Appendix 1 – Population Genetics and Colony Assignments

1 Assessing locus F_{is}, F_{st} and linkage disequilibrium

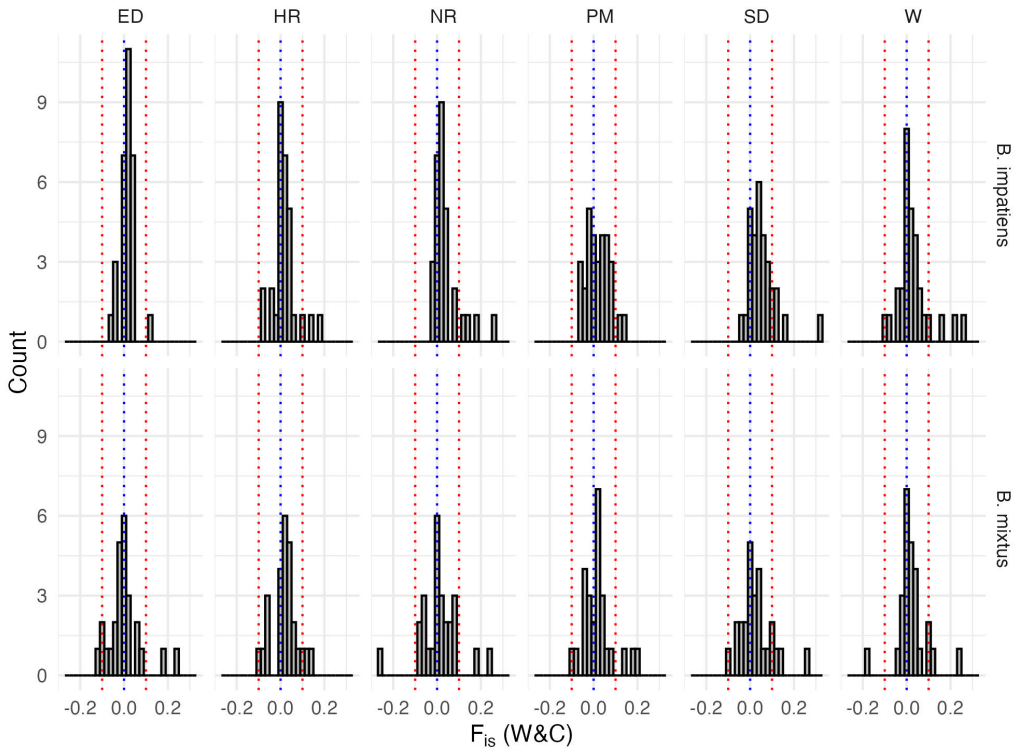

Figure A1: Estimates of F_{is} for each locus in each subpopulation. Estimates from 2022 and 2023 were calculated separately but are shown together for each site x species combination. Blue dotted lines indicates $F_{is} = 0$ and red dotted lines indicate $F_{is} = \pm 0.1$.

