Ant phylogenomics reveals a natural selection hotspot preceding the origin of complex eusociality

Jonathan Romiguier, Marek Borowiec, Arthur Weyna, Quentin Helleu, Etienne Loire, Christine Mendola, Christian Rabeling, Brian Fisher, Philip Ward, Laurent Keller

Ant phylogenomics reveals a natural selection hotspot preceding the origin of complex eusociality

Jonathan Romiguier[1,2], Marek L. Borowiec[3], Arthur Weyna[1], Quentin Helleu[2], Etienne Loire[4], Christine La Mendola[2], Christian Rabeling[5], Brian L. Fisher[6], Philip S. Ward[7], Laurent Keller[2]

1. CNRS, University of Montpellier (UMR-5554 - ISEM), 34095, Montpellier, France.

2. Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

3. Department of Entomology, Plant Pathology and Nematology, University of Idaho, Moscow, ID, USA

4. ASTRE, Cirad, INRAE, University of Montpellier, Montpellier, France

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

6. Department of Entomology, California Academy of Sciences, San Francisco, CA 94118, USA

7. Department of Entomology and Nematology, University of California, Davis, CA 95616, USA

* Lead Contact and Corresponding Author: Jonathan Romiguier (jonathan.romiguier@umontpellier.fr)

Twitter: @SelfishMeme

Summary

The evolution of eusociality has allowed ants to become one of the most conspicuous and ecologically dominant groups of organisms in the world. A large majority of the current ~14,000 ant species belong to the formicoids[1], a clade of nine subfamilies which exhibit the most extreme forms of reproductive division of labour, large colony size[2], worker polymorphism[3] and extended queen longevity[4]. The eight remaining non-formicoid subfamilies are less well studied, with few genomes having been sequenced so far and unclear phylogenetic relationships[5]. By sequencing 65 genomes, we provide a robust phylogeny of the 17 ant subfamilies, retrieving high support to the controversial leptanillomorph clade (Leptanillinae and Martialinae) as the sister-group to all other extant ants. Moreover, our genomic analyses revealed that the emergence of the formicoids was accompanied by an elevated number of positive selection events. Importantly, the top three gene functions under selection are linked to key features of complex eusociality with histone acetylation being implicated in caste differentiation, gene silencing by RNA in worker sterility and autophagy in longevity. These results show that key pathways associated with eusociality have been under strong selection during the Cretaceous, suggesting that the molecular foundations of complex eusociality may have evolved rapidly in less than 20 Ma.

Results and discussion

A reference tree of ant subfamilies

To build a comprehensive phylogenetic tree including representatives of all extant ant subfamilies, we conducted two main types of analyses on the 4,300,911 amino acids from 4,151 single-copy protein-coding genes that we generated from 83 species (including 9 hymenopteran outgroups, and 65 newly sequenced genomes; see Material and Methods). First, we performed supermatrix approaches, where all genes were concatenated for estimating a single species tree. Second, we performed supertree approaches, where the species trees were estimated from all gene trees. The two best resulting trees from each approach are summarized in Figure 1. Each node of the tree is supported with maximal support by the supermatrix analyses, and nearly all nodes (78/82) are congruently supported by both the supermatrix and supertree approaches. Importantly, the deepest and most important nodes of the ant phylogeny, including every relationship among the 17 ant subfamilies are maximally supported by both the supermatrix and supertree approaches.

Our results confirm that the so-called poneroid subfamilies (Ponerinae, Paraponerinae, Agroecomyrmecinae, Proceratinae, Apomyrminae, Amblyoponinae) are monophyletic[6–9], rather than paraphyletic[10–12]. Both the supermatrix and supertree provide maximum support for the poneroid monophyly. The relationships among poneroid subfamilies are also all congruently supported by both approaches (Figure 1), including for the Paraponerinae (one living species) and Agroecomyrmecinae (two living species) which are inherently difficult to relate with other subfamilies as their deep phylogenetic divergence resulting in long branches. An analysis of ultraconserved elements (UCE) markers for an increased data set of 166 taxa (see STAR Methods) retrieved the same subfamily relationships (Figure S1), further supporting the view that our phylogeny (Figure 1) is robust.

