Social complexity, life-history and lineage influence the molecular basis of caste in a major transition in evolution

Christopher Wyatt (✉ cw13722@gmail.com)

University College London

Michael Bentley

University College London

Daisy Taylor

University College London

Emeline Favreau

University College London <https://orcid.org/0000-0002-0713-500X>

Ryan Brock

University of East Anglia <https://orcid.org/0000-0003-2977-1370>

Benjamin Taylor

University College London

Emily Bell

School of Biological Sciences, University of Bristol <https://orcid.org/0000-0002-7309-5456>

Ellouise Leadbeater

Royal Holloway University of London <https://orcid.org/0000-0002-4029-7254>

Seirian Sumner

UCL <https://orcid.org/0000-0003-0213-2018>

Article

Keywords: Superorganismality, Castes, Wasps, Transcriptomics

Posted Date: August 27th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-835604/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Social complexity, life-history and lineage influence the molecular basis of caste in a major transition in evolution

Christopher D. R. Wyatt^{1*}, Michael Bentley^{1,†}, Daisy Taylor^{1,2,†}, Emeline Favreau¹, Ryan E. Brock^{2,3}, Benjamin A. Taylor¹, Emily Bell², Ellouise Leadbeater⁴ & Seirian Sumner^{1*}

¹ Centre for Biodiversity and Environment Research, University College London, London, UK.

² School of Biological Sciences, University of Bristol, United Kingdom, BS8 1TQ.

³ School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk, NR4 7TJ, UK

⁴ Department of Biological Sciences, Royal Holloway University of London, Egham, UK.

* Corresponding authors

† Equal contribution

Christopher Douglas Robert Wyatt

Centre for Biodiversity and Environment Research,
University College London, London, UK.

E-mail: c.wyatt@ucl.ac.uk; Phone +447307390637

Seirian Sumner

Centre for Biodiversity and Environment Research,
University College London, London, UK.

E-mail: s.sumner@ucl.ac.uk

Key words: Superorganismality, Castes, Wasps, Transcriptomics

Abstract

Major evolutionary transitions describe how biological complexity arises; e.g. in the evolution of complex multicellular bodies, and superorganismal insect societies. Such transitions involve the evolution of division of labour, e.g. among queen and worker castes in insect societies. A key mechanistic hypothesis for the evolution of division of labour is that a shared set of genes co-opted from a common solitary ancestral ground plan - a so-called genetic toolkit for sociality - regulates insect castes across different levels of social complexity. The vespid wasps represent an excellent system in which to test this. Here, using conventional and machine learning analyses of brain transcriptome data from nine species of vespid wasps, we find evidence of a shared genetic toolkit across species representing different levels of social complexity, with a large suite of genes classifying castes correctly across species. However, we also found evidence of additional fine-scale differences in predictive gene sets, functional enrichment and rates of gene evolution that were related to level of social complexity, but also life-history traits (e.g. mode of colony founding). Thus, there appear to be shifts in the gene networks regulating social behaviour and rates of gene evolution that are influenced by innovations in both social complexity and life-history. These results suggest that the concept of a shared genetic toolkit for sociality may be too simplistic to fully describe the process of the major transition to sociality, even within a single lineage. Diversity in lineage, social complexity and life-history traits must be taken into account in the quest to uncover the molecular bases of the major transition to sociality.

Main

The major evolutionary transitions span all levels of biological organisation, facilitating the evolution of life's complexity on earth via cooperation between single entities (e.g. genes in a genome, cells in a multicellular body, insects in a colony), generating fitness benefits beyond those attainable by a comparable number of isolated individuals¹. The evolution of sociality is one of the major transitions and is of general relevance across many levels of biological organisation from genes assembled into genomes, single-cells into multi-cellular entities, and insects cooperating in superorganismal societies. The best-studied examples of sociality are in the hymenopteran insects (bees, wasps and ants) - a group of over 17,000 species, exhibiting different forms of sociality across the transition from simple sociality (with small societies where all group members are able to reproduce and switch roles in response to opportunity), through to complex societies (consisting of thousands of individuals, each committed during development to a specific cooperative role and working for a shared reproductive outcome within the higher-level 'individual' of the colony, known as the 'superorganism'²). Recent analyses of the molecular mechanisms of insect sociality have revealed how conserved suites of genes, networks and functions are shared among independent evolutionary events of insect superorganismality³⁻⁸. An outstanding question is whether there are fine-scale differences in the genetic mechanisms operating at different levels of complexity across the spectrum of the major transition – from the simplest to most complex societies.

A key step in the evolution of sociality is the emergence of a reproductive division of labour, where some individuals commit to reproductive or non-

76 reproductive roles, known as queens and workers respectively in the case of insect
77 societies. An overarching mechanistic hypothesis for social evolution is that the
78 repertoire of behaviours typically exhibited in the life cycle of the solitary ancestor of
79 social species were uncoupled to produce a division of labour among group
80 members with individuals specialising in either the reproductive ('queen') or
81 provisioning ('worker') phases of the solitary ancestor⁹. Such phenotypic decoupling
82 implies that there will be a conserved mechanistic toolkit that regulates queen and
83 worker phenotypes in species representing different levels of social complexity
84 across the spectrum of the major transition (reviewed in¹⁰). An alternative to the
85 shared toolkit hypothesis is that the molecular processes regulating social
86 behaviours in non-superorganismal societies (where caste remains flexible, and
87 selection acts primarily on individuals) differ fundamentally from those processes that
88 regulate social behaviours in superorganismal societies^{11,12}. Phenotypic innovations
89 across the animal kingdom have been linked to genomic evolution: taxonomically-
90 restricted genes^{13–16}, rapid evolution of proteins^{17,18} and regulatory elements^{17,19}
91 been found in most lineages of social insects²⁰. Indeed, some recent studies have
92 suggested that the processes regulating different levels of social complexity may be
93 different^{8,17,19,21}. The innovations in social complexity and the shift in the unit of
94 selection (from individual- to group-level²²) that occur in a major transition may
95 therefore be accompanied by shifts in genomic processes and evolution, throwing
96 into question whether the concept of a universal conserved genomic toolkit that
97 regulates social behaviours across the spectrum of the major transition may be too
98 simplistic²³. The roles of conserved and novel processes are not necessarily
99 mutually exclusive; novel processes may coincide with phenotypic innovations, whilst

conserved mechanisms may regulate core processes at all stages of social evolution.

Until recently, the data available to test these hypotheses were largely limited to species that represent either the most complex – superorganismal - forms of sociality (e.g. most ants, honeybees²⁴), or the simplest forms of social complexity as non-superorganisms (e.g. *Polistes* wasps^{7,25–27} and incipiently social bees^{28–31}), the latter of which likely represent the first stages in the major transition. Recent studies have identified core gene sets underlying caste-differentiated brain gene expression across a range of ants⁵ and bees⁸; ants lack ancestrally non-superorganismal representatives and so may not be informative for studying simpler societies and, putatively, the process of social evolution³². Analyses of molecular evolution across two independently evolved lineages of sociality in bees showed increased levels of regulatory complexity with social complexity. These insights highlight the need for a fine-scale dissection of the molecular processes regulating social behaviour for species representing different forms of social complexity⁸, but also the need to consider the influence of life-history traits on gene evolution, as a selective process that is independent of complexity *per se*.

A promising group for unpicking the molecular basis of social complexity are the vespid wasps³³, which exhibit the full diversity of social complexity among some 1,100 species, as well as diversity in life-history traits such as mode of colony founding. Aculeate wasps are of paramount importance in studying the evolution of sociality in the Hymenoptera as they represent the evolutionary root to both bees and ants³⁴. However, to date sociogenomic studies of caste evolution in wasps have been limited to *Polistes*^{7,25,35}. Here we generated brain transcriptomic data of caste-

specific phenotypes for nine species of social wasps, representing diversity in social complexity and mode of colony founding, a life-history trait that is independent of social complexity (Fig. 1). We employed a machine-learning algorithm - support vector machines (SVMs) – along-side more conventional gene expression analyses to conduct fine-scale analysis of the molecular processes associated with caste, to determine whether there is a conserved genetic toolkit for social behaviour in species displaying different forms of sociality and that are putative proxy representatives of the process of the major transition from non-superorganismal (simple) to superorganismal (complex) species in Vespid wasps (Aim 1). We then further interrogate these data to identify how transcriptomic signatures of social behaviour are influenced by social complexity (i.e. simpler versus the more complex societies), mode of colony foundation (i.e. independent-founding versus swarm-founding) as key life-history trait in vespid, and lineage (vespines versus polistines). Our results provide the first evidence of a conserved genetic toolkit across forms of social complexity in the spectrum of the major transition to sociality in wasps; they also reveal how there are fine-scale differences in the molecular patterns and processes between simple and complex societies, and with life-history traits. Finally, we discuss how patterns of gene evolution are contrasting with those reported in bees and ants. These results suggest the concept of a shared genetic toolkit across the spectrum of sociality and across lineages may be too simplistic, and that level of social complexity and life-history traits influence the molecular processes underpinning social behaviour.

Results

We chose one species from each of nine different genera of social wasps from among the two main subfamilies - Polistinae and Vespinae; collectively these species represent the full spectrum of social diversity (detail on species choice given in Fig. 1 and Methods). For each species, we sequenced RNA extracted from whole brains of adults to construct *de novo* brain transcriptomes for the two main social phenotypes – adult reproductives (defined as mated females with developed ovaries, henceforth referred to as ‘queens’ for simplicity; see Supplementary Table S1) and adult non-reproductives (defined as unmated females with no ovarian development, henceforth referred to as ‘workers’; see Supplementary Methods). Using these data we could reconstruct a phylogenetic tree of the Hymenoptera using single-copy orthologous genes (Orthofinder³⁶), resulting in expected patterns of phylogenetic relationships (Supp. Fig. 1). This dataset allows us to test the extent to which the same molecular processes underpin the evolution of social phenotypes of different forms of sociality, as extant proxies in the major transition to superorganismality in wasps.

Aim 1: Is there a shared genetic toolkit for caste among species across the major transition from non-superorganismality to superorganismality in Vespid wasps?

We found several lines of evidence for a shared genetic toolkit for caste across the wasp species using two different analytical approaches.

