Tests of hybridisation in *Tetragonula* stingless bees using multiple genetic markers.

Authors: James P. Hereward[1], Tobias J. Smith[1], Ros Gloag[2], Dean R. Brookes[1] & Gimme H. Walter[1].

Affiliations:

[1]School of Biological Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia.

[2]School of Life and Environmental Sciences, The University of Sydney, Sydney, New South Wales 2006, Australia.

Corresponding author: James Hereward, School of Biological Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia

E-mail: j.hereward@uq.edu.au

Keywords: Hybridisation, gene flow, species, genetic marker, stingless bee

Abstract

Discrepancies in mitochondrial and nuclear genetic data are often interpreted as evidence of hybridisation. We re-examined reports of hybridisation in three cryptic stingless bee species in the genus *Tetragonula* in South East Queensland, Australia (*T. carbonaria*, *T. davenporti*, and *T. hockingsi*). Previous studies on this group using microsatellite markers proposed that occasional hybrids are found. In contrast, we find that allele frequencies at neutral regions of the nuclear genome, both microsatellites and random *snps*, reliably separated the three species, and thus do not support hybridisation. We found no inter-species variation in PCR amplicons of the nuclear gene *EF1alpha*, but low and moderate species-specific polymorphisms in the nuclear gene *Opsin* and the mitochondrial *16S* respectively, with no cases of mito-nuclear discordance at these genes. We confirm that nuclear divergence between these species is low, based on 10-26kb of non-coding sequence flanking *EF1alpha* and *Opsin* (0.7-1% pairwise difference between species). However, we find mitogenomes to be far more diverged than nuclear genomes (21.6-23.6% pairwise difference between species). Based on these comprehensive analyses of multiple marker types, we conclude that there is no ongoing gene flow in the *Tetragonula* species of South East Queensland, despite their high morphological similarity to one another and the low nuclear divergence among them. The mitogenomes and draft nuclear genomes provided for these species will be a resource for further molecular studies on this group, which are important pollinators in Australian natural and agroecosystems.

Introduction

Allopatry is key to speciation, but species boundaries may be tested when populations that diverged in allopatry come back into secondary contact through changes in habitat or climate, or as a result of anthropogenic movement over geographic boundaries (Sánchez-Guillén *et al*., 2013, Gloag *et al*. 2017; Brookes *et al.* 2020). The process of hybridisation is closely tied to the designation of species in population genetic terms (Paterson 1978). When species are considered conceptually as recombining gene pools (Ayala *et al.* 1994), hybridisation is the process of gene flow across two such species gene pools. High levels of hybridisation in local areas of overlap indicate that the species in question would be better viewed as subspecies, which, by definition, have independent geographical distributions but hybridise freely where and when they do overlap (Walter 2003, Ford 1974). Such situations are often seen when closely related lineages of the same species are brought back into sympatry (Feder *et al*. 2003), but the pattern is also sometimes seen in more distantly related lineages (e.g. Shipman *et al.* 2019). The frequency of hybridisation, or gene flow, between gene pools is therefore key to understanding the limits of a given species, and to our understanding of speciation.

Accurate assessment of the frequency of hybridisation under natural conditions is therefore the first step in shedding light on the process of speciation. The more recently two populations have diverged, the more likely they have not yet become separate species. When two subspecies come back into secondary contact, gene flow is generally re-established (Feder *et al*. 2003). Hybrid detection between close relatives is especially sensitive to the molecular methodologies used (e.g. number and type of markers), and more likely to be susceptible to sampling error because parent species are similar in appearance with respect to one another.

An Australian species complex of stingless bees (Meliponini) in the genus *Tetragonula* illustrates these challenges well. The genus *Tetragonula* includes three cryptic sister species (*T. carbonaria. T. hockingsi* and *T. davenporti*), indigenous to eastern Australia, that are part of a monophyletic clade known as the "Carbonaria group" (Franck *et al*., 2004). These bees can be kept readily in wooden hives, are popular pets, and are effective pollinators of some tropical and subtropical crops (Dyer *et al.* 2016, Heard 1994). This has led to many thousands of colonies being kept by beekeepers in Eastern Australia. Today, the distributions of all three species overlap in parts of south-east Queensland (Fig. 1). Beekeeping and other human activities, such as land use change, have greatly increased the overlap of their distributions in recent decades, particularly through the southward spread of *T. hockingsi*. Unfortunately, the historical distributions of the species are poorly known due to their cryptic morphology.

Queens, males and workers (females) cannot be reliably identified based on morphology alone (Dollin *et al.* 1997). Rather, species are typically identified by the structure of their brood comb, with *T. carbonaria* producing a neat spiral of adjacent brood cells in a layered comb and *T. hockingsi* producing clusters of cells in a form known as semi-comb (Fig. 1). *Tetragonula davenporti* shares the semi-comb brood structure of *T. hockingsi*, and is known from only a handful of managed colonies in a sub-coastal pocket of southern Queensland (Franck *et al*. 2004, Brito *et al.* 2014), but is believed to persist in the wild in nearby forested areas (P. Davenport pers. Comm 2019; R. Gloag unpublished data).

Relatively little is known about the mating behaviour of these bees, because it occurs during the nuptial flight. Genetic data indicates that queen's mate only once (Green & Oldroyd 2002), and observations of tethered virgins indicate that male genital detachment is likely the mechanism by which multiple mating is avoided (Smith 2020). Male mating congregations, which form outside colonies with virgin queens preparing for nuptial flights, can contain a mix of both *T. carbonaria* and *T. hockingsi* males (Cunningham *et al.* 2014). These two species are

also known to engage in interspecific nest usurpations, in which an invading colony kills a resident queen and installs its own in the nest (Cunningham *et al.* 2014), a behaviour which may increase the chances that males are attracted to interspecific queens.

Previous efforts to establish if, and how regularly, gene flow occurs between these species, have returned equivocal results. First, microsatellite markers have been used to look for evidence of hybrids. Frank *et al.* (2004) identified some South-East Queensland colonies of *Tetragonula* as being of hybrid origin based on microsatellite data, where hybrids were defined as samples having a posterior probability of assignment of less than 90% to a cluster in an analysis using the genetic assignment program *STRUCTURE* (Pritchard *et al.* 2000). The same three Australian *Tetragonula* species were re-investigated by Brito *et al.* (2014), with microsatellite data and two mitochondrial genes. Three instances of hybridisation in South-East Queensland were reported within the zone in which *T. carbonaria* and *T. hockingsi* overlap. This was based on a threshold of 95% posterior probability of belonging to a cluster in a Bayesian clustering analysis of the microsatellite data. Brito *et al*. (2014) also included two hives reported to be *T. davenporti* in their study. One of the *T. davenporti* hives clustered with *T. carbonaria* in a Principal Coordinate Analysis (PCoA) of genetic distance based on the microsatellite data, and the other clustered slightly outside of the '*T. carbonaria*' cluster. They suggested that *T. hockingsi* represents two genetic clusters, one each in northern and southern Queensland, based on their microsatellite and mitochondrial data. Their reported *T. hockingsi x T. carbonaria* hybrids were from South East Queensland and they suggested that this hybridisation was due to the anthropogenic movement of northern *T. hockingsi* into the southern region.

Second, both nuclear and mitochondrial markers have been used to look for evidence of introgression and thus past interbreeding across Australia's East Coast *Tetragonula* species. These efforts were hampered by the amplification of nuclear copies of mitochondrial genes

117 (numts) (e.g. for *cytochrome b* (*cytb*); Franck *et al.* 2004). Recently, several numts of
118 Cytochrome Oxidase Subunit I (*COI*) were identified in *T. hockingsi* and *T. carbonaria*
119 (Françoso *et al*. 2019), which cast into question even the designation of *T. davenporti* as a true
120 species. Surprisingly, the genuine COI sequences of *T. hockingsi* and *T. carbonaria* are highly
121 diverged (16.5 %), relative to available nuclear genes (Cunningham 2014, Francosco 2019).
122 This pattern of high mitochondrial divergence but low nuclear divergence might be consistent
123 with recent or ongoing gene flow and highlights the need for new analyses to understand the
124 species status of these bees. This will in turn reveal the likely impact of continued movement
125 of managed *Tetragonula* hives around Australia.

126 To resolve whether ongoing gene flow occurs between *T. carbonaria, T. hockingsi* and
127 *T. davenporti* we sampled these species in sympatry in South East Queensland. We compared
128 microsatellite data, using the same markers as previous studies, to a *genotyping by sequencing*
129 approach that yielded 2,049 *snp* markers with less than 5% missing data. We also look for
130 evidence of mitochondrial introgression, by considering two nuclear gene sequences and one
131 mitochondrial gene sequence of each species. We then use Illumina sequencing to reconstruct
132 mitochondrial genomes and confirm that our mitochondrial sequences are not pseudogenes.
133 Finally, we assembled draft nuclear genomes for these three species as a resource for further
134 genetic analysis of the species complex. This comprehensive approach leads to new insights
135 about the extent of hybridisation in these bees, and their species status.

