bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 # Multiple lineages, same molecular basis: task specialization is
2 # commonly regulated across all eusocial bee groups

3

4 Natalia de Souza Araujo[1,2*] and Maria Cristina Arias[1]

5

6 [1] Department of Genetics and Evolutionary Biology, Universidade de São Paulo. Rua do Matão, 277,
7 05508-090 São Paulo - SP, Brazil
8 [2] [current address] Interuniversity Institute of Bioinformatics in Brussels and Department of
9 Evolutionary Biology & Ecology, Université Libre de Bruxelles. Avenue F.D. Roosevelt, 50. B-1050
10 Brussels, Belgium

11

12 * Corresponding author: N.S. Araujo – souza.nataliaa@gmail.com

13

14 ## Abstract

15 A striking feature of advanced insect societies is the existence of workers that forgo
16 reproduction. Two broad types of workers exist in eusocial bees: nurses which care
17 for their young siblings and the queen, and foragers who guard the nest and forage for
18 food. Comparisons between this two worker subcastes have been performed in
19 honeybees, but data from other bees are scarce. To understand whether similar
20 molecular mechanisms are involved in nurse-forager differences across distinct
21 species, we compared gene expression and DNA methylation profiles between nurses
22 and foragers of the buff-tailed bumblebee *Bombus terrestris* and of the stingless bee
23 *Tetragonisca angustula.* These datasets were then discussed comparatively to
24 previous findings on honeybees. Our analyses revealed that although the expression
25 pattern of genes is often species-specific, many of the biological processes and
26 molecular pathways involved are common. Moreover, DNA methylation and gene
27 expression correlation were dependent on the nucleotide context.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

28 **Introduction**

29 Caste specialization in eusocial insects is a notorious example of polyphenism,
30 where multiple morphological and behavioural phenotypes emerge from the same
31 genotype[1,2]. In social Hymenoptera (bees, wasps and ants), queen and worker
32 reproductive castes perform distinct functions in the colony. While queens undertake
33 reproductive duties, workers perform all the other tasks necessary for nest
34 maintenance and growth[3]. Two broad categories of workers exist in eusocial bees:
35 nurses and foragers[4,5]. Nurses are responsible for comb construction, offspring/queen
36 care and internal colony maintenance, while foragers perform tasks related to external
37 colony defence and resources provisioning[5,6]. In advanced eusocial bee species, such
38 as honeybees, worker subcastes are mainly age determined; younger bees are nurses
39 and when they become older, they switch to being foragers[7,8]. In primitively eusocial
40 species, such as the social bumblebees, specialization in worker subcastes is not so
41 straightforward[9,10].

42 To investigate differences in bee worker subcastes, many studies have been
43 conducted in the highly eusocial honeybee (*Apis*). Gene expression comparisons have
44 identified expression changes between worker behaviours[1,5,7,11,12], which could even
45 be used to predict neurogenomic states in individual bees[13]. Similarly, profiles of
46 DNA methylation, an epigenetic marks that likely underpins gene expression
47 differences, were additionally shown to directly correlate with worker task[14,15].
48 Certain genes were differentially methylated according to the worker subcaste and
49 foragers that were forced to revert to nursing restored more than half of the nursing-
50 specific DNA methylation marks[16,17].

51 Many of the molecular differences between honeybee workers and nurses
52 could have arisen later in the evolution of this lineage. To broadly understand how

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

53 subcastes evolved it is necessary to differentiate such more recent changes – that
54 could be species-specific – from those shared across species, and thus likely ancestral.
55 The highly eusocial stingless bees have age-based division of labour[18], similarly to
56 that of honeybees despite their most common ancestor being 50 to 80 million years
57 ago[19,20]. To date, no global expression or epigenetic studies have been performed in
58 stingless bees to understand worker task specialization. Similarly, while primitively
59 eusocial bumblebees are largely studied ecological biological models and important
60 wild and managed pollinators, we know comparatively little about the molecular
61 underpinnings of differences between its worker subcastes. Indeed, studies have been
62 restricted to few genes, leaving many open questions[21–23]. A major limiting element
63 for these studies is that this species display a somewhat fluctuating division of labour
64 with indistinctive separation between subcastes[10,21,23].
65 We aim to fill in this knowledge gap through the analyses of the global gene
66 expression differences between nurses and foragers, and the characterization of nurses
67 DNA methylation profile in two eusocial bee species, the primitively eusocial buff-
68 tailed bumblebee, *Bombus terrestris*, and the highly eusocial stingless bee,
69 *Tetragonisca angustula.* Combined, these two bee species and the honeybee represent
70 the three evolutionary branches of eusocial corbiculates sharing a common social
71 origin[24]. Hence, in addition to using the generated datasets to uncover unique and
72 more recent molecular traits linked to task division in *B. terrestris* and *T. angustula*,
73 we also verified whether common genes and pathways could be involved in task
74 specialization across all the eusocial bee groups.
75
76 **Results**
77

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

78 *Reference transcriptome assemblies*

79 For both species we used as reference a transcriptome set of
80 superTranscripts[25], in which multiple transcripts from the same gene are represented
81 in a single sequence. *B. terrestris* workers had 27,987 superTranscripts of which 431
82 are potentially lncRNAs and 21,638 (77,3%) were annotated. The final *T. angustula*
83 assembly had 33,065 superTranscripts, and was largely complete. Indeed, 26,623
84 superTranscripts (80.5%) had a high sequence similarity to known protein-coding
85 genes from other species in the UniRef90 database, and 347 were considered
86 lncRNAs (transcriptomes available at
87 https://github.com/nat2bee/Foragers_vs_Nurses). A summary of major quality
88 parameters from the two species datasets can be found on Table SI.
89
90 *Differential expression analyses in* <u>*Bombus terrestris*</u>

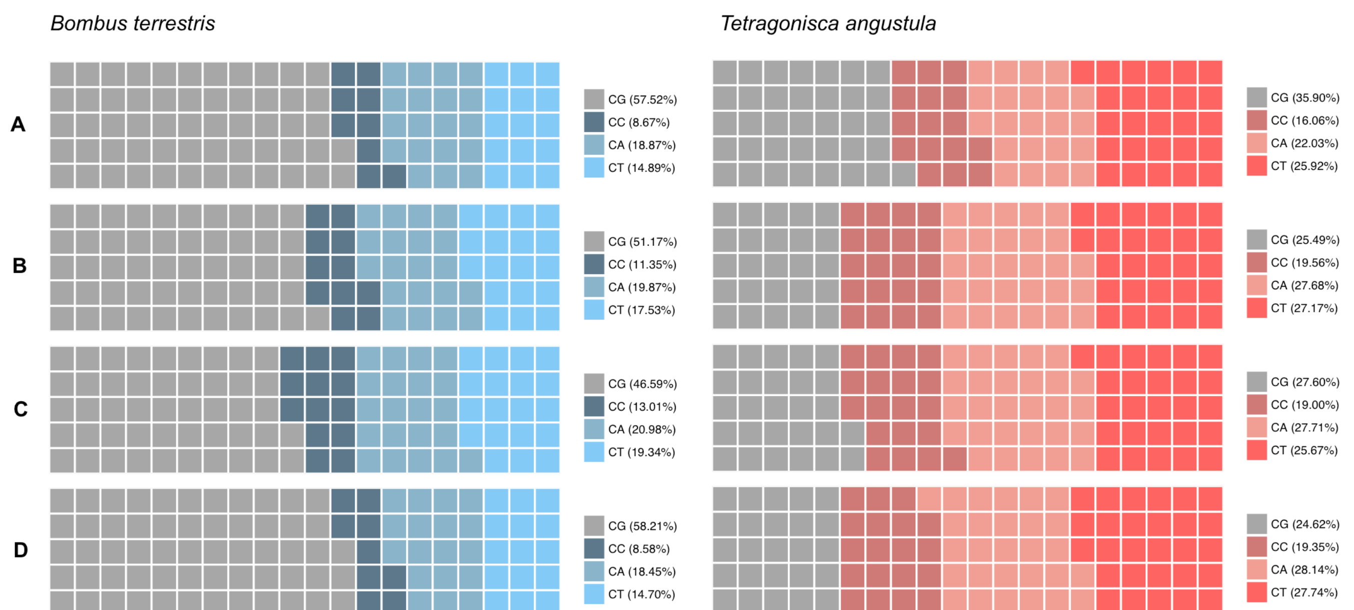
91 Since task division in *B. terrestris* workers is a plastic behaviour[9,21], we
92 performed a principal component analysis of the normalized read counts as an
93 additional verification step to validate our sampling method. The main components
94 clearly clustered nurses and foragers samples separately (Figure S1) indicating that
95 our sampling method was efficient to obtain two distinct groups in bumblebee
96 workers, here considered as nurses and foragers due to the activities they were
97 performing when sampled. We found 1,203 differentially expressed superTranscripts
98 between the two worker groups (Figure S2), whereby 436 superTranscripts were more
99 highly expressed in nurses (Supplementary file S2) and 767 were more highly
100 expressed in foragers (Supplementary file S3). The majority of these superTranscripts
101 (77.3% and 72.6% respectively) have similarity to known protein-coding genes, while
102 respectively three and one are possible long non-coding RNAs (lncRNAs). Five Gene

