Supplementary Material

Males are capable of long-distance dispersal in a social bee

Francisco Garcia Bulle Bueno, Bernardo Garcia Bulle Bueno, Gabriele Buchmann, Tim Heard, Tanya Latty, Benjamin P. Oldroyd, Anette E. Hosoi, Rosalyn Gloag

In this document:

See "Supplementary Information Table S1.xlxs" for Table S1

Model Description

1.1 Probability of male sibship

We used a Bayesian model to estimate the probability that two male *Tetragonula carbonaria* in our dataset were brothers, given their alleles at seven microsatellite loci and the allele frequencies of our sampled population. Under this model, pairwise sibship probability is calculated as follows:

$$\boldsymbol{P\left(b_{ij} = 1 \middle| G_i = g_i, G_j = g_j\right)} \\ = \frac{\boldsymbol{P(G_i = g_i, G_j = g_j, b_{ij} = 1)}}{\boldsymbol{P(G_i = g_i, G_j = g_j)}} \tag{1}$$

where b_{ij} is a variable indicating that males *i* and *j* are brothers, G_i and G_j are random variables which take values from the set of all possible genotypes for males *i* and *j* respectively, and g_i and g_j are their actual genotypes.

Equation (1) is equal to:

$$= \frac{P\left(G_i = g_i, G_j = g_j \middle| b_{ij} = 1\right) * P(b_{ij} = 1)}{P\left(G_i = g_i, G_j = g_j \middle| b_{ij} = 1\right) * P\left(b_{ij} = 1\right) + P\left(G_i = g_i, G_j = g_j \middle| b_{ij} = 0\right) * P(b_{ij} = 0)}$$

We can then consider each term in this equation in turn. First, the probability that males *i* and *j* have genotypes g_i and g_j if they are brothers can be expressed as:

$$P\left(G_i = g_i, G_j = g_j \middle| b_{ij} = 1\right) = \Pi_{k=1}^{7} P\left(G_{ik} = g_{ik}, G_{jk} = g_{jk} \middle| b_{ij} = 1\right) \tag{2}$$

where G_{ik} is the random variable taking value from the set of all possible alleles at the *k*-th locus for male *i* , $G_{jk}$is the random variable taking value from the set of possible alleles at the *k*-th locus for male *j,* and g_{ik} and g_{jk} are the actual alleles of each male at that locus.

Haploid male bees inherit **one** of their diploid mother's two alleles at each locus. If we imagine a mother's genotype to be A_1A_2, then the probability that brothers inherit the same allele (say, A_1) is 0.5, and we denote the event as $b_{ij} = 1$. Likewise, the probability that brothers inherit different alleles (one brother inherits A_1 and the other A_2) is 0.5 and we denote the event as $b_{ij} = 0$. Therefore:

$$P\left(G_{ik} = g_{ik}, G_{jk} = g_{jk} | b_{ij} = 1\right) = \left(\frac{1}{2}\right) * P\left(G_{ik} = g_{ik}, G_{jk} = g_{jk} | b_{ij} = 1,\ b_{ij}^{k} = 1\right) + \left(\frac{1}{2}\right) * P\left(G_{ik} = g_{ik}, G_{jk} = g_{jk} | b_{ij} = 1,\ b_{ij}^{k} = 0\right) \tag{3}$$

Note that the alleles at the *k*-th locus of brothers may have the same value even under the condition $b_{ij}^{k} = 0$. This happens if the mother is homozygous at that locus (i.e. $A_1 = A_2$).

We then estimate the probability of each condition in equation (3) based on the frequency of alleles in the total population, such that:

$$P\left(G_{ik} = g_{ik}, G_{jk} = g_{jk} | b_{ij} = 1, b_{ij}^{k} = 1\right) = P(G_{ik} = g_{ik})\ if\ g_{ik} = g_{jk}, 0\ otherwise \tag{4}$$

And

$$P\left(G_{ik} = g_{ik}, G_{jk} = g_{jk} | b_{ij} = 1, b_{ij}^{k} = 0\right) = P(G_{ik} = g_{ik}) * P\left(G_{jk} = g_{jk}\right)$$

where the probability of carrying a given allele at the k-th locus (that is, $P(G_{ik} = g_{ik})$) is equal to that allele's frequency in our total sampled population:

$$P(G_{ik} = g_{ik}) = \sum_{l=1}^{n} I_{g_{lk=gik}} / n \tag{5}$$

Where $I_{g_{lk=gik}}$ is an indicator variable taking the value of 1 if $g_{lk=}g_{jk}$, and 0 otherwise.

