Biological Journal of the Linnean Society, 2021, **XX**, 1–10. With 2 figures.

Connecting social polymorphism to single nucleotide polymorphism: population genomics of the small carpenter bee, *Ceratina australensis*

BROCK A. HARPUR[1,] and SANDRA M. REHAN[2,*,]

[1]*Department of Entomology, Purdue University, West Lafayette, IN, USA*
[2]*Department of Biology, York University, Toronto, ON, Canada*

Received 30 September 2020; revised 29 December 2020; accepted for publication 8 January 2021

How do social insects expand and adapt to new ranges and how does sociality per se contribute to their success (or failure)? These questions can become tractable with the use of population genomics. We explored the population genomics of the socially polymorphic small carpenter bee, *Ceratina australensis*, across its range in eastern and southern Australia to search for evidence of selection and identify loci associated with social nesting. We sampled and sequenced fully the genomes of 54 socially and solitarily nesting *C. australensis* within Queensland, Victoria and South Australia, yielding 2 061 234 single nucleotide polymorphisms across the genome. We found strong evidence of population-specific selection and evidence of genetic variants associated with social nesting behaviour. Both the sets of associated loci and differentially expressed 'social' genes had evidence of positive selection, suggesting that alleles at genes associated with social nesting might provide different fitness benefits.

ADDITIONAL KEYWORDS: facultative sociality – population genomics – social evolution.

INTRODUCTION

Social insects are among the most successful animals on the planet; they have expanded onto and established on every continent, except for Antarctica. Despite their success, there are few examples of how social species have adapted genetically to their ranges, nor the adaptive value of sociality per se. Although sociality can be adaptive (Leadbeater *et al.*, 2011), its rarity across the animal kingdom (Wilson, 1971) suggests that there are high tolls along the road to becoming eusocial and that the benefits of being social might be context specific (Rehan *et al.*,2014; Rehan & Toth, 2015; Kennedy *et al.*, 2018).

The questions of how adaptive sociality is and what costs it might have become tractable through the careful selection of model species (Rehan & Toth, 2015) and the combined insights of behavioural ecology and population genomics (Sugg *et al.*, 1996; Harpur *et al.*, 2014, 2017). The small carpenter bee, *Ceratina australensis*, provides a useful model system to explore the adaptive value of social vs. solitary living (Rehan, 2020). *Ceratina australensis* is socially polymorphic, living either solitarily or in small social groups (Michener, 1974). It colonized Australia ≥ 18 000 years ago and since then has expanded to occupy a range that constitutes wet subtropical forests, semi-arid scrub and southern temperate coastal dunes (Fig. 1A; Dew *et al.*, 2016; Oppenheimer *et al.*, 2018).

Across this range, *C. australensis* is socially polymorphic (Rehan *et al.*, 2010). Approximately 13% of all nests in a given year will be social across the species range (Rehan *et al.*, 2010). This is surprising given that social nesters have lower lifetime reproductive success than solitary nesters (Rehan *et al.*, 2010, 2014). How then has sociality been maintained across such a wide range of habitats? It has been hypothesized that social nesters have higher fitness than solitary nesters during periods of high parasitism (Rehan *et al.*, 2010). In this case, parasites are small chalcid wasps that enter the nest while females are away foraging for provisions and lay their eggs in brood cells (Rehan *et al.*, 2014). In solitary nests, there is greater per capita brood production but also elevated parasitism rates, and some nests experience total nest failure owing to complete parasitism of all brood