Caste explains gene expression variation, after species-normalisation

The main factor explaining gene expression variation (using single copy orthogroups) was species identity, and to some extent subfamily, as all the Polistinae clustered tightly (Fig. 2a). However, since we are interested in determining whether there is a shared toolkit of caste-biased gene expression across species, we needed to control for the effect of species in our data. To do this, we performed a between-species normalisation on the transcript per million (TPM) score, scaling the variation of gene expression to a range of -1 to 1 (see Supplementary Methods). After species-normalisation, the samples separate mostly into queen and worker phenotypes in the top two principle components (Fig. 2b). This suggests that subsets of genes (a potential toolkit) are shared across these species and are representative of caste differences. However, there were outliers: *Brachygastra* did not cluster with any of the other samples (despite correct PC1 queen direction); *Agelaia* showed little caste-specific separation and in the opposite direction to the other species (Fig. 2b). These two outlier species may not share the same caste-specific patterns as the other species, but we cannot rule out data and/or sampling anomalies.

Next, we used inferred lists of orthologs allowing some flexibility in numbers of isoforms (maximum of three isoforms, taking the most highly expressed) for each orthogroup and missing data (maximum of two species), allowing larger matrices of genes to be tested for differences between castes (Supplementary Table S2). Using this approach we detected 57 orthologous genes with four or more caste-biased species [queen or worker biased] (Fig. 3a; Supplementary Table S3) and 353 orthologous genes (Supplementary Table S3) with caste-biased expression in at least two of nine species in the same direction (either queen- or worker-biased only).

There was overrepresentation of pheromone, chitin/odorant binding and muscle related GO terms using these two cut-offs (Fig. 3b/c; Supplementary Table S3). No ortholog showed consistent caste-biased differential expression across all nine wasp species (Fig. 3a; Supplementary Table S3; using unadjusted <0.05 p values). Despite this, notable signatures of caste regulation were apparent across the species, with the exception of *Brachygastra*. For example, orthogroup OG0001418 was upregulated in queens from eight of the nine species and is predicted to belong to the vitellogenin gene family (using the *Metapolybia* protein sequence to represent the orthogroup), a transcriptional target and mediator of the insect physiology-regulator juvenile hormone (JH). JH is known as a pleiotropic master regulator of social traits, such as aggression, foraging behaviour and reproduction; in both Vespinae and Polistinae JH has been shown experimentally to regulate reproduction and is thought to be an honest fertility signal^{37,38}. A second gene of interest that was consistently queen-biased was oocyte zinc finger 22. Zinc finger proteins have diverse functions but broadly they regulate gene expression by binding to DNA or RNA and controlling transcription; they are caste-biased in a range of social insects and have also been identified as targets for positive selection in social evolution²¹ and so could be a key transcription factor driving wide expression patterns among castes across our species.

A toolkit of many genes with small effects predicts caste across different forms of sociality in wasps

Conventional differential expression analyses (e.g. edgeR) require a balance of P value cut-offs and fold change requirements to reduce false-positive and false-

negative errors³⁹. Therefore, consistent patterns of many genes with smaller effect sizes may be missed when applying strict statistical measures. Support vector machine (SVM) learning approaches use a supervised learning model capable of detecting subtle but pervasive signals in differential expression between conditions (e.g. for classification of single cells^{40,41}, cancer cells⁴² and in social insect castes⁴³). We used this approach to test whether gene expression can successfully classify caste identity for unknown samples; accurate classification of samples as queens or workers based on their global transcription patterns would be evidence for a genetic toolkit underpinning social phenotypes.

Starting with a “leave-one-out” SVM approach, we classified queen and worker samples, using a predictive gene set generated from a model trained on caste-specific gene expression from eight of the nine species, with the ninth species being the test sample. We applied three different iterations of the leave-one-out approach using different Orthofinder conversion tables (Supp. Table S2) and a radial SVM model (see Methods; Supplementary Fig. 2). Our first approach used the “true-single-copy” orthologues that could be detected with sufficient expression (Fig. 4a-left; Supp. Table S4) across all nine species (n=1221 genes); after progressive feature selection (based on linear regression, from left to right; see Methods) the top 259 genes (<0.05 from linear regression) could predict the queen caste correctly in seven of the nine species (Fig. 4a- left; with ≥ 0.6 likelihood in the queen sample within the top 20% of genes); the two outlier species (*Agelaia* and *Brachygastra*) showed generally lower predictions of queen likelihood (<0.5) and had previously been highlighted as outliers (Figs. 2 & 3). We explored two other orthology lists in order to explore whether, if the single-copy ortholog rule was relaxed and more

orthogroups were included, caste predictions would remain accurate: in one model we allowed up to 3 isoforms per gene (Fig. 4a-middle: Supp. Table S4; n=1831 genes), and in the third we allowed up to two out of nine species to have a missing gene representative of the orthogroup, plus up to three isoforms per gene (Fig. 4a-right: Supp. Table S4; n=5536 genes). Even with these relaxed conditions, caste was classified accurately for all species except for the aforementioned outliers. This suggests that large numbers of genes with caste-biased expression exist across social wasps.

We found some level of functional enrichment among the predictive genes that remained in the models after feature selection. There were consistent patterns of enrichment for GO terms in the two more relaxed models (Fig 4b- middle & 4b-right), with genes enriched for mostly ionic transport and visual/sensory perception (Supplementary Table S4). Expression of the top 50 genes (from linear regression; chosen from Fig 4a- left) allow us to identify putatively most important genes (Fig. 4c); some were previously been identified as associated with caste differentiation in social insects (e.g. vitellogenin; zinc finger (as mentioned above) and ubiquitin); others have not been associated with caste previously but may have relevant functions in regulating caste. For example, striatin-4 (OG0005432) is involved in calmodulin binding and thus may play role in activating (or inactivating) calcium-binding; adenosylhomocysteinase (OG0002120) is one of the most conserved proteins in living organisms facilitates local transmethylation and thus may be important in active transcription and epigenetic processes⁴⁴. Interestingly, the vast majority of the putatively important genes are downregulated in queens, and upregulated in workers, suggesting that perhaps the transcriptomic demands of

worker behaviour may be more diverse than those involved in being a queen, whilst queens express a more restricted molecular toolkit as expected if they are specialised in reproduction.

We compared the genes and GO terms obtained from the SVM and edgeR methods (Fig. 3c). There was little overlap in enrichment of GO terms between the two datasets. Moreover, there was less overlap than expected by chance between the 1420 SVM predictor genes and the 353 genes identified as differentially expressed (edgeR; <0.05 P value): in comparisons that included least two species in the same direction, only 35 genes overlapped (Hypergeometric: $1.89\text{e-}14$, under-enriched 2.59 fold; Supplementary Table S4). Of the genes that are present in both sets, some have previously been identified as having relevance to social evolution and caste differentiation; these include Vitellogenin (as above)⁴⁵ (two protein isoforms: OG0001119 and OG0001418), zinc finger (as above) (OG0005731) and gonadotrophin-releasing hormone, which is related to the regulation of reproduction. ATP-synthase has caste specific expression in the weaver ant⁴⁶ (OG0006255 and OG0006612) and esterase E4-like (OG0000783) is upregulated in young honeybee queens compared to nurses at the proteomic level⁴⁷. There are also other genes of interest, which to our knowledge have not previously associated with caste, including Toll-like receptor 8 (OG0001441) (see Supplementary Table S4). Finally, 40% of the overlapping genes ($n=14$) are uncharacterised proteins, suggesting a role for potentially novel genes; novel (or taxon-restricted) genes have been previously highlighted for their potential role in social evolution across bees and wasps⁴⁸.

Aim 2: How does level of social complexity, phylogeny and life-history influence the molecular basis of castes?

Level of social complexity may influence the genes regulating caste.

To explore whether there are any finer-scale differences in among the different levels of social complexity, we trained an SVM model using *Mischocyttarus*, *Polistes*, *Angiopolybia* and *Metapolybia* as representatives of the simpler societies in the major transition (see Fig. 1), and tested how well this gene set classified castes for the four species representing the more complex societies in the major transition (*Polybia*, *Agelaia*, *Vespa* and *Vespula*; see Fig. 1). *Brachygastra* was excluded due to its poor performance overall (see Fig. 4) and to balance sample sizes in training and test sets. If castes in the test species classify well, this would suggest that the processes regulating castes in the simpler societies are equally important in the more complex societies (i.e. there is no specific caste toolkit for social behaviour in simple societies, which is absent (putatively lost) in more complex societies). Conversely, if the test species do not classify well, this would suggest that there are distinct processes regulating caste in the simpler societies that are less important in more complex forms of sociality.

When training the model on the four species with simple societies, we found that both vespines (*Vespa* and *Vespula*) had good classification confidences at around 0.8 (Fig. 5a), comprising around 800 genes after feature selection. Similarly, *Polybia* also classified well, though with lower confidence. *Agelaia* could not classify castes correctly; however, this species was identified as an outlier in our first analyses and this may reflect some facet of their biology that we do not yet understand, or an anomaly with the data. Considering these four tests together, a set

of 277 gene orthologs were found (Supplementary Table S5; p value <0.05 in all four species tests), that represent a core set of genes that classify caste correctly in simple societies. Based on these species, these results suggest that the genetic toolkit for simple societies is well conserved in the more complex societies that we sampled, although future work should confirm this by expanding the repertoire of species.

The reciprocal test performed less well, suggesting there may be different processes regulating more complex sociality, which is absent from the simpler societies. We trained the SVM using the four species with the more complex societies (*Polybia*, *Agelaia*, *Vespa* and *Vespula*) and tested it on the four species with simpler societies (*Mischocyttarus*, *Polistes*, *Metapolybia* and *Angiopolybia*). After feature selection, around 350 genes were found in each of the four tests, and 289 were found in all four tests, describing castes in these more complex societies (Supplementary Table S5). In general, these genes were successful in classifying castes for the simpler societies (Fig. 5a-lower), although none of the species matched the high classification confidence of the vespines in the reciprocal test. The lower overall success of the gene set trained on the more complex species in classifying castes in species from the simpler societies suggests there may be different (additional) processes regulating castes at the more complex levels of sociality.

To explore this idea further, we tested whether there was significant overlap among the three putative toolkits, by comparing the identities of genes predictive of 'simpler societies' (n=277), 'more complex societies' (n=289), and 'full spectrum' (Fig 4a- Left; n=259) toolkits. We found significant overlap between the combined genes

in the simpler and the more complex ‘toolkits’ with the full spectrum set (Fig 5b grey vs pink; blue vs pink), but not between the simpler and more complex societies (grey vs blue). This suggests that there may be distinct differences in the genes regulating castes at different levels of social complexity. We found further suggestive evidence for this at the putative functional level: there were no consistent patterns of GO terms across the two toolkits, with only the complex society toolkit having significant levels of enrichment after correction (Fig. 5c- mostly cation/ion transport; Supplementary Table S5).