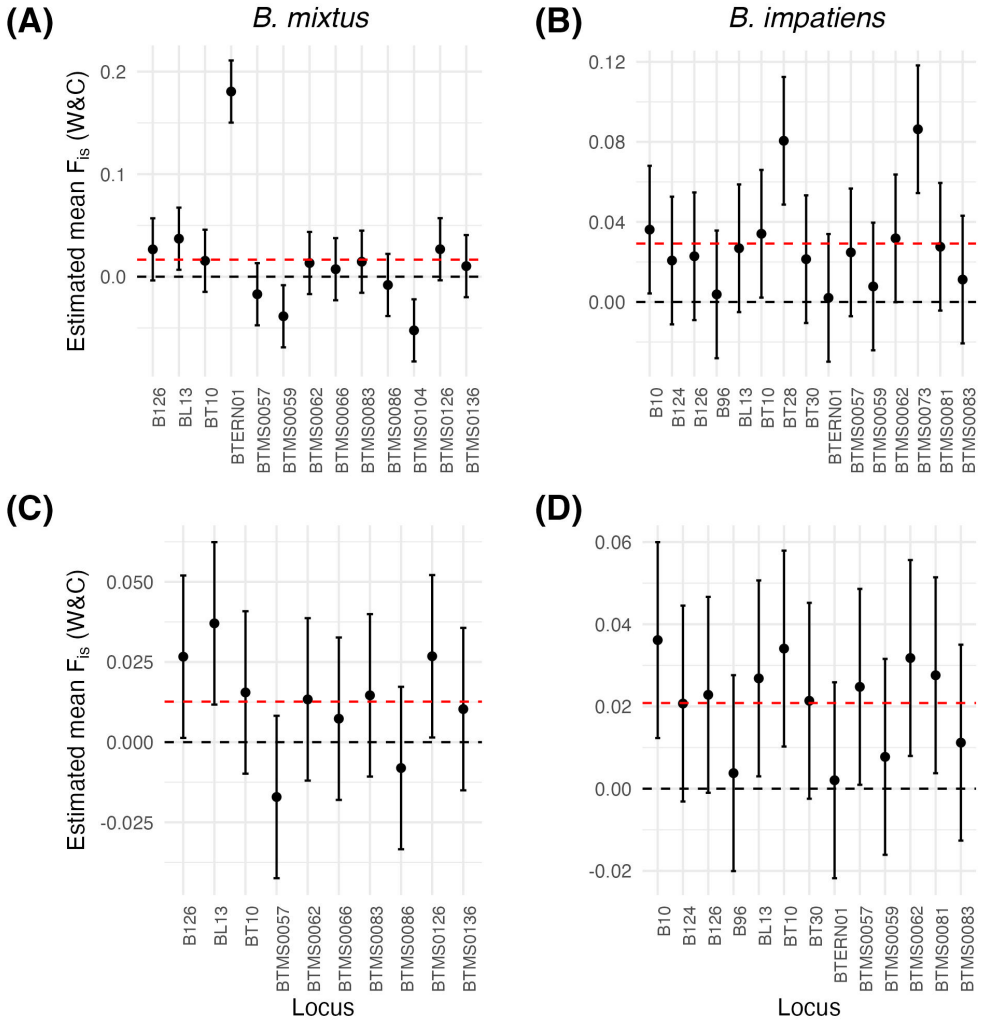

Figure A2: Locus-specific F_{is} marginal means. A) *B. mixtus* all loci; B) *B. impatiens* all loci; C) *B. mixtus* loci following iterative removal of loci which differed significantly from global mean F_{is}; D) *B. impatiens* loci following iterative removal of loci which differed significantly from global mean F_{is}. Dashed black line denotes $F_{is} = 0$, dashed red line denotes global mean F_{is} for each species.

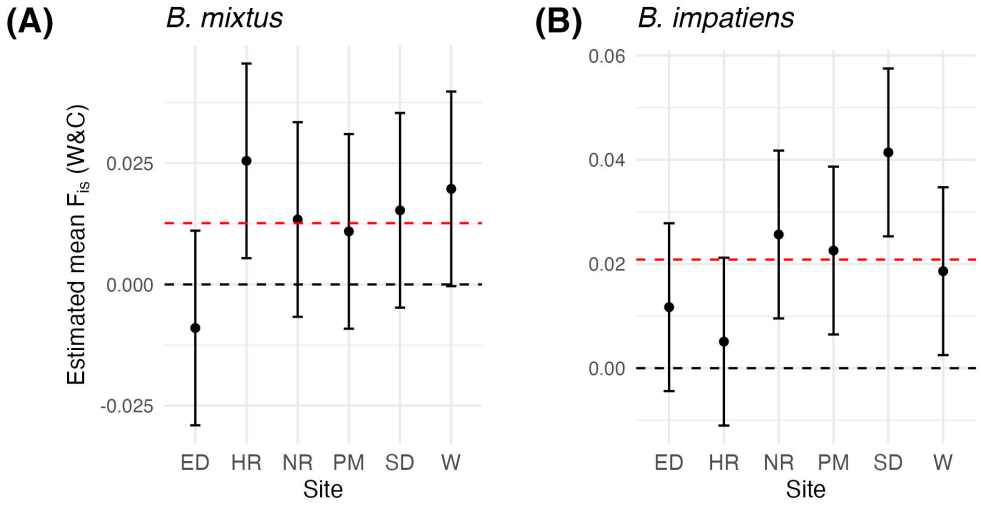

Figure A3: Site-specific F_{is} marginal means following removal of low-quality loci for A) *B. mixtus* and B) *B. impatiens*. Dashed black line denotes $F_{is} = 0$, dashed red line denotes global mean F_{is} for each species.

