ARTICLES
https://doi.org/10.1038/s41559-018-0689-x
nature ecology & evolution

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*] and Guojie Zhang[1,3,4*]

Specialized queens and life-time unmated workers evolved once in the common ancestor of all ants, but whether caste development across ants continues to be at least partly regulated by a single core set of genes remains obscure. We analysed brain transcriptomes from five ant species (three subfamilies) and reconstructed the origins of genes with caste-biased expression. Ancient genes predating the Neoptera were more likely to regulate gyne (virgin queen) phenotypes, while the caste differentiation roles of younger, ant-lineage-specific genes varied. Transcriptome profiling showed that the ancestral network for caste-specific gene regulation has been maintained, but that signatures of common ancestry are obscured by later modifications. Adjusting for such differences, we identified a core gene-set that: (1) consistently displayed similar directions and degrees of caste-differentiated expression; and (2) have mostly not been reported as being involved in caste differentiation. These core regulatory genes exist in the genomes of ant species that secondarily lost the queen caste, but expression differences for reproductive and sterile workers are minor and similar to social paper wasps that lack differentiated castes. Many caste-biased ant genes have caste-differentiated expression in honeybees, but directions of caste bias were uncorrelated, as expected when permanent castes evolved independently in both lineages.

In all but a few evolutionarily derived ants, female eggs develop into either gynes (future reproductive queens) or life-time unmated workers. Such differentiated castes are comparable to germ-lines and soma—an analogy that motivated Wheeler[1] to characterize ant colonies as superorganisms. Irreversible transitions to superorganismality have also evolved in corbiculate bees, vespine wasps and evolutionarily derived termites, always from ancestors where colonies have a single strictly monogamous queen, similar to obligatory multicellular eukaryotes having evolved from ancestors initiating new individuals from a single zygote[2] (reviewed by ref. [3]). Because all metazoans have a common ancestor, their fundamental cell types are homologous, just like queen and worker castes should be homologous when all ants have a single common ancestor[4]. This topic has been exhaustively explored in metazoans[5] where homologous character development has now been confirmed to be encoded and largely regulated by the ancestral gene regulatory network (GRN), consisting of a set of orthologues that interact in shaping homologous phenotypic traits[6]. However, the ontology of traits that characterize social insect superorganisms has received much less attention, and evidence for the existence of GRNs that shaped their independent origins has remained ambiguous[7–9].

Detecting ancestral GRNs for superorganismal caste phenotypes is likely to be difficult because these genes will have become extensively rewired over evolutionary time. They will also have diversified when subfamilies and genera evolved, analogous to later adjustments in metazoan developmental integration and cell type diversification[10,11]. Other genes that presently affect caste differentiation and social behaviour may be specific for subfamilies or genera, and thus without sister lineage orthologues[12–16]. Such lineage-specific genes may have facilitated new adaptations for work coordination, communication or foraging[9,15,16], which is again analogous to how we understand evolutionary elaborations across lineages of multicellular animals[17,18].

Previous studies in honeybees and ants have shown that both ancestral and novel genes have shaped caste differentiation. Many genes upregulated in honeybee worker larvae have no *Drosophila* homologues, suggesting they are novel[12]—similar to whole-body transcriptome comparisons of adult *Temnothorax longispinosus* ants and larvae of *Monomorium pharaonis*[14,19]. However, these single-species studies could not capture the diversity in life history and social structure across species[20], and the use of whole-body transcriptomes precludes detecting caste differences at the homologous organ level. The power of these studies to reconstruct GRNs underlying social traits therefore remained limited[21,22] compared with what has been possible for mammalian organs[23,24]. The only comparative ant study so far used a whole-body approach and detected only a single gene with caste-biased expression across 16 species[8]. The use of whole-body transcriptomes without reference genomes[8] also implies that gene orthology across species cannot be established (Supplementary Information) and that caste-specific gene expression may seem insignificant just because differences in opposite directions across organs are averaged out[25].

The objective of this study was to assess the origin of genes with caste-biased expression and identify the ancestral GRN in the brains of five ant species from different genera and three different subfamilies. We used re-annotated genomes to identify conserved genes, and developed a method to normalize for subfamily/genus-specific effects and colony identity before evaluating caste-biased gene expression across ant species. This allowed us to identify a number of conserved pathways and candidate genes that have previously remained hidden. We validated our approach by examining gene expression in the brains of reproductive and sterile workers of two ants that secondarily lost the queen caste. This allowed us to test the expectation that their GRNs for reproductive division of labour should be less differentiated because fertile and sterile work-

[1]Centre for Social Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [2]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan. [3]China National GeneBank, BGI-Shenzhen, Shenzhen, China. [4]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. *e-mail: jjboomsma@bio.ku.dk; guojie.zhang@bio.ku.dk

ers represent variation within a single caste phenotype, similar to fertile and sterile females of *Polistes* wasps that never evolved differentiated castes with 100% predictable mating status. We also examined brain gene expression data of honeybee queens and workers to assess whether some of the same genes were involved in these two independent origins of superorganismality, and whether directionality of caste-biased expression for shared genes was correlated across ants and bees.

Results

Reconstructing the evolutionary origin of ant genes with caste-biased expression. We produced and re-analysed brain transcriptomes for five ant species with normal caste differentiation—*Acromyrmex echinatior*, *M. pharaonis*, *Linepithema humile*, *Solenopsis invicta* and *Lasius niger*[26]—as well as the queenless ants *Ooceraea biroi*[22] and *Dinoponera quadriceps*[27], which are, respectively, clonal with egg laying and sterile workers and sexual with secondarily evolved dominance hierarchies where one worker becomes inseminated to serve as the (gamergate) queen[28] (Supplementary Table 1). We recovered all orthologous genes across 23 re-annotated insect genomes and traced the evolutionary history of genes for all five species with normal caste differentiation (see Methods) to assign gene origins in the neopteran phylogenetic tree and later gene emergences in the ant tree. We then identified genes with caste-specific brain expression in each species and found that 42–62% of them had early neopteran origins and 8–25% were ant-lineage specific, either at the subfamily or genus level. These percentages were similar but somewhat lower than those for the whole genomic background, which gave 45% (40–51%) of genes with ancient homologues and 26% (18–35%) without (Supplementary Table 2).

We then used log-likelihood ratios to evaluate differences between proportions of caste-biased gene origins at each node and proportions of extant genes with caste-biased brain expression, which allowed direct comparison between caste-biased genes of different evolutionary ages. Likelihood ratios of gyne-biased expression increased with evolutionary age in all five ants (all $P<1\times10^{-3}$; Fig. 1), implying that early neopteran genes are significantly more likely to have gyne-biased expression than genes of more recent origin. However, likelihood ratios of worker-biased genes varied, having similar frequencies as gyne-biased genes in *A. echinatior* (gynes versus small workers), *S. invicta* and *L. niger*, but slightly or steeply decreasing frequencies with the age of gene origin in *M. pharaonis* and *L. humile*, respectively (Fig. 1). In the first three species, ancient genes had varying degrees of higher prevalence than lineage-specific genes (*S. invicta*: 10.9 versus 2.6%; *L. niger*: 4.9 versus 1.9% (both $P<1\times10^{-3}$); *A. echinatior*: 0.3 versus 0.1% (not significant, $P=0.29$)). However, in *L. humile* and *M. pharaonis*, ancient genes were significantly less likely to have worker-biased expression than lineage-specific genes (*L. humile*: 10.4 versus 40.3%; *M. pharaonis*: 2.9 versus 3.7% (both $P<1\times10^{-3}$)) (Fig. 1 and Supplementary Table 3).

Relative effects of species and caste identity on gene expression in ant brains. We identified 6,672 orthologues that were shared across all 7 ant species. Overall expression similarity matrices for the five ants with typical gyne–worker caste differentiation showed that brain samples from the same ant species clustered together regardless of caste (Fig. 2a). Principal component analysis (PCA) of gene expression data clearly separated samples according to species identity, with the first two axes jointly explaining 37% of the total variance in gene expression (Fig. 2b). However, within each species, samples of the same caste clustered more closely than samples from the same colony (Fig. 2a), suggesting that brain gene expression patterns might predict gyne and worker phenotypes, provided lineage-specific determinants of gene expression could be removed.

