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

Authors: James P. Hereward[1], Tobias J. Smith[1], Dean R. Brookes[1], Ros Gloag[2], 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 microsatellites proposed that rare hybrids do occur. 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 revealed no evidence of hybridisation. We found no inter-species variation in 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 discord 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 and low nuclear divergence. The mitochondrial genomes indicate that divergence could be deep between these lineages but the low divergence in the nuclear genomes means that further investigation is required before divergence times can be estimated accurately. The mitogenomes, and draft nuclear genomes provided for these species will be a resource for further molecular studies on this group.

Introduction

Allopatry is a potent driver of speciation, but species boundaries may be tested when populations that have diverged in allopatry come back into secondary contact due to changes in habitat or climate, or as a result of dispersal by humans over geographic boundaries (Mallet *et al*., 2011; Sánchez-Guillén *et al*., 2013). The process of hybridisation is closely tied to species concepts and their designation in population genetic terms (Paterson 1978); by definition, hybridisation cannot occur within a species. When species are considered conceptually as recombining species gene pools (Ayala *et al.* 1994), hybridisation is a 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 are subspecies, which, by definition, have independent geographical distributions but hybridise freely where and when they do overlap (Walter 2003, Ford 1974). Such situations are usually seen when closely-related but divergent lineages of the same species are brought back into sympatry (Feder *et al*. 2003), but sometimes this pattern is seen across 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 particular species and their limits, and is key to our understanding of speciation.

Understanding hybridisation relies not only on our concept of species generally and on definitions of particular species, but also on the methods that are available to detect it. Accurate assessment of hybridisation is the first step in shedding light on the mechanisms that might prevent gene flow between diverged populations. The more recently two populations have diverged, the more likely they have not yet become separate species, and thus the more likely that gene flow is re-established (Feder *et al*. 2003). Hybrid detection between close-relatives is expected to be 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 or cryptic in appearance. A species complex of stingless bees (Meliponini) in the genus *Tetragonula* from Australia illustrates this well.

The stingless bee (Meliponini) genus *Tetragonula* includes three cryptic sister species (*T. carbonaria. T. hockingsi* and *T. davenporti*) native to eastern Australia that are part of a monophyletic clade known as the "Carbonaria group" (Franck *et al*., 2004). These bees can be readily kept in wooden hives, are popular pets, and are effective pollinators of some tropical and subtropical crops (Dyer *et al.* 2016, Heard 1994), leading to many thousands of colonies kept by beekeepers in Eastern Australia. Today, the distributions of all three species overlap in parts of south-east Queensland (Fig 1). However, the historical distributions of the species are poorly known due to their cryptic morphology. It is likely that beekeeping and other human activities, such as land use change, have greatly increased the overlap of their distributions in recent decades, particularly via the reported southward spread of *T. hockingsi*. 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 and *T. hockingsi* producing clusters of cells in a form known as semi-comb (Fig. 2). *T. 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). No behavioural barriers to reproduction between the species are known. Indeed, male mating congregations, which form outside colonies with virgin queens, 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.

Several previous efforts have been made to establish if, and how regularly, gene flow occurs between these species, with equivocal results. First, microsatellite markers have been used to look for evidence of hybrids. Frank *et al.* (2004) identified some Queensland colonies of *Tetragonula* as being of hybrid origin based on microsatellite data and defined as a sample having a posterior probability of assignment of less than 90% to a cluster in the *STRUCTURE* (Pritchard *et al.* 2000) analysis. Franck *et al.* (2004) also attempted to characterise eight other species of *Tetragonula* genetically, from Australia and Thailand, based on the *cytochrome b* (*cytb*) mitochondrial gene, but the outcome was poor because they detected issues with pseudogenes. 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 southeast Queensland were reported (within the distribution overlap zone of *T. carbonaria* and *T. hockingsi*), using a cut-off of 95% posterior probability of belonging to a cluster in a *STRUCTURE* 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. In interpreting these results, Brito *et al.* (2014) suggested that *T. hockingsi* represents two genetic clusters (based on their microsatellite and mitochondrial data), one each in northern and southern Queensland. Their *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, but nuclear copies of mitochondrial genes (numts) have regularly hampered these attempts. Most recently, several numts of Cytochrome Oxidase Subunit I (*COI*) were identified in *T. hockingsi* and *T. carbonaria* (Françoso *et al*. 2019). This cast into question earlier work on hybrids that incorporated COI-numts, and even the designation of *T. davenporti* as a true species. Surprisingly, the genuine COI sequences of these species are highly diverged (16.5 %) relative to available nuclear genes (Cunningham 2014, Francosco 2019). This pattern of high mitochondrial divergence but low nuclear divergence might be consistent with recent or ongoing gene flow, and further shows that new analyses are needed to understand the extent to which species in this complex are discrete, and the potential impact of continuing movement of hives around Australia.

Here we aim to resolve whether there is evidence of ongoing gene flow between *T. carbonaria, T. hockingsi* and *T. davenporti* kept by beekeepers in sympatry in South East Queensland, using both microattelites and a *genotyping by sequencing* approach that assayed 2,049 *snp* markers. We also look for evidence of mitochondrial introgression, by considering two nuclear gene sequences and one mitochondrial gene sequence of each species, and use Illumina sequencing to reconstruct mitochondrial genomes and thus confirm that mitochondrial sequences are not pseudogenes. Finally, we assembled draft nuclear genomes for these three species as a resource for further genetic analysis of the species complex. This comprehensive approach leads to new insights about the extent of hybridisation in these bees, and their species status.