These analyses are important because the rooting of the ant phylogeny has been a controversial issue since the discovery of *Martialis heureka*, an extremely rare and morphologically divergent ant species that was initially inferred as the sister-group of all other ants[10]. Some studies suggested that Leptanillinae is the subfamily sister to all other ants[11,13], while a recent study suggested that Leptanillinae and Martialinae may form a monophyletic group[9]. Our analyses support this last hypothesis, with Leptanillinae+Martialinae forming a clade (hereafter referred to as the leptanillomorph clade) that is the sister-group to all other ants (see Figure 1). This conclusion was also supported by two further analyses controlling for outgroup composition, which has been suggested to affect the rooting of the ant tree[9]. First, we built an alternative supermatrix of 2,343 genes (983,951 amino acids) containing many outgroups, with 115 non-ant aculeate species borrowed from published transcriptomes[14]. These analyses revealed that the rooting of the ant phylogeny and subfamily relationships are not affected by the inclusion of all these outgroups (see Figure S1B). Second, using random combinations of outgroups (see STAR Methods for details), we always found the same rooting, with strong support for the leptanillomorph clade being sister to all other ants (bootstrap values: 100 for 109 trees and 99 for the remaining 6 trees). Because all species of the leptanillomorph clade are pale, blind and have a similar hypogeic ecology, this suggests that some early ants may have escaped extinction by retreating to these stable, subterranean habitats before other lineages diversified by developing novel morphological and behavioral adaptations[10]. According to our divergence date estimates, the common ancestor of the leptanillomorphs lived around the Jurassic-Cretaceous boundary (~145 Ma) shortly after the common

ancestor of all extant ants (~150 Ma). This suggests the possibility that subterranean lifestyles existed in the ancestors of extant ants or, more likely, that a hypogeic lifestyle originated at an early stage in the history of leptanillomorphs. This result contrasts with the fossil evidence, because the earliest-known fossilized crown ants were not specialized to subterranean habitats and come from Burmese amber deposits that are approximately 99 million years old[15,16]. Set against our divergence date estimates, this indicates a gap in the ant fossil record of ~50 million years, further emphasizing an existing discrepancy between fossils and molecular data when it comes to the question of ant origins[17]. This is reminiscent of the debate on the origin of placental mammals, estimated be in the middle Cretaceous by molecular data, whereas there is no fossil record before the K-T crisis[18]. It has been suggested that the lack of fossils may stem from the occurrence of only a few lineages of placental mammals and perhaps small populations sizes during the Cretaceous[19]. Similarly, it is possible that the abundance of crown ants was low at first, and only sufficiently increased with the rise of angiosperm[6] to be represented in the fossil record. Alternatively, there may be methodological biases leading to overestimation of divergence dates[20,21] or incorrect phylogenetic placement of early ant fossils[16].

Pervasive positive selection associated with the origin of the socially diverse formicoid clade

To investigate the molecular changes associated with the evolution of complex eusociality, we conducted positive selection analyses on the 4,151 ortholog genes of the 75 ant genomes (see STAR Methods). The percentage of genes under positive selection varied greatly among the 38 branches ranging from 0 to 2.6% in a single branch (Figure 2). Strikingly, the branch leading to the formicoid clade stands out as a clear outlier with a 30-fold higher rate of positive selection compared to the average of other tree branches. There were 110 positively selected genes on the branch subtending the formicoid clade while the average number of genes with positive selections was only 3.1 in other branches (maximum value = 20 genes). This finding is particularly remarkable given that the genes considered in our analysis are highly conserved universal orthologs across Hymenoptera[22]. This indicates that extensive molecular changes in well-conserved core genes occurred along the branch giving rise to the formicoids. By contrast, there was no evidence of a further burst of positive selection later in the evolutionary history of the formicoids, including in the multiple branches leading to the most complex eusocial species (Figure 2). This suggests that most of the genetic innovations specific to complex eusociality in formicoids occurred in less than ~20 Ma during the early Cretaceous (see Figures 1 and 2).

Functional enrichment analyses for the formicoid branch revealed that histone acetylation was the most significantly over-represented function among the 110 positively selected genes. Histone acetylation is well-known for controlling transcriptional activity[23], reprograming foraging behaviour of the major worker caste into the minor worker caste[24], colony activity rhythms[25], and the longevity/fecundity trade-off in workers[26]. Histone acetylation is also involved in caste determination of honeybees through the effect of royal jelly[27] and has been identified as a key caste-specific enhancer of transcription regulating the differential larval development of queens and workers[28]. Interestingly, our analyses revealed positive selection on *histone acetyltransferase* (Data S1), a gene previously linked to functions potentially relevant to eusociality, such as the regulation of worker polymorphism[29]. The second most significant function was autophagy. Autophagy has repeatedly been shown to be essential for queen lifespan extension[30,31] and the caste-specific programmed cell death responsible for the divergent ovary development in queen and worker honeybees[32]. Finally, the third most significant function was gene silencing by

RNA (Figure 2). From our results, we retrieved the gene *Tudor-SN*, which is a candidate for controlling worker sterility in honeybees[33]. Altogether,these results reveal that the common ancestor of the formicoid ants underwent important genomic changes relative to the regulation of gene expression (e.g., histone acetylation and gene silencing RNA) and soma maintenance (e.g., autophagy).