2 Testing COLONY on simulated data

To test the informativeness of our genetic loci and to validate the accuracy of COLONY2.0 (**jonesCOLONYProgramParentage2010**) for accurately detecting siblingships amongst our samples, we performed simulations using realistic family size distributions and the allelic frequencies present in our real data.

We approached this simulations with four objectives:

(i) To determine false positive and false negative siblingship assignment rates, given the informativeness of our microsatellite datasets,

(ii) To inform an appropriate strategy (probability threshold, number of runs of the software) for maintaining or rejecting each sib-pair;

(iii) To select suitable software parameters, and in particular to evaluate the usefulness of siblingship size priors and exclusion of across-site siblingships for reducing false positive rates as sample size increases;

(iv) To assess whether modelling female polygamy would improve family reconstruction in the case of sibling genotypes simulated under varying

rates of multiple paternity.

2.1 Simulation strategy

2.1.1 Spatially explicit siblingships

We first simulated spatially explicit siblingships following **popeInferringForagingRanges2017<** We began by simulating six 5 x 5 trapping grids (locations $k \in \kappa$) on a single raster surface comprised of cells $j \in \mathbb{J}$. Colonies ($i \in \mathbb{C}$) were distributed uniformly at random throughout the "landscape."

We sampled individuals from colonies $i \in \mathbb{C}$ captured at traps $k \in \mathbb{K}$ from the joint distribution $\Pr(s, c \mid s \in \kappa)$, where $\{s, c\}$ are the indices of a random visitation event of an individual from colony $c \in \mathbb{C}$ to grid cell $s \in \mathbb{J}$.

To do this, we first sampled a trap (k) from

$$\Pr(s = k \mid s \in \kappa) \;=\; \frac{\Pr(s = k)}{\Pr(s \in \kappa)} \tag{1}$$

where

$$\Pr(s = k) \;=\; \sum_{i \in C} \Pr(s = k \mid c = i)\ \Pr(c = i)$$

and

$$\Pr(s \in \kappa) \;=\; \sum_{i \in \mathbb{C}} \Pr(s \in \kappa \mid c = i)\ \Pr(c = i) \;=\; \sum_{i \in \mathbb{C}} \sum_{k \in \kappa} \Pr(s = k \mid c = i)\ \Pr(c = i)$$

Combining these statements gives a probability of sampling from trap k of:

$$\Pr(s = k \mid s \in \kappa) \;=\; \frac{\sum_{i \in C} \Pr(s = k \mid c = i)\ \Pr(c = i)}{\sum_{i \in \mathbb{C}} \sum_{k \in \kappa} \Pr(s = k \mid c = i)\ \Pr(c = i)} \tag{2}$$

We then sampled a colony (i) from

$$\Pr(c = i \mid s = k) \;=\; \frac{\Pr(s = k \mid c = i)\Pr(c = i)}{\Pr(s = k)} \;=\; \frac{\Pr(s = k \mid c = i)\Pr(c = i)}{\sum_{i \in \mathbb{C}} \Pr(s = k \mid c = i)\ \Pr(c = i)} \tag{3}$$

We define the foraging kernel of workers from colony i as:

$$\Pr(s = k \mid c = i) \;=\; \frac{\lambda_i(k)}{\sum_{j \in J} \lambda_i(j)} \tag{4}$$