Materials and Methods

141 *Sampling and DNA extraction*

142 We sampled four workers per colony from hived colonies of each species between
143 February 2017 and April 2019 in south-east Queensland (*T. carbonaria*: N=39, *T. hockingsi*,
144 N=21, *T. davenporti*, N=3); (Fig 1). Most *T. carbonaria* and *T. hockingsi* colonies were
145 sampled in the greater Brisbane region. We opened each hive and assigned it to species based
146 on brood nest architecture (spiral comb or semi-comb). The three *T. davenporti* colonies were
147 sampled from the border area of Queensland and New South Wales and were the same colonies
148 used for past genetic work on this species (P. Davenport, pers. comm. 2019). A further two *T.*
149 *carbonaria* colonies and one *T. hockingsi* colony were sampled from the same meliponary as
150 the three *T. davenporti* colonies. We extracted DNA using a silica spin-column method (Ridley
151 *et al* 2016) and normalized concentrations to 5ng/ul per sample following quantification
152 (PicoGreen, Life Technologies, California, USA).

153

154 *Neutral nuclear markers*

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

165 Exonuclease I and one unit of Antarctic Phosphatase as above, and genotyped by Macrogen
166 Inc. (Seoul, Republic of Korea). Microsatellite peaks were analysed using the microsatellite
167 plugin in Geneious v. 2019.2.1 (http://www.geneious.com, Kearse *et al*. (2012)).

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

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

186 *Analyses*: For both our microsatellite and *snp* datasets, we performed principal component
187 analyses (PCA) using the adegenet package in R (Jombart, 2008) and *STRUCTURE* analyses
188 (Pritchard 2000). *STRUCTURE* is an individual-based clustering algorithm that assigns

189 individuals to each of *K* population clusters using Hardy-Weinberg equilibrium and linkage
190 information in genetic markers. To assess the effect of replication and missing data on
191 hybridisation inferences from the *STRUCTURE* output, we repeated the analyses on three
192 versions of each of our two datasets: (i) a dataset including one worker per colony, (ii) a dataset
193 including all four workers per colony, and (iii) the dataset with 20% randomly deleted. For
194 the microsatellite data, data was removed evenly across markers until 20% missing was
195 reached, for *snps* we changed the missing data threshold in the filtering step to 20%, and this
196 increased the number of markers to 13,409 loci. For each *STRUCTURE* run, we used the
197 admixture model and ran 500,000 iterations of burn-in followed by 1 million iterations. We
198 assumed both two or three populations (K=2 and K=3) and performed multiple runs that were
199 then permuted and plotted using Clumpak server (Kopelman *et al.* 2015).

200

201 *Gene sequencing*

202 To confirm whether common marker genes showed species-specific polymorphisms,
203 we sequenced two nuclear genes (*EF1alpha* and *Opsin*) and one mitochondrial gene (16S).
204 These genes were selected as they have previously been used for a phylogenetic study of
205 stingless bees (Rasmussen and Cameron, 2009). For *EF1alpha* we used primers F2-ForH
206 (GGRCAYAGAGATTTCATCAAGAAC) and F2-RevH2
207 (TTGCAAAGCTTCRKGATGCATTT) (Hines 2006). For *Opsin* we used LWRhF
208 (AATTGCTATTAYGARACNTGGGT) and LWRhR
209 (ATATGGAGTCCANGCCATRAACCA) (Mardulyn & Cameron 1999). For 16S we used
210 primers 874-16S1R (TATAGATAGAAACCAAYCTG) (Cameron 1992) and 16SWb
211 (CACCTGTTTATCAAAAACAT) (Dowtown & Austin 1992). We performed PCRs in 25µl
212 reactions containing 2 µl DNA template, 1U MyTaq polymerase (Bioline, Australia), 0.2 µM

213 of each PCR primer, and 1x buffer. For amplification of *EF1alpha* and 16S we performed PCR
214 with 35 cycles at 52ºC annealing, and for *Opsin* we used five cycles at 45ºC annealing followed
215 by 35 cycles at 50ºC. PCR products were cleaned using 1U of Exonuclease I and Antarctic
216 Phosphatase (New England Biolabs, Ipswich, Mass., USA) and incubating at 37 °C for 20 min
217 followed by 10 min enzyme denaturation at 80 °C. Sequencing was conducted using the same
218 forward and reverse primers used for PCR, by Macrogen Inc. (Seoul, Republic of Korea). We
219 edited sequence data with CodonCode aligner, and aligned them using MAFFT in Geneious
220 Prime 2019.2.1 (www.geneious.com, Kearse *et al*. 2012). We generated haplotype networks
221 using the TCS algorithm (Clement *et al.* 2000) in PopART version 1.7
222 (http://popart.otago.ac.nz/) (Leigh and Bryant 2015).

223

224 *Mitogenome sequencing and assembly*

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

233

234 Sequences were checked with FastQC (Andrews, 2019), then we removed adaptors,
235 and quality-trimmed the end of the sequences of all samples to q10 using bbduk from the
236 bbtools package (version 36, Bushnell, 2019), with a kmer of 8. We binned the sequences by

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

251

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

257

258 *Nuclear genome assembly*

259 To further investigate the low nuclear divergence suggested by our two nuclear genes
260 (*Opsin* and *EF1alpha*) we considered the remaining genome-wide data generated from our

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

269

270 Results

271 *Tests of hybridization using microsatellites and snps*

272 PCAs using both microsatellite and *snp* datasets separated samples into three non-
273 overlapping clusters consistent with their species assignment ((<5% missing data, Fig. 3A).
274 The microsatellite PCA placed *T. davenporti* closer to *T. carbonaria* than to *T. hockingsi* (Fig.
275 3A), but the *snp* PCA indicated an equal genetic distance between the three clusters (20% of
276 the variance explained by each of the first two components, Fig. 3B).

277

278 *STRUCTURE* analyses using both microsatellite and *snp* datasets (N= X colonies, 1
279 worker per colony) returned no evidence of hybrids, with all individuals showing close to 100%
280 assignment to a species' cluster (Fig. 4). Notably however, STRUCTURE results were
281 sensitive to duplicated and missing data. The same analyses using datasets comprising four
282 workers per colony (akin to resampling the same individual multiple times), assigned X-X
283 individuals at less than 90% posterior probability of belonging to their species' cluster, and

284 thus putative hybrids (Fig. 4). Likewise, subsampling to 20% missing microsattelite data
285 resulted in two *T. hockingsi* individuals being designated as hybrids with *T. davenporti* Missing
286 data had no impact on the *snp* dataset (i.e. every individual was still assigned 100% to its cluster
287 when lower threshold for missing data shifted marker number from 1,935 to 13,409).