We similarly use population allele frequencies to calculate the probability of males *i* and *j* carrying their observed genotypes if they are not brothers:

$$P\left(G_i = g_i, G_j = g_j \middle| b_{ij} = 0\right) = P(G_i = g_i) * P(G_j = g_j) \tag{6}$$

where

$$P(G_i = g_i) = \Pi_{k=1}^{7} P(G_{ik} = g_{ik})$$

and likewise, for $P\big(G_j = g_j\big)$.

Finally, we assume that the prior probability of two males being brothers, independent of any genotype information, is proportional to the total number of colonies contributing males to the sample set, *S*. Thus:

$$P\big(b_{ij} = 1\big) = 1/S \tag{7}$$

And

$$P\big(b_{ij} = 0\big) = 1 - 1/S$$

1.2 Simulations

We used a simulation-based approach to estimate the typical natal dispersal distances of male *T. carbonaria*. Males dispersed from their natal nests according to the exponential function:

$$P(d) = \lambda e^{-\lambda d}$$

where $d =$ metres flown, λ = 1/mean dispersal distance. For each of sample Sets 2 and 3, we ran simulations for 30 values of λ that represented mean male dispersal distances between 500m and 6500m. Each simulation followed these steps:

i. **Location of male-producing colonies**. We generated a colony at a random site within a virtual study area. The virtual study area was a rectangle overlaid on the map of our actual study site, with all boundaries at least 30km away from any collection site (Fig S1). Colony locations were determined according to wr, where a random state wr∈1,2,3,…,100. We used onwater.io (https://onwater.io/) to assess whether simulated colonies fell onto water and reassigned them if so (**Fig. S1**).

ii. **Male genotypes.** For each colony, we first assigned a queen genotype with independent random sets of alleles at each of seven loci, based on population allele frequencies. We then generated 3000 males per colony where males were randomly assigned one allele per locus from their mother. This number approximates total males produced by a strong colony in Sydney during spring in one month (see Results, this study).

iii. **Male dispersal**. The distance flown by each male (*d*) was generated according to the distribution *P(d)* above. The final destination of each male was uniformly selected as a random point on the circle of circumference *d*, centred on the natal colony. That is, we assumed males were equally likely to fly away from the colony in all directions. Any males whose final destination was above water were allocated another final destination.

iv. **Male collection**. Any males with final destinations within *M* metres of a collection site was added to that collection (Set 3, *M* = 500m, Set 2, *M* = 300m). A sensitivity analysis of *M* is provided in Supp. Material 1.3)

v. **End collection**. We continued to simulate virtual colonies until the number of represented families in our virtual collection was equal to Nf (the number of males sampled in our actual dataset). If the number of virtually collected bees is larger than the number collected in the experiment, we randomly keep only a number of bees equal to the collected sample.

vi. **Cumulative Distribution Functions** Finally, we calculated the sibship of each male pair in our simulated collection (as done for actual data above, **Supplementary Material, 1.1**) and obtained the cumulative distribution function, binned by the probability of sibship (0-0.05, 0.05-0.15, 0.15-0.25, ..., 0.85-0.95, 0.95-1). We calculated such CDFs for each of five values of *S*, representing different assumptions about the number of total colonies represented in our sample (Set 2, S = 201, 380, 475, 570; Set 3, S=100, 300, 450, 600, 750, 900).

vii. **Simulations vs Observed data** We assessed which values of λ (i.e which dispersal distributions) gave simulated datasets of male collections that most closely matched our actual datasets. For each λ, we took the geometric average of the area between the simulated and observed CDF curves. The lower the area, the more closely the simulation matched our real data.