These changes may have been important in allowing the evolution of extreme division of labour in formicoids, the ant clade comprising the vast majority of species exhibiting extreme forms of complex eusociality (e.g. maximum colony size of 3 million polymorphic workers in formicoid *Dorylus* species, compared to a maximum of 50,000 monomorphic workers in some poneroid species of the genus *Leptogenys*[34]). However, given that the nine formicoid subfamilies also display some species with less complex levels of eusociality, this implies that while the genomic changes that occurred during the early Cretaceous may have favoured the emergence of extreme division of labour and more overtly complex forms of eusociality, they did not necessarily lead to such changes in social organisation. Knowing which selective pressures triggered these dramatic molecular changes remains an open and intriguing question.

Conclusions

By providing genome-wide data for all ant subfamilies, this study infers the leptanillomorph clade as the sister-clade to all other extant ants and clarifies controversial subfamily relationships which will be important for further comparative studies in ants. The comparative genome analysis also reveals important changes in key molecular pathways implicated in the differential gene expression of queens and workers. This burst of molecular innovations, which occurred over ~20 million years in the early Cretaceous, possibly played an important role in facilitating the evolution of complex eusociality including the large colony sizes, extensive caste polymorphism and extreme fecundity/longevity of queens which characterize multiple lineages of formicoids.

Acknowledgements

We thank the FEBS (Federation of European Biochemical Societies) for the long-term fellowship granted to Jonathan Romiguier and funding from the French ANR (t-ERC grant RoyalMess) and ERC (RoyalMess, grant 948688), ERC and Swiss NSF to Laurent Keller, a grant from US National Science Foundation (CAREER DEB-1943626) to Christian Rabeling, and three reviewers for very useful comments. We thank the following for helping us to collect samples: C.S. Moreau, F.A. Esteves, S. Van Noort, N. Gunawardene, C. Poteaux, T. Bruscht, J.M. Gómez Durán, M. Molet, E. Vargo, Bui Tuan Viet, M.D. Goodisman, T. Delsinne and J. Orivel.

Author contributions

JR and LK conceived the study, JR, MB, CR, BLF, PSW and LK coordinated sample collection efforts, CLM performed DNA extractions, JR, MB, QH, AW and EL performed analyses, JR, MB, CR, BLF, PSW and LK wrote the manuscript.

Declaration of interests

The authors declare no competing interests.

Figures

Figure 1: Phylogeny and timeline of ant evolution based on whole-genome data. The topology is inferred according to the main supermatrix analysis (1,692,052 amino acid sites after cleaning, maximum likelihood, IQ-TREE PMSF C20 profile mixture model, Figure S1D), ultrafast bootstrap support is displayed first when node support is not maximal. Node support of the main supertree analyses (from gene trees of the 1,552 alignments of more than 500 amino acid sites, gene trees inferred with model search in IQ-TREE, species tree with ASTRAL, Figure S1E) are displayed second when not maximal (dash when node shows incongruence with the supermatrix analysis). Time divergence has been estimated using chronos with 12 calibration nodes (Table S2). Ant images from Alex Wild, used with permission. See also Figure S1, Table S1 and Table S2.

Figure 2: Branch subtending socially diverse formicoid ants shows increased rates of positive selection. Colours indicate the percentage of genes significantly under positive selection (aBSREL analysis) on each branch. Colony size and polymorphism data represent the maximum observed value in the genus and have been extracted from the literature[34]. Significantly enriched functional categories (Biological Process) under positive selection are represented as a word cloud for the formicoid branch. The size of the font is proportional to the p-value (Fisher's exact test, larger font indicating the most significant ones, from 0.011 to 0.047). Darker colours indicate the three functions with the highest numbers of significant annotated genes under positive selection. See also Data S1.

STAR Methods

Resource availability

Lead contact

Further information and requests for resources be directed to and will be fulfilled by the lead contact, Jonathan Romiguier (jonathan.romiguier@umontpellier.fr).

Material availability

This study did not generate new unique reagents.

