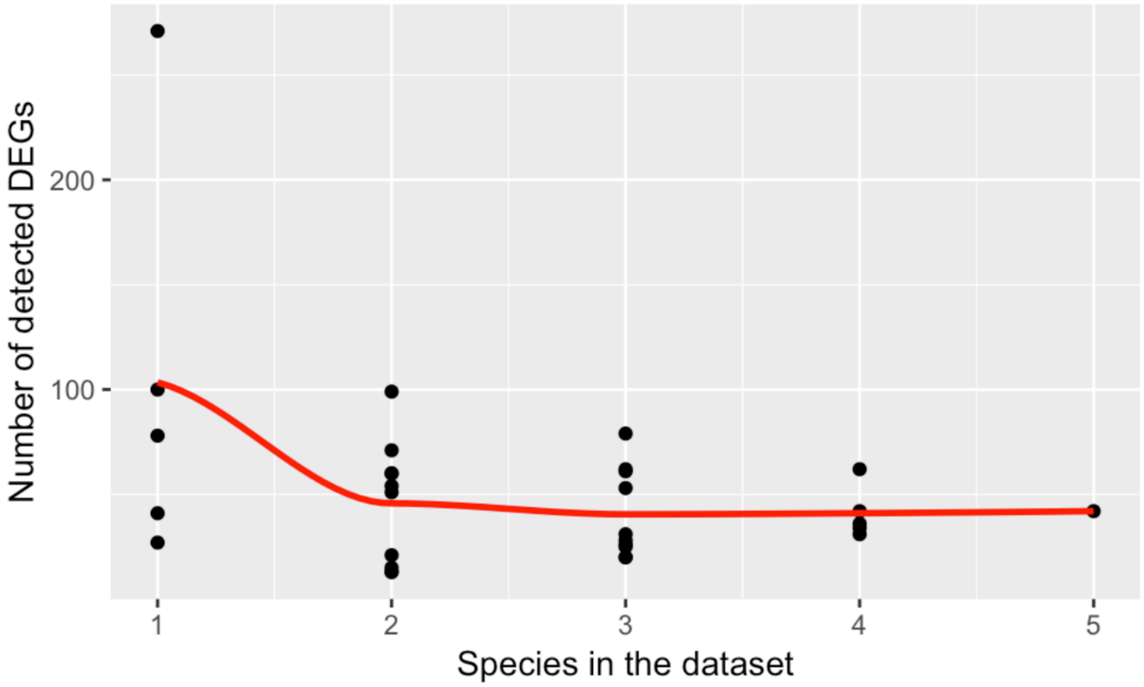
Using whole body transcriptomes will preclude the identification of many differentially expressed genes (DEGs), because the same gene might have different functions in different cell-types/tissues/organs [1]. Such pleiotropic effect of genes are the rule rather than exceptional. When measuring transcriptomes from whole body samples one obtains averages of gene-expression levels across all tissues, so that tissue-specific caste-biased expression patterns will often average out and thus be missed as DEGs.

Our study only included species of ants and other social insects with assembled reference genomes that we re-annotated for the seven ant species. We thus excluded species for which only assembled transcriptomes were available. This decision was based on two further drawbacks of transcriptomes without reference genomes: 1. transcriptomes do not contain synteny information (location information of genes), precluding that one can differentiate between orthologs and paralogs; 2. the statistical power for reconstructing annotated transcripts of the best transcriptome-assembly software so far (Trinity) is ca. 90% [2]. While this is a decent percentage in single species studies, the number of genes being missed will grow exponentially in multiple-species comparisons for every additional species included. This implies that Type II error increases multiplicatively, so detection power for orthologous genes would have dropped to 60% when including 5 species, as we did in the first part of our study focusing on ants with normal caste differentiation. Thus, the combined effect of these two drawbacks is that any transcriptome-based method comparable to our approach would suffer from a very high rate of false negatives in orthologous gene detection.

2. Technical advantages for using Generalized linear model (GLM) methods to detect caste-biased differentially expressed genes across-species (DEGs)

Using a GLM approach to detect cross-species caste-biased DEGs allowed us to avoid the multiple testing problem outline above. Where other studies (e.g. [3]) first examined DEGs for each species independently and then identified cross-species DEGs based on the overlap of DEGs, our approach allowed us to identify cross-species DEGs with a single generalized linear model (GLM) that included the additive effect of species identify, colony of origin, and phenotypes (caste) on gene expression. While the first approach (finding DEGs for each species separately) may be more stringent at the single species level, the rampant (exponential) increase in Type II error (rate of false negatives) as outlined above quickly compromises cross-species analysis of DEGs. Our single GLM approach thus has much higher degrees of freedom and overall detection power [4].

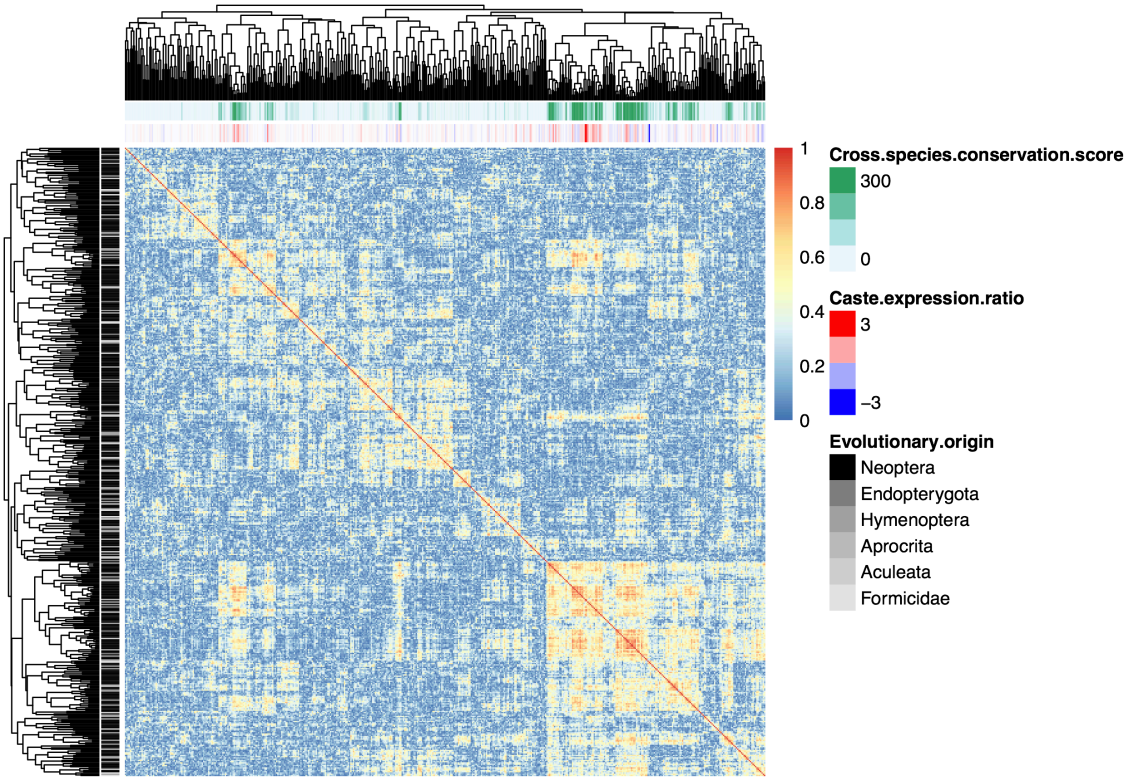
To illustrate the difference between these two analytical approaches, we plotted the number of detected DEGs against the number of species analysed in our study (**Supplementary Figure I**). This shows that: (1) the number of detected DEGs with our method (GLM) decreased when increasing the number of species from one to two, but (2) that number remained stable with a decreasing variance when the number of species increased to and beyond N = 3. The fact that there is no further decrease and actually an increase in reliability by variance reduction, represents a significant improvement relative to state of the art (e.g. **Figure 4** in Morandin et al. 2016 [3]), where exponential decrease of DEG detection efficiency continues with every species added.