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

103 Ontology (GO) biological processes terms ("transposition", "DNA-mediated,
104 transposition"; "DNA integration"; "DNA recombination"; and "pseudouridine
105 synthesis") were overrepresented among the differentially expressed superTranscripts
106 ($p < 0.01$; Table SII).
107
108 *Differential expression analyses in <u>Tetragonisca angustula</u>*
109 In workers of *T. angustula* 241 superTranscripts were differentially expressed
110 between nurses and foragers (Figure S2). Among these, 179 had higher levels of
111 expression in nurses, being 157 genes with a significant blast hit to protein databases
112 (Supplementary file S4). Foragers had 62 superTranscripts reported as more highly
113 expressed than in nurses of which 59 were annotated (Supplementary file S5).
114 Enrichment analyses revealed 30 GO terms for biological process (BP) as enriched in
115 the tested set of differentially expressed superTranscripts when compared to the entire
116 transcriptome ($p < 0.01$; Table SII), including processes related to mitochondrial
117 metabolism ("aerobic respiration"; "respiratory electron transport chain"; "oxidative
118 phosphorylation" and "mitochondrial ATP synthesis coupled electron transport") and
119 other metabolic process ("lipid metabolic process" and "carbohydrate metabolic
120 process").
121
122 *DNA methylation in worker genes*
123 Whole bisulfite sequencing (WBS) from *B. terrestris* and *T. angustula* nurses
124 were used to screen DNA methylation patterns in the entire transcriptome and among
125 the differentially expressed superTranscripts. Because *T. angustula* lacks a reference
126 genome and most DNA methylation reported in bees occur within gene exons[14], we
127 performed methylation analyses by mapping bisulfite sequenced reads to the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

128 transcriptomes and not genomes (complete estimations available at
129 https://github.com/nat2bee/Foragers_vs_Nurses) . In *B. terrestris* 23.14 % of all
130 cytosine sites are in CG (cytosine/guanine) context. This is a higher proportion than in
131 *T. angustula* where 15.44 % of all C sites available occur in CG context. We find this
132 explains the higher proportion of CG methylation observed in the bumblebee (Figure
133 1). Nevertheless, in both species DNA methylation in CG context was enriched, that
134 is there was more DNA methylation in CG context than it would be expected simply
135 based on the proportion of sites available. Furthermore, superTranscripts general
136 methylation (mC) levels are higher in *T. angustula* (mean mC 1.24 %) than in *B.*
137 *terrestris* (mean mC 0.66 %) (Figure 2).

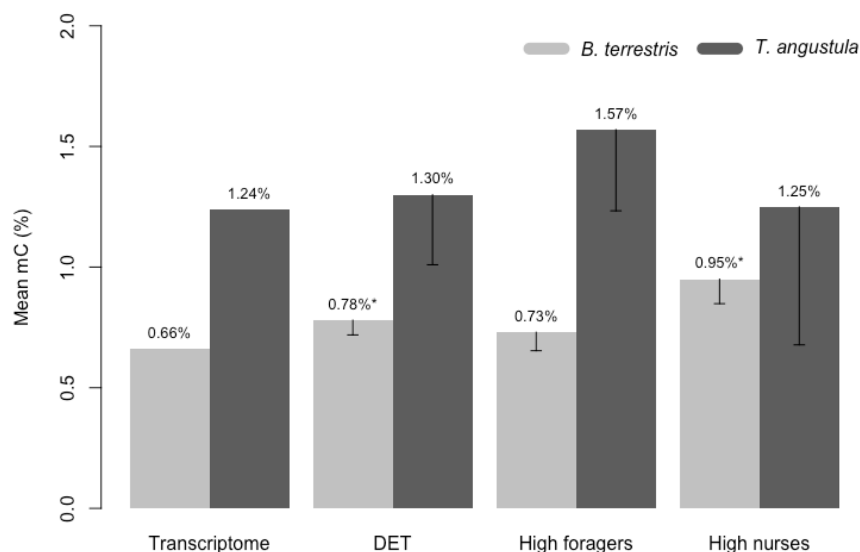


139 **Figure 1** Nucleotide context in which the methylated cytosines occur proportionally to all
140 methylated cytosines reported in nurses of *B. terrestris* and *T. angustula*, in distinct gene sets.
141 **A** – in the entire transcriptome; **B** –in the differentially expressed superTranscripts between
142 foragers and nurses; **C** – in the superTranscripts with higher expression levels in foragers; **D** –
143 in the superTranscripts with higher expression levels in nurses. Grey squares represent
144 methylation at CG context; methylation in non-CG context is illustrated in different shades of
145 blue for *B. terrestris* and in red shades for *T. angustula*. One square ≈ 1%, considering all mC
146 reported sums up to 100%.

147

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

148 In both species the differentially expressed superTranscripts had higher levels
149 of methylation than the overall transcriptomic mean (Figure 2), however only in *B.*
150 *terrestris* this difference was significant (*B. terrestris* p = 6.267e-4, *T. angustula* p =
151 0.3669 at 95% CI). Interestingly, while in *B. terrestris* this increase was mostly due to
152 the greater methylation level of superTranscripts highly expressed in nurses; the mean
153 mC level of the highly expressed superTranscripts in *B. terrestris* nurses was 43.93%
154 higher than the global transcriptomic mean (p = 1.339e-06 at 95% CI). In *T.*
155 *angustula* superTranscripts highly expressed in foragers were the more methylated
156 ones (Figure 2), although still not at a significant level when compared to the general
157 mean (p = 0.05355 at 95% CI). The nucleotide context in which the methylated
158 cytosines occurred also varied in each gene subset (Figure 1). There was an overall
159 reduction in the contribution of CG methylation in the subset of differentially
160 expressed superTranscripts when compared to the entire transcriptome, except for
161 superTranscripts highly expressed in *B. terrestris* nurses (Figure 1D).



162

163 **Figure 2** Mean mC levels in distinct gene sets of *B. terrestris* and *T. angustula* nurses.
164 Transcriptome – refers to the values observed in the complete transcriptome; DET –
165 differentially expressed superTranscripts between nurses and foragers; High foragers –

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

166 superTranscripts with higher expression levels in foragers when compared to nurses; High
167 nurses – superTranscripts with higher expression levels in nurses when compared to foragers.
168 * significantly different from the global transcriptomic mean, with $p < 0.01$ at 95% CI;
169 confidence interval bars are shown.
170

171 Combined these findings suggest a correlation between mC and gene
172 expression depending on the methylation context. Indeed, we identified a positive
173 correlation between global transcript expression levels and CG methylation at both
174 species (*B. terrestris* $r_s = 0.23$ and *T. angustula* $r_s = 0.24$) but not with CW (CA –
175 cytosine/adenine or CT – cytosine/thymine) methylation (*B. terrestris* $r_s = 0.08$ and *T.*
176 *angustula* $r_s = -0.07$). Curiously, when we used only the set of differentially expressed
177 superTranscripts, no correlation was found between gene expression and mC in *B.*
178 *terrestris*, neither in CG ($r_s = 0.08$) nor in CW ($r_s = -0.06$) context. However, in *T.*
179 *angustula*, both types of methylation correlated negatively with gene expression in
180 this scenario (CG $r_s = -0.31$; CW $r_s = -0.35$). This suggests that DNA methylation
181 indeed plays a role in subcaste task division of other eusocial bee species, as in
182 honeybees, but in a more complex way than previously recognized.
183

184 *Comparative analyses of genes involved in task division in the two species*

185 In order to recognize species-specific from shared molecular mechanisms,
186 different strategies were used. First, we asked whether the exact same genes were
187 commonly involved in the observed subcaste differences of *T. angustula* and *B.*
188 *terrestris*. Comparing the two sets of differentially expressed superTranscripts we
189 identified 15 genes in common (Table I; Figure 3C), which is significantly more than
190 it would be expected by chance ($p = 6e-04$, mean number of genes expected 7.04,
191 SD=2.58). Interestingly, the expression pattern of these genes was not always
192 equivalent in both species (Table I). Seven genes were commonly highly expressed in
193 nurses of both species when compared to foragers, ($p = 4e-04$, mean number of genes

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

194 expected 2.16, SD=1.45), but only two genes were commonly highly expressed in

195 foragers (p = 0.3062, mean number of genes expected 1.1, SD=1.03).

196

197 **Table I** List of genes in common to the sets of differentially expressed superTranscripts
198 between nurses and foragers of *T. diversipes* and *B. terrestris*. † Indicate genes commonly
199 highly expressed in nurses of the two species when compared to foragers. ᐞ Indicate genes
200 commonly highly expressed in foragers of the two species when compared to nurses.