136

137 Materials and Methods

138 *Sampling and DNA extraction*

139 We sampled four workers per colony from hived colonies of each species between
140 February 2017 and April 2019 in south-east Queensland (*T. carbonaria*: N=33, *T. hockingsi*,

141 N=21, *T. davenporti*, N=3) and we also included two *Tetragonula hockingsi* hives from
142 Rockhampton (Queensland), two hives of *Tetragonula carbonaria* from Newcastle (NSW),
143 and two from Sydney (NSW) (Fig. 1). Most *T. carbonaria* and *T. hockingsi* colonies were
144 sampled in the greater Brisbane region. We opened each hive and assigned it to species based
145 on brood nest architecture (spiral/layered comb or semi-comb). The three *T. davenporti*
146 colonies were sampled from the border area of Queensland and New South Wales and were
147 the same colonies used for past genetic work on this species (P. Davenport, pers. comm. 2019).
148 A further two *T. carbonaria* colonies and one *T. hockingsi* colony were sampled from the same
149 meliponary as the three *T. davenporti* colonies. We extracted DNA using a silica spin-column
150 method (Ridley *et al.* 2016) and normalized concentrations to 5ng/ul per sample following
151 quantification (PicoGreen, Life Technologies, California, USA).

152

153 *Neutral nuclear markers*

154 *Microsatellites*: We first assessed species status and evidence of hybridisation using seven
155 microsatellite loci (Green *et al.* 2001), as used in previous molecular studies of this group
156 (Franck *et al.* 2004, Brito *et al.* 2014). We used PCR reactions of 2 µl DNA template, 0.5 U
157 MyTaq polymerase (Bioline, Australia), 0.1 µM of forward primer, 0.2 µM of reverse primer,
158 0.2 µM M13 labelled primer with different fluorescent dyes (6-FAM, VIC, PET or NED) and
159 1x buffer, with reaction conditions of an initial denaturation at 95 °C for 3 min, followed by 10
160 cycles of 10 s at 95 °C, annealing at 45 °C for 25 s, and 1 min extension at 72 °C, then 40
161 cycles of 10 s at 95 °C, annealing at 55 or 52°C for 25 s, and 1 min extension at 72 °C, and the
162 final extension was at 72 °C for 2 min (see Cunningham *et al.* (2014) for locus-specific
163 annealing temperatures). Loci were then pooled with one locus per dye, cleaned with 1U
164 Exonuclease I and one unit of Antarctic Phosphatase, as above, and genotyped by Macrogen

165 Inc. (Seoul, Republic of Korea). Microsatellite peaks were analysed using the microsatellite
166 plugin in Geneious v. 2019.2.1 (http://www.geneious.com, Kearse *et al*. (2012)).

167 *Single nucleotide polymorphisms:* We generated genome-wide *snp* data using a genotyping-
168 by-sequencing method. We adapted a protocol and adaptor regime based on the methods of
169 Elshire *et al.* (2011), Poland *et al.* (2012) and Petersen *et al*. (2012), with barcodes based on
170 Caporaso *et al.* (2012). Full details of the method are provided in supplementary data and can
171 be found online at http://www.jameshereward.org/GBS.html. We pooled 288 individuals per
172 sequencing lane, and sequenced the libraries with PE150 Illumina sequencing at Novogene
173 (Beijing, China).

174 The sequence data were demultiplexed, assembled, and *snps* called using STACKS
175 (Catchen *et al.* 2013). We then filtered the vcf using vcftools, with the data first being filtered
176 to a minor allele count of three (one heterozygote and one homozygote), which allows the
177 conservative removal of singleton *snps* that are likely to be errors, without discarding rare
178 alleles (Linck & Battey 2019). We set a minimum depth of 5 and kept only biallelic *snps*. We
179 filtered the data for missing data in three steps. First, any marker missing more than 50% data
180 was discarded (i.e. 50% of individuals were not genotyped at that marker), to remove the
181 markers most affected by missing data. Second, any individual that had missing data at more
182 than 50% of the markers was discarded, to remove the individuals that had bad quality
183 genotyping. Finally, any marker missing more than 5% data was discarded to produce a final
184 dataset with relatively little missing data (~3%).

185 For both our microsatellite and *snp* datasets, we performed principal component
186 analyses (PCA) using the adegenet package in R (Jombart, 2008) and *STRUCTURE* analyses
187 (Pritchard 2000). For the PCA we used the full *snp* dataset (2,049 *snps*), but for *STRUCTURE*
188 only one *snp* per locus was retained (1,935 *snps*). *STRUCTURE* is an individual-based

189 clustering algorithm that assigns individuals to each of *K* population clusters using Hardy-
190 Weinberg equilibrium and linkage information in genetic markers. To assess the effect of
191 replication and missing data on hybridisation inferences from the *STRUCTURE* output, we
192 repeated the analyses on three versions of our microsatellite and *snp* datasets : (i) a dataset
193 including one worker per colony, (ii) a dataset including all four workers per colony, and (iii)
194 a dataset with 20% randomly deleted. To achieve 20% missing data in the the microsatellite
195 dataset, data was removed evenly across markers until a 20% reduction was achieved. For the
196 *snp* dataset, we changed the missing data threshold in the filtering step to 20%, and this
197 increased the number of markers to 13,409 loci. Having more markers than the original dataset
198 clearly makes for an uneven comparison, so we simply deleted every marker after the first
199 1,935 *snp* markers to make this dataset the same size as the original one with no missing data.
200 For each *STRUCTURE* run, we used the admixture model and ran 500,000 iterations of burn-
201 in followed by 1 million iterations. We assumed both two or three populations (K=2 and K=3)
202 and performed multiple runs that were then permuted and plotted using Clumpak server
203 (Kopelman *et al.* 2015).

204

205 *Gene sequencing*

206 To confirm whether common marker genes showed species-specific polymorphisms,
207 we sequenced two nuclear genes (*EF1alpha* and *Opsin*) and one mitochondrial gene (*16S*).
208 These genes were selected as they have previously been used for a phylogenetic study of
209 stingless bees (Rasmussen and Cameron, 2009). For *EF1alpha* we used primers F2-ForH
210 (GGRCAYAGAGATTTCATCAAGAAC) and F2-RevH2
211 (TTGCAAAGCTTCRKGATGCATTT) (Hines 2006). For *Opsin* we used LWRhF
212 (AATTGCTATTAYGARACNTGGGT) and LWRhR

213 (ATATGGAGTCCANGCCATRAACCA) (Mardulyn & Cameron 1999). For *16S* we used
214 primers 874-16S1R (TATAGATAGAAACCAAYCTG) (Cameron 1992) and 16SWb
215 (CACCTGTTTATCAAAAACAT) (Dowton & Austin 1992). We performed PCRs in 25µl
216 reactions containing 2 µl DNA template, 1U MyTaq polymerase (Bioline, Australia), 0.2 µM
217 of each PCR primer, and 1x buffer. For amplification of *EF1alpha* and *16S* we performed PCR
218 with 35 cycles at 52ºC annealing, and for *Opsin* we used five cycles at 45ºC annealing followed
219 by 35 cycles at 50ºC. PCR products were cleaned using 1U of Exonuclease I and Antarctic
220 Phosphatase (New England Biolabs, Ipswich, Mass., USA) and incubating at 37 °C for 20 min
221 followed by 10 min enzyme denaturation at 80 °C. Sequencing was conducted using the same
222 forward and reverse primers used for PCR, by Macrogen Inc. (Seoul, Republic of Korea). We
223 edited sequence data with CodonCode aligner, and aligned them using MAFFT in Geneious
224 Prime 2019.2.1 (www.geneious.com, Kearse *et al*. 2012). We generated haplotype networks
225 using the TCS algorithm (Clement *et al.* 2000) in PopART version 1.7
226 (http://popart.otago.ac.nz/) (Leigh and Bryant 2015).

227

228 *Mitogenome sequencing and assembly*

229 The "Carbonaria group" of *Tetragonula* are now known to have several pseudogenes
230 (or "numts", nuclear insertions of mitochondrial genes) (Françoso *et al.* 2019; see also Franck
231 *et al.* 20014, Brito *et al.* 2014). To ensure we were amplifying only the true *16S* gene and to
232 provide a reference genome for each species for future work, we conducted whole genome
233 sequencing on one individual of each of the three species and reconstructed the complete
234 mitochondrial sequence. Sequencing libraries were made using the NebNext UltraII DNA kit
235 for Illumina (New England Biolabs, Ipswich, Mass., USA), and PE150 sequencing conducted
236 by Novogene (Beijing, China).