Supplementary Figure I: The number of detected DEGs plotted against the number of species analyzed. The data used in this study, fitted with a local polynomial regression (red line), showing that the efficiency of detecting DEGs drops when going from one to two species, due to multiple species having different sets of lineage-specific genes, but remained stable for N ≥ 3 and with sampling variance becoming less with every species added.

3. Cross-species co-expression network-analysis for caste-specific expression levels

To reveal the identity of the ancestral GRN, we performed co-expression network analysis for the five ant species with normal caste differentiation. Using caste-specific transcriptomes normalized for colony-level variation (see *Methods*), we constructed a co-expression matrix for the 6672-orthologous across all samples for gynes and workers. We then evaluated the extent of conserved directions of regulation (i.e. evolutionary connectedness), based on the absolute values of Spearman correlation coefficients for each pair of genes, assuming that gene-pairs with high positive or negative correlation coefficients for caste-specific expression are likely to have maintained their co-regulation, whereas genes with uncorrelated expression have not. We identified clusters of co-expressed genes across the five ant species, and these gene clusters exhibited significant higher levels of connectedness (reflecting conservation of co-regulation) than the overall transcriptomic background (shades of green in the top horizontal bar; FDR adjusted $P < 1e^{-10}$; one-sided paired Wilcox test) (**Supplementary Figure II**). Furthermore, genes that were identified as ancestral caste DEGs (blue or red in the second horizontal bar) were nested within these green clusters. These results suggest that the ant GRN genes are structured around a core-set of genes that have maintained their conserved co-regulation functions and thus represent the ancestral GRN that shaped the origin of caste-based superorganismality in ants.



Supplementary Figure II. Evolutionary connectedness and level of conservation for 6672 1:1 orthologous genes across the five ant species with normal differentiation between gyne and worker castes. The cross-species co-expression network was constructed from correlation matrices based on the species-specific caste-specific expression levels normalized for colony-variation for all gyne and worker samples, so that each cell represents the absolute value of a pair-wise Spearman's correlation coefficient between each gene-pair. Heat-map colours range from blue (absolute Spearman correlation coefficient = 0, no correlation) to red (absolute Spearman correlation coefficient = 1, perfect positive or negative correlation), and blocks of genes that clustered together represent genes that are highly co-regulated with each other across the five ant species. The gene-specific degrees of cross-species conservation scores in the network are given as green shades in the horizontal bars above the plot and are based on the $\log_{10}$ transformed adjust *P* values obtained by comparing the connectedness distribution of each target gene with that of all other genes (one-sided paired Wilcoxon test), ranging from 0 (no difference to background distribution) to 300 (connectedness distribution significantly higher than background). White indicates a gene that is least conserved for its expression regulation with other genes, whereas dark green indicates highly conserved genes that are part of the ancestral GRN. The second horizontal bar summarize the extent of cross-species caste-bias for each gene based on average $\log_2$ gyne/worker expression ratios across the five ant species, ranging from extreme worker-biased expression (blue) to extreme gyne-biased expression (red). The black/white bar towards the left of the plot indicates the evolutionary age of each gene, ranging from (pre-)Neopteran ancestry (black) to ant-specific genes (light grey). To keep plotting manageable, only 1000 of the 6672 orthologous genes were randomly sampled and plotted, but the overall pattern and proportional participation in the GRN should statistically remain the same.

4. Effects of *P* value thresholds for DEGs detection in the honeybee

Because our corbiculate bee expression data only contained a single species (the honeybee), that data set must have included both conserved (superorganismal corbiculate) caste-biased genes and honeybee-specific caste-biased genes. The latter category would have been filtered out if additional corbiculate bee species would have been available for simultaneous analysis. In contrast, our ant data contained five species, which allowed us to identify conserved caste-biased genes in ants as an insect family. To avoid biased inferences due to this asymmetry, we used a more stringent criterion for DEG detection in the honeybee (1.5 fold-change expression difference between castes as the null model; adjusted *P*-values < 1e^{-3}), whereas in ants we used 1.5 fold-change and adjusted *P*-values < 1e^{-2} as thresholds for accepting cross-species DEGs.

*Adjusted P-value threshold **	*# (%) of DEGs in the honeybee; # of DEGs with orthologs in ants*	*# (%) of DEGs across five-ant species; # of DEGs with orthologs in the honeybee*	*# of DEGs shared between ants and the honeybee (regardless of caste-biased direction)*	*# of DEGs shared between ants and the honeybee (same direction of caste-biased expression)*
P < 1e-2	1602 (14.9%) 818	42 (0.6%) 39	16	6
P < 1e-3	1405 (13.0%); 733	33 (0.5%) 30	11	4

* Using 1.5 fold-change expression difference between castes as null-model threshold

Using either the 1e^{-3} or the 1e^{-2} *P*-value threshold leads to the same conclusion that there is significantly higher overlap between ant and honeybee DEGs (16/39 (41%) and 11/30 (37%)) compared to overall background (14.9% and 13.0%) (*P* = 8e^{-5} and 3e^{-3}), whereas overlap of DEGs with the same direction of caste-biased expression was not significant (6/39 (15%) and 4/30 (13%); *P* = 0.82 and 1). This is consistent with the results of our PCA showing that genes contributing to caste phenotypes in ants and the honey overlap, but that the directions of caste-biased expression are uncorrelated and thus independent (**Supplementary Figure 9**).