Further support for there being different molecular processes shaping caste transcription at different levels of social complexity was found in the patterns of gene evolution. Overall, the highest rates of dN/dS were in the species with simpler societies; this was true across all nine species (dN/dS mean of orthogroups in foreground branch with simple species = 0.1639218; with complex species = 0.1015024, Wilcoxon test $W = 307318$, $p\text{-value} < 2.2e-16$) and also just within the Polistinae (dN/dS mean of orthogroups in foreground branch with simple species = 0.180918; with complex species = 0.1327845; Wilcoxon test $W = 119615$, $p\text{-value} = 2.824e-11$). Some orthogroups had experienced significant positive selection (i.e. on the given foreground branch, null model rejected, chi-square test $p < 0.05$, $dN/dS \Rightarrow 1$; a total of eleven orthogroups). Six of these orthogroups were under positive selection in the simplest polistines relative to the more complex polistines (Supplementary Table S7-ST15 $n = 6$; OG0002957 (transmembrane 234 homolog isoform X1), OG0003204 (mitochondrial ribonuclease P 3, upregulated in zebrafish behavioural experiments⁴⁹), OG0003159 (lysozyme c-1-like, in ant venom⁵⁰), OG0003184 (doublesex isoform X1, female-specific gene expression in ant⁵¹),

361 OG0005376 (uncharacterized protein LOC107063793), OG0005104 (box C D
362 snoRNA 1)). Four were among the more complex polistines, relative to the simpler
363 species (Supplementary Table S7-ST16: OG0003362 (UMP-CMP kinase isoform
364 X1), OG0003491 (uncharacterized protein LOC107067502), OG0004867
365 (uncharacterized protein LOC106789623), OG0005164 (probable tubulin
366 polyglutamylase TTLL9)). However, when the vespines were also included, only one
367 orthogroup was under positive selection (Supplementary Table S7-ST14;
368 OG0003694 (transmembrane 59-like isoform X1)), suggesting there may be some
369 lineage effects here, although wider taxonomic sampling of vespines would be
370 required to confirm this. Of the 11 significant orthogroups identified across these
371 analyses, only one was differentially expressed with respect to caste in two species
372 (OG0004867, uncharacterized protein LOC106789623), and four other orthogroups
373 (OG0003362: UMP-CMP kinase isoform X1; OG0003491: uncharacterized protein
374 LOC107067502; OG0003159: lysozyme c-1-like; OG0003184: doublesex isoform
375 X1) were differentially expression in just one species (Supplementary Table S7-
376 ST27).

377 Taken together, these results provide evidence that the molecular processes
378 regulating caste differentiation vary depending on the level of social complexity. The
379 finding of more rapid rates of protein evolution in the simpler societies
380 (Supplementary Figure 3c) was unexpected since studies of gene evolution in social
381 bees have found the higher rates of evolution in the more complex societies⁸.
382 Molecular processes shaping the evolution of sociality may differ with lineage and/or
383 life-history. Patterns of molecular evolution should be further explored with broader

analysis of species representing different levels of complexity within lineages, but also with wider taxonomic breadth when comparing across lineages.

No evidence that the genes regulating castes vary among subfamilies

Using the same reciprocal SVM approach, we found little evidence that sub-family influences the molecular processes regulating caste. Using the seven polistines as the training set, queens and workers in the two vespines classified with 100% confidence (Supplementary Table 6 [tab *Vespula*_Polistinae + *Vespa*_Polistinae]; the reverse of this test could not be performed due to low sample sizes for a Vespine training set). Despite the small number of vespines sampled, these results suggests that the genes regulating castes are shared across these two subfamilies.

Life-history traits may influence the genes regulating caste

A key life-history trait that varies in social wasps is their mode of colony founding. Some species form new colonies alone ('independent founders') whilst others form new colonies as a swarm containing queens and workers ('swarm founders'). Founding behaviour is independent of level of complexity: e.g. the species with the simplest (e.g. *Polistes* and *Mischocyttarus*) and the most complex (*Vespa* and *Vespula*) societies are all independent founders; equally the swarm-founders (*Angiopolybia*, *Metapolybia*, *Agelaia*, and *Polybia*) exhibit a spectrum of social complexity (see Fig. 1). It is possible therefore, that the molecular basis of caste reflects this key life-history trait, and not just social complexity. We found some evidence to support this. The gene set obtained from an SVM trained to classify castes in the swarm-founders was very successful in classifying castes for the

independent founders, with classifications between 0.7-1 for all test species. Conversely, the reciprocal analysis performed poorly: the gene set obtained from an SVM trained on the independent-founders was unable to predict castes at >0.7 for three of the four species (Supplementary Figure 4; Supplementary Table 6 [Swarm/Independent]). The influence of colony founding behaviour on molecular processes was further supported when comparing the rates of protein evolution in the swarm-founders relative to the independent founding species (Supplementary Table 7 - ST12): the rates of protein evolution are on average higher among the swarm-founders (mean = 0.1461955) than the species with complex species but smaller than simple species (Wilcoxon tests $W = 384912$, $p\text{-value} = 1.847\text{e-}07$, and $W = 126972$, $p\text{-value} = 0.0001657$ respectively). These results suggest that there may be additional caste-related molecular processes involved in the evolution of innovations in life-history traits (in this case, mode of colony founding) that are independent of social complexity.

Discussion

Major transitions in evolution provide a conceptual framework for understanding the emergence of biological complexity. Discerning the processes by which such transitions arise provides us with critical insights into the origins and elaboration of the complexity of life. In this study we explored the evidence for two key hypotheses on the molecular bases of social evolution by analysing caste transcription in nine species of wasps. As predicted, we find evidence of a shared genetic toolkit across

the spectrum of social complexity in wasps; importantly, using machine learning we reveal that this toolkit likely consists of many hundreds of genes of small effect (Fig. 4). However, in sub-setting the data by level of social complexity and by a key life-history trait (colony founding), several important new insights are revealed. Firstly, there appears to be a putative toolkit for castes in the simpler societies that largely persists across the major transition, through to superorganismality. Secondly, different (additional) processes appear to become important at more complex levels of sociality. Further sampling is required to determine the extent to which the role of these additional processes is driven by the evolution of superorganismality, and specific innovations that have been key in the major transition to sociality. Thirdly, the so-called toolkit for sociality is also influenced by the mode of colony founding; the influence of ecology and life-history is largely overlooked in the current literature on the molecular basis of sociality. Finally, we discuss apparent differences among lineages, highlighting the importance of wider taxonomic consideration in the quest to understand the molecular basis of sociality and the nature of the major transition. Taken together, we suggest that the concept of a shared molecular toolkit for sociality may be too simplistic and that future studies should consider the evolutionary impact of other traits on molecular processes, specifically life-history and ecology, alongside castes.

The first important finding is that we identified a substantial set of genes that consistently classify caste across most of the species, irrespective of the level of social complexity. The taxonomic range of samples used meant we were able to confirm that specific genes are consistently differentially expressed, with respect to

caste, across the species. These patterns would be difficult to detect if only looking at a few species within a lineage, or species across several lineages, or species representing only a limited selection of social complexities. In addition to typical caste-biased molecular processes, we also identified that genes related to synaptic vesicles are different between castes; this is interesting as the regulation of synaptic vesicles affects learning and memory in insects⁵². To our knowledge, this is the first evidence of what may be a conserved genetic toolkit for sociality, from the simplest stages of social living through to true superorganismality, including intermediate stages of complexity in wasps. Greater taxonomic sampling will help recover the full spectrum of genes involved that are regulated across the major transition and may have been important in the evolution of sociality.

The underlying assumption, based on the conserved toolkit hypothesis, has been that whatever processes regulate castes in complex societies must also regulate castes in simpler societies^{6,23,53}. Unexpectedly, our analyses suggest that different molecular processes underpinning castes come into play at the more complex levels of sociality compared to species with more simple societies. The predictive gene set identified in the SVM trained on more complex species performed less well in classifying caste than the predictive toolkit derived from the simpler species. There may be fundamental differences discriminating (near) superorganismal societies from non-superorganismal societies⁵⁴. This highlights the importance of examining different stages in the major transition when attempting to elucidate its patterns and processes^{8,10,33}.

Vespinae and Polistinae display a wide diversity in traits other than social complexity. One of these in the species we sampled was the mode of colony

foundation. Some species start new colonies alone, or with a small group of reproductively potent females (independent founders) whilst others form a new colony as a group of queen(s) and workers (swarm founders). Among the Polistinae, mode of colony founding has a phylogenetic basis, with the Polistini and Mischocyttarini tribes being independent founders, whilst the Epiponini tribe are all swarm founders (see Fig. 1)^{55,56}. Within the Polistinae, therefore, there is a tendency to consider swarm founding as a trait that is associated with social complexity. However, the Vespinae sampled in this study are also independent founders, and yet display the most complex forms of sociality. As such, our sampling afforded the opportunity to determine whether this important life-history trait influenced brain transcription among social phenotypes, independently of the form of social complexity. Consistent with the idea that independent founding is an ancestral trait, we detected additional brain transcriptomic processes at play among castes of the swarm founders, that were did not distinguish castes among the independent founders. Clearly, wider taxonomic sampling is required to explore this further, but these analyses bring into focus how the transcriptomic basis of castes is not a simple reflection of social complexity; ecology and life-history deserve more attention in explaining variation in the molecular bases of social phenotypes⁵⁷. This is likely to be especially important when seeking to compare patterns of caste transcription *across* lineages with contrasting life-histories and ecologies.