Data and code availability

- Raw reads of sequenced genomes data have been deposited at ENA (European Nucleotide Archive). Accession numbers are listed in the key resources table. Genome assemblies, data, raw results and command lines for reproducibility are available in the following Zenodo repository and are publicly available as of the date of publication https://doi.org/10.5281/zenodo.5705739. DOIs are listed in the key resource table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Experimental model and subject details

We sampled 65 species (64 ants, 1 jewel wasp *Ampulex compressa*) across all ant subfamilies. We collected specimen from various sources and collectors (details and full overview of samples available in Table S1).

Method details

Sequencing and genome assembly

DNA extractions have been performed based on a high salt method[63] on samples conserved in ethanol or in -80 °C freezers.
We sequenced the genomes of 65 samples (64 ants and the jewel wasp *Ampulex compressa*) using Illumina Hiseq technology (paired-end, 150 bp reads). Reads were cleaned using *Trimmomatic*[35] and first assembled using *AbySS* 2.0.2[36] with kmer size set on 61 or the optimal value as estimated by KmerGenie v1.7016[37] . We removed potential contaminations using the blobtools pipeline[38], with the exclusion of contigs that blasted on any non-arthropod phylum against the Genebank database (*nt*). Reads mapping on these contaminant contigs were filtered and remaining reads were used for a second genome assembly using SPAdes 3.9.0[39]. For each species, we selected the best assembly between *SPAdes* and *AbySS* based on the number of complete and single copy ortholog genes using BUSCO[22] with a 4,415 Hymenoptera ortholog dataset from OrthoDB v9[40] (see Table S1 for assembler, BUSCO scores and N50 of each species).

Phylogeny

We built our phylogenomic dataset by complementing the 4,415 ortholog genes of our 65 species with 17 (9 ants and 8 hymenopteran outgroups) supplemental reference genomes from OrthoDB v9[40]. We aligned the amino-acid sequences using mafft[42] and cleaned the resulting alignments with Spruceup[43] and *trimal*[64] with the “automated1” option.

We concatenated the alignments in a supermatrix that we analysed with two different substitution models with IQ-TREE v 2.0.5[45] and 1,000 ultrafast bootstraps[65]. First, we performed a gene partitioned analysis (all partitions share the same set of branch lengths but have their own evolution rate) with LG+F+G4 models after having removed partitions that failed at symmetry tests testing stationarity and homogeneity assumptions[66] (resulting tree is presented in Figure S1C and has been used as the *guide tree* for the next analysis).

Second, because the most controversial ant subfamily relationships are expected to be affected by long-branch-attraction artefacts (Martialinae and Paraponerinae are monotypic subfamilies with long branches), we used the posterior mean site frequency model (PMSF model, LG+C20+F+G) which has been designed to correct for such artefacts by modeling site heterogeneity[67,68]. The resulting tree is presented in Figure 1 + Figure S1D and is referred in the results as the main supermatrix analysis. We also used this tree for estimating divergence times via penalised maximum likelihood approaches[69] and a set of 12 node calibrations (details of calibrations and references in Table S2).

A coalescent-based species tree analysis was performed by first producing gene trees using IQ-TREE[45] with the substitution model selected for each alignment by the built-in ModelFinder option MFP+MERGE. Because coalescent-based species tree approaches are sensitive to inaccurate gene trees[70], we only kept gene trees from long alignments of more than 500 amino-acids (n=1,366) and used ASTRAL 5.7.4[46] for producing the supertree (Figure S1E).

In case outgroup composition affected our results[9], we built an alternative dataset containing 115 outgroups of Aculeate species by matching our OrthoDB IDs with the IDs provided by the alignments of an Hymenoptera phylogenomic dataset[14], resulting in a 2,343 gene supermatrix (983,951 sites) after an automated *trimal* cleaning. We first performed the same PMSF analysis as described above (see tree in Figure S1B). Second, we produced a reduced supermatrix by keeping only alignments with at least 90% of species and sites with 90% of non-ambiguous characters, resulting in a supermatrix with 271,959 sites. We then removed from 0 to 115 random outgroups from this supermatrix, producing 116 new supermatrices for testing the effect of random outgroup removal. The same PMSF tree inference as described above was then performed, resulting in 116 phylogenetic trees (available in the Zenodo repository).