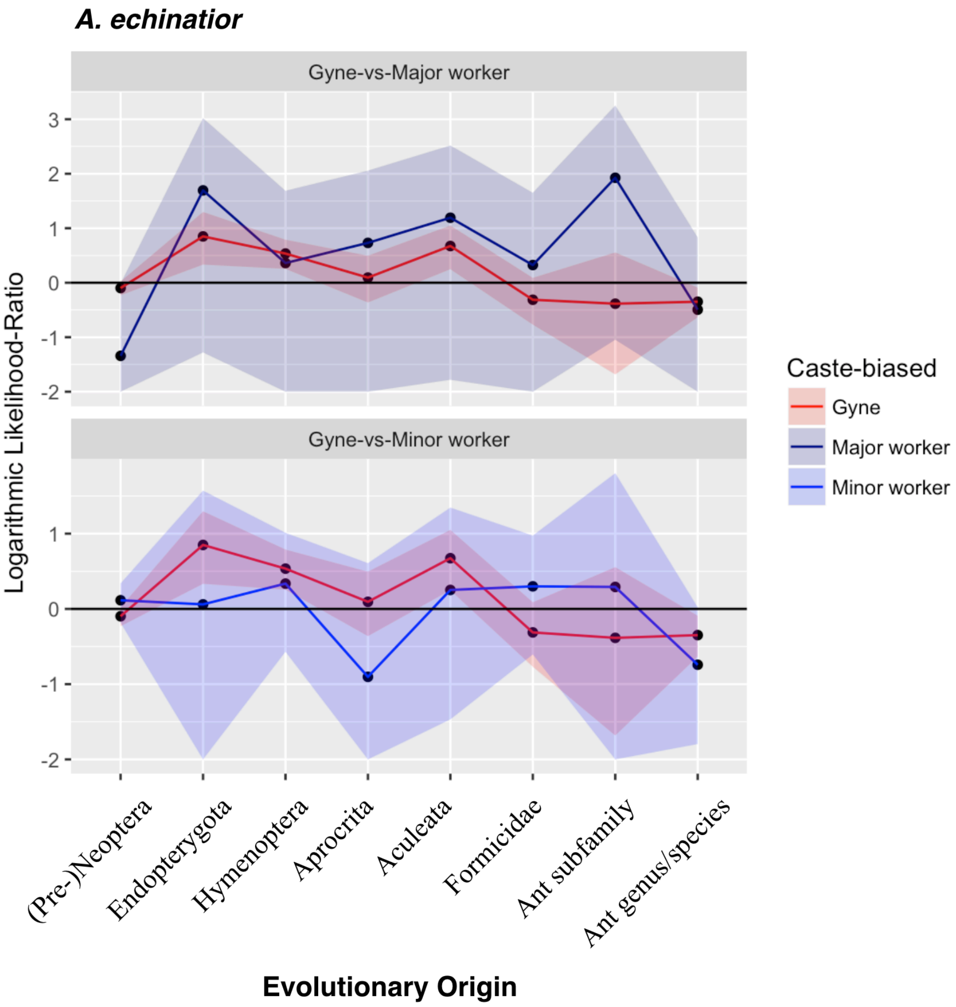
5. The role of ancestral and lineage-specific genes in caste regulation

Our study shows that both ancestral and lineage-specific genes are involved in the regulation of queen and worker caste phenotypes of ants, and that the likelihood of having been recruited for caste-biased expression varies with gene age and type of caste. In the five ant species with standard queen-worker caste differentiation that we analysed, ancestral genes were more likely to be over-expressed in the brains of gynes (winged virgin prospective queens), consistent with earlier whole body transcriptome studies in the honeybee and the ants *Temnothorax longispinosus* and *M. pharaonis* [5-7]. However, while other studies found that novel genes are more likely to be recruited for worker-biased expression, we found that this is not a universal pattern across ant species. In both *M. pharaonis* and *L. humile*, novel lineage-specific genes were more likely to be expressed in a worker-biased manner, whereas in *A. echinatior*, *S. invicta* and *L. niger* the ancestral genes had higher probabilities to have been recruited for worker-biased expression. An interesting coincidence is that *M. pharaonis* and *L. humile* are highly polygynous ants where each colony has many cooperatively breeding queens. In comparison *A. echinatior* and *L. niger* have single queen (monogynous) colonies by default and are at best facultatively polygynous [8-10]; *S. invicta* can have both monogynous and polygynous colonies, but samples for this study came exclusively from monogynous colonies. Polygynous ants usually have much larger effective population sizes than monogynous ants [11,12], and the so-called unicolonial invasive ants to which both the *M. pharaonis* and *L. humile* belong [12] also have much larger colonies. As obligate polygyny is an evolutionary derived state in ants [13], a more pronounced role for novel genes in the regulation of gene expression in the worker caste might be what one would expect, but more ant species with contrasting types of social organization will have to be analysed before we can draw any firm conclusions.

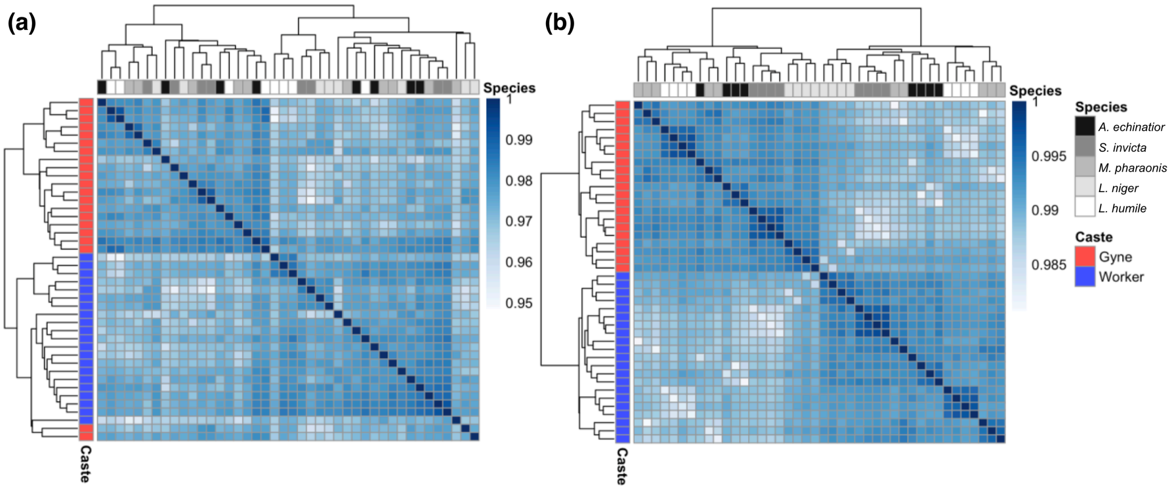
In a functional sense, the evolution of permanent adult castes must imply that workers are the evolutionary derived caste, because queens continue to be fertile and there is no ancestral sterility. However, as soon as the point of no return towards superorganisnality was passed, both caste-specific developmental pathways will have recruited or dropped genes if that made them more efficient queens or workers in coevolution with each other. Based on the number of lineage specific genes, a previous study has corroborated that ant workers are the more derived state [7]. However, our study also indicates that lineage-specific genes have shaped caste regulation in both gynes and workers (**Figure 3; Supplementary Table 3**), consistent with evolutionary derived elaborations in both castes. That gains and losses are both likely to be important is illustrated by the magnitude of gene expression difference between castes being species/lineage specific, with the number of DEGs ranging from a few hundred in *A. echinatior* to several thousands in *L. humile* (**Supplementary Table 7**). Future studies including more ant species with a wider spread of phylogenetic histories will likely be able to explain the underlying pattern of these differences and to explore possible differences with other life history traits.