237 Sequences were checked with FastQC (Andrews, 2019), then we removed adaptors,
238 and quality-trimmed the end of the sequences of all samples to q10 using bbduk from the
239 bbtools package (version 36, Bushnell, 2019), with a kmer of 8. We binned the sequences by
240 depth using a kmer-based approach in bbnorm (Bushnell, 2019). We put everything under 500x
241 coverage into one file (this includes the nuclear genome, which had between 23x and 51x
242 coverage for each species, Table 1), everything between 500x and 3,000x into another file (this
243 is repetitive sequence and other artefacts below the depth of the mitochondria), everything over
244 3,000x coverage was placed in a third file, and this contained mitochondrial sequence and
245 anything else with a depth over 3,000. This process excluded nuclear-mitochondrial
246 pseudogenes and also reduced the amount of data going into subsequent steps. We then
247 reconstructed mitogenomes by mapping reads to the mitogenome of the stingless bee *Melipona*
248 *bicolor* (AF466146.2, Araujo & Arias 2019), followed by *de novo* assembly of the mapped
249 reads, mapping and extension to the longest contigs, and then joining of contigs and manual
250 verification of the final sequence by read-mapping (Geneious version 2019.2., Kearse *et al*.
251 2012). The final mitogenome assemblies were annotated using DOGMA (Wyman *et al*. 2004)
252 and MITOS (Bernt *et al.* 2013), with reference to the *M. bicolor* mitogenome and manual
253 editing in Geneious.

254 We aligned these three new bee mitogenomes with three more unpublished *Tetragonula*
255 mitogenomes and the *M. bicolor* mitogenome (AF466146.2) using MAFFT (Katoh and
256 Standley 2013). The most appropriate evolutionary model for each gene was estimated using
257 jModelTest 2.1.10 with the best model chosen based on Bayesian Information Criterion (BIC).
258 We then constructed a bayesian tree with MrBayes (Huelsenbeck & Ronquist 2001).

259

260 *Nuclear genome assembly*

261 To investigate further the low nuclear divergence suggested by our two nuclear genes
262 (*Opsin* and *EF1alpha*), we considered the remaining genome-wide data generated from our
263 Illumina sequencing. We took the lower genomic bin of the Illumina data (everything below
264 500x coverage based on kmer binning; Bushnell 2019), and assembled each species using
265 SPADES (Bankevich *et al.* 2012), which was run with default settings, including the in-built
266 error correction module. We assessed the assemblies and generated basic statistics using
267 QUAST (Gurevich *et al.* 2013). We then took the 768bp *EF1alpha* PCR product and searched
268 it against custom-blast databases for each genome. We did the same for *Opsin*. These larger
269 sequences were aligned with MAFFT and a Bayesian tree was inferred from this alignment
270 with MrBayes.

271

272 Results

273 *Tests of hybridisation using microsatellites and snps*

274 PCAs using both microsatellite and *snp* datasets separated samples into three non-
275 overlapping clusters consistent with their species assignment (<5% missing data, Fig. 2A). The
276 microsatellite PCA placed *T. davenporti* closer to *T. carbonaria* than to *T. hockingsi* (Fig. 2A),
277 but the *snp* PCA indicated an equal distance between the three clusters (20% of the variance
278 explained by each of the first two components, Fig. 2B). The loading plot of the *snp* data
279 indicated that many of the 2,043 *snp* markers were contributing to the clustering in the PCA,
280 rather than a few specific loci (Fig. 2C).

281 *STRUCTURE* analyses using both microsatellite and *snp* datasets (N= 63 colonies, one
282 worker per colony) returned no evidence of hybrids, with all individuals showing close to 100%
283 assignment to a species' cluster (Fig. 3). Notably, *STRUCTURE* results were sensitive to

284 sibling-sampling and missing data. The same analyses using datasets comprising four workers
285 per colony (akin to resampling the same individual multiple times), assigned four individuals
286 (microsatellites) and two individuals (*snps*) at less than 90% posterior probability of belonging
287 to their species' cluster, thus identifying them as putative hybrids (Fig. 3). Likewise,
288 subsampling to 20% missing data resulted in two *T. hockingsi* individuals being designated as
289 hybrids with *T. davenporti*, but only with the microsatellite data. Missing data had no impact
290 on the *snp* dataset, i.e. every individual was still assigned 100%.

291

292 *Species-specificity of PCR-amplified gene regions*

293 The mitochondrial gene *16S*, and the nuclear gene *Opsin*, each grouped samples into
294 three haplotype clusters consistent with species assignment (Fig. 4), although pairwise
295 divergence between species was greater for *16S*, with ~6% between *T. carbonaria* and *T.*
296 *hockingsi* based on 229 individuals representing all 63 colonies than for *Opsin*, with ~0.5%
297 difference between *T. carbonaria* and *T. hockingsi* (116 individuals representing all 63
298 colonies). The interspecific sequence difference in the *16S* fragment was considerably lower
299 than that of the complete mitogenome (see below), because half of the amplified region was
300 highly conserved (no variation). The nuclear gene *EF1alpha* had identical sequence for all
301 samples (245 individuals representing all 63 colonies). No individuals showed mitochondrial-
302 nuclear discordance between *Opsin* and *16S*. That is, if an individual had a *16S* sequence
303 representative of *T. carbonaria* it also had an *Opsin* sequence representative of *T. carbonaria*
304 (Fig. 4). However, one colony that was labelled *T. carbonaria* during field collections returned
305 sequences and nuclear markers consistent with *T. hockingsi* and was assumed to be the result
306 of either labelling error, or a recent interspecific colony takeover. Further, one colony suspected

307 of being *T. davenporti* was genetically assigned to the *T. carbonaria* cluster, highlighting the
308 difficulty of identifying them without genetics.

309

310 *Mitogenomes and genomes*

311 We were unable to fully resolve the AT-rich control region of the mitochondria, but we
312 did recover all mitochondrial genes. Our assemblies of the mitogenomes for these three species
313 indicated 22.2% pairwise difference across *T. carbonaria* and *T. hockingsi*, 21.6% across *T.*
314 *carbonaria* and *T. davenporti,* and 23.6% across *T. davenporti* and *T. hockingsi* (Fig. 5). Using
315 these mitogenomes we were able to determine that the *16S* mitochondrial sequences were of
316 mitochondrial origin and not nuclear pseudogenes. An obvious gene inversion was detected,
317 relative to *M. bicolor* and all other bees for which mitogenome sequences are available
318 (Supplementary Fig. 1) with *12S*, *16S*, *NAD1*, *CYTB* and *NAD6* reversed and inserted in
319 between *NAD3* and *NAD5*.

320 Nuclear genomes returned a mean *n50* of 14,954 across the three species (Table 1).
321 From this sequencing, we investigated further the lack of interspecific difference at *EF1alpha*.
322 We took the *EF1alpha* sequence and used it as a 'bait' to pull out the regions of the genome
323 surrounding the 768bp PCR product. We managed to select at least 10kb in each species that
324 aligned around the PCR product using MAFFT (Katoh and Standley 2013). Duplications were
325 not present as only one hit was found in each assembly. Variation was detected around the
326 *EF1alpha* locus and this separated the three species, based on the one individual of each that
327 we sequenced. We thus confirmed that the pairwise difference across these three bee species
328 was indeed extremely low in this region of the nuclear genome (0.7-1%, Fig. 5), relative to that
329 detected in the mitochondrial genome. Similarly, we recovered 26kb around the *Opsin* gene,
330 which returned the same low nucleotide difference between species (0.7-1%).

331 A Bayesian phylogenetic analysis based on either mitogenomes, or the flanking regions
332 of *EF1alpha*, resolved the evolutionary relationships between the *Tetragonula* species in the
333 same way, with *T. carbonaria* and *T. davenporti* diverging from each most recently, and *T.*
334 *hockingsi* as their sister group (Fig 5).

335

336 Discussion

337 *Testing for hybridisation*

338 Hybridisation and gene flow are central to our understanding of species and speciation,
339 but accurately detecting hybridisation depends on the markers and analyses used. In the case
340 of Australia's *Tetragonula* stingless bees, two previous studies concluded a low incidence of
341 hybrid colonies in an area of sympatry between sister species (2-3% of all sampled colonies;
342 Franck *et al.* 2004, Brito *et al.* 2014). In this study, we use multiple additional molecular
343 markers to analyse a new sample of colonies from the same area of sympatry (South-East
344 Queensland) and find all markers reliably segregate the three species, *T. carbonaria*, *T.*
345 *hockingsi* and *T. davenporti*. We cannot discount rare occurrences of hybrid colonies based on
346 63 hive samples, but we can conclude that there is no gene flow between the species in this
347 region beyond the F1 generation from our data.