We built an alternative dataset of ultra-conserved elements loci (UCE) from our genomes and merged the data with the phylogenetic dataset of Branstetter et al.[8]. We used phyluce[71] to extract UCE loci from our genome assemblies. 2,510 loci were retrieved but we only kept the 1,855 that were common with the Branstetter et al.[8] dataset. We aligned the data using mafft[42], cleaned the alignment using trimal[64] with the *-automated1* option and removed alignments that contained fawer than 75% of the total number of taxa (n = 166). We concatenated the 1,230 remaining alignments in a supermatrix with 426,015 sites and inferred a tree using IQ-TREE with a locus partitioned analysis (all partitions share the same set of branch lengths but have their own evolution rate) with GTR+I+G4 models and 1000 ultrafast bootstraps (Figure S1A).

Divergence dating analyses

First, we used a simple approach exploiting the largest supermatrix by using the topology and branch lengths of Figure S1D (supermatrix of 1,692,052 amino acids, PMSF model LG+C20+F+G) to estimate divergence times via penalised maximum likelihood approaches[69] and a set of 12 node calibrations (details of calibrations and references in Table S2). To confirm the retrieved estimations (presented in Figure 1), we analysed a reduced dataset using a Bayesian approach, as implemented in MCMCTree, a part of the PAML package, v4.10[72]. MCMCTree utilizes rapid approximate likelihood computation[73], which makes it suitable for divergence dating of genome-scale data sets[18]. Due to computational constraints, we used an alignment with loci containing a minimum of 95% of our 83 taxa, totalling 182,809 amino acid sites. We fixed the topology to be the same as our analysis of the full alignment. We constrained our root node with a soft bound

maximum age of 236 Ma, corresponding to the lower bound of the 95% highest posterior density (HPD) interval for that split in Hymenoptera tree estimations[14]. We also set soft bounds on the root of the Formicidae to be 103 Ma and 169 Ma, corresponding to the upper 95% bound of HPD in Borowiec et al.[9] and lower bound in Economo et al.[74], the most divergent of recent estimates for the crown age of the family[17]. We also used minimum node age constraints based on fossils presented in Table S2. We ran each analysis unpartitioned, under the LG model for 5 million generations. We examined each run's statistics in Tracer[75] and confirmed convergence and sufficient effective sample sizes (>>200) for all parameters. Retrieved estimations were close to those retrieved with penalised likelihood on the whole dataset (Figure 1) and are available with all output files in the zenodo repository (https://doi.org/10.5281/zenodo.5705739).

Positive selection analysis

We performed a positive selection detection analysis by using 4,415 nucleotide alignments. Nucleotide alignments have been produced and refined from amino-acid alignments using the command *reportGapsAA2NT* and *refineAlignment* from macse v1.2[47]. We cleaned the alignments of potential errors further by using *hmmcleaner v1.8,* a tool that has been reported as especially effective for reducing false positives for detection of positive selection[48]. We used the value of 5 for the threshold parameter then removed every species with fewer than 20% nucleotides remaining after the cleaning. To ensure that gap-rich regions did not bias our analyses, we applied two different supplemental cleaning treatments by keeping only codons shared with more than 50 and 75% of the species of the alignment. All of the following analyses were performed with the three cleaning strategies (hmmcleaner only; hmmcleaner+50% complete codons; hmmcleaner+75% complete codons) and retrieved consistent results regarding the relatively high percentage of genes under positive selection in the Formicoid branch compared to other branches (38.13, 36.26 and 30.58-fold increases, respectively, see Data S1). Only the results of the "hmmcleaner+more than 75% complete codons" treatment are presented in the main text.

We used the aBSREL method (adaptive Branch-Site Random Effects Likelihood) from the HyPhy package[76], an improved implementation of the branch-site model typically used to test whether positive selection has occurred on some branches via the estimation of dN/dS (non-synonymous substitution rate over synonymous substitution rate)[77]. All branches were tested for positive selection for each gene, with p-values corrected for multiple testing on multiple branches (using the built-in correction in aBSREL.An additional correction was conducted for multiple testing on multiple genes[78]. For each internal branch in Figure 2, we only considered alignments containing at least one species for each of its three connected clades, ensuring that the positive selection test reflects this exact part of the evolutionary history of the gene. For each gene, we reported the gene ontology function of the *Apis mellifera* ortholog gene as available in OrthoDB[40]. Gene ontology enrichment analyses have been performed using topGO[50] with Fisher exact tests and the default weight01 algorithm.

GC-content variations are known to potentially bias detection of positive selection methods via the process of biased gene conversion[79,80]. Particularly for our results, strong GC-content variations among Formicoid and Poneroid species could lead to an overestimation of positive selection in the branch leading to Formicoids. To ensure that it is not the case, we measured the average GC-content of all species and retrieved no significant difference between Formicoids and Poneroids when considering all genes (45.43% vs 46.14% in Formicoids and Poneroids, p-value = 0.14 from Welch two sample test) or only considering

the 110 genes retrieved as positively selected from the main analysis (48.10% vs 48.92% in Formicoids and Poneroids, p-value = 0.15 from Welch two sample test).