6. Implications and possible limitations of using transcriptome data from different tissues or developmental/behavioural stages

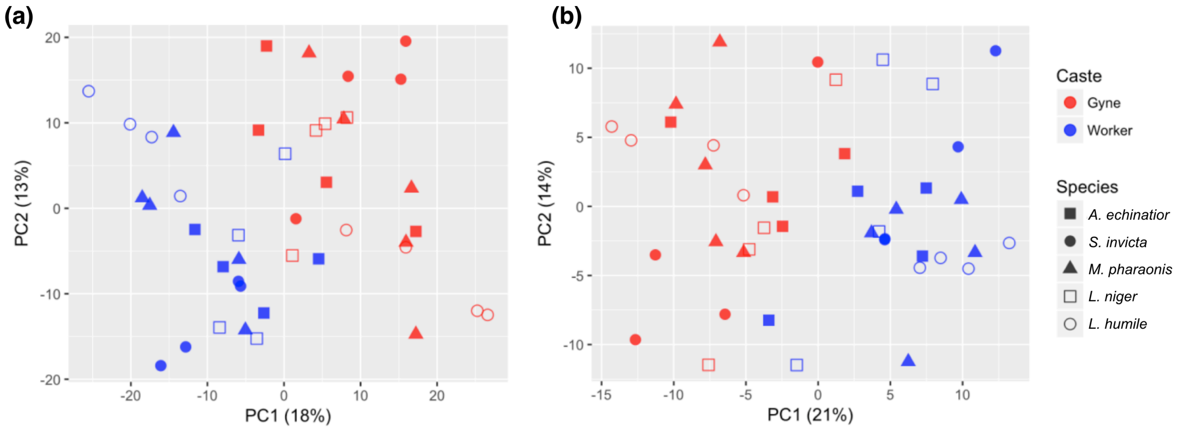
Overall caste-specific expression patterns in our study revealed conserved GRNs across the five ant species with normal caste differentiation. However, our study used brain samples from adult gynes and randomly picked workers, which might have reduced the GRN signals. Future studies will thus have to focus more explicitly on comparing caste-specific transcriptomes from other tissues, such as the reproductive organs (i.e. ovaries) and organs where caste differences are not expected (e.g. legs) so they could serve as controls. We expect that queen-worker expression difference in reproductive tissues (ovaries) will be more dramatic in superorganismal social insects than in cooperatively-breeding social insects without permanent castes, and that ovary GRNs should be evolutionarily conserved. Previous studies in *M. pharaonis* and *L. niger* have shown that age has a strong influence on brain gene expression pattern [14,15], so we might also expect that GRN signatures will be different across different developmental stages and different ages of adults in each caste.



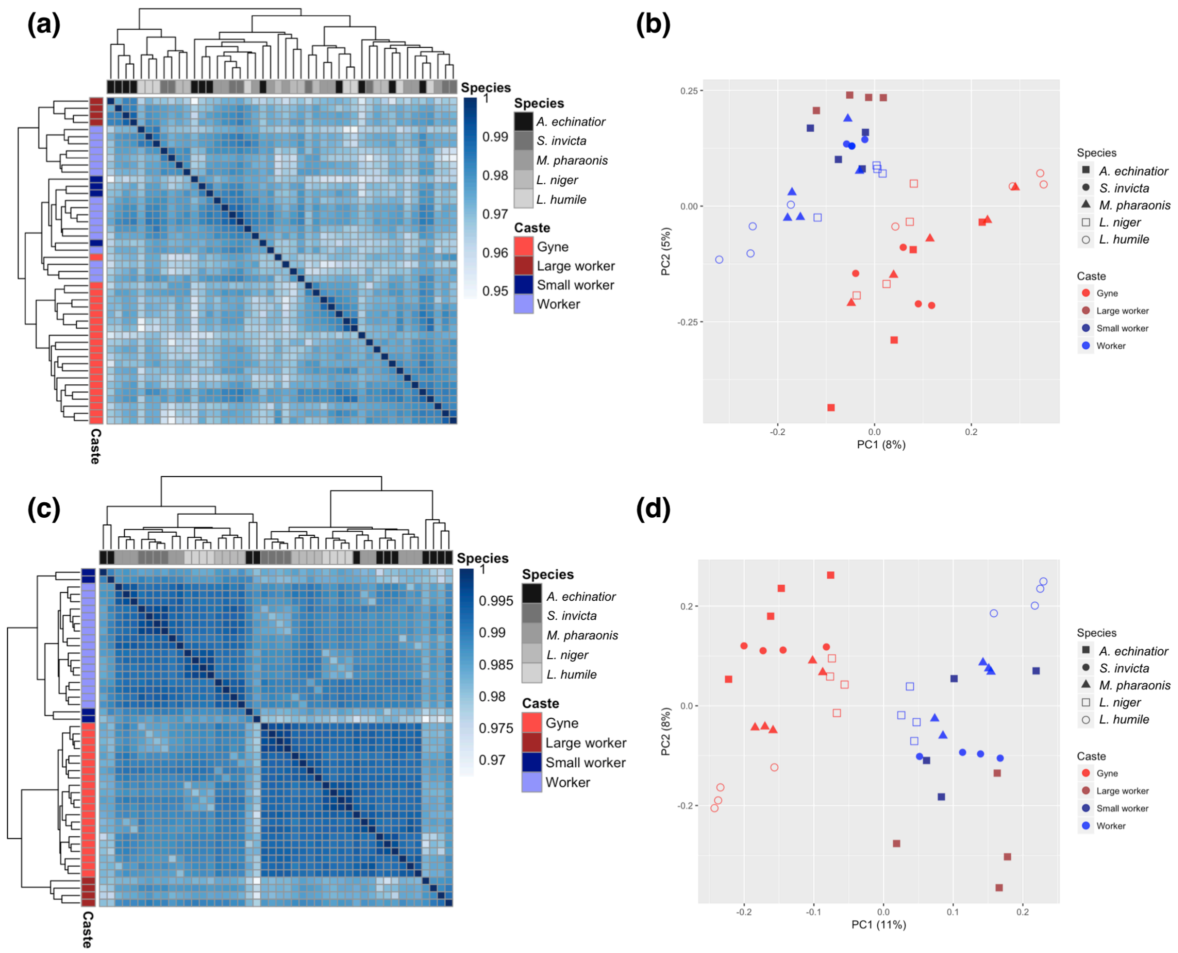
Supplementary Figure 1. Likelihood-ratios of genes with caste-biased expression in the brains of *Acromyrmex echinatior*, comparing gynes-vs-major workers and gynes-vs-minor workers. For each plot, log likelihood-ratios were calculated with the same method as in **Figure 1**, except that caste DEGs were calculated based on the comparison between gyne and major worker samples or between gyne and minor worker samples. Colors and legend are the same as **Figure 1**, except for the dark blue dots and connecting lines in the A-panel representing major workers of *A. echinatior.*