Gene	GO terms associated
mucin-2-like	-
basement membrane-specific heparan sulfate proteoglycan core protein isoform x1	-
† *transposable element tc1 transposase*	GO:0015074 [DNA integration]; GO:0006313 [transposition, DNA-mediated]; GO:0003677 [DNA binding]
urea transporter 1-like	-
† *chymotrypsin-2-like*	-
† *cytochrome c oxidase subunit 1*	GO:0009060 [aerobic respiration]; GO:0006119 [oxidative phosphorylation]; GO:0004129 [cytochrome-c oxidase activity]; GO:0020037 [heme binding]; GO:0005506 [iron ion binding]
fatty acyl-coa reductase cg5065	-
ᐞ *at-rich interactive domain-containing protein 2*	-
ᐞ *cytochrome p450*	GO:0020037 [heme binding]; GO:0005506 [iron ion binding]; GO:0004497 [monooxygenase activity]; GO:0016705 [oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen]
† *cathepsin l*	-
† *retrovirus-related pol polyprotein from transposon tnt 1-94*	GO:0015074 [DNA integration]; GO:0003676 [nucleic acid binding]; GO:0008270 [zinc ion binding]
sec23-interacting protein	GO:0046872 [metal ion binding]
† *targeting protein for xklp2*	GO:0032147 [activation of protein kinase activity]; GO:0060236 [regulation of mitotic spindle organization]
tubulin glycylase 3a	GO:0006464 [cellular protein modification process]
† *histone h3*	GO:0003677 [DNA binding]

201

202 Secondly, we investigated whether the same molecular pathways could be

203 involved in the task division of the two species. For this, we searched for similarities

204 among the biological processes to which the differentially expressed superTranscripts

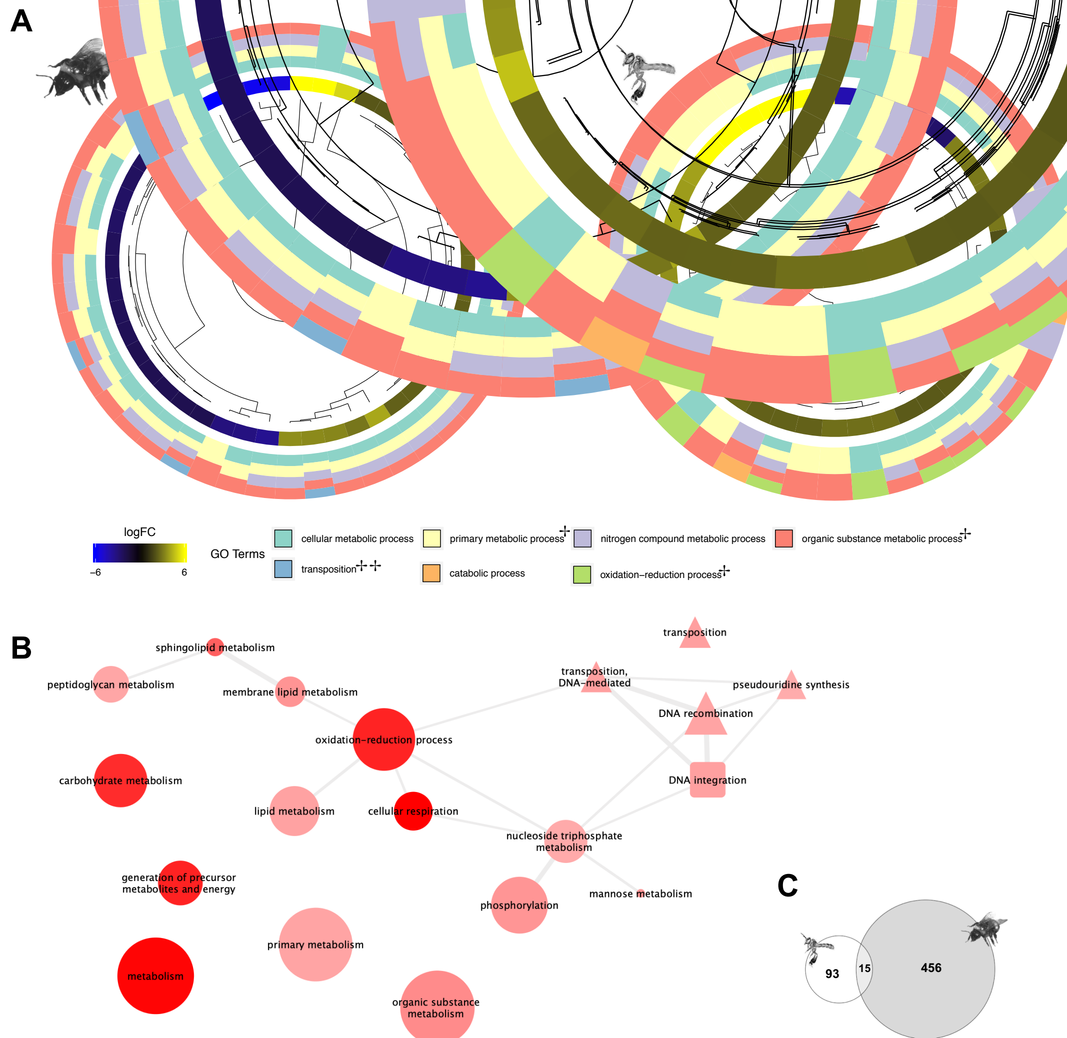
205 were related. We used a comparative approach based on GO subgraphs of the

206 enriched terms. This type of subgraph relies on the hierarchical graphic structure

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

207 among GO terms, where parent terms are more general and less specialized than child
208 terms[26,27]. Consequently, using subgraphs it is possible to compare not only the
209 enriched terms themselves but also their hierarchical connections, reducing gene
210 annotation bias[28]. In this comparison (Figures S3, S4 and S5) we found that the
211 enriched GO terms of the two species were associated and eventually all of them
212 nested under two main processes (Figure S3): "metabolic process" (GO:0008152) and
213 "cellular process" (GO:0009987). Thus, although specific enriched terms are distinct
214 in both species (only "DNA integration" is commonly enriched), this divergence
215 disappears at the parental levels of the topology and almost all terms in *B. terrestris*
216 subgraph are also contained in *T. angustula* subgraph (Figure S3). At the third
217 hierarchical level (Figure 3A), lineage specific GO processes start to emerge such as
218 "transposition" (GO:0032196) in *B. terrestris*, and "catabolic process" (GO:0009056)
219 and "oxidation-reduction process" (GO:0055114) in *T. angustula*. Nevertheless,
220 superTranscripts showing the greatest differences in expression within species (i.e.
221 higher absolute mean logFC between groups) are not the ones related to these species-
222 specific processes (Figure 3A). The connection between the enriched GO terms in the
223 set of differentially expressed superTranscripts of *B. terrestris* and that of *T.*
224 *angustula* can also be visualized on semantic similarity-based clusters (Figure 3B).
225
226

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



227

228

229 **Figure 3** Comparisons between *B. terrestris* and *T. angustula* GO processes involved in task
230 specialization. **A** – Hierarchical clustering of the differentially expressed superTranscripts
231 with the third hierarchical level of GO annotation organized by their mean logFC difference
232 between nurses and foragers. Outer circle colours show to which GO term the gene could be
233 associated to. ++ BP term enriched in the set of differentially expressed superTranscripts of *B.*
234 *terrestris*; + BP term enriched in the set of differentially expressed superTranscripts of *T.*
235 *angustula*. **B** – Similarity network of the enriched GO terms, after semantic similarity-based
236 reduction. GO terms that are very similar to each other are linked and the line width indicates
237 the degree of similarity. Edge shape indicates whether the shown term is enriched in *B.*
238 *terrestris* (circle), in *T. angustula* (triangle) or in both species (square). Edge colour intensity
239 indicates the p-value in the enrichment test (the darker the colour tone, the smaller the p-
240 value). Edge size indicate the frequency of the GO term in the entire UniProt database. **C** –
241 Euler diagram showing the number of genes in common between the set of differentially
242 expressed superTranscripts of each species.

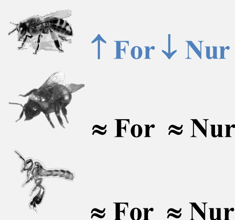
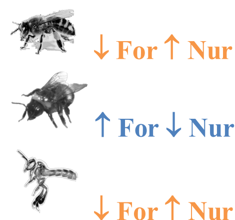
243