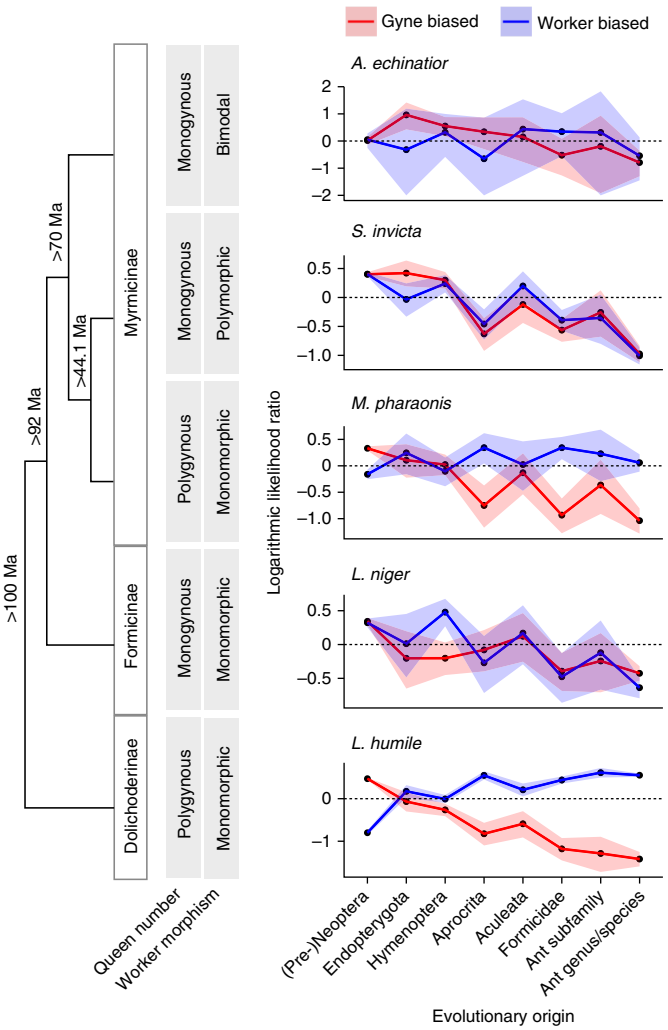

Fig. 1 | Likelihood ratios of genes with caste-biased expression in the brains of five ant species originating at subsequent phylogenetic nodes. All $\log_2$ likelihood ratios are relative to the total number of genes with caste-biased expression in either direction, with positive and negative values indicating higher and lower likelihoods, and shaded areas representing 90% CIs based on binomial distributions. The ant subfamily tree on the left has been supplemented with key characteristics of typical social organization (monogynous = single-queen colonies; polygynous = multiple-queen colonies; monomorphic = unimodal size distribution of workers; polymorphic = skewed size distribution of workers (as in *S. invicta*) or bimodal distribution (as in *A. echinatior*)). For *S. invicta*, we pooled equal numbers of large and small workers, so gene expression refers to average expression levels among different worker sizes. For *A. echinatior*, we present likelihood ratios between gynes and minor workers here, and those between gynes and major workers in Supplementary Fig. 1. Estimated divergence dates are given along the tree, based on the earliest fossil records and estimated phylogenetic divergence of ant clades[59,60].

To separate the effects of genus or subfamily identity and caste on brain transcriptomes, we calculated species-specific *z* scores by subtracting the mean expression levels of each gene and normalizing also for variation in gene expression across samples. This produced a matrix of directly comparable residuals across species, which showed that the brains of same-caste individuals have similar patterns of gene expression (Supplementary Fig. 2a). The first two principal component axes explained 15% of the overall variance and completely separated all samples according to caste (Fig. 3a). Similar patterns were found when transcriptome data were prepro-

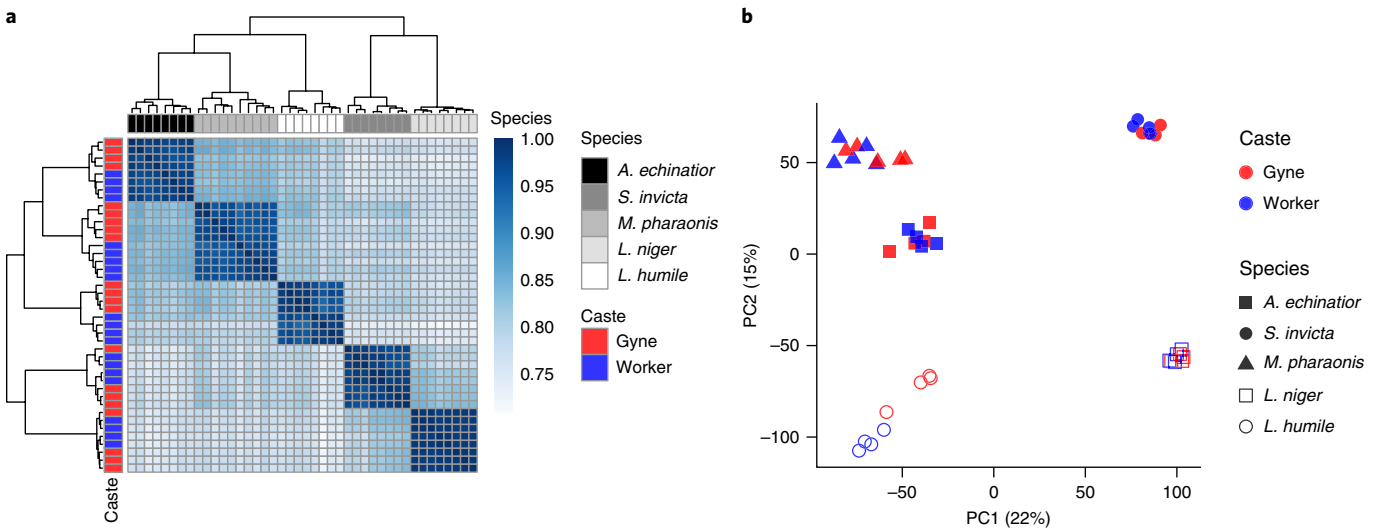

Fig. 2 | Gene expression signatures of species identity and caste type, calculated from 6,672 one-to-one orthologues across 7 ant species. a, Expression similarity matrix for brain samples. Each cell represents overall transcriptomic expression similarity between a pair of samples based on Spearman's correlation coefficients across all orthologous genes, ranging from 0 (white, no correlation) to 1 (dark blue, perfect positive correlation). The type of caste represented by each sample is illustrated in the red/blue left vertical bar and species identities are given in the grey-scale horizontal bar above the plot. **b**, First two principal components (PC1 and PC2) explaining transcriptomic variation across brain samples, with red/blue colours for castes and different symbols for species. Filled symbols refer to the three species of the subfamily Myrmicinae and open symbols refer to the remaining two species belonging to the subfamilies Formicinae (*L. niger*) and Dolichoderinae (*L. humile*). All 6,672 one-to-one orthologues were used for both plots, and expression levels were normalized by $\log_2$ transformation of the number of transcripts per million kilobases and then by quantile normalization across samples.

cessed with other methods, such as regularized log transformation or variance-stabilizing transformation (Supplementary Fig. 3)[29]. A coherent signature of caste-specific gene expression in ant brains thus appears to exist, even though we could not adjust for epistatic and environmental confounders that are likely to also be important. Normalizing for colony-level effects further improved the overall resolution for caste-specific gene expression, with the percentage of variation explained by the two principal component axes increasing to 24% (Fig. 3c and Supplementary Fig. 2b; see Methods).

We used two-way analysis of variance (ANOVA) to unravel the confounding effects of ant subfamily (Fig. 1) using the scores along the first two principal component axes describing transcriptome variation after species-level normalization. This showed that caste remains the dominant predictor along both principal component axes, explaining 44.5 and 42.8% of the respective variances ($P = 8 \times 10^{-9}$ and 6×10^{-8}, respectively; Fig. 3a). Subfamily did not appear to have a directly significant effect (explained variance $= 1 \times 10^{-4}$ for principal component 1 and 1×10^{-5} for principal component 2; both $P = 0.99$), which was expected because normalization for species effects also normalized for subfamily effects, but 26.5 and 23.4% of the respective variances along the first and second principal component axes were explained by the interaction term between caste and subfamily identity ($P = 8 \times 10^{-6}$ and 7×10^{-5}, respectively). Using colony-level normalization produced a similar result (Supplementary Table 4), suggesting that future comparisons across a phylogenetically more diverse set of ants may reveal clear subfamily effects on brain gene expression across castes in addition to the genus/species-level effects that we document.

Brain transcriptomes in ants that secondarily gained a worker caste or lost the queen caste. Additional worker and soldier castes have arisen in many ant lineages, normally with body sizes intermediate between queens and small workers[30–32]. Within the attine ants, which evolved 55–60 million years ago (Ma), only the evolutionarily derived *Atta* and *Acromyrmex* leaf-cutting ants that originated around 15 Ma[33] have such additional castes as secondary innovations. We compared the brain transcriptome similarity (normalizing again for species identity) of large and small *A. echinatior* workers relative to workers of the other four species. This showed that large-worker brain transcriptomes were distinct from those of gynes, but that their expression profiles were more distant from those of the monomorphic workers of the four other ant species (Supplementary Fig. 4), as expected when large workers are evolutionarily derived. Also, these resolutions further improved when we normalized for colony-level variation (Supplementary Fig. 4).

The queen caste has secondarily disappeared in several ant lineages. This happened before workers had irreversibly lost the sperm storage organ (spermatheca), so mated workers could evolve to be ergonomically cheaper egg layers when colonies were small[34,35]. Rarely, ant lineages abandoned sexual reproduction altogether[36,37] to produce new workers from unfertilized worker eggs[22,27]. Using the published brain transcriptomes of reproductive and non-reproductive workers (Supplementary Table 1), we found (after species-level normalization) that clonal reproductive and non-reproductive *O. biroi* workers were separated by the first principal component axis and that separation once more improved after normalizing for colony identity. Brain gene expression differences were less pronounced in *D. quadriceps* where colony-level normalization was needed to make the two worker phenotypes segregate in most but not all colonies (Supplementary Fig. 5a). In both species, the expression patterns of fertile and sterile workers segregated along the same principal component axis, suggesting there may have been convergent evolutionary responses to losing the queen caste (Supplementary Fig. 5b).