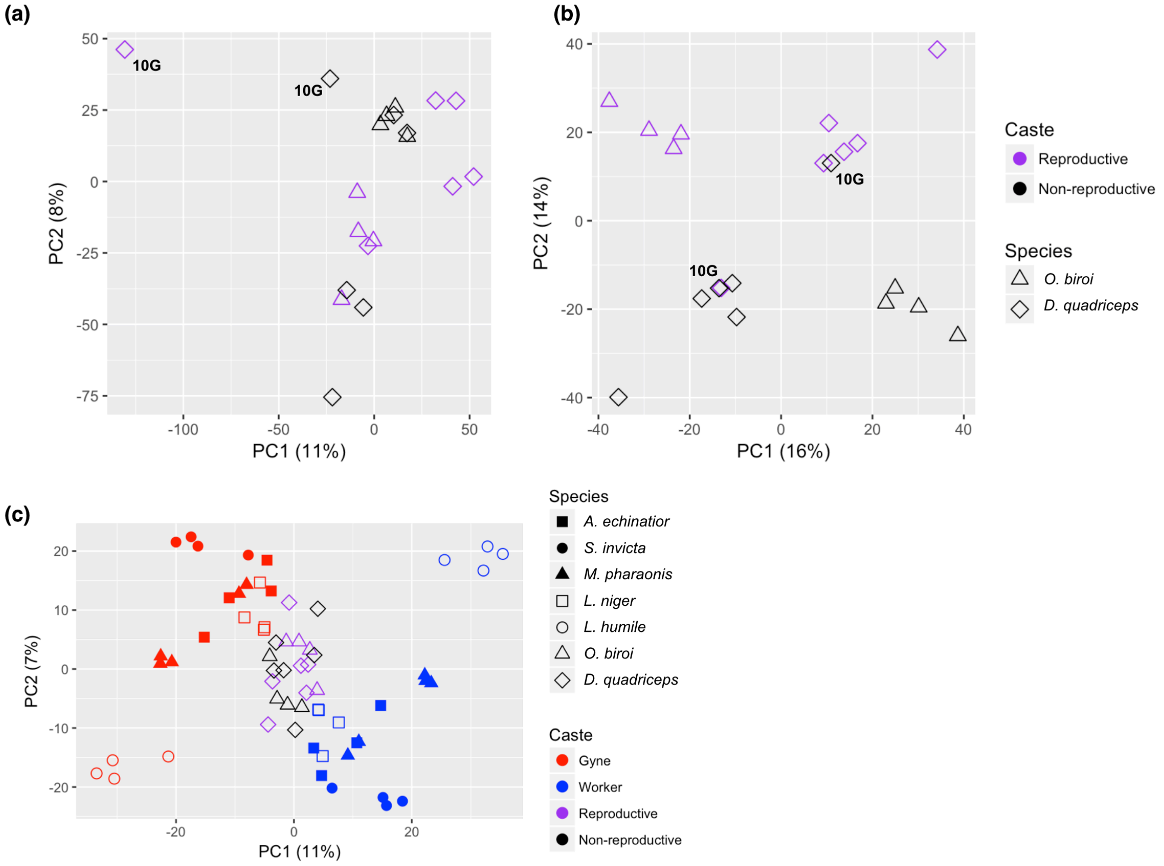
Supplementary Figure 2: Brain transcriptomes for the five ant species with normal caste differentiation, adjusted for additive effects of species identity (a) or colony identity (b). (a) Expression similarity matrix for brain transcriptomes after partialing out additive effects of species identity across the same five ant species as in **Figure 2**, with each cell representing an expression similarity between a pair of samples based on Spearman correlation coefficients across all orthologous genes. (b) Expression similarity matrix for brain transcriptomes after partialing out additive effects of colony identity. Legends are the same as in **Figure 2a**.



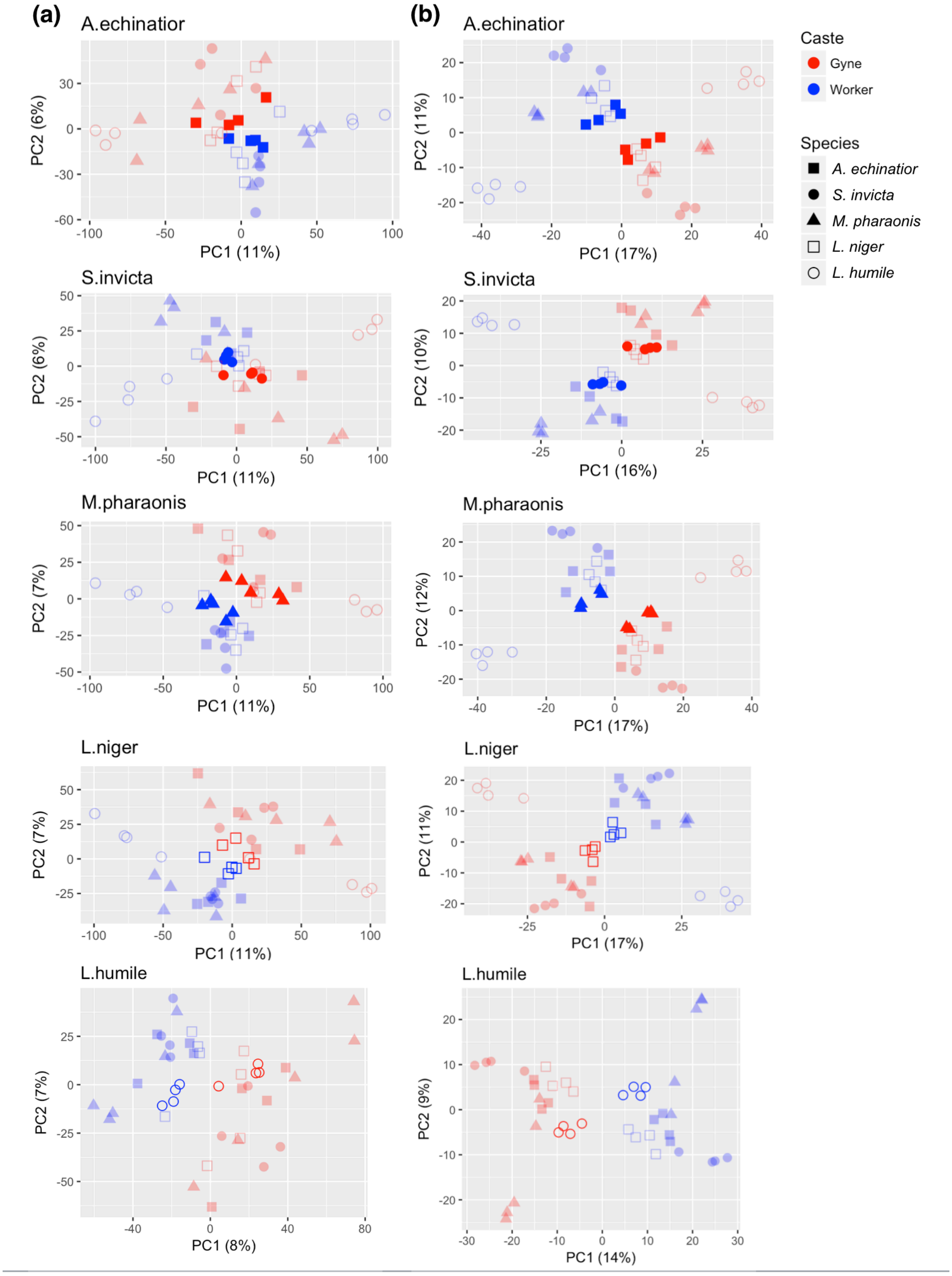
Supplementary Figure 3. PCA of brain transcriptomes normalized with two methods. (a) regularized log transformation (rlog function in DESeq2 [16]) to minimize gene expression differences across samples; this method is particular effective for genes with low expression levels and also normalizes with respect to sequencing library size, and (b) a variance stabilizing transformation (VST function in DESeq2 [16]) that yields a matrix of normalized gene expression values that have approximately constant expression variances across the ranges of mean gene expression levels). For both (a) and (b), we used the raw count data (number of reads mapped to the corresponding ant genomes) of brain transcriptomes of the five ant species with typical caste differentiation as input before normalizing as in (a) or (b). We further adjusted for the effects of species identity by normalizing mean expression level differences across species for each gene. Species identity adjustment method and legends are the same as in **Figure 3a**.