244 Discussion

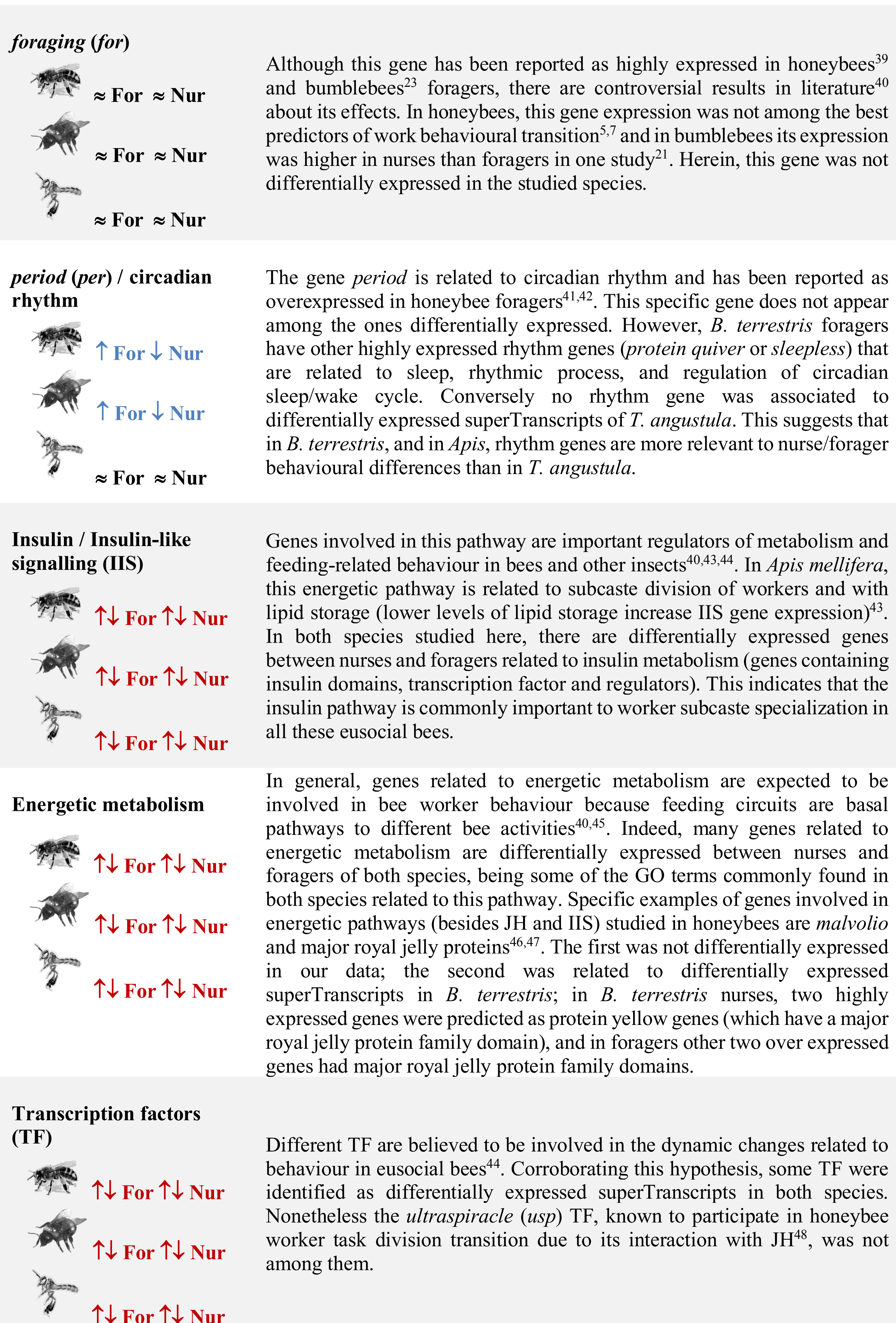
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The comparison of present findings in *B. terrestris* and *T. angustula* with previously published information about task specialization in *Apis* workers are summarized on Table II. Because the literature about this topic in honeybees is extensive and these studies applied distinct methodologies of sampling, expression estimation and data analyses, we restricted our comparisons to a review of genes and molecular pathways commonly highlighted across studies.

Table II Genes and molecular pathways commonly highlighted in literature as involved in honeybee worker task division compared to present findings in *B. terrestris* and *T. angustula*. For – foragers; Nur – Nurses. Symbols indicate if: evidences suggest that the expression is higher (↑) or lower (↓) in one group compared to the other, in blue if foragers have higher expression levels than nurses and in orange if the opposite occurs; (≈) in black, no changes identified or controversial evidences; and (↑↓) in red, indicate a mixed pattern, with some genes in the pathway highly expressed and others with reduced expression level in one of the two subcastes.

juvenile hormone (JH)  **↑ For ↓ Nur** **≈ For ≈ Nur** **≈ For ≈ Nur**	These hormones are important regulators in honeybee maturation affecting task division system in workers[29]. In honeybees, foragers have higher levels of JH than nurses[4,5,29] but in primitively eusocial bees, changes in JH appear not to affect worker behaviour[22]. This led to the hypothesis that JH might only be involved with age related task division[30,31]. In the present dataset we did not find any direct evidence of the involvement of JH in age related task division of *T. angustula* workers. This agrees with previous studies about JH in stingless bees, which have demonstrated that JH expression differences are important in differentiating queens and workers but not nurses and foragers, although titter levels of JH are significantly reduced in foragers[32]. One transcript in our dataset, highly expressed in *B. terrestris* foragers, was indirectly related to JH pathways, a gene predicted as "takeout-like". This gene family has been associated with multiple processes in insects, including eusocial insects, in which it has been shown to be strongly sensitive to queen pheromone[33].
***vitellogenin* (*vg*)**  **↓ For ↑ Nur** **↑ For ↓ Nur** **↓ For ↑ Nur**	This yolk precursor protein is related to egg production in many insects[34]. In honeybees it interacts with JH in a double repressor network, and its expression is reduced in foragers[4,5,34]. For bumblebees this double repressor network apparently does not exist, instead this protein gene has been associated with worker aggression[22] and reproductive status when expressed in the fat body[35]. In our *B. terrestris* data, two genes highly expressed in foragers have *vg* transcription factor domains. As a primitively eusocial species, bumblebee workers may dispute reproductive status with queens in later stages of the colony cycle[36]. Therefore, it would be interesting to further investigate if the higher expression of these vg associated genes in foragers could be related to this behaviour. In *T. angustula*, we found in nurses a higher expression of one *vg* receptor gene indicating the relevance of this protein in this subcaste, as in honeybees. Nevertheless, it is worth noticing that stingless bees workers usually produce trophic eggs[37], so *vg* might be involved in this process or even have alternative unknown roles, as suggested in[38].

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



***foraging* (*for*)** ≈ For ≈ Nur ≈ For ≈ Nur ≈ For ≈ Nur	Although this gene has been reported as highly expressed in honeybees[39] and bumblebees[23] foragers, there are controversial results in literature[40] about its effects. In honeybees, this gene expression was not among the best predictors of work behavioural transition[5,7] and in bumblebees its expression was higher in nurses than foragers in one study[21]. Herein, this gene was not differentially expressed in the studied species.
***period* (*per*) / circadian rhythm** ↑ For ↓ Nur ↑ For ↓ Nur ≈ For ≈ Nur	The gene *period* is related to circadian rhythm and has been reported as overexpressed in honeybee foragers[41,42]. This specific gene does not appear among the ones differentially expressed. However, *B. terrestris* foragers have other highly expressed rhythm genes (*protein quiver* or *sleepless*) that are related to sleep, rhythmic process, and regulation of circadian sleep/wake cycle. Conversely no rhythm gene was associated to differentially expressed superTranscripts of *T. angustula*. This suggests that in *B. terrestris*, and in *Apis*, rhythm genes are more relevant to nurse/forager behavioural differences than in *T. angustula*.
Insulin / Insulin-like signalling (IIS) ↑↓ For ↑↓ Nur ↑↓ For ↑↓ Nur ↑↓ For ↑↓ Nur	Genes involved in this pathway are important regulators of metabolism and feeding-related behaviour in bees and other insects[40,43,44]. In *Apis mellifera*, this energetic pathway is related to subcaste division of workers and with lipid storage (lower levels of lipid storage increase IIS gene expression)[43]. In both species studied here, there are differentially expressed genes between nurses and foragers related to insulin metabolism (genes containing insulin domains, transcription factor and regulators). This indicates that the insulin pathway is commonly important to worker subcaste specialization in all these eusocial bees.
Energetic metabolism ↑↓ For ↑↓ Nur ↑↓ For ↑↓ Nur ↑↓ For ↑↓ Nur	In general, genes related to energetic metabolism are expected to be involved in bee worker behaviour because feeding circuits are basal pathways to different bee activities[40,45]. Indeed, many genes related to energetic metabolism are differentially expressed between nurses and foragers of both species, being some of the GO terms commonly found in both species related to this pathway. Specific examples of genes involved in energetic pathways (besides JH and IIS) studied in honeybees are *malvolio* and major royal jelly proteins[46,47]. The first was not differentially expressed in our data; the second was related to differentially expressed superTranscripts in *B. terrestris*; in *B. terrestris* nurses, two highly expressed genes were predicted as protein yellow genes (which have a major royal jelly protein family domain), and in foragers other two over expressed genes had major royal jelly protein family domains.
Transcription factors (TF) ↑↓ For ↑↓ Nur ↑↓ For ↑↓ Nur ↑↓ For ↑↓ Nur	Different TF are believed to be involved in the dynamic changes related to behaviour in eusocial bees[44]. Corroborating this hypothesis, some TF were identified as differentially expressed superTranscripts in both species. Nonetheless the *ultraspiracle* (*usp*) TF, known to participate in honeybee worker task division transition due to its interaction with JH[48], was not among them.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DNA methylation / epigenetic modifications

↑↓ For ↑↓ Nur

↑↓ For ↑↓ Nur

↑↓ For ↑↓ Nur

DNA methylation is known to participate in nursing to foraging transition in honeybees[16,17]. In the two species investigated in the present study genes possibly related to epigenetic changes were also differentially expressed. In *T. angustula* histone genes (H3 and H2B) and a methyltransferase were differentially expressed and in *B. terrestris* histone H3-K4 demethylation and lncRNAs were reported. All these genes were highly expressed in nurses, except for one lncRNA over expressed in *B. terrestris* foragers.