The visitation intensity of individuals from colony i to location j is defined as:

$$ln(\lambda_i(j)) = \frac{-\|x_j - \delta_i\|}{\rho} \tag{5}$$

where x_j are the spatial coordinates of any grid cell in the raster, and δ_i are the spatial coordinates of colony i. The foraging kernel in this example is therefore assumed to be symmetrical and exponentially decaying as a function of distance from the colony location. This means that the total visitation of each colony across the landscape ($\sum_{j \in J} \lambda_i(j)$) is the same for all colonies, and can be represented using the constant $\mathbb{D}$. $\Pr(c = i)$ is the proportion of all bees in the landscape originating from colony i, e.g., $\Pr(c = i) = \frac{n_i}{N}$ where n_i is the number of bees from colony i, and $N = \sum_{i \in \mathbb{C}} n_i$ is the total number of bees in the landscape.

Combining (4) with (2) and (3) gives the probability of sampling an individual from trap k

$$\Pr(s = k \mid s \in \kappa) \;=\; \frac{\sum_{i \in \mathbb{C}} \lambda_i(k) \frac{n_i}{N}}{\sum_{k \in \kappa} \sum_{i \in \mathbb{C}} \lambda_i(k) \frac{n_i}{N}}$$

and the probability that the individual originates from colony i

$$\Pr(c = i \mid s = k) \;=\; \frac{\lambda_i(k) \frac{n_i}{N}}{\sum_{i \in \mathbb{C}} \lambda_i(k) \frac{n_i}{N}}$$

For each simulation, samples are drawn from $\Pr(s, c \mid s \in \kappa)$ until a stopping point (desired number of samples) is reached. n_i is updated after each "sampling event" to prevent oversampling from colonies located very close to traps.

To verify that the size of sampled siblingships (e.g., number of siblings per sibling group) accurately mirrors the distribution of siblingship sizes in real data, we compared our simulated distributions to the distribution of siblingshp sizes in our real data (**??**). For this simulation strategy, we found that moderating the background density of colonies (i.e., the total number of colonies simulated on the landscape) was the most effective strategy for moderating average siblingship size. A larger number of simulated colonies resulted in a higher proportion of singleton colonies (colonies represented by only a single individual).

2.1.2 Multilocus genotypes

We simulated multilocus genotypes for each sampled individual under several mating scenarios. In the simplest case, we assume monogamy for both males and queens. The majority of the simulation results presented below follow this assumption. In a second set of simulations, we assumed varying rates of multiple paternity (e.g., queen polygamy) to assess the impact of this assumption on siblingship inferences. For each simulation we used the following heuristic:

(i) Simulate parental genotypes for each siblingship based on the allele frequencies present in our real data;

(ii) Randomly draw offspring genotypes from the set of possible parental alleles at each locus.

We performed simulations based on allele frequencies for both species (*B. mixtus* and *B. impatiens*) because variation in marker number and/or polymorphic information content could lead to differing results. We used inferred allele frequencies from an earlier run of COLONY2.0, which accounts for heightened frequency of alleles present in large families; although raw allele frequencies would have likely been sufficient, given that average family size was small (<2 individuals) and families with > 4 individuals were rare for both datasets.

In the case of monogamous matings, male genotypes were assigned di-

rectly to all offspring in the siblingship; in families which were assigned multiple paternity, we assumed two fathers and assigned inheritance of paternal multilocus genotypes from $\Pr(father_1, father_2) = (0.7, 0.3)$ following the proportions observed for *B. impatiens* in **birdMatingFrequencyEstimation2024<emp**

After assigning a multilocus genotype to each individual, we introduced errors and data missingness based on observed error and missingness rates in our real datasets. To introduce errors, we mutated each allele with a probability equal to the rate of errors for that locus and species; we assumed that most errors would be due to contamination, rather than allele dropout, and drew new (erroneous) alleles from the same allele frequencies described above. In the case of data missingness, we observed that individuals which were missing data for *one* copy of a locus were more likely to be missing data for *both* copies than if missingness were distributed uniformly at random. This is likely because there were two primary missingness-generating processes in real data: amplification failure (both alleles missing for an individual) and binning failure (one or both alleles missing for an individual). (In cases were only one copy of a locus failed to amplify, heterozygous individuals would be falsely classified as homozygous—an error, rather than missing data). To account for the structure of missingness, we first calculated the proportion of missing data for each marker ($P_{missing}$) and then removed data for (i) both alleles of each individual with probability $1/3 * P_{missing}$ and for (ii) a single allele per individual with probability $1/3 * P_{missing}$.