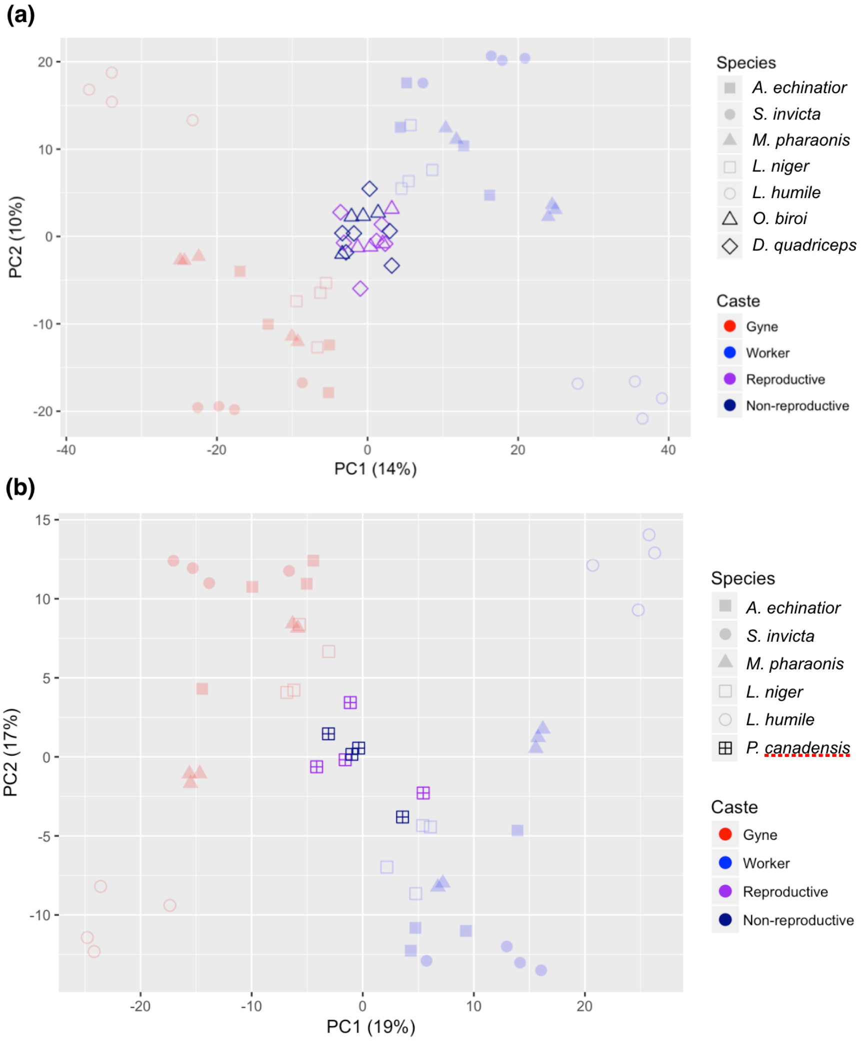
Supplementary Figure 4: Brain transcriptomes for ants belonging to the large and small worker caste in *A. echinatior*, adjusted for confounding effects of either species-level or colony-level identity across the five ant species with normal caste differentiation. Expression similarity matrix for brain samples with each cell representing overall transcriptomic expression similarity between a pair of samples based on Spearman correlation coefficients across all orthologous genes, after normalization for species identity (a) or colony identity (c). The first two principle components of a PCA of transcriptomes obtained from brain samples, normalized for species (b) or colony (d) identity, with species having different symbols and castes having different colours. For both plots, quantile normalized brain transcriptomes were adjusted for effects of species (a and b) or colony (c and d) identity by normalizing expression level differences across species/colonies for each gene. The legends are the same as in **Figure 2**, except for including new colours to highlight large and small workers in *A.echinatior*.



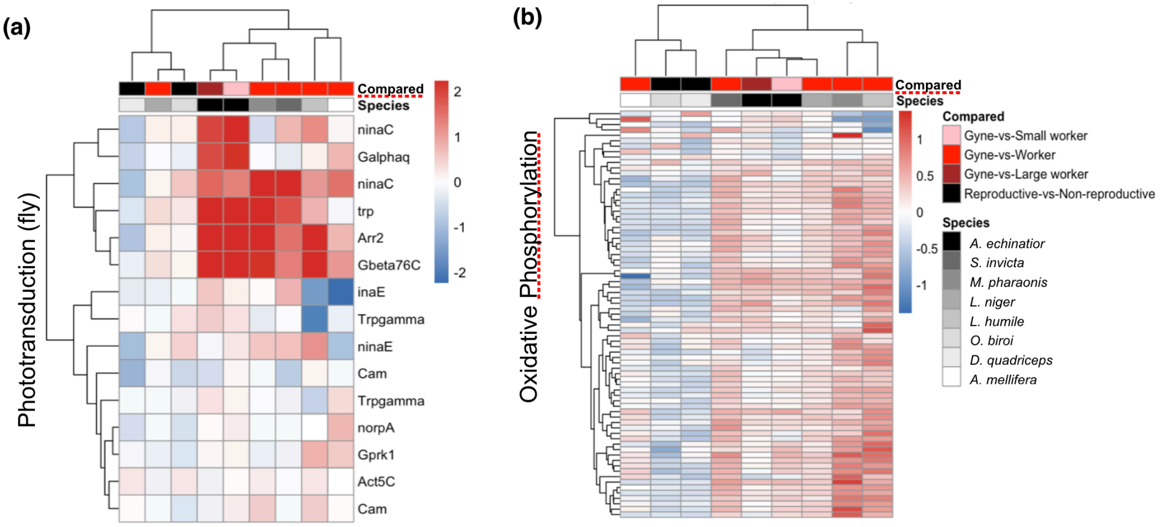
Supplementary Figure 5. PCAs for brain transcriptomes of the two queenless ants, *C. biroi* and *D. quadriceps*. PCAs are based on quantile normalized brain transcriptomes adjusted for (a) the effects of species identity and (b) the effects of colony identity. (c) PCA of the two queenless ants together with the five ant species with typical caste differentiation, using quantile normalized brain transcriptomes adjusted for the additive effects of colony identity. For *D. quadriceps*, samples from colony 10G were specifically labelled to illustrate that fertile and sterile workers in this colony failed to cluster according to general fertility and sterility status in spite of stringent normalization. Colour coding and symbols are identical to **Figure 3**, but additional symbols and colours were used to differentiate between reproductive and non-reproductive workers in the queenless ants *O.biroi* and *D.quadriceps.*