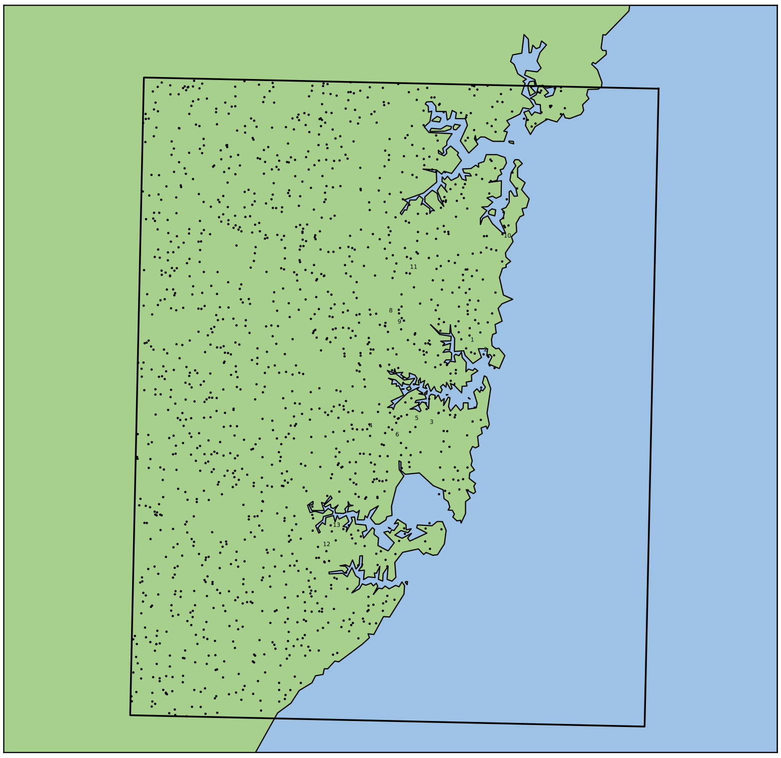

Fig S1. A map of our Set 3 collection site showing virtual simulated colonies (black dots). The simulated colony locations were determined according to wr, where a random state $wr \in 1,2,3,\ldots,100$.

1.3 Sensitivity analysis of *M*

In our simulations, *M* is the distance that a male must pass within a requeening colony (i.e. a collection site) for the male to be included in our sample. In biological terms, it represents the range at which *T. carbonaria* males can detect a virgin queen's pheromone (or other signal emitted by colonies in the requeening process). In honey bees, this distance has been estimated at 100m (Brockmann, Dietz, Spaethe, & Tautz, 2006), but for stingless bees it is unknown. We chose values ranging between 300 and 500m for Sets 2 and 3 respectively, which represent the largest possible area without causing overlap in the detection radius of our collection sites. To check that these values of *M* did not introduce significant variability in our results, we tested how the results for Set 3 (Sydney 2018) would change if *M* took different values (200m, 300m or 500m); **Fig. S2** below. As each value of *M* gave similar mean dispersal distances for males, we conclude that our simulation results are robust within a reasonable range of possible values of *M*.

