Genetic toolkit for sociality predicts castes across the spectrum of social complexity in wasps

Christopher D. R. Wyatt[1*], Michael Bentley[1,†], Daisy Taylor[1,2,†], Ryan E. Brock[2,3], Benjamin A. Taylor[1], Emily Bell[2], Ellouise Leadbeater[4] & 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

Seirian Sumner

Centre for Biodiversity and Environment Research,

University College London, London, UK.

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

Key words: Superorganismality, Major evolutionary transitions, Castes, Wasps,

Abstract

Major evolutionary transitions describe how biological complexity arises; e.g. in evolution of complex multicellular bodies, and superorganismal insect societies. Such transitions involve the evolution of division of labour, e.g. as queen and worker castes in insect societies. Castes across different evolutionary lineages are thought to be regulated by a conserved genetic toolkit. However, this hypothesis has not been tested thoroughly across the complexity spectrum of the major transition. Here we reveal, using machine learning analyses of brain transcription, evidence of a shared genetic toolkit across the spectrum of social complexity in Vespid wasps. Whilst molecular processes underpinning the simpler societies (which likely represent the origins of social living) are conserved throughout the major transition, additional processes appear to come into play in more complex societies. Such fundamental shifts in regulatory processes with complexity may typify other major evolutionary transitions, such as the evolution of multicellularity.

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 individuals[1]. 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 levels 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 superorganismality[3–7]. An outstanding question is to what extent are genomic mechanisms operating *across levels* of complexity in the major transition – from simple to complex sociality – conserved[8]? A lack of data from representatives across any one lineage of the major transition have limited our ability to address this question.

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-reproductive roles, known as queens and workers respectively in the case of insect societies. An overarching mechanistic hypothesis for social evolution is that the repertoire of behaviours typically exhibited in the life cycle of the solitary ancestor were uncoupled to produce a division of labour among group members with individuals specialising in either the reproductive ('queen') or provisioning ('worker') phases of the solitary ancestor[9]. Such phenotypic decoupling implies that there will be a conserved mechanistic toolkit that regulates queen and worker phenotypes in species representing different levels of social complexity across the spectrum of the major transition (reviewed in[10]). An alternative to the shared toolkit hypothesis is that the molecular processes regulating social behaviours in non-superorganismal societies (where caste remains flexible, and selection acts primarily on individuals) differ fundamentally from those processes that regulate social behaviours in superorganismal societies [11,12]. Phenotypic innovations across the animal kingdom have been linked to genomic evolution: taxonomically-restricted genes[13–16], rapid evolution of proteins[17,18] and regulatory elements[17,19] been found in most lineages of social insects[20]. Indeed, some recent

76 studies have suggested that the processes regulating different levels of social complexity
77 may be different[17,19,21]. The innovations in social complexity and the shift in the unit of
78 selection (from individual- to group-level[22]) that accompany the major transition may
79 therefore be accompanied by genomic evolution, throwing into question whether a universal
80 conserved genomic toolkit regulates social behaviours across the spectrum of the major
81 transition[8]. The roles of conserved and novel processes are not necessarily mutually
82 exclusive; novel processes may coincide with phenotypic innovations, whilst conserved
83 mechanisms may regulate core processes at all stages of social evolution.

84 Currently, data are largely limited to species that represent either the most complex –
85 superorganismal - levels of sociality (e.g. ants or honeybees[23]), or the simplest levels of
86 social complexity as non-superorganisms that likely represent the first stages in the major
87 transition (e.g. *Polistes* wasps[7,24–26] and incipiently social bees[27–30]). We lack data on the
88 intermediary stages of the major transition and thus lack a comprehensive analysis of if and
89 how molecular mechanism change *across* any single evolutionary transition to
90 superorganismality. One exception is a recent study that identified a core gene set that
91 consistently underlie caste-differentiated brain gene expression across five species of ants[5];
92 however, this study lacked ancestrally non-superorganismal representatives (one species
93 had secondarily lost the queen caste but evolved from a superorganismal ancestor[7]).

94 A promising group for exploring these questions are the social wasps[31], with some
95 1,100 species exhibiting the full spectrum of sociality. We generated brain transcriptomic
96 data of caste-specific phenotypes for nine species of social wasps, representing a range of
97 levels of social complexity in the transition to superorganismality (Fig. 1). Using machine-
98 learning algorithms we exploited these datasets to determine whether there is a conserved
99 genetic toolkit for social behaviour across the major transition from non-superorganismal
100 (simple) to superorganismal (complex) species within the same lineage (Aim 1). We then
101 further interrogate these data to identify whether there are any key discernible differences in

102 the molecular bases of social behaviour in the simpler versus the more complex societies
103 (Aim 2). Accordingly, we provide the first evidence of a conserved genetic toolkit across the
104 spectrum of the major transition to sociality in wasps; we also reveal novel insights into the
105 molecular patterns and processes at a key transitional point of the major transition from
106 simple sociality to complex superorganismality.

107

108 Results

109 We chose one species from each of nine different genera of social wasps representing the
110 full spectrum of social diversity within the Polistinae and Vespinae (see Figure 1;
111 Supplemental Table S1). For each species, we sequenced RNA extracted from whole brains
112 of adults to construct *de novo* brain transcriptomes for the two main social phenotypes –
113 adult reproductives (defined as mated females with developed ovaries, henceforth referred
114 to as 'queens' for simplicity; see Supplementary Table S2) and adult non-reproductives
115 (defined as unmated females with no ovarian development, henceforth referred to as
116 'workers'; see Supplementary Methods). Using these data we could reconstruct a
117 phylogenetic tree of the Hymenoptera using single orthologous genes (Orthofinder[32]),
118 resulting in expected patterns of phylogenetic relationships (Supp. Fig. 1). This dataset
119 provides coverage across the spectrum of the major transition to sociality (see Fig. 1;
120 Supplementary Table S1), and provides us with the opportunity to test the extent to which
121 the same molecular processes underpin the evolution of social phenotypes across the
122 spectrum of the major transition to superorganismality in wasps.

123

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

We found several lines of evidence of 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 individual-level gene expression variation was species identity (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; *Agelaia* showed little caste-specific separation; and *Vespa* phenotypes clustered in the opposite direction to phenotypes in the other species. These initial data visualisations suggest that these species may not share the same caste-specific patterns as the other species, but we cannot rule out data and/or sampling anomalies, especially since gynes (unmated, newly-emerged queens) were included in the queen sample for *Vespa*.

Analyses of orthologous genes found in all nine species (Supplementary Table S3) revealed sets of caste-biased orthologous genes among the nine species; however, no orthologous DEGs showed consistent caste-biased differential expression across all nine wasp species (Fig. 3a; Supplementary Table S4; using unadjusted <0.05 p values). Depsite this, notable

149 signatures of caste regulation were apparent, across the species set: e.g. orthogroup
150 OG0002698 was differentially express across six of the nine species and is predicted to
151 belong to the vitellogenin gene family (79.0% identity; using the *Metapolybia* protein
152 sequence to represent the orthogroup), a well-known regulator of social behaviour in
153 insects[33]. When the analysis was limited to caste-specific DEGs found in at least two species
154 (n=95; Supplementary Table S4), there was overrepresentation of catabolic and metabolic
155 GO terms (Fig. 3b; Supplementary Table S4).

156

157 A toolkit of many genes with small effects predicts caste across the spectrum
158 of sociality in wasps

159 Conventional differential expression analyses (e.g. edgeR) require a balance of *P* value cut-
160 offs and fold change requirements to reduce false-positive and false-negative errors[34].
161 Therefore, consistent patterns of many genes with smaller effect sizes may be missed when
162 applying strict statistical measures. Support vector machine (SVM) learning approaches use
163 a supervised learning model capable of detecting subtle but pervasive signals in differential
164 expression between conditions (e.g. for classification of single cells[35,36], cancer cells[37] and in
165 social insect castes[38]). We used this approach to test whether gene expression can
166 successfully classify caste identity for unknown samples; accurate classification of samples
167 as queens or workers based on their global transcription patterns would be evidence for a
168 genetic toolkit underpinning social phenotypes.

169 Starting with a "leave-one-out" SVM approach, we attempted to classify samples of a test
170 species as queens or workers, using a predictive gene set generated from a model trained
171 on caste-specific gene expression from eight of the nine species, with the ninth species
172 being the test sample. The analysis was repeated until each of the species had been 'left
173 out' and their caste classification tested. Using 3486 single copy orthologues, and removing

174 orthogroups with low expression (n=2020), we could filter the matrix by progressive feature
175 selection (based on linear regression, to refine the gene sets to those that are informative;
176 see Suppl. Methods), which reduces the number of genes used in the SVM, focussing on
177 those genes informative for caste. When testing each left-out species, we largely attain
178 accurate caste predictions for seven of the nine species, across most feature selection filters
179 (Fig. 4a; > 0.5 likelihood in queen sample); the same two outlier species from Fig. 2 (*Agelaia*
180 and *Brachygastra*) showed generally lower predictions of queen likelihood for the queen
181 sample (<0.5). This suggests that many hundreds of genes may be caste-biased to some
182 degree.

183 Within the SVM model of nine species (Fig. 4), we found 400 significant orthologs (genes)
184 after feature selection with a *P* value of less than 0.05 (Supplementary Table S5; top 53
185 genes (p <0.001) shown in Fig. 4b). These 400 genes were enriched for neural vesicle
186 transport related signalling functions (Fig. 4c; Supplementary Table S5), and may form the
187 most important constituents of a shared toolkit for social behaviour across non-
188 superorganismal and super-organismal social wasps.

189 Using Gene Set Enrichment Analysis (GSEA), we could compare the genes discovered in
190 the two methods (edgeR and SVM), finding enrichment in the gene sets identified as
191 important for social behaviour (Supplementary Figure 2). However, only ten genes were
192 identified as significantly caste-biased in both methods (Supplementary Table S6). Of these,
193 some have previously been identified as having relevance to social evolution and caste
194 differentiation; these include Vitellogenin (mentioned earlier; OG0002698) and Cytochrome
195 P450 (OG0000434)[10,39], thought to be involved in chemical signalling between castes and
196 associated with expression of juvenile hormone[39]. Further, UDP-glucuronosyltransferase 2C-
197 like (OG0001554), downregulated in virgin versus mated fire ant queens[40]; esterase E4-like
198 (OG0000645) upregulated in young honeybee queens compared to nurses at the proteomic
199 level[41]; neprilsin-1 (OG0004128) is differentially expressed in major/minor *C. floridanus*

200 workers and after caste reprogramming[42], which could be involved in caste memory
201 formation[43]. There are also other genes of interest, which to our knowledge have not
202 previously associated with caste, including Toll-like receptor 8 (OG0002639) (see
203 Supplementary Table S6).