*Corresponding author. E-mail: sandra.rehan@gmail.com

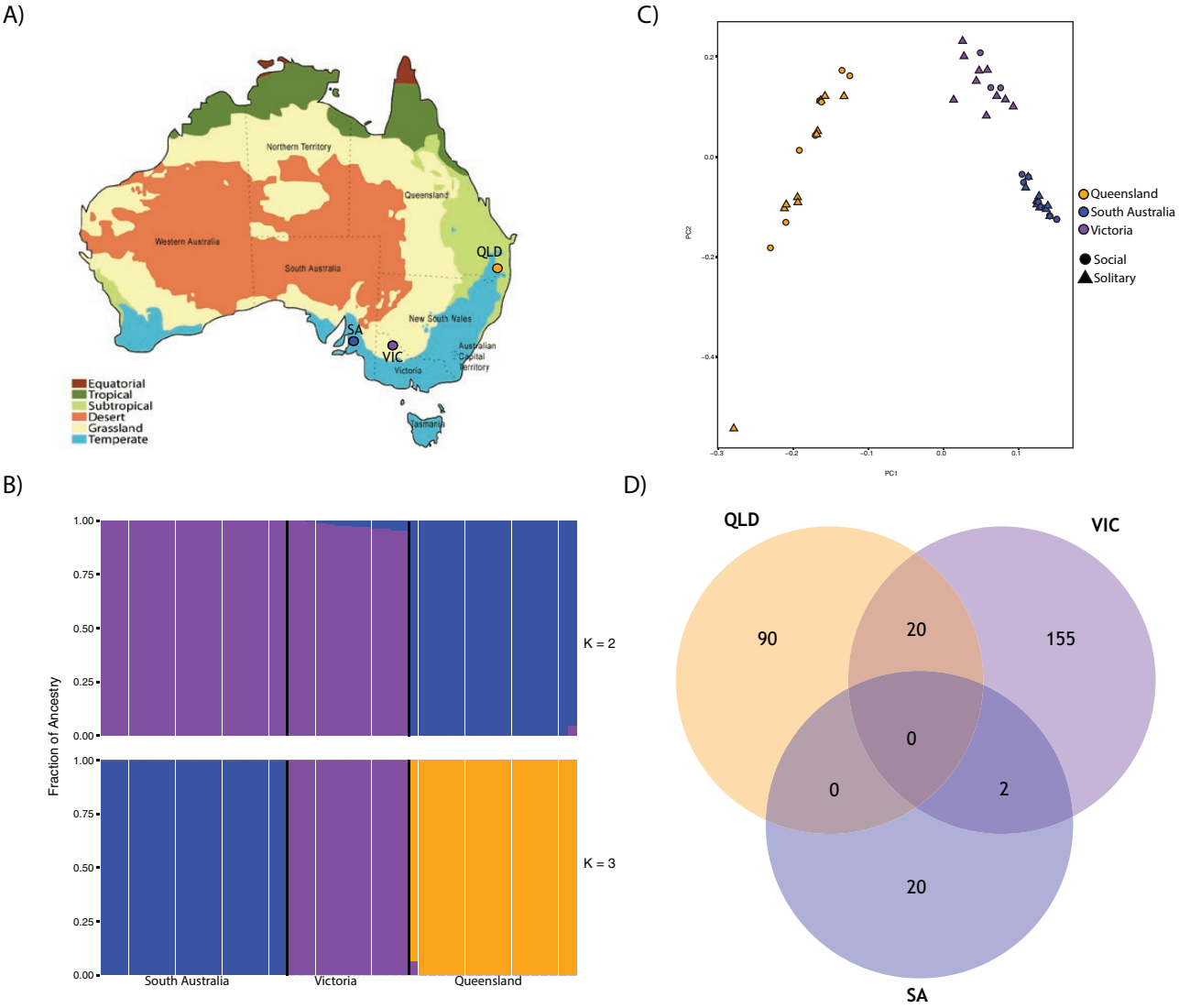

Figure 1. A, sampling locations for *Ceratina australensis* used in this study and the bioclimatic regions from which they were sampled. The map depicts the Australian Bureau of Meteorology climate classification, a modification of Köppen's classification. Data are from bom.gov.au (image credit: Martyman, CC BY-SA 3.0). B, population-level admixture among *C. australensis* populations as determined through ADMIXTURE at $K = 2$ and $K = 3$ (K = populations or genetic groups). For all three independent runs, the most likely $K = 3$ (average coefficion of variation error = 0.281). Black lines separate sampling populations. C, a principal components (PC) analysis supports the division of *C. australensis* into three distinct populations. D, Venn diagram of overlapping selection on genes within the analysis.

within the nest. Social nests have a secondary female that remains in the nest to guard while the primary female forages. Social nests are protected from total nest failure during high parasitism and thus have a fitness advantage in unpredictable parasitism rates in the environment (Rehan *et al.*, 2014). This bet-hedging strategy of maintaining both solitary and social reproductive strategies in sympatry is thought to persist via balancing selection (Kennedy *et al.*, 2018). Solitary nesting is advantageous in periods of low to moderate parasitism, but social nesting is maintained owing to the survival advantages during stochastically high parasitism experienced by populations (Rehan *et al.*, 2014). If true, this hypothesis predicts selection acting on 'social loci' (loci that contribute additive genetic variation to the expression of social nesting) in years of high parasitism and negative selection in years of low parasitism.

Here, we have taken a bottom-up, population genetic approach to understand where and how selection has acted on the genome of *C. australensis* across its Australian range. We created a large population

genomic dataset with representative samples from each of three environmental zones within Australia. Within each population, we sampled individuals from solitary and social nests. Using this dataset, and integrating previously collected transcriptomic evidence (Rehan *et al.*, 2018), we uncovered a candidate set of 'social loci', quantified how selection has acted on these loci and found sets of genes with evidence of population-specific selection.

MATERIAL AND METHODS

Sampling

Ceratina australensis nests were collected from dead, broken stems in Warwick, QLD (28.24°S, 152.09°E), Mildura, VIC (34.15°S, 142.16°E) and Adelaide, SA (34.94°S, 138.50°E) during the summer reproductive season in January 2016. Nests were split lengthwise and adult individuals flash frozen in liquid nitrogen in the field and transported to the laboratory for future DNA extraction. Nests included in this study were ≥ 5 km apart.

We also collected two female *Ceratina japonica* samples for genome sequencing to be used as a sister species for phylogenetic comparison. These samples were collected from dead broken stems of hydrangea bushes in Sapporo, Japan in July 2015.

Sequencing, variant calling, filtration and functional prediction

Whole bodies of 54 adult *C. australensis* females were used for DNA extraction. DNA was extracted using a Qiagen Genomic-tip 20/G kit (Valencia, CA, USA), following the standard protocol. Genomic DNA was then used to generate 150 bp, paired-end (PE) libraries for Illumina sequencing, performed at Genome Quebec. Fifty-four PE libraries were constructed and sequenced on Illumina HiSeq 2500, producing 328 145 Mb read data. Reads were trimmed of the adapter sequence with TRIMMOMATIC v.0.35 (Bolger *et al.*, 2014), aligned with BWA v.0.7.12 (Li & Durbin, 2009) to the *C. australensis* genome (Rehan *et al.*, 2018) and re-aligned around indels with GATK v.3.5 (McKenna *et al.*, 2010), with duplicate reads removed by PICARD v.1.123 MarkDuplicates. Single nucleotide polymorphisms (SNPs) were detected and genotyped using the GATK HaplotypeCaller pipeline.