2.2 Determining an appropriate heuristic for maintaining or rejecting inferred siblingships

It has been previously noted (or speculated) in the literature than COLONY is prone to inferring siblingships between non-siblings, referred to hereafter as *false positive sibships*. Our own preliminary data analyses suggested that this was the case (e.g., a high number of inferred siblingships between individuals separated by >20 km, when individuals from all study sites were permitted to form siblingships). While the biologically of bumblebees does not unilaterally exclude the possibility of such distant relationships, the likelihood of observing such separation distances is extremely small (see discussion below, section "Observing colonymates at multiple sites").

A commonly used strategy to deal with false positives is to repeat multi-

ple "runs" (usually 2-5) of the COLONY software on the same dataset, and maintain family groups which are inferred in all runs at or above some confidence threshhold (usually $P \geq 0.95$, but sometimes $P \geq 0.8$). See, for example, **carvellMolecularSpatialAnalyses2012**; **raoBumbleBeeHymenoptera2012**; **dreierFinescaleSpatialGenetic2014a**; **geibBumbleBeeNest2015a**; **carvellBumblebeeFamilyLin** **molaWildfireRevealsTransient2020a<empty citation>**. However, we are not aware of any studies which give support for a particular threshhold probability or number of runs necessary to reach a particular confidence level in assignments, or to achieve a satisfactory balance between false positive siblingships and missed (false negative) siblingships. Indeed, the desireable threshhold is likely to vary as a function of the number and informativeness of markers for a given population.

To overcome these limitations, we tested probability exclusion criteria from $P = 0.95$ to $P = 1$, for 1 or 5 runs of COLONY version 2.0.6.5. Further, we compared the use of family cluster probabilities (COLONY output file .BestCluster—hereafter referred to as the family method) to full sibling dyad probabilities (COLONY output .FullSibDyad—hereafter referred to as the dyad method).

We began by simulating 5 datasets (e.g., different siblingship arrangements with unique parental genotypes) consisting of n = 1200 individuals each, which was roughly the midpoint of population sizes for our real data. For each dataset we performed 5 runs of COLONY (see Table **??** for a summary of COLONY software settings).

We first note that repeated runs of the software appear to serve almost no purpose—in most cases, false positive and false negative rates are either identical or nearly overlapping (Fig **??**).

We additionally found that the family method was less reliable than dyad method for excluding erroneous siblingships; false negative rates were similar for both strategies, but 17-36% of all inferred pairwise relationships were false positives when using the family method, even under the most stringent probability threshold (Fig **??**). The dyad method, in contrast, resulted in $\leq$ 11% false positives for all conditions tested, and achieved a false positive rate of around 5% for $P \geq 0.99$ (Fig **??**.

We note that for very high values of P (e.g., $P = 1$) the dyad method leads to a higher proportion of false negatives ($\geq$ 5%). For this reason, we

chose a probability threshold of $P = 0.99$ for our analysis, as we can be reasonably confident that this threshold will result in both false positive and false negative rates at or around 5%.