204

205 Aim 2: Are there different fine-scale toolkits that reflect different levels of social 206 complexity?

207 To explore differential patterns *within* the conserved predictive toolkit for caste differentiation
208 identified in Aim 1, we trained an SVM model using the four species with the simplest
209 societies (*Mischocyttarus, Polistes, Metapolybia* and *Angiopolybia*) as representatives of the
210 earlier stages in the major transition (see Fig. 1), and tested how well this gene set classified
211 castes for the four species with the more complex societies as representatives of the later
212 and superorganismal stages of the major transition (*Polybia, Agelaia, Vespa* and *Vespula;*
213 see Fig. 1). *Brachygastra* was excluded due to its poor performance overall (see Fig. 4) and
214 to ensure we compared the same number of training sets in each case. If castes in the test
215 species classify well, this would suggest that the processes regulating castes in the simpler
216 societies are also important in the more complex societies (i.e. there is no specific toolkit for
217 simple sociality, which is then lost in the evolution of social complexity). Conversely, if the
218 test species do not classify well, this would suggest that there are distinct processes
219 regulating caste in the simpler societies that are lost (or become less important) in the
220 evolution of more complex forms of sociality.

221 The putative toolkit for castes in the simplest societies consisted of ~1021 genes after
222 feature selection (Supplementary Table S7 [Simple]). *Vespula* and *Polybia queens* classified
223 extremely well (Fig. 5-upper); importantly, classifications for both these species improved
224 with progressive feature selection. *Vespa* classified correctly but less well (likely because the

queens included gynes); *Agelaia* classified to the wrong caste (consistent with results from Aim 1). Overall, 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.

We next conducted the reciprocal analysis, training the SVM using the four species with the more complex societies (*Polybia, Agelaia, Vespa* and *Vespula*) and testing it on the four species with simpler societies (*Mischocyttarus, Polistes, Metapolybia* and *Angiopolybia)*. The toolkit for castes in these more complex societies was much smaller than the one for simple societies, consisting of ~464 genes after feature selection (Supplementary Table 7 [Complex]), possibly due to the greater taxonomic distances involved (inc. Polistinae and Vespinae). This putative toolkit for castes in more complex societies was less successful in classifying castes for the simpler societies (Fig. 5a-lower), than the reciprocal analysis (above; Fig. 5a-upper): although two species classified in the right direction (*Metapolybia* and *Angiopolybia),* their classifications have much lower confidence than in the reciprocal test; furthermore, for the two simplest societies, *Polistes* queens were classified close to 0.5 (meaning the gene sets were uncertain between queen/worker) and *Mischocyttarus* classified in the wrong direction (Fig. 5a-lower). These results raise the interesting idea that the processes regulating caste differentiation in species with more complex societies may be unimportant (or absent) in the simpler societies. In further support of this, the putative 'simple society toolkit' overlapped to a greater extent with the overall toolkit found across all species (Fig. 4) than those of the putative 'complex society toolkit' (Fig. 5b), hypergeometric overlap shown for both comparisons). Gene ontology results are similar between the two sets, and are composed of synaptic and membrane related terms (Fig. 5c; Supplementary Table 8); however the 'simple society toolkit' contains enrichment for metabolic/cellular respiration and ion/cation transport which are missing in the 'complex society toolkit'.

We conducted additional tests to determine whether other factors could better explain the

molecular basis of caste, besides level of social complexity, and to verify that our reciprocal SVM approach was valid given the small sample sizes. Using the same reciprocal SVM approach, we found that the molecular basis of a key life-history trait - nest founding behaviour - are largely conserved across species (Supplementary Figure 3; Supplementary Table 7[Swarm/Independent], Supplementary Table 8). From a biological perspective this suggests there is no specific genomic innovation associated with this life-history innovation that interacts with caste, as caste was correctly predicted in all species, with the exception of *Agelaia*. From a methodological perspective this indicates that the SVMs can perform well even using this small number of species unlikely to be affecting the performance of our social complexity SVM. Likewise, we tested for an effect of phylogeny, testing how well castes in the Vespines (*Vespa, Vespula*) classified using a putative Polistinae caste toolkit as the training set; there was little influence of subfamily on performance of the SVMs, with queens and workers being classified with 70-80% confidence (Supplementary Figure 3; Supplementary Table 7; the reverse of this test could not be performed due to low sample sizes for a Vespine training set). This suggests that the genes important for caste identity are shared across these two subfamilies.

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

276 hundreds of genes of small effect (Fig. 4). However, in sub-setting the data by level of social
277 complexity, two important new insights are revealed. Firstly, there appears to be a putative
278 toolkit for castes in the simpler societies that largely persists across the major transition,
279 through to superorganismality. Secondly, different (additional) processes appear to become
280 important at more complex levels of sociality. Further sampling is required to determine the
281 extent to which the role of these additional processes is driven by the evolution of
282 superorganismality, and the point of no return in the major transition to sociality.

283

284 The first important finding is that we identified a substantial set of genes that consistently
285 classify caste across most of the species, irrespective of the level of social complexity. The
286 taxonomic range of samples used meant we were able to confirm that specific genes are
287 consistently differentially expressed, with respect to caste, across the species. These
288 patterns would be difficult to detect if only looking at a few species, species across several
289 lineages, or species representing only a limited range of social complexity. In addition to
290 typical caste-biased molecular processes, we also identified that genes related to synaptic
291 vesicles are different between castes; this is interesting as the regulation of synaptic vesicles
292 affects learning and memory in insects[44]. To our knowledge, this is the first evidence of what
293 may be a conserved genetic toolkit for sociality, from the first stages of social living to true
294 superorganismality, including intermediate stages of complexity, which putatively represent
295 different points in the major transition. Greater taxonomic sampling will allow further
296 exploration of how these genes and their regulation change across the major transition, and
297 help recover the full spectrum of genes that may have been important in the evolution of
298 sociality.

299 The underlying assumption, based on the conserved toolkit hypothesis, has been that
300 whatever processes regulate castes in complex societies must also regulate castes in
301 simpler societies. Unexpectedly, our analyses suggest there may be additional molecular

302 processes underpinning castes that become important in the more complex levels of
303 sociality. The predictive gene set identified in the SVM trained on more complex species
304 performed less well in classifying caste than the predictive tool kit derived from the simpler
305 species. There may be fundamental differences discriminating (near) superorganismal
306 societies from non-superorganismal societies. This highlights the importance of examining
307 different stages in the major transition when attempting to elucidate its patterns and
308 processes.

309 There were two consistent outlier species in every stage of our analyses: *Agelaia* and
310 *Brachygastra.* Although we cannot rule out issues with the data, all samples underwent the
311 same rigorous QC testing at the lab, sequencing and bioinformatics stages and so are
312 unlikely to fully explain these patterns. Another explanation is that they are genuinely
313 biologically different to the other species. One of the most profound phenotypic innovations
314 in social insect evolution is when caste becomes irreversibly committed during
315 development[11,22,45]; this has been referred to as 'the point of no return' in evolutionary terms,
316 as once a society is comprised of workers and queens who are mutually dependent on each
317 other for colony function (like different cogs in the same machine), it is difficult to revert to
318 independence[12]. After this point, the society can be considered as a definitive superorganism
319 – with a new level of individuals and unit of selection[12]. Intriguingly, these two species are
320 putatively at this point in the transition to super-organismality (Fig. 1). *Vespa* also failed to
321 classify well in some analyses, but this is likely explained by the fact that the sample of
322 queens included some gynes (unmated newly-emerged future-queens). Our morphometric
323 analyses of *Brachygastra* (Supplementary Table S1) detected possible evidence of pre-
324 imaginal caste determination, suggesting it is on the cusp of becoming superorganismal.
325 Similarly, subtle differences in morphology among queens and workers of *Agelaia* suggests
326 they too may have some level of pre-imaginal caste determination[46]. We speculate that the
327 evolution of irreversible caste commitment (in superorganisms) is accompanied by a

328 fundamental shift in the underlying regulatory molecular machinery such that species
329 undergoing the transition to superorganismality may have to rewire the core set of genes
330 involved in regulating caste.

331

332 Despite being able to extract consistent SVM predictions, our models are only as good as
333 the initial data used to train them. Our study suffers from a few limitations. Firstly, the sample
334 size (number of species) is relatively low; SVMs are generally used on very large datasets
335 such as clinical trials in the medical sciences[47]. Although our models did perform well, the
336 analyses would be more robust by using more species in the training datasets. Indeed, we
337 observed reduced performance in our model predictions when fewer species were included
338 in the training set. Secondly, by comparing across multiple species, we can only train our
339 model on genes that have a single representative isoform per species in each separate test.
340 This reduces the numbers of genes we can test in each SVM model, especially where more
341 distantly related species are included. We overcame this limitation by merging gene isoforms
342 within the same orthogroup (potential gene duplications), yet this comes with some
343 additional costs as some genes are discarded in this process. Finally, genomes are not
344 available for most of the species we tested; our measurements are based on *de novo*
345 sequenced transcriptomes, which potentially contain misassembled transcripts, which could
346 reduce the ability to find single copy orthologs across species. For these reasons, the
347 numbers of genes detected in our putative toolkit for sociality is likely to be conservative and
348 modest (potentially by several fold). These limitations are likely to apply to many similar
349 studies, due to the difficulty and expense of obtaining high quality genomic data for specific
350 phenotypes for non-model organisms. Our study illustrates the power of SVMs in detecting
351 large suites of genes with small effects, which largely differ from those identified from
352 conventional differential expression analysis[38]. We advocate the use of the two methods in
353 parallel: our conventional analyses suggested that metabolic genes appear to be responsible

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

360 In conclusion, our analysis of brain transcriptomes for castes of social wasps suggest that
361 the molecular processes underpinning sociality are conserved throughout the major
362 transition to superorganismality. However, additional processes may come into play in more
363 complex societies, putatively driven by selection happening at the point-of-no-return, where
364 societies transition to become committed superorganisms. Importantly, this suggests there
365 may be fundamental differences in the molecular machinery that discriminates
366 superorganismal societies from non-superorganismal societies. The evolution of irreversible
367 caste commitment (in superorganisms) may require a fundamental shift in the underlying
368 regulatory molecular machinery. Such shifts may be apparent in the evolution of sociality at
369 other levels of biological organisation, such as the evolution of multicellularity, taking us a
370 step closer to determining whether there is a unified process underpinning the major
371 transitions in evolution.