We used singular-value decomposition to assess overall gene-expression similarity between the two queenless and five other ants. We first validated this method for the normalized brain transcriptome data of the five ants via leave-one-out jack-knife resampling,

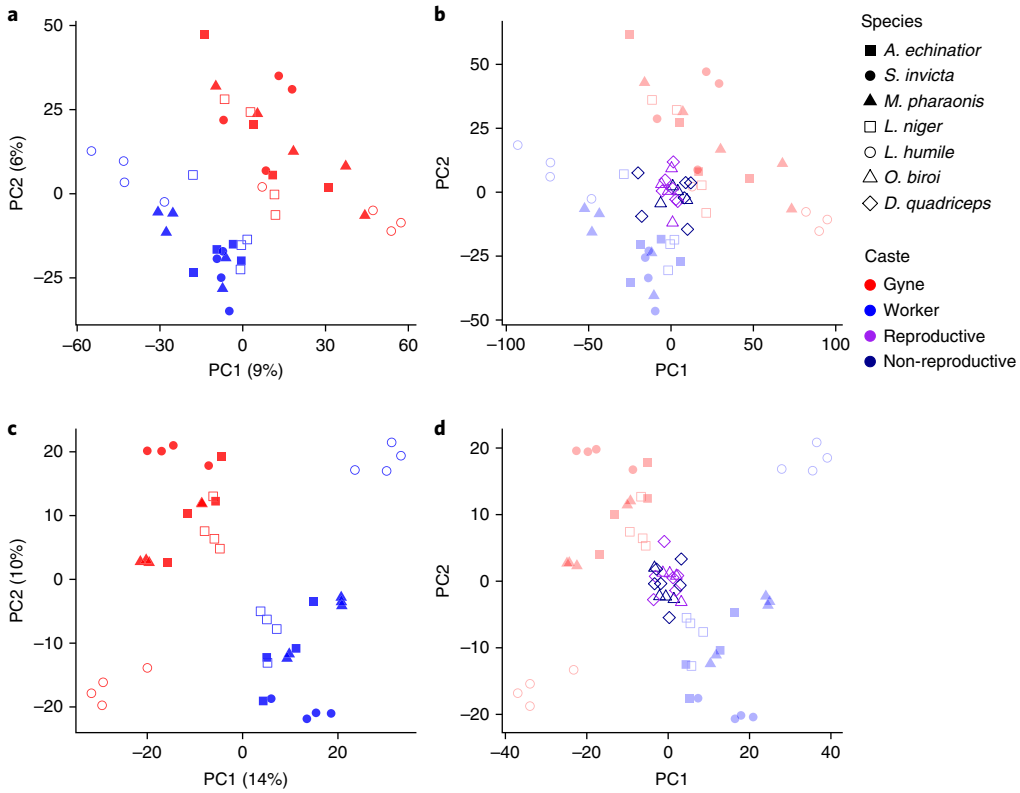

Fig. 3 | PCAs for brain transcriptomes calculated from 6,672 one-to-one orthologues across seven ant species, after adjusting for either species- or colony-level variation in the mean and variance of gene expression. a,**c**, PCAs for brain transcriptomes after partialling out the additive effects of species identity (**a**) and species plus colony identity (**c**) in the same five ant species as in Figs. 1 and 2. **b**,**d**, Projection of similarly normalized brain transcriptome scores for the queenless ants *O. biroi* and *D. quadriceps* (foreground darker symbols) on the two principal components of **a** and **c**, respectively. All 6,672 one-to-one orthologues were used in both plots, and the colour coding and symbols are identical to Fig. 2b for the ant species included in both analyses. Additional symbols and colours were used to differentiate between reproductive and non-reproductive workers of the two queenless ants.

which assigned four ants as training data and then tested the match between expected and observed caste-specific expression in the fifth species. This showed that segregated gene-expression profiles of queens and workers could be retrieved for all five combinations, and again more consistently after normalizing for colony-level variation (Supplementary Fig. 6). Projecting the queenless ant data onto the five normal ant species PCA matrix showed that brain transcriptomes of their reproductive and non-reproductive workers fell in the centre of the plot and failed to segregate, both after normalizing for species- and colony-level differences (Fig. 3b,d). With the first two principal component axes probably representing the ancestral GRN for differential queen–worker gene expression, this result is consistent with both queenless ants having lost the ancestral caste GRN (Supplementary Fig. 5), although the genes composing the GRN are still present. We further tested whether phenotypic plasticity for reproductive roles within a single female bauplan would generally fail to produce segregation along our principal component axes for genetically hard-wired castes by plotting brain gene-expression data for breeders and helpers of *Polistes canadensis* wasps[27]. This showed a similar lack of segregation, consistent with these wasps lacking permanent castes because they never passed the point of no return to superorganismality[3] (Supplementary Fig. 7).

Key ancestral genes in the genetic regulatory network mediating queen–worker caste differentiation in ants. To elucidate functional categories of gene-sets regulating gyne–worker caste differentiation, we identified pathways with consistent caste-biased expression across the five ant species using enrichment analysis[38]. After further colony-level normalization, we found that metabolism-associated genes (for example, oxidative phosphorylation and carbon metabolism) were significantly upregulated in gynes across the five ant species. Also, the phototransduction pathway was upregulated in gynes except for *L. niger* where a same-direction difference was not significant. Such enrichments were not detected in the reproductive workers of both queenless ants, as expected when they lack specialized queen phenotypes that need vision during mating flights (Supplementary Table 5 and Supplementary Fig. 8).

Using a generalized linear model (GLM) to account for additive effects of caste and species identity on the expression of the 6,672 orthologous genes, we identified 42 genes with significant differential expression between gynes and workers across all 5 ant species (38 upregulated in gynes, 4 upregulated in workers; fold change > 1.5; false discovery rate (FDR) < 0.01). We interpreted these as potential key genes in the conserved ancestral GRN for queen–worker differentiation (Supplementary Table 6). Consistent with the gene-set/pathway enrichment results, several of these genes are involved in vision, including *neither inactivation nor afterpotential protein C* (*ninaC*) and the protein *bride of sevenless* (*boss*). Also, genes associated with hormonal and insulin systems, including *neuroparsin-A-like*, *vitellogenin-3-like*, *eclosion hormone*, *locusta insulin-related peptide* (*LIRP*) and *insulin-like growth factor-binding protein complex acid labile subunit* (*IGFALS*), were significantly upregulated in gynes across the five ant species (Fig. 4), whereas *Ras-related and oestrogen-regulated growth inhibitor-like protein* (*RERGL*) was significantly upregulated in workers (Fig. 4).

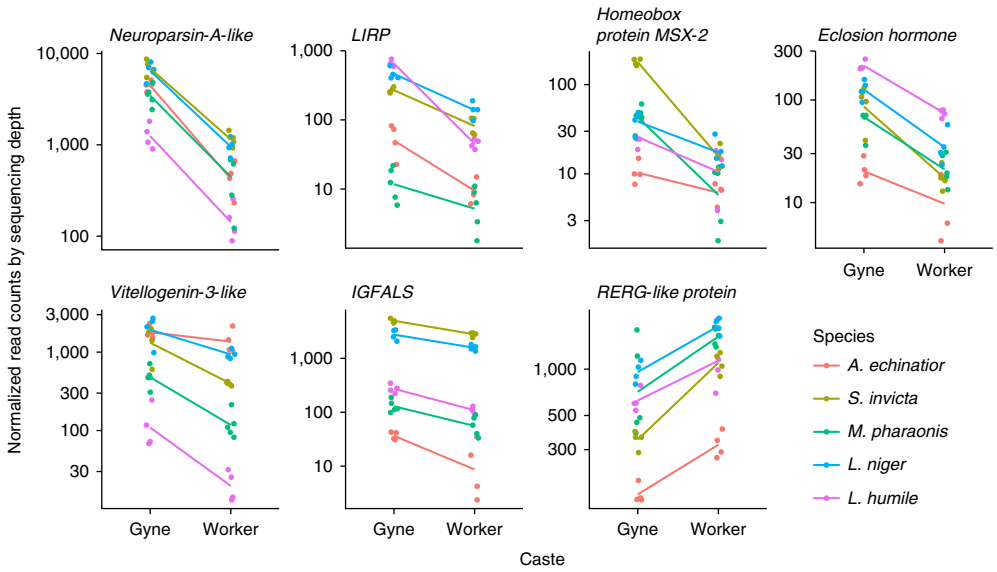

Fig. 4 | Expression levels for conserved caste regulatory genes expressed in the brains of gynes and workers across the five ant species with typical queen–worker differentiation. Dots show the gene expression levels for each brain sample. Lines connect the average brain gene expression values for gynes and workers (both coloured according to species identity). All expression differences between gynes and workers were significant (FDR < 0.01, two-sided Wald tests).

To tentatively explore the identity of the ancestral ant GRN components, we constructed a cross-species gene co-expression matrix (see Methods), which produced clusters of genes exhibiting conserved co-expression with the key genes of the conserved ancestral GRN that were nested within these clusters (Supplementary Information). These networks seemed consistent with the extant ant GRN having maintained overall aspects of co-regulation, but this result will need validation and refinement with species-specific transcriptome data for time series of developmental phenotypes.