348 The lack of gene flow between *Tetragonula* species indicates that these species either
349 do not mate, or that interspecific mating produces inviable or infertile offspring. Beekeepers
350 regularly keep multiple species in close proximity, providing opportunities for interspecific
351 mating; for example, at one property that we sampled, colonies of all three species had resided
352 together for at least nine years. In stingless bees, mating occurs on the wing close to the queen's
353 colony. Males typically aggregate out the front of a colony that is undergoing the re-queening
354 process (i.e. has a virgin queen), presumably attracted by a pheromone (Vollet-Neto *et al.*

355 2018). Male aggregations may attract dozens to hundreds of unrelated males from the
356 surrounding area (Vollet-Neto *et al*. 2018; Cameron *et al.* 2004). Virgin queens then fly into
357 the aggregation and mate with just one male, before returning to the nest (Green & Oldroyd
358 2002). Even if aggregations sometimes contain males of more than one *Tetragonula* species, it
359 may be that males are only sufficiently attracted to conspecific queens to attempt actual mating.
360 Alternatively, queens may be crucial to the recognition process at the point a male makes
361 contact with them, as in *Aphytis lingnanensis* parasitoids (Fernando and Walter 1997). Further,
362 interspecific mating may occur, but produce inviable diploid brood due to genetic
363 incompatibility. This latter scenario has been reported in sister species of honeybee that have
364 come into sympatry recently (*Apis cerana* and *Apis mellifera*). Interspecific mating between
365 these honeybees is common, but all hybrid offspring die at the early larval stage (Remnant *et*
366 *al*. 2014; Gloag *et al*. 2017). In this case, mis-mated queens may be quickly replaced. For
367 example, workers of at least one stingless bee species (*Scaptotrigona depilis*) will execute
368 laying queens that have made inbred matings and are producing diploid male brood (Vollet-
369 Neto *et al*. 2017). Clearly, further investigation into the mating of *Tetragonula* species is required
370 to clarify the interactions between males and females in mixed species swarms.

371 Australian *Tetragonula* highlight two common reasons why it can be difficult to
372 identify hybridisation between sister taxa. First, *Tetragonula* have nuclear pseudogenes of
373 several common mitochondrial marker genes (Françoso *et al.* 2019; Franck *et al*. 2004). Such
374 pseudogenes can complicate efforts to assess hybridisation through mito-nuclear discordance,
375 because they may be mistaken for valid haplotypes that are shared across species and thus
376 mistakenly taken as evidence of introgression. Even where researchers aim to avoid
377 pseudogenes, by checking for stop codons or frame-shifting indels in their sequences, a subset
378 of pseudogenes may still be missed. The reconstruction of whole mitogenomes from short read
379 data, using a depth cut-off to avoid nuclear copies, as in this study, may be necessary to identify

380 true mitochondrial genes. A further issue is that the pattern expected from mito-nuclear
381 discordance (a sample returning one mitochondrial sequence and a different nuclear one) is the
382 same as one gets when samples are mislabelled. Rates of mislabelling have been estimated at
383 around 4% in published genetic datasets (Zych *et al.* 2017), so any cases of mito-nuclear
384 discordance at these rates probably deserve additional scrutiny. We detected one instance of
385 mislabelling across two of our datasets, by virtue of having conducted multiple genetic tests on
386 the same individuals. Both pseudogenes and "mis-labelling introgression" have likely led to
387 mistaken assignments of hybridisation in many taxa, and inflated perceptions of the extent to
388 which species hybridise.

389 Second, where nuclear markers suggest low divergence between species, the
390 interpretation of *STRUCTURE* analyses must take care to avoid biasing group assignment or
391 over-interpreting structure plots (Lawson *et al.* 2018). In our dataset, the inclusion of four
392 individuals per colony (full sisters) in a *STURCTURE* analysis led to the assignment of some
393 individuals as putative hybrids. When this pseudo replication was removed, however, by
394 including just one individual per colony, no hybrids were evident. Likewise, missing data had
395 a large effect on the likelihood of inferring *Tetragonula* hybridisation in our dataset. Our
396 original microsatellite dataset had low levels of missing data (<5%) and every individual was
397 assigned at close to 100% into a species group. When we subsampled this dataset to 20%
398 missing data, however, some samples appeared to be hybrids, indicating that *STRUCTURE*
399 analyses can be sensitive to such levels of missing data. Notably, in this test on the effect of
400 missing data, we subsampled our microsatellite data randomly across markers. In real datasets,
401 microsatellites tend to go missing in a much less random way, generally with one or two loci
402 most affected. The impact of missing microsatellite data is therefore likely to be even greater
403 in practice than in our simulation. *STRUCTURE* assignments using microsatellites were the
404 basis of hybrid identification in previous studies of Australian *Tetragonula* and may have been

405 sensitive to these issues (Franck *et al.* 2004; Brito *et al.* 2014). In our study, *snp* data was much
406 less sensitive to the effect of missing data and vastly outperformed microsatellite data to test
407 for hybridisation. Indeed, in taxa where nuclear genome divergence is low, using many markers
408 may be the only way to confidently assess hybridisation rates. In the absence of direct parentage
409 analysis to assess hybridisation, such a multi-marker nuclear genome assay is the best approach
410 to infer whether gene flow between species is ongoing (Twyford & Ennos 2011).

411

412 *Mitochondrial divergence between Tetragonula species*

413 Our mitogenome data support previous studies indicating a high mitochondrial
414 sequence divergence between species of the 'Carbonaria complex' (Cunningham *et al.* 2012;
415 Francosco *et al.* 2019). This mitochondrial divergence contrasts with the low divergence we
416 observed in two nuclear genes and their flanking regions, around 0.7-1% bp difference *vs.* 21-
417 23% in the mitogenomes. We investigated the lack of variation in the *EF1alpha* PCR product
418 further to see whether it was restricted to this 700bp fragment. We were able to recover the
419 sequence flanking this region from our whole genome assemblies. These larger fragments did
420 exhibit variation and recovered the same phylogeny as the mitogenomes. The discrepancy in
421 divergence might indicate a particularly high mutation rate for mitochondria in this clade. The
422 ratio of mitochondrial to nuclear mutation rate varies substantially across metazoans (from at
423 least 0.8x to 40x, Allio *et al*. 2017). A mutation rate in *Tetragonula* mitochondria that was
424 around 20 times that of their nuclear genome would be within the range reported for insects,
425 albeit at the high end (Allio *et al*. 2017; Francosco *et al.* 2019). Several other Hymenoptera
426 have high rates of mitochondrial divergence, compared to other insects (Castro *et al.* 2002;
427 Kaltenpoth 2012). This pattern of mito-nuclear substitution rate difference could also be
428 interpreted as being evidence for a relatively deep divergence time for these species (evidenced

429 in the divergent mitochondria) that was later followed by a period of nuclear gene flow, erasing
430 nuclear divergence. Under this scenario, past gene flow between diverged populations must
431 have occurred in a similar fashion between all three species, because they share a common ratio
432 of mitochondrial to nuclear divergence. Distinguishing between these two alternatives requires
433 a better understanding of rate variation in mitochondrial and nuclear genomes of stingless bees
434 relative to other Hymenoptera, and population genomic studies to elucidate the evolutionary
435 history of this group.

436 *Tetragonula davenporti* remains the most mysterious of the three *Tetragonula* species
437 considered in this study. It has been confirmed from only one property in the Gold Coast
438 hinterland (originally sourced from forested areas nearby) and one wild specimen from the
439 Queensland-New South Wales border (R. Gloag, unpublished data). It is essentially cryptic in
440 morphology with *T. carbonaria* and in nest structure with *T. hockingsi*. Nevertheless, the
441 distinctive mitogenome of *T. davenporti* reported here lends strong support to its status as a
442 valid species. The natural range of *T. davenporti* is currently unknown, and the difficulty in
443 identifying it morphologically to species means that its presence among managed colonies is
444 also unknown. South-east Queensland is an area of high endemicity, so it may well be endemic
445 in this area.