372

373 Acknowledgements

374 We would like to thank Laura Butters for her help with the morphometric analyses of
375 *Brachygastra*, James Carpenter at the American Natural History Museum and Christopher K.
376 Starr for confirming species identification, and the Bristol Genomics Facility for their
377 assistance with library preparations and sequencing. This work was conducted under
378 collection and export permits for Trinidad (Forestry Division Ministry of Agriculture: #001162)

and Panama (Autoridad nacional del ambiente (ANAM) SE/A-55-13). It was funded by the Natural Environment Research Council (NE/M012913/2; NE/K011316/1) awarded to SS, and a NERC studentship and Smithsonian Tropical Research Institute pre-doctoral fellowship awarded to 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;* 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[48–51]. 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 individual[21]. 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 differentiation[52]. As examples of species for which there is little or no evidence of developmental (morphological) caste determination, we chose *Angiopolybia pallens* which is phylogenetically basal in the Epiponines[53,54] and *Metapolybia cingulata (*Fabricius)[53,54]. We confirmed the lack of clear caste allometric differences in *M. cingulata* as data were lacking (see Morphometrics methods (below) and Supplementary File S1).

As examples of species showing subtle, colony-stage-specific caste allometry, we chose a species of *Polybia.* The social organisation of *Polybia* spp is highly variable, ranging from complete absence of morphological queen-worker differentiation[55]. *Polybia quadricincta* is a relatively rare (and little studied) epiponine wasp which can be found across Bolivia, Brazil,

Columbia, French Guiana, Guyana, Peru, Suriname and Trinidad (Richards, 1978). Our morphometric analyses found some evidence of subtle allometric morphological differentiation in this species, but with variation through the colony cycle (Supplementary File S1); this suggest it is a representative species for the evolution of the first signs of pre-imaginal caste differentiation.

Many species of the genera *Agelaia* and *Brachygastra* appear to show pre-imaginal caste determination with allometric morphological differences between adult queens and workers[53]. We chose one species from each of these genera as representatives of the most socially complex Polistine wasps. Although no morphological data were available for *Agelaia cajennensis* (Fabricius) all species of *Agelaia* studied show some level of preimaginal caste determination[53,56]. *Brachygastra* exhibit a diversity of caste differentiation[53,57]; our morphological analysis of caste differentiation *B. mellifica* confirms that this species is highly socially complex, with large colony sizes[53] and pre-imaginal caste determination resulting in allometric caste differences (Supplementary File S1).

All species of Vespines are independent nest founders and superorganisms, with a 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 mating[58], which have likely evolved under different selection pressures to the major transition itself[59]. The European hornet, *Vespa crabro,* 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* with larger colony sizes than *Vespa*[58], suggesting a more complex level of social organisation.

432 ## Sample collection

433 Where possible, we sampled from colonies representing different stages in the colony cycle,
434 as caste differentiation can vary as the colony matures in some species (Supplementary File
435 S2). *Metapolybia cingulata* (6 colonies), *Polistes canadensis* (3 colonies), *Agelaia*
436 *cajennensis* (1 colony) and *Mischocyttarus basimacula basimacula* (3 colonies) were
437 collected from wild populations in Panama in June 2013. *Brachygastra mellifica* (4 colonies)
438 were collected from populations in Texas, USA in June 2013. *Angiopolybia pallens* (2
439 colonies) and *Polybia quadricincta* (2 colonies) were collected from Arima Valley, Trinidad in
440 July 2015. *Vespa crabro* (4 colonies) and *Vespula vulgaris* (4 colonies) were collected from
441 various locations in South East England, UK in 2017. Queens and workers were collected
442 directly from their nests during the daytime, placed immediately into RNA*later* (Ambion,
443 Invitrogen) and stored at -20°C until further use. An exception was that gynes (newly-
444 emerged, unmated queens) in addition to queens were used for *V. crabro* due to difficulty in
445 obtaining samples of mature queens. Samples were ultimately pooled *within castes* for
446 bioinformatics analyses, such that each informatic pool consisted of 3-6 individual brains
447 from wasps sampled across 2-4 colonies to capture individual-level and colony-level
448 variation in gene expression (see Supplementary Table S2). Samples of *M. cingulata, A.*
449 *cajennensis, M. basimacula and B. mellifica* were sent to James Carpenter at the American
450 Natural History Museum for species verification. *A. pallens* and *P. quadricincta* were
451 identified by Christopher K. Starr, at University of West Indies, Trinidad and Tobago.

452 ## Morphometrics

453 Data on morphological differentiation among colony members (and thus information on
454 whether pre-imaginal (developmental) caste determination was present) was lacking for *M.*
455 *cingulata, P. quadricinta* and *B. mellifica*; therefore, we conducted morphometric analyses on
456 these three species in order to ascertain the level of social complexity. Morphometric
457 analyses were carried out using GXCAM-1.3 and GXCapture V8.0 (GT Vision) to provide

458 images for assessing morphology. We measured 7 morphological characters using ImageJ
459 v1.49 for queens and workers for each species. The body parts measured were: head length
460 (HL), head width (HW), minimum interorbital distance (MID), mesoscutum length (MSL),
461 mesoscutum width (MSW), mesosoma height (MSH) and alitrunk length (AL) (for
462 measurement details, see [60]). Abdominal measurements were not recorded as ovary
463 development could alter the size of abdominal measurements, therefore biasing the results.
464 The morphological data were analysed to determine whether the phenotypic classification,
465 as determined from reproductive status, could be explained by morphological differences.
466 ANOVA was used to determine size differences between castes for each morphological
467 characteristic. A linear discriminant analysis was also employed to see if combinations of
468 characters were helpful in discriminating between castes. The significance of Wilks' lambda
469 values were tested to determine which morphological characters were the most important for
470 caste prediction. All statistical analyses were carried out using SPSS v23.0 or Exlstat 2018.
471 Data and analyses given in Supplementary Table S1.

472

473 ## Dissections & RNA extractions

474 Individual heads were stored in RNAlater for brain dissections; abdomens were removed and
475 dissected to determine reproductive status. Ovary development was scored according to [31,61]
476 and the presence/absence of sperm in the spermathecae was identified to determine
477 insemination. Inseminated females with developed ovaries were scored as 'queens'; non-
478 inseminated females with undeveloped ovaries were scored as workers. Brains were
479 dissected directly into RNAlater; RNA was extracted from individuals and then pooled after
480 extraction into caste-specific pools; pooling after RNA extraction allowed for elimination of
481 any samples with low quality RNA. Pooling individuals was generally necessary to ensure
482 sufficient RNA for analyses, as well as accounting for individual variation to ensure
483 expression differences are due to caste or species, and not dependent on colony or random

484 differences between individuals. One exception to this was the *V. vulgaris* samples which
485 were sequenced as individual brains and pooled bioinformatically after sequencing.
486 Individual sample sizes per species are given in Supplementary Table S2.
487
488 Total RNA was extracted using the RNeasy Universal Plus Mini kit (Qiagen, #73404),
489 according to the manufacturer's instructions, with an extra freeze-thawing step after
490 homogenization to ensure complete lysis of tissue, as well as an additional elution step to
491 increase RNA concentration. RNA yield was determined using a NanoDrop ND-8000
492 (Thermo Fisher Scientific); all samples showed A260/A280 values between 1.9 and 2.1. An
493 Agilent 2100 Bioanalyser was used to determine RNA integrity. Samples of sufficient quality
494 and concentration were pooled and sent for sequencing. Libraries were prepared using
495 Illumina TruSeq RNAseq sample prep kit at the University of Bristol Genomics Facility. Five
496 samples were pooled per lane to give ~ 50M read per sample. Paired-end libraries were
497 sequenced using an Illumina HiSeq 2000. Descriptions of pooling of individuals and pooled
498 sets into single representatives of caste are shown in (Supplementary table S2). Raw reads
499 are available on SRA/GEO (GSE159973).
500

501 ### Preparation of *de novo* transcriptomes

502 Transcriptomes of *Agelaia, Angiopolybia, Metapolybia, Brachygastra, Polybia, Polistes* and
503 *Mischocyttarus* were assembled using the following steps. First, reads were first filtered for
504 rRNA contaminants using tools from the BBTools (version:BBMap_38) software suite
505 (https://jgi.doe.gov/data-and-tools/bbtools/). We then used Trimmomatic v0.39[62] to trim reads
506 containing adapters and low-quality regions. Using these filtered RNA sequences, we could
507 assemble a *de novo* transcriptome for each species (merging queen and worker samples)
508 using Trinity v2.8[63] and filter protein coding genes to retain a single transcript (most

509 expressed) for each gene and transcript per million value (TPM), which we use for the rest of
510 the analyses.

511 For *Vespula* and *Vespa*, reads from both queen and worker samples were assembled into
512 *de novo* transcriptomes using a Nextflow pipeline
513 (github:biocorecrg/transcriptome_assembly). This involved read adapter trimming with
514 Skewer[64], *de-novo* transcriptome assembly with Trinity v2.8.4[63] and use of TransDecoder
515 v5.5.0[63] to identify likely protein-coding transcripts, and retain all translated transcripts.
516 These were further filtered to retain the largest open read frame-containing transcript, which
517 we listed as the major isoform of each protein. Trinity assembly statistics are shown in
518 Supplementary Table S2.

519

520 Measuring gene expression within-species.

521 We calculated abundances of transcripts within queen and worker samples using
522 "align_and_estimate_abundance.pl" within Trinity, using estimation method RSEM v1.3.1[65],
523 "trinity_mode" and bowtie2[66] aligner. We then used edgeR v3.26.5 [67] (R version 3.6.0) to
524 compare gene expression between queens and workers. Because we were comparing a
525 single sequencing pool of several individuals per caste, we used a hard-coded dispersion of
526 0.1 and the robust parameter set to true to account for n = 1. Raw *P* values for each gene
527 were corrected for multiple testing using a false discovery rate (FDR) cut-off value of 0.05.
528 We did not take advantage of genome data (where available), as only two of the species had
529 published genomes at the time of analysis; using transcriptome-only analyses makes the
530 analysis more consistent across species. Trinity assemblies and RSEM counts are available
531 on GEO/SRA (GSE159973)

532

533 Identification of orthologs.

534 To identify gene-level orthologs, we used Orthofinder v.2.2.7[68] with diamond blast[32,69],
535 multiple sequence alignment program MSA[65] and tree inference using FastTree v2.1.10[70].
536 for our focal nine species, plus four out-group Hymenoptera species (Supplementary Fig. 1)
537 and *Drosophila melanogaster*. The largest spliced isoform per gene (from Trinity) was
538 designated the representative sequence for each gene. For subsequent analyses using the
539 orthofinder table of genes, we allowed the merging of genes belonging to the same species
540 in a single orthogroup (potential duplications). This decision has consequences for the
541 number of genes we can use to test in our models, as the more species used will reduce the
542 numbers of genes (with 1 to 1 orthology across all the species used in Orthofinder and SVM
543 models). In order to get a sufficient number of single-copy gene orthogroups, we merged the
544 genes in one species where there were three or less representative isoforms, only keeping
545 the gene most highly expressed.