260

261 Our comparisons sought to differentiate species-specific from common
262 molecular mechanisms involved in worker task division across all eusocial lineages of
263 corbiculate bees. Species-specific mechanisms were mostly related to the expression
264 pattern of genes. Many of the differentially expressed genes were not common to all
265 species, and among the ones that were, the pattern of expression was not necessarily
266 the same. Genes highly expressed in one species subcaste were often down expressed
267 in the same subcaste of the other species. For instance, genes related to circadian
268 rhythm are highly expressed in foragers of *B. terrestris* and *Apis*[41,42], but not in *T.*
269 *angustula* foragers. Moreover, genes related to yolk production, such as *vg* related
270 genes, are commonly highly expressed in nurses of *T. angustula* and *Apis* [4,5,34], but
271 not in *B. terrestris* nurses. This is not unexpected since each lineage forgo unique
272 selective pressures, despite presenting similar behaviours[49]. Even closely related
273 species (within the same taxonomic genus) are known to differ in the expression
274 pattern of certain genes[11]. This implies that the expression profile of particular genes,
275 identified through the study of one single species, should not be directly extrapolated
276 to explain other species responses.
277 A clear illustration of how misleading these assumptions can be is the *vg*/JH
278 network, which has been largely studied in honeybees. Honeybee nurses present
279 higher levels of *vg* and lower levels of JH when compared to foragers. Once workers
280 become foragers their levels of JH increases, which in turn represses the *vg*

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

281 expression in a double repressor network[4,48]. In bumblebees, as demonstrated
282 previously[22] and corroborated by our data, this network is not regulated in the same
283 manner. In this bee, genes related to JH and *vg* are both highly expressed in foragers.
284 For *T. angustula* we found supporting evidence of the involvement of *vg* in nursing
285 behaviour, nonetheless JH genes were not highly expressed in foragers. This supports
286 the hypothesis that in stingless bees the typical *vg/*JH double repressor network
287 observed in honeybees is also not functional, and *vg* is distinctly regulated[32,38].
288 Nevertheless, gene expression dynamics in worker behaviour is not
289 completely unrelated across eusocial bees. Beyond the literal expression trend, we
290 still found a significant number of genes commonly differentially expressed between
291 nurses and foragers of *B. terrestris* and *T. angustula*. Interestingly, some of these
292 common genes were also shown to be responsive to queen pherormone[33]. Genes like
293 *cytochrome p450*, *fatty acyl-CoA*, mitochondrial and histone related were found to be
294 sensitive to queen pheromone in ants and bees[33]. Moreover, biological processes
295 terms enriched in each species set of differentially expressed superTranscripts were
296 highly comparable. Our comparisons of enriched GO terms subgraphs highlighted
297 broader similarities between *B. terrestris* and *T. angustula*, indicating how distinct
298 GO terms (and genes) were involved in similar biological processes. In general,
299 biological terms related to energetic and metabolic processes ("organic substance
300 metabolic process"; "primary metabolic process"; "nitrogen compound metabolic
301 process"; and "cellular metabolic process") were central to subcaste differentiation of
302 both species.
303 The relevance of metabolic pathways to insect sociality has been demonstrated
304 in many studies over the years[45,50–52] and it is most certainly not a species-specific
305 trait. These pathways are affected by queen pheromone in different species and are

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

306 involved with caste determination of multiple hymenopteran lineages, including bees,
307 ants and wasps[33,53]. Given the central role of energetic and metabolic maintenance in
308 any living animal it is not surprising that changes in these pathways will affect a
309 number of features, including behavioural phenotypes. It is however fascinating to
310 observe how plastic and dynamic, in terms of gene regulation, these networks can be,
311 with different lineages frequently evolving unique responses but still being sensible to
312 similar cues (like queen pheromone).

313 This mosaic pattern of species-specific features involved in common
314 molecular processes is also observed in the epigenetic machinery. Transcriptomic and
315 WBS data support the involvement of DNA methylation and other epigenetic factors
316 in worker specialization of the two analysed species. Genes involved in epigenetic
317 alterations were found among the differentially expressed superTranscripts of *T.*
318 *angustula* and *B. terrestris*, and the species global methylation patterns were distinct
319 from that of their differentially expressed superTranscripts. The differentially
320 expressed superTranscripts had overall more mC (Figure 2) and less CG methylation
321 (Figure 1). Still, a closer investigation revealed distinct specific epigenetic
322 mechanisms in each species. To begin with, epigenetic related genes that are
323 differentially expressed in each species are different. Likewise, only in *T. angustula*,
324 the genes highly expressed in foragers were more methylated at CG context and had
325 higher mean mC levels than genes overexpressed in nurses. The opposite was found
326 in *B. terrestris*. Considering that WBS data was obtained from nurses in both species,
327 these distinct patterns are quite unexpected.

328 Although in the studied species DNA methylation was frequent in CG context,
329 methylation within other nucleotides contexts also occurred (i.e. non-CG or non-CpG
330 methylation). Non-CG DNA methylation is frequently associated with a number of

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

331 processes in plants[54,55] and only recently its function in other eukaryotes have gained
332 more attention[56]. Still, the effects of differential DNA methylation contexts in most
333 organisms are poorly understood and underestimated (reviewed in[56,57]). Methylation
334 in CG and non-CG sites are typically mediated by distinct mechanisms[58]; CG
335 methylation is constitutively maintained by DNA methyltransferase 1 (Dnmt1)[56,57],
336 while non-CG methylation are kept by mechanisms of *de novo* methylation involving
337 the DNA methyltransferase 3 (Dnmt3)[59]. Therefore, non-CG methylation is majorly
338 related to new and more variable epigenetic alterations[57]. Supporting evidence for the
339 existence of non-CG methylation in social insects was previously reported for ants[60]
340 and honeybees[59]. In honeybees, non-CG methylation seems to be involved with
341 alternative mRNA splicing and is especially enriched in genes related to behavioural
342 responses. However, no direct correlation with sociality could be stablished[59]. Herein,
343 such correlation is demonstrated with the different proportions of CG and non-CG
344 methylation observed in the set of differentially expressed superTranscripts when
345 compared to the general transcriptomic profile. This indicate that both CG and non-
346 CG methylation interplay in worker task division. Further data is needed to infer how
347 specific methylation contexts could affect certain behavioural changes but based on
348 the results gathered so far, we hypothesize that non-CG methylation dynamics is
349 relevant to task division and possibly to other social traits.
350 Higher levels of mC in bees have been associated to an increase in gene
351 expression, i.e. genes with more methylation also have higher expression levels[14]. We
352 found this correlation to be true for CG methylation in both species tested, but not for
353 methylation in non-CG context. In fact, among the differentially expressed
354 superTranscripts of *T. angustula*, where greater levels of non-CG methylation are
355 observed, we found a negative correlation between gene expression and DNA

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

356 methylation. This suggests that the effect of mC in bee gene expression might also be
357 dependent of the methylation context; CG methylation seems to increase gene
358 expression while non-CG methylation might supress it.
359 Finally, it is important to consider some of the limitations of the present study.
360 First, aiming to obtain a global overview of gene expression and DNA methylation
361 differences we used full bodies for the transcriptomic and bisulfite sequencings. Since
362 we know that different body parts, tissues and even cells have unique gene expression
363 dynamics[12] it is likely that our approach had reduced our power to detect small scale
364 alterations and specific contexts. Moreover, to facilitate the comparisons between the
365 two bees we used similar pipelines for them. This means that sometimes we
366 compromised the bumblebee analysis to match it with the analysis of the species with
367 no reference genome available. For example, we annotated both species
368 transcriptomes based on search similarities to databases instead of using *B. terrestris*
369 genome for its annotation. This approach might especially affect GO enrichment
370 analysis. Differently from genome annotation, transcriptomic annotation is redundant,
371 i.e. multiple transcripts (or superTranscripts in our case) may annotate to the same
372 gene and this affects the frequency of GO terms in the dataset. To deal with this, we
373 kept the frequency of GO terms proportional in the enrichment test by using
374 the appropriate background list (in our case the complete transcriptome set), which is
375 the used and recommended approach for GO enrichment tests[61]. However, our
376 enrichment stats might still be biased by the chosen approach. Nevertheless, since GO
377 annotations are dynamic and always biased by database representation[62], we have
378 chosen to keep the same methodological approach for both species. In this manner, if
379 the enrichment test is biased it will be equally biased in both species facilitating
380 comparisons. Finally, we did not validate our gene expression results with an

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

381 alternative independent method (such as real time reverse polymerase chain reaction).
382 Given due consideration, the present study can only describe broad patterns and
383 conclusions regarding the species general expression and methylation profiles.
384 Further works should address detailed and more subtle differences.

385 Through the analyses of the global transcriptomic and DNA methylation
386 profiles of subcastes from two eusocial bee lineages, we gather an important dataset
387 for the study of social behaviour evolution. These data aligned to a review of the
388 honeybee literature, allowed comparisons among all eusocial corbiculate bee groups;
389 Apini, Bombini and Meliponini. Main findings support the hypothesis that common
390 and more ancient molecular mechanisms are involved in worker task division across
391 these species, standing as central among them energetic and metabolic pathways, and
392 epigenetic factors. However, despite these similarities, particular gene expression
393 patterns tend to be species-specific. This scenario could be explained by later
394 specialization of species-specific molecular responses to ancient social cues which
395 left a mosaic profile in worker task division, where unique and shared features are
396 found. Moreover, results indicate that non-CG methylation is relevant to worker
397 behavioural dynamics and that it might affect gene expression differently from CG
398 methylation. As a result, the involvement of non-CG methylation in other social traits
399 should be further investigated.
400
401 **Material and Methods**
402
403 *Sample collection and sequencing*
404 Bee species were chosen based on their behaviour (primitively eusocial and
405 highly eusocial), phylogenetic relationship (corbiculate bees[24]), and sampling

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

406 convenience. Samples were from three colonies per species. *B. terrestris* colonies
407 were obtained from commercial suppliers (Biobest®) and kept in lab condition at
408 Queen Mary University of London (England). All bees in the colonies were marked
409 and housed in wood boxes attached to foraging arenas. After 16 days of adaptation all
410 recently born workers received an individual number tag; individuals used in the
411 analyses were all tagged. Bumblebee workers usually do not forage right after
412 emergency[63], therefore we waited for five more days before start sampling. For *T.*
413 *angustula*, colonies regularly kept at the Laboratório de Abelhas (University of São
414 Paulo – Brazil) were used for sample collection.

