Reconstructing queen genotypes by pool sequencing colonies in eusocial insects: Statistical Methods and their application to honeybee

Sonia E Eynard, Alain Vignal, Benjamin Basso, Kamila Canale-tabet, Yves Le Conte, Axel Decourtye, Lucie Genestout, Emmanuelle Labarthe, Fanny Mondet, Bertrand Servin

Received: 13 December 2021 | Revised: 29 June 2022 | Accepted: 4 July 2022

DOI: 10.1111/1755-0998.13685

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Reconstructing queen genotypes by pool sequencing colonies in eusocial insects: Statistical Methods and their application to honeybee

Sonia E. Eynard[1,2] | Alain Vignal[1] | Benjamin Basso[3,4] | Kamila Canale-Tabet[1] | Yves Le Conte[3] | Axel Decourtye[4] | Lucie Genestout[2] | Emmanuelle Labarthe[1] | Fanny Mondet[3] | Bertrand Servin[1]

[1]GenPhySE, INRAE, INP, ENVT, Université de Toulouse, Castanet-Tolosan, France

[2]LABOGENA DNA, Jouy-en-Josas, France

[3]Abeilles et Environnement, INRAE, Avignon, France

[4]ITSAP, Avignon, France

Correspondence
Sonia E. Eynard and Bertrand Servin, GenPhySE, INRAE, INP, ENVT, Université de Toulouse, Castanet-Tolosan, France.
Emails: sonia.eynard@inrae.fr; bertrand.servin@inrae.fr

Funding information
Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt, Grant/Award Number: CASDAR MOSAR RT 2015-776; Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt, Grant/Award Number: Investissement d'avenir (ICF2A) BeeStrong Ps2A

Handling Editor: C. Alex Buerkle

Abstract

Eusocial insects are crucial to many ecosystems, and particularly the honeybee (*Apis mellifera*). One approach to facilitate their study in molecular genetics, is to consider whole-colony genotyping by combining DNA of multiple individuals in a single pool sequencing experiment. Cheap and fast, this technique comes with the drawback of producing data requiring dedicated methods to be fully exploited. Despite this limitation, pool sequencing data have been shown to be informative and cost-effective when working on random mating populations. Here, we present new statistical methods for exploiting pool sequencing of eusocial colonies in order to reconstruct the genotypes of the queen of such colony. This leverages the possibility to monitor genetic diversity, perform genomic-based studies or implement selective breeding. Using simulations and honeybee real data, we show that the new methods allow for a fast and accurate estimation of the queen's genetic ancestry, with correlations of about 0.9 to that obtained from individual genotyping. Also, it allows for an accurate reconstruction of the queen genotypes, with about 2% genotyping error. We further validate these inferences using experimental data on colonies with both pool sequencing and individual genotyping of drones. In brief, in this study we present statistical models to accurately estimate the genetic ancestry and reconstruct the genotypes of the queen from pool sequencing data from workers of an eusocial colony. Such information allows to exploit pool sequencing for traditional population genetics analyses, association studies and for selective breeding. While validated in *Apis mellifera*, these methods are applicable to other eusocial hymenopterans.

KEYWORDS
Apis mellifera, eusocial insects, genotype, pool sequencing

1 | INTRODUCTION

Some species of eusocial organisms live in large colonies produced by a single individual, the queen, having a specific mating system in which she is mated to a cohort of males. In the case of the honeybee, *Apis mellifera*, a colony is always composed of a single queen, a large number (up to tens of thousands) of workers and hundreds of males. The queen is usually the key reproducing individual and all individuals present in the colony are its offspring. In the wild, after mating with a cohort of 10–20 males, a virgin queen will return to her colony and maintain its population, throughout her life, by continuously laying eggs. Fertilised eggs will produce diploid worker females, while unfertilised eggs will produce haploid males. Males are therefore a direct sample of the queen genome and can be considered as flying gametes. The mosaic genetic composition of a colony, chromosomes coming from a queen and multiple inseminating drones, makes standard genomic analyses complex especially when making breeding decisions (Brascamp & Bijma, 2014; Uzunov et al., 2017). In eusocial populations, each worker performs individual tasks participating to the collective phenotype of the colony. However, although the phenotype of the colony is collective, the queen contributes more than half of the genetics of the colony (through diploid female and haploid male offspring) that will be passed on to next generations. Thus, the queen's genotype itself is an essential piece of information for genetic analyses aimed at studying the evolution of populations or at performing selective breeding. Even though the field of insect genomics has boomed in the past decades, there still is a need to expand traditional approaches of population genetics for this specific kind of organisms (Toth & Zayed, 2021). Contrary to large animal species, sampling the queen for genotyping without threatening its integrity is risky. Non-destructive methods to genotype the honeybee queen have been proposed by Gregory and Rinderer (2004), Su et al. (2007) or Bubnič et al. (2020) but can rarely be performed in routine beekeeping practices. Another possible approach is to perform individual or pool genotyping of a set of males (Petersen et al., 2020), gametes of the queen, and reconstruct its genotypes. However, this implies an increased manipulation effort to sample the individual males and of sequencing cost. Advances in sequencing technologies have brought new opportunities to develop tools for genomics. Among these, parallel sequencing allows for counting sequencing reads at positions along the genome, granting for the development of pool sequencing for the estimation of allele frequencies (Schlotterer et al., 2014). By combining DNA from multiple individuals into a unique sequencing experiment, pool sequencing allows for cheap and fast data acquisition, especially for non-model organisms for which resources are limited. However, pool sequencing outcomes, allele counts in the pool instead of genotypes, are more difficult to use in practice and require specific software to perform SNP calling, mainstream population genetics analyses, association testing (Bansal, 2010; Chang et al., 2015; Kofler et al., 2011; Purcell et al., 2007; Speed et al., 2020; Zhou & Stephens, 2012) and more. Additionally, traditional pool sequencing is performed on a group of unrelated individuals representing a population often linked by an environmental factor (e.g. a population in a specific location, an ecotype). This is not the case with honeybee colony pool sequencing as the pooled individuals are related.

In this study, we propose a new application of pool sequencing to multiple individuals from a single colony in the context of eusocial insects. Contrastingly to standard pool experiments, representing a population of individuals, pool experiments on colonies can be seen as sequencing of a meta-individual, the colony. Using this specificity, we introduce dedicated statistical methods to estimate the genetic ancestry of the queen and reconstruct its genotypes from the pool sequencing of its worker. The acquisition of genotype data will (i) provide information on the queen that can further benefit breeding decisions and (ii) allow for the use of standard programs and softwares for population genetics analyses such as admixture or association studies. Two models are proposed and evaluated. The first model estimates the genetic ancestry of the queen, based on data from a single colony but assuming information on the allele frequencies of markers in reference populations. The second model exploits information available across multiple colonies to reconstruct the queen genotypes. Performances of the models are evaluated through simulations, including some based on real data from an *Apis mellifera* diversity panel (Wragg et al., 2022). Using these simulations, we show that the genetic ancestry of a queen estimated from pool sequencing data matches results from standard population genetics methods based on queen genotypes data. Additionally, we show that the genotypes of the queen can be reconstructed with an error rate limited to a few percent. To evaluate the interest of pool sequencing compared to individual genotyping, we applied our genotypes reconstruction models to honeybee data from a field experiment. For multiple colonies, we have both pool sequences of workers and individual sequences of male offspring. We showed that inference of the genetic ancestry and the genotypes of the queen based on pool sequencing data matches results obtained from individual data on male offspring.

Models introduced in this study can be used sequentially to first estimate the genetic ancestry of a population of colonies, then use this information to cluster the data set into homogeneous populations and finally infer queen genotypes by considering colonies jointly. Finally, we discuss the interpretation of the results obtained with the models proposed, their applicability and possible extensions.

2 | MATERIALS AND METHODS

For the sake of understanding, statistical models are presented here from the most simple to the most complex even though they can be used independently in the rest of the paper.