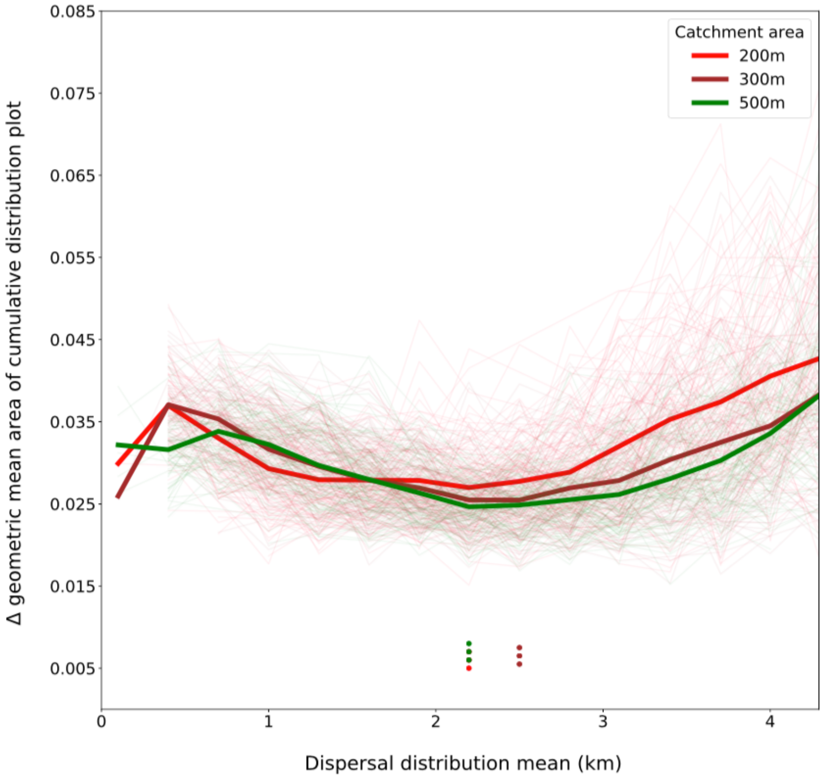

Fig. S2. Cumulative distribution plots of the difference between simulated data and observed data based on Set 3 collections for three values of *M* (200, 300, 500m). These simulations ran with 600 total families, and minimum and maximum flight means of 100m and 4.2km, respectively.

2.1 Detection error of the number of colonies contributing to male aggregations

T. carbonaria male aggregations vary in size, but in some cases are very large. In our study, we typically genotyped only 200 males from large aggregations. To estimate what proportion of total colonies (families) contributing males to an aggregation would be detected from a sample of 200 males, we genotyped additional males for three large aggregations sampled in Set 1 (507±28 males genotyped per aggregation). We then calculated the number of families represented in our sample using COLONY (Wang, 2004) for increasing intervals of 100 males and plotted the number of samples vs number of detected families (Python (Sanner, 1999). Based on these plots (Fig. S3), 200 males typically detected around 80% of the families contributing to large swarms.

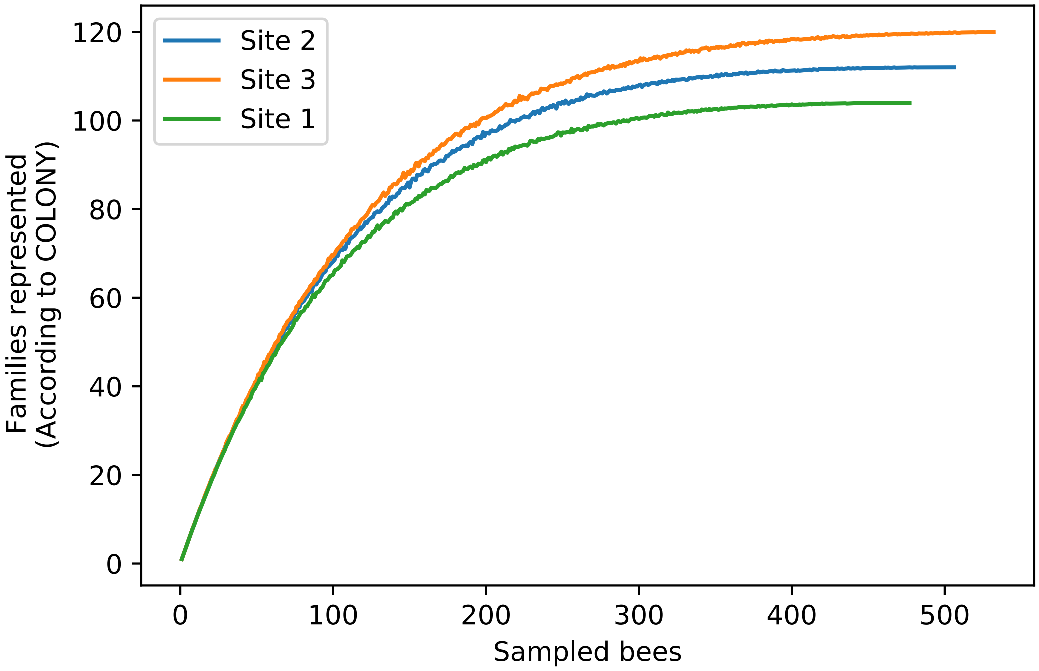

Fig. S3 Cumulative sampling distributions of estimated number of families contributing males to a mating aggregation in large swarms (Set 1, Sites 1-3).