Gene predictions and gene family analyses

We performed gene predictions for our 65 genomes by using MAKER2 v 2.31.8[51] , a pipeline for genome annotation using ncbi-blast v 2..2.28[52] , RepeatMasker v 4.0.5[53] , exonerate v 2.2.0[54] , snap v 2013.11.29[55], augustus v 3.2.2[56], tRNAscan-SE v1.3.1 and snoscan 0.9[57]. We filtered genes with fewer than 1000 nucleotides and individual protein sets were blasted against each other as well as against 28 additional Hymenoptera protein sets (detailed list available in output files available on 1https://doi.org/10.5281/zenodo.5705739) using orthofinder v 2.2.1[58] . The resulting gene count data file was then used for a gene family evolution analysis with CAFE v5[59] with base model default setting values. After trying several filtering methods, we removed gene families with difference in gene number larger than 50 to prevent "-inf" likelihood scores.

We assigned GO terms for our 65 protein sets using eggNOG v5[60] and used it to assign GO terms to gene families analysed with CAFE. To identify gene functions over-represented in gene families that underwent significant expansion/contraction, GO term enrichment analyses were performed using topGO[50], with Fisher exact tests and the default weight01 algorithm. Gene functions significantly over-represented and potentially related to eusociality include *autophagy*, *determination of adult lifespan, oogenesis, detection of chemical stimulus involved in sensory perception of smell, maintenance of chromatin silencing, olfactory receptor activity, histone deacetylase binding, histone kinase activity* (see the topGOresults tables in the Zenodo repository for the whole list). However, these results should be taken with caution because our analyses showed that the genomes available in public databases tended to exhibit greater increases/decreases of gene families than our 65 genomes. This is probably due to the fact that we had to use short-read sequencing technologies to be able to analyse low amounts of DNA and/or degraded DNA for some species that are rare and very difficult to collect. We therefore chose not to present these results in the main text but instead made them available in a Zenodo repository (https://doi.org/10.5281/zenodo.5705739) where we provide MAKER2 control file and options, resulting protein fasta files, orthofinder main output files, CAFE5 input and output files, eggNOG annotations and topGO analysis input/output.

Supplemental information titles and legends

Table S1: Detailed informations for each species included in this study, Related to Figure 1.

Table S2: Fossil calibrations, Related to Figure 1.

Data S1: Positive selection analyses details, Related to Figure 2. A) Branch ID correspondances. **B)** Number of genes under positive selection and average dN/dS for each branch and cleaning strategy. **C)** Positive selection test results for the formicoid branch (branch id 25) and GO-terms. **D)** Description of positively selected genes in the formicoid branch (branch id 25). **E)** Significant GO-terms for the formicoid branch

(Biological Process). **F)** Significant GO-terms for the formicoid branch (Molecular Function). **G)** Significant GO-terms for the formicoid branch (Cellular componment).

References

1. Bolton, B. (2021). An online catalog of the ants of the world. https://antcat.org.

2. Dornhaus, A., Powell, S., and Bengston, S. (2012). Group size and its effects on collective organization. Annu. Rev. Entomol. *57*, 123–41.

3. Oster, G.F., and Wilson, E.O. (1978). Caste and ecology in the social insects (Princeton University Press).

4. Genoud, M. (1997). Extraordinary lifespans in ants: a test of evolutionary theories of ageing. Nature *389*, 3–5.

5. Ward, P.S. (2014). The phylogeny and evolution of ants. Annu. Rev. Ecol. Evol. Syst. *45*, 23–43.

6. Moreau, C.S., Bell, C.D., Vila, R., Archibald, S.B., and Pierce, N.E. (2006). Phylogeny of the ants: diversification in the age of angiosperms. Science *312*, 101–104.

7. Ward, P.S., and Fisher, B.L. (2016). Tales of dracula ants: the evolutionary history of the ant subfamily Amblyoponinae (Hymenoptera: Formicidae). Syst. Entomol. *41*, 683–693.

8. Branstetter, M.G., Longino, J.T., Ward, P.S., and Faircloth, B.C. (2017). Enriching the ant tree of life: enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. Methods Ecol. Evol. *8*, 768–776.