2.1 | Models

We consider data coming from colony pool sequencing experiments. For each colony, whole-genome sequencing is assumed to be

performed on DNA extracted from the mix of a large number of worker bees. For a colony c, the raw data consists of the reference allele counts and sequencing depths at a fixed set of L bi-allelic loci. At a locus l, with observed reference allele count x_l^c and sequencing depth d_l^c, we have:

$$x_l^c \mid d_l^c, f_l^c, g_l^c \sim Binomial\left(\frac{f_l^c + g_l^c}{2}, d_l^c\right) \tag{1}$$

where g_l^c is the (unknown) queen genotypes expressed as the frequency of the reference allele (i.e. 0, 0.5 or 1) and f_l^c is the (unknown) reference allele frequency in the males that mated with the queen. We are interested in reconstructing genotypes of the queen $g_l^c \forall l \in [1 \dots L]$. As f_l^c and g_l^c both contribute to the allele counts in the pool, it is clear that these parameters are unidentifiable without more information. To separate them, we thus need external information on f_l^c and/or g_l^c. We now discuss two possibilities to incorporate such information and the associated inferences that can be drawn.

2.1.1 | Homogeneous population model (HPM)

In this approach, we add to model (1) the hypothesis that queen and males of all colonies come from the same random mating population. Under this hypothesis, (i) the allele frequency at a given locus is the same for all colonies and (ii) genotypes at a locus are sampled according to this frequency, so we have for a locus l:

$$\forall c, f_l^c = f_l$$
$$g_l^c \mid f_l \sim \frac{1}{2} Binomial(f_l, 2) \quad i.e. \begin{cases} P(g_l^c = 0) = (1 - f_l)^2 \\ P(g_l^c = 0.5) = 2f_l(1 - f_l) \\ P(g_l^c = 1) = f_l^2 \end{cases} \tag{2}$$

This new model has only one parameter per locus (f_l) and the likelihood is

$$P(x_l^c, d_l^c \mid f_l) = \sum_{G \in \{0, 0.5, 1\}} P(x_l^c, d_l^c, f_l \mid g_l^c) P(g_l^c = G \mid f_l)$$
$$\mathscr{L}(f_l, \boldsymbol{x_l}; \boldsymbol{d_l}) = \prod_c P(x_l^c, d_l^c \mid f_l) \tag{3}$$

where $\boldsymbol{x_l}$ is the vector of reference allele counts in all colonies and $\boldsymbol{d_l}$ the corresponding vector of sequencing depths. The likelihood (3) is maximized numerically for f_l on $[0, 1]$. The maximizing value (called the Maximum Likelihood Estimate, MLE) $\hat{f}_l$ can be used for inference of $\boldsymbol{g_l}$ based on the posterior distribution $P\left(\boldsymbol{g_l} \mid \boldsymbol{x_l}, \boldsymbol{d_l}, \hat{f}_l\right) \propto P\left(\boldsymbol{x_l} \mid \boldsymbol{d_l}, \boldsymbol{g_l}, \hat{f}_l\right) P\left(\boldsymbol{g_l} \mid \hat{f}_l\right)$.

The HPM should only be applied when the set of colonies have a similar genetic background. We, therefore, developed another approach, the admixture model, aimed at estimating the genetic ancestry of a single colony from pool sequencing data.

2.1.2 | Admixture model (AM)

The objective of this model is to describe the 'genetic background', the subspecies, of a colony. To do so, we will adopt the widely used modelling framework introduced by Pritchard et al. (2000) and define the genetic background of a colony as *the proportions of the queen genome coming from a set of pre-defined reference populations*. In our applications below, the reference populations considered are *Apis mellifera mellifera*, *Apis mellifera ligustica & carnica* and *Apis mellifera causasia*, the three main genetically distinct populations found in Western Europe (Wragg et al., 2022). We will do that in a supervised manner and therefore, we will assume that we are provided with allele frequencies in a set of K reference populations at the L loci: this takes the form of an $L \times K$ matrix $\boldsymbol{F}$ where F_{lk} is the frequency of the reference allele at locus l in population k. Here we are interested in inferring $\boldsymbol{q}$, the K-vector of admixture proportions for the queen: q_k is the proportion of alleles over all loci that come from population k. Dropping the c index as the model is fitted for each colony independently, the likelihood for $\boldsymbol{q}$ is:

$$P(x_l \mid d_l, \boldsymbol{F_l}, \boldsymbol{q}) = \sum_g \int_0^1 P(x_l \mid d_l, g_l, f_l) P(g_l \mid \boldsymbol{q}, \boldsymbol{F}) P(f_l \mid \boldsymbol{F}) \mathrm{d}f_l$$
$$\mathscr{L}(\boldsymbol{q}; \boldsymbol{x}, \boldsymbol{d}) = \prod_l P(x_l \mid d_l, \boldsymbol{F}, \boldsymbol{q}) \tag{4}$$

In order to compute likelihood (4), we need to specify $P(g_l \mid \boldsymbol{q}, \boldsymbol{F})$, the prior distribution for g_l given the admixture proportions, and $P(f_l \mid \boldsymbol{F})$ the prior for the allele frequency at locus l. To perform inference, we need to devise a way of maximizing the likelihood (4). We now explain how we addressed these two issues.

2.1.3 | Priors

To specify the prior $P(g_l \mid \boldsymbol{q}, \boldsymbol{F})$, we use the classical approach of introducing latent variables $\boldsymbol{Z}_l = (z_l^1, z_l^2)$ at each locus l that denotes the origins (in terms of reference populations) of the two alleles carried by the queen. Then we can write:

$$P(g_l \mid \boldsymbol{q}, \boldsymbol{F}) = \sum_{Z_l} P(g_l \mid \boldsymbol{Z}_l, \boldsymbol{F}) P(\boldsymbol{Z}_l \mid \boldsymbol{q}) \tag{5}$$

where $P(g_l \mid \boldsymbol{Z}_l, \boldsymbol{F})$ is the probability of the queen genotypes given the origins of the two alleles, which is a function of the allele frequencies in the K reference populations (e.g. $P(g_l = 0.5 \mid \boldsymbol{Z}_l = (2, 2), \boldsymbol{F}) = 2F_{2l}(1 - F_{l2})$), and $P(\boldsymbol{Z}_l \mid \boldsymbol{q})$ is the probability of the pair of origins that depends on the admixture proportions $\boldsymbol{q}$ (e.g. $P(\boldsymbol{Z}_l = (0, 0)) = q_0^2$).

For $P(f_l \mid \boldsymbol{F})$, the prior on the allele frequency in males mated to the queen, we use an informative prior based on allele frequencies in the reference populations:

$$\log\left(\frac{f_l}{1 - f_l}\right) = logit(f_l) \sim \mathcal{N}\left(\overline{logit(\boldsymbol{F}_l)}, Var(logit(\boldsymbol{F}_l))\right) \tag{6}$$

where the logit of the allele frequency in males is following a Gaussian distribution with parameters $\overline{logit(\boldsymbol{F}_l)}$ and $Var(logit(\boldsymbol{F}_l))$, the mean and variance of allele frequencies in the reference population. This prior is informative if all reference populations have similar allele frequencies and more diffuse if allele frequencies in reference populations differ greatly (Figure S1). Finally, the estimation of vector $\boldsymbol{q}$ is performed using and EM algorithm. One can note that this is similar to the supervised version of the estimation procedure of the Pritchard et al. (2000) model as the matrix of allele frequencies $\boldsymbol{F}$ is considered known a priori.

2.2 | Simulations

To evaluate the performance of the two models, we simulated data as obtained from a pool sequencing experiment. We assume these data come in the form of the reference allele counts x_l^c and sequencing depths d_l^c at each locus l, knowing the queen genotypes g_l^c and allele frequencies in the inseminating drones f_l^c. To further condition our simulations on what can be expected from real data, we exploited information available in a reference population of *Apis mellifera* (Wragg et al., 2022). These data consist of 628 European samples of haploid drones (Table S2) with genotypes available at 6,914,704 single nucleotide polymorphisms (SNPs). Wragg et al. (2022) showed that this panel is structured into three main genetic backgrounds for which unadmixed (reference) individuals can be identified, with a threshold of 99% of their genetic background being from a unique type: the **M** background (*Apis mellifera mellifera*) with 85 reference individuals, the **L** background (*Apis mellifera ligustica & carnica*) with 44 reference individuals and the **C** background (*Apis mellifera caucasia*) with 16 reference individuals (Table S3). In the simulations described below, the reference panel information used was either the allele frequencies in the three main backgrounds ($\boldsymbol{F} = (F_{lk}) \in [0,1]^{L\times 3}$, where the columns contain the allele frequencies of all L markers in genetic backgrounds **L**, **M** and **C** in this order) and/or the genotypes of the reference individuals.