Our samples had an average of 12× coverage across the entire genome (range of 5×–17×). After SNP calling and genotyping with GATK, we removed all sites flagged as 'LowQual' and obtained 4 075 576 predicted sites. We applied several manual filters to identify and remove low-quality sites and genotype calls. Sites that deviate significantly from Hardy–Weinberg equilibrium can be indicative of a poor genotype prediction owing to low depth or complexity of the sequence (Moskvina *et al.*, 2006). In addition, sites in highly repetitive regions of the genome (and difficult to genotype) can have high sequence variability among samples (Magi *et al.*, 2013). In our dataset, the 29 275 sites that deviated significantly from the expectations of Hardy–Weinberg equilibrium ($P < 1 \times 10^{-9}$) had significantly greater depth and greater variation in depth among individuals than the rest of the genome ($P < 0.000001$); therefore, we flagged them for removal. Given that these sites had elevated variation in depth across individuals, we flagged all 275 202 sites that had high variation in depth among individuals (> 90% of site–depth variability across the genome; SD > 35) for downstream removal, as an added precaution. We removed 113 782 sites with very high variation (> 95% of our data; 23 reads) and 185 716 sites from the 9788 scaffolds with < 500 variable sites because, upon visual inspection, alignment appeared poor in these gene-sparse scaffolds. We removed all of these sites along with indels and sites that were not biallelic using VCFtools (Danecek *et al.*, 2011) and included filters to remove low-frequency alleles (--maf 0.01) and genotype calls with low average depth and with fewer than five reads (--min-meanDP 5 --minDP 5). We did not allow genotyping calls at sites with genotyping quality < 20 (--minGQ 20), and we did not allow sites with > 10 individuals missing a genotyping call. Our final dataset of 2 061 234 sites had an average of 14.7 reads per site and contained no sites with more than two individuals missing coverage. We used SNPEFF v.3.1 (Cingolani *et al.*, 2012) to identify predicted amino acid-changing mutations (non-synonymous mutations) across all protein-coding genes within the genome. All sequencing data generated for this study can be accessed via NCBI SRA (BioProject: PRJNA407923).

Relatedness, population genetic metrics and population structure

We ran ADMIXTURE v.1.22 using an unsupervised model for $K = 1$–5 (K = populations or genetic groups) three times with 10 000 randomly selected SNPs and 20 cross-validation steps. For all three independent runs, the most likely $K = 3$ [average coefficient of variation (CV) error = 0.28]. We performed principal components analysis using EIGENSOFT (Price *et al.*, 2006). For all the following analyses, unless otherwise stated, we treated populations individually. We estimated relatedness across the genome, within populations (ϕ; the probability of two alleles being identical by descent), using the VCFtools --relatedness2 option (Manichaikul *et al.*, 2010).

We estimated the average differentiation among populations using F_{ST}, the pairwise fixation index (Weir & Cockerham, 1984), as executed in VCFTOOLS (*--weir-fst-pop*). We also used VCFTOOLS to estimate Tajima's *D* (Tajima, 1989) in 1000 bp windows across the genome. To search for evidence of local adaptation within each of the three populations, we used a haplotype-based approach: the integrated haplotype score (iHS; Voight *et al.*, 2006). The iHS measures the amount of extended haplotype homozygosity along ancestral and derived alleles. Alleles that arise and fix owing to selection will have extended haplotype homozygosity. The iHS is a standardized measure across the genome and provides a means to identify unusually large haplotypes around a given ancestral or derived SNP. We estimated iHS and its significance using the R package REHH (Gautier *et al.*, 2016). The ancestral allele was identified by integrating genomic data from *C. japonica* and assuming that the allele with the highest frequency within this population is the ancestral allele. Phasing was performed with SHAPEIT v.2.12 (Delaneau *et al.*, 2014) for all sites with an allele frequency > 0.05 and which were independently segregating within a 50 bp window (plink --indep 50 5 2). A reference panel is not available for *C. australensis*, but phasing in this way is regularly used for non-model bee genomes (Chen *et al.*, 2018) and should not be biased by relatedness (Delaneau *et al.*, 2014). We used phased data only to estimate iHS. We estimated decay of linkage disequilibrium using POPLDDECAY (Zhang *et al.*, 2019).

LOCI ASSOCIATED WITH SOCIALITY

We estimated aspects of the genetic basis of social nesting (as a binary trait: social vs. solitary nesting) through the use of a multilocus Bayesian sparse linear mixed model (BSLMM) and a standard linear mixed model (LMM) using GEMMA v.0.98.1 (Zhou & Stephens, 2012) for all sites with an allele frequency > 0.1. GEMMA includes a kinship matrix as a random effect in order to control for linkage disequilibrium (LD) and relatedness among individuals. It also allows for estimation of the amount of total phenotypic variation explained by all SNPs, the number of SNPs (*N*) where the genotypic effect on the phenotype (β) is non-zero, and the effect size of all significantly associated SNPs as βγ. The BSLMM model was run with 5 000 000 burn-in iterations followed by 50 000 000 sampling iterations, recording every 100 iterations (*-bslmm 1 -km 2 -w 5 000 000 -s 50 000 000 -rpace 100 -wpace 100*). We ran an LMM with GEMMA that included all samples across all sites. We repeated this same analysis for samples only from Queensland. We found significant correlation in the associated *P*-values for both models ($r = 0.58$; $P < 0000001$). We assumed all associations with $P < 0.001$ to be significant.

GENE ONTOLOGY

We determined the putative functional role of locally adapted genes (any with iHS $P < 0.05$) and genes associated with social nesting using the R package topGO (Alexa & Rahnenfuhrer, 2010) using Gene Ontology (GO) for the most recent build of the *C. australensis* genome (Rehan *et al.*, 2018). We examined enrichment of all ontology terms (cell component 'CC', molecular function 'MF' and biological process 'BP') using the default parameters and Fisher's exact test. We assumed all associations with $P < 0.05$ to be significant.