546 Comparing gene expression between species.

547 To compare gene expression between species, we focused on our set of shared one-to-one
548 orthologs (merging 3 or less isoforms per species). We began by computing log transformed
549 TPMs (transcripts per million reads) for each gene in each sample from the raw counts,
550 followed by quantile normalisation. Next, we normalised for species, using an approach that
551 is comparable to calculating a species-specific z-score for each sample. Specifically, we
552 transformed the expression scores calculated above by subtracting the species mean and
553 dividing by the species mean for each sample within a species. This calculation has two
554 important effects. Firstly, subtracting the species mean from each sample within a species
555 centres the mean expression of each species on zero, making the units of expression more
556 comparable across species. Secondly, dividing by the species mean from each sample
557 standardises the expression scores, producing a measure that is independent of the units of

558 measurement, so that the magnitude of difference between queens and workers in each
559 sample is no longer important. The transformed expression score thus allows us to focus on
560 the relative expression in queens versus workers across species. Finally, we removed
561 orthogroups where the counts per million were below 10 in both Queen and Worker samples
562 of each species, to remove lowly expressed genes that may contribute noise to subsequent
563 analyses. We then performed principal component analysis (PCA) in R on the raw TPM
564 values and those with species scaling.

565

566 ### Machine learning (support vector machines)

567 Support vector machines (SVM) were used to classify caste across the species. In brief,
568 starting with a matrix of gene expression values, we performed pre-filtering steps (feature
569 selection), before training a model and testing this on an additional dataset. The code to run
570 these steps is shown on github (https://github.com/Sumner-lab/Multispecies_paper_ML). In
571 summary, this involved taking species-scaled, logged and normalised matrix (from RSEM),
572 with filtering of lowly expressed genes (as above), then invoking SVM predictions (radial
573 model) and plotting; code was executed in perl or R. The full detail of these steps are
574 outlined below.

575 To perform feature selection we identified only those orthologs that showed some
576 association with caste across species. For this we used linear regression of each gene on
577 caste: lm(caste ~ expr, data), using the training data only. With regression beta coefficients
578 per orthologous gene, we could then rank genes by their statistical association with caste
579 (Supp. Table 4 using the absolute values of the regression coefficients). This enabled us to
580 measure how the classification certainty changed as we filtered out genes statistically un-
581 associated with reproductive division of labour (Figure 4a). This basic feature selection
582 approach is widely used to filter large datasets in the machine learning models [71].

583 Classification certainty of 0.5 would indicate the SVM could not tell the difference between
584 the two castes (maximal uncertainty), and a classification certainty of 0/1 (worker or queen)
585 would indicate that the SVM could predict caste accurately every time (maximal certainty).

586 After identifying candidate toolkit genes of reproductive division of labour, we tested whether
587 or not they could be used to predict caste in unseen data. To do this, we trained support
588 vector machines (SVM) using the R package e1071[72]. Radial kernels were chosen for the
589 svm, which had better error statistics. We used a "leave-one-out" cross validation procedure
590 to see how well an SVM could predict the castes of our samples, where the model is trained
591 on all but one species and tested on the removed species.

592 GO/GSEA enrichment and BLAST

593 To perform GO enrichment tests, we used the R package TopGO v2.42.0[73], using Bonferroni
594 cut-off *P* values of <=0.05. In order to assign gene ontology terms to genes in our new
595 species, we used our Orthofinder homology table with annotations to *Drosophila*
596 *melanogaster* (downloaded from Ensembl Biomart 1.10.2019). Within species, we calculated
597 enrichment of each species' gene to a background of all the genes expressed above a mean
598 of 1 TPM. When comparing across the orthogroups (OG), we used *Metapolybia* GO
599 annotations (derived from homology to *Drosophila*), with a background of those genes that
600 have a mean >1 TPM in all species orthologues. GO comparisons were similar using other
601 species as a database of gene to GO terms.

602 Using default settings in GSEA v4.0.3[74] we compared the lists derived from the SVM
603 experiment and conventional differential expression analysis (using the preranked mode).
604 First (Supplementary Figure 2a), using the list of 2020 SVM (9 species) orthologs (excluding
605 low-expression genes) ranked from 1 to 2020 based on the linear regression *P* values we
606 could derive enrichment scores from the DEGs (n=95; from edgeR), where the total were
607 reduced to 19 genes that were present in both analyses. Second (Supplementary Figure 2b),

608 we ranked Vespula (Trinity) differentially expressed genes by log fold change, deriving
609 enrichment scores with the 400 SVM genes significant in the nine species SVM, after linear
610 regression p.value cutoff of 0.05. Blast2GO v 1.4.4[75] using Metapolybia gene sequences
611 using was used to annotate sequences, along with manual use of NCBI blastn[76] suite online.

612 Abbreviations

613 DOL = division of labour

614 ORF = open reading frame

615 MT = major transition

616 GO = gene ontology

617 SRA = sequence read archive

618 NCBI = National Centre for Biotechnology Information

619 BUSCO = Benchmarking set of Universal Single-Copy Orthologs

620 SVM = support vector machine

621 PCA = principal components analysis

622 Blast = Basic local alignment search tool

623 nr = non-redundant

624

625 Author's contributions

626 SS conceived the study and supervised the project; SS, EL, EB and BT collected the
627 samples; DT, EB, BT and RB performed molecular lab work; DT & RB carried out the

morphological work; MB & CW executed the bioinformatics pipelines, performed the statistical analyses; CW & SS drafted the manuscript, with input from all authors.

References

1. Szathmáry, E. & Maynard Smith, J. The major evolutionary transitions. *Nature* **374**, 227–232 (1995).

2. Kennedy, P. *et al.* Deconstructing Superorganisms and Societies to Address Big Questions in Biology. *Trends in Ecology and Evolution* **32**, 861–872 (2017).

3. Toth, A. L. & Robinson, G. E. Evo-devo and the evolution of social behavior. *Trends Genet.* **23**, 334–341 (2007).

4. Berens, a. J., Hunt, J. H. & Toth, a. L. Comparative Transcriptomics of Convergent Evolution: Different Genes but Conserved Pathways Underlie Caste Phenotypes across Lineages of Eusocial Insects. *Mol. Biol. Evol.* **32**, 690–703 (2014).

5. Qiu, B. *et al.* Towards reconstructing the ancestral brain gene-network regulating caste differentiation in ants. *Nat. Ecol. Evol.* **2**, 1782 (2018).

6. Warner, M. R., Qiu, L., Holmes, M. J., Mikheyev, A. S. & Linksvayer, T. A. Convergent eusocial evolution is based on a shared reproductive groundplan plus lineage-specific plastic genes. *Nat. Commun.* **10**, 1–11 (2019).

7. Patalano, S. *et al.* Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13970–5 (2015).

8. Toth, Amy L. Rehan, S. . Climbing the social ladder: The molecular evolution of

650 sociality. (2015).

651 9. West-Eberhard MJ. Wasp societies as microcosms for the study of development and
652 evolution. in *Natural history and evolution of paper wasps.* (eds. Turillazzi, S. & West-
653 Eberhard, M. J.) 290–317 (Oxford University Press, 1996).

654 10. Toth, A. L. & Rehan, S. M. Molecular Evolution of Insect Sociality: An Eco-Evo-Devo
655 Perspective. *Annual Review of Entomology* (2017). doi:10.1146/annurev-ento-
656 031616-035601

657 11. Boomsma, J. J. Lifetime monogamy and the evolution of eusociality. *Philos. Trans. R.*
658 *Soc. Lond. B. Biol. Sci.* **364**, 3191–207 (2009).

659 12. Boomsma, J. J. & Gawne, R. Superorganismality and caste differentiation as points of
660 no return: how the major evolutionary transitions were lost in translation. *Biol. Rev.*
661 (2017). doi:10.1111/brv.12330

662 13. Ferreira, P. G. *et al.* Transcriptome analyses of primitively eusocial wasps reveal
663 novel insights into the evolution of sociality and the origin of alternative phenotypes.
664 *Genome Biol.* **14**, R20 (2013).

665 14. Sumner, S. The importance of genomic novelty in social evolution. *Mol. Ecol.* **23**,
666 (2014).

667 15. Feldmeyer, B., Elsner, D. & Foitzik, S. Gene expression patterns associated with
668 caste and reproductive status in ants: Worker-specific genes are more derived than
669 queen-specific ones. *Mol. Ecol.* **23**, 151–161 (2014).

670 16. Johnson, B. R. & Tsutsui, N. D. Taxonomically restricted genes are associated with
671 the evolution of sociality in the honey bee. *BMC Genomics* **12**, 164 (2011).

672 17. Simola, D. F. *et al.* Social insect genomes exhibit dramatic evolution in gene
673 composition and regulation while preserving regulatory features linked to sociality.

674 *Genome Res.* **23**, 1235–1247 (2013).

675 18. Harpur, B. a *et al.* Population genomics of the honey bee reveals strong signatures of
676 positive selection on worker traits. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2614–9 (2014).

677 19. Rubin, B. E. R., Jones, B. M., Hunt, B. G. & Kocher, S. D. Rate variation in the
678 evolution of non-coding DNA associated with social evolution in bees. *Philos. Trans.*
679 *R. Soc. B Biol. Sci.* **374**, (2019).

680 20. Kapheim, K. M. Genomic sources of phenotypic novelty in the evolution of eusociality
681 in insects. *Curr. Opin. Insect Sci.* 1–9 (2015). doi:10.1016/j.cois.2015.10.009

682 21. Dogantzis, K. A. *et al.* Insects with similar social complexity show convergent patterns
683 of adaptive molecular evolution. 1–8 (2018). doi:10.1038/s41598-018-28489-5

684 22. Taylor, B. A., Reuter, M. & Sumner, S. Patterns of reproductive differentiation and
685 reproductive plasticity in the major evolutionary transition to superorganismality. *Curr.*
686 *Opin. Insect Sci.* (2019).

687 23. Branstetter, M. *et al.* Genomes of the Hymenoptera. *Curr. Opin. Insect Sci.* **25**, 65–75
688 (2017).

689 24. Standage, D. S. *et al.* Genome, transcriptome and methylome sequencing of a
690 primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social
691 insect. *Mol. Ecol.* **25**, 1769–1784 (2016).

692 25. Toth, A. L. *et al.* Shared genes related to aggression, rather than chemical
693 communication, are associated with reproductive dominance in paper wasps (Polistes
694 metricus). *BMC Genomics* **15**, 75 (2014).

695 26. Bluher, S. E., Miller, S. E. & Sheehan, M. J. Fine-scale population structure but limited
696 genetic differentiation in a cooperatively breeding paper wasp. *Genome Biol. Evol.*
697 (2020).