Phenotypic evolution is often accompanied by increased rates of gene evolution. Indeed, high rates of protein evolution have been detected in worker-biased genes of more complex societies like the honeybee, and some ants^{16–19}. This makes sense as the worker caste (with altruism and partial/complete sterility) is the

503 dominant phenotypic innovation in social evolution and has led to the suggestion that
504 rapid gene evolution is expected in the most complex forms of sociality, where
505 phenotypic specialisations are most pronounced^{17,20}. Our dataset afforded a first test
506 of this idea in the wasps. We were surprised to find that the highest rates of gene
507 evolution were detected in the wasp species with the simpler societies rather than
508 the more complex societies; these results are in stark contrast with the findings of a
509 recent study on a range of social complexity in bees⁸ where (in keeping with
510 previous studies), rapid gene evolution was most apparent among species with the
511 most complex societies. There are several interpretations of these contrasting
512 results. First, the level of social complexity displayed by the simplest species in Shell
513 et al 2021 represents a much more rudimentary form of sociality compared to the
514 simplest societies in our study, including four solitary species, one facultatively social
515 species (e.g. females of *Ceratina australensis* can nest either alone or in a small
516 group) and some barely exhibiting any reproductive division of labour (e.g. *Ceratina*
517 *calcarata*). Facultative sociality is rare in the vespid wasps, being largely limited to a
518 few species of Stenogastrinae; the simplest Polistinae are represented in our sample
519 and all are obligately social (meaning they always nest in a group). It is possible
520 therefore, that the rapid gene evolution occurs only when a species is committed to
521 group living, with a clear division of reproductive labour. An alternative explanation
522 lies in ecology and life-history: bees and wasps differ fundamentally in how they
523 provision brood, with wasps hunting prey and bees collecting pollen. These findings
524 add to the complexities of explaining the molecular basis of social evolution, and
525 highlight a greater need to emphasis and account for variation in natural history,
526 ecology and life-history as well as social behaviour per se^{10,24,33,54}.

Despite being able to extract consistent SVM predictions, our models are only as good as the initial data used to train them. Our study suffers from a few limitations. Firstly, the sample size (number of species) is good for a sociogenomics study, but relatively low for machine learning studies; SVMs are generally used on very large datasets such as clinical trials in the medical sciences⁵⁸. Although our models did perform well, the analyses would be more robust by using more species in the training datasets. Indeed, we observed reduced performance in our model predictions when fewer species were included in the training set. Secondly, by comparing across multiple species, we can only train our model on genes that have a single representative isoform per species in each separate test. This reduces the numbers of genes we can test in each SVM model, especially where more distantly related species are included. We overcame this limitation by merging gene isoforms within the same orthogroup (potential gene duplications); yet this comes with some additional costs as some genes are discarded in this process. Finally, genomes are not available for most of the species we tested; our measurements are based on *de novo* sequenced transcriptomes, which potentially contain misassembled transcripts, which could reduce the ability to find single copy orthologs across species. For these reasons, the numbers of genes detected in our putative toolkit for sociality is likely to be conservative and modest (potentially by several fold). These limitations are likely to apply to many similar studies, due to the difficulty and expense of obtaining high quality genomic data for specific phenotypes for non-model organisms. Our study illustrates the power of SVMs in detecting large suites of genes with small effects, which largely differ from those identified from conventional differential expression analysis⁴³. We advocate the use of the two methods in parallel: our conventional

analyses suggested that metabolic genes appear to be responsible for the differences between castes, whereas the SVM genes were mostly enriched in neural vesicle transportation genes, which have not previously been connected with caste evolution. SVMs may therefore reveal new target for genes involved in the evolution of sociality. We anticipate that machine learning approaches, as demonstrated here, may become a useful tool in a wide range of ecological and evolutionary studies on the molecular basis of phenotypic diversity.

In conclusion, our analyses of brain transcriptomes for castes of social wasps suggest that the molecular processes underpinning sociality are conserved throughout the major transition to superorganismality. However, different (putatively additional) processes may come into play in more complex societies. There may be fundamental differences in the molecular machinery that discriminates key points of innovation in the major transition; for example, the evolution of irreversible caste commitment (in superorganisms) may require a fundamental shift in the underlying regulatory molecular machinery. Such shifts may be apparent in the evolution of sociality at other levels of biological organisation, such as the evolution of multicellularity, taking us a step closer to determining whether there is a unified process underpinning the major transitions in evolution.

Acknowledgements

We would like to thank Laura Butters for her help with the morphometric analyses of *Brachygastra*, Robin Southon, Daniel Fabbro, Liam Crowley and Sam Morris for their assistance in the field, James Carpenter at the American Natural History Museum

575 and Christopher K. Starr for confirming species identification, and the Bristol
576 Genomics Facility for their assistance with library preparations and sequencing. This
577 work was conducted under collection and export permits for Trinidad (Forestry
578 Division Ministry of Agriculture: #001162) and Panama (Autoridad nacional del
579 ambiente (ANAM) SE/A-55-13). It was funded by the Natural Environment Research
580 Council (NE/M012913/2; NE/K011316/1) awarded to SS, and a NERC studentship
581 and Smithsonian Tropical Research Institute pre-doctoral fellowship awarded to
582 E.F.B.

Methods

Study Species

Nine species of vespid wasps were chosen to represent different levels of social complexity across the major transition (Fig. 1). The simplest societies in our study are represented by *Mischocyttarus basimacula basimacula* (Cameron) and *Polistes canadensis* (Linnaeus); wasps in these two genera are all independent nest founders and lack morphological castes (defined as allometric differences in body shape, rather than overall size) or any documented form of life-time caste-role commitment^{59–62}. They live in small family groups of reproductively totipotent females, one of whom usually dominates reproduction (the queen); if the queen dies she is succeeded by a previously-working individual²¹. As such, these societies represent some of the earliest stages in the major transition, where caste roles are least well defined, and where individual-level plasticity is advantageous for maximising inclusive fitness.

The Neotropical swarm-founding wasps (Hymenoptera: Vespidae; Epiponini) include over 20 genera with at least 229 species, exhibiting a range of social complexity measures, from complete absence of morphological caste (pre-imaginal) determination to colony-stage specific morphological differentiation, through to permanent morphological queen-worker differentiation⁵⁶. As examples of species for which there is little or no evidence of developmental (morphological) caste determination, we chose *Angiopolybia pallens* (Lepeletier) which is phylogenetically basal in the Epiponines^{55,63} and *Metapolybia cingulata* (Fabricius)^{55,63}. We confirmed

605 the lack of clear caste allometric differences in *M. cingulata* as data were lacking
606 (see Morphometrics methods (below) and Supplementary Table S8).

607 As examples of species showing subtle, colony-stage-specific caste allometry, we
608 chose a species of *Polybia*. The social organisation of *Polybia* spp is highly variable,
609 ranging from complete absence of morphological queen-worker differentiation⁶⁴.

610 *Polybia quadricincta* (Saussure) is a relatively rare (and little studied) epiponine
611 wasp which can be found across Bolivia, Brazil, Columbia, French Guiana, Guyana,
612 Peru, Suriname and Trinidad (Richards, 1978). Our morphometric analyses found
613 some evidence of subtle allometric morphological differentiation in this species, but
614 with variation through the colony cycle (Supplementary Table S8); this suggest it is a
615 representative species for the evolution of the first signs of pre-imaginal caste
616 differentiation.

617 Many species of the genera *Agelaia* and *Brachygastra* appear to show pre-imaginal
618 caste determination with allometric morphological differences between adult queens
619 and workers⁵⁵. We chose one species from each of these genera as representatives
620 of the most socially complex Polistine wasps. Although no morphological data were
621 available for *Agelaia cajennensis* (Fabricius) all species of *Agelaia* studied show
622 some level of preimaginal caste determination^{55,65}. *Brachygastra* exhibit a diversity of
623 caste differentiation^{55,66}; our morphological analysis of caste differentiation *B.*
624 *mellifica* (Say) confirms that this species is highly socially complex, with large colony
625 sizes⁵⁵ and pre-imaginal caste determination resulting in allometric caste differences
626 (Supplementary Table S8).

627 All species of Vespines are independent nest founders and superorganisms, with a
628 single mated queen establishing a new colony alone and with morphological castes

that are determined during development. However, some species exhibit derived superorganismal traits, such as multiple mating⁶⁷, which have likely evolved under different selection pressures to the major transition itself⁶⁸. The European hornet, *Vespa crabro* (Linnaeus), exhibits the hallmarks of superorganismality (see Fig. 1) but little evidence of more derived superorganismal traits, such as high levels of multiple mating. Conversely, multiple mating is common in *Vespula* species, including *V. vulgaris* (Linnaeus) with larger colony sizes than *Vespa*⁶⁷, suggesting a more complex level of social organisation.

Sample collection

Where possible, we sampled from colonies representing different stages in the colony cycle, as caste differentiation can vary as the colony matures in some species (Supplementary Table S1. *Metapolybia cingulata* (6 colonies), *Polistes canadensis* (3 colonies), *Agelaia cajennensis* (1 colony) and *Mischocyttarus basimacula basimacula* (3 colonies) were collected from wild populations in Panama in June 2013. *Brachygastra mellifica* (4 colonies) were collected from populations in Texas, USA in June 2013. *Angiopolybia pallens* (2 colonies) and *Polybia quadricincta* (2 colonies) were collected from Arima Valley, Trinidad in July 2015. *Vespa crabro* (4 colonies) and *Vespula vulgaris* (4 colonies) were collected from various locations in South East England, UK in 2017. Queens and workers were collected directly from their nests during the daytime, placed immediately into RNA^{later} (Ambion, Invitrogen) and stored at -20°C until further use. Samples were ultimately pooled *within castes* for all analyses, such that each pool consisted of 3-6 individual brains from wasps sampled across 2-4 colonies to capture individual-level

and colony-level variation in gene expression (see Supplementary Table S1).
Samples of *M. cingulata*, *A. cajennensis*, *M. basimacula* and *B. mellifica* were sent to
James Carpenter at the American Natural History Museum for species verification.
A. pallens and *P. quadricincta* were identified by Christopher K. Starr, at University of
West Indies, Trinidad and Tobago.

Morphometrics

Data on morphological differentiation among colony members (and thus information
on whether pre-imaginal (developmental) caste determination was present) was
lacking for *M. cingulata*, *P. quadricincta* and *B. mellifica*; therefore, we conducted
morphometric analyses on these three species in order to ascertain the level of
social complexity. Morphometric analyses were carried out using GXCAM-1.3 and
GXCapture V8.0 (GT Vision) to provide images for assessing morphology. We
measured 7 morphological characters using ImageJ v1.49 for queens and workers
for each species. The body parts measured were: head length (HL), head width
(HW), minimum interorbital distance (MID), mesoscutum length (MSL), mesoscutum
width (MSW), mesosoma height (MSH) and alitrunk length (AL) (for measurement
details, see ⁶⁹). Abdominal measurements were not recorded as ovary development
could alter the size of abdominal measurements, therefore biasing the results. The
morphological data were analysed to determine whether the phenotypic
classification, as determined from reproductive status, could be explained by
morphological differences. ANOVA was used to determine size differences between
castes for each morphological characteristic. A linear discriminant analysis was also
employed to see if combinations of characters were helpful in discriminating between
castes. The significance of Wilks' lambda values were tested to determine which

morphological characters were the most important for caste prediction. All statistical analyses were carried out using SPSS v23.0 or Exlstat 2018. Data and analyses given in Supplementary Table S8.