RESULTS

VARIANT IDENTIFICATION, POPULATION GENETIC PARAMETERS AND STRUCTURE

Our analysis revealed distinct populations of *C. australensis* across Australia (Fig. 1; Supporting Information, Fig. S1). We estimated F_{ST} at all 2 061 234 sites across the genome among populations and between social and solitary groups within and between each population (Table 1). We identified mutations within 19 269 protein-coding gene regions and identified 170 028 predicted amino acid-changing (non-synonymous) mutations in the genome and 307 056 synonymous mutations. Mean F_{ST} was highest between Queensland and South Australia ($F_{ST} = 0.083$) and lowest between South Australia and Victoria ($F_{ST} = 0.045$). Nucleotide diversity was highest within Queensland ($\pi = 0.11$) and lowest in South Australia ($\pi = 0.052$; Table 1). We found significantly lower relatedness among samples within Queensland ($\phi = 0.094 \pm 0.0045$ SE) and Victoria ($\phi = 0.059 \pm 0.0016$ SE) relative to South Australia ($\phi = 0.12 \pm 0.00091$ SE; ANOVA $F_{2415} = 76$, $P < 0.00001$; Tukey's HSD for all comparisons < 0.05). Additionally, we found evidence of inbreeding within both Victoria ($F_{is} = 0.17 \pm 0.24$ SE) and South Australia ($F_{is} = 0.46 \pm 0.007$ SE), but not within Queensland ($F_{is} = -0.079 \pm 0.018$ SE). Differences in F_{is} among populations were all significant (pairwise Wilcoxon rank test;

Table 1. Pairwise fixation index (mean F_{ST}) among *Ceratina australensis* populations, with nucleotide diversity (π) within populations on the diagonal

	Queensland	Victoria	South Australia
Queensland	π = 0.11		
Victoria	0.054	π = 0.085	
South Australia	0.082	0.045	π = 0.052

$P < 1 \times 10^{-7}$). Finally, we found that LD decayed very rapidly in all three populations (Supporting Information, Fig. S1), which is likely to reflect the high recombination rates observed in bees more generally (Wilfert *et al.*, 2007). We found the highest levels of LD within South Australia and the lowest in Queensland (Supporting Information, Fig. S1).

We ran ADMIXTURE for $K = 1$–5 on three randomly drawn SNP datasets. For all three independent runs, the most likely $K = 3$ (average CV error = 0.281; Fig. 1B). As K decreased to $K = 2$, Queensland remained distinct from Victoria and South Australia. We never observed social populations segregating from solitary populations (Fig. 1B; Supporting Information, Fig. S1). A principal components analysis-based approach supported the division of *C. australensis* into three distinct populations and the close relationship between the two southernmost populations relative to Queensland (Fig. 1C).

Evidence of population-specific natural selection

Each sampled population resides in a distinct microclimatic region: Queensland being subtropical, Victoria grassland and South Australia temperate (Fig. 1). Each population might therefore harbour distinct adaptations. To detect putative genes that have contributed to divergence between *C. australensis* populations, we used iHS across the genome within each population individually and extracted significant outlier iHS regions as those putatively acted on by selection (see Material and Methods). We found that the most highly significant iHS regions of the genome ($P < 1 \times 10^{-3}$) also had lower Tajima's D than all non-significant regions of the genome (mean Tajima's D −1.7 vs. −1.3; $P = 4.4 \times 10^{-14}$), indicative of positive selection. In total, we found 287 sites across 140 scaffolds with evidence of positive selection (iHS $P < 1 \times 10^{-3}$; Supporting Information, Table S1) and an average of 96 within each population.

There was evidence of population-specific enrichment for GO terms within each population, largely for protein binding (Supporting Information, Table S2). Finally, positively selected variants were distributed evenly among functional classes (e.g. non-synonymous, introns; Fisher's exact test, $P > 0.1$ for all comparisons; Supporting Information, Table S1).

Variants associated with nesting conditions are plentiful and of weak effect

Several lines of evidence suggest that genetic variation contributes to the expression of social nesting within *C. australensis*. First, using a genome-wide association approach, we identified 350 sites significantly associated with social nesting (Fig. 2). Second, we found that these SNPs had higher F_{ST} (mean = 0.21) between solitary and social nesters in Queensland than all SNPs across the genome (mean = 0.029; $P < 0.00001$). Finally, we found that 31% of the phenotypic variation was explained by all SNPs, which is significantly more variation than a permuted-phenotype dataset (consisting of the same genotypic data but with phenotypes exchanged among samples, randomly; $F_{1,675\,098} = 24\,207$; $P < 2 \times 10^{-16}$).

Genes within nesting strategy-associated regions have distinct function and evidence of selection

The 350 variants associated with nesting strategy overlapped with 281 protein-coding genes (Supporting Information, Table S3). We found no evidence of enrichment within any particular functional mutational class (e.g. non-synonymous; Fisher's exact test $P > 0.1$). We did, however, find that associated genes overlapped with those found to be expressed differentially between naturally occurring social and solitary nesting individuals (Rehan *et al.*, 2018). This was significantly more than expected by chance (Fisher's exact test $P = 0.026$). Ten of the candidates were found to be upregulated in social females, and one was found to be upregulated within solitary nesting females (Supporting Information, Table S3). One candidate, Caust.v2_004417, a core promoter and histone-binding protein (NSL1), was significantly associated with social nesting in our study and was also found to be upregulated in social females relative to solitary females (Rehan *et al.*, 2018). NSL1 is differentially expressed between socially founding and solitarily founding queen ants (Manfredini *et al.*, 2013). As a set, the social-associated genes were significantly enriched for 26 GO terms (Supporting Information, Table S4). Previous investigations discovered 337 significantly enriched GO terms for genes expressed differentially between and among social and solitary behavioural classes in *C. australensis* (Rehan *et al.*, 2018).

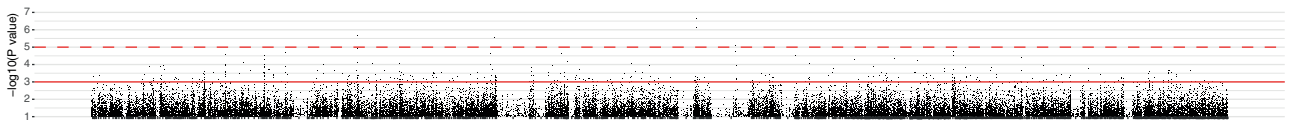

Figure 2. Manhattan plot showing genome-wide association analyses identifying significant SNP sites associated with social nesting behaviour.

Of the 26 significant GO terms we discovered in our study, six (GO:0004252, GO:0005201, GO:0005509, GO:0005581, GO:0005992 and GO:0008021) overlapped with the 337 found previously (Supporting Information, Table S4). These terms have been associated with social behaviour across many studies (Supporting Information, Table S4).