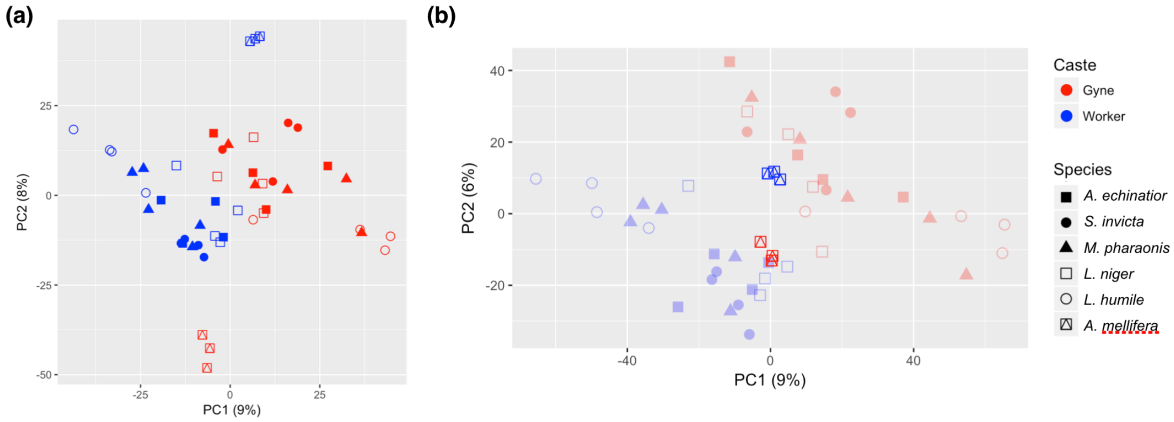
Supplementary Figure 6. Validation of the Singular-Value Decomposition (SVD) method for extracting the first two PC axes that characterize GRN similarity in ants. We first used SVD to extract the first two PC axes of brain transcriptomes adjusted for (a) species or (b) colony identity from any four of the five ant species with typical caste differentiation to obtain training data (blurred data points) and then projected the brain transcriptomes of the remaining ant species as test data (clear data points) onto the extracted PC axes, independently for *A. echinatior*, *L. humile*, *L. niger*, *M. pharaonis* and *S. invicta*. This showed that the first two PC axes extracted in SVD-training can indeed separate caste-specific transcriptome data in subsequent tests of single additional species with normal caste differentiation.



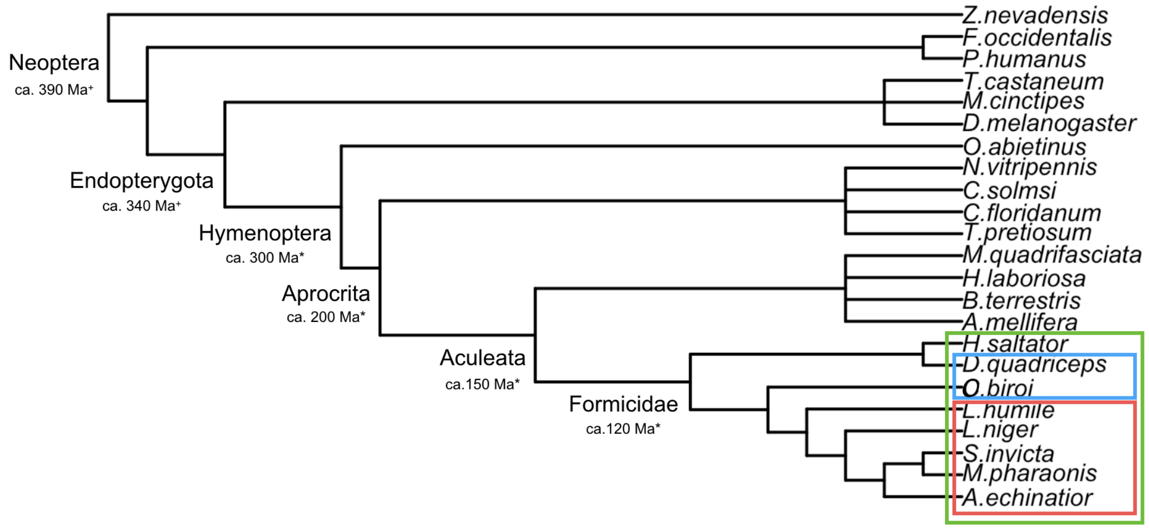
Supplementary Figure 7. Projection of brain transcriptomes adjusted for effect of colony identity from the two queenless ants and the non-superorganismal social wasp *Poliste canadensis*, onto the extracted first two PC axes from the five ant species with typical caste differentiation. For both (a) and (b), the first two PC axes were extracted with SVD from the quantile normalized brain transcriptomes (blurred data points; same method as in **Supplementary Figure 6** but now for all five ant species with normal caste differentiation**)**, after which brain transcriptomes of the queenless ants (a) or *P. canadensis* (b) (clear data points) were projected onto the extracted PC axes. This shows that reproductive and non-reproductive phenotypes in both queenless ants (where both are workers) and *P. canadensis* (where both are totipotent females) cannot be distinguished according to the joint characteristics that specify the GRN divergence for permanent, morphologically differentiated castes in ants with typical caste differentiation.