Dissections & RNA extractions

Individual heads were stored in RNAlater for brain dissections; abdomens were removed and dissected to determine reproductive status. Ovary development was scored according to ^{33,54} and the presence/absence of sperm in the spermathecae was identified to determine insemination. Inseminated females with developed ovaries were scored as 'queens'; non-inseminated females with undeveloped ovaries were scored as workers. Brains were dissected directly into RNAlater; RNA was extracted from individuals and then pooled after extraction into caste-specific pools; pooling after RNA extraction allowed for elimination of any samples with low quality RNA. Pooling individuals was generally necessary to ensure sufficient RNA for analyses, as well as accounting for individual variation to ensure expression differences are due to caste or species, and not dependent on colony or random differences between individuals. One exception to this was the *V. vulgaris* samples which were sequenced as individual brains and pooled bioinformatically after sequencing. Individual sample sizes per species are given in Supplementary Table S1.

Total RNA was extracted using the RNeasy Universal Plus Mini kit (Qiagen, #73404), according to the manufacturer's instructions, with an extra freeze-thawing step after homogenization to ensure complete lysis of tissue, as well as an additional

elution step to increase RNA concentration. RNA yield was determined using a NanoDrop ND-8000 (Thermo Fisher Scientific); all samples showed A260/A280 values between 1.9 and 2.1. An Agilent 2100 Bioanalyser was used to determine RNA integrity. Samples of sufficient quality and concentration were pooled and sent for sequencing. Libraries were prepared using Illumina TruSeq RNA-seq sample prep kit at the University of Bristol Genomics Facility. Five samples were pooled per lane to give ~ 50M read per sample. Paired-end libraries were sequenced using an Illumina HiSeq 2000. Descriptions of pooling of individuals and pooled sets into single representatives of caste are shown in (Supplementary Table S1). Raw reads are available on SRA/GEO (GSE159973).

Preparation of *de novo* transcriptomes

Transcriptomes of *Agelaia*, *Angiopolybia*, *Metapolybia*, *Brachygastra*, *Polybia*, *Polistes* and *Mischocyttarus* were assembled using the following steps. First, reads were first filtered for rRNA contaminants using tools from the BBTools (version:BBMap_38) software suite (<https://jgi.doe.gov/data-and-tools/bbtools/>). We then used Trimmomatic v0.39⁷⁰ to trim reads containing adapters and low-quality regions. Using these filtered RNA sequences, we could assemble a *de novo* transcriptome for each species (merging queen and worker samples) using Trinity v2.8⁷¹ and filter protein coding genes to retain a single transcript (most expressed) for each gene and transcript per million value (TPM), which we use for the rest of the analyses.

For *Vespula* and *Vespa*, reads from both queen and worker samples were assembled into *de novo* transcriptomes using a Nextflow pipeline

(github:biocorecrg/transcriptome_assembly). This involved read adapter trimming with Skewer⁷², *de-novo* transcriptome assembly with Trinity v2.8.4⁷¹ and use of TransDecoder v5.5.0⁷¹ to identify likely protein-coding transcripts, and retain all translated transcripts. These were further filtered to retain the largest open read frame-containing transcript, which we listed as the major isoform of each protein. Trinity assembly statistics are shown in Supplementary Table S2.

Identification of orthologs

To identify gene-level orthologs, we used Orthofinder v.2.2.7³⁶ with diamond blast⁷³, multiple sequence alignment program MSA⁷⁴ and tree inference using FastTree v2.1.10⁷⁵, for our nine species (Supplementary Fig. 1). The largest spliced isoform per gene (from Trinity) was designated the representative sequence for each gene. In addition for specific iterations, we allowed the existence of multiple gene isoforms belonging to the same species in a single orthogroup (potential duplications); using up to 3 gene isoforms, only keeping the gene most highly expressed. This strategy allows us to test a greater number of orthogroups, where a single duplication in one species would otherwise remove an entire orthogroup from the study. The drawback is that some gene isoforms are removed from the study that could have been informative. Similarly, we were able to include orthogroups that had missing gene information (NAs) for up to two species, filling in these missing genes with read counts of 10 for both queen and worker, thereby allowing us to test orthogroups where some species data are missing. Missing gene orthologs in a single species could be due to a true absence (likely gene loss) in a species, or due to assembly weaknesses, where the use of non-genome guided RNA-Seq assemblies from just

brain tissue, could lead to missing orthologs. Scripts to perform these steps are included in the Github (https://github.com/Sumner-lab/Multispecies_paper_ML/tree/master/Scripts_to_improve_orthogroups). Orthofinder was also used to create the phylogenetic tree in Supplementary Figure 1. Using the single copy orthogroups only, and with the additional of *Apis mellifera* (GCF_000002195.4), *Nomia melanderi* (GCA_003710045.1), *Dinoponera quadriceps* (GCA_001313825.1) and *Solenopsis invicta* (GCF_016802725.1).

Gene expression and differential expression

We calculated abundances of transcripts within queen and worker samples using “align_and_estimate_abundance.pl” within Trinity, using estimation method RSEM v1.3.1⁷⁴, “trinity_mode” and bowtie2⁷⁶ aligner. We then used edgeR v3.26.5⁷⁷ (R version 3.6.0) to compare gene expression between queens and workers. Given that we were comparing a single sequencing pool of several individuals per caste, we used a hard-coded dispersion of 0.1 and the robust parameter set to true to account for $n = 1$. Raw P values for each gene were corrected for multiple testing using a false discovery rate (FDR) cut-off value of 0.05. We did not take advantage of genome data (where available), as only two of the species had published genomes at the time of analysis; using transcriptome-only analyses makes the analysis more consistent across species. Trinity assemblies and RSEM counts are available on GEO/SRA (GSE159973). To compare differential expression across the nine species (Figure 3), we used the orthology list based on three splice isoforms represented as the most expressed and using only orthogroups with a single entry in all nine species after the merging of three isoforms.

Building a normalised gene expression matrix

Focused on our set of one-to-one orthologs (merging 3 or less isoforms per species), we began by computing log transformed TPMs (transcripts per million reads) for each gene in each sample from the raw counts, followed by quantile normalisation. Next, we normalised for species, using an approach that is comparable to calculating a species-specific z-score for each sample. Specifically, we transformed the expression scores calculated above by subtracting the species mean and dividing by the species mean for each sample within a species. This calculation has two important effects. Firstly, subtracting the species mean from each sample within a species centres the mean expression of each species on zero, making the units of expression more comparable across species. Secondly, dividing by the species mean from each sample standardises the expression scores, producing a measure that is independent of the units of measurement, so that the magnitude of difference between queens and workers in each sample is no longer important. The transformed expression score thus allows us to focus on the relative expression in queens versus workers across species. Finally, we removed orthogroups where the counts per million were below 10 (with some deviations in specific iteration; listed with run plots) in both Queen and Worker samples of each species, to remove lowly expressed genes that may contribute noise to subsequent analyses. We then performed principal component analysis (PCA) in R on the raw TPM values and those with the normalisation steps outlines above (see: https://github.com/Sumner-lab/Multispecies_paper_ML/blob/master/Scripts_to_run_svm/template_scripts/getExpressionAllOrthofinder.noNameReplace.R for full R code of this section).

796

797 **Machine learning (support vector machines)**

798 Support vector machine (SVM) classification was used to classify caste across the
799 species. In brief, this analysis involved taking species-scaled, logged and normalised
800 gene expression data from the nine species, filtering the lowly expressed genes,
801 splitting into training and testing datasets (8 training, 1 testing), creating an SVM
802 classifier that best separates caste in the training species (find best gamma and C
803 tuning variables) and testing one of species left out, given the classification. The
804 code to run these steps is shown on github ([https://github.com/Sumner-](https://github.com/Sumner-lab/Multispecies_paper_ML)
805 [lab/Multispecies_paper_ML](https://github.com/Sumner-lab/Multispecies_paper_ML)), and is executed in perl using subsidiary R scripts. The
806 full detail of these steps are outlined below.

807 The first step is to normalise the expression matrix, given the svm classifier
808 expects the fully normalised and scaled data. This has been outlined in the previous
809 section, and involves orthologous TPMs being logged, quantile and species-scaled;
810 as well as filtering of lowly expressed genes. Second, data were split into test and
811 train sets, so each species was addressed in turn, with queen and worker samples
812 registered based on sample names (e.g. `caste <- grepl("Worker ",`
813 `colnames(data.train))`), and converted into a binary format. Third, a svm classifier
814 could be tuned using the train data (8 of 9 species), using the R package `e1071`⁷⁸,
815 with optional grid search ranges of $\gamma = 10^{(-8:-3)}$ and $\text{cost} = 2^{(1:10)}$. These
816 settings allow the classifier to explore the best gamma and cost settings to use in the
817 model. Fourth, based on the classifier and the best model it predicted (always the
818 radial model, and variable C and gamma values), we could predict the likelihood of
819 the nines sample being a queen or a worker. To explore which kernel was most

appropriate, we ran the script using radial, polynomial, linear and sigmoidal kernels (Supplementary Figure 2), which showed consistent changes in prediction certainty among the radial, polynomial and sigmoidal kernels but more erratic predictions in the linear kernel. We decided to continue with the radial kernel for the rest of our predictions, as it has decent prediction accuracy once the dataset is filtered, and has progressive improvement across the filtering spectrum (Supplementary Figure 2; top left), unlike the linear kernel (bottom left).

Given the majority of genes are likely not to contribute to caste differentiation, on top of the basic svm test, we performed feature selection, which is commonly used in machine learning to increase the power of the classifier⁷⁹. For this we used linear regression of each gene on caste ($\text{lm}(\text{caste} \sim \text{expr}, \text{data})$), using the training data only. With regression beta coefficients per orthologous gene, we could then rank genes by their statistical association with caste (Supp. Table 4 using the absolute values of the regression coefficients). This enabled us to measure how the classification certainty changed as we filtered out genes statistically un-associated with reproductive division of labour (Figure 4a).