446 This study investigated managed colonies mostly from Brisbane and SE Queensland
447 because this is where hybridisation is most likely between these three species and has been
448 reported previously. However, *T. carbonaria* and *T. hockingsi* also co-occur in parts of Far
449 North Queensland, and *T. hockingsi* is increasingly being transported by beekeepers into more
450 southern parts of *T. carbonari*a's range (pers. obs. TJS). Examining wild populations of this
451 genus may reveal additional species within the '*carbonaria* complex'. Comparing wild
452 populations to those that have been transported anthropogenically will provide further insight
453 into the impact of human movement on the evolution of this group. Finally, we hope the

genomic resources developed here will enable further investigation of this fascinating group of social bees.

Acknowledgements

We would like to acknowledge all the beekeepers and hive owners who donated bees to this study, especially Tim Heard, Peter Davenport and Alan Lashmar. We thank the University of Queensland for financial support.

Data Availability

The PCR amplicon data can be found on Genbank; acessions MN659448 - MN659676 for 16S, MN659677 - MN659921 for EF1alpha and MN659922 - MN660037 for Opsin.

Individuals used for whole genome sequencing have NCBI Biosample numbers as shown in Table 1. Illumina data has been deposited at the Sequence Read Archive (SRA), *T. carbonaria* = SRX7089578, *T. hockingsi* = SRX7104558, *T. davenporti* = SRX7104558. Genome assemblies can be found on Genbank, *T. carbonaria* = GCA_010645115.1, *T. hockingsi* = GCA_010645185.1, *T. davenporti* = GCA_010645165.1. These can all be found on the University of Queensland espace as well (Hereward et al. 2020, doi: https://doi.org/10.14264/84de1f8).

Microsatellite data, raw *snp* data, vcf files, R scripts, additional genome assembly data, and locations and dates of samples can be found on the University of Queensland espace (Hereward et al. 2020, doi: https://doi.org/10.14264/84de1f8) The locations are only accurate to around 1km to protect the privacy of beekeepers.

476

477

478 ## References

479 Allio R, Donega S, Galtier N, *et al*. 2017. Large variation in the ratio of mitochondrial to
480 nuclear mutation rate across animals: implications for genetic diversity and the use of
481 mitochondrial DNA as a molecular marker. Molecular biology and evolution 34(11): 2762–
482 2772.

483 Andrews S. 2019. FastQC: a quality control tool for high throughput sequence data [Online].
484 Available: http://www.bioinformatics.babraham.ac.uk?/projects/fastqc/.

485 Araujo N de S and Arias MC. 2019. Mitochondrial genome characterization of *Melipona*
486 *bicolor*: Insights from the control region and gene expression data. Gene 705: 55–59.

487 Ayala, FJ, Escalante, A, O'Huigino, C, & J Klein. 1994. Molecular genetics of speciation and
488 human origins. Proceedings of the National Academy of Science USA. 91: 6787-6794.

489 Bankevich, A, Nurk, S, Antipov, D, Gurevich, AA, Dvorkin, M, Kulikov, AS, Lesin, VM, *et*
490 *al.* 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell
491 sequencing. Journal of Computational Biology: A Journal of Computational Molecular Cell
492 Biology 19 (5): 455–77.

493 Barreto FS, Watson ET, Lima TG, *et al*. (2018) Genomic signatures of mitonuclear coevolution
494 across populations of *Tigriopus californicus*. Nature Ecology & Evolution 2(8): 1250–1257.

495 Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M,
496 Stadler PF. MITOS: improved de novo metazoan mitochondrial genome annotation. Molecular
497 Phylogenetics and Evolution. 69(2):313-9

498 Brito, RM, Francisco, FO, Ho, SYW, & BP Oldroyd. 2014. Genetic architecture of the
499 Tetragonula carbonaria species complex of Australian stingless bees (Hymenoptera: Apidae:
500 Meliponini). Biological Journal of the Linnean Society. 113(1) 149-161.

501 Brookes DR, Hereward JP, Wilson LJ, & GH Walter. 2020. Multiple invasions of a generalist
502 herbivore—Secondary contact between two divergent lineages of *Nezara viridula* Linnaeus in
503 Australia. Evolutionary applications 244: 171.

504 Bushnell B. 2019. Bbmap – a suite of bioinformatics tools. [Online]. Available:
505 https://sourceforge.net/projects/bbmap/

506 Cameron, SA, Austin, AD, James Derr, & RA Wharton. 1992. The application of nucleotide
507 sequence data to phylogeny of the Hymenoptera: a review. Journal of Hymenoptera Research.
508 1:63-79.

509 Caporaso, J, Gregory, CL, Lauber, WA, Walters, D, Berg-Lyons, J Huntley, N, Fierer, S & M
510 Owens, *et al.* 2012. Ultra-High-Throughput Microbial Community Analysis on the Illumina
511 HiSeq and MiSeq Platforms. The ISME Journal 6 (8): 1621–24.

Castro LR, Austin AD and Dowton M. 2002. Contrasting rates of mitochondrial molecular evolution in parasitic Diptera and Hymenoptera. Molecular biology and evolution 19(7): 1100–1113.

Catchen, J, Hohenlohe, P, Bassham, S, Amores A, & W Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology, 22(11):3124-3140

Clement M, Posada D, & Crandall, KA. 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology. 9(10):1657-9.

Cunningham, JP, Hereward, JP, Heard, TA, De Barro, PJ, & SA West. 2014. Bees at war: interspecific battles and nest usurpation in stingless bees. American Naturalist 184 (6): 777–86.

Dollin AE, Dollin LJ & SF Sagakami. 1997. Australian stingless bees of the genus *Trigona* (Hymenoptera:Apidae). Invertebrate Taxonomy. 11, 861-896.

Dowton, M & AD Austin. 1994. Molecular phylogeny of the insect order Hymenoptera: Apocritan relationships. Proc Natl Acad Sci USA. 91(21): 9911–9915.

Dyer AG, Streinzer M & J Garcia. 2016. Flower detection and acuity of the Australian native stingless bee Tetragonula carbonaria Sm. Journal of Comparative Physiology. 202: 629-639.

Elshire, RJ, Glaubitz, JC, Sun, Q, Poland, JA, Kawamoto, K, Buckler, ES & SE Mitchell. 2011. A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. PloS One 6 (5): e19379.

Feder, JL *et al*. 2003. 'Allopatric genetic origins for sympatric host-plant shifts and race formation in Rhagoletis', Proceedings of the National Academy of Sciences of the United States of America. National Acad Sciences, 100(18), pp. 10314–10319.

Fernando, LCP & Walter, GH. 1997. Species status of two host-associated populations of *Aphytis lingnanensis* (Hymenoptera: Aphelinidae) in citrus. Bulletin of Entomological Research, 87, 137-144.

Ford, J. 1974. Concepts of subspecies and hybrid zones, and their application in australian ornithology. Emu - Austral Ornithology 74 (3): 113–23.

Françoso, E, Zuntini, AR, Ricardo, PC, Silva, JPN, Brito, R, Oldroyd, BP & MC Arias. 2019. Conserved numts mask a highly divergent mitochondrial-coi gene in a species complex of australian stingless bees *Tetragonula* (Hymenoptera: Apidae). Mitochondrial DNA. Part A, DNA Mapping, Sequencing, and Analysis 30 (7): 806–17.

Franck P, Cameron E, Good G, Rasplus JY, & BP Oldroyd. 2004. Nest architecture and genetic differentiation in a species complex of Australian stingless bees. Molecular Ecology. 13(8) 2317-31.

Gloag R, Tan K, Wang Y, *et al*. 2017. No evidence of queen thelytoky following interspecific crosses of the honey bees *Apis cerana* and *Apis mellifera*. Insectes sociaux 64(2). Springer: 241–246.

549 Good, JM, Vanderpool, D, Keeble, S, & K Bi. 2015. Negligible nuclear introgression despite
550 complete mitochondrial capture between two species of chipmunks. Evolution; International
551 Journal of Organic Evolution 69 (8): 1961–72.

552 Green, CL, Franck, P & BP Oldroyd. 2001. Characterization of microsatellite loci for *Trigona*
553 *carbonaria*, a stingless bee endemic to Australia. Molecular Ecology Notes 1 (1-2): 89–92.

554 Green, CL, & BP Oldroyd. 2002 Queen mating frequency and maternity of males in the
555 stingless bee *Trigona carbonaria* Smith. Insectes sociaux 49(3): 196–202.

556 Gurevich, A, Saveliev, V, Vyahhi, N, & G Tesler. 2013. QUAST: Quality assessment tool for
557 genome assemblies. Bioinformatics. 29 (8): 1072–75.

558 Heard TA. 1994. Behaviour and pollinator efficiency of stingless bees and honeybees on
559 macadamia flowers. Journal of Apicultural Research 33(4):191–198.