415 Workers subcaste were determined in two different ways. First, for *B.*
416 *terrestris*, nurses were selected based on observation. Colonies were observed for one
417 day during all their active foraging period (6 hours uninterruptedly). Tagged bees who
418 stayed inside of the nest during the entire period, never entering the foraging arenas,
419 were considered nurses. In the following day, nurses and foragers were collected and
420 immediately frozen in liquid nitrogen. Foragers were sampled first, while collecting
421 nectar in the foraging arena. Nurses were posteriorly collected inside of the colonies.
422 Then, for *T. angustula*, nurses were defined by age. Brood cells (close to emergency)
423 were removed from the colonies and transferred to an incubator with controlled
424 temperature and humidity. Upon emergency, female workers were marked with
425 specific colours using a water-based ink and immediately returned to the colony. Ten
426 to twelve days after their emergency and reintroduction, colonies were opened and
427 marked individuals were sampled. During this age worker bees from *T. angustula*
428 present nursing behaviour[37]. Foragers were collected while leaving and returning to
429 the colonies from foraging trips. To prevent sampling of guard workers[2], bees
430 standing in front of the colony entrance were avoided. Some of the foragers were

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

431 collected before nurse sampling and others after this period, but no foragers were
432 sampled while nurses were marked and collected so as to avoid effects of colony
433 disturbance in the worker behaviour. Nurses from different colonies were collected in
434 different days.

435 All individuals were sampled between 10h-12h for both species, and entire
436 worker bodies were used for RNA and DNA extraction. For RNA-Seq, six *T.*
437 *angustula* workers, from the same colony and subcaste, were pooled as one sample. *B.*
438 *terrestris* samples were a pool of RNA extractions from three workers per subcaste/
439 colony. Each colony was considered as one sample replicate. Total RNA was
440 extracted from workers using Qiagen® extraction kit (RNeasy Mini Kits). RNA
441 quality and quantification were verified using the Bionalyzer®, Nanodrop® and
442 Qubit®. Samples were posteriorly used for RNA sequencing on Illumina® HiSeq
443 2000. Library preparation was performed by sequencing providers. *B. terrestris*
444 workers were sequenced by the Genome Center at Queen Mary University of London,
445 and *T. angustula* samples were sequenced at LACTAD (Unicamp). RNA sequencing
446 generated 30-50 million paired reads (100bp) per colony replicate. For whole bisulfite
447 sequencing, total DNA from one nurse (whole body) per species was used for the
448 phenol-chloroform DNA extraction[64]. WBS were performed following the protocol
449 described in[65] using the Illumina® NextSeq500. WBS returned 60-70 million single
450 reads (150 bp) per sample. Sequencing and library preparation were performed at
451 University of Georgia. All sequenced reads are available at BioProject ID
452 PRJNA615177.
453

454 *Transcriptome assembly and differential expression analyses and comparisons*

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

455 Reads quality assessment was performed using the FastQC program[66]
456 (v0.11.2) before and after cleaning. The FASTX Toolkit[67] (v0.0.14) was used to trim
457 the first 14 bp of all reads because an initial GC bias[68] was detected. Low quality
458 bases (phred score below 30) and small reads (less than 31 bp) were removed using
459 SeqyClean[69] (v1.9.3). Samples from nurses and foragers were combined for the
460 assemblies. To increase *de novo* transcriptome assembly efficiency, cleaned reads
461 were digitally normalized[70] (20x coverage). Transcriptome assembly were performed
462 differently for each species. For *B. terrestris*, its genome[71] was used as reference in
463 two approaches. First, using HISAT2[72] (v2-2.0.3) and StringTie[73] (v1.2.2) a regular
464 reference assembly was obtained. Secondly, the Trinity[74] (v2.1.1) program was used
465 to perform a reference guided *de novo* assembly. The two resulting assemblies were
466 merged using CD-Hit[75] (v4.6), Corset[76] (v1.05) and Lace[25] (v0.80) to cluster
467 transcripts into superTranscripts. We have chosen to use this combined approach for
468 *B. terrestris* for two reasons. First, to optimized the transcriptome assembly based on
469 our dataset, a recommended procedure even for species with well-annotated reference
470 genome and transcriptome[77]. Second, to make *B. terrestris* and *T. angustula* datasets
471 more comparable since for the later we have used the clustering method. There is no
472 reference genome for *T. angustula*, therefore we performed a combined *de novo*
473 assembly using two strategies with the Trinity pipeline: a reference guided *de novo*
474 assembly, based on the genome of another stingless bee, *Melipona quadrifasciata*[78];
475 and a complete *de novo* assembly. Afterwards, the two assemblies were merged as in
476 the bumblebee. Assemblies used programs default recommended parameters, CD-Hit
477 was used to merge transcripts with more than 95% similarity, Corset was set to keep
478 transcripts with a minimum of 50x coverage, and Lace was used to obtain the
479 superTranscripts.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

480 SuperTranscripts were then annotated with Annocript[79] (v1.2) using the
481 UniProt Reference Clusters (UniRef90) and the UniProtKB/Swiss-Prot databases[80]
482 (June 2016 version). SuperTranscripts with significant blast hits (e-value < 1e-5)
483 against possible contaminants (plants, fungus, mites and bacteria) in the UniRef90
484 were removed from the final datasets. Finally, only potentially coding
485 superTranscripts (based on blast results and ORF analysis) or possible lncRNAs were
486 kept. This annotation pipeline was used for both species. Quality parameters from the
487 transcriptomes were analysed using QUAST[81] (v4.0), BUSCO[82] (v2), TransRate[77]
488 (v1.0.3) and Qualimap[83] (v2.2).
489 Differential expression analyses were performed in each species
490 independently and compared posteriorly, as illustrated in Figure S6. Bowtie2[84]
491 (v2.2.5), RSEM[85] (v1.2.22) and DESeq2[86] (p-value < 1e-3) were used to identify
492 differentially expressed superTranscripts, using scripts from the Trinity package – just
493 figure parameters were adapted. During analyses we identified a possible batch effect
494 in samples from *T. angustula*: one nurse and one forager replicate were sequenced in
495 different lanes and it seemed to affect sample correlation. This effect was corrected
496 during differential expression analyses following the suggested protocol in DESeq2
497 documentation. No batch effect was identified in *B. terrestris* samples. To test
498 whether any GO term was enriched in a set of differentially expressed
499 superTranscripts compared to the total transcriptome, a classical Fisher's exact test
500 was performed using the R package TopGO[28]. Species comparisons of differentially
501 expressed genes was based on gene annotation, using only unique and non-redundant
502 terms (i.e. those genes not containing "uncharacterized protein" in their annotation).
503 The list of overlapping genes was then manually curated to remove annotation
504 incoherencies not detected computationally, i.e. when gene lists from both species

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

505 were compared with our R script 18 terms were common, after manual curation we
506 removed three genes from this list because of partial or redundant annotation matches
507 ("transposase", "transporter" and "cytochrome c oxidase subunit [fragment]"), leaving
508 15 genes in common. In the random sampling statistics this manual filtering
509 correction was not used, so the numbers of common genes obtained with the
510 computational comparison were used. Comparisons between the set of GO enriched
511 terms and subgraphs was manual. The similarity network parameters was estimated
512 with REVIGO[87] using Medium (0.7) similarity threshold. In the interactive network
513 mode of this program, the input data for Cytoscape[88] was downloaded for further
514 figure edition. Statistical tests of significance for comparisons were based on random
515 sampling using R[89] scripts, p-value smaller than 0.01 were considered significant.
516 Scripts used are available at https://github.com/nat2bee/Foragers_vs_Nurses.
517
518 *DNA methylation analysis*
519 Cleaning and adapter trimming of the bisulfite converted reads were
520 performed using Trim Galore[90] (v 0.4.3) wrapper script with default parameters.
521 Complete transcriptome assemblies were used as reference so DNA methylation of
522 coding regions could be analysed, since these regions are the main methylation targets
523 in bees and other Hymenoptera[14]. PCR bias filtering, alignment of the cleaned reads
524 and methylation call were performed using the BS-Seeker2[91] (v 2.1.0), because this
525 program allows the use of Bowtie2 in local alignment mode, which was necessary to
526 properly align WBS reads to a transcriptome. CGmapTools[92] (v 0.0.1) was used to
527 filter low coverage methylated sites (< 10x) and to obtain DNA methylation statistics,
528 including context use. Remaining statistical tests were performed using R, as follows:
529 a random sampling test was used to verify whether the proportion of CG methylation

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

530 found deviated from what was expected by chance; one-tailed z-test was used to test
531 whether differences between the mean methylation observed in the set of
532 superTranscripts was different from the general transcriptomic mean; and the
533 correlation between methylation and gene expression was calculated using
534 Spearman's correlation coefficient between the superTranscript mean methylation and
535 its normalized read count. Scripts used are available at
536 https://github.com/nat2bee/Foragers_vs_Nurses.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