2.2.1 | Independent markers

To evaluate the performance of the models proposed, a first set of simulations was performed on 1000 independent SNPs chosen to be common in all three populations and ancestry informative with respect to the **L**, **M** and **C** genetic backgrounds. To this effect, the 1000 SNPs were randomly sampled from the 722,170 SNPs out of the 6,914,704 that had a minor allele frequency (MAF) > 0.1 and a variance across genetic backgrounds > 0.1. For this first set of simulations, only the allele frequencies in the reference panel at the 1000 SNPs were used.

First, for each colony c, the proportions of the genome coming from each of the genetic backgrounds (termed *genetic ancestry* from now on) of the queen ($\boldsymbol{q}_q^c$) and the inseminating drones ($\boldsymbol{q}_d^c$) were sampled from a Dirichlet distribution:

$$\begin{aligned}\boldsymbol{q}_q^c &= \left[q_{q,L}^c q_{q,M}^c q_{q,C}^c\right] \sim Dir\left(\left[\alpha_L^c,\alpha_M^c,\alpha_C^c\right]\right)\\ \boldsymbol{q}_d^c &= \left[q_{d,L}^c q_{d,M}^c q_{d,C}^c\right] \sim Dir\left(\left[\alpha_L^c,\alpha_M^c,\alpha_C^c\right]\right)\end{aligned} \tag{7}$$

Different values were given to the α parameters to consider different levels of admixed ancestries for the colony (Table 1). Simulated genetic ancestries are represented in Figure S2.

Second, the allele frequencies of each SNP l in the cohort of inseminating drones were simulated as:

$$f_l^c \sim \frac{1}{n_d} Binomial\left(n_d \boldsymbol{F}_{l,\bullet}, \boldsymbol{q}_d^c\right) \tag{8}$$

where $\boldsymbol{F}_{l,\bullet}$ is the l-th line of the $\boldsymbol{F}$ matrix and n_d is the number of inseminating drones, here fixed at 15 (Estoup et al., 1994; Palmer & Oldroyd, 2000; Tarpy & Nielsen,2002; Tarpy et al., 2004; Bastin et al., 2017).

Third, the genotypes of the queen at a SNP l was simulated by first drawing the population of origin of each of the two allele of the queen ($\boldsymbol{Z}_l = (z_l^1, z_l^2)$) from a multinomial distribution with parameter $\boldsymbol{q}_q^c$. The genotypes of the queen was finally obtained as $g_l^c = \frac{a_{l1}^c + a_{l2}^c}{2}$ where:

$$\begin{cases} a_{l1}^c \sim Bernoulli\left(F_{l,z_l^1}\right) \\ a_{l2}^c \sim Bernoulli\left(F_{l,z_l^2}\right) \end{cases} \tag{9}$$

Finally, pool sequencing data were simulated as

$$x_l^c \sim Binomial\left(\frac{f_l^c + g_l^c}{2}, d^c\right) \tag{10}$$

where d_c is the sequencing depth, which was fixed at 30 unless otherwise specified in the Results section.

2.2.2 | Linked markers

Pool sequencing experiments provide information on a large number of markers distributed throughout the genome. To evaluate the performance of the models in realistic conditions for the distribution of allele frequencies and the genetic structure, a second set of simulations was performed using individual genotypes of 628 individuals from the diversity panel previously described in Wragg et al. (2022) and used beforehand to define reference genetic backgrounds. First, individuals were clustered into seven groups, of all potential combinations of admixture between the three genetic backgrounds, using hard thresholds on their initial vectors of genetic ancestries estimated with ADMIXTURE (Alexander et al., 2009) (Figure S3).

It has been said that honeybee queens are inseminated by 15 drones on average (Estoup et al., 1994; Palmer & Oldroyd, 2000;

TABLE 1 Simulated genetic ancestries for queen and drones under Dirichlet distribution

Number of simulated colonies	Queen genetic ancestry	Dirichlet alpha parameters for queen	Drones genetic ancestry	Dirichlet alpha parameters for drones
100	LMC	10,10,10	LMC	10,10,10
40/30/30	L__/_M_/__C	(10,0.5,0.5)/(0.5,10,0.5)/(0.5,0.5,10)	LMC	10,10,10
40/30/30	L__/_M_/__C	(10,0.5,0.5)/(0.5,10,0.5)/(0.5,0.5,10)	L__/_M_/__C	(10,0.5,0.5)/(0.5,10,0.5)/(0.5,0.5,10)
100	LM_	10,10,0.5	LMC	10,10,10
100	LM_	10,10,0.5	LM_	10,10,0.5
100	L__	10,0.5,0.5	L__	10,0.5,0.5
50/50	L__/_M_	(10,0.5,0.5)/(0.5,10,0.5)	L__/_M_	(10,0.5,0.5)/(0.5,10,0.5)
100	_MC	0.5,10,10	LMC	10,10,10
100	_MC	0.5,10,10	_MC	0.5,10,10
100	_M_	0.5,10,0.5	_M_	0.5,10,0.5
50/50	_M_/__C	(0.5,10,0.5)/(0.5,0.5,10)	_M_/__C	(0.5,10,0.5)/(0.5,0.5,10)
100	L_C	10,0.5,10	LMC	10,10,10
100	L_C	10,0.5,10	L_C	10,0.5,10
100	__C	0.5,0.5,10	__C	0.5,0.5,10
50/50	L__/__C	(10,0.5,0.5)/(0.5,0.5,10)	L__/__C	(10,0.5,0.5)/(0.5,0.5,10)

Note: For each of the 15 scenarios designed for simulations we present the number of simulated colonies, the queen's genetic ancestry in term of *Apis m. ligustica & carnica* **L**, *Apis m. mellifera* **M** and *Apis m. caucasia* **C**, the associated Dirichlet alpha vectors, and the same information for the inseminating drones.

Tarpy & Nielsen, 2002; Tarpy et al., 2004; Bastin et al., 2017). Therefore, each colony was simulated by sampling haploid genotypes of 17 individuals. Two were united to create the genotype of the queen (replacing step (9) above) and the remaining 15 were used as inseminating drones under different scenarios of admixture between the three populations, replacing step (8). Then pool sequencing data x_i^c was simulated as in (10). The simulated scenarios are the same as for independent markers, despite that only 20 colonies are simulated per scenario because of sampling limitation due to the restricted number of individuals to select from. As an example, let us describe the sampling of individuals when the queen of the colony is **L** genetic ancestry and the inseminating drones are **LMC** genetic ancestry. The two individuals to make the queen were sampled from the group of 'pure' **L** and the 15 inseminating drones were sampled from all the possible groups, as their combination will create a mixture of genetic backgrounds.

2.3 | Evaluation of statistical models

2.3.1 | Genetic ancestry

For each colony and for each set of simulations, the queen genetic ancestry q^c was estimated using the AM. For independent marker simulations, the estimates were compared to the true simulated value. For linked marker simulations, they were compared to the estimates obtained by running ADMIXTURE (Alexander et al., 2009) on the simulated queen genotypes. All simulated colonies were analysed jointly with the AM and thereafter clustered into seven groups based on their ancestry vectors. Hence, each cluster was a group of colonies with homogeneous genetic ancestry.

2.3.2 | Genotype reconstruction

The HPM was used to reconstruct the queen genotypes, within each of the ancestry clusters described above, in the linked marker simulations. Criteria for evaluating the model were as follows:

- the genotyping error rate measured as the proportion of errors in the reconstructed genotypes among all markers. We measured the genotyping error rate for different calling probability thresholds (see Results).
- the calibration of the posterior genotype probabilities. For each locus and each simulated colony, the HPM provides the posterior probabilities of the three possible genotypes. Because in the simulations the true genotype is known, we can evaluate in which proportion of the simulations (π) a genotype with posterior probability P is the true genotype. If the model is perfectly calibrated $\pi = P$. Hence, the calibration of the model was measured as

$$AUC = \int_0^1 | P - \pi | \tag{11}$$

In practice, we estimated π by grouping genotype probabilities in bins of size 0.05.