We found some evidence that alleles within associated genes might be acted on by positive selection. Of the set of 281 genes we identified tentatively to be associated with social nesting, 18 also had significant evidence of positive selection in at least one population (Supporting Information, Tables S1 and S3). Additionally, a single gene (Caust.v2_003367), which is expressed differentially between social and solitary nesters (Rehan *et al.*, 2018), shows evidence of positive selection (Supporting Information, Table S1).

DISCUSSION

Our bottom-up, population genomic approach allowed us to examine patterns of genetic diversity, identify putatively selected loci and identify a tentative list of loci contributing variation to the expression of social nesting within populations of *C. australensis*. We have provided gene lists for future functional analyses to target and explore both the fitness effects of candidates and their role in social behaviour. The latter list is a particularly important contribution because it represents the first genome-wide association for nesting behaviour in small carpenter bees. Most importantly, we have demonstrated that by careful selection of a model species, integrating natural history and novel genome-wide approaches, we can begin to disentangle the role of genetics and ecology in the evolution of facultative sociality.

Population structure, diversity and natural selection in *C. australensis*

Here, we study three genetically distinct populations of *C. australensis* within Australia. Populations currently present in Victoria and South Australia are likely to be younger and/or have experienced greater demographic changes than in Queensland (Fig. 1). Our data, along with those of two previous studies using mitochondria- (Dew *et al.*, 2016) and microsatellite-based methods (Oppenheimer *et al.*, 2018), strongly support a migration of *C. australensis* out of Queensland and into Victoria and South Australia. The three populations are likely to represent a large, core population in Queensland and two marginal populations in South Australia and Victoria. As expected of marginal populations, both Victoria and South Australia have reduced genetic diversity and elevated population differentiation relative to their source population (Le Corre & Kremer, 1998; Eckert *et al.*, 2008).

The marginal populations in Victoria and South Australia also suffer from reduced genetic diversity and inbreeding, which might reduce their likelihood of persistence (Zayed & Packer, 2005). The effects of population expansion might have been exacerbated further in South Australia by a recent bottleneck. Previous studies have linked this bottleneck to habitat fragmentation (Oppenheimer *et al.*, 2018). Regardless of the cause, evidence of inbreeding within the southern margins of *C. australensis* (the first documented evidence to date) might have important implications for the persistence of the species. Inbreeding can be particularly harmful to bee populations owing to their mechanism of sex determination (Van Wilgenburg *et al.*, 2006; Harpur *et al.*, 2013). In many bees, sex is determined by the genotype at autosomal sex loci (a system call the complementary sex determination pathway), with eggs heterozygous at sex loci becoming female and hemizygous eggs developing into males (Whiting, 1933, 1943). When a diploid egg is homozygous at sex loci, a diploid male develops. Across many species diploid males are sterile, and their presence increases the probability of a population becoming locally extirpated or extinct without an influx of genetic diversity (Zayed & Packer, 2005; Harpur *et al.*, 2013). In large populations, the number of alleles at complementary sex determination loci should be sufficiently large to make it unlikely to produce homozygosity (Yokoyama & Nei, 1979); however, small bee populations can be at risk of both inbreeding and diploid male production (Zayed & Packer, 2005; Van Wilgenburg *et al.*, 2006). Recent observations of diploid males in South Australia (Oppenheimer & Rehan, 2020), along with our genome-wide evidence of inbreeding (above) and annotated versions of complementary sex determiner (csd) and feminizer in the *C. australensis* genome (Rehan *et al.*, 2018), suggest that this population might be at risk of extirpation without an influx of new genetic diversity. Future laboratory-based breeding trials in this species would help to elucidate the risk of diploid male production in this group and the number and identity of sex-determining loci (Van Wilgenburg *et al.*, 2006; Harpur *et al.*, 2013).

Variants associated with nesting condition

The three populations of *C. australensis* each harbour social and solitary nests; ~13% of nests in a given year will be social (Rehan *et al.*, 2010). We have uncovered the first evidence that social nesting behaviour in *C. australensis* might be associated with genetic variation. There was no single large-effect locus that could provide the sole explanation for the variation

in social nesting. Although social nesting might be a heritable trait, its expression is influenced by several hundred loci across the genome, each with a very small effect.

We note that the genome association we performed, as with all genome scans, is highly tentative. Although we used some of the best available techniques to control for relatedness among samples and population structure, our sample size is smaller than that of a typical genome-wide association study (Nishino *et al.*, 2018; Pardiñas *et al.*, 2018). Unfortunately, it is biologically unrealistic to obtain sample sizes even an order of magnitude larger for *Ceratina* because their cavities can be difficult to find and are often limited to one or a few per site (Rehan *et al.*, 2010). We believe that our data present a good starting point for further genetic analysis. The genes we identified overlap with previously reported genes expressed differentially in this species (Rehan *et al.*, 2018; Steffen & Rehan, 2020) and they overlap with GO terms associated with social behaviours in several other species (Supporting Information, Table S4). Future functional genomic work along with controlled crosses will be able to test the hypothesis that these genes are indeed associated with social nesting.