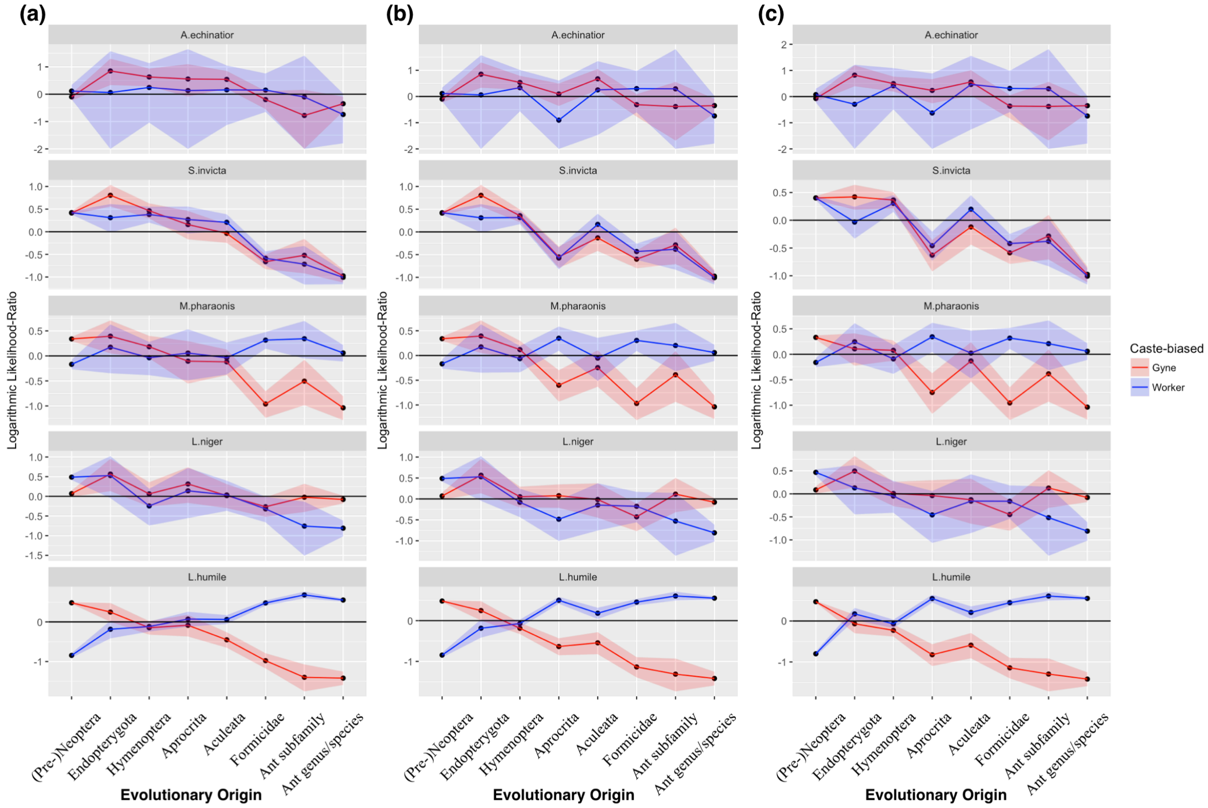
Supplementary Figure 8. Gene expression differences between gyne (reproductive) and worker (non-reproductive) castes in the Oxidative Phosphorylation and Phototransduction pathways across the seven ant species included in our study and the honey bee. Columns represent aligned comparisons between castes across the eight species, and rows represent genes involved in the two selected pathways sorted according to their annotations in *Drosophila melanogaster*. Cell-specific shades of red and blue represent log2 gene expression differences between caste phenotypes, with red indicating gyne(reproductive)-biased expression and blue indicating worker(non-reproductive)-biased expression. Species identities and caste-comparisons are highlighted in the top horizontal bars. Note there are nine columns for eight species because *A. echinatior* provided two types of comparison.



Supplementary Figure 9. Brain transcriptomes adjusted for effects of species identity for the five ant species with normal caste differentiation and the honey bee. (a) PCA of brain transcriptomes showing that the first PC axis separates gynes and workers across all five ant species while the second PC axis separates gynes and workers in the honey bee. (b) Projection of the brain transcriptomes of honey bee queen and worker castes on the extracted first and second PC axes for brain transcriptomes of the five ant species (blurred data points) with normal caste differentiation, which we inferred jointly represent the genetic caste regulatory network of ants (same plot as **Supplementary Figure 7**). This shows that the caste GRN across ants is very different from the caste GRN in the honey bee, because expression bias is in opposite direction across the two PC axes in (a) and for the second PC axis in (b). For both plots, PCAs were obtained from quantile normalized brain transcriptomes adjusted for the additive effect of species identity. Colour coding and symbols are identical to **Figure 3**, but additional symbols were used to characterize the honey bee (*Apis mellifera*).



Supplementary Figure 10. Phylogeny of the 23 insect species used for assessing the phylogenetic bifurcation at which a gene-ortholog first appeared. The five ant species with permanent physical caste differentiation among adult queens and workers are highlighted with a red box and the two ant species that secondarily lost the queen caste and transitioned to either parthenogenesis (*O. biroi*) or having inseminated (gamergate) workers (*D. quadriceps*) are highlighted with a blue box. For species outside the ant family Formicidae (green box), phylogenetic relationships within major branches were not used for any of our analyses, so polytomies did not affect our estimates of the origins of orthologous genes. Estimated divergence times are given along the tree, based on the dated phylogenetic tree of insect (Ma: Million years ago; $^{+}$ from [17], and * from [18]).



Supplementary Figure 11. Testing the robustness of phylogenetic likelihood-ratios for caste-biased genes with different gene loss rate settings: 5% (a), 10% (b), and 20% (c). Colours and legends are the same as in **Figure 1**.

Reference:

1. Mank, J. E. The transcriptional architecture of phenotypic dimorphism. *Nature Ecology & Evolution* **1,** 0006 (2017).
2. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29,** 644–652 (2011).
3. Morandin, C. *et al.* Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. *Genome Biol* **17,** 1 (2016).
4. Hox, J. *Multilevel Analysis: Techniques and Applications, Second Edition*. (Taylor & Francis, 2010).
5. Warner, M. R., Mikheyev, A. S. & Linksvayer, T. A. Genomic Signature of Kin Selection in an Ant with Obligately Sterile Workers. *Mol Biol Evol* **34,** 1780–1787 (2016).
6. Barchuk, A. R. *et al.* Molecular determinants of caste differentiation in the highly eusocial honeybee Apis mellifera. *BMC Developmental Biology 2007 7:1* **7,** 70 (2007).
7. Feldmeyer, B., Elsner, D. & Foitzik, S. Gene expression patterns associated with caste and reproductive status in ants: worker-specific genes are more derived than queen-specific ones. *Mol. Ecol.* **23,** 151–161 (2014).
8. Keller, L. & Ross, K. G. Selfish genes: a green beard in the red fire ant. *Nature* **394,** 573–575 (1998).
9. Bekkevold, D., Frydenberg, J. & Boomsma, J. J. Multiple mating and facultative polygyny in the Panamanian leafcutter ant Acromyrmex echinatior. *Behav Ecol Sociobiol* **46,** 103–109 (1999).
10. Van der Have, T. M., Boomsma, J. J. & Menken, S. B. J. Sex-Investment Ratios and Relatedness in the Monogynous Ant *Lasius Niger* (L.). *Evolution* **42,** 160–172 (1988).
11. Boulay, R., Arnan, X., Cerdá, X. & Retana, J. The ecological benefits of larger colony size may promote polygyny in ants. *J. Evol. Biol.* **27,** 2856–2863 (2014).
12. Hölldobler, B. & Wilson, E. O. *The Ants*. (Harvard University Press, 1990).
13. Hölldobler, B. & Wilson, E. O. The number of queens: An important trait in ant evolution. *Naturwissenschaften* **64,** 8–15 (1977).
14. Mikheyev, A. S., Linksvayer, T. A. & Khaitovich, P. Genes associated with ant social behavior show distinct transcriptional and evolutionary patterns. *eLife Sciences* **4,** e04775 (2015).
15. Lucas, E. R., Romiguier, J. & Keller, L. Gene expression is more strongly influenced by age than caste in the ant Lasius niger. *Mol. Ecol.* **25,** 3389–5073 (2017).
16. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014).
17. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346,** 763–767 (2014).
18. Peters, R. S. *et al.* Evolutionary History of the Hymenoptera. *Current Biology* **0,** 1013–1018 (2017).