Toolkit genes involved in convergent caste differentiation in ants and corbiculate bees. The corbiculate bees (honeybees, stingless bees and bumblebees) and ants independently evolved the permanent castes[39] that define superorganismality[3]. It has been hypothesized that there exists a toolkit of genes that are differentially expressed in breeder and helper phenotypes across insect lineages[9,40], but whether such genes play analogous roles in independent transitions to superorganismality remains unknown. We therefore compared the underlying GRNs for caste differentiation between ants and honeybees using pupal honeybee brain transcriptome data of orthologous genes[41]. Among the 42 conserved caste GRN genes in ants, 39 have orthologues in the honeybee and 15 of these (38.5%) had significant differential expression between honeybee queens and workers (fold change > 1.5; $P < 1 \times 10^{-3}$; Supplementary Information). This percentage is higher than the overall proportion of genes with any differential expression across honeybee gynes and workers (733/6,036 = 12.1%; $P < 1 \times 10^{-4}$), consistent with an ancestral hymenopteran toolkit of genes mediating reproductive division of labour[9]. However, only 5 of these 15 genes had the same direction of caste-biased expression between ants and honeybees (an uncharacterized protein *LOC105150705*, *elongation of very long chain fatty acids protein AAEL008004-like*, *ninaC*, *neuroparsin-A-like* and *homeobox protein MSX-2*; Supplementary Table 6), a ratio not significantly different from the 12.1% overall proportion ($P = 0.58$). This lack of directional consistency in caste-biased expression implies that the ancestral GRNs of ants and honeybees are uncorrelated despite including some of the same toolkit genes, as expected when irreversible transitions to superorganismality evolved convergently.

Independent recruitment of genes with caste-biased expression was confirmed when we considered all orthologous genes between ants and the honeybee. After normalization, ant gynes and workers segregated along the first principal component axis and honeybee gynes and workers segregated along the second principal component axis (Supplementary Fig. 9a). When we projected the honeybee caste-specific brain transcriptomes on the first two principal component axes for ant transcriptomes, the second principal component axis separated gyne and worker brain samples of the honeybee, but with expression bias in the opposite direction compared with ants (Supplementary Fig. 9b).

Discussion

Our study reconstructs the contours of the ancestral brain GRN in queens and workers[20,42] that established permanent reproductive division of labour in ants[43,44]. We found that later subfamily/genus effects blur this conserved ancestral GRN[23], as expected after around 120 Myr of adaptive radiation[20,43]. This finding underlines the necessity of elaborate normalization procedures, which possibly explains why our study, based on a single-matrix approach across species, achieved higher resolution than Morandin et al.[8] (Supplementary Information). Some ant GRN genes, such as *vitellogenin-3-like* and *neuroparsin-A-like*, are homologues of juvenile hormone genes[40,45] involved in the regulation of facultative and obligate reproductive division of labour across social Hymenoptera[40,46,47]. However, the *myosin light chain* gene, reported as differentially expressed between castes across 16 ant species[8], was not part of our GRN. This gene was slightly overexpressed in gynes (rather than underexpressed, as in ref. [8]) in four ant species (not in *L. niger*), suggesting it exemplifies the problematic resolution of whole-body transcriptomes, because it must have had opposite caste-biased expression in other body parts. This previous study[8] used overlapping species comparisons to identify common genes with caste-biased expression, which implied that type 2 errors increased proportionally. Our GLM approach avoids this problem by integrating all cross-species expression levels into a single data matrix after multiple normalization procedures, so

that sampling variance in the number of detected ancestral differentially expressed genes becomes gradually less when more species are included in a comparative analysis (Supplementary Information). Three-quarters of the genes with caste-biased expression across the five ants that our study discovered have not been identified in this role before, corroborating the improved resolution that our methodological approach offers for future comparative transcriptomics studies of social evolution questions.

Our results support the notion that permanent reproductive division of labour between specialized queens and life-time unmated workers was a major evolutionary transition in the ancestral ants[1,3], involving a single genetically hard-wired point of no return that established the brain GRN for caste. Our analyses beyond the five ants with normal caste differentiation are consistent with genetically hard-wired caste differentiation being fundamentally different from what phenotypic plasticity can achieve, both in ants that secondarily lost the queen caste and in dominance hierarchies of cooperatively breeding *Polistes* paper wasps that never evolved permanent castes. As envisaged by Wheeler[1], only superorganismality based on permanently differentiated castes invites analogies with metazoan multicellular development. One of these is serial homology, as we found for the brain GRNs of the two distinct worker castes of *A. echinatior*—a pattern functionally reminiscent of transcriptomes specifying the upper and lower first molars of mice that originated from a single tooth[48]. Our results also confirm that the point-of-no-return transition to superorganismality in the corbiculate bee lineage (bumblebees, stingless bees and honeybees)[39,49] was independent from the ants. Our study predicts that brain GRNs for permanent castes between these three corbiculate bee lineages should be mutually correlated because they share a recent common ancestor, but uncorrelated with the queen and worker brain GRN in vespine wasps, which evolved superorganismality independent of ants and bees. Larger-scale brain GRN studies across the superorganismal lineages as defined by Wheeler[1,3] will thus be illuminating to further explore lineage-specific elaborations of social organization, such as the tentative link that we report between the recruitment of novel genes for worker functions and obligate polygyny in ants (Fig. 1 and Supplementary Information).

Independence of gene-expression directionality in ants and the honeybee does not preclude that orthologous genes are part of ancestral genetic toolkits across convergently evolved superorganismal lineages[9]. The toolkit concept is about participation[9,23], whereas the ancestral GRN concept is about correlated co-expression across castes. The existence of a toolkit is purely an empirical issue, but an ancestral GRN has predictive elements because co-expression should be similar across lineages derived from the same major transition to superorganismality, but not across independent, irreversible transitions of this kind[3]. Our findings were consistent with this expectation, but the details of this inference remain somewhat provisional because transcriptomes were from adult brains in ants and from pupae in the honeybee. Although we partly validated these results by comparing pupal transcriptomes from the honeybee and fire ant (see Methods), direct comparisons with transcriptomes of virgin adult honeybee queens are needed. More comparative brain transcriptome studies of ancestral GRNs across the ant, bee and wasp lineages with permanently unmated workers will undoubtedly add to our understanding of the selection forces that convergently produced analogous forms of superorganismal social organization.

Methods

Sample collection. *A. echinatior.* Five colonies of *A. echinatior* were collected in Panama in 2004 (Ae263), 2014 (Ae704) and 2016 (Ae747, Ae764 and Ae767) and reared at the Centre for Social Evolution, University of Copenhagen under a constant temperature of around 25 °C and around 70% humidity. Gynes and minor workers were collected within the fungus garden, whereas major workers were collected while they were foraging outside, always during daytime (10:00 to 16:00), and isolated in groups of 6–10 individuals of the same caste in small fluon-coated plastic boxes. Ants were then put on ice to reduce activity before their brains were dissected in a dissection dish with diethyl pyrocarbonate-treated phosphate-buffered saline on ice. Dissection was done under a stereomicroscope with sterile (heat-treated) forceps, and only samples consisting of the complete and undamaged anatomical structure of a brain were retained. Each brain dissection was completed within 5 min and dissected brains were transferred to 1 ml RNAlater (pooling a maximum of 12 brain samples of the same caste), kept at 4 °C overnight and then transferred to a −20 °C freezer for long-term storage.

M. pharaonis. Five colonies of *M. pharaonis* (3rd Room X1/B, CS10, Donor 3rd, Donor BQ- and 3rd Room X1/A) were reared at the Centre for Social Evolution, University of Copenhagen under a constant temperature of 27 °C and 50% humidity from a stock collected in 2008[50]. Gynes were separated from males at the pupal stage and reared with 10–15 workers in fluon-coated petri dishes to be collected within 3 d after they hatched as adults, whereas workers were collected directly from the source colonies when they were foraging. All collections were done during daytime (10:00 to 16:00) and adults were kept in groups of 6–10 of the same caste in small fluon-coated boxes. Dissection procedures were the same as in *A. echinatior*, except that dissected brains were transferred into RNAlater with a droplet of phosphate-buffered saline because the brains were too small to be handled directly with forceps. On storage, each 1 ml RNAlater sample contained at most 6 brain samples of the same caste to ensure the RNAlater concentration remained high enough.

L. humile. Five colonies of *L. humile* were collected in Caldes d'Estrac, Spain (Catalan3b) and Castell d'Aro, Spain (Main 3a, Main 4a, Main 5b and Main 5d) in 2016 and reared at the Centre for Social Evolution, University of Copenhagen under a constant temperature of 27 °C and 50% humidity. Gyne isolation, worker ant sample collection, brain dissection and storage proceeded in the same way as for *M. pharaonis*.

S. invicta. Four monogynous (single-queen) colonies were collected in Taoyuan, Taiwan—two in October 2012 and two in April 2014—and transferred to a fluon-coated plastic box in the laboratory. Approximately 120 gynes and 200 workers from each colony were randomly selected for dissection. An equal number of small and large workers were selected. Ants were cooled on ice, then soaked into RNAlater and dissected with forceps in RNAlater on ice. Forceps were dipped into ethanol and flame-sterilized between samples. Each dissection was completed within 5 min, then transferred into tubes with 50 µl TRIzol (maximum 12 brain samples of the same castes per tube) and stored at −80 °C. All sampling and dissections were done within one month after collection in the field.