288

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

290 The mitochondrial gene *16S*, and the nuclear gene *Opsin*, each grouped samples into three
291 haplotypes consistent with species assignment (Fig. 2) although pairwise divergence between
292 species was greater for 16S (229 individuals representing all 63 colonies ; ~6% between *T.*
293 *carbonaria* and *T. hockingsi*) than *Opsin* (116 individuals representing all 63 colonies ; ~0.5%
294 difference between *T. carbonaria* and *T. hockingsi*). The percent difference in the *16S* gene
295 region amplified was lower than that of the complete mitochondria (see below), because one
296 half of the PCR region showed no variation. The nuclear gene *EF1alpha* was identical in
297 sequence for all samples (245 individuals representing all 63 colonies). No individuals showed
298 mitochondrial-nuclear discordance between *Opsin* and *16S*. That is, if an individual had a *16S*
299 sequence representative of *T. carbonaria* it also had an *Opsin* sequence representative of *T.*
300 *carbonaria* (Fig. 2). However, one colony that was labelled *T. carbonaria* during field
301 collections returned sequences (and nuclear markers) consistent with *T. hockinsgsi*, and this
302 was assumed to be the result of either labelling error, or a recent interspecific colony takeover
303 (1 of 69, 1.5% sample ID error rate). We also detected one instance of miss-labelling between
304 the sequence dataset and the microsatellite dataset because the same plate of individuals had
305 been extracted twice and two samples had switched positions in one of these plates.

306

307

308 *Mitogenomes and genomes*

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

318

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

330

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

Discussion

Testing for hybridisation

Understanding hybridisation is crucial to understanding species and speciation, but detecting hybridisation relies critically on the methods used to infer it, as illustrated here using a range of techniques. Incorrect inference of hybridisation can lead to the species status of an organism being questioned, as was the case with these stingless bees. Initial reports of hybridisation (Franck *et al.* 2004 & Brito *et al.* 2014) would suggest these species were actually subspecies that had recently come into secondary contact, especially given their close ecological interaction in SE Queensland. We find, conversely, that these three species, *T. carbonaria*, *T. hockingsi* and *T. davenporti*, are good species with no evidence for hybridisation in the areas of sympatry that we sampled (including in colonies of all three species residing together at one property for at least nine years). The mitogenome data indicate that the divergence between these species is deep (21.6 to 23.6% difference, and at least 8 Mya for the closest relatives). Although the nuclear genome does show much less variation (0.7-1%), this ratio of nuclear to mitochondrial divergence (~22x) is within the known natural range (0.8-40x) described in a recent comparative study across metazoans, and within the range reported specifically for insects (Allio *et al.* 2017).

354 In this study we analysed multiple genetic markers to assess the extent of hybridisation
355 (if any) in these closely related sympatric species. Each marker set was found to have its own
356 set of issues that can lead to incorrect inferences about the extent of hybridisation. We discuss
357 issues highlighted by this study in the following sections.

358

359 *Extremely low levels of nuclear divergence*

360 We found extremely low levels of nuclear divergence in this study relative to the
361 divergence rates of the mitochondrial sequence. This pattern has often been considered
362 evidence for ongoing nuclear geneflow despite different mitochondrial backgrounds. A recent
363 study suggests that this ratio varies substantially across taxa (Allio *et al*. 2017). The nuclear-
364 mitochondrial ratio in this group of bees does seem to be extremely high relative to that reported
365 for other insects, but the pattern we observed when considering the larger fragments of the
366 nuclear genome was consistent and one cannot automatically infer hybridisation from this
367 pattern. Given the data we have generated it is not possible to determine whether nuclear gene
368 flow has occurred across divergent mitochondrial lineages for some period of time, but our
369 data indicates that no gene flow is occurring across the nuclear genome now.

370

371 We investigated the lack of variation in the *EF1alpha* PCR product further to see
372 whether it was restricted to the 700bp *pcr* product. We were able to recover more sequence
373 from the whole genome assemblies. These larger fragments did exhibit variation, and from
374 that, we recovered the same phylogeny (Fig. 5) as from the mitogenomes. The ratio of mito-
375 nuclear divergence is the same across the three species included in this analysis. Hymenoptera
376 are regarded as having a higher rate of mitochondrial divergence compared to other insects
377 (Kaltenpoth 2012). The extremely large discrepancy across the mitochondrial and nuclear rates

378 in these species requires further work to establish whether this derives from a fast
379 mitochondrial rate, a slow nuclear rate, or a combination of these two. This further investigation
380 is required before divergence times can be estimated more realistically in this group of bees.

381

382 *Mito-nuclear pseudogenes*

383 Two earlier studies on the three *Tetragonula* species we investigated amplified nuclear
384 pseudogenes instead of the true mitochondrial gene region. This is a common problem in
385 mitochondrial DNA based tests, but is more prevalent in some insect groups than others.
386 Recently, Françoso *et al.* (2019) successfully re-designed primers for COI that avoid picking
387 up pseudogenes in these *Tetragonula* species. Researchers are now much more aware of the
388 issue of pseudogenes and it is common to check for stop codons or frame-shifting indels, but
389 this scrutiny detects only a subset of pseudogenes. The ability to reconstruct whole
390 mitogenomes from short read data (and using a depth cut-off, as we did, to avoid nuclear
391 copies) provides a way to detect pseudogenes more comprehensively. There are likely many
392 reports of hybridisation based on undetected pseudogenes in the literature, inflating perceptions
393 of the extent of hybridisation between species, and contributing misinformation to the debate
394 on species concepts.