Although it is common to score social behaviour as binary (social vs. solitary), it is crucially important to recognize that sociality is a complex trait. In *C. australensis*, social colonies contain a primary female, who forages and reproduces, and a secondary female, who remains within the nest and delays reproduction until the following season (Rehan *et al.*, 2010). Unlike other social insect species, there are no differences in size between primary and secondary nest mates (Sakagami & Maeta, 1984, 1987; Maeta, 1995; Rehan *et al.*, 2009, 2015). However, the social secondary female does not forage and has smaller ovaries than her social primary (Rehan *et al.*, 2010). It has been suggested that a reproductive primary might only 'tolerate' a non-reproductive, non-foraging nest mate (Rehan *et al.*, 2010) and that a secondary female might have different dispersal abilities from a solitary female or primary (Rehan *et al.*, 2014). The genetic influence on social nesting could therefore be multifactorial. In the social secondary, loci might contribute to variation in ovarian development, dispersal or general flight ability. In the primary, loci might contribute to variation in the tolerance of accepting a social secondary. It therefore seems unsurprising in terms of both quantitative genetic expectations and our understanding of the natural history of this species that many loci exhibit genetic variation associated with social nesting behaviour. Furthermore, in at least one other socially polymorphic bee species the nesting strategy seems to be underpinned by many loci with weak effects (Kocher *et al.*, 2018). At least in bees (for ants, see Wang *et al.*, 2013; Purcell *et al.*, 2014), social nesting conforms to the fourth law of behavioural genetics: for a particular behavioural trait, there will be very many genetic variants, each of which accounts for a very small proportion of the heritable variation (Valdar *et al.*, 2006; Park *et al.*, 2011; Savolainen *et al.*, 2013; Chabris *et al.*, 2015).

Our candidates provide a useful list for further functional validation. Of the numerous genes associated with social nesting, few stand out as candidates for further functional validation. Caust. v2_004417 (NSL1) is associated with social nesting in our study and is upregulated in social females (Rehan *et al.*, 2018) and ant queens that found colonies socially (Manfredini *et al.*, 2013). There has been no functional analysis of NSL1 within bees, but within *Drosophila*, NSL1 is required for germline cell maintenance and oogenesis (Yu *et al.*, 2010). Common variants in this and other candidate genes might act to increase the likelihood of a female being of reduced reproductive quality in her first year and push her into a non-foraging, non-reproducing social secondary role. The action of those variants is yet to be understood but could occur through apoptosis in ovarian tissue of a social secondary. Apoptosis of ovarian tissue in honeybee workers is crucial to regulating their non-reproductive status (Ronai *et al.*, 2016). How exactly that occurs mechanistically will be an exciting avenue for future research.

Our work provides further useful insights into the evolution of social nesting in *C. australensis* and a novel hurdle in the evolution of eusociality in general. Across multiple years, *C. australensis* has been observed to nest socially in 13% of nests (Rehan *et al.*, 2010). Social nesting individuals have lower fitness than their solitary nesting conspecifics (Rehan *et al.*, 2010, 2014). We found tentative evidence of positive selection at loci associated with social nesting, supporting previous bet-hedging hypotheses regarding the adaptive mechanisms through which social nesting is maintained in *C. australensis* (Rehan *et al.*, 2010, 2014). Why then does social nesting persist across these populations? The genetic architecture might contribute to the persistence of the social polymorphism. The strength of selection acting on a QTL allele is proportional to its effect size (Lande, 1976). Loci with a small effect are unlikely to reach fixation unless acted on by strong positive selection for many generations. Selection against (or for) social nesting would therefore be inefficient at fixing the phenotype. This effect, coupled with the possibility of solitary nesters proliferating during years of low parasitism, suggests that social nesting is very unlikely to fix in *C. australensis*.

Conclusions

Social species have colonized every continent successfully. They are argued to be among the most successful organisms on the planet (Wilson, 1971). Despite this, we know little about the genetics of social behaviour nor how 'social genes' contribute to the adaptive divergence of lineages. Here, we used the socially polymorphic small carpenter bee, *C. australensis*, to carry out a tentative exploration of the genetic architecture of social behaviour and quantify adaptive selection on those genes contributing variation to the expression of social behaviour. We put forward that in this species, social behaviour is likely to be controlled by many loci of very small effect. If incipient sociality is generally controlled by many loci of small effect (Kocher *et al.*, 2018), this might provide another impasse to the evolution of obligate sociality in general, because only very strong selection events can fix alleles in these cases. In the case of *C. australensis*, empirical evidence and mathematical models suggest that social nesting is maintained weakly during years of high parasitism and difficult to purge during years of low or no parasitism (Rehan *et al.*, 2014; Kennedy *et al.*, 2018). This raises an intriguing question about the evolution of social behaviour: how much (or how little) does 'social genetic architecture' facilitate the evolution of social behaviour? This question will benefit greatly from the continued exploration of socially polymorphic species and their social and solitary relatives.

ACKNOWLEDGEMENTS

We thank Sarah Lawson, Sean Lombard and Wyatt Shell for assistance with field collections. This research was supported by National Geographic (9659-15) and National Science and Engineering Research Council Discovery grants to S.M.R.

REFERENCES

Alexa A, Rahnenfuhrer J. 2010. *topGO: enrichment analysis for gene ontology. R package version 2: 2010. R package version 2.42.0*. Available at: https://bioconductor.org/packages/release/bioc/html/topGO.html

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120.

Chabris CF, Lee JJ, Cesarini D, Benjamin DJ, Laibson DI. 2015. The fourth law of behavior genetics. *Current Directions in Psychological Science* **24:** 304–312.

Chen C, Wang H, Liu Z, Chen X, Tang J, Meng F, Shi W. 2018. Population genomics provide insights into the evolution and adaptation of the eastern honey bee (*Apis cerana*). *Molecular Biology and Evolution* **35:** 2260–2271.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3. *Fly* **6:** 80–92.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* **27:** 2156–8.

Delaneau O, Marchini J, The 1000 Genomes Project Consortium. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications* **5:** 3934.

Dew RM, Rehan SM, Schwarz MP. 2016. Biogeography and demography of an Australian native bee *Ceratina australensis* (Hymenoptera, Apidae) since the last glacial maximum. *Journal of Hymenoptera Research* **49:** 25–41.

Eckert CG, Samis KE, Lougheed SC. 2008. Genetic variation across species' geographical ranges: the central–marginal hypothesis and beyond. *Molecular Ecology* **17:** 1170–1188.

Gautier M, Klassmann A, Vitalis R. 2016. REHH 2.0: a reimplementation of the R package REHH to detect positive selection from haplotype structure. *Molecular Ecology Resources* **17:** 78–90.