537 **References**

538

539 1. Grozinger, C. M., Fan, Y., Hoover, S. E. R. & Winston, M. L. Genome-wide analysis reveals
540 differences in brain gene expression patterns associated with caste and reproductive status in
541 honey bees (Apis mellifera). *Mol. Ecol.* **16**, 4837–4848 (2007).
542 2. Grüter, C. *et al.* Repeated evolution of soldier sub-castes suggests parasitism drives social
543 complexity in stingless bees. *Nat. Commun.* **8**, e4 (2017).
544 3. Robinson, G. E., Fahrbach, S. E. & Winston, M. L. W. Insect societies and the molecular
545 biology of social behavior. *Bioessays* **19**, 1099–1108 (1997).
546 4. Guidugli, K. R. *et al.* Vitellogenin regulates hormonal dynamics in the worker caste of a
547 eusocial insect. *FEBS Lett.* **579**, 4961–4965 (2005).
548 5. Whitfield, C. W. *et al.* Genomic dissection of behavioral maturation in the honey bee. *Proc.*
549 *Natl. Acad. Sci. U. S. A.* **103**, 16068–16075 (2006).
550 6. Engels, W. & Imperatriz-Fonseca, V. L. Caste Development, Reproductive Strategies, and
551 Control of Fertility in Honey Bees and Stingless Bees. in *Social Insects: An Evolutionary*
552 *Approach to Castes and Reproduction* (ed. Engels, P. D. W.) 167–230 (Springer Berlin
553 Heidelberg, 1990).
554 7. Whitfield, C. W., Cziko, A.-M. & Robinson, G. E. Gene Expression Profiles in the Brain
555 Predict Behavior in Individual. *Science (80-.).* **296**, 296–299 (2003).
556 8. Hrncir, M., Jarau, S. & Barth, F. G. Stingless bees (Meliponini): senses and behavior. *J. Comp.*
557 *Physiol. A Neuroethol. Sensory, Neural, Behav. Physiol.* 1–5 (2016). doi:10.1007/s00359-016-
558 1117-9
559 9. Goulson, D. *et al.* Can alloethism in workers of the bumblebee, *Bombus terrestris*, be explained
560 in terms of foraging efficiency? *Anim. Behav.* **64**, 123–130 (2002).
561 10. Couvillon, M. J., Jandt, J. M., Bonds, J., Helm, B. R. & Dornhaus, A. Percent lipid is
562 associated with body size but not task in the bumble bee *Bombus impatiens*. *J. Comp. Physiol.*
563 *A* **197**, 1097–1104 (2011).
564 11. Sen Sarma, M., Whitfield, C. W. & Robinson, G. E. Species differences in brain gene
565 expression profiles associated with adult behavioral maturation in honey bees. *BMC Genomics*
566 **8**, (2007).
567 12. Cervoni, M. S. *et al.* Mitochondrial capacity, oxidative damage and hypoxia gene expression
568 are associated with age-related division of labor in honey bee (Apis mellifera L.) workers . *J.*
569 *Exp. Biol.* **220**, 4035–4046 (2017).
570 13. Chandrasekaran, S. *et al.* Behavior-specific changes in transcriptional modules lead to distinct
571 and predictable neurogenomic states. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 18020–18025 (2011).
572 14. Yan, H. *et al.* DNA Methylation in Social Insects: How Epigenetics Can Control Behavior and
573 Longevity. *Annu. Rev. Entomol.* **60**, 435–452 (2015).
574 15. Cardoso-Júnior, C. A. M., Guidugli-Lazzarini, K. R. & Hartfelder, K. DNA methylation affects
575 the lifespan of honey bee (Apis mellifera L.) workers – Evidence for a regulatory module that
576 involves vitellogenin expression but is independent of juvenile hormone function. *Insect*
577 *Biochem. Mol. Biol.* **92**, 21–29 (2018).
578 16. Lockett, G. A., Kucharski, R. & Maleszka, R. DNA methylation changes elicited by social
579 stimuli in the brains of worker honey bees. *Genes, Brain Behav.* **11**, 235–242 (2012).
580 17. Herb, B. R. *et al.* Reversible switching between epigenetic states in honeybee behavioral
581 subcastes. *Nat. Neurosci.* **15**, 1371–1373 (2012).
582 18. Mateus, S., Ferreira-Caliman, M. J., Menezes, C. & Grüter, C. Beyond temporal-polyethism:
583 division of labor in the eusocial bee Melipona marginata. *Insectes Soc.* **66**, 317–328 (2019).
584 19. Peters, R. S. *et al.* Evolutionary History of the Hymenoptera. *Curr. Biol.* **27**, 1013–1018
585 (2017).
586 20. Martins, A. C., Melo, G. A. R. & Renner, S. S. The corbiculate bees arose from New World
587 oil-collecting bees: Implications for the origin of pollen baskets. *Mol. Phylogenet. Evol.* **80**,
588 88–94 (2014).
589 21. Kodaira, Y., Ohtsuki, H., Yokoyama, J. & Kawata, M. Size-dependent foraging gene
590 expression and behavioral caste differentiation in Bombus ignitus. *BMC Res. Notes* **2**, (2009).
591 22. Amsalem, E., Malka, O., Grozinger, C. & Hefetz, A. Exploring the role of juvenile hormone
592 and vitellogenin in reproduction and social behavior in bumble bees. *BMC Evol. Biol.* **14**, 1–13
593 (2014).
594 23. Tobback, J., Mommaerts, V., Vandersmissen, H. P., Smagghe, G. & Huybrechts, R. Age- and
595 task-dependent *foraging* gene expression in the bumblebee *Bombus terrestris*. *Arch. Insect*
596 *Biochem. Physiol.* **76**, 30–42 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

597 24. Bossert, S. *et al.* Combining transcriptomes and ultraconserved elements to illuminate the
598 phylogeny of Apidae. *Mol. Phylogenet. Evol.* **130**, 121–131 (2019).
599 25. Davidson, N. M., Hawkins, A. D. K. & Oshlack, A. SuperTranscripts: A data driven reference
600 for analysis and visualisation of transcriptomes. *Genome Biol.* **18**, 148 (2017).
601 26. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* **25**,
602 25–29 (2000).
603 27. Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids*
604 *Res.* **47**, D330–D338 (2019).
605 28. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. (2016).
606 29. Robinson, G. E., Strambi, C., Strambi, A. & Feldlaufer, M. F. Comparison of juvenile hormone
607 and ecdysteroid haemolymph titres in adult worker and queen honey bees (Apis mellifera). *J.*
608 *Insect Physiol.* **37**, 929–935 (1991).
609 30. Cameron, S. A. & Robinson, G. E. Juvenile hormone does not affect division of labor in
610 bumble bee colonies (Hymenoptera, Apidae). *Ann. Entomol. Soc. Am.* **83**, 626–631 (1990).
611 31. Hartfelder, K. Insect juvenile hormone: From 'status quo' to high society. *Brazilian J. Med.*
612 *Biol. Res.* **33**, 157–177 (2000).
613 32. Cardoso-Júnior, C. A. M. *et al.* Methyl farnesoate epoxidase (mfe) gene expression and
614 juvenile hormone titers in the life cycle of a highly eusocial stingless bee, Melipona scutellaris.
615 *J. Insect Physiol.* **101**, 185–194 (2017).
616 33. Holman, L., Helanterä, H., Trontti, K. & Mikheyev, A. S. Comparative transcriptomics of
617 social insect queen pheromones. *Nat. Commun.* **10**, 1–12 (2019).
618 34. Nelson, C. M., Ihle, K. E., Fondrk, M. K., Page, R. E. & Amdam, G. V. The gene *vitellogenin*
619 has multiple coordinating effects on social organization. *PLoS Biol.* **5**, 0673–0677 (2007).
620 35. Lockett, G. A., Almond, E. J., Huggins, T. J., Parker, J. D. & Bourke, A. F. G. Gene expression
621 differences in relation to age and social environment in queen and worker bumble bees. *Exp.*
622 *Gerontol.* **77**, 52–61 (2016).
623 36. Bloch, G. Regulation of queen-worker conflict in bumble-bee (Bombus terrestris) colonies.
624 *Proc. R. Soc. B Biol. Sci.* **266**, 2465–2469 (1999).
625 37. Koedam, D. & Tienen, P. G. M. Van. The regulation of worker-oviposition in the stingless bee.
626 *Insectes Soc.* **44**, 229–244 (1997).
627 38. Dallacqua, R. P., Simões, Z. L. P. & Bitondi, M. M. G. Vitellogenin gene expression in
628 stingless bee workers differing in egg-laying behavior. *Insectes Soc.* **54**, 70–76 (2007).
629 39. Ben-Shahar, Y., Robichon, A., Sokolowski, M. B. & Robinson, G. E. Influence of Gene Action
630 Across Different Time Scales on Behavior. *Science (80-.).* **296**, 741–744 (2002).
631 40. Weitekamp, C. A., Libbrecht, R. & Keller, L. Genetics and Evolution of Social Behavior in
632 Insects. *Annu. Rev. Genet.* **51**, 219–239 (2017).
633 41. Toma, D. P., Bloch, G., Moore, D. & Robinson, G. E. Changes in period mRNA levels in the
634 brain and division of labor in honey bee colonies. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6914–
635 6919 (2000).
636 42. Bloch, G., Rubinstein, C. D. & Robinson, G. E. *period* expression in the honey bee brain is
637 developmentally regulated and not affected by light, flight experience, or colony type. *Insect*
638 *Biochem. Mol. Biol.* **34**, 879–891 (2004).
639 43. Ament, S. A., Corona, M., Pollock, H. S. & Robinson, G. E. Insulin signaling is involved in the
640 regulation of worker division of labor in honey bee colonies. *Proc. Natl. Acad. Sci. U. S. A.*
641 **105**, 4226–4231 (2008).
642 44. Dolezal, A. G. & Toth, A. L. Honey bee sociogenomics: A genome-scale perspective on bee
643 social behavior and health. *Apidologie* **45**, 375–395 (2014).
644 45. Fischer, E. K. & O'Connell, L. A. Modification of feeding circuits in the evolution of social
645 behavior. *J. Exp. Biol.* **220**, 92–102 (2017).
646 46. Ben-Shahar, Y., Dudek, N. L. & Robinson, G. E. Phenotypic deconstruction reveals
647 involvement of manganese transporter *malvolio* in honey bee division of labor. *J Exp Biol* **207**,
648 3281–3288 (2004).
649 47. Buttstedt, A., Moritz, R. F. A. & Erler, S. Origin and function of the major royal jelly proteins
650 of the honeybee (*Apis mellifera*) as members of the *yellow* gene family. *Biol. Rev.* **89**, 255–
651 269 (2014).
652 48. Ament, S. A. *et al.* The Transcription Factor *Ultraspiracle* Influences Honey Bee Social
653 Behavior and Behavior-Related Gene Expression. *PLoS Genet.* **8**, e1002596 (2012).
654 49. Harpur, B. A. *et al.* Queens and Workers Contribute Differently to Adaptive Evolution in
655 Bumble Bees and Honey Bees. **9**, 2395–2402 (2017).
656 50. Fischman, B. J., Woodard, S. H. & Robinson, G. E. Molecular evolutionary analyses of insect