3.1. Evidence for male dispersal distances based on males with identical genotypes

We manually sorted genotypes of all the males from set 2 and 3 to identify males that shared alleles at all seven loci. For the genotypes of each of these conservative sets of possible brothers, we then calculated the probability that two males in our population would share that genotype by chance, using the formula:

$$p = f_1 * f_2 * f_3 \dots * f_7$$

where f_1 is the frequency of the observed allele at locus 1, f_2 is the frequency of the observed allele at locus 2, etc. We then calculated the probability that the males' shared a genotype by descent (i.e. were brothers), rather than chance, according to:

$$p_s = (1 - p)^n$$

where *n* is the total sample size of all males sampled. In this way, we identified pairs of males that were carrying rare alleles in combinations that made it highly unlikely that they shared genotypes by chance alone. We considered pairs of males with $p_s > 0.85$ to be likely brothers, and $p_s > 0.95$ to be highly likely brothers. We then plotted the distance separating the sample location of these brother pairs (**Fig S4**). As for our sibship assignment using models and simulations (**Supplementary Material, 1.1-1.3**), this estimate revealed that the great majority of likely brothers were collected from the same or nearby aggregations, 0-7km apart ($p_s > 0.85$, N=310; $p_s > 0.95$, N=199), but a small number were sampled at aggregations separated by >10km ($p_s > 0.85$, N=15; $p_s > 0.95$, N=1); Fig S4.