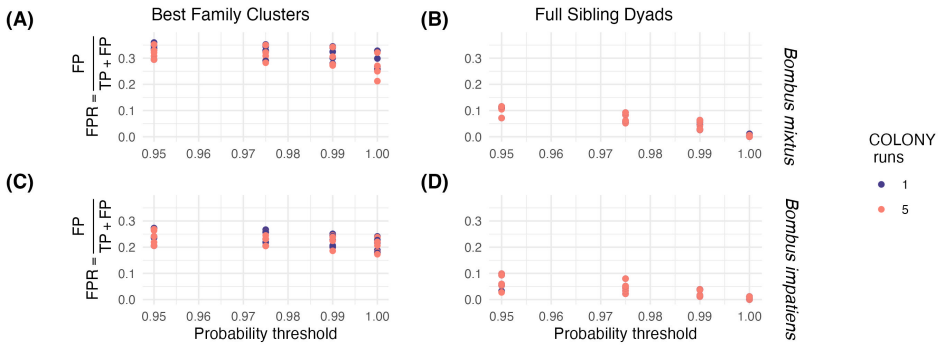

Figure A4: caption

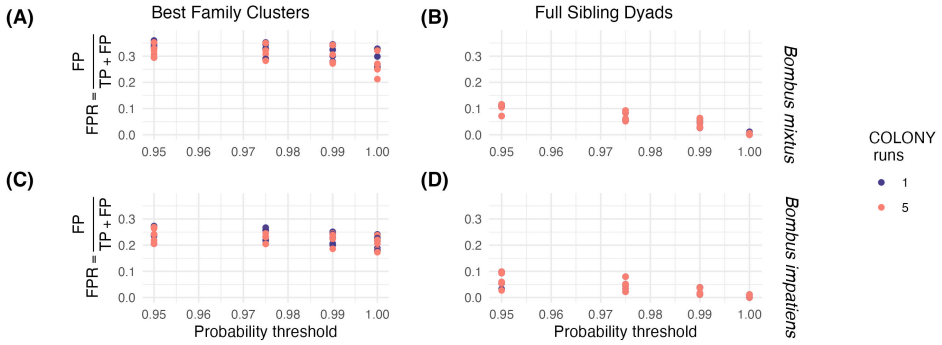

Figure A5: caption

One benefit of using family clusters rather than sibling dyads to determine sibling relationships arises in the case of "non-circular families." These families represent cases in which some (but not all) individuals in a group are inferred to be full siblings (e.g., A related to B, B related to C, A not related to C). Such cases are rare, but are handled internally by COLONY such that all inferred family clusters are circular. Because we choose here to refer to dyad pairs rather than family clusters, we are left the task of deciding how to resolve non-circular families—indeed, resolution of non-circular families could be one process which leads to a higher false positive rate.

Because of this uncertainty, we explored several heuristics for resolving non-circular families. We started by exploring the structure of non-circular families in our simulated datasets, to determine whether non-circularity is more frequently the result of false positives (e.g., a third individual being erroneously added to a sibling pair) or false negatives (e.g., failure to detect a sibling relationship between any pair of siblings in a triad).

To do this, we plotted non-circular families from all five simulations, using a threshold of $P = 0.99$ for inclusion of pairwise relationships.

b

2.3 Assessing the use of cross-site sibling exclusion for reducing false positive rates

We next systematically evaluated the use of sibling exclusion criteria and siblingship size priors to determine whether these software specifications have any impact on siblingship inference accuracy. Previous studies on *Bombus* have varied in their approaches to the scale at which possible siblingships are evaluated, with some studies performing separate runs of the software for populations sampled at different sites/regions (CITE: JHA?) while others group populations at larger scales, permitting the discovery of long distance foraging or dispersal events between sites (CITE: LEPAIS, MOLA, ETC. RAO?). While identifying the maximum foraging or dispersal range for different species or landscape contexts is an important goal, we consider two challenges to this method, related to (i) the rarity of capturing individuals engaged in such long distance events, and (ii) the statistical challenges introduced by a very high number of pairwise comparisons between individuals at different sites. For a more thorough discussion of the first point, we direct the reader to **lepaisEstimationBumblebeeQueen2010a<empty c**
Ultimately, we posit that the rate at which such events should be captured is much lower than both the false positive and false negative rates of the COLONY software, due to the quadratically increasing search area over which foragers will be dispersed as distance from their nest increases (see discussion in **osborneBumblebeeFlightDistances2008<empty citation>**). Furthermore, given the large sample sizes of our populations, and the fact that the majority of pairwise comparisons (without exclusion) will occur between individuals at different sites, we hypothesized that allowing for siblingship assignments between all individuals would result in a high percentage of false positive relationships that would severely bias esti-

mates of colony locations and foraging behaviour.