RNA extraction, complementary DNA (cDNA) library construction and RNA sequencing. *A. echinatior, M. pharaonis* and *L. humile* brain samples were retrieved from their RNAlater storage tubes, after which 10–12 brain samples for each caste and colony were pooled before RNA extraction. Total RNA was extracted using a Qiagen RNeasy Plus Micro Kit (catalogue number 74034) according to the manufacturer's protocol (2013), and quality was evaluated with an Experion HighSens RNA analysis assay (Bio-Rad) before library preparation. Total amounts of RNA were quantified using an Invitrogen Qubit RNA High Sensitivity assay. Based on the RNA sample concentration, Ambion ERCC RNA Spike-In Mix (catalogue number 4456740) was added to each sample according to the manufacturer's instructions. cDNA libraries were then constructed using a Clontech SMARTer PCR cDNA Synthesis Kit (catalogue number 634925), following the protocol described in the manufacturer's user manual (protocol number PT4097-1). The numbers of PCR cycles for the second-strand cDNA synthesis reactions were optimized as described in the user manual. RNA sequencing was then conducted at BGI's Sequencing Centre, Wuhan on an Illumina HiSeq 4000 platform with 150-base pair (bp) paired-ends reads. We generated around 3–5 Gbp of RNA sequencing (RNA-Seq) data for each sample (Supplementary Table 1).

For *S. invicta*, brain samples (already in TRIzol) of the same caste and colony were first pooled and then homogenized with ceramic beads in a bead shaker (FastPrep-24; MP Biomedicals), after which chloroform was added to separate the liquid phase according to a standard TRIzol protocol. Afterwards, the aqueous phase containing the total RNA was directly applied to the GE Healthcare Illustra RNAspin Midi Kit for RNA purification according to the manufacturer's protocol. RNA sequencing was conducted at the Academia Sinica High Throughput Genomics Core, Taipei, Taiwan. All samples were sequenced on an Illumina HiSeq 2500 platform with 101-bp paired-ends reads. For one set of biological replicates, additional sequencing was obtained on an Illumina Genome Analyzer platform with 96-bp paired-end reads and an Illumina MiSeq platform with 250-bp paired-end reads. This generated 4–10 Gbp of RNA-Seq data for each sample of *S. invicta* (Supplementary Table 1).

We further included published brain transcriptome data for *L. niger*[26], *O. biroi*[22] and *D. quadriceps*[27] in our comparative analyses, each with four to six colony replicates (see 'Sample collection' and Supplementary Table 1).

Quality checks of the RNA-Seq data. For the three sets of RNA samples with ERCC spike-in, the data quality was checked with mapped reads ($\log_2$ transformed

and with one pseudo read to avoid $\log_2[0]$), so we could remove outlier samples if their spike-in Pearson correlation coefficients across samples were < 0.95. Similarities in caste-biased gene expression were also used to check data quality. For each ant species, mapped transcriptome reads ($\log_2$ transformed as mentioned above) were compared among samples within the same caste of a focal species and checked for potential biological or technical biases when producing within-caste Pearson correlation coefficients < 0.9, as suggested by the ENCODE RNA-Seq guidelines[51]. These checks implied that we excluded seven outlier samples collected from deviating locations (assuming biological reasons for being outliers) or being prepared with somewhat deviating experimental procedures (assuming technical reasons for being outliers). If no associations with potential biological or technical biases could be identified from our notebooks, samples were retained for downstream analysis.

Improved genome annotation for all seven ant species based on RNA-Seq data. The genomes of the seven ant species included in this study (that is, *A. echinatior*, *M. pharaonis*, *L. humile*, *S. invicta*, *L. niger*, *D. quadriceps* and *O. biroi*) were re-annotated based on the National Center for Biotechnology Information genome annotations of all of these seven species plus *Drosophila melanogaster* and *Nasonia vitripennis*. The re-annotations were done with GeMoMa[52]—a method based on the assumption that amino acid and intron positions are conserved across the insect phylogeny. This tool was chosen as particularly suitable for our project because the explicit use of exon–intron-structure information was shown to improve annotation accuracy[52], and because integrating genome annotations across ant species increases the discovery rate of orthologous genes in each ant lineage. For each target ant species, the genome was first re-annotated by independent comparison with the genome annotations of the other ant species and the two non-social insects. Then, RNA-Seq data for the target ant species were used to pursue re-annotation for each of the ant genomes. The re-annotated genomes of the target ant species were finally filtered for the combined analysis, using the genome annotation filter provided by GeMoMa to remove redundant or overlapping exons before they were used as reference genomes for RNA-Seq analysis. Overall, these re-annotations increased the total number of annotated genes across all ant genomes by 27% (from 90,856 to 115,739), which increased the accuracy of our orthology inferences.

Improved genome annotations for 16 other social and non-social insect species. To infer the origins of orthologous genes in the most accurate way possible, we also re-annotated the genomes of *Harpegnathos saltator*, *Apis mellifera*, *Bombus terrestris*, *Habropoda laboriosa*, *Melipona quadrifasciata*, *Trichogramma pretiosum*, *Copidosoma floridanum*, *Ceratosolen solmsi*, *N. vitripennis*, *Orussus abietinus*, *D. melanogaster*, *Mochlonyx cinctipes*, *Tribolium castaneum*, *Pediculus humanus*, *Frankliniella occidentalis* and *Zootermopsis nevadensis* (Supplementary Fig. 10). For each of these insect species, we used the six improved ant genomes as references (excluding *L. niger* because we added this species at a later stage of analysis), and re-annotated the insect genomes using GeMoMa as above, except for the integration of RNA-Seq data as they were only available for the seven ant species.

Finding gene-orthologues across the 23 insect genomes. Orthologous genes of the seven ant species were identified in four steps. For the orthologues among the 16 re-annotated insect genomes other than the 7 focal ant species of our study we: (1) retained the longest protein sequence for each gene when we obtained more than one transcript; and (2) performed reciprocal best-hit BLASTP with an *E* value < 1×10^{-5} and sequence coverage >50%. For the 7 ants, we then: (3) used the homologues that both fell within the 95% confidence interval (CI) of being orthologues in reciprocal BLASTP searches, and confirmed assignments by additional synteny information because only orthologues should have remained co-linear with at least their immediately neighbouring genes; and (4) retained only genes with exactly one orthologue for each of the 7 ant species so that interspecific transcriptome comparisons were always based on one-to-one orthologue matches. Steps (2) and (3) were done using Proteinortho 5.16 with default settings[53].

Determining the origins of orthologous genes. Using the package Proteinortho[53], the origin of each gene was determined by integrating the orthologue information and phylogenetic relationships among species (Supplementary Fig. 10) using three criteria. First, if a reference gene had one or more orthologues in a branch with multiple species, we considered this to be sufficient evidence for the occurrence of a target gene's orthologue in a branch. Second, the occurrence of a target gene's orthologues along phylogenetic branches was conceptualized as a binomial process, and the probability of missing an orthologue (that is, the expected rate of loss of a gene) on a branch was set to be 10%. This threshold is somewhat arbitrary, but allowed missing orthologues (either due to gene loss or miss-annotation) in recent branches to have a somewhat higher probability while still considering orthologue losses to be independent among branches. This procedure thus controlled for the number of missing orthologues along branches (that is, for every candidate orthologue occurring in the earliest branch, we examined the inferred orthologue's distribution along later branches and discarded a distribution as pseudo-orthologous if the probability was <0.01, after which we continued the same procedure for the next oldest branch). This refers to situations where orthologues occurred in an ancient phylogenetic node without finding corresponding orthologues in multiple recent lineages. Third, the origin of a focal gene was considered to be unambiguously identified with the oldest paired orthologue occurrence in the phylogenetic tree.

To illustrate the robustness of our approach, we present results for a 10% rate of gene loss in the main manuscript and provide complementary analyses for putative rates of 5 and 20% in Supplementary Fig. 11.

Likelihood estimation of the nodes where caste-biased gene expression evolved. The likelihood (LR) of caste-biased gene expression at each phylogenetic node (*i*) was calculated as the log ratio between the proportion of genes with specific caste-biased expression in a focal evolutionary node (*n*) relative to the overall proportion of genes with the same type of caste-biased expression in the transcriptome (*N*),

$$\mathrm{LR}(i)_{\mathrm{caste}} = \log_2\left[\frac{n(i)_{\mathrm{caste}}}{n(i)} : \frac{N_{\mathrm{caste}}}{N}\right] \quad (1)$$

where 'caste' can either be gynes or (subcategories of) workers. The ratio of $n(i)_{\mathrm{caste}}/n(i)$ and N_{caste}/N also produced a CI of the likelihood ratio estimates when assuming that recruitment of genes for caste-biased expression follows a binomial distribution.

Transcriptome quantification. Transcriptome quantifications were done with the Salmon pipeline[54]. In brief, RNA-Seq data for the seven ant species were quasi-mapped to the corresponding transcriptome with improved annotation (as described above), after which bias-correction options were turned on to account for guanine–cytosine bias and sequence-specific bias. For the RNA-Seq data of *A. echinatior*, *M. pharaonis* and *L. humile,* where ERCC spike-in was added before cDNA library construction, additional quantifications were done with corresponding transcriptome + ERCC pseudo-scaffolds.