560 Hines, HM, Cameron SA & PH Williams. 2006. Molecular phylogeny of the bumble bee
561 subgenus *Pyrobombus* (Hymenoptera:Apidae:Bombus) with insights into gene utility for
562 lower-level analysis. Invertebrate Systematics 20(3) 289-303 https://doi.org/10.1071/IS05028

563 Huelsenbeck, JP & F Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny.
564 Bioinformatics 17:754-755.

565 Jombart, T. 2008. adegenet: an R package for the multivariate analysis of genetic markers.
566 Bioinformatics 24: 1403-1405.

567 Kaltenpoth M, Showers Corneli P, Dunn DM, *et al*. 2012. Accelerated evolution of
568 mitochondrial but not nuclear genomes of Hymenoptera: new evidence from crabronid wasps.
569 PloS one 7(3): e32826.

570 Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7:
571 improvements in performance and usability. Molecular Biology and Evolution. 30(4): 772-780.

572 Kearse, M, Moir, R, Wilson, A, Stones-Havas, S, Cheung, M, Sturrock, S, Buxton, S, Cooper,
573 A, Markowitz, S, Duran, C, Thierer, T, Ashton, B, Meintjes, P, & A Drummond. 2012.
574 Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization
575 and Analysis of Sequence Data. Bioinformatics, 28, 1647-1649.

576 Kopelman, NM, Mayzel, J, Jakobsson, M, Rosenberg NA and I Mayrose. 2015. clumpak: a
577 program for identifying clustering modes and packaging population structure inferences across
578 K. Molecular Ecology Resources. 15(5): 1179–1191.

579 Lawson, DJ, van Dorp, L & D Falush. 2018. A Tutorial on How Not to over-Interpret
580 STRUCTURE and ADMIXTURE Bar Plots. Nature Communications 9 (1): 3258.

581 Leigh, JW, & D Bryant. 2015. Popart: full-feature software for haplotype network construction.
582 Edited by Shinichi Nakagawa. Methods in Ecology and Evolution / British Ecological Society
583 6 (9): 1110–16.

584 Linck, E, and Battey, CJ. 2019. Minor allele frequency thresholds strongly affect population
585 structure inference with genomic data sets. Molecular Ecology Resources 19 (3): 639–47.

586 Mallet J, Wynne IR, & CD Thomas. 2011. Hybridisation and climate change: brown argus
587 butterflies in Britain (*Polyommatus* subgenus *Aricia*). Insect Conservation and Diversity. 4 (3)
588 192-199.

589 Mardulyn, P, & SA Cameron. 1999. The Major Opsin in Bees (Insecta: Hymenoptera): A
590 Promising Nuclear Gene for Higher Level Phylogenetics. Molecular Phylogenetics and
591 Evolution. 12(2):168-76.

592 Oliveira DCSG, Raychoudhury R, Lavrov DV, *et al*. 2008. Rapidly evolving mitochondrial
593 genome and directional selection in mitochondrial genes in the parasitic wasp nasonia
594 (hymenoptera: pteromalidae). Molecular biology and evolution 25(10): 2167–2180.

595 Perea, S, Vukić, J, Šanda, R & I Doadrio. 2016. Ancient mitochondrial capture as factor
596 promoting mitonuclear discordance in freshwater fishes: A case study in the genus *Squalius*
597 (Actinopterygii, Cyprinidae) in Greece. PloS One 11 (12): e0166292.

598 Peterson, BK, Weber, JN, Kay, EH, Fisher, HS & HE, Hoekstra. 2012. Double digest RADseq:
599 an inexpensive method for de novo snp discovery and genotyping in model and non-model
600 species. PloS One 7 (5): e37135.

601 Poland, JA, Brown, PJ, Sorrells, ME, & JL Jannink. 2012. Development of high-density genetic
602 maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach.
603 PloS One 7 (2): e32253.

604 Pritchard, JK, Stephens, M, & P Donnelly. 2000. Inference of population structure using
605 multilocus genotype data. Genetics 155 (2): 945–59.

606 Rasmussen C & SA Cameron. 2009. Global stingless bee phylogeny supports ancient
607 divergence, vicariance, and long distance dispersal. Biological journal of the Linnean Society.
608 Linnean Society of London 99(1): 206–232.

609 Remnant EJ, Koetz A, Tan K, *et al.* 2014. Reproductive interference between honeybee species
610 in artificial sympatry. Molecular ecology 23(5): 1096–1107.

611 Ridley, AW, Hereward, JP, Daglish, GJ, Raghu, S, McCulloch, GA & GH Walter. 2016. Flight
612 of *Rhyzopertha* Dominica (Coleoptera: Bostrichidae)-a spatio-temporal analysis with
613 pheromone trapping and population genetics. Journal of Economic Entomology 109 (6): 2561–
614 71.

615 Sánchez-Guillén RA, Muñoz J, Rodríguez-Tapia G, Arroyo TPF, & A Córdoba-Aguilar. 2013.
616 Climate-induced range shifts and possible hybridisation consequences in insects. PloS One.
617 8(11) e80531.

618 Schuelke, M. 2000. An economic method for the fluorescent labelling of pcr fragments. Nature
619 Biotechnology 18 (2): 233–34.

620 Shipham, A, Joseph, L, Schmidt, DJ, Drew, A, Mason, I & JM Hughes. 2019. Dissection by
621 genomic and plumage variation of a geographically complex hybrid zone between two
622 australian non-sister parrot species, *Platycercus adscitus* and *Platycercus eximius*. Heredity
623 122 (4): 402–16.

624 Smith TJ. 2020. Evidence for male genitalia detachment and female mate choice in the
625 Australian stingless bee *Tetragonula carbonaria*. Insectes Soc 67: 189–193.

626 Toon, A, Daglish, GJ, Ridley, AW, Emery, RN, Holloway JC, & GH Walter. 2015. Random
627 mating between two widely divergent mitochondrial lineages of *Cryptolestes ferrugineus*

(Coleoptera: Laemophloeidae): a test of species limits in a phosphine-resistant stored product pest. Journal of Economic Entomology. 109(5):2221-8.

Twyford, AD & RA Ennos. 2011. Next-generation hybridisation and introgression. Heredity 108: 179-189.

Vollet-Neto A, Koffler S, dos Santos CF, *et al*. 2018 Recent advances in reproductive biology of stingless bees. Insectes sociaux 65(2). Springer: 201–212.

Vollet-Neto A, Oliveira RC, Schillewaert S, *et al*. Diploid male production results in queen death in the stingless bee *Scaptotrigona depilis*. Journal of Chemical Ecology. 43(4):403-410.

Walter GH. 2003. Insect pest management and ecological research. Cambridge University Press, Cambridge, p 162-165.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 20(17):3252-3255.

Xiao J-H, Jia J-G, Murphy RW, *et al*. 2011 Rapid evolution of the mitochondrial genome in Chalcidoid wasps (Hymenoptera: Chalcidoidea) driven by parasitic lifestyles. PloS one 6(11): e26645.

Zych, K, Basten, LS, Elvin, M, Rodriguez, M, Van der Velde, KJ, Arends, D, Westra, HJ *et al.* 2017. reGenotyper: Detecting Mislabeled Samples in Genetic Data. PloS One 12 (2): e0171324.

Author Contributions

All authors conceived the project, TJS obtained samples, JPH conducted molecular lab work, JPH and DRB analysed the genetic data. All authors wrote the manuscript.

Supplementary data

Supplementary Figure 1. Mauve alignment of the mitogenomes showing the inversion between *Tetragonula* and other bees. Supplementary Figure 2. PCA plot of the *snp* data highlighting the samples from Rockhampton (Queensland), Newcastle, and Sydney (both New South Wales). Supplementary Data 1. Protocol for the genotyping by sequencing assay. Supplementary Table 1. Spreadsheet of the GBS adaptors.

Figures and Tables

Table 1. Characteristics of the genomes and their draft assemblies for three species of *Tetragonula* stingless bees, including genome size, coverage, and % repeat, which were estimated with a kmer analysis in bbnorm. The mitogenome coverage was assessed by read mapping, and the assembly statistics (length, longest contig, n50, number of contigs) were calculated in Quast. The genome assemblies can be found at genbank under the provided bioproject and biosample IDs.