698 27. Rehan, S. M. *et al.* Conserved Genes Underlie Phenotypic Plasticity in an Incipiently
699 Social Bee. *Genome Biol. Evol.* **10**, 2749–2758 (2018).

700 28. Kocher, S. D. *et al.* The genetic basis of a social polymorphism in halictid bees. *Nat.*
701 *Commun.* **9**, (2018).

702 29. Shell, W. A. & Rehan, S. M. Social modularity: Conserved genes and regulatory
703 elements underlie caste-antecedent behavioural states in an incipiently social bee.
704 *Proc. R. Soc. B Biol. Sci.* (2019). doi:10.1098/rspb.2019.1815

705 30. Kapheim, K. M. *et al.* Developmental plasticity shapes social traits and selection in a
706 facultatively eusocial bee. *Proc. Natl. Acad. Sci.* 202000344 (2020).
707 doi:10.1073/pnas.2000344117

708 31. Taylor, D., Bentley, M. A. & Sumner, S. Social wasps as models to study the major
709 evolutionary transition to superorganismality. *Curr. Opin. Insect Sci.* **28**, 26–32
710 (2018).

711 32. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
712 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*
713 (2015). doi:10.1186/s13059-015-0721-2

714 33. Roy-Zokan, E. M., Cunningham, C. B., Hebb, L. E., McKinney, E. C. & Moore, A. J.
715 Vitellogenin and vitellogenin receptor gene expression is associated with male and
716 female parenting in a subsocial insect. *Proc. R. Soc. B Biol. Sci.* (2015).
717 doi:10.1098/rspb.2015.0787

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

721 35. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data.

722 *Genome Biol.* (2016). doi:10.1186/s13059-016-0888-1

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

726 37. Furey, T. S. *et al.* Support vector machine classification and validation of cancer tissue
727 samples using microarray expression data. *Bioinformatics* (2000).
728 doi:10.1093/bioinformatics/16.10.906

729 38. Taylor, B. A., Cini, A., Wyatt, C. D. R., Reuter, M. & Sumner, S. The molecular basis
730 of socially-mediated phenotypic plasticity in a eusocial paper wasp. *bioRxiv*
731 2020.07.15.203943 (2020). doi:10.1101/2020.07.15.203943

732 39. Hoffmann, K., Gowin, J., Hartfelder, K. & Korb, J. The Scent of Royalty: A P450 Gene
733 Signals Reproductive Status in a Social Insect. *Mol. Biol. Evol.* **31**, 2689–2696 (2014).

734 40. Calkins, T. L. *et al.* Brain gene expression analyses in virgin and mated queens of fire
735 ants reveal mating-independent and socially regulated changes. *Ecol. Evol.* (2018).
736 doi:10.1002/ece3.3976

737 41. Iovinella, I. *et al.* Antennal protein profile in honeybees: Caste and task matter more
738 than age. *Front. Physiol.* (2018). doi:10.3389/fphys.2018.00748

739 42. Glastad, K. M. *et al.* Epigenetic Regulator CoREST Controls Social Behavior in Ants.
740 *Mol. Cell* (2020). doi:10.1016/j.molcel.2019.10.012

741 43. Turrel, O., Goguel, V. & Preat, T. Drosophila neprilysin 1 rescues memory deficits
742 caused by amyloid-β peptide. *J. Neurosci.* (2017). doi:10.1523/JNEUROSCI.1634-
743 17.2017

744 44. Yanay, C., Morpurgo, N. & Linial, M. Evolution of insect proteomes: Insights into
745 synapse organization and synaptic vesicle life cycle. *Genome Biol.* (2008).

doi:10.1186/gb-2008-9-2-r27

45. Helanterä, H. An organismal perpective on the evolution of insect societies. *Front. Ecol. Evol.* **4**, 1–12 (2016).

46. Noll, F. B., Zucchi, R. & Simões, D. Morphological caste differences in the neotropical swarm-founding polistinae wasps: Agelaia m. a. multipicta and a. p. pallipes (hymenoptera vespidae). *Ethol. Ecol. Evol.* (1997). doi:10.1080/08927014.1997.9522878

47. Kohannim, O. *et al.* Boosting power for clinical trials using classifiers based on multiple biomarkers. *Neurobiol. Aging* (2010). doi:10.1016/j.neurobiolaging.2010.04.022

48. Montagna, T. S. & Antonialli, W. F. Morphological differences between reproductive and non-reproductive females in the social wasp Mischocyttarus consimilis Zikán (Hymenoptera: Vespidae). *Sociobiology* **63**, 693–698 (2016).

49. Jeanne, R. L. Social Biology of the neotropical wasp, Mischocyttarus drewseni. *Bull. Museum Comp. Zool.* **144**, 63–150 (1972).

50. Murakami, A. S. N., Shima, S. N. & Desuó, I. C. More than one inseminated female in colonies of the independent-founding wasp Mischocyttarus cassununga von Ihering (Hymenoptera, Vespidae). *Rev. Bras. Entomol.* **53**, 653–662 (2009).

51. Reeve, H. K. Polistes. in *The Social Biology of Wasps* (ed. Ross KG, M. R.) 99–148 (Cornell University Press, 1991).

52. Richards, O. W. *The social wasps of the Americas*. (1978).

53. Noll, F. B., Wenzel, J. W. & Zucchi, R. Evolution of Caste in Neotropical Swarm-Founding Wasps (Hymenoptera : Evolution of Caste in Neotropical Swarm-Founding Wasps (Hymenoptera : Vespidae ; Epiponini). *Am. Museum Novit.* **3467**, 1–24

770 (2004).

771 54. Gelin, L. F. F. *et al.* Morphological Caste Studies In The Neotropical Swarm-Founding
772 Polistinae Wasp Angiopolybia pallens (Lepeletier) (Hymenoptera : Vespidae).
773 *Neotrop. Entomol.* **37**, 691–701 (2008).

774 55. West-Eberhard, M. J. Temporary Queens in Metapolybia Wasps : Nonreproductive
775 Helpers Without Altruism ? *Science (80-.).* **200**, 441–443 (1978).

776 56. Sakagami, S. F., Zucchi, R., Yamane, S., Noll, F. B. & Camargo, J. M. P.
777 Morphological caste differences in Agelaia vicina, the neotropical swarm-founding
778 polistine wasp with the largest colony size among social wasps (Hymenoptera:
779 Vespidae). *Sociobiology* **28**, 207–223 (1996).

780 57. V., B. M., Noll, F. B. & Zucchi, R. Morphological Caste Differences and Non-Sterility of
781 Workers in Brachygastra augusti (Hymenoptera , Vespidae , Epiponini), a
782 Neotropical Swarm-Founding Wasp Author (s): *J. New York Entomol. Soc.* **111**,
783 242–252 (2003).

784 58. Loope, K. J., Chien, C. & Juhl, M. Colony size is linked to paternity frequency and
785 paternity skew in yellowjacket wasps and hornets. *BMC Evol. Biol.* **14**, 1–12 (2014).

786 59. Hastings, M. D., Queller, D. C., Eischen, F. & Strassmann, J. E. Kin selection ,
787 relatedness , and worker control of reproduction in a large-colony epiponine wasp ,
788 Brachygastra mellifica. *Behav. Ecol.* **9**, 573–581 (1998).

789 60. Gobbi, N., Noll, F. B. & Penna, M. A. H. 'Winter' aggregations, colony cycle, and
790 seasonal phenotypic change in the paper wasp Polistes versicolor in subtropical
791 Brazil. *Naturwissenschaften* (2006). doi:10.1007/s00114-006-0140-z

792 61. Kronauer, D. J. & Libbrecht, R. Back to the roots: the importance of using simple
793 insect societies to understand the molecular basis of complex social life. *Curr. Opin.*

794 *Insect Sci.* **28**, 33–39 (2018).

795 62. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina
796 sequence data. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu170

797 63. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the
798 Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494 (2013).

799 64. DI Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature*
800 *Biotechnology* (2017). doi:10.1038/nbt.3820

801 65. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
802 DIAMOND. *Nat. Methods* **12**, 59 (2014).

803 66. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data
804 with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

805 67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat.*
806 *Methods* (2012). doi:10.1038/nmeth.1923

807 68. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
808 differential expression analysis of digital gene expression data. *Bioinformatics* **26**,
809 139–140 (2010).

810 69. Emms, D. M. & Kelly, S. OrthoFinder2: fast and accurate phylogenomic orthology
811 analysis from gene sequences. *bioRxiv* (2018). doi:10.1101/466201

812 70. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-
813 likelihood trees for large alignments. *PLoS One* (2010).
814 doi:10.1371/journal.pone.0009490

815 71. Karagiannopoulos, M., Anyfantis, D., Kotsiantis, S. B. & Pintelas, P. E. Feature
816 Selection for Regression Problems. *8th Hell. Eur. Res. Comput. Math. its Appl.*

817 *HERCMA 2007* (2007). doi:10.1109/ICDM.2014.63

818 72. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. *e1071: Misc*
819 *Functions of the Department of Statistics (e1071), TU Wien. R package version*
820 (2011).

821 73. Alexa, A. & Rahnenführer, J. Gene set enrichment analysis with topGO. *Bioconductor*
822 *Improv.* (2007).

823 74. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach
824 for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*
825 (2005). doi:10.1073/pnas.0506580102

826 75. Götz, S. *et al.* High-throughput functional annotation and data mining with the
827 Blast2GO suite. *Nucleic Acids Res.* (2008). doi:10.1093/nar/gkn176

828 76. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local
829 alignment search tool. *J. Mol. Biol.* (1990). doi:10.1016/S0022-2836(05)80360-2

830 77. Piekarski, P. K., Carpenter, J. M., Lemmon, A. R., Lemmon, E. M. & Sharanowski, B.
831 J. Phylogenomic evidence overturns current conceptions of social evolution in wasps
832 (vespidae). *Mol. Biol. Evol.* **35**, 2097–2109 (2018).