Identification of genes that are differentially expressed among castes. For each ant species, genes with differential expression among castes were identified using the RNA-Seq data from the samples that had passed quality checking. Analyses were done with the DESeq2 package in R[29] as follows: (1) we used transcript-level read counts from Salmon as input for DESeq2; (2) we aggregated the transcript-level read counts to become gene-level read counts using tximport; (3) we modelled the gene-level read counts as: count ~ caste + colony, assuming that the influence of caste and colony on gene expression levels is additive (enabling us to include the possibly confounding effects of colony origin on caste-specific gene expression levels of colony members); and (4) we tested for gene expression differences between castes using DESeq2 and adjusted *P* values after independent hypothesis filtering, which controls for false discovery due to multiple testing while increasing the detection power[55]. Genes were identified as being differentially expressed between castes when the adjusted *P* value was < 0.05 based on a two-sided Wald test.

For the honeybee data in our subsequent comparison with ant data, we used recently published brain transcriptome data of the P4 stage of worker pupae and the P3 stage of gyne pupae, as these two stages represent the same developmental age[41]. We used the same procedure to detect differentially expressed genes between castes, as in our ant comparisons, except for using a more stringent fold-change expression difference between castes as a threshold for acceptance of differentially expressed genes (that is, we kept a 1.5-fold-change expression difference between castes as the null model and used a more stringent *P* value < 1×10^{-3} instead of 1×10^{-2} used in our ant species comparisons). See Supplementary Information for a comparative evaluation of *P* value cut-offs.

While the choice of *P* value or fold-change cut-off will affect the number of genes identified as differentially expressed (detection power), other technical and experimental design differences will also affect the detection power, which makes it difficult to remain fully consistent when using data from multiple sources as we do in the present study. However, as our PCA and clustering analyses used whole transcriptomes (that is, the expression level of all genes in brains), we did not rely on specific *P* values or fold-change cut-offs for our overall comparisons of caste-specific transcriptomes among ant and other social insect species. This implies that our main conclusions (of consistent directional gene expression differences in queens versus workers across ant species, and of genetic regulatory networks between ants and the honeybee being uncorrelated in their direction of expression bias) are independent of the choice of *P* value and fold-change cut-off.

We used pupal brain transcriptome data of honeybee queens and workers for comparison with ant transcriptomes because adult brain transcriptomes (RNA-Seq data) were not available. Using only adult caste data would have been preferable, but we partially validated our comparisons by repeating our analyses (data not included) using pupal brain transcriptome data of gynes and workers in *S. invicta*, which showed that they segregated for the same two principal component axes as the brain transcriptomes of adult gyne and worker castes in the five ant species. Also, direct pupal comparisons between *S. invicta* and the honeybee recovered the distinctly opposite patterns of directionality in gene expression that we report. We therefore assume that the conclusions of our comparisons of gene-expression bias in ants and the honeybee are reasonably robust, but we have added a caveat on this in the Discussion.

Normalization of gene-expression scores across samples. To compare transcriptomes across samples, expression levels for genes with one-to-one orthologues in all seven ant species ($n = 6{,}672$) were normalized across samples. Normalizations were done as follows: (1) for each sample, transcripts per million kilobases for each gene were increased with 1×10^{-5} pseudo-transcripts (to avoid log[0] scores) and then $\log_2$ transformed; and (2) the $\log_2$-transformed transcriptome data were then quantile normalized among samples (that is, by ranking the expression levels for all genes in each sample and then replacing gene-specific expression levels with rank numbers across samples to equalize the average expression-level scores). This quantile normalization ensured that the overall distributions of gene expression levels remained the same across samples so that the effects of any technical artefacts were minimized[56].

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses and Gene Ontology analyses. Gene Ontology and KEGG annotations for ant genomes were done by comparing homologous annotations between ants and *D. melanogaster* and integrating this information in three steps. First, genes in the ant genomes were compared by BLASTP with those in *D. melanogaster*. Second, gene-pair values with BLAST *E* values $< 1 \times 10^{-5}$ and query coverage $> 30\%$ were retained. Third, ant genes were annotated with Entrez Gene ID from their fruit fly homologues and linked with Entrez Gene ID values from the Gene Ontology and KEGG annotations.

To find gene sets with the most consistent caste differences in gene expression across ant species, we examined brain gene expression differences between gyne and worker samples across colony replicates in the five ant species with typical caste differentiation: *A. echinatior*, *S. invicta*, *M. pharaonis*, *L. niger* and *L. humile*. We omitted *O. biroi* and *D. quadriceps* from this analysis because they lack the queen caste and have secondarily evolved egg laying by parthenogenetic or inseminated workers (gamergates), respectively. We used the quantile-normalized expression values of the 6,672 orthologous genes (see the section 'Normalization of gene expression scores across samples' above) as input for gene-set enrichment analyses using the General Application Gene-set Enrichment (GAGE) package in R[38]. Colony-specific gyne samples were always marked as 'target' and worker samples as 'reference', and we fixed the 'compare' option as paired when comparing differences in gene expression between gyne–worker pairs for colony replicates. Enrichment analyses were done using KEGG and Gene Ontology annotations from corresponding homologues in the fruit fly. We fixed the 'same expression direction' option as true in both KEGG and Gene Ontology enrichment analyses, which allowed us to detect gene sets that are consistently up- or downregulated in gynes relative to workers. Gene sets were identified as being significantly enriched when FDRs were $< 1 \times 10^{-2}$ in two-sided non-parametric Kolmogorov–Smirnov tests.

Adjusting for species and colony identity to obtain a general assessment of caste-specific gene expression in ants. A transcriptome with a total number of n genes in any sample can be regarded as an n dimensional vector $\vec{\mathbf{X}}$, where $\vec{\mathbf{X}} = (X_1, X_2, \ldots, X_n)$ and X_i represents the expression level of gene i in a particular sample. This expression level X_i is then affected both by its species-specific background (for example, caused by species-specific developmental constraints, adaptations or environmental conditions) and its actual caste state, which should require the phenotypic expression of specific morphological and/or physiological traits. It is therefore useful to decompose vector $\vec{\mathbf{X}}$ as: $\vec{\mathbf{X}} = \vec{\mathbf{G}} + \vec{\mathbf{C}} + \vec{\mathbf{S}}$, where $\vec{\mathbf{G}}$ represents the influence of the lineage (species)-specific genetic background on gene expression, $\vec{\mathbf{C}}$ represents the influence of the actual caste state and $\vec{\mathbf{S}}$ represents any other effects on gene expression that are sample specific, such as environmentally induced colony-level differences or any technical biases that might have affected single samples. To compare $\vec{\mathbf{C}}$ across multiple species, it is then essential to subtract $\vec{\mathbf{G}}$ and $\vec{\mathbf{S}}$ from $\vec{\mathbf{X}}$. Assuming that the effect of colony is small compared with the effect of caste, and because samples of the same species were always prepared with the same experimental procedures, $\vec{\mathbf{S}}$ can be assumed to be the same within species and thus to group together with $\vec{\mathbf{G}}$ in a new vector $\vec{\mathbf{G}'}$. Thus, the task then becomes to obtain an estimate of $\vec{\mathbf{G}'}$ for each species that can be used to partial out any confounding effects on caste-biased gene expression. Because C_i, the influence of the caste phenotype on the expression of gene i in a particular sample, is either negative or positive and we expect a bimodal distribution of these values, we can assume that the distribution of C_i across samples of the same species is symmetrical around zero, also because we collected equal numbers of samples of different caste phenotypes. The sum of the C_i scores across multiple samples of the same ant species will then be 0 and the same applies to the sum of $\vec{\mathbf{C}}$. Then, the sum of $\vec{\mathbf{X}}$ across multiple samples of the same species can be expressed as: $\mathrm{sum}(\vec{\mathbf{X}}) = \mathrm{sum}(\vec{\mathbf{G}'}) + \mathrm{sum}(\vec{\mathbf{C}}) = \mathrm{sum}(\vec{\mathbf{G}'}) + 0$, so the $\mathrm{average}(\vec{\mathbf{X}}) = \vec{\mathbf{G}'}$. This implies that $\vec{\mathbf{C}} = \vec{\mathbf{X}} - \vec{\mathbf{G}'} = \vec{\mathbf{X}} - \mathrm{average}(\vec{\mathbf{X}})$, so that it is legitimate to compare $\vec{\mathbf{C}}$ across different species.

Based on this principle, we subtracted the mean expression levels within species for each gene to obtain quantile-normalized orthologous transcriptome data controlled for species-level differences in the overall levels of gene expression. This was done with the SVA package in R[57], using the Combat function by setting species identity as a batch covariate and setting the mean.only option as 'false' to also correct for differences in the total ranges of gene expression, which are affected by absolute expression-level difference across species. This normalization method thus removed all interspecies differences in gene expression profiles, so the mean and variance of expression levels for each gene became the same across species. The remaining residuals then reflected sample (colony-level) differences within species so that samples with similar residuals now represented similar caste-biased gene-expression values across species.