Harpur BA, Dey A, Albert JR, Patel S, Hines HM, Hasselmann M, Packer L, Zayed A. 2017. Queens and workers contribute differently to adaptive evolution in bumble bees and honey bees. *Genome Biology and Evolution* **9:** 2395–2402.

Harpur BA, Kent CF, Molodtsova D, Lebon JM, Alqarni AS, Owayss AA, Zayed A. 2014. Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences of the United States of America* **111:** 2614–2619.

Harpur BA, Sobhani M, Zayed A. 2013. A review of the consequences of complementary sex determination and diploid male production on mating failures in the Hymenoptera. *Entomologia Experimentalis et Applicata* **146:** 156–164.

Kennedy P, Higginson AD, Radford AN, Sumner S. 2018. Altruism in a volatile world. *Nature* **555:** 359–362.

Kocher SD, Mallarino R, Rubin BER, Yu DW, Hoekstra HE, Pierce NE. 2018. The genetic basis of a social polymorphism in halictid bees. *Nature Communications* **9:** 4338.

Lande R. 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution; international journal of organic evolution* **30:** 314–334.

Le Corre V, Kremer A. 1998. Cumulative effects of founding events during colonisation on genetic diversity and differentiation in an island and stepping-stone model. *Journal of Evolutionary Biology* **11:** 495–512.

Leadbeater E, Carruthers JM, Green JP, Rosser NS, Field J. 2011. Nest inheritance is the missing source of

direct fitness in a primitively eusocial insect. *Science* **333:** 874–876.

Li H, **Durbin R. 2009.** Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Maeta Y. 1995. Task allocation in artificially induced colonies of a basically solitary bee *Ceratina (Ceratinidia) okinawana*, with a comparison of sociality between *Ceratina* and *Xylocopa*: Hymenoptera, Anthophoridae, Xylocopinae. *Japanese Journal of Ecology* **63:** 115–150.

Magi A, **Tattini L**, **Cifola I**, **D'Aurizio R**, **Benelli M**, **Mangano E**, **Battaglia C**, **Bonora E**, **Kurg A**, **Seri M**, **Magini P**, **Giusti B**, **Romeo G**, **Pippucci T**, **De Bellis G**, **Abbate R**, **Gensini GF. 2013.** EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biology* **14:** R120.

Manfredini F, **Riba-Grognuz O**, **Wurm Y**, **Keller L**, **Shoemaker D**, **Grozinger CM. 2013.** Sociogenomics of cooperation and conflict during colony founding in the fire ant *Solenopsis invicta*. *PLoS Genetics* **9:** e1003633.

Manichaikul A, **Mychaleckyj JC**, **Rich SS**, **Daly K**, **Sale M**, **Chen WM. 2010.** Robust relationship inference in genome-wide association studies. *Bioinformatics* **26:** 2867–2873.

McKenna A, **Hanna M**, **Banks E**, **Sivachenko A**, **Cibulskis K**, **Kernytsky A**, **Garimella K**, **Altshuler D**, **Gabriel S**, **Daly M**, **DePristo MA. 2010.** The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20:** 1297–303.

Michener CD. 1974. *The social behavior of the bees; a comparative study*. Cambridge: Belknap Press of Harvard University Press.

Moskvina V, **Craddock N**, **Holmans P**, **Owen MJ**, **O'Donovan MC. 2006.** Effects of differential genotyping error rate on the type I error probability of case-control studies. *Human Heredity* **61:** 55–64.

Nishino J, **Ochi H**, **Kochi Y**, **Tsunoda T**, **Matsui S. 2018.** Sample size for successful genome-wide association study of major depressive disorder. *Frontiers in Genetics* **9:** 227.

Oppenheimer RL, **Rehan SM. 2020.** Inclusive fitness of male and facultatively social female nesting behavior in the socially polymorphic bee, *Ceratina australensis*. *Annals of the Entomological Society of America*. In press. doi: 10.1093/aesa/saaa022.

Oppenheimer RL, **Shell WA**, **Rehan SM. 2018.** Phylogeography and population genetics of the Australian small carpenter bee, *Ceratina australensis*. *Biological Journal of the Linnean Society* **124:** 747–755.

Pardiñas AF, **Holmans P**, **Pocklington AJ**, **Escott-Price V**, **Ripke S**, **Carrera N**, **Legge SE**, **Bishop S**, **Cameron D**, **Hamshere ML**, **Han J**, **Hubbard L**, **Lynham A**, **Mantripragada K**, **Rees E**, **MacCabe JH**, **McCarroll SA**, **Baune BT**, **Breen G**, **Byrne EM**, **Dannlowski U**, **Eley TC**, **Hayward C**, **Martin NG**, **McIntosh AM**, **Plomin R**, **Porteous DJ**, **Wray NR**, **Caballero A**, **Geschwind DH**, **Huckins LM**, **Ruderfer DM**, **Santiago E**, **Sklar P**, **Stahl EA**, **Won H**, **Agerbo E**, **Als TD**, **Andreassen OA**, **Bækvad-Hansen M**, **Mortensen PB**, **Pedersen CB**, **Børglum AD**, **Bybjerg-Grauholm J**, **Djurovic S**, **Durmishi N**, **Pedersen MG**, **Golimbet V**, **Grove J**, **Hougaard DM**, **Mattheisen M**, **Molden E**, **Mors O**, **Nordentoft M**, **Pejovic-Milovancevic M**, **Sigurdsson E**, **Silagadze T**, **Hansen CS**, **Stefansson K**, **Stefansson H**, **Steinberg S**, **Tosato S**, **Werge T**, **Collier DA**, **Rujescu D**, **Kirov G**, **Owen MJ**, **O'Donovan MC**, **Walters JTR**; **GERAD1 Consortium; CRESTAR Consortium**. **2018.** Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics* **50:** 381–389.

Park JH, **Gail MH**, **Weinberg CR**, **Carroll RJ**, **Chung CC**, **Wang Z**, **Chanock SJ**, **Fraumeni JF Jr**, **Chatterjee N. 2011.** Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 18026–18031.