2.4 | Validation on experimental data

2.4.1 | Data set

In order to evaluate the performance of the genotyping by pool sequencing approach, we produced a new data set where colonies

were on the one hand sequenced in pools and on the other hand individual drones were sampled and sequenced. Thirty-four colonies, present at an experimental apiary and representing the diversity of French honeybee populations, were sampled in 2016. For each colony, between approximately 300 and 500 worker bees were collected and pooled for sequencing purposes. DNA extraction was performed from a blended solution of all the workers of the colony with 4 M urea, 10 mM Tris–HCl pH 8, 300 mM NaCl and 10 mM EDTA. The elution was centrifuged for 15 min at 3500*g*, and 200 µl of supernatant was preserved with 0.5 mg proteinase K and 15 µl of DTT 1 m for incubation overnight at 56 °C. After manual DNA extraction and DNA Mini Kit (Qiagen) a volume of 100 µl was used to perform pair-end sequencing. DNA sequencing was performed at the GeT-PlaGe core facility, INRAE Toulouse. DNA-seq libraries were prepared according to Illumina's protocols using the Illumina TruSeq Nano DNA HT Library Prep Kit. Briefly, DNA was fragmented by sonication, size selection was performed using SPB beads (kit beads) and adaptors were ligated to be sequenced. Library quality was assessed using an Advanced Analytical Fragment Analyzer and libraries were quantified by QPCR using the Kapa Library Quantification Kit. DNA-seq experiments were performed on an Illumina HiSeq3000/NovaSeq6000 using a paired-end read length of 2x150 pb with the dedicated reagent Kits. The final aim was to obtain approximately 30× raw sequencing data per sample. Raw reads were then aligned to the honeybee reference genome Amel-HAV3.1, Genbank assembly accession GCA_003254395.2 (Wallberg et al., 2019), using BWA-MEM (v0.7.15; Li, 2013). For pool sequencing experiments, the resulting BAM files were converted into pileup files using Samtools mpileup (Li & Durbin, 2009) with the parameters: -C 50 coefficient of 50 for downgrading mapping quality for reads with excessive mismatches, −q 20 minimum mapping quality of 20 for an alignment, −Q 20 and minimum base quality of 20, following standard protocols. This procedure was applied exclusively to the 6,914,704 single nucleotide polymorphisms (SNPs) identified in (Wragg et al. (2022)) as polymorphic in the European honeybee population. The pileup files were interpreted by the PoPoolation2 utility mpileup2sync (Kofler et al., 2011) for the Sanger Fastq format with a minimum quality of 20. They were finally converted to allele counts and sequencing depth files using a custom-made script. In addition, for each of these 34 colonies, four male offspring of the queen, genetically equivalent to queen gametes, were individually sequenced as in Wragg et al. (2022) (Table S4). In order to reduce computation time, this analysis was performed on a subset of about 50,000 markers. These markers were selected following the criteria: (1) maximum of two polymorphic sites within a 100 base pair window, (2) only one representative marker per linkage disequilibrium block with r^2 higher than 0.8, (3) variance between allele frequencies in the different genetic backgrounds higher than zero, to allow for population identification and (4) sampled so that the MAF follows a uniform distribution. This selection led to exactly 48,334 markers in the experimental data set.

2.4.2 | Genetic ancestry

For each colony, using pooled sequencing data, the queen genetic ancestry q^c was estimated using the AM as described above. For the male offspring data, for each colony two ways to estimate the genetic ancestries were considered:

1. By averaging the genetic ancestry vectors of the four males as estimated by ADMIXTURE.
2. By first reconstructing the queen genotypes from the male offspring data (see below) and then analysing the resulting genotypes with ADMIXTURE.

2.4.3 | Genotype reconstruction

For pool sequencing data, queen genotypes were reconstructed using the HPM, considering the 34 colonies jointly. For the male offspring data, queen genotypes were reconstructed by first estimating the genotype probabilities at each locus from individual data of the four sequenced male offspring. Our goal is to reconstruct the genotype of a parent at a locus (g_l) (here the queen) from the haploid genotypes of a set of n_g gametes (here the male offspring). Let R be the random variable of the number of reference alleles observed in the offspring and assume that there is a per allele sequencing error equal to ϵ, then the genotype likelihoods can be computed from the sampling distributions:

$$\begin{cases} R \mid g_l = 0 \sim Binomial(n_g, \epsilon) \\ R \mid g_l = 0.5 \sim Binomial(n_g, 1/2) \\ R \mid g_l = 1 \sim Binomial(n_g, 1-\epsilon) \end{cases} \quad (12)$$

To compute the genotype posterior probabilities when r_l reference alleles are observed at a locus, we specify a uniform prior on the three possible genotypes, so that $P(g_l = x | R = r_l) = P(R = r_l | g_l = x) / \sum_{x' \in [0,0.5,1]} P(R = r_l | g_l = x')$. For our application, we fixed $\epsilon = 10^{-3}$ and n_g is four as described above. Because we have only four drones per colony in this data set, there is still some uncertainty in the genotypes of the queen. For example the highest posterior probability achievable for a genotype with $n_g = 4$ is ≈ 0.94. When comparing the genotypes reconstructed from the offspring data and from the pool sequencing data, the concordance between the two approaches has to be measured with respect to what is expected between the true genotypes of the queen and the one reconstructed from noisy data (either offspring or pool sequencing). Unfortunately, we do not know the true genotypes of the queen in our data set. However, we can measure the concordance between the genotypes reconstructed with four male offspring to the true genotypes of the queen using data from Liu et al. (2015). In this data set, genotypes of the queen and of 13 to 15 offspring is available for three

colonies. With that many offspring, the genotypes of the queen can be reconstructed with certainty and be compared to the one obtained by down-sampling the data to four offspring per colony. For each of the three colonies in Liu et al. (2015), we called the offspring genotypes at the set of markers present in the diversity panel, reconstructed the queen genotypes using (i) all offspring ($n_g = 15$ or 13) and (ii) a 100 randomly down-sampled data sets consisting of four offspring only.

3 | RESULTS

In this study, we developed statistical models to estimate genetic ancestry and queen genotypes from pool sequencing data from workers of the colony. Simulations, from independent and linked markers, were performed to evaluate the performance of our models in terms of queen genetic ancestry inference and genotypes reconstruction. The scenarios are described in Figure S2. Moreover, these models were applied to an experimental data set composed of both pool sequenced data and individual male offspring of the queen.

3.1 | Validation on simulations

3.1.1 | Genetic ancestry

For independent markers, correlations between simulated genetic ancestries and estimated genetic ancestries using the AM ranged between 0.88 and 0.9 depending on the genetic background. For linked markers, correlations between genetic ancestries estimated using ADMIXTURE (Alexander et al., 2009) on the queen genotypes simulated from real data and estimated by the AM ranged between 0.93 and 0.95 depending on the genetic background (Figure 1).

In addition to the 15 scenarios listed, we also estimated genetic ancestry by the AM in scenarios in which queen and drones had divergent ancestries (Table S1). We observed that shifting from the initial hypothesis that queen and drones come from the same origin led to highly biased genetic ancestry estimations with the AM (Figure S4). It should be noted that the statistical model under the AM is based on the assumption that markers are independent. To match this assumption a subset of 1000 markers, rather than the whole genome, was used to estimate genetic ancestry for simulations with linked markers. These results show that the AM outputs accurate genetic ancestry estimates. Moreover, it shows high agreement to a standard population genetics model such as ADMIXTURE (Alexander et al., 2009), under the assumption that queen and drones are of same origin. The observed results confirm that using only a subset of ancestry informative markers, here 1000 from the whole genome, is sufficient to accurately estimate genetic ancestries using the AM.

3.1.2 | Genotype reconstruction

One major assumption of the HPM is that colonies within the population are of homogeneous genetic ancestries. Therefore, using simulations for linked markers across the whole genome, we compared and clustered all the simulated colonies based on their genetic ancestries estimated by the AM. In our study, we assume that colonies come from a mixture of three main genetic backgrounds (as described in Wragg et al. (2022)), we thus clustered our simulated colonies in seven groups from pure to hybrid genetic types (Figure 2).