395

396 *Mito-nuclear discordance*

397 Hybridisation (and thus gene flow) between population samples is often inferred from
398 discordance across mitochondrial and nuclear gene sequences. This pattern may well be
399 representative of recent hybridisation, but may also be caused by rare events, as seen with
400 examples of mitochondrial capture (Good *et al.* 2015; Perea *et al.* 2016). Another issue with
401 this measure of hybridisation is that the pattern (a sample returning one mitochondrial sequence

402 and a different nuclear one) is the same as one gets when samples are mislabelled. A recent
403 study indicates that rates of mislabelling are around 4% in published genetic datasets (Zych *et*
404 *al.* 2017), so any cases of mito-nuclear discordance at these rates probably deserve additional
405 scrutiny. We detected one instance of mis-labelling across two of our datasets, by virtue of
406 having conducted multiple genetic tests on the same individuals. To detect this kind of
407 'accidental introgression' any samples showing mito-nuclear discordance should be re-
408 amplified and sequenced again.

409

410 *Incorrect inference from STRUCTURE plots*

411 The two earlier papers that report hybridisation among the three *Tetragonula* species
412 inferred it on the basis of individuals having less than 90% assignment to a cluster in a
413 *STRUCTURE* plot (Franck *et al.* 2004; Brito *et al.* 2014). However, *STRUCTURE* quite often
414 mis-assigns individuals and this can be influenced by many factors. Warnings have been
415 published about over-interpreting structure plots (Lawson *et al.* 2018), but it is still a prevalent
416 issue. With bees, living in colonies of closely-related individuals, the inclusion of four
417 individuals per hive (in our analysis), caused *STRUCTURE* to assign several individuals in a
418 manner consistent with hybridisation. Clearly, in this case, this is simply a consequence of a
419 biased sample of the population having been presented to the algorithm, because the analysis
420 on single hive representatives did not show any suggestions of hybridisation.

421

422 We investigated the effect of missing data on inferring hybridisation from
423 *STRUCTURE* plots. Our two original datasets had very low levels of missing data (<5%) and
424 every individual was assigned at close to 100% when only one bee per hive was used (Fig. 4).
425 When we subsampled the two datasets to 20% missing data, the microsatellites indicated

426 hybrids, but the *snp* dataset did not. This was probably because the '20% missing' *snp* dataset
427 had many more markers than the '5% missing' microsatellite dataset. The microsatellite results
428 illustrate how missing data can lead to incorrect inferences of hybridisation from population
429 assignment tests like STRUCTURE. Including more markers appears to compensate for the
430 effect of missing data, at least based on our *snp* results (Fig. 4).

431

432 *Using multiple nuclear markers*

433 In cases where the nuclear genome divergence is low, using multiple markers, and
434 including more rapidly evolving markers are two ways in which to assess hybridisation rates
435 more accurately. In this study we used 7 microsatellites (more variable than *snps*), and also
436 assayed 2,049 *snps* (using more markers). Both of these marker systems revealed no evidence
437 for hybridisation in the PCA plots, with no individuals intermediate between clusters (Fig. 3).
438 This kind of multi-marker nuclear genome assay is the most accurate way in which to infer
439 hybridisation (Twyford & Ennos 2011), with a parentage analysis providing the most
440 convincing evidence of recent hybridisation, where it is known to have occurred.

441 Our results indicated much tighter clustering of the *snp* data than the microsatellite data
442 in the PCA plots, and less ambiguous assignments in *STRUCTURE*. This observation indicates
443 that misidentification of hybridisation will be less common in *snp* datasets or around a couple
444 of thousand markers than might be the case when using relatively few microsatellites.

445

446 *Tetragonula davenporti and the 'carbonaria complex'*

447 Using multiple markers, we have comprehensively tested for hybridisation in three
448 *Tetragonula* bees that overlap in their geographic distribution. Although previous studies have

449 reported that these species hybridise in these areas, our approach found no evidence for it.
450 Despite swarms of these bees taking over each other's hives and having mixed species drone
451 aggregations, the three species are all good species, likely diverged from one another millions
452 of years ago, and do not hybridise even when in close sympatry. *Tetragonula davenporti* has
453 been confirmed from only one property in the Gold Coast hinterland. Its species status had
454 been questioned recently (Françoso *et al.* 2019), but we find strong evidence for the validity of
455 *T. davenporti* as a species. The natural range of this species is currently unknown, and the
456 difficulty in identifying adults (which are almost identical to those of *T. carbonaria*) means
457 that their distribution in managed colonies is also relatively unknown. South-east Queensland
458 is an area of high endemicity, so it may well be endemic in this area. This study investigated
459 managed colonies mostly from Brisbane and SE Queensland because this is where
460 hybridisation is most likely between these three species. Examining wild populations of this
461 genus may reveal additional species within the '*carbonaria* complex' (perhaps the 'northern'
462 *T. hockingsi*).

463

464 *Conclusions*

465 We have highlighted several ways in which hybridisation might be incorrectly inferred
466 when using certain marker types and certain types of inference. It is quite likely that many
467 published records of hybridisation might be affected by some of these issues. In particular,
468 hybridisation inferred from a low frequency (<5%) of mitochondrial-nuclear discordance. This
469 inference comes from the same pattern as miss-labelling, which is also usually around this
470 frequency. We suggest these considerations should inform the design of genetic studies that
471 investigate hybridisation and also inform the meta-analysis of genetic studies of hybridisation.
472 We have also shown how whole-genome data can provide additional resolution when the rate

473 of nuclear mutation is low, and this should allow further resolution of closely related lineages.
474 The genomic resources developed here will enable further molecular investigation of the
475 evolution of this group of social, stingless bees, and enable further analysis of the divergence
476 times and ratio of mitochondrial to nuclear divergence rates.

477

478 Acknowledgements

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

482

483 Data Availability

484 Genome assemblies and mitogenomes have been deposited at genbank (pending
485 accession numbers).

486 Illumina data has been deposited at the Short Read Archive, (awaiting final SRA
487 numbers.)

488 Microsatellite data and *snp* vcf files are to be deposited at the University of Queensland
489 espace (doi to follow).