Species	Genome size estimate	Nuclear genome coverage	Mitogenome coverage	% repeat	Assembly length	Longest contig	N50	N contigs	BioProject	BioSample
T. carbonaria	457,237,826	23x	6,899	33%	284,886,424	187,309	14,476	156,012	PRJNA578948	SAMN13088838
T. hockingsi	376,762,141	35x	11,089	26%	288,263,186	200,717	11,462	225,627	PRJNA578994	SAMN13089661
T. davenporti	342,107,567	38x	26,907	17%	288,842,417	271,029	18,924	72,464	PRJNA578998	SAMN13089823

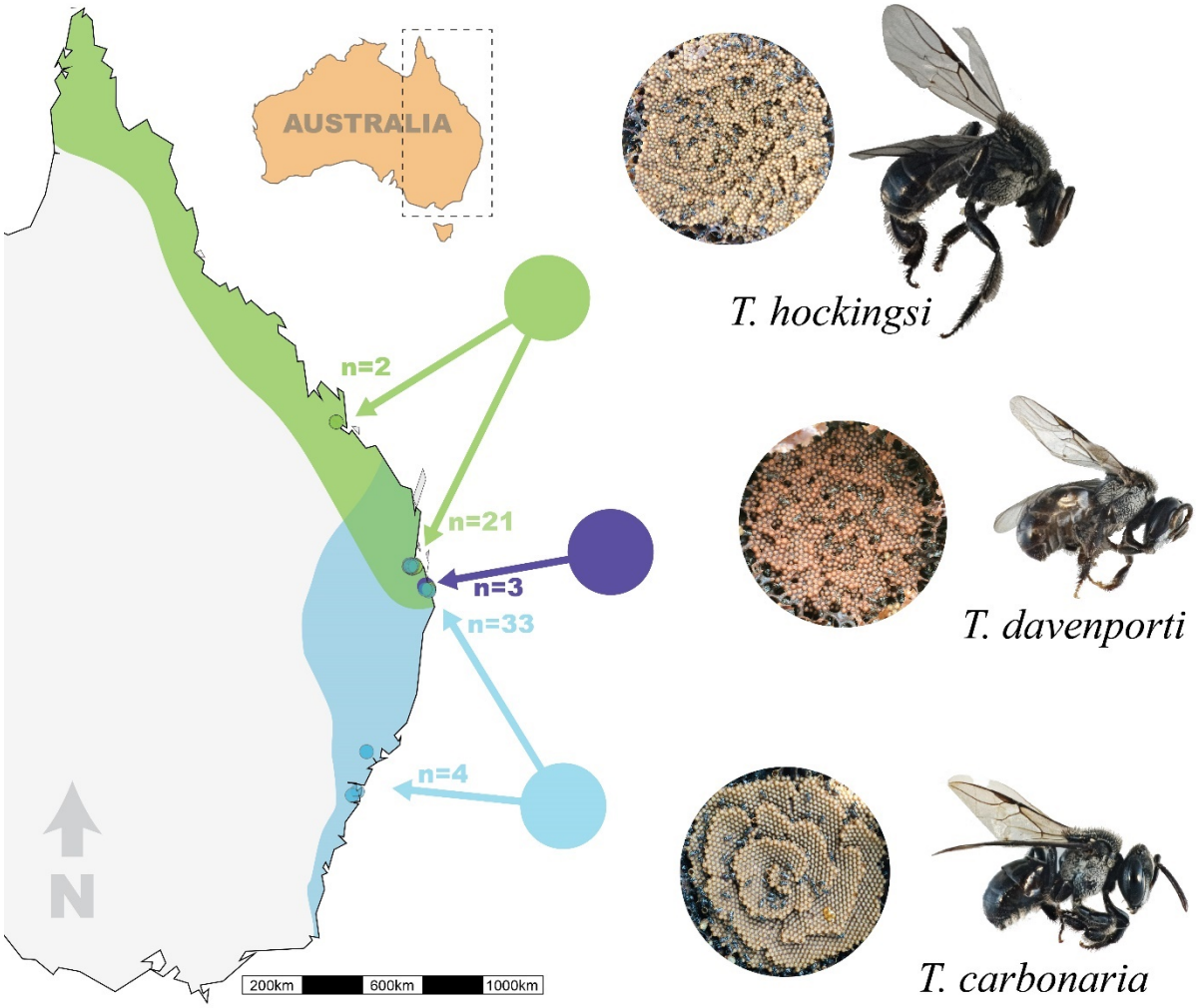

Figure 1. Map showing the known distributions of three species of *Tetragonula* in Australia, sampling sites are shown as small coloured circles, and numbers used in this study denoted by different colours for each species. The brood structure and the bees are shown on the right, and the distributions are based on Heard (2016).

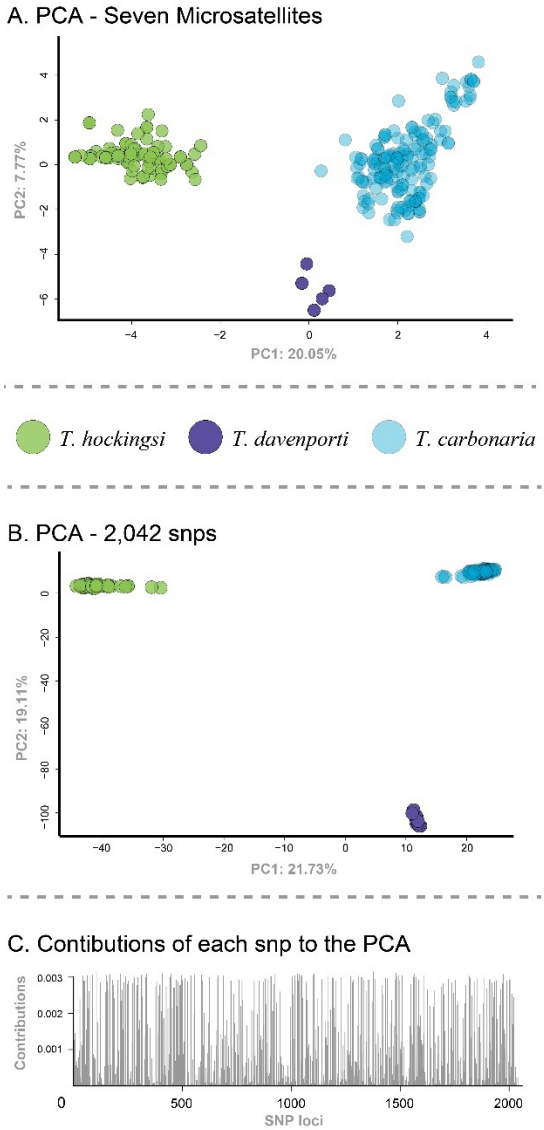

Figure 2. Principal Components Analysis (PCA) of genetic distance based on nuclear DNA markers. Plot showing principal components one and two based on the seven microsatellite markers (A), the same plot based on the 2,049 nuclear *snp* markers (B), and a loading plot showing the relative contributions of each of the 2,049 *snp* markers to the PCA (C).

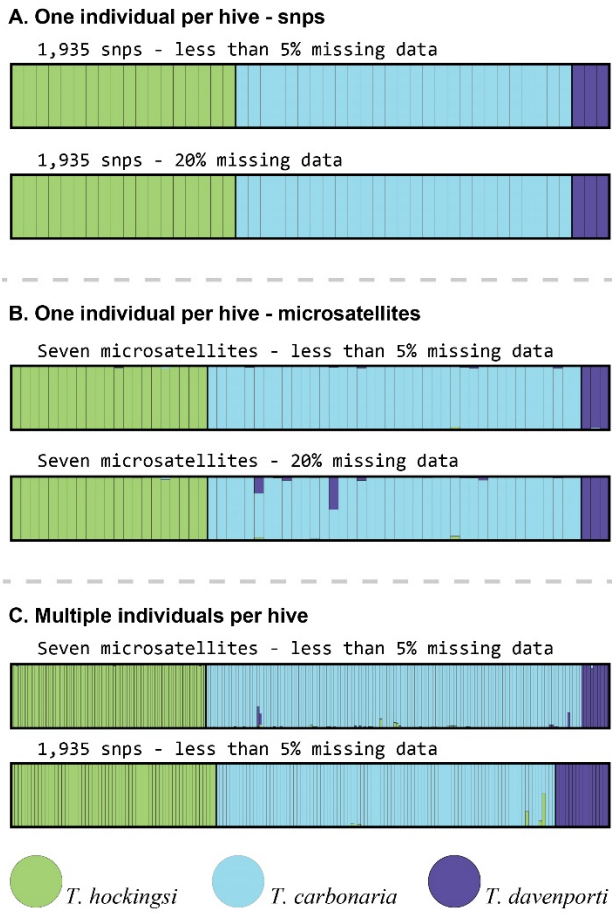

Figure 3. Structure plots assuming three hypothetical populations (K = 3), each bar represents an individual and the proportions of the three colours represent the posterior probability of that individual belonging to each of the three hypothetical (K) populations. Analyses were conducted for two *snp* datasets, one with 1,935 *snps* and less than 5% missing data, and one with the same number of *snps* with 20% missing data, using one individual per hive (A). The same analyses were conducted on the two datasets based on seven microsatellite loci and one individual per hive, one with less than 5% missing data and one with 20% missing data (B). The analyses were also run with multiple individuals per hive, mostly four but sometimes fewer, to examine the effect of including closely related siblings (C).