Thereafter, to evaluate queen genotypes reconstruction performance we implemented the HPM on our seven groups of homogeneous colonies for linked markers. As the HPM does not make the assumption of independence of markers the inference could be performed on the whole genome, approximately 7 million markers. Across all simulations and all scenarios, we observed a strong correlation between the rate of genotype agreement between simulated and estimated genotypes and the associated estimated genotype probability. The rate of agreement is a measure of the genotypes identity between simulated and modelled estimates. Genotypes inferred with a high probability are often correctly predicted by the HPM, whereas genotypes inferred with a low probability are often wrongly predicted by the HPM, making genotypes with a probability close to 0.5 the hardest to infer precisely. The calibration of the HPM for genotypes reconstruction, measured as the Area under the Curve between agreement rates and probabilities, was 0.055 (Figure 3a). AUC ranges between 0, for perfect correlation, and 0.5, for completely imperfect correlation. A large proportion of the markers have genotype probabilities close to zero or to one, making the genotypes drawn for these markers almost certain (Figure 3a). As expected, we observed that the genotyping error rate decreases slightly when the best genotype probability threshold increases. Filtering for markers with higher best genotype probability leads to more accurate genotypes reconstruction. However, such filtering is accompanied with a small reduction in genotype call rate. For example if no filtering on best genotype probability is applied, 100% of the genome will be reconstructed with an average genotyping error rate of 4%. If filtering for markers with best genotype probabilities above 0.9 is applied about 95% of the whole genome will be reconstructed with an average genotyping error rate as little as 2% (Figure 3b). Additionally, we observed that genotyping error rate increased when the MAF threshold increased meaning that filtering on MAF might cause an increase in genotyping error, accompanied by a drastic reduction in genotype call rate (Figure 3c). Minor allele frequency and best genotype probability are highly linked as markers with low MAF tend to be easier to infer with a high probability. In our simulations, a large proportion, more than 50%, of the whole genome is composed of markers with MAF below 0.05. Yet applying a filter on best genotype probability does not seem to highly impact the distribution of MAF on the whole genome (Figure S5). Rather than

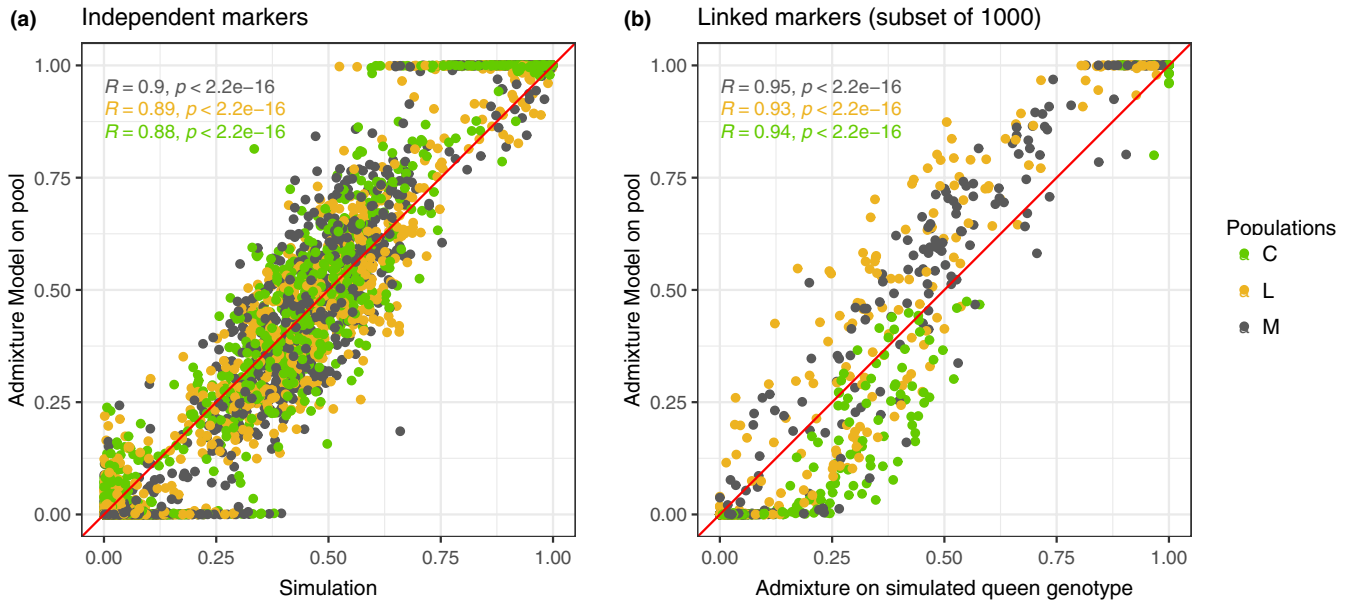

FIGURE 1 Genetic ancestry comparison. Regression of the genetic ancestry vectors (one value for each of the genetic background for every simulated colony) estimated by the Admixture Model (AM) on pool sequencing experiment against genetic ancestry vectors for all colonies (15 * 100) simulated for independent markers (a) or estimated by ADMIXTURE for all colonies (20 * 15, the number of simulated colonies is lower due to limitation in the number of individuals to sample from in the real data set) simulations for linked markers (subset of 1000) (b). The red line represents the regression with intercept 0 and slope 1, meaning perfect agreement between the two estimates. Values for Spearman's rank correlations between ancestry vectors are shown in the top left corner for each of the three genetic ancestries yin yellow *Apis m. ligustica & carnica* L, in grey *Apis m. mellifera* M and in green *Apis m. caucasia* C.

filtering on MAF we suggest to filter on best genotype probability, for example equal to or greater than 0.95. Indeed, such filtering will improve the queen genotypes reconstruction accuracy without heavily impacting the allele frequency distribution of the markers genotyped on the whole genome. In fact, we observed that genotyping error, on the whole genome and without filtering, is on average about 4% (Figure 3d). After applying a filter on best genotype probability equal to or greater than 0.95 genotyping error becomes on average as low as about 2%.

These results show average estimates across all simulation scenarios and colonies after grouping based on genetic ancestry. Detailed results for calibration and genotyping error are presented Figure S6.

To conclude, using simulations we confirm that the statistical model AM performs similarly to ADMIXTURE leading to highly accurate genetic ancestry inference. A small set of markers, as low as 1000 in our example where genetic ancestry differentiation is strong, seems sufficient to accurately estimate genetic ancestry with the AM. Using simulation of linked markers across the whole genome, we confirmed that the HPM reconstructed queen genotypes with high accuracy. Furthermore, we inferred the impact of MAF and best genotype probability thresholds on the genotype call rate and the associated genotyping error rate. Our advice is to filter on best genotype probability equal to or greater than 0.95 to reduce genotyping error. Such filtering goes without drastic loss of predicted markers and while preserving allele frequency distribution across the genome.

3.2 | Application on experimental data

To further evaluate the performance of the AM and HPM, we analysed real data on honeybee colonies for which four individual drones and a pool of workers were sequenced (see Materials and Methods).

3.2.1 | Genetic ancestry

For each colony, the genetic ancestry of the queen was estimated either from the group of male offspring or from the pool of workers. Genetic ancestries from worker pool sequences were estimated using the AM. For male offspring, it was estimated with ADMIXTURE (Alexander et al., 2009) either using the male offspring directly (admix_males) or from the genotypes of the queen reconstructed using male offspring (admix_proba), as described in the Material and Methods section. Both approaches gave virtually equal genetic ancestry values from ADMIXTURE, with a Mean Squared Difference (MSD) of 1.4×10^{-3} (standard deviation 1.1×10^{-3}). When comparing these estimates with those obtained on the worker pool sequences, MSD were slightly higher with 0.024 and 0.026 with standard errors of 0.025 and 0.021 for admix_males and admix_proba respectively (Table 2). For the 34 experimental colonies, most of the genetic ancestries estimated using either queen reconstructed genotypes from worker pool sequencing data, male offspring, or using individual sequencing of male offspring, gave very similar q vectors (Figure S7).