490

491

492

493

494

495 References

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

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

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

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

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

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

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

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

516 Cameron, SA, Austin, AD, James Derr, & RA Wharton. 1992. The application of nucleotide
517 sequence data to phylogeny of the Hymenoptera: a review. Journal of Hymenoptera Research.
518 1:63-79.

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

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

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

526 Cunningham, JP, Hereward, JP, Heard, TA, De Barro, PJ, & SA West. 2014. Bees at war:
527 interspecific battles and nest usurpation in stingless bees. American Naturalist 184 (6): 777–
528 86.

529 Dollin AE, Dollin LJ & SF Sagakami. 1997. Australian Stingless Bees of the Genus *Trigona*
530 (Hymenoptera:Apidae). Invertebrate Taxonomy. 11, 861-896.

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

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

535 Elshire, RJ, Glaubitz, JC, Sun, Q, Poland, JA, Kawamoto, K, Buckler, ES & SE Mitchell. 2011.
536 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species.
537 PloS One 6 (5): e19379.

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

541 Ford, J. 1974. Concepts of Subspecies and Hybrid Zones, and Their Application in Australian
542 Ornithology. Emu - Austral Ornithology 74 (3): 113–23.

543 Françoso, E, Zuntini, AR, Ricardo, PC, Silva, JPN, Brito, R, Oldroyd, BP & MC Arias. 2019.
544 Conserved Numts Mask a Highly Divergent Mitochondrial-COI Gene in a Species Complex
545 of Australian Stingless Bees Tetragonula (Hymenoptera: Apidae). Mitochondrial DNA. Part
546 A, DNA Mapping, Sequencing, and Analysis 30 (7): 806–17.

547 Franck P, Cameron E, Good G, Rasplus JY, & BP Oldroyd. 2004. Nest architecture and genetic
548 differentiation in a species complex of Australian stingless bees. Molecular Ecology. 13(8)
549 2317-31.

550 Good, JM, Vanderpool, D, Keeble, S, & K Bi. 2015. Negligible Nuclear Introgression despite
551 Complete Mitochondrial Capture between Two Species of Chipmunks. Evolution;
552 International Journal of Organic Evolution 69 (8): 1961–72.

553 Green, CL, Franck, P & BP Oldroyd. 2001. Characterization of Microsatellite Loci for Trigona
554 Carbonaria, a Stingless Bee Endemic to Australia. Molecular Ecology Notes 1 (1-2): 89–92.

555 Gurevich, A, Saveliev, V, Vyahhi, N, & G Tesler. 2013. QUAST: Quality Assessment Tool
556 for Genome Assemblies. Bioinformatics. 29 (8): 1072–75.

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

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

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

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

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

569 Katoh K, Standley DM. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
570 Improvements in Performance and Usability. Molecular Biology and Evolution. 30(4): 772-
571 780.

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
582 Construction. Edited by Shinichi Nakagawa. Methods in Ecology and Evolution / British
583 Ecological Society 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
602 Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing
603 Approach. 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 Ridley, AW, Hereward, JP, Daglish, GJ, Raghu, S, McCulloch, GA & GH Walter. 2016. Flight
607 of Rhyzopertha Dominica (Coleoptera: Bostrichidae)-a Spatio-Temporal Analysis With
608 Pheromone Trapping and Population Genetics. Journal of Economic Entomology 109 (6):
609 2561–71.

610 Sánchez-Guillén RA, Muñoz J, Rodríguez-Tapia G, Arroyo TPF, & A Córdoba-Aguilar. 2013.
611 Climate-Induced Range Shifts and Possible Hybridisation Consequences in Insects. PloS One.
612 8(11) e80531.

613 Schuelke, M. 2000. An Economic Method for the Fluorescent Labeling of PCR Fragments.
614 Nature Biotechnology 18 (2): 233–34.

615 Shipham, A, Joseph, L, Schmidt, DJ, Drew, A, Mason, I & JM Hughes. 2019. Dissection by
616 Genomic and Plumage Variation of a Geographically Complex Hybrid Zone between Two
617 Australian Non-Sister Parrot Species, *Platycercus adscitus* and *Platycercus eximius*. Heredity
618 122 (4): 402–16.

619 Toon, A, Daglish, GJ, Ridley, AW, Emery, RN, Holloway JC, & GH Walter. 2015. Random
620 Mating Between Two Widely Divergent Mitochondrial Lineages of *Cryptolestes ferrugineus*
621 (Coleoptera: Laemophloeidae): A Test of Species Limits in a Phosphine-Resistant Stored
622 Product Pest. Journal of Economic Entomology. 109(5):2221-8.

623 Twyford, AD & RA Ennos. 2011. Next-generation hybridization and introgression. Heredity
624 108: 179-189.

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

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

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

632 Zych, K, Basten, LS, Elvin, M, Rodriguez, M, Van der Velde, KJ, Arends, D, Westra, HJ *et al.*
633 2017. reGenotyper: Detecting Mislabeled Samples in Genetic Data. PloS One 12 (2):
634 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.