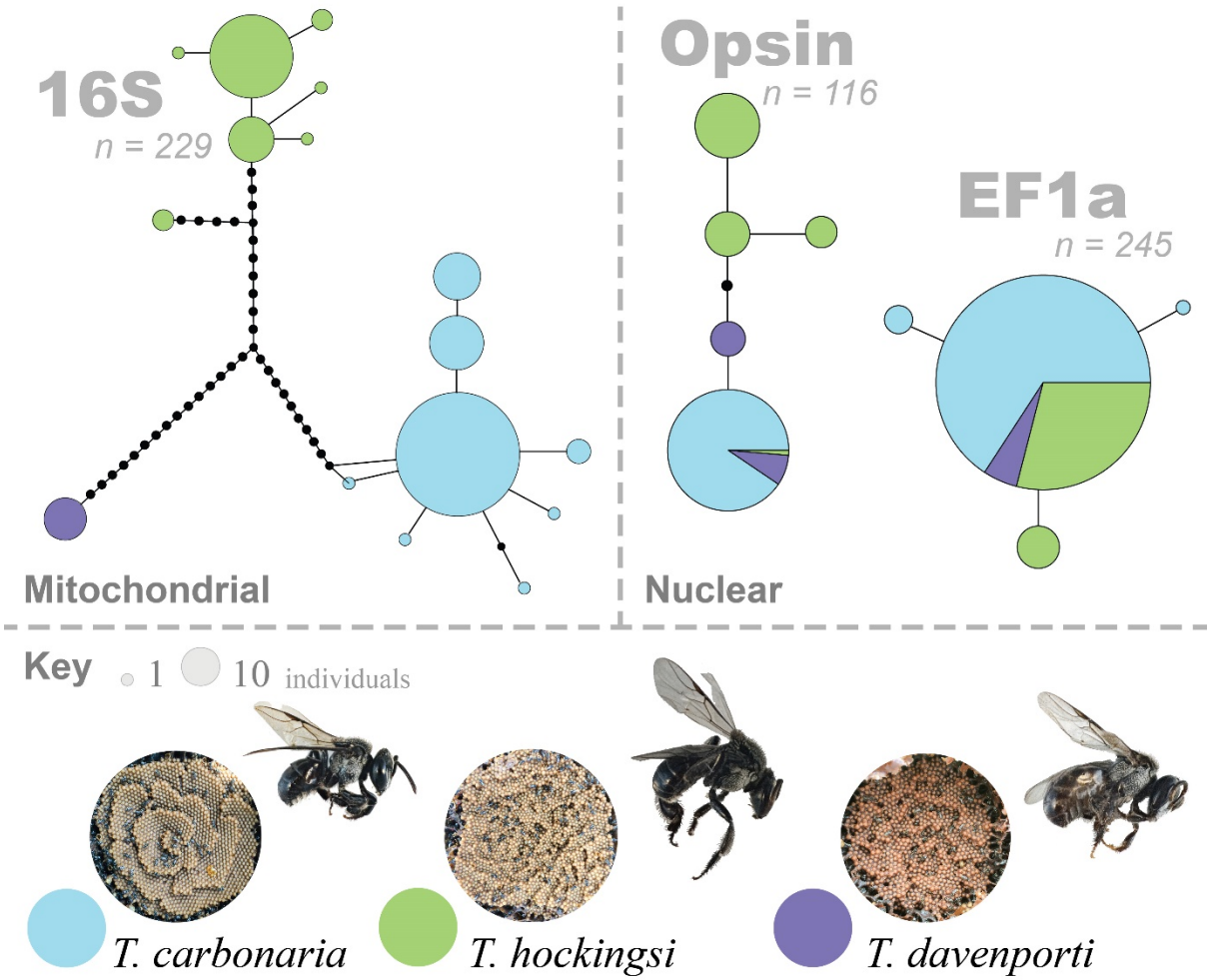

Figure 4. Haplotype networks for the mitochondrial *16S* gene, and nuclear *Opsin* and *EF1alpha* (top). Brood of the three species we investigate is illustrated as are representative bees of each of the three *Tetragonula* species in the key (bottom).

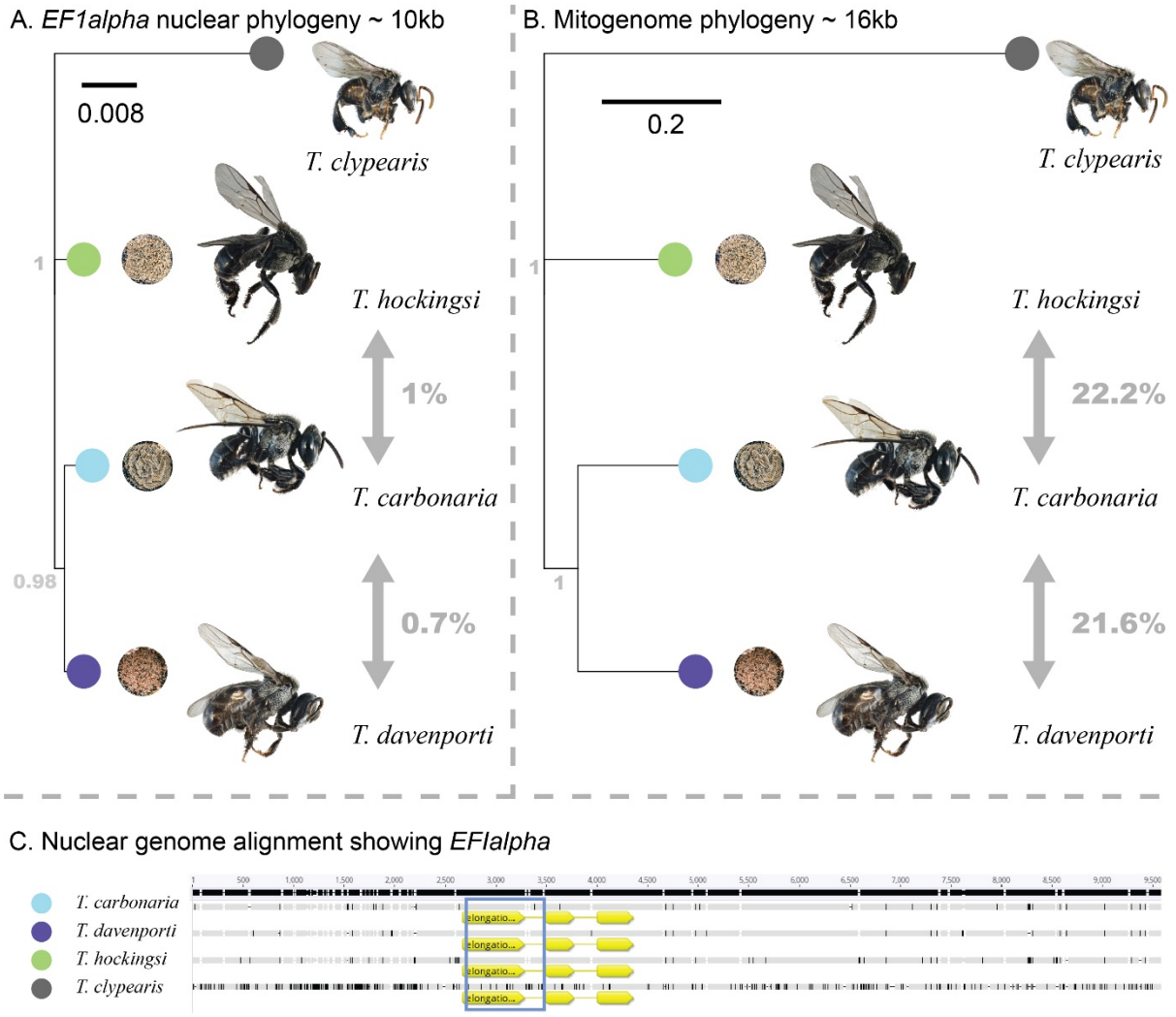

Figure 5. Bayesian tree of four Australia *Tetragonula* sp generated from a 10kb surrounding the nuclear *EF1alpha* gene (A). Bayesian tree constructed from whole mitogenomes (B). The nuclear genome alignment of *EF1alpha* and flanking regions; the blue rectangle represents the PCR product (C).