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

657 societies. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl**, 10847–54 (2011).
658 51. Morandin, C. *et al.* Comparative transcriptomics reveals the conserved building blocks
659 involved in parallel evolution of diverse phenotypic traits in ants. *Genome Biol.* **17**, 1–19
660 (2016).
661 52. Chandra, V. *et al.* Social regulation of insulin signaling and the evolution of eusociality in ants.
662 *Science (80-.).* **361**, 398–402 (2018).
663 53. Berens, A. J., Hunt, J. H. & Toth, A. L. Comparative transcriptomics of convergent evolution:
664 Different genes but conserved pathways underlie caste phenotypes across lineages of eusocial
665 insects. *Mol. Biol. Evol.* 1–14 (2014). doi:10.1093/molbev/msu330
666 54. Henderson, I. R. & Jacobsen, S. E. Epigenetic inheritance in plants. *Nature* **447**, 418–424
667 (2007).
668 55. Lister, R. *et al.* Highly Integrated Single-Base Resolution Maps of the Epigenome in
669 Arabidopsis. *Cell* **133**, 523–536 (2008).
670 56. Jang, H. S., Shin, W. J., Lee, J. E. & Do, J. T. CpG and non-CpG methylation in epigenetic
671 gene regulation and brain function. *Genes (Basel).* **8**, 2–20 (2017).
672 57. Lyko, F. The DNA methyltransferase family: A versatile toolkit for epigenetic regulation. *Nat.*
673 *Rev. Genet.* **19**, 81–92 (2018).
674 58. Bernatavichute, Y. V., Zhang, X., Cokus, S., Pellegrini, M. & Jacobsen, S. E. Genome-wide
675 association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis
676 thaliana. *PLoS One* **3**, (2008).
677 59. Cingolani, P. *et al.* Intronic Non-CG DNA hydroxymethylation and alternative mRNA splicing
678 in honey bees. (2013).
679 60. Bonasio, R. *et al.* Genome-wide and caste-specific DNA methylomes of the ants camponotus
680 floridanus and harpegnathos saltator. *Curr. Biol.* **22**, 1755–1764 (2012).
681 61. Timmons, J. A., Szkop, K. J. & Gallagher, I. J. Multiple sources of bias confound functional
682 enrichment analysis of global -omics data. *Genome Biol.* **16**, 15–17 (2015).
683 62. Gaudet, P. & Dessimoz, C. Gene Ontology: Pitfalls, Biases and Remedies. in *The Gene*
684 *Ontology Handbook* (eds. Dessimoz, C. & Škunca, N.) **1446**, 189–205 (Springer New York,
685 2017).
686 63. Woodard, S. H., Bloch, G. M., Band, M. R. & Robinson, G. E. Molecular heterochrony and the
687 evolution of sociality in bumblebees (*Bombus terrestris*). *Proc. R. Soc. B Biol. Sci.* **281**, (2014).
688 64. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium
689 thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).
690 65. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylC-seq library
691 preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10**, 475–483
692 (2015).
693 66. Andrews, S. FastQC: A quality control tool for high throughput sequence data. (2010).
694 67. FASTX-Toolkit. (2009).
695 68. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused
696 by random hexamer priming. *Nucleic Acids Res.* **38**, 1–7 (2010).
697 69. Zhbannikov, I. SeqyClean. (2013).
698 70. Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B. & Brom, T. H. A reference-free algorithm
699 for computational normalization of shotgun sequencing data. *Genome Announc.* **2**, 1–18
700 (2012).
701 71. Sadd, B., Barribeau, S. & Bloch, G. The genomes of two key bumblebee species with primitive
702 eusocial organization. *Genome Biol.* **16**, 1–32 (2015).
703 72. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory
704 requirements. *Nat. Methods* **12**, 357–360 (2015).
705 73. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq
706 reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
707 74. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome
708 from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2013).
709 75. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: A web server for clustering and
710 comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
711 76. Davidson, N. M. & Oshlack, A. Corset: enabling differential gene expression analysis for de
712 novo assembled transcriptomes. *Genome Biol.* **15**, 1–14 (2014).
713 77. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: Reference-free
714 quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
715 78. Kapheim, K. M. *et al.* Genomic signatures of evolutionary transitions from solitary to group
716 living. *Science (80-.).* **348**, 1139–1143 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

717 79. Musacchia, F., Basu, S., Petrosino, G., Salvemini, M. & Sanges, R. Annocript: A flexible
718 pipeline for the annotation of transcriptomes able to identify putative long noncoding RNAs.
719 *Bioinformatics* **31**, 2199–2201 (2015).
720 80. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a
721 comprehensive and scalable alternative for improving sequence similarity searches.
722 *Bioinformatics* **31**, 926–932 (2015).
723 81. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for
724 genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
725 82. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO:
726 Assessing genome assembly and annotation completeness with single-copy orthologs.
727 *Bioinformatics* **31**, 3210–3212 (2015).
728 83. García-Alcalde, F. *et al.* Qualimap: evaluating next-generation sequencing alignment data.
729 *Bioinformatics* **28**, 2678–2679 (2012).
730 84. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,
731 357–359 (2012).
732 85. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or
733 without a reference genome. *BMC Bioinformatics* **12**, 1–16 (2011).
734 86. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for
735 RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–34 (2014).
736 87. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO Summarizes and Visualizes Long
737 Lists of Gene Ontology Terms. *PLoS One* **6**, e21800 (2011).
738 88. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular
739 interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
740 89. R Core Team. R: A language and environment for statistical computing. (2017).
741 90. Krueger, F. Trim Galore. (2012).
742 91. Guo, W. *et al.* BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. *BMC*
743 *Genomics* **14**, 1–8 (2013).
744 92. Guo, W. & Zhu, P. CG-maptools: Command-line Toolset for Bisulfite Sequencing Data
745 Analysis. (2017).
746
747

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020461; this version posted April 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

748 **Acknowledgements**

749 Authors would like to especially thank Dr. Yannick Wurm from Queen Mary
750 University of London for his suggestions on this study and manuscript, they contributed
751 immensely for this work. We also thank Dr. Isabel Alves-dos-Santos and Dr. Sheina
752 Koffler from the Laboratório de Abelhas (University of Sao Paulo) for the support
753 during *T. angustula* sampling, and Dr. Lars Chittka and Dr. Stephan Wolf from Bee
754 Sensory and Behavioural Ecology Lab (Queen Mary University of London) for their
755 support during *B. terrestris* sampling. Additionally, we would like to thank Dr. Bob
756 Schmitz from the Schmitz lab (University of Georgia) for the support with DNA
757 methylation sequencing, and to Susy Coelho from the University of Sao Paulo for
758 technical assistance.
759

760 **Additional Information**

761 This study was financed by FAPESP (São Paulo Research Foundation,
762 processes numbers 2013/12530-4, 2012/18531-0 and 2014/04943-0), by CNPq
763 (Conselho Nacional de Desenvolvimento Científico e Tecnológico, sponsorship to
764 MCA), by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil
765 (CAPES) [Finance Code 001], and developed at the Research Centre on Biodiversity
766 and Computing (BioComp) of the University of Sao Paulo, supported by the university
767 Provost's Office for Research. Part of the bioinformatic analyses was performed at the
768 cloud computing services of the University of Sao Paulo and of the Queen Mary
769 University of London.

770 Authors declare they have no competing financial interests.