Price AL, **Patterson NJ**, **Plenge RM**, **Weinblatt ME**, **Shadick NA**, **Reich D. 2006.** Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38:** 904–909.

Purcell J, **Brelsford A**, **Wurm Y**, **Perrin N**, **Chapuisat M. 2014.** Convergent genetic architecture underlies social organization in ants. *Current Biology: CB* **24:** 2728–2732.

Rehan SM. 2020. Small Carpenter Bees (*Ceratina*). In: Starr C. ed. *Encyclopedia of social insects*. Cham: Springer.

Rehan SM, **Bulova SJ**, **O'Donnell S. 2015.** Cumulative effects of foraging behavior and social dominance on brain development in a facultatively social bee (*Ceratina australensis*). *Brain, Behavior and Evolution* **85:** 117–124.

Rehan SM, **Glastad KM**, **Steffen MA**, **Fay CR**, **Hunt BG**, **Toth AL. 2018.** Conserved genes underlie phenotypic plasticity in an incipiently social bee. *Genome Biology and Evolution* **10:** 2749–2758.

Rehan SM, **Richards MH**, **Adams M**, **Schwarz MP. 2014.** The costs and benefits of sociality in a facultatively social bee. *Animal Behaviour* **97:** 77–85.

Rehan SM, **Richards MH**, **Schwarz MP. 2009.** Evidence of social nesting in the *Ceratina* of Borneo (Hymenoptera: Apidae). *Journal of the Kansas Entomological Society* **82:** 194–209.

Rehan SM, **Richards M**, **Schwarz M. 2010.** Social polymorphism in the Australian small carpenter bee, *Ceratina (Neoceratina) australensis*. *Insectes Sociaux* **57:** 403–412.

Rehan SM, **Toth AL. 2015.** Climbing the social ladder: the molecular evolution of sociality. *Trends in Ecology & Evolution* **30:** 426–433.

Ronai I, **Oldroyd BP**, **Vergoz V. 2016.** Queen pheromone regulates programmed cell death in the honey bee worker ovary. *Insect Molecular Biology* **25:** 646–652.

Sakagami S, **Maeta Y. 1987.** Multifemale nests and rudimentary castes of an "almost" solitary bee *Ceratina flavipes*, with additional observations on multifemale nests of *Ceratina japonica* (Hymenoptera, Apoidea). *Kontyu* **55:** 391–409.

Sakagami SF, **Maeta Y. 1984.** Multifemale nests and rudimentary castes in the normally solitary bee *Ceratina japonica* (Hymenoptera, Xylocopinae). *Journal of the Kansas Entomological Society* **57:** 639–656.

Savolainen O, **Lascoux M**, **Merilä J. 2013.** Ecological genomics of local adaptation. *Nature Reviews. Genetics* **14:** 807–820.

Steffen MA, **Rehan SM. 2020.** Genetic signatures of dominance hierarchies reveal conserved cis-regulatory and brain gene expression underlying aggression in a facultatively social bee. *Genes, Brain, and Behavior* **19:** e12597.

Sugg DW, **Chesser RK**, **Dobson FS**, **Hoogland JL. 1996.** Population genetics meets behavioral ecology. *Trends in Ecology & Evolution* **11:** 338–342.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.

Valdar W, **Solberg LC**, **Gauguier D**, **Burnett S**, **Klenerman P**, **Cookson WO**, **Taylor MS**, **Rawlins JN**, **Mott R**, **Flint J. 2006.** Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics* **38:** 879–887.

Voight BF, **Kudaravalli S**, **Wen XQ**, **Pritchard JK. 2006.** A map of recent positive selection in the human genome. *PLoS Biology* **4:** 446–458.

Wang J, **Wurm Y**, **Nipitwattanaphon M**, **Riba-Grognuz O**, **Huang YC**, **Shoemaker D**, **Keller L. 2013.** A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493:** 664–668.

Weir BS, **Cockerham CC. 1984.** Estimating *F*-statistics for the analysis of population structure. *Evolution; international journal of organic evolution* **38:** 1358–1370.

Whiting PW. 1933. Selective fertilization and sex determination in Hymenoptera. *Science* **78:** 537–538.

Whiting PW. 1943. Multiple alleles in complementary sex determination of Habrobracon. *Genetics* **28:** 365–382.

van Wilgenburg E, **Driessen G**, **Beukeboom LW. 2006.** Single locus complementary sex determination in Hymenoptera: an "unintelligent" design? *Frontiers in Zoology* **3:** 1.

Wilfert L, **Gadau J**, **Schmid-Hempel P. 2007.** Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity* **98:** 189–197.

Wilson EO. 1971. *The insect societies*. Cambridge: Harvard University Press.

Yokoyama S, **Nei M. 1979.** Population dynamics of sex-determining alleles in honey bees and self-incompatibility alleles in plants. *Genetics* **91:** 609–626.

Yu L, **Song Y**, **Wharton RP. 2010.** E(nos)/CG4699 required for *nanos* function in the female germ line of *Drosophila*. *Genesis:The Journal of Genetics and Development* **48:** 161–170.

Zayed A, **Packer L. 2005.** Complementary sex determination substantially increases extinction proneness of haplodiploid populations. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 10742–10746.

Zhang C, **Dong S-S**, **Xu J-Y**, **He W-M**, **Yang T-L. 2019.** PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35:** 1786–1788.

Zhou X, **Stephens M. 2012.** Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44:** 821–824.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1. Genes acted on by positive selection in *Ceratina australensis*.
Table S2. Gene Ontology of genes acted on by positive selection in *Ceratina australensis*.
Table S3. Single nucleotide polymorphisms associated with social nesting in *Ceratina australensis*.
Table S4. Gene Ontology of genes associated with social nesting in *Ceratina australensis*.
Figure S1. Linkage disequilibrium (LD) decay across three populations.