Figures and Tables

Table 1. Characteristics of the genomes and their draft assemblies for three species of 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

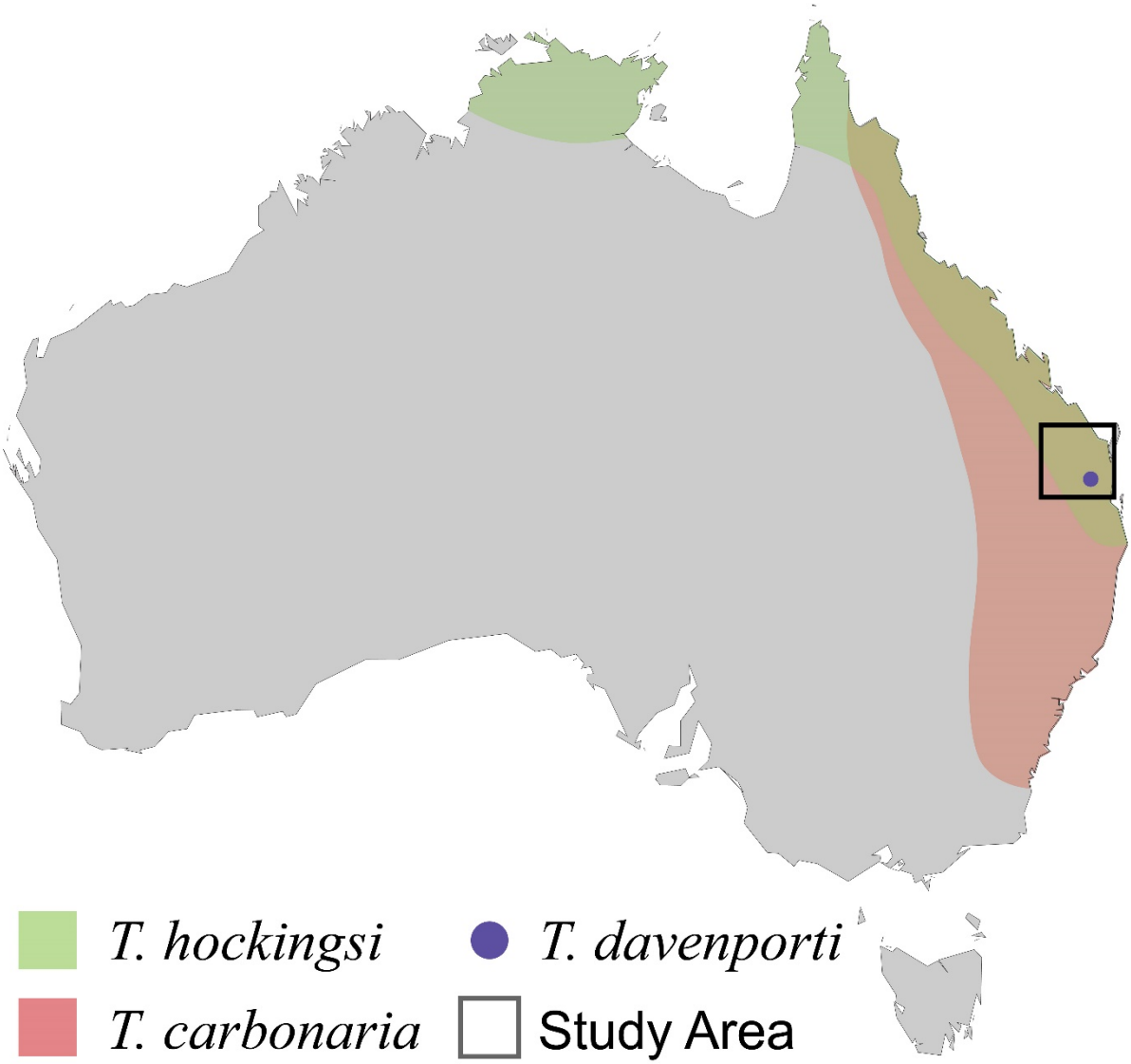

Figure 1. Map showing the known distributions of three species of *Tetragonula* in Australia, the study area is marked with a black rectangle.