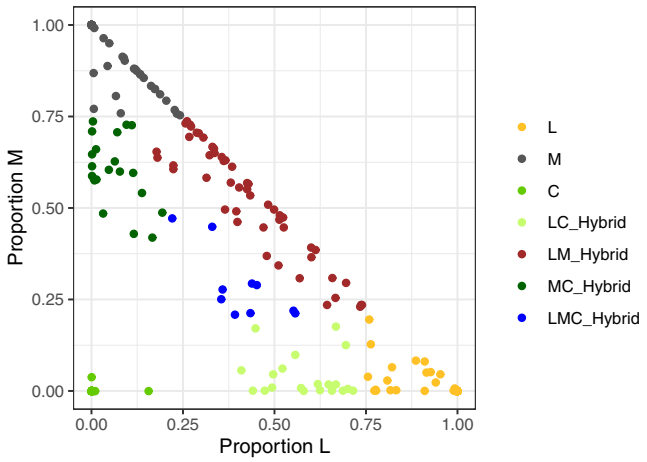

FIGURE 2 Genetic ancestries for the simulated colonies as estimated by the Admixture Model (AM). Representation in two dimensions of genetic ancestry (*x* and *y* axis give the genetic ancestry value in two of the three populations of honeybee in our data set) for all the colonies (20 * 15) simulated for linked markers after estimation of their genetic ancestry vectors by the AM model. Individuals can be grouped by genetic ancestry. Here we decided on seven groups, each in a different colour, in yellow *Apis m. ligustica* + *Apis m. carnica* L, in grey *Apis m. mellifera* M, in green *Apis m. caucasia* C, in light green hybrids *Apis m. ligustica* + *Apis m. carnica* and *Apis m. caucasia*, in brown hybrids *Apis m. ligustica* + *Apis m. carnica* and *Apis m. mellifera*, in dark green hybrids *Apis m. mellifera* and *Apis m. caucasia* and in blue the three ways hybrids

3.2.2 | Genotype reconstruction

To validate queen genotypes reconstruction from worker pool sequence on our experimental data set, we used publicly available data from Liu et al. (2015). Three colonies for each of which both the queen and 13 to 15 male offspring were individually sequenced were available. Of our 50,000 selected markers, only 14,988 were available, as polymorphic SNPs, on the data set from Liu et al. (2015). This reduction in the number of markers available for the analysis can be explained as the population used for SNP calling was composed of fewer individuals from a unique and uniform origin in the data set from Liu et al. (2015). To validate genotypes reconstruction, we compared (i) queen genotypes reconstructed from worker pool sequence and queen genotypes reconstructed on probabilities from four male offspring (pool/offspring) on the experimental data set; (ii) genotypes from individually sequenced queens and queen genotypes reconstructed on probabilities from four male offspring (queen/offspring) and (iii) pairs of queen genotypes reconstructed on probabilities from four independent male offspring (offspring/offspring) on the data set from Liu et al. (2015). Genotype concordance was on average 0.94 (standard deviation 0.03), 0.96 (standard deviation 0.01) and 0.92 (standard deviation 0.01) for pool/offspring, queen/offspring and offspring/offspring respectively (Figure 4).

The highest concordance is observed between the actual queen genotypes and its reconstruction from four male offspring. Queen genotypes reconstruction from pooled workers and from male offspring seem to present similar concordance than when pairs of independent male offspring are compared. A few colonies show some discrepancy between genetic ancestry estimates. Indeed, we observed a genetic ancestry from worker pool sequences mostly divergent from the estimates based on males, despite having high concordance between queen genotypes reconstruction. This can be due to multiple possible limitations. First, the AM might struggle when it comes to disentangling queen genotypes from cohort of inseminating drones in the worker pool sequencing data. Second, sampling only four male offspring might not be sufficient to accurately represent the queen genetic ancestry. Third, there might be a genetic contradiction between the queen that produced the male offspring and the one that produced the workers (e.g. in case of requeening). Finally, there might be a bias in the markers used for the AM. However, this validation confirms that queen genotypes reconstructed using worker pool sequencing data perform as well as individually sequenced multiple male offspring. Additionally we showed, on the data from Liu et al. (2015), that increasing the number of male offspring individually sequenced improved genotype concordance quite substantially (Figure S8). Eight and 10 male offspring show a concordance between reconstructed and real genotypes close to one.

To summarise, the difference between genetic ancestries estimated from male offspring or worker pool sequencing data, using the AM, was small. Queen genotypes reconstruction from worker pool sequence data was in agreement with queen genotypes reconstructed from male offspring. This value was slightly lower than when comparing queen reconstructed genotypes from male offspring with the real queen genotypes and slightly higher than when comparing queen reconstructed genotypes from different sets of male offspring of the same queen. The HPM on worker pool sequencing data is an accurate alternative to individually sequencing a limited number of male offspring of the queen when one wants to access the queen genotypes.

4 | DISCUSSION

The past decade has seen the growth of the molecular genomics era with the development of new sequencing platforms and technologies, one of them being pool sequencing. This technology allows for the combination of multiple individuals in one sequencing experiment, reducing drastically preparation and sequencing costs and therefore making high-depth sequencing available for a wide variety of samples. Traditionally, pool sequencing is used to perform analyses on multiple individuals from a population. It has also been frequently used, in plant and animal, for genotyping pools of individuals from the same population (Arca et al., 2020; Johnston et al., 2013). Additionally, pool sequencing might be of interest when group level information is desired as, for example in the context of eusocial organisms. In such a case, the pool will represent a meta-individual of the colony rather than a population. One pitfall of using such a sequencing method it that the outcome is in the form of allele read counts and sequencing depths rather than diploid genotype

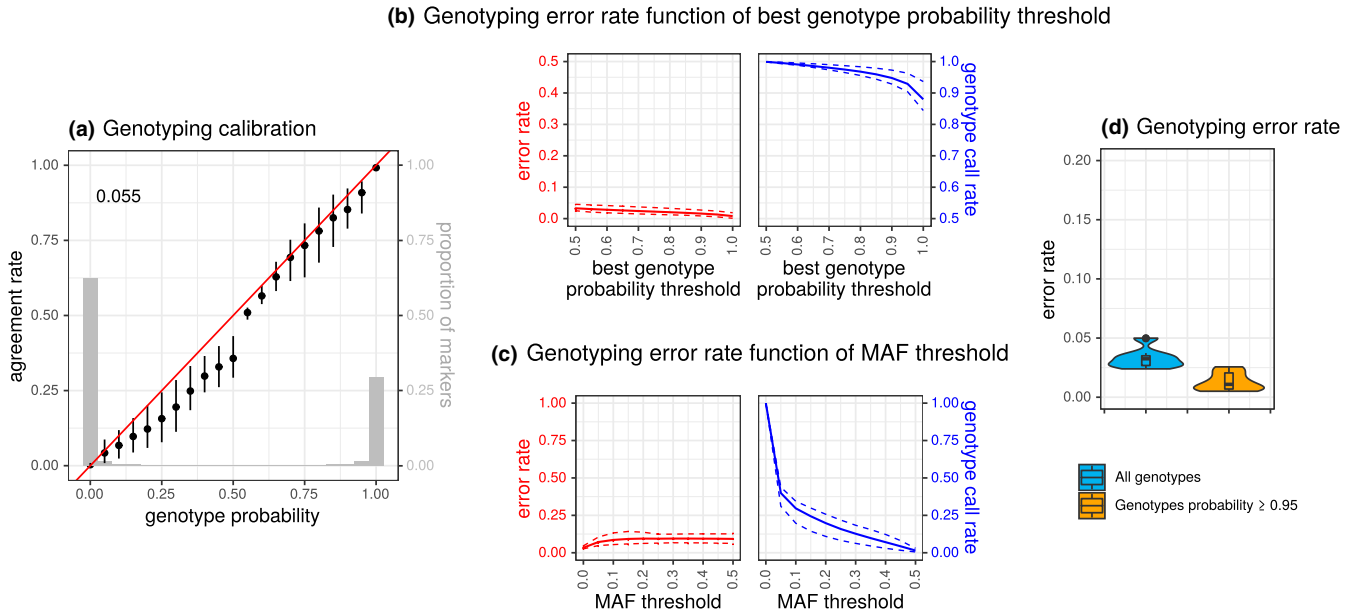

FIGURE 3 Queen genotypes reconstruction. For simulations from linked markers across the whole genome, all values are averaged across all colonies for each scenarios. (a) Genotyping calibration, each point represents the genotype agreement rate per genotype probability value with bars representing the quantiles 95% to 5%. The red line is the regression with intercept 0 and slope 1. Value for Area Under the Curve between perfect and observed calibration is shown in the top left corner. The grey histogram represents the proportions of markers in each bin of genotype probability. (b) Genotyping error rate as function of best genotype probability threshold. In red, the solid line represents the average genotyping error rate across all scenarios as a function of the best genotype probability, the dotted lines are the quantiles 95% and 5%. In blue, the solid line represents the average genotype call rate, across all scenarios, if thresholds were applied on the best genotype probability, the dotted lines are the quantiles 95% and 5% for genotype call rate. (c) Genotyping error rate as function of minor allele frequency (MAF) threshold. As for (b), in red, the solid line represents the average genotyping error rate across all scenarios as a function of the MAF threshold, the dotted lines are the quantiles 95% and 5% for genotyping error rate. In blue, the solid line represents the average genotype call rate, across all scenarios, as a function of the MAF threshold, the dotted lines are the the quantiles 95% and 5% for genotype call rate. (d) Violin plot of the genotyping error, for all markers or filtering on best genotype probability equal to or greater than 0.95