9. Borowiec, M.L., Rabeling, C., Brady, S.G., Fisher, B.L., Schultz, T.R., and Ward, P.S. (2019). Compositional heterogeneity and outgroup choice influence the internal phylogeny of the ants. Mol. Phylogenet. Evol. *134*, 111–121.

10. Rabeling, C., Brown, J.M., and Verhaagh, M. (2008). Newly discovered sister lineage sheds light on early ant evolution. Proc. Natl. Acad. Sci. *105*, 14913–14917.

11. Kück, P., Hita Garcia, F., Misof, B., and Meusemann, K. (2011). Improved phylogenetic analyses corroborate a plausible position of *Martialis heureka* in the ant tree of life. PLoS ONE *6*, e21031.

12. Brady, S.G., Schultz, T.R., Fisher, B.L., and Ward, P.S. (2006). Evaluating alternative hypotheses for the early evolution and diversification of ants. Proc Natl Acad Sci U A *103*, 18172–18177.

13. Moreau, C.S., and Bell, C.D. (2013). Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. Evolution *67*, 2240–2257.

14. Peters, R.S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., et al. (2017). Evolutionary history of the Hymenoptera. Curr. Biol. *27*, 1013–1018.

15. Barden, P., and Grimaldi, D.A. (2016). Adaptive radiation in socially advanced stem-group ants from the Cretaceous. Curr. Biol. *26*, 515–521.

16. Boudinot, B.E., Khouri, Z., Richter, A., Griebenow, Z.H., Kamp, T. van de, Perrichot, V., and Barden, P. (2022). Evolution and systematics of the Aculeata and kin (Hymenoptera), with emphasis on the ants (Formicoidea: †@@@idae fam. nov., Formicidae). bioRxiv, 2022.02.20.480183.

17. Borowiec, M.L., Moreau, C.S., and Rabeling, C. (2020). Ants: Phylogeny and classification. In Encyclopedia of social insects, C. K. Starr, ed. (Springer International Publishing), pp. 1–18.

18. dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J., Donoghue, P.C.J., and Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc. Biol. Sci. *279*, 3491–3500.

19. Cooper, A., and Fortey, R. (1998). Evolutionary explosions and the phylogenetic fuse. Trends Ecol. Evol. *13*, 151–156.

20. Springer, M.S., Foley, N.M., Brady, P.L., Gatesy, J., and Murphy, W.J. (2019). Evolutionary models for the diversification of placental mammals across the KPg boundary. Front. Genet. *10*.

21. Bromham, L., Duchêne, S., Hua, X., Ritchie, A.M., Duchêne, D.A., and Ho, S.Y.W. (2018). Bayesian molecular dating: opening up the black box. Biol. Rev. *93*, 1165–1191.

22. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: assessing genome assembly and annotation completeness. In Gene prediction (Springer), pp. 227–245.

23. Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. Genes Dev. *12*, 599–606.

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

25. Libbrecht, R., Nadrau, D., and Foitzik, S. (2020). A role of histone acetylation in the regulation of circadian rhythm in ants. iScience *23*, 100846.

26. Choppin, M., Feldmeyer, B., and Foitzik, S. (2021). Histone acetylation regulates the expression of genes involved in worker reproduction and lifespan in the ant *Temnothorax rugatulus*. Authorea Prepr.

27. Spannhoff, A., Kim, Y.K., Raynal, N.J.-M., Gharibyan, V., Su, M.-B., Zhou, Y.-Y., Li, J., Castellano, S., Sbardella, G., Issa, J.-P.J., et al. (2011). Histone deacetylase inhibitor activity in royal jelly might facilitate caste switching in bees. EMBO Rep. *12*, 238–243.

28. Wojciechowski, M., Lowe, R., Maleszka, J., Conn, D., Maleszka, R., and Hurd, P.J. (2018). Phenotypically distinct female castes in honey bees are defined by alternative chromatin states during larval development. Genome Res. *28*, 1532–1542.

29. Alvarado, S., Rajakumar, R., Abouheif, E., and Szyf, M. (2015). Epigenetic variation in the Egfr gene generates quantitative variation in a complex trait in ants. Nat. Commun. *6*, 6513.

30. Madeo, F., Zimmermann, A., Maiuri, M.C., and Kroemer, G. (2015). Essential role for autophagy in life span extension. J. Clin. Invest. *125*, 85–93.

31. Nakamura, S., and Yoshimori, T. (2018). Autophagy and Longevity. Mol. Cells *41*, 65–72.

32. Barchuk, A.R., Cristino, A.S., Kucharski, R., Costa, L.F., Simões, Z.L., and Maleszka, R. (2007). Molecular determinants of caste differentiation in the highly eusocial honeybee *Apis mellifera*. BMC Dev. Biol. *7*, 70.