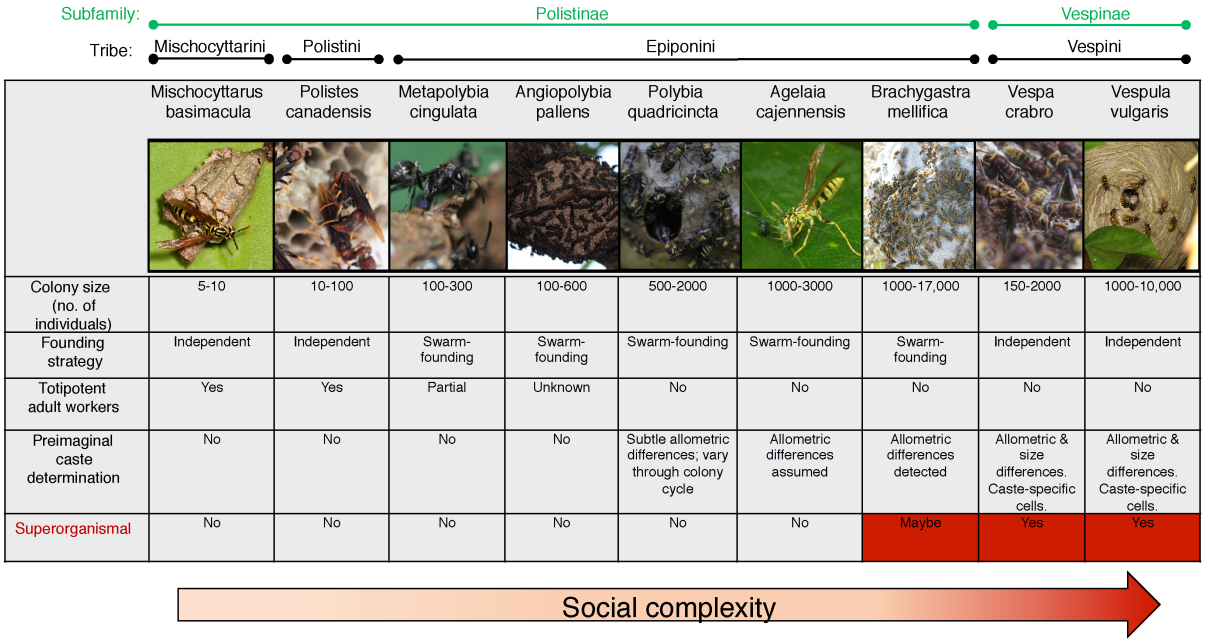
833 78. O'Donnell, S. Reproductive caste determination in eusocial wasps (Hymenoptera:
834 Vespidae). *Annu. Rev. Entomol.* **43**, 323–46 (1998).

835 79. Matsuura, M. & Yamane, S. *Biology of the vespine wasps.* (Springer Verlag, 1990).

836 80. Menezes, R. S. T., Lloyd, M. W. & Brady, S. G. Phylogenomics indicates Amazonia as
837 the major source of Neotropical swarm-founding social wasp diversity. *Proc. R. Soc. B*
838 (2020).

839

841 # Main Figures



842

843 **Fig. 1 |** Social wasps as a model group. The nine species of social wasps used in this
844 study, and their characteristics of social complexity. The Polistinae and Vespinae are two
845 subfamilies comprising 1100+ and 67 species of social wasp respectively, all of which share
846 the same common non-social ancestor, an eumenid-like solitary wasp[77]. The Polistinae are
847 an especially useful subfamily for studying the process of the major transition as they include
848 species that exhibit simple group living comprised of small groups (<10 individuals) of
849 totipotent relatives, as well as species with varying degrees of more complex forms of
850 sociality, with different colony sizes, levels of caste commitment and reproductive
851 totipotency[78]. The Vespinae include the yellow-jackets and hornets, and are all
852 superorganismal, meaning caste is determined during development in caste-specific brood
853 cells; they also show species-level variation in complexity, in terms of colony size and other
854 superorganismal traits (e.g. multiple mating, worker policing)[79]. Ranked in order of increasing
855 levels of social complexity, from simple to more complex, these species are: *Mischocyttarus*
856 *basimacula basimacula*, *Polistes canadensis*, *Metapolybia cingulata, Angiopolybia pallens,*
857 *Polybia quadricincta, Agelaia cajennensis, Brachygastra mellifica, Vespa crabro* and
858 *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 S1). Image credits: *M. basimacula* (Stephen Cresswell). *A. cajennesis (*Gionorossi; Creative Commons); *V. vulgaris (*Donald Hobern; Creative Commons). *V. crabro* (Patrick Kennedy); *P. canadensis; M. cingulata, A. pallens, P. quadricinta,* (Seirian Sumner), *B. mellifica* (Amante Darmanin; Creative Commons).

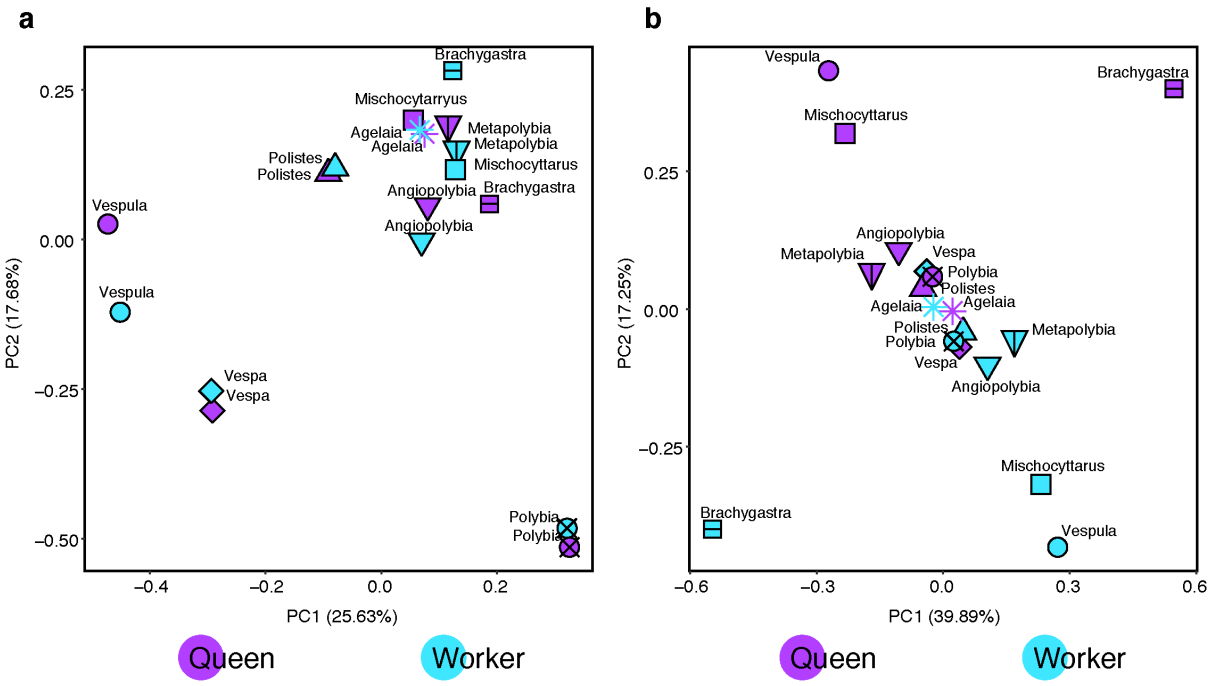


Fig. 2 | Principal component analyses of orthologous gene expression before and after between-species normalisation. a) Principal component analyses performed using log2 transcript per million (TPM) gene expression values. This analysis used single-copy orthologs (using Orthofinder), allowing up to three 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). Species are denoted by symbols.

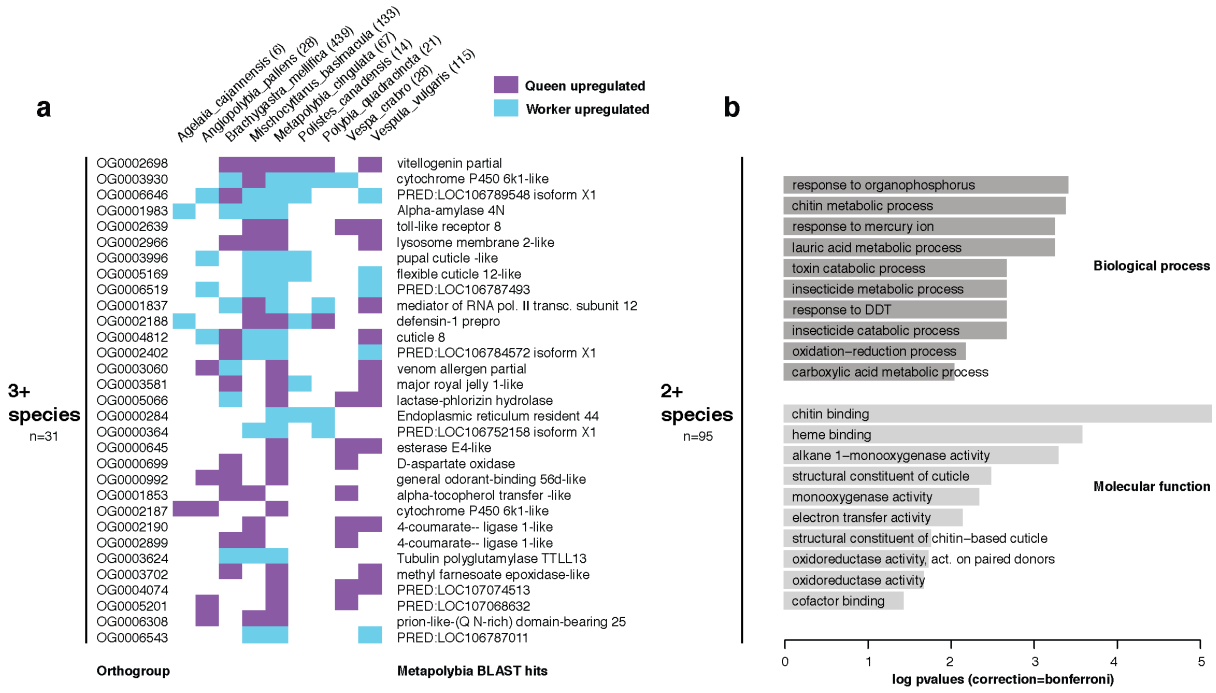


Fig. 3 | Overlap of differential caste-biased genes (queen vs worker) and their functions across eusocial wasp species. a) Heatmap showing the differential genes that are caste-biased in at least three species (identified using edgeR) using the orthologous genes present in the nine species. Listed for each species, is the total number of differentially expressed genes per species (orthologous-one copy only). *Metapolybia* Blast hits are listed. **b)** Gene ontology histogram of overrepresented terms of genes found differentially expressed in at least two out of the nine species (n=95 genes; in either queen or worker [not both]), with a background of those expressed in each species above 1 TPM. *P* values are single-tailed and were not corrected, given the low levels of enrichment generally and are therefore not significant for multiple testing.