TABLE 2 Genetic ancestry Mean Squared Difference between data and models

	Queen from males vs. males	Queen from pool vs. males	Queen from pool vs. queen from males
Model_i	Admix_proba	AM	AM
Model_j	Admix_males	Admix_males	Admix_proba
Min	1.36E-05	2.94E-04	1.35E-03
Mean	1.43E-03	0.024	0.026
Median	1.15E-03	0.014	0.02
Max	4.19E-03	0.085	0.082
SD	1.16E-03	0.025	0.021

Note: Mean Squared Differences between genetic ancestry estimated by different data and different models on the experimental colonies. Minimum, average, median, maximum and standard deviations are calculated for each combination.

observations. From allele read counts and sequencing depths, it is possible to estimate allele frequencies, either for reference versus alternative alleles or for ancestral versus derived alleles thus giving potential information on the evolutionary history of the pool. However, this datatype is not in the standard format for downstream analyses.

So far only a few programs can handle allele counts data, for example PoPoolation (Kofler et al., 2011) and CRISP (Bansal, 2010) for SNP calling, Plink (Chang et al., 2015; Purcell et al., 2007) and the R package poolfstat (Gautier et al., 2022; Hivert et al., 2018) for population genetics or GEMMA (Zhou & Stephens, 2012) and LDAK (Speed et al., 2020) for association studies. However, when considering eusocial insects from the same colony as a pool we might break underlying assumptions made by these models. In fact, eusocial insects present characteristics deviating from what could be expected in a standard population used for pool sequencing experiment. First,

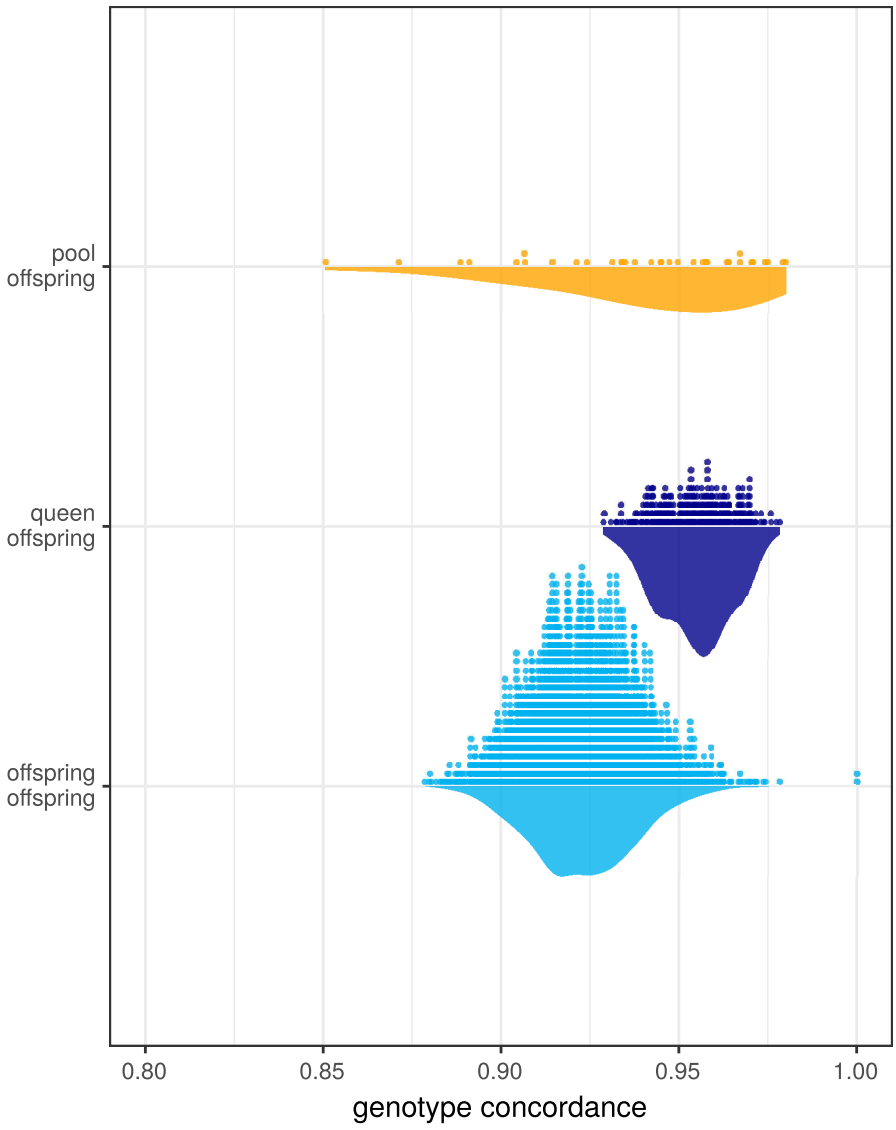

FIGURE 4 Concordance between queen genotypes reconstruction based on different data. The plot represents the concordance, only for markers after filtering for best genotype probability equal to or greater than 0.94, between (i) queen genotypes reconstructed from worker pool sequence experiments using the HPM and queen genotypes reconstructed from genotype probabilities (pool/offspring), based on four male offspring for experimental colonies, in orange (ii) queen genotypes reconstructed from genotype probabilities based on four male offspring for a 100 sampling events and actual queen genotypes from the Liu et al. (2015) (queen/offspring), in dark blue and (iii) pairs of queen genotypes reconstructed from genotype probabilities based on four male offspring for independent sets of individuals with the data from Liu et al. (2015) (offspring/offspring), in light blue. Concordance values for each test are represented as dots, top, and as density distribution, bottom

in hymenopterans, reproductive systems are often polyandric, leading to non-standard genetic relationships across individuals in the colony. Second, traits of interest are likely to be measured at the colony level. Therefore, in order to avoid computational limitations and biases that could be brought by the use of pool sequencing with non-adapted models one may want to infer individual genetic information (e.g. ancestry and genotypes) from pool genotyping.

In honeybee for instance, a colony can be considered as a polyploid organism (with two major chromosomes, coming from the queen and present in the whole population, and about 15 chromosomes coming from inseminating drones randomly distributed in workers and daughter queens). Indeed a colony is composed of haploid males and diploid females. The male offspring of the queen that can be described as 'flying gametes' as they come from queen unfertilized eggs. The worker bees descend from the mating of a queen with a cohort of about 15 inseminating drones. This leads to complex genetic relationships between individuals in the colony (Barron et al., 2001; Oxley & Oldroyd, 2010). The number of inseminating drones, as a parameter of the model, might impact computation. In fact, with our models, we make the assumption that the queen has been mated to a large number of drones, of similar genetic ancestry. It is clear that developing a method to infer the number of inseminating drones from pool sequence data would be crucial to optimize our models, especially if the number of inseminating drones is limited. The honeybee queen carries the key part of the genetic information of the colony and is the reproducing organ of the next generation, thus making it a favoured pathway for selection. In addition, honeybee populations used by breeders and beekeepers are often highly structured, with vast differences between genetically pure and highly admixed colonies. Honeybee populations have been influenced by domestication, hybridization and selection performed by beekeepers on traits usually measured at the colony level, thus making the use of pools of workers highly relevant. These features are also in favour of using *Apis mellifera* as model organism to develop statistical models to use pool sequencing data. Moreover, we also benefit from the available knowledge on the organism compared to other eusocial insects. For example we can exploit the diversity panels, such as built in Wragg et al. (2022), as prior in our models

to facilitate inference. In this context the methods developed here are expected to be easily applicable to organisms with lower level of population stratification, as should be the case for most other eusocial insects.