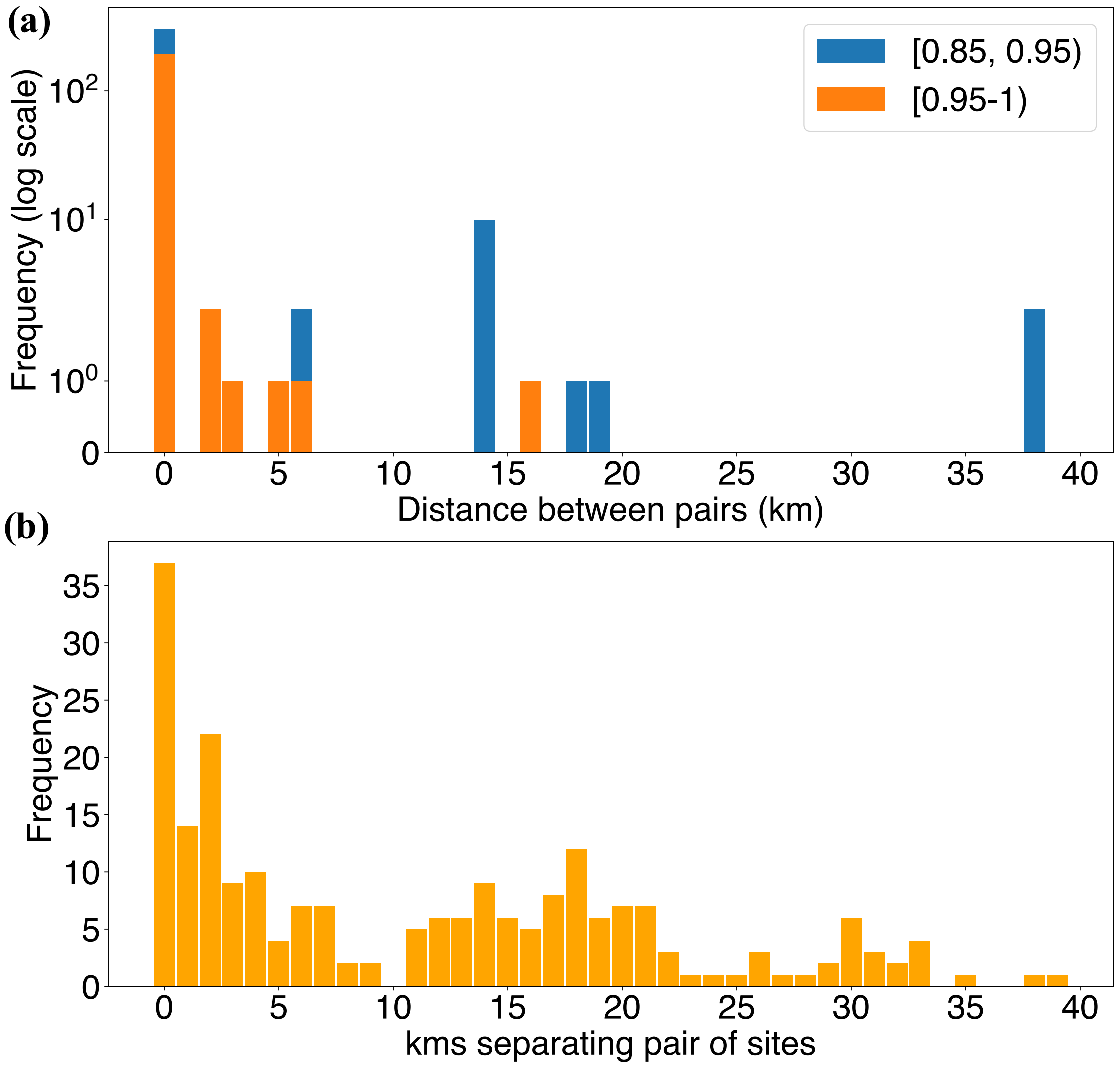

Fig S4. Distances between pairs of males sampled at mating aggregations (Sets 2 and 3), for males that share alleles at all seven loci analysed: (a) The number of these males with high probability of sibship sampled from mating aggregations separated by different distances ($p_s >$ 0.85, blue bars, n=325; p_s> 0.95 (orange bars, n=200). Most of these likely pairs of brothers were collected from the same or nearby aggregations, but a small number were sampled at aggregations separated by >10km ($p_s > 0.85$, N=15; $p_s > 0.95$, N=1). (b) the distance between pairs of sampled aggregations (sites) in our collections).

Table S2. Four replicates of mark-recapture of males at mating aggregations. A small proportion of our marked males reappeared at target mating aggregations within 48 hours when released at distances of 1-4.5km from the target aggregation (0.2 - 2.5% per replicate). In Replicates 1-3, we also released a subset of males directly underneath the target aggregation, of which 17-27% were then sampled in the aggregation. For Replicate 4, we monitored two target swarms at different distances from the release sites.

Replicates	Total males marked		Males released per distance	Distance from target mating aggregation/s (km)	N males recaptured at target swarm		
					24h	36h	48h
1	1490		630	1 km	3	2	2
			600	1 km	2	1	0
			260	At target	34	13	10
2	900		400	1 km	7	6	0
			400	1 km	4	3	0
			100	At target	17	10	0
3	2100		1000	2.5 km	0	0	2
			1000	4.5 km	0	0	2
			100	At target	0	0	17
4	1600	target 1	800	2 km	0	1	0
			800	1.5 km	0	2	1
		target 2	800	1 km	0	16	4
			800	3 km	0	0	0

References for Supplementary Material

Brockmann, A., Dietz, D., Spaethe, J., & Tautz, J. (2006). Beyond 9-ODA: sex pheromone communication in the European honey bee Apis mellifera L. *Journal of chemical ecology, 32*(3), 657-667.

Sanner, M. F. (1999). Python: a programming language for software integration and development. *J Mol Graph Model, 17*(1), 57-61.

Wang, J. (2004). Sibship reconstruction from genetic data with typing errors. *Genetics, 166*(4), 1963-1979.