One of the key assumptions in the procedure outlined above is that the effect of the colony of origin on brain gene expression is small compared with the effect of caste phenotype. However, because comparisons in the two queenless ants were among plastic phenotypes (egg laying or not) within a single morphological caste rather than between two permanently differentiated castes, the colony-level variation was expected to be relatively more important. Thus, in these species, we subtracted mean colony-expression levels while also adjusting for unequal variation in gene expression for each gene, to obtain quantile-normalized orthologous transcriptome data controlled for colony-level differences in overall gene expression. However, this made us realize that comparison with the five other ant species that had both caste phenotypes should ideally be based on samples that had undergone the same normalization procedures. We therefore always present results after species- and colony-level normalization side by side. The colony-level normalizations normally explained higher proportions of the variance, but this was due to each colony only having one sample for pooled gynes and one for pooled workers, whereas species-level normalizations had five colony-level replicates. The higher percentages of explained variance after colony-level normalization therefore appear to mostly reflect a reduction in the error term and should therefore carry similar but not quite identical weight as the species-level normalized results. This is because species-level normalization also removes any technical noise across pooled samples of ten individuals, due, for example, to RNA extraction, PCR cycles and sequencing methodology, but variation of this kind should have remained very minor.

Constructing similarity matrices across samples, castes and species. Sample similarity matrices were calculated using $1 - r_s$ (Spearman's correlation coefficients) as a measure of the gene expression distance between samples after quantile normalization procedures. Overall caste-specific expression similarity matrices were thus constructed according to the principles outlined above, using all orthologous gene data and the hierarchical clustering (hclust) function in R.

ANOVA to examine variation along the first two principal component axes for caste-specific gene expression. We used PCA to evaluate the similarity of genome-wide caste-specific gene expression patterns across the different ant species. Because each PCA axis obtained is a composite of many gene-level expression values, we examined the distribution of these scores using two-way ANOVAs with subfamily identity and caste state as main factors and their interaction term, giving the model:

$$\text{Principal component value} \sim \text{subfamily identity} + \text{caste state} + \text{subfamily identity} \times \text{caste state} + \varepsilon$$

This analysis allowed us to estimate the proportion of variance in principal component scores that could be explained by subfamily identity—a component that should be zero after adjusting for species-level differences. The other main factor, caste state (morphologically differentiated gynes or workers) was also categorical, and the statistical interaction between these two main effects can thus be understood as a subfamily-specific overall effect on observed patterns of caste-biased gene expression.

Identifying caste-biased differentially expressed genes across the five ant species. Genes that are differentially expressed between gynes and workers across the five ant species were identified with a GLM in DESeq2 (ref. [29]). RNA read counts for each sample were imported, after which we estimated the library size for each sample within the same ant species to adjust for library size differences, and subsequently modelled the expression levels of the 6,672 orthologous genes as:

$$\text{Read count} \sim \text{species identity} + \text{caste state}$$

This assumes that the effect of species identity and caste state are additive, and will thus identify genes with consistent caste-biased expression across the five ant species with normal caste differentiation. We used a 1.5-fold-change difference in expression between castes as the null model, and genes were scored as significantly caste-biased if they passed an FDR of 0.01 (that is, 1.5 not overlapping with the 99% CI) in two-sided Wald tests.

Constructing the cross-species co-expression network. A cross-species co-expression network was constructed using all caste-specific transcriptomes for the 6,672 orthologous genes, normalized for colony-level variation. The co-expression level for each gene pair was then calculated based on the absolute values of Spearman's correlation coefficients among all samples for gynes and workers, and

the use of absolute values ensured that co-expression levels captured both positive and negative regulation.

Projection of caste-specific single-species transcriptomes on the principle component axes obtained for the five ant species with normal caste differentiation. After completing our analyses of caste-biased gene expression in the brains of the five ant species with normal queen–worker caste differentiation, we examined whether and to what extent the ant caste GRN might be shared with other social insect species by projecting brain published gene expression data of two queenless ants[22,27], the honeybee[41] and the paper wasp *P. canadensis*[27] onto the PCA axes obtained for our five focal ant species.

To validate this method, we re-ran the PCAs for different combinations of four of the five ants to test whether the principal component axes produced would be consistent enough to also separate caste-biased gene expression in the fifth ant species. We did so by extracting the first two eigenvectors of the (quantile normalized and adjusted for species identity) transcriptomes of the four reference ant species using singular-value decomposition, which produced the first two principal components that separated castes across the four reference species. We then projected the normalized transcriptomes of the fifth ant species onto the PCA by matrix multiplication with the extracted eigenvectors from the four reference species. This procedure is comparable to using a leave-one-out jack-knife resampling approach[58]. Because it produced satisfactory results, we applied this procedure throughout comparisons between the five ant species with normal caste differentiation and the two queenless ants, the honeybee and the paper wasp *P. canadensis*.

Code availability. The Python and R scripts used to process the data are available at https://github.com/StanQiu/ant_brain_comparative.

Data availability

RNA-Seq data have been deposited under BioProject accession number PRJNA427677 (https://www.ncbi.nlm.nih.gov/sra/SRP127971).

Received: 10 January 2018; Accepted: 6 September 2018;
Published online: 22 October 2018

References

1. Wheeler, W. M. The ant-colony as an organism. *J. Morphol.* **22**, 307–325 (1911).
2. Fisher, R. M., Cornwallis, C. K. & West, S. A. Group formation, relatedness, and the evolution of multicellularity. *Curr. Biol.* **23**, 1120–1125 (2013).
3. Boomsma, J. J. & Gawne, R. Superorganismality and caste differentiation as points of no return: how the major evolutionary transitions were lost in translation. *Biol. Rev.* **93**, 28–54 (2018).
4. Ward, P. S. Ants. *Curr. Biol.* **16**, R152–R155 (2006).
5. Gould, S. J. *Ontogeny and Phylogeny* (Harvard Univ. Press, Cambridge, 1977).
6. Wagner, G. P. *Homology, Genes, and Evolutionary Innovation* (Princeton Univ. Press, Princeton, 2014).
7. Simola, D. F. et al. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res.* **23**, 1235–1247 (2013).
8. 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).
9. Toth, A. L. & Robinson, G. E. Evo-devo and the evolution of social behavior. *Trends Genet.* **23**, 334–341 (2007).
10. Carroll, S. B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).
11. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nat. Rev. Genet.* **9**, 868–882 (2008).
12. Barchuk, A. R. et al. Molecular determinants of caste differentiation in the highly eusocial honeybee *Apis mellifera*. *BMC Dev. Biol.* **7**, 70 (2007).
13. Johnson, B. R. & Tsutsui, N. D. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC Genomics* **12**, 164 (2011).
14. 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).
15. Sumner, S. The importance of genomic novelty in social evolution. *Mol. Ecol.* **23**, 26–28 (2014).
16. Johnson, B. R. & Linksvayer, T. A. Deconstructing the superorganism: social physiology, groundplans, and sociogenomics. *Q. Rev. Biol.* **85**, 57–79 (2010).
17. Ding, Y., Zhou, Q. & Wang, W. Origins of new genes and evolution of their novel functions. *Annu. Rev. Ecol. Evol. Syst.* **43**, 345–363 (2012).
18. Chen, S., Krinsky, B. H. & Long, M. New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* **14**, 645–660 (2013).
19. 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).
20. Ward, P. S. The phylogeny and evolution of ants. *Annu. Rev. Ecol. Evol. Syst.* **24**, 2047–2052 (2014).
21. Mank, J. E. The transcriptional architecture of phenotypic dimorphism. *Nat. Ecol. Evol.* **1**, 0006 (2017).
22. Libbrecht, R., Oxley, P. R. & Keller, D. J. C. Robust DNA methylation in the clonal raider ant brain. *Curr. Biol.* **26**, 391–395 (2016).
23. Wagner, G. P. The developmental genetics of homology. *Nat. Rev. Genet.* **8**, 473–479 (2007).
24. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).
25. Roux, J., Rosikiewicz, M. & Robinson-Rechavi, M. What to compare and how: comparative transcriptomics for evo-devo. *J. Exp. Zool. B Mol. Dev. Evol.* **324**, 372–382 (2015).
26. 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**, 5058–5073 (2017).
27. Patalano, S. et al. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *Proc. Natl Acad. Sci. USA* **112**, 13970–13975 (2015).
28. Monnin, T., Ratnieks, F., Jones, G. R. & Beard, R. Pretender punishment induced by chemical signalling in a queenless ant. *Nature* **419**, 61–65 (2002).
29. 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).
30. Frumhoff, P. C. & Ward, P. S. Individual-level selection, colony-level selection, and the association between polygyny and worker monomorphism in ants. *Am. Nat.* **139**, 559–590 (1992).
31. Schwander, T., Rosset, H. & Chapuisat, M. Division of labour and worker size polymorphism in ant colonies: the impact of social and genetic factors. *Behav. Ecol. Sociobiol.* **59**, 215–221 (2005).
32. Trible, W. & Kronauer, D. J. C. Caste development and evolution in ants: it's all about size. *J. Exp. Biol.* **220**, 53–62 (2017).
33. Nygaard, S. et al. Reciprocal genomic evolution in the ant–fungus agricultural symbiosis. *Nat. Commun.* **7**, 12233 (2016).
34. Monnin, T. & Peeters, C. How many gamergates is an ant queen worth? *Naturwissenschaften* **95**, 109–116 (2007).
35. Cronin, A. L., Molet, M., Doums, C., Monnin, T. & Peeters, C. Recurrent evolution of dependent colony foundation across eusocial insects. *Annu. Rev. Entomol.* **58**, 37–55 (2013).
36. Heinze, J. The demise of the standard ant (Hymenoptera: Formicidae). *Myrmecol. News* **11**, 9–20 (2008).
37. Rabeling, C. & Kronauer, D. J. C. Thelytokous parthenogenesis in eusocial Hymenoptera. *Annu. Rev. Entomol.* **58**, 273–292 (2013).
38. Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinform* **10**, 161 (2009).
39. Peters, R. S. et al. Evolutionary history of the Hymenoptera. *Curr. Biol.* **27**, 1013–1018 (2017).
40. Corona, M. et al. Vitellogenin, juvenile hormone, insulin signaling, and queen honey bee longevity. *Proc. Natl Acad. Sci.* **104**, 7128–7133 (2007).
41. Vleurinck, C., Raub, S., Sturgill, D., Oliver, B. & Beye, M. Linking genes and brain development of honeybee workers: a whole-transcriptome approach. *PLoS ONE* **11**, e0157980 (2016).
42. LaPolla, J. S., Dlussky, G. M. & Perrichot, V. Ants and the fossil record. *Annu. Rev. Entomol.* **58**, 609–630 (2013).
43. Barden, P. & Grimaldi, D. A. Adaptive radiation in socially advanced stem-group ants from the Cretaceous. *Curr. Biol.* **26**, 515–521 (2016).
44. Peeters, C. in *The Evolution of Social Behaviour in Insects and Arachnids* (eds Crespi, B. J. & Choe, J. C.) 372–391 (Cambridge Univ. Press, Cambridge, 1997).
45. Girardie, J., Boureme, D., Couillaud, F., Tamarelle, M. & Girardie, A. Anti-juvenile effect of neuroparsin A, a neuroprotein isolated from locust corpora cardiaca. *Insect Biochem.* **17**, 977–983 (1987).
46. Toth, A. L. et al. Brain transcriptomic analysis in paper wasps identifies genes associated with behaviour across social insect lineages. *Proc. R. Soc. B* **277**, 2139–2148 (2010).
47. Mikheyev, A. S., Linksvayer, T. A. & Khaitovich, P. Genes associated with ant social behavior show distinct transcriptional and evolutionary patterns. *eLife* **4**, e04775 (2015).
48. Pantalacci, S. et al. Transcriptomic signatures shaped by cell proportions shed light on comparative developmental biology. *Genome Biol.* **18**, 29 (2017).
49. Romiguier, J. et al. Phylogenomics controlling for base compositional bias reveals a single origin of eusociality in corbiculate bees. *Mol. Biol. Evol.* **33**, 670–678 (2016).
50. Pontieri, L., Schmidt, A. M., Singh, R., Pedersen, J. S. & Linksvayer, T. A. Artificial selection on ant female caste ratio uncovers a link between female-biased sex ratios and infection by *Wolbachia* endosymbionts. *J. Evol. Biol.* **30**, 225–234 (2017).
51. Conesa, A. et al. A survey of best practices for RNA-Seq data analysis. *Genome Biol.* **17**, 13 (2016).