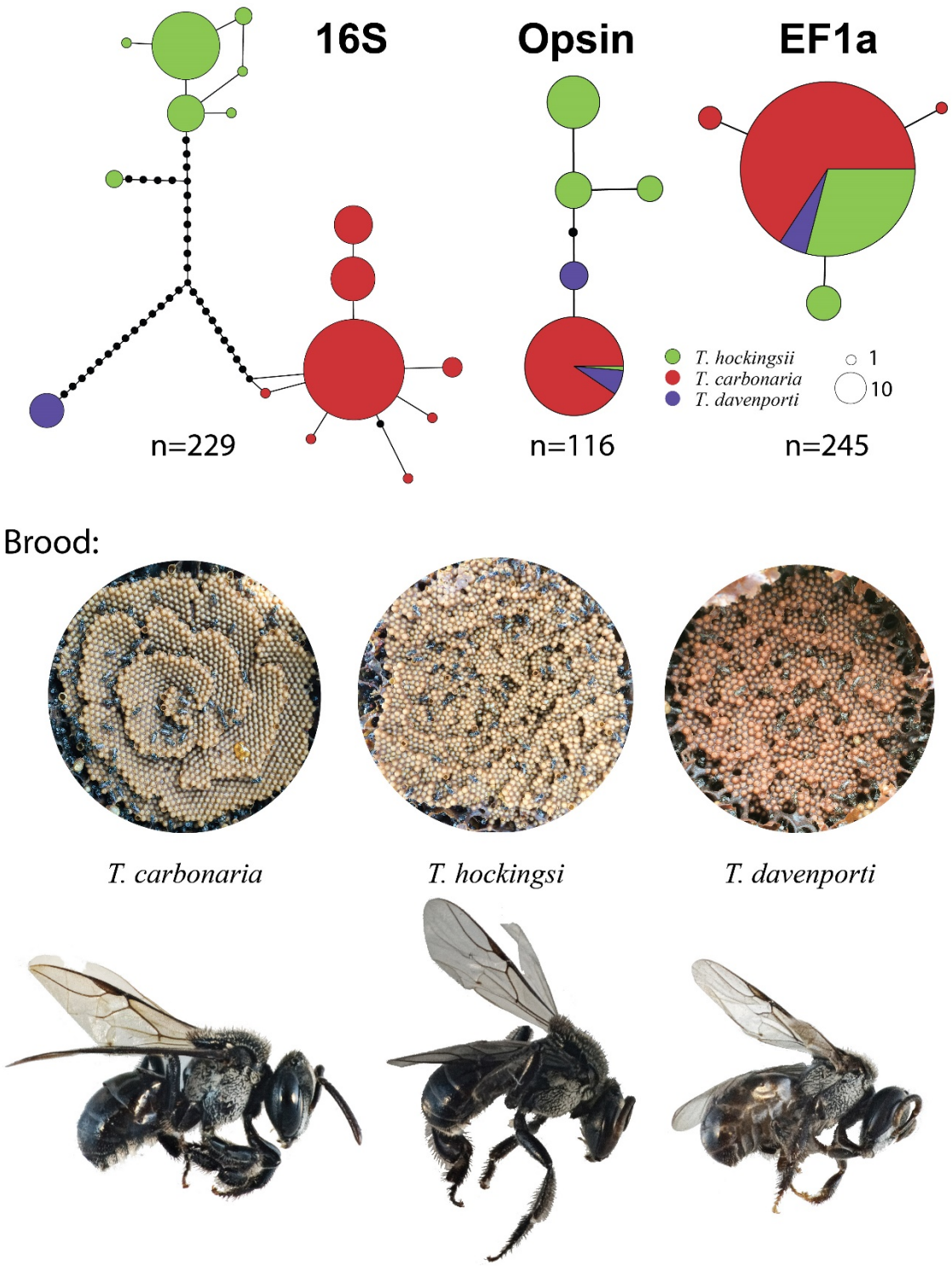

Figure 2. Haplotype networks for the mitochondrial *16S* gene, and nuclear *Opsin* and *EF1alpha* (top). Brood of the three species we investigate is illustrated (middle), as are representative bees of each species (bottom).

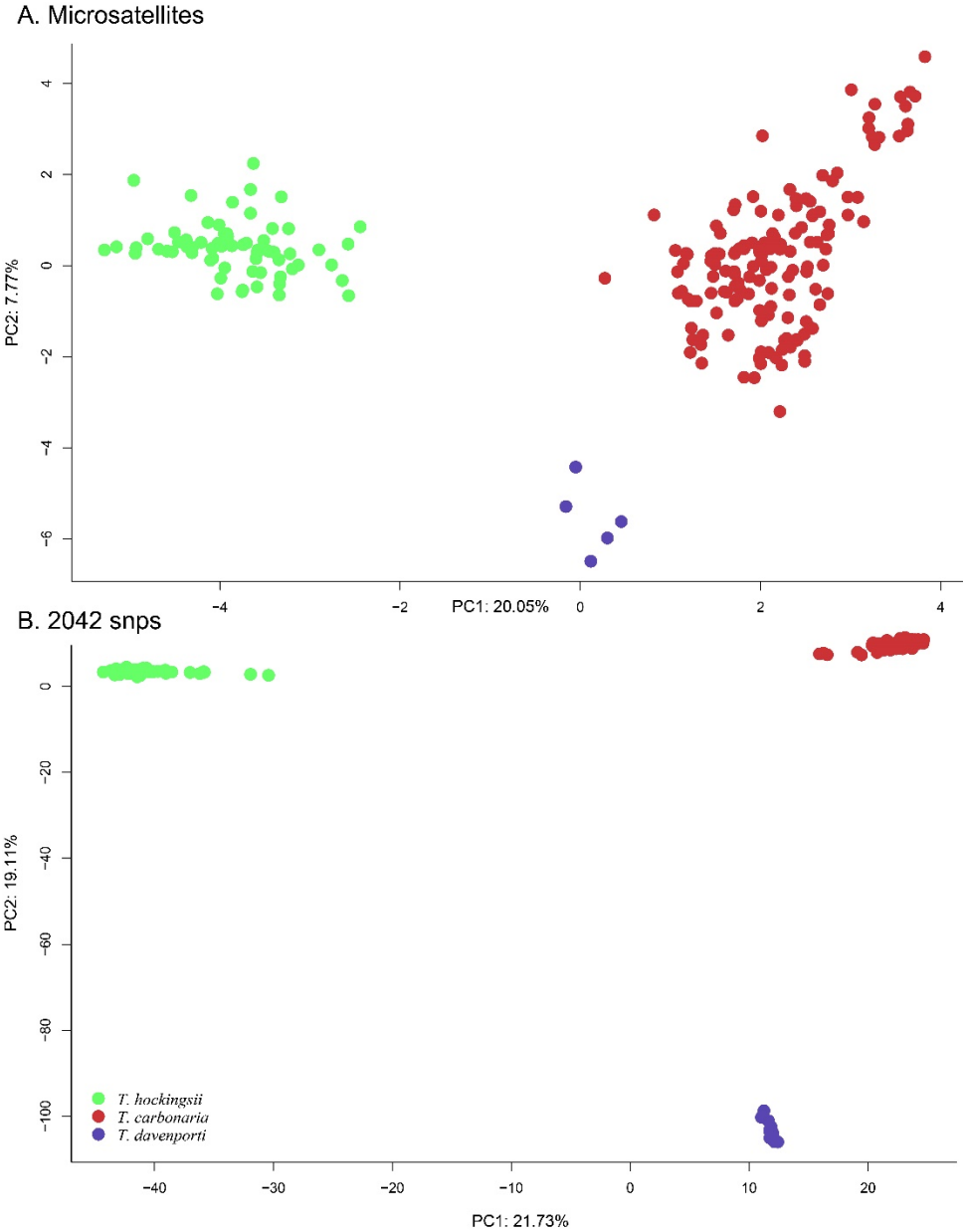

Figure 3. PCA of nuclear genetic data, 7 microsatellites (A) and 2049 *snp*s (B).