Classification certainty of 0.5 would indicate the SVM could not tell the difference between the two castes (maximal uncertainty), a classification certainty of 0 would indicate maximal certainty of a worker and 1 a maximal certainty of being a queen.

dNdS scores

Using the single-copy orthologous genes for the nine species, we aligned nucleotide sequences for each of the 1,971 orthogroups (single-copy allowing for 3 isoforms)

that successfully went through PRANK (version v.151120⁸⁰). For each species and for each orthogroup, we then calculated dn/ds ratios (omega) in branch models (Supplementary Table S7 [ST2-ST16] and Supplementary Figure 3; null hypothesis model in PAML codeml (model = 0, NSSites = 0, tree file without foreground) and in an alternative hypothesis model (model = 2, NSSites = 2, tree file without some taxa as foreground)); as well as in branch-site models (ST17-ST26; null hypothesis model (omega fixed at 1, model = 2, NSSites = 2, tree file with one species branch as foreground) and in an alternative hypothesis model (omega not fixed, all other parameters set as the null model) (version v. 4.8⁸¹)). We assessed log likelihoods using a chi-square test and reported the adjusted p-values using R (version 3.6.3) and Benjamini and Hochberg adjustment⁸².

GO enrichment and BLAST

To perform GO enrichment tests, we used the R package TopGO v2.42.0⁸³, using Bonferroni cut-off *P* values of ≤ 0.05 . In order to assign gene ontology terms to genes in our new species, we used our Orthofinder homology table with annotations to *Drosophila melanogaster* (downloaded from Ensembl Biomart 1.10.2019). Within species, we calculated enrichment of each species' gene to a background of all the genes expressed above a mean of 1 TPM. When comparing across the orthogroups (OG), we used *Metapolybia* GO annotations (derived from homology to *Drosophila*), with a background of those genes that have a mean >1 TPM in all species orthologues. GO comparisons were similar using other species as a database of gene to GO terms.

867 Blast2GO v 1.4.4⁸⁴ with default settings was used to annotate Metapolybia
868 genes.

869

870 **Abbreviations**

871 DOL = division of labour

872 ORF = open reading frame

873 MT = major transition

874 GO = gene ontology

875 SRA = sequence read archive

876 NCBI = National Centre for Biotechnology Information

877 BUSCO = Benchmarking set of Universal Single-Copy Orthologs

878 SVM = support vector machine

879 PCA = principal components analysis

880 Blast = Basic local alignment search tool

881 nr = non-redundant

882

883 **Author's contributions**

884 SS conceived the study and supervised the project; SS, EL, EB and BT collected the
885 samples; DT, EB, BT and RB performed molecular lab work; DT & RB carried out the
886 morphological work; EF, MB & CW executed the bioinformatics pipelines, performed
887 the statistical analyses; CW & SS drafted the manuscript, with input from all authors.

888

889 **References**

890

- 891 1. Szathmáry, E. & Maynard Smith, J. The major evolutionary transitions. *Nature*
892 **374**, 227–232 (1995).
- 893 2. Kennedy, P. *et al.* Deconstructing Superorganisms and Societies to Address
894 Big Questions in Biology. *Trends in Ecology and Evolution* **32**, 861–872
895 (2017).
- 896 3. Toth, A. L. & Robinson, G. E. Evo-devo and the evolution of social behavior.
897 *Trends Genet.* **23**, 334–341 (2007).
- 898 4. Berens, a. J., Hunt, J. H. & Toth, a. L. Comparative Transcriptomics of
899 Convergent Evolution: Different Genes but Conserved Pathways Underlie
900 Caste Phenotypes across Lineages of Eusocial Insects. *Mol. Biol. Evol.* **32**,
901 690–703 (2014).
- 902 5. Qiu, B. *et al.* Towards reconstructing the ancestral brain gene-network
903 regulating caste differentiation in ants. *Nat. Ecol. Evol.* **2**, 1782 (2018).
- 904 6. Warner, M. R., Qiu, L., Holmes, M. J., Mikheyev, A. S. & Linksvayer, T. A.
905 Convergent eusocial evolution is based on a shared reproductive groundplan
906 plus lineage-specific plastic genes. *Nat. Commun.* **10**, 1–11 (2019).
- 907 7. Patalano, S. *et al.* Molecular signatures of plastic phenotypes in two eusocial
908 insect species with simple societies. *Proc. Natl. Acad. Sci. U. S. A.* **112**,
909 13970–5 (2015).
- 910 8. Shell, W. A. *et al.* Sociality sculpts similar patterns of molecyalr evolution in

- 911 two independently evolved lineages of eusocial bees. *Commun. Biol.* **4**, 1–9
912 (2021).
- 913 9. West-Eberhard MJ. Wasp societies as microcosms for the study of
914 development and evolution. in *Natural history and evolution of paper wasps*.
915 (eds. Turillazzi, S. & West-Eberhard, M. J.) 290–317 (Oxford University Press,
916 1996).
- 917 10. Toth, A. L. & Rehan, S. M. *Molecular Evolution of Insect Sociality: An Eco-Evo-*
918 *Devo Perspective. Annual Review of Entomology* **62**, 419–442 (Annual
919 Reviews, 2017).
- 920 11. Boomsma, J. J. Lifetime monogamy and the evolution of eusociality. *Philos.*
921 *Trans. R. Soc. Lond. B. Biol. Sci.* **364**, 3191–207 (2009).
- 922 12. Boomsma, J. J. & Gawne, R. Superorganismality and caste differentiation as
923 points of no return: how the major evolutionary transitions were lost in
924 translation. *Biol. Rev.* (2017). doi:10.1111/brv.12330
- 925 13. Ferreira, P. G. *et al.* Transcriptome analyses of primitively eusocial wasps
926 reveal novel insights into the evolution of sociality and the origin of alternative
927 phenotypes. *Genome Biol.* **14**, R20 (2013).
- 928 14. Sumner, S. The importance of genomic novelty in social evolution. *Mol. Ecol.*
929 **23**, (2014).
- 930 15. Feldmeyer, B., Elsner, D. & Foitzik, S. Gene expression patterns associated
931 with caste and reproductive status in ants: Worker-specific genes are more
932 derived than queen-specific ones. *Mol. Ecol.* **23**, 151–161 (2014).
- 933 16. Johnson, B. R. & Tsutsui, N. D. Taxonomically restricted genes are associated
934 with the evolution of sociality in the honey bee. *BMC Genomics* **12**, 164

- 935 (2011).
- 936 17. Simola, D. F. *et al.* Social insect genomes exhibit dramatic evolution in gene
937 composition and regulation while preserving regulatory features linked to
938 sociality. *Genome Res.* **23**, 1235–1247 (2013).
- 939 18. Harpur, B. a *et al.* Population genomics of the honey bee reveals strong
940 signatures of positive selection on worker traits. *Proc. Natl. Acad. Sci. U. S. A.*
941 **111**, 2614–9 (2014).
- 942 19. Rubin, B. E. R., Jones, B. M., Hunt, B. G. & Kocher, S. D. Rate variation in the
943 evolution of non-coding DNA associated with social evolution in bees. *Philos.*
944 *Trans. R. Soc. B Biol. Sci.* **374**, (2019).
- 945 20. Kapheim, K. M. Genomic sources of phenotypic novelty in the evolution of
946 eusociality in insects. *Curr. Opin. Insect Sci.* 1–9 (2015).
947 doi:10.1016/j.cois.2015.10.009
- 948 21. Dogantzis, K. A. *et al.* Insects with similar social complexity show convergent
949 patterns of adaptive molecular evolution. 1–8 (2018). doi:10.1038/s41598-018-
950 28489-5
- 951 22. Taylor, B. A., Reuter, M. & Sumner, S. Patterns of reproductive differentiation
952 and reproductive plasticity in the major evolutionary transition to
953 superorganismality. *Curr. Opin. Insect Sci.* (2019).
- 954 23. Toth, Amy L. Rehan, S. . Climbing the social ladder: The molecular evolution
955 of sociality. (2015).
- 956 24. Branstetter, M. *et al.* Genomes of the Hymenoptera. *Curr. Opin. Insect Sci.* **25**,
957 65–75 (2017).
- 958 25. Standage, D. S. *et al.* Genome, transcriptome and methylome sequencing of a

- 959 primitively eusocial wasp reveal a greatly reduced DNA methylation system in
960 a social insect. *Mol. Ecol.* **25**, 1769–1784 (2016).
- 961 26. Toth, A. L. *et al.* Shared genes related to aggression, rather than chemical
962 communication, are associated with reproductive dominance in paper wasps
963 (*Polistes metricus*). *BMC Genomics* **15**, 75 (2014).
- 964 27. Bluher, S. E., Miller, S. E. & Sheehan, M. J. Fine-scale population structure but
965 limited genetic differentiation in a cooperatively breeding paper wasp. *Genome*
966 *Biol. Evol.* (2020).
- 967 28. Rehan, S. M. *et al.* Conserved Genes Underlie Phenotypic Plasticity in an
968 Incipiently Social Bee. *Genome Biol. Evol.* **10**, 2749–2758 (2018).
- 969 29. Kocher, S. D. *et al.* The genetic basis of a social polymorphism in halictid
970 bees. *Nat. Commun.* **9**, (2018).
- 971 30. Shell, W. A. & Rehan, S. M. Social modularity: Conserved genes and
972 regulatory elements underlie caste-antecedent behavioural states in an
973 incipiently social bee. *Proc. R. Soc. B Biol. Sci.* (2019).
974 doi:10.1098/rspb.2019.1815
- 975 31. Kapheim, K. M. *et al.* Developmental plasticity shapes social traits and
976 selection in a facultatively eusocial bee. *Proc. Natl. Acad. Sci.* 202000344
977 (2020). doi:10.1073/pnas.2000344117
- 978 32. Linksvayer, T. A. & Johnson, B. R. Re-thinking the social ladder approach for
979 elucidating the evolution and molecular basis of insect societies. *Curr. Opin.*
980 *Insect Sci.* (2019). doi:10.1016/j.cois.2019.07.003
- 981 33. Taylor, D., Bentley, M. A. & Sumner, S. Social wasps as models to study the
982 major evolutionary transition to superorganismality. *Curr. Opin. Insect Sci.* **28**,

983 26–32 (2018).