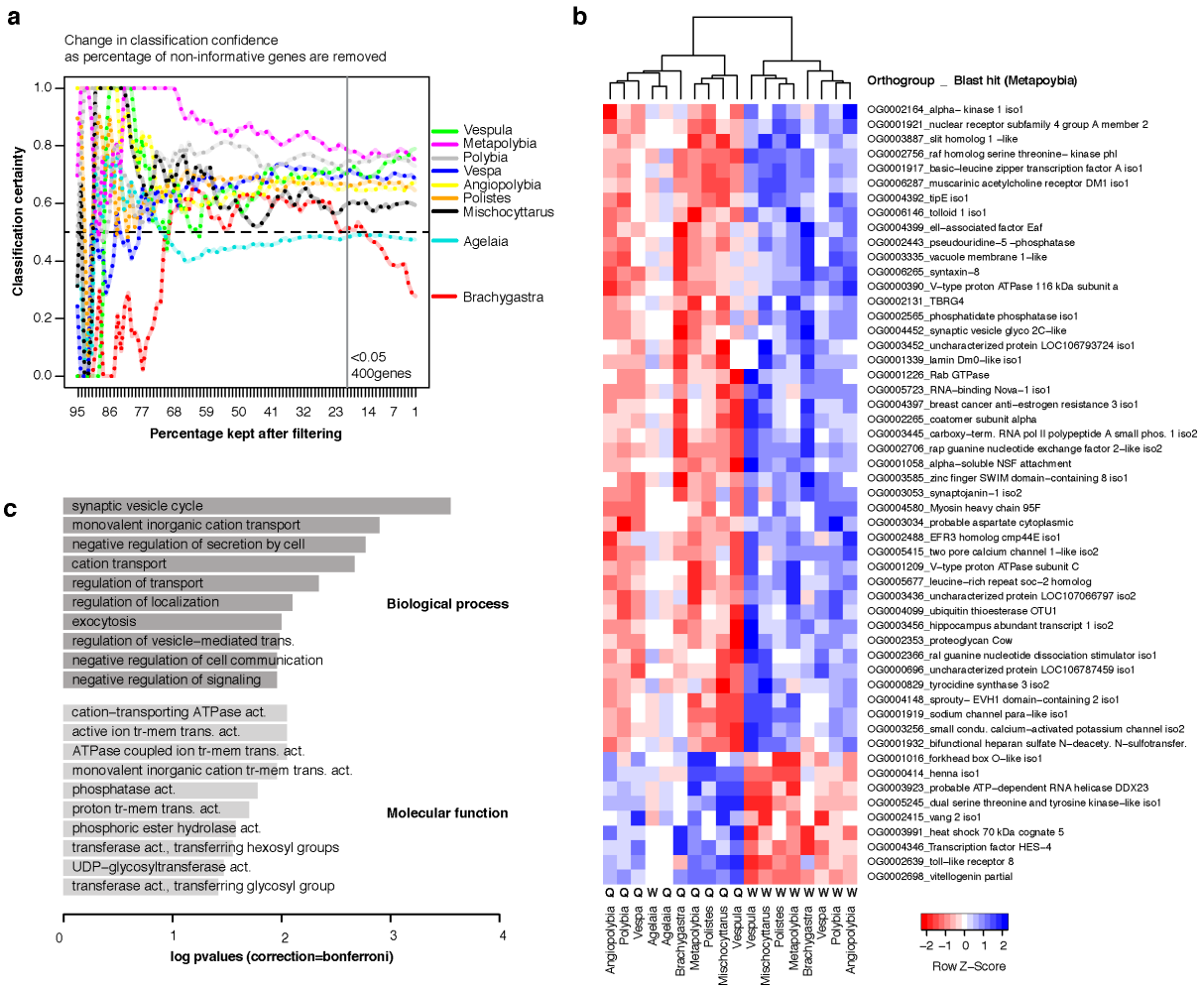


Fig. 4 | A genetic toolkit for social behaviours across eusocial wasps. a) Change in certainty of correct classifications through progressive feature selection. Models were trained on eight species and tested on the ninth species. Features (a.k.a. genes) were sorted by linear regression with regard to caste identity, beginning at 95% where almost all genes were used for the predictions of caste, to 1% where only the top one percent of genes from the linear regression (sorted by *P* value) were used to train the model. '1' equates to high classification certainty. **b)** Heatmap of the top 53 species-normalised gene expression levels in the nine species with queen/worker indicated. Genes selected using linear regression (*P* value < 0.001) used in the SVM model, showing orthogroup name and top *Metapolybia* BLAST hit. **c**) Gene Ontology for the top 400 orthologous genes predictive of caste across species (linear regression *P* value < 0.05), and a background of all genes used

910 in the SVM model (i.e. with a single gene representative for all species in the test). *P* values
911 are bonferroni corrected and single-tailed. Abbreviations, “tr-mem”:transmembrane, “act.”:
912 activity, “trans.”:transporter.

913

914

915

916

917

918

919

920

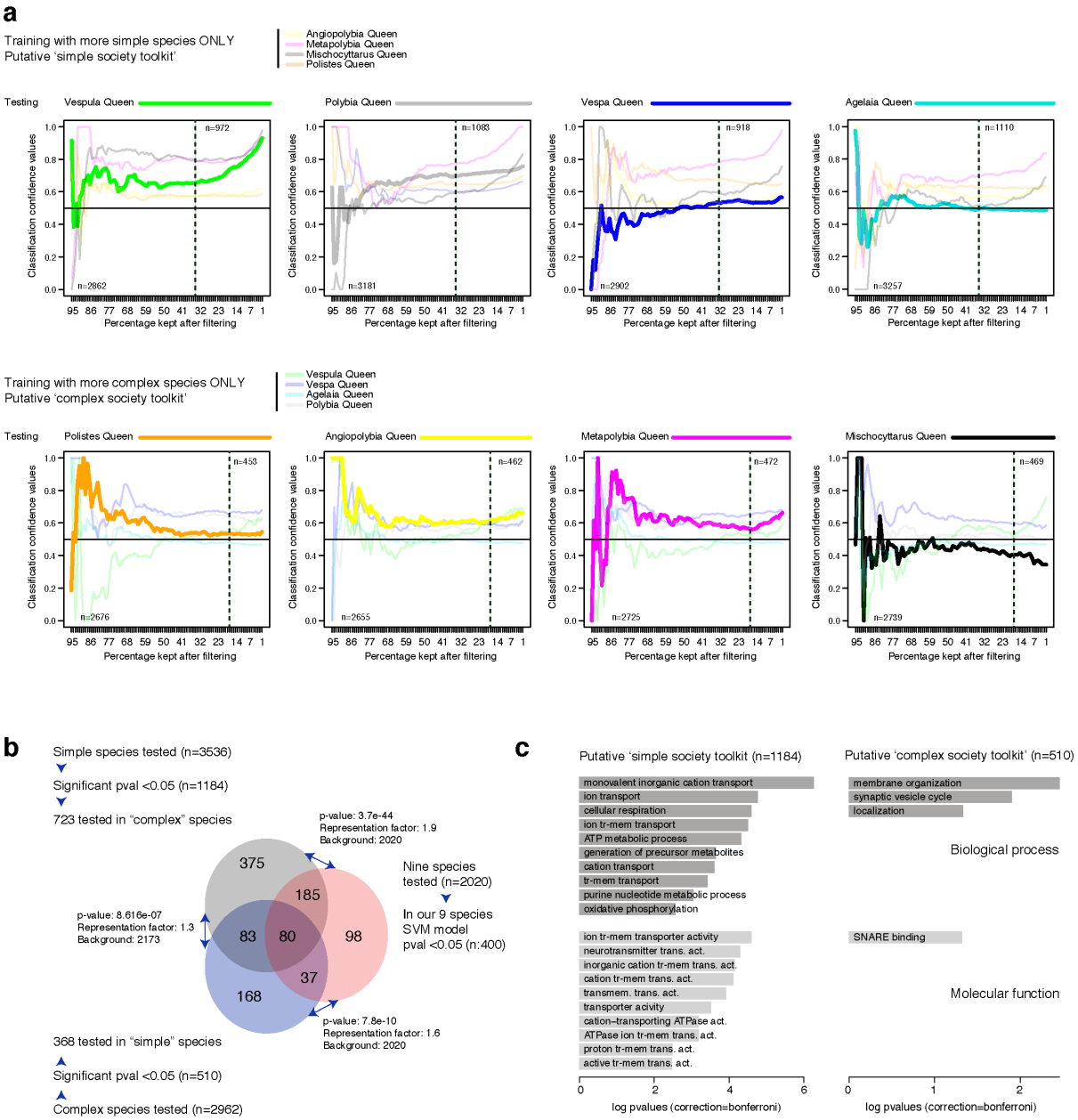
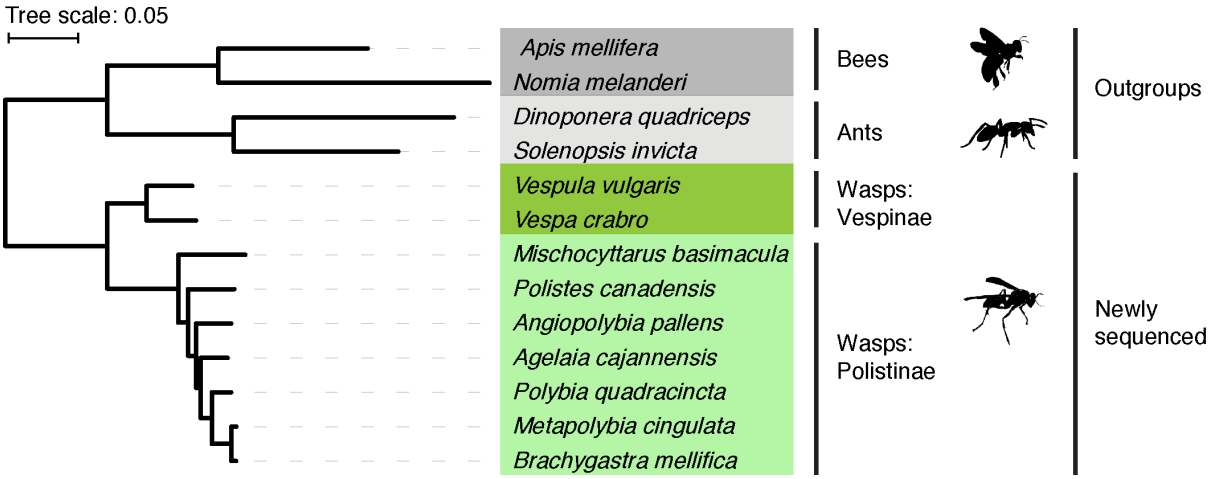


Fig.5 | Testing for the presence of a defined 'simple society toolkit' and a 'complex society toolkit'. a): Using the four species with the more complex societies, or the four with the more simple societies, we trained and tested an SVM model, using progressive filtering of genes (based on the linear regression). Likelihood of being a queen from zero to 1 is plotted for each species across the progressively filtered sets. The number of genes used in the SVM model are shown for each test (bottom left of each panel), of which the total number of genes left after the regression filter are shown (top right of each panel), using genes with a *P* value < 0.05. 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 (maximum of 3 isoforms merged). **b)** Overlap of significant genes in the different sets, compared to the 400 found using all nine species. For each experiment, the number of genes (orthogroups) tested is listed, then the number of genes significant after linear regression, and finally the number of genes that were also tested in the other two experiments. Significant overlap is shown using hypergeometric tests (one-tailed). Blue represents genes expressed in the four species with the most complex societies; grey those expressed in the four species with the most simple societies; pink are those expressed across all nine species. **c)** Enriched gene ontology terms (TopGO) using a background of all genes tested in each individual experiment, using a bonferroni corrected single-tailed *P* values.