To test this hypothesis, and to determine whether total sample size had an effect on the utility of excluding cross-site siblingships, we simulated five datasets of n = 2000 individuals, and subsetted each data set to contain 20, 40, 60, 80, or 100% of the initial samples. These data were simulated to reflect trapping grids which were arranged in a 3 x 2 grid with traps in adjacent grids at least 6km apart. The minimum distance between adjacent trapping grids in our real data was 5km (also separated by a large river, expected to limit dispersal), and all other sites were at least 7km apart. The mean foraging distance of colonies in our simulation was set to 1km (99% of all visitations within 3.32km). This would allow for colonies located midway between trapping grids to sampled at two sites, while reflecting the fact that the majority of bumblebee foraging is thought to occur within (at *most*) a few kilometers of the nest. We therefore believe that the simulated data would represent a relatively optimistic view of the number of cross-site siblingships which could be present in the real data.

We created siblingship exclusion tables for COLONY by excluding (for each individual) all potential siblingships with individuals captured at different trapping grids (sites). We ran COLONY on each dataset with and without incorporation of the siblingship exclusion table. Further software specifications for these runs can be found in Table **??**. We used the dyad method and an exclusion threshold of $P = 0.99$ as described in the previous subsection.

We found that for both species (*B. mixtus* and *B. impatiens*) and for all tested sample sizes (n = 400-2000 individuals), cross-site sibling exclusion resulted in a lower false positive rate (Fig **??** A-B). The variation in false positive rates between independent simulations decreased with increasing sample size. When using cross-site exclusion, mean false positive rates were fairly consistent across sample sizes, but without cross-site exclusion, the mean false positive rate tended to decrease with increasing sample size.

False negative rates were similar for both methods, indicating that cross-site exclusion did not cause us to lose a high proportion of real cross-site siblingships (Fig **??** C-D). Indeed, manual inspection of datasets revealed that cross-site sampling was extremely rare or non-existant, given our parameterization and sample size.

Table A1: Description of software settings (COLONY 2.0.6.5 (**jonesCOLONYProgramParentage2010**)) for simulations.

Simulation	Comparison	Sample Size	Sibship Size Prior	Cross-site Exclusion	Runs
Subsection 1	Number of COLONY runs	1200	yes	yes	1-5
Subsection 1	Probability threshold	1200	yes	yes	1-5
Subsection 1	Families vs dyads	1200	yes	yes	1-5
Subsection 2	Siblingship size prior	400-2000	yes/no	yes/no	1
Subsection 2	Cross-site exclusion	400-2000	yes/no	yes/no	1
Subsection 3	Mating system	1000	no	no	1
Subsection 3	Mating system (augmented data)	1000	yes	no	1

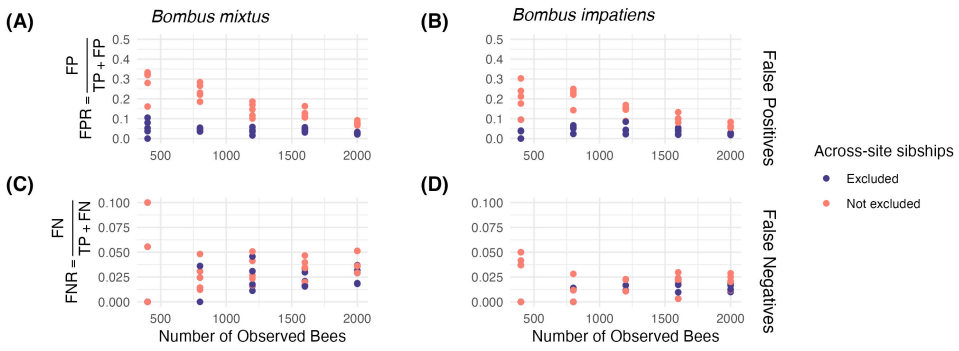

Figure A6: caption

2.4 Assessing the use of siblingship size priors for reducing false positive rates

2.5 Evaluating the effects of multiple paternity on siblingship inference

3 Observing colonymates at multiple sites