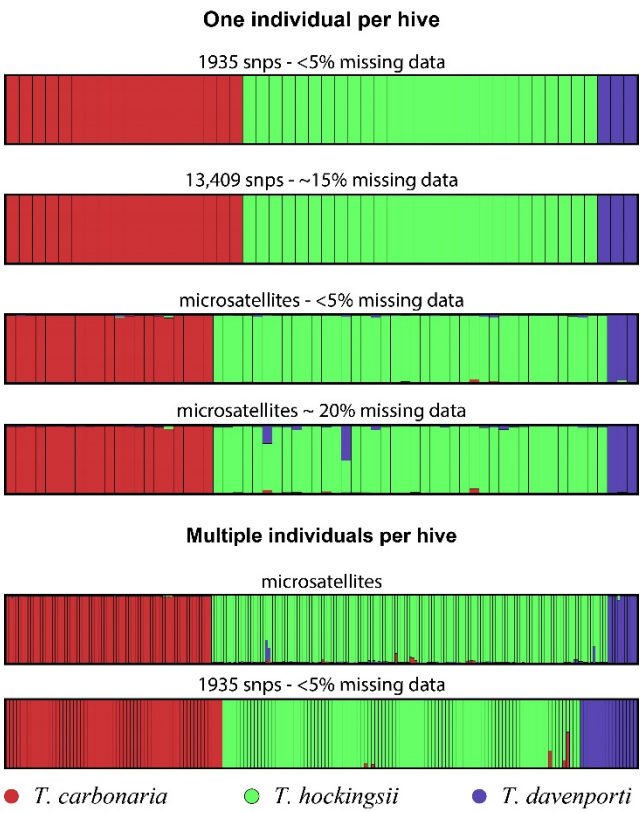

Figure 4. Structure plots for 7 microsatellite loci (with < 5% missing data and 20% missing data) and the *snp* dataset (for the 1,935 *snps* with less than 5% missing data, and the 13,409 snps that were retained when the missing data limit was set to 20%). When the missing data limit for *snps* was raised to 20% the *snp* dataset had a final missing data rate of 15%. A. The analyses were run with multiple individuals per hive (mostly four but sometimes fewer). B. The same analyses were run with only one individual bee per hive.

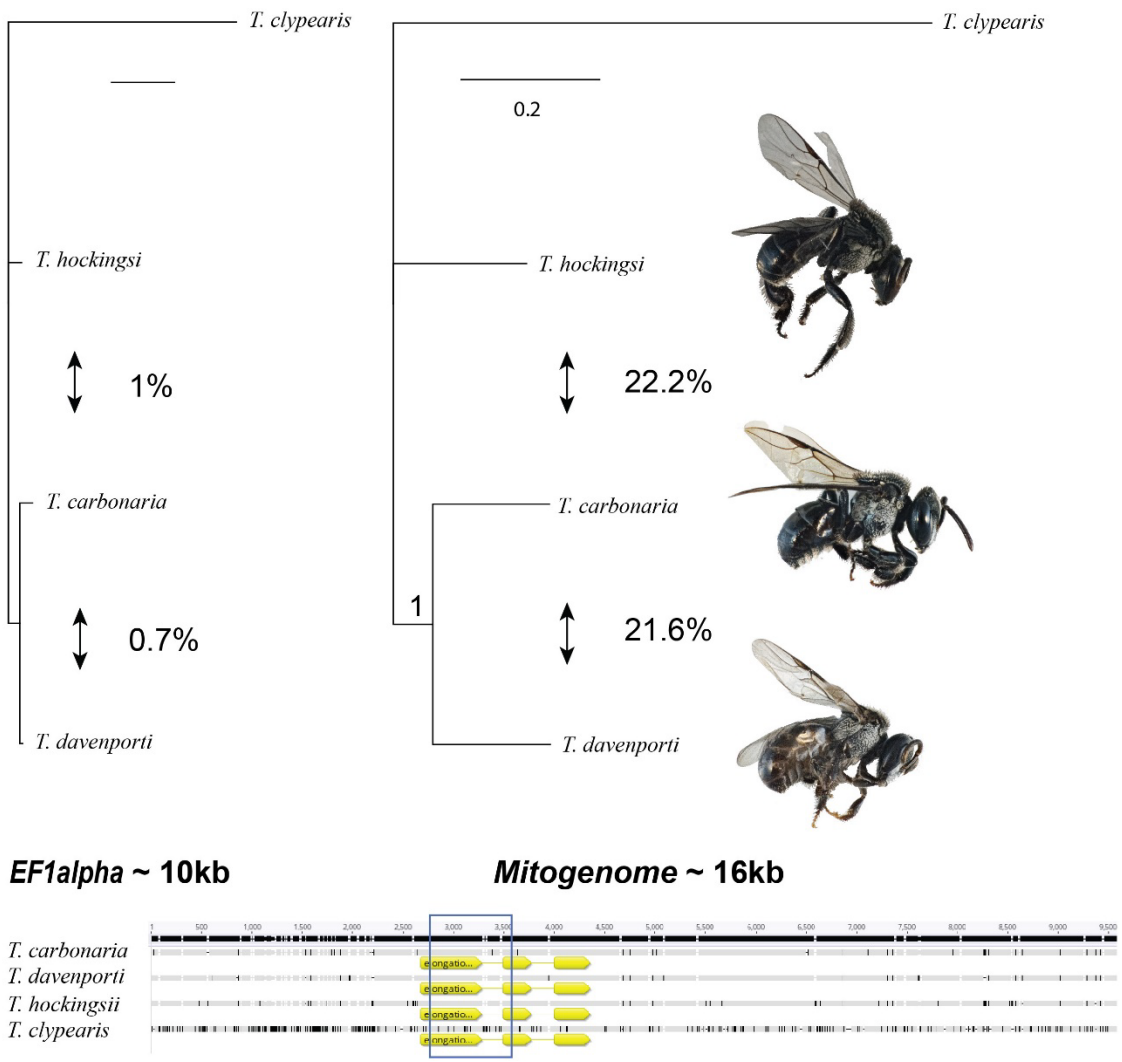

Figure 5. Bottom. Alignment of 10kb surrounding the *EF1alpha* gene, blue rectangle represents the PCR product, (Left) Bayesian tree of this region. Right. A Bayesian tree was also constructed from the mitogenome alongside which the bees are illustrated.