957 # Supplementary Figures



958

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

960 related hymenopterans generated using Orthofinder (SpeciesTree_rooted.txt). *Drosophila*

961 was chosen as the root of the species tree (not shown), with two representative ants

962 (*Dinoponera quadriceps* and *Solenopsis invicta*) and two bees (*Apis mellifera* and *Nomia*

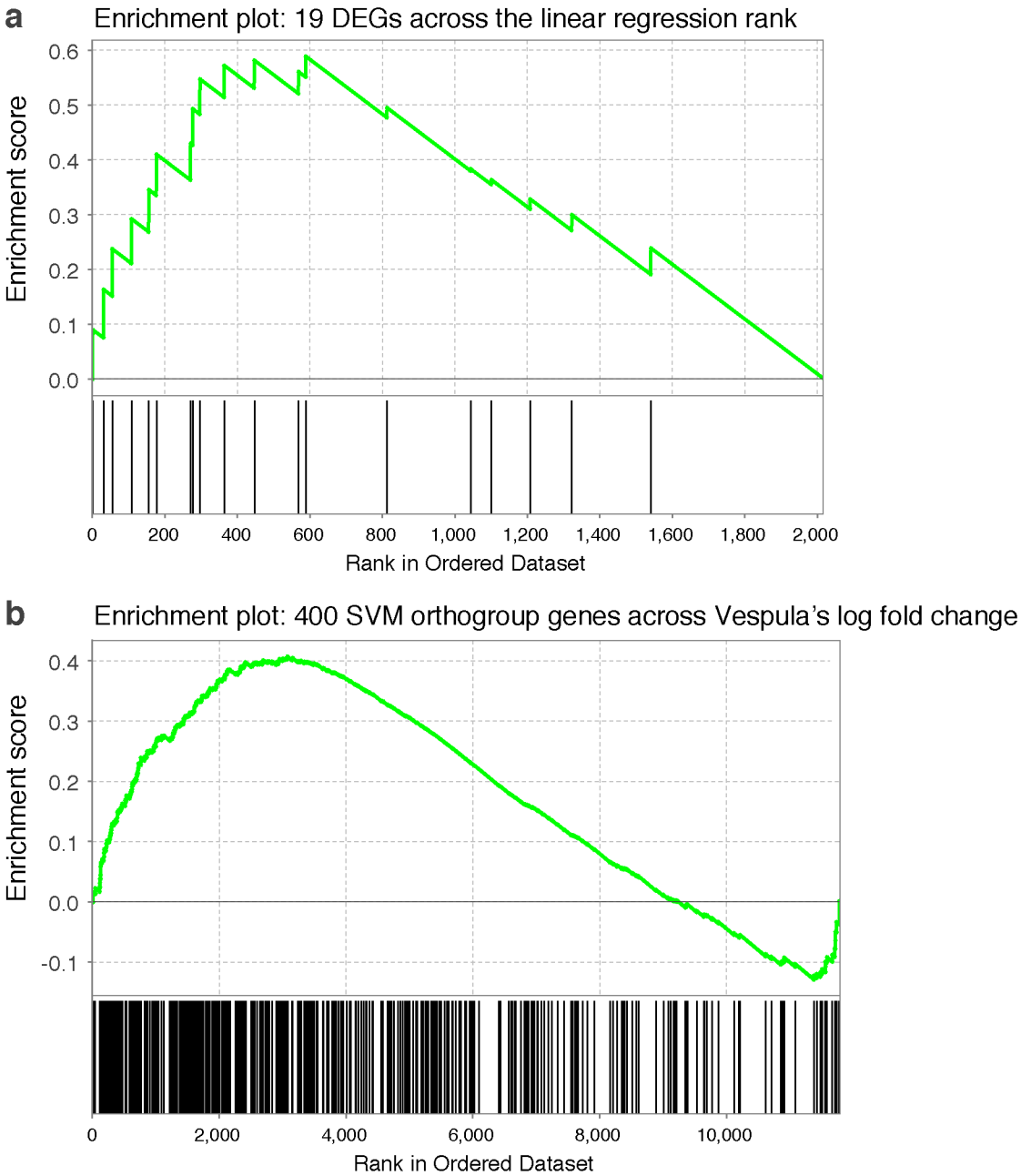
963 *melanderi*). Colours show groupings of ants, bees and wasps (Vespinae or Polistinae). For

964 the nine wasp species (this study) we have sequenced adult caste-specific brain

965 transcriptomic data (queen and worker). As expected: Vespinae are clearly separated from

966 the Polistinae; *Angiopolybia* is basal to the other Epiponini; independent-founding, non-

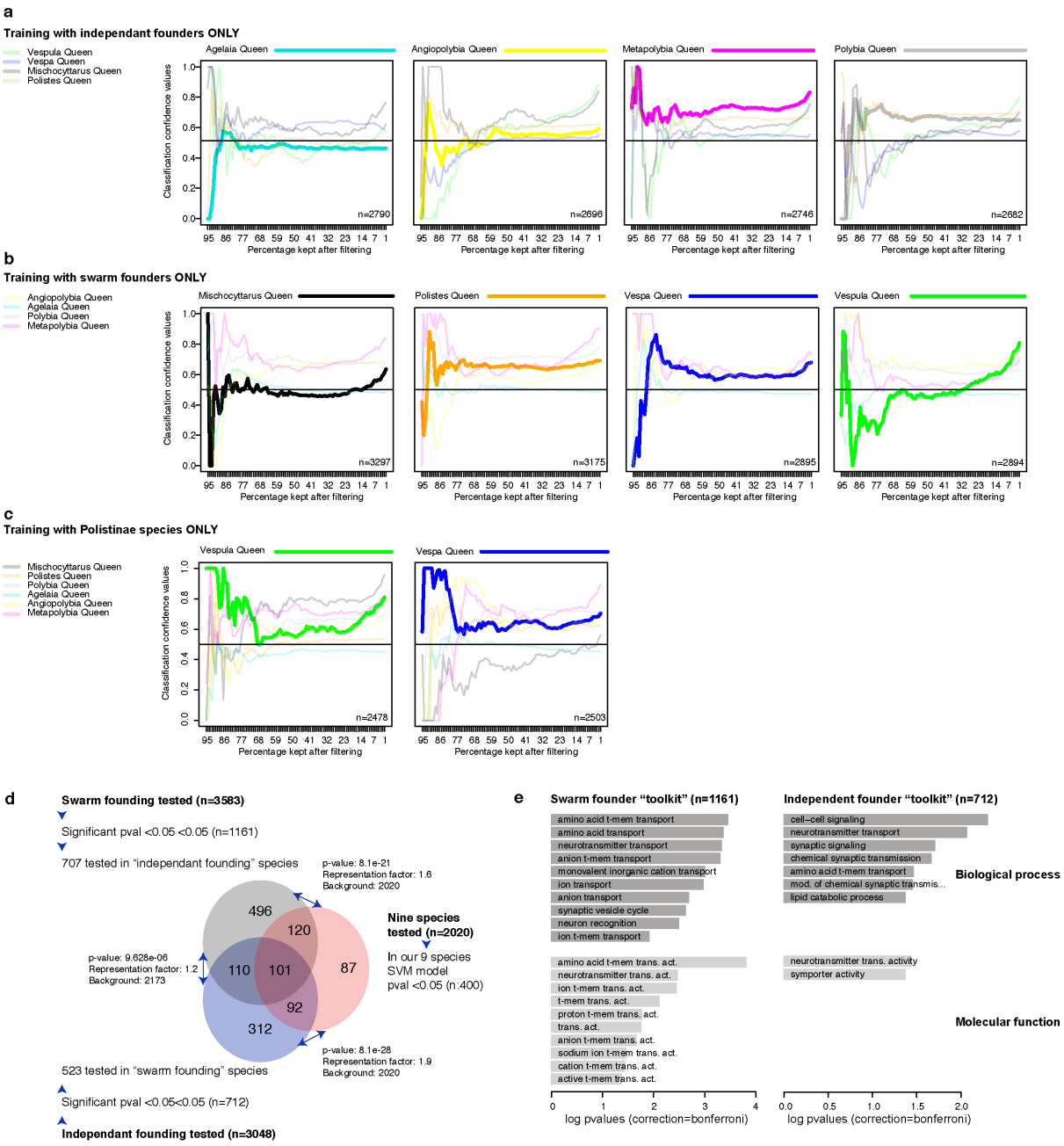
967 superorganismal wasps (*Polistes* and *Mischocyttarus*) are basal to the Polistinae[77,80].



Supp. Figure. 2. | GSEA –Gene Set Enrichment Analysis comparing overlap of the orthogroups discovered with differentially expressed genes.

a): Enrichment scores using the 95 orthogroups differentially expressed in at least two wasp species and their ordering across the ranked SVM genes derived from linear regression (2020 orthogroups; where rank 0 is the orthogroup most associated with caste). Of 95 genes, only 19 orthogroups were found in the linear regression sorted set. The upper plot shows the enrichment scores, while the lower plot shows the position of the 19 orthogroup

genes across the ranked SVM list. Enrichment was found toward the higher ranks (nearer 0), suggesting that there is some overlap in genes found using linear regression (SVM)/edgeR approaches. **b):** Enrichment scores using the 400 significant SVM genes (from linear regression, converted to Vepsula IDs) across the log fold changes of Vespula genes from edgeR. Enrichment in both queen and worker biased ends (see two peaks: left/right) were detected, again suggesting limited overlap in the genes found using the two approaches.



Supp. Figure. 3. | SVM predictions using founding behaviour and phylogeny to subset 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)** The reverse of 'a'. **c)** Training with Polistine species only, and testing on the Vespines (*Vespa* and *Vespula*). The

995 reverse was not possible, as the SVM requires a minimum number of samples to work. **d)**
996 Overlap of genes found using the swarm/independent founding toolkits, along with the
997 overlap with the 400 genes discovered using all nine species. Hypergeometric p.values and
998 representation factors are shown. For each set we show the total number of genes tested in
999 each experiment, followed by the number significant at the <0.05 pvalue cutoff, and then of
1000 these, the number that were also tested in the other two comparisons. These genes
1001 numbers are then overlapped in the Venn. **e)** Enriched gene ontology terms (TopGO) using
1002 a background of all genes tested in each individual experiment, using a bonferroni corrected
1003 single-tailed *P* values.