33. Hartfelder, K., Tiberio, G.J., Lago, D.C., Dallacqua, R.P., and Bitondi, M.M.G. (2018). The ovary and its genes—developmental processes underlying the establishment and function of a highly divergent reproductive system in the female castes of the honey bee, *Apis mellifera*. Apidologie *49*, 49–70.

34. Blanchard, B.D., and Moreau, C.S. (2017). Defensive traits exhibit an evolutionary trade-off and drive diversification in ants. Evolution *71*, 315–328.

35. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

36. Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., and Warren, R.L. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. *27*, 768–777.

37. Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. Bioinformatics *30*, 31–37.

38. Laetsch, D.R., and Blaxter, M.L. (2017). BlobTools: Interrogation of genome assemblies. F1000Research *6*, 1287.

39. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., and Prjibelski, A.D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.

40. Zdobnov, E.M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R.M., Simao, F.A., Ioannidis, P., Seppey, M., Loetscher, A., and Kriventseva, E.V. (2017). OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. *45*, D744–D749.

41. Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V., and Zdobnov, E.M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. *35*, 543–548.

42. Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. *33*, 511–518.

43. Borowiec, M.L. (2019). Spruceup: fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments. J. Open Source Softw. *4*, 1635.

44. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

45. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

46. Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics *19*, 15–30.

47. Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E.J.P. (2011). MACSE: Multiple alignment of coding sequences accounting for frameshifts and stop codons. PLoS ONE *6*, e22594.

48. Di Franco, A., Poujol, R., Baurain, D., and Philippe, H. (2019). Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. BMC Evol. Biol. *19*, 21.

49. Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S.L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol. Biol. Evol. *32*, 1342–1353.

50. Alexa, A., and Rahnenführer, J. (2009). Gene set enrichment analysis with topGO. Bioconductor Improv *27*, 1–26.

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

52. Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezhuk, Y., et al. (2013). BLAST: a more efficient report with usability improvements. Nucleic Acids Res. *41*, W29–W33.

53. Chen, N. (2004). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr. Protoc. Bioinforma. *5*, 4.10.1-4.10.14.

54. Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

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

56. Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. *33*, W465–W467.

57. Schattner, P., Brooks, A.N., and Lowe, T.M. (2005). The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. *33*, W686–W689.

58. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

59. Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. (2020). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics *36*, 5516–5518.

60. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG 5.0: a hierarchical,

functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. *47*, D309–D314.

61. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

62. Yu, G. (2020). Using ggtree to visualize data on tree-like structures. Curr. Protoc. Bioinforma. *69*, e96.

63. Aljanabi, S.M., and Martinez, I. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Res. *25*, 4692–4693.

64. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

65. Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. *35*, 518–522.

66. Naser-Khdour, S., Minh, B.Q., Zhang, W., Stone, E.A., and Lanfear, R. (2019). The prevalence and impact of model violations in phylogenetic analysis. Genome Biol. Evol. *11*, 3341–3352.

67. Wang, H.-C., Minh, B.Q., Susko, E., and Roger, A.J. (2018). Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. *67*, 216–235.

68. Tihelka, E., Cai, C., Giacomelli, M., Lozano-Fernandez, J., Rota-Stabelli, O., Huang, D., Engel, M.S., Donoghue, P.C.J., and Pisani, D. (2021). The evolution of insect biodiversity. Curr. Biol. *31*, R1299–R1311.

69. Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528.

70. Gatesy, J., and Springer, M.S. (2014). Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. *80*, 231–266.

71. Faircloth, B.C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics *32*, 786–788.

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

73. dos Reis, M., and Yang, Z. (2011). Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol. Biol. Evol. *28*, 2161–2172.

74. Economo, E.P., Narula, N., Friedman, N.R., Weiser, M.D., and Guénard, B. (2018). Macroecology and macroevolution of the latitudinal diversity gradient in ants. Nat. Commun. *9*, 1778.

75. Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst. Biol. *67*, 901–904.

76. Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S.L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol. Biol. Evol. *32*, 1342–1353.

77. Yang, Z., and Dos Reis, M. (2010). Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol. *28*, 1217–1228.

78. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

79. Galtier, N., and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. *23*, 273–277.

80. Romiguier, J., and Roux, C. (2017). Analytical biases associated with GC-content in molecular evolution. Front. Genet. *8*, 16.