984 34. Branstetter, M. G. *et al.* Phylogenomic Insights into the Evolution of Stinging
 985 Wasps and the Origins of Ants and Bees. *Curr. Biol.* **27**, 1019–1025 (2017).

986 35. Ferreira, P. G. *et al.* Transcriptome analyses of primitively eusocial wasps
 987 reveal novel insights into the evolution of sociality and the origin of alternative
 988 phenotypes Transcriptome analyses of primitively eusocial wasps reveal novel
 989 insights into the evolution of sociality. *Genome Biol.* **14**, R20 (2013).

990 36. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole
 991 genome comparisons dramatically improves orthogroup inference accuracy.
 992 *Genome Biol.* (2015). doi:10.1186/s13059-015-0721-2

993 37. Oi, C. A. *et al.* Effects of juvenile hormone in fertility and fertility-signaling in
 994 workers of the common wasp *Vespula vulgaris*. *PLoS One* **16**, 1–13 (2021).

995 38. Kelstrup, H., Hartfelder, K., Nascimento, F. S. & Riddiford, L. M. Frontiers in
 996 Zoology The role of juvenile hormone in dominance behavior , reproduction
 997 and cuticular pheromone signaling in the caste-flexible epiponine wasp ,
 998 *Synoeca surinama*. *Front. Zool.* **11**, (2014).

999 39. De Smet, F. *et al.* Balancing false positives and false negatives for the
 1000 detection of differential expression in malignancies. *Br. J. Cancer* (2004).
 1001 doi:10.1038/sj.bjc.6602140

1002 40. Illicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data.
 1003 *Genome Biol.* (2016). doi:10.1186/s13059-016-0888-1

1004 41. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E.
 1005 ScPred: Accurate supervised method for cell-type classification from single-cell
 1006 RNA-seq data. *Genome Biol.* (2019). doi:10.1186/s13059-019-1862-5

- 1007 42. Furey, T. S. *et al.* Support vector machine classification and validation of
1008 cancer tissue samples using microarray expression data. *Bioinformatics*
1009 (2000). doi:10.1093/bioinformatics/16.10.906
- 1010 43. Taylor, B. A., Cini, A., Wyatt, C. D. R., Reuter, M. & Sumner, S. The molecular
1011 basis of socially mediated phenotypic plasticity in a eusocial paper wasp. *Nat.*
1012 *Commun.* 2020.07.15.203943 (2020). doi:10.1101/2020.07.15.203943
- 1013 44. Vizan, P., Di Croce, L. & Aranda, S. Functional and pathological roles of
1014 AHCY. *Front. Cell Dev. Biol.* **9**, 654344 (2021).
- 1015 45. Hoffmann, K., Gowin, J., Hartfelder, K. & Korb, J. The scent of royalty: A P450
1016 gene signals reproductive status in a social insect. *Mol. Biol. Evol.* **31**, 2689–
1017 2696 (2014).
- 1018 46. Sheeja, C. C., Thushara, V. V. & Divya, L. Caste-Specific Expression of Na +
1019 /K + -ATPase in the Asian Weaver Ant, *Oecophylla smaragdina* (Fabricius,
1020 1775). *Neotrop. Entomol.* (2018). doi:10.1007/s13744-018-0598-3
- 1021 47. Iovinella, I. *et al.* Antennal protein profile in honeybees: Caste and task matter
1022 more than age. *Front. Physiol.* (2018). doi:10.3389/fphys.2018.00748
- 1023 48. Sumner, S. The importance of genomic novelty in social evolution. *Mol. Ecol.*
1024 **23**, (2014).
- 1025 49. Simões, J. M. The social brain : how social stimuli are translated into
1026 neuroendocrine signals. (University of Lisboa, 2014).
- 1027 50. Mariano, D. O. C. *et al.* Bottom-up proteomic analysis of polypeptide venom
1028 components of the giant ant *Dinoponera quadriceps*. *Toxins (Basel)*. (2019).
1029 doi:10.3390/toxins11080448
- 1030 51. Sasa, C., Miyazaki, S., Higashi, S. & Miura, T. Gene expressions for the

- 1031 sexually-dimorphic antennae in a ponerine ant. in (IUSSI).
- 1032 52. Yanay, C., Morpurgo, N. & Linial, M. Evolution of insect proteomes: Insights
1033 into synapse organization and synaptic vesicle life cycle. *Genome Biol.* (2008).
1034 doi:10.1186/gb-2008-9-2-r27
- 1035 53. Rehan, S. M., Berens, A. J. & Toth, A. L. At the brink of eusociality:
1036 transcriptomic correlates of worker behaviour in a small carpenter bee. *BMC*
1037 *Evol. Biol.* **14**, 260 (2014).
- 1038 54. Kronauer, D. J. & Libbrecht, R. Back to the roots: the importance of using
1039 simple insect societies to understand the molecular basis of complex social
1040 life. *Curr. Opin. Insect Sci.* **28**, 33–39 (2018).
- 1041 55. Noll, F. B., Wenzel, J. W. & Zucchi, R. Evolution of Caste in Neotropical
1042 Swarm-Founding Wasps (Hymenoptera : Evolution of Caste in Neotropical
1043 Swarm-Founding Wasps (Hymenoptera : Vespidae ; Epiponini). *Am. Museum*
1044 *Novit.* **3467**, 1–24 (2004).
- 1045 56. Richards, O. W. *The social wasps of the Americas.* (1978).
- 1046 57. Sumner, S., Bell, E. & Taylor, D. A molecular concept of caste in insect
1047 societies. *Curr. Opin. Insect Sci.* 42–50 (2017). doi:10.1016/j.cois.2017.11.010
- 1048 58. Kohannim, O. *et al.* Boosting power for clinical trials using classifiers based on
1049 multiple biomarkers. *Neurobiol. Aging* (2010).
1050 doi:10.1016/j.neurobiolaging.2010.04.022
- 1051 59. Montagna, T. S. & Antonialli, W. F. Morphological differences between
1052 reproductive and non-reproductive females in the social wasp *Mischocyttarus*
1053 *consimilis* Zikán (Hymenoptera: Vespidae). *Sociobiology* **63**, 693–698 (2016).
- 1054 60. Jeanne, R. L. Social Biology of the neotropical wasp, *Mischocyttarus drewseni*.

- 1055 *Bull. Museum Comp. Zool.* **144**, 63–150 (1972).
- 1056 61. Murakami, A. S. N., Shima, S. N. & Desuó, I. C. More than one inseminated
1057 female in colonies of the independent-founding wasp *Mischocyttarus*
1058 *cassununga* von Ihering (Hymenoptera, Vespidae). *Rev. Bras. Entomol.* **53**,
1059 653–662 (2009).
- 1060 62. Reeve, H. K. *Polistes*. in *The Social Biology of Wasps* (ed. Ross KG, M. R.)
1061 99–148 (Cornell University Press, 1991).
- 1062 63. Gelin, L. F. F. *et al.* Morphological Caste Studies In The Neotropical Swarm-
1063 Founding Polistinae Wasp *Angiopolybia pallens* (Lepeletier) (Hymenoptera :
1064 Vespidae). *Neotrop. Entomol.* **37**, 691–701 (2008).
- 1065 64. West-Eberhard, M. J. Temporary Queens in *Metapolybia* Wasps :
1066 Nonreproductive Helpers Without Altruism ? *Science* (80-.). **200**, 441–443
1067 (1978).
- 1068 65. Sakagami, S. F., Zucchi, R., Yamane, S., Noll, F. B. & Camargo, J. M. P.
1069 Morphological caste differences in *Agelaia vicina*, the neotropical swarm-
1070 founding polistine wasp with the largest colony size among social wasps
1071 (Hymenoptera: Vespidae). *Sociobiology* **28**, 207–223 (1996).
- 1072 66. V., B. M., Noll, F. B. & Zucchi, R. Morphological Caste Differences and Non-
1073 Sterility of Workers in *Brachygastera augusti* (Hymenoptera , Vespidae ,
1074 Epiponini), a Neotropical Swarm-Founding Wasp Author (s): *J. New York*
1075 *Entomol. Soc.* **111**, 242–252 (2003).
- 1076 67. Loope, K. J., Chien, C. & Juhl, M. Colony size is linked to paternity frequency
1077 and paternity skew in yellowjacket wasps and hornets. *BMC Evol. Biol.* **14**, 1–
1078 12 (2014).

- 1079 68. Hastings, M. D., Queller, D. C., Eischen, F. & Strassmann, J. E. Kin selection ,
1080 relatedness , and worker control of reproduction in a large-colony epiponine
1081 wasp , *Brachygastra mellifica*. *Behav. Ecol.* **9**, 573–581 (1998).
- 1082 69. Gobbi, N., Noll, F. B. & Penna, M. A. H. ‘Winter’ aggregations, colony cycle,
1083 and seasonal phenotypic change in the paper wasp *Polistes versicolor* in
1084 subtropical Brazil. *Naturwissenschaften* (2006). doi:10.1007/s00114-006-0140-
1085 z
- 1086 70. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for
1087 Illumina sequence data. *Bioinformatics* (2014).
1088 doi:10.1093/bioinformatics/btu170
- 1089 71. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq
1090 using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**,
1091 1494 (2013).
- 1092 72. DI Tommaso, P. *et al.* Nextflow enables reproducible computational workflows.
1093 *Nature Biotechnology* (2017). doi:10.1038/nbt.3820
- 1094 73. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-
1095 life scale using DIAMOND. *Nat. Methods* (2021). doi:10.1038/s41592-021-
1096 01101-x
- 1097 74. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
1098 DIAMOND. *Nat. Methods* **12**, 59 (2014).
- 1099 75. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-
1100 likelihood trees for large alignments. *PLoS One* (2010).
1101 doi:10.1371/journal.pone.0009490
- 1102 76. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq

- 1103 data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
- 1104 77. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2.
1105 *Nat. Methods* (2012). doi:10.1038/nmeth.1923
- 1106 78. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. *e1071: Misc*
1107 *Functions of the Department of Statistics (e1071), TU Wien. R package*
1108 *version* (2011).
- 1109 79. Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B. & Pintelas, P. E. Feature
1110 Selection for Regression Problems. *8th Hell. Eur. Res. Comput. Math. its Appl.*
1111 *HERCMA 2007* (2007). doi:10.1109/ICDM.2014.63
- 1112 80. Löytynoja, A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol.*
1113 (2014). doi:10.1007/978-1-62703-646-7_10
- 1114 81. Yang, Z. Paml: A program package for phylogenetic analysis by maximum
1115 likelihood. *Bioinformatics* (1997). doi:10.1093/bioinformatics/13.5.555
- 1116 82. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical
1117 and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
- 1118 83. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO.
1119 *Bioconductor Improv.* (2007).
- 1120 84. Götz, S. *et al.* High-throughput functional annotation and data mining with the
1121 Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–35 (2008).
- 1122 85. Piekarski, P. K., Carpenter, J. M., Lemmon, A. R., Lemmon, E. M. &
1123 Sharanowski, B. J. Phylogenomic evidence overturns current conceptions of
1124 social evolution in wasps (vespidae). *Mol. Biol. Evol.* **35**, 2097–2109 (2018).
- 1125 86. O'Donnell, S. Reproductive caste determination in eusocial wasps
1126 (Hymenoptera: Vespidae). *Annu. Rev. Entomol.* **43**, 323–46 (1998).