52. Keilwagen, J. et al. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* **44**, e89 (2016).
53. Lechner, M. et al. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinform.* **12**, 124 (2011).
54. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
55. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA* **107**, 9546–9551 (2010).
56. Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **107**, 9546–9551 (2003).
57. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161 (2014).
58. Efron, B. in *Breakthroughs in Statistics* (eds Kotz, S. & Johnson, N. L.) 569–593 (Springer, New York, 1992).
59. Moreau, C. S., Bell, C. D., Vila, R., Archibald, S. B. & Pierce, N. E. Phylogeny of the ants: diversification in the age of angiosperms. *Science* **312**, 101–104 (2006).
60. Ward, P. S., Brady, S. G., Fisher, B. L. & Schultz, T. R. The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae). *Syst. Entomol.* **40**, 61–81 (2015).

Acknowledgements

This work was supported by grants from the Lundbeck Foundation (R190-2014-2827 to G.Z.), Carlsberg Foundation (CF16-0663 to G.Z.), Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13000000 to G.Z.) and Biodiversity Research Center, Academia Sinica (100-2311-B-001-015-MY3, 103-2311-B-001-018-MY3 and 104-2314-B-001-009-MY5 to J.W.), as well as an Academia Sinica Career Development Grant (to J.W.) and an ERC Advanced Grant (323085 to J.J.B.). We thank C. Guo, H. Yu and Q. Li for coordination of the sequencing at BGI.

Author contributions

G.Z., J.J.B., J.W. and B.Q. designed the experiments. R.S.L., N.-C.C. and B.Q. reared and isolated the ant colonies in the laboratory. B.Q. and N.-C.C. collected the ants, dissected the ant brains and extracted RNA. B.Q. constructed the cDNA libraries. N.-C.C. and J.W. generated the transcriptome data for *S. invicta*. B.Q. analysed the data. B.Q., G.Z. and J.J.B. interpreted the data and wrote and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0689-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.J.B. or G.Z.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Guojie Zhang

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	We used GeMoMa (v1.4.2) to perform genome re-annotations, Proteinortho (v5.16) to identify orthologous genes across ant species, Hisat2 (v2.0.5), StringTie (v1.3.3) and Salmon (v0.8.2) to perform RNAseq alignment and quantification, DESeq2 (R package; v1.16) to identify differentially expressed genes, GAGE (R package; v2.27.3) to do gene set enrichment, and SVA (R package; v3.26.0) to adjust for species or colony identify for gene expression. For gene age determination, Singular Value Decomposition (SVD) and visualization, we wrote our codes with R and python. All codes used in this study are provided in Github: https://github.com/StanQiu/ant_brain_comparative

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-seq data have been deposited under Bioproject accession number PRJNA427677. The Python and R scripts used to process the data are available on https://github.com/StanQiu/ant_brain_comparative

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We decided sample size based on the detection power of RNAseq (both https://doi.org/10.1186/s13059-016-0881-8 and our computational simulation result). With minimum of four replicates, we had over 90% detection power for genes with 2-fold expression changes.
Data exclusions	We excluded 7 samples from our analyses . For the three sets of RNA samples with ERCC-spike-in, Pearson correlations of spike-in reads (log transformed) across samples of < 0.95 were considered as outliers and removed. Similarities in caste-biased gene expression were also used to check data quality: For each ant species, mapped reads for the genome (log transformed) were compared among samples within the same caste, and samples different from other same-caste samples of the focal species (within-caste Pearson correlations < 0.9) were checked for potential biological or technical deviations, as suggested by the ENCODE RNAseq guideline [goo.gl/d9DvyV]. For example, we excluded potential outlier samples if they were collected from different locations (biological explanation) or were prepared with different experimental procedures (technical explanation). If no potential biological or technical explanations could be identified, samples were retained for downstream analysis.
Replication	Our study initially only included four ant species: A.echinatior, L.humile, M.pharaonis and S.invicta, and we found that samples of the same caste can be separated with PCA after normalizing for species identity. We then further included recent published brain transcriptome data from L.niger, and we identified the same pattern. Therefore, we believe our study is reproducible.
Randomization	Gynes (future queens) and workers samples were collected from four to five different colonies. To minimize/randomize batch effect of caste and colony identity in preparation, RNA extraction and cDNA library construction were done with the same procedure and within the same run for each species, and sequencing were done with multiplexed bar-coding.
Blinding	The investigators were not blinded to phenotype (caste), colony and species information during collection and analysis procedures. But clustering analysis were done with blinded phenotype information.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☐	☒ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Ants are available in labs of Centre for Social evolution, University of Copenhagen or Biodiversity Research Center, Academia Sinica, Taoyuan, Taiwan.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Five colonies of A. echinatior were collected in Panama in 2004 (Ae263), 2014 (Ae704) and 2016 (Ae747, Ae764, Ae767) and reared at the Centre for Social Evolution, University of Copenhagen under a constant temperature of ca. 25ºC and ca. 70% humidity. Gynes and minor workers were collected within the fungus garden, whereas major workers were collected while they were foraging outside, always during daytime (10 am to 4 pm), and isolated in groups of 6-10 individuals of the same caste in small fluon-coated plastic boxes. Five colonies of M. pharaonis (3rd Room X1/B, CS10, Donor 3rd, Donor BQ-, 3rd Room X1/A) were reared at the Centre for Social Evolution, University of Copenhagen under constant temperature of 27ºC and 50% humidity from a stock collected in 2008. Gynes were separated from males at the pupal stage and reared with 10–15 workers in fluon-coated petri-dishes to be collected within three days after they hatched as adults, whereas workers were collected directly from the source colonies when they were foraging. All collections were done during daytime (10 am to 4 pm) and adults were kept in groups of 6-10 of the same caste in small fluon-coated boxes. Five colonies of L. humile were collected in Caldes d'Estrac, Spain (Catalan3b) and Castell d'Aro, Spain (Main 3a, Main 4a, Main 5b, Main 5d) in 2016 and were reared at the Centre for Social Evolution, University of Copenhagen under constant temperature of 27ºC and 50% humidity. Gyne isolation, worker ant sample collection, brain dissection, and storage proceeded in the same way as for Monomorium pharaonis. Four monogynous (single queen) colonies were collected in Taoyuan, Taiwan, two in October 2012 and two in April 2014, and transferred to a fluon-coated plastic box in the lab. Approximately 120 gynes and 200 workers from each colony were randomly selected for dissection. For workers, an equal number of small and large workers were selected. Ant sample collection and dissection procedures for L. niger, the clonal and queenless ant O. biroi, and the queenless ant D. quadriceps were comparable to our own procedures and described in original research papers.
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*