- 1127 87. Matsuura, M. & Yamane, S. *Biology of the vespine wasps*. (Springer Verlag,
1128 1990).
- 1129 88. Menezes, R. S. T., Lloyd, M. W. & Brady, S. G. Phylogenomics indicates
1130 Amazonia as the major source of Neotropical swarm-founding social wasp
1131 diversity. *Proc. R. Soc. B* (2020).
1132
1133

1134

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

cingulata, *Angiopolybia pallens*, *Polybia quadricincta*, *Agelaia cajennensis*,
Brachygastera mellifica, *Vespa crabro* and *Vespula vulgaris* (see Supplementary
Methods for further details of species choice). Where data on evidence of
morphological castes was not available from the literature, we conducted
morphometric analyses of representative queens and workers from several colonies
per species (see Supplementary Methods; Supplementary Table S8). Image credits:
M. basimacula (Stephen Cresswell). *A. cajennensis* (Gionorossi; Creative Commons);
V. vulgaris (Donald Hobern; Creative Commons). *V. crabro* (Patrick Kennedy); *P.*
canadensis; *M. cingulata*, *A. pallens*, *P. quadricincta*, (Seirian Sumner), *B. mellifica*
(Amante Darmanin; Creative Commons).

Fig. 2 | Principal component analyses of orthologous gene expression before and after between-species normalisation. a) Principal component analyses performed using log2 transcript per million (TPM) gene expression values. This analysis used single-copy orthologs (using Orthofinder), allowing up to three gene isoforms in a single species to be present, whereby we took the most highly expressed to represent the orthogroup, as well as filtering of orthogroups which have expression below 10 counts per million. **b)** Principal component analysis of the species-normalised and scaled TPM gene expression values using same filters as (a). Caste denoted by purple (queen) or blue (worker). Each species has a different shape/symbol.

Fig. 3 | Overlap of differential caste-biased genes (queen vs worker) and their functions. **a)** Heatmap showing the differential genes that are caste-biased in at least four species (identified using edgeR) using the orthologous genes present in the nine species. In cases where three isoforms exist for a single species orthogroup, only the isoform with greatest expression was considered, plus at least 2 species could have missing gene data for each orthogroup. Listed for each species (in brackets), is the total number of orthologous differentially expressed genes per species. *Metapolybia* Blast hits are listed along with orthogroup name. **b)** Gene ontology histogram of overrepresented terms of genes found differentially expressed

1192 in at least four out of the nine species (n=57 genes; in either queen or worker [not
1193 both]), using a background of all genes tested across the nine species. *P* values are
1194 single-tailed and bonferroni corrected. **c)** Gene ontology for orthogroups with at least
1195 two species with caste-biased gene expression in the same direction (n=353 genes).

Fig. 4 | A genetic toolkit for social behaviours across eusocial wasps. a)
Change in certainty of correct caste classifications through progressive feature

1199 selection using Support Vector Machine (SVM) approaches. Models were trained on
 1200 eight species and tested on the ninth species. Features (a.k.a. orthologous genes)
 1201 were sorted by linear regression with regard to caste identity within the 8 training
 1202 species, beginning at 99% where almost all genes were used for the predictions of
 1203 caste, to 1% where only the top one percent of genes from the linear regression
 1204 (sorted by P value) were used to train the model. '1' equates to high queen
 1205 classification certainty. This test was repeated using various filters, showing: LEFT:
 1206 use of only 1to1 orthogroups, where no gene isoforms or missing values were
 1207 accepted into the model, MIDDLE: allowing three gene isoforms (using the most
 1208 highly expressed as the representative) and RIGHT: allowing three gene isoforms
 1209 and missing values (NAs; up to 2 species with no ortholog), filling in with non-caste
 1210 biased gene values for queen and worker. Horizontal guide markers are placed at
 1211 0.4, 0.5 and 0.6 respectively indicating the confidence score for classifying queens,
 1212 and the vertical dotted line marks the percentage at which the regression score
 1213 becomes < 0.05 . **b)**: Histograms indicate the Bonferroni-corrected enriched gene
 1214 ontology (GO) terms for the top orthologous genes from the corresponding model
 1215 above (linear regression P value < 0.05 ; shown as dotted line in upper plots), tested
 1216 against a background of all genes used in the corresponding SVM model (i.e. with a
 1217 single gene representative for all species in the test). P values are single-tailed. **c)**
 1218 Heatmap of the top 50 genes after sorting by regression, in SVM with 3 merged
 1219 isoforms and 2 NAs; showing species-normalised gene expression levels in the nine
 1220 species for queen (Q) and worker (W) samples (where red indicated down-regulation
 1221 and blue up; Row Z-score), orthogroup name and top *Metapolybia* BLAST hit.

1222

Fig.5 | Testing for the presence of a distinct 'simple society toolkit' and a

'complex society toolkit'. a): Using either the four species with the simpler

societies or the four with the more complex societies, we trained SVM models, using

progressive filtering of genes (based on the linear regression); we then tested how

well these models classified caste for the opposite set of species. UPPER panel:

training set was the four simpler species, and the test set was the four more complex

species. LOWER panel: training set was the four more complex species, and the test

set was the four simpler species. Likelihood of being a queen from 0 to 1 is plotted

for each species across the progressively filtered sets. The total number of

orthologous genes used in each SVM model are shown (bottom left of each panel); the total number of these genes that were significant predictors of caste (p value < 0.05) is listed at the dotted line. For each test (pair of Queen/Worker in each species), the SVM model was run using genes with only 1 homologous gene copy per species. **b)** Overlap of significant genes among the three putative toolkits, comparing those genes for 'simpler species' (Fig 5a UPPER panel), 'more complex species' (Fig 5a LOWER panel), and 'full spectrum' (Fig 4a- Left) toolkits. For each experiment, the number of genes (orthogroups) tested and the total number that were significant after linear regression are given outside of the Venn diagram. Inside the Venn diagram are the numbers of these significant genes shared (or not) between the three groups. Significant overlap is shown using hypergeometric tests (one-tailed). The colours represent the genes that predict caste in: the more complex societies (blue), the simpler societies (grey) and across the full spectrum (pink). There is significant overlap between the simpler and the more complex genes with the full spectrum set, but not between the simpler and complex societies, suggesting that there may be distinct differences in the genes regulating castes at different levels of social complexity. **c)** Enriched gene ontology terms (TopGO) using a background of all genes tested in each individual experiment, using an uncorrected single-tailed p values.

Supplementary Figures

Supp. Figure. 1 | Phylogenetic tree of social wasps used in this study, and

related hymenopterans generated using Orthofinder (SpeciesTree_rooted.txt).

Drosophila was chosen as the root of the species tree (not shown), with two representative ants (*Dinoponera quadricaps* and *Solenopsis invicta*) and two bees (*Apis mellifera* and *Nomia melanderi*). Colours show groupings of ants, bees and wasps (Vespinae or Polistinae). For the nine wasp species (this study) we have sequenced adult caste-specific brain transcriptomic data (queen and worker). As expected: Vespinae are clearly separated from the Polistinae; *Angiopolybia* is basal to the other Epiponini; independent-founding, non-superorganismal wasps (*Polistes* and *Mischocyttarus*) are basal to the Polistinae^{85,88}.

1270

1271

Supp. Figure. 2 | SVM kernel type iterations. For each of the four kernel types, a

1272

filtered SVM plot shows the leave one out experiment. Each coloured line represents

1273

the prediction of the queen sample at various regression filters, from 99% data

1274

inclusion to 1% inclusion of only the top genes in the regression analysis. Species

1275

are listed in their Genus (capital letter) and species (first three lower case letters of

1276

each species).

Supp. Figure 3 | dN/dS. a) All orthogroups that had at least one gene significantly under selection are listed (branch-site models). b) Overlap of positive selection

1280 genes and various toolkit gene lists. Top, shows the differential gene list overlap.
1281 Middle, shows the SVM gene list <0.05 for singleton orthogroups. Bottom, shows the
1282 three gene isoform merge and 2 NA gene list. The number of genes overlapping are
1283 in the middle and in each sphere shows the total number of genes compared for
1284 each list (given that the two backgrounds do not overlap). A mutual background is
1285 listed showing the total number of genes that were possible to have been significant
1286 in the two compared lists.

1287

1288

1289

1290

Supp. Figure. 4. | SVM predictions using founding behaviour and phylogeny to subsets of the training data. SVM predictions are given for each single species given training on species listed to left, showing the prediction after progressive feature selection from 95 to 1% of genes remaining after selection. Numbers of genes in each test are indicated in the lower left part of each plot. **a)** Training with independent-founders only, tested on the four swarm-founders (see Figure 1 for species). **b)** Training with swarm-founders and testing on the four independent-founders. **c)** Training with Polistine species only, and testing on the Vespines (*Vespa* and *Vespa*). The reverse was not possible, as the SVM requires a minimum number of samples to work.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supp.Table.S1RNAseqsamples.xlsx](#)
- [Supp.Table.S2.OrthologyplusMerge.xlsx](#)
- [Supp.Table.S3.OrthogroupDEGs.v2.xlsx](#)
- [Supp.Table.S4.NormalisedSVMandresult.xlsx](#)
- [Supp.Table.S5.SVMSuperNotSuper.xlsx](#)
- [Supp.Table.S6.SVMNestTypePolistinae.xlsx](#)
- [Supp.Table.S7.DnDs.xlsx](#)
- [Supp.Table.S8Morphometricdataresults.xlsx](#)