Here we present two statistical models to infer queen information from pool experiment data. First, the AM allows to infer queen genetic ancestries from worker pool sequencing data, knowing expected allele frequencies estimated in reference populations. The correlation between predicted and expected ancestries was high (about 0.9). Moreover, this process can be run rapidly on a small subset of markers, making it computationally efficient, and for each colony independently, allowing parallelisation. Second, the HPM allows for an accurate queen genotypes reconstruction with as little as 2% genotyping errors. This model takes advantage of the information from other colonies of the same group to complete genotypes reconstruction, making the assumption that colonies within a group are of homogeneous genetic ancestries. When genetic ancestry is unknown prior to the analysis, we suggest to first infer genetic ancestry using the AM for all the colony DNA pools of interest. It is then possible to group them based on similarities in their ancestries. Finally, one can perform genotypes reconstruction on these groups separately with the HPM. Therefore, we propose to use our statistical models sequentially to reach highly accurate queen genotypes reconstruction. To date, a common way to infer honeybee queen genotypes without manipulating and sacrificing this queen is to perform pool sequencing on multiple male offspring (Petersen et al., 2020). For this purpose Jones et al. (2020) suggests, using theoretical estimations, to sequence at least 30 individuals. This procedure requires to be able to identify and sample enough male offspring from the colony, which is not always easy depending on the season, the colony and the time available for sampling. An alternative is to individually sequence multiple male offspring. In such a case, the number of individual sequences is the limiting parameter for accurate queen genotypes reconstruction. At least eight to ten individuals are needed to accurately deduce queen haplotypes, which we cannot obtain from a worker pool experiment, and to lower the risk of incorrect genotypes reconstruction (Figure S8). Using real data, we saw that our statistical models, based on pool sequence experiment, reconstructed queen genotypes at least as well as sequencing four individual male offspring. Queen genotypes reconstruction from pool sequencing data from workers of the colony appears to be a relevant alternative, cheaper as only one sequencing experiment needs to be performed. Simulations, of independent and linked markers, and the experimental field data set concluded that we could estimate honeybee queen genetic ancestry and genotypes accurately and efficiently using our methods.

Despite the efficiency of the statistical models described in this study some limitations have been identified and further improvements can be conducted. One crucial assumption of our model is that honeybee queen and inseminating drones have similar genetic ancestry, which is often true with natural breeding. However, this assumption does not necessarily hold with artificial insemination, in extremely controlled breeding environments, or even when the breeding environment is 'polluted' by an unexpected genetics. In fact, when queen and inseminating drones have highly divergent ancestries our models will estimate biased genetic ancestry and queen genotypes (Figure S4). Additional external information is necessary to account for heterogeneity in the origin of breeding parents of the pool. One way to do so would be by implementing a two-step reconstruction algorithm focusing first on the inseminating drones' allele frequencies. For example we could use information on the breeding practices or from sampling drones from the environment as a representation of the mating cohort. Once information on the mating cohort is available it can be easily implemented in our model by adapting the prior in Equation (6). In this study, we performed simulations of pool experiments with a sequencing depth of 30×. In practice, and especially in the context of non-model organisms, such sequencing depth might be difficult to reach either due to sequencing cost or to genetic material availability. Therefore, we also tested the simulations with a depth of 10 or 100. We compared our results in terms of genotyping error rate and genotype call rate on the genome after filtering for best genotype probability. In Figure S9, we can see that increasing sequencing depth from 10 to 30 improved the accuracy of genotype inference and the genotype call rate. At high sequencing depth, 100, we observed a higher genotyping error rate overall and limited improvement in the fraction of markers inferred with certainty. It is likely that some level of heterogeneity within the groups used to reconstruct queen genotypes led to wrong decisions at higher sequencing depth. Increasing sequencing depth seems to cause a higher sensitivity to the hypothesis of homogeneous populations by the statistical HPM. One option to reduce this impact would be to group colonies based on their genetic ancestries to a more refined scale. Indeed, further developments in the HPM could allow to take into account a level of heterogeneity in the population to reduce the sensitivity of the model to the homogeneity assumption.

We observed that the HPM performed better, had a lower genotyping error rate, if inferred genotypes along the genome were filtered based on their certainty, measured as a probability. In our simulations, such filtering did not affect the allele frequency distribution and reduced only slightly the number of inferred markers along the genome while reducing genotyping error rate (Figure S5). This filtering procedure can then be followed by an imputation step to assign markers that had lower probabilities. Also, taking into consideration linkage disequilibrium along the genome to refine the genotypes inferred by the HPM, could be adapted in our statistical model. Such developments would benefit from identification of haplotype blocks in the honeybee genome (Saelao et al., 2020; Wallberg et al., 2017; Wragg et al., 2016, 2022) tagging the different *Apis mellifera* populations. An efficient strategy would be to reconstruct queen genotypes with the HPM, filter on genotype probability to retain only markers from which reconstruction is satisfying and then apply an imputation step taking into account known haplotype blocks and LD between markers.

To conclude, colony pool sequencing data can be used to infer queen genetic ancestries when knowing allele frequencies in reference populations present in the environment are available. Using pool sequencing data across multiple colonies of homogeneous genetic

ancestry in which queen and inseminating drones come from a similar origin, it is possible to reconstruct honeybee queen genotypes accurately. Such genotypes are valuable for population genetics analyses and association studies with mainstream models currently available, and genetic ancestry estimates can be useful for selective breeding purposes. Additional developments to take into consideration some level of heterogeneity, discrepancy of origins between queen and inseminating drone cohort and linkage disequilibrium along the genome will help further increase genotypes reconstruction accuracy. The statistical models described in the study have been designed within the context of eusocial hymenopterans but tested solely on *Apis mellifera*. Such models could be tested within the framework of studies on other eusocial species with multiple mating of a single queen (Micheletti & Narum, 2018) and having known genetic diversity panels to estimate prior for allelic frequencies.

AUTHOR CONTRIBUTIONS

A.V., B.S., F.M., B.B., Y.L.C. and A.D. designed the data collection. F.M., B.B. and Y.L.C. performed the data collection. K.T. and E.L. performed the laboratory preparation of the samples, DNA extraction, library preparation and sequencing. B.S. developed the methods and wrote the models. S.E.E. designed and performed the simulations and model comparisons. S.E.E., B.S. and A.V. designed the study, interpreted the results and drafted and reviewed the manuscript. F.M., Y.L.C., L.G. and A.D. contributed to the discussion. All authors have read and approved the manuscript.

ACKNOWLEDGEMENTS

This study was performed with the support of the ITSAP team for the maintenance of the honeybee colonies and the data collection, the sequencing platform GeT-PlaGe, Toulouse (France), a partner of the France Génomique National Infrastructure, funded as part of 'Investissement d'avenir' program managed by Agence Nationale pour la Recherche (ANR-10-INBS-09), for the sequencing and especially Olivier Bouchez. Bioinformatics analyses were performed on the computing facility Genotoul. This research was funded by the Ministère de l'Agriculture de l'Agroalimentaire et de la Forêt within the framework of MOSAR RT 2015-776 project and the Ministère de l'Agriculture de l'Agroalimentaire et de la Forêt and Investissement d'avenir (call ICF2A) for BeeStrong Ps2A project. Thanks to Claude Chevalet for the initial discussions on the idea, the members of the BeeStrong project, Florence Phocas and François Guillaume, for their contributions to the discussion during the development of this study. Thanks to the two reviewers for their constructive comments that highly improved the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

Statistical models are available on GitHub https://github.com/BertrandServin/beethoven and scripts developed to perform the simulation on GitHub https://github.com/seynard/pool2geno. The vcf file containing the filtered SNPs and the complete diversity panel can be found in (Wragg et al., 2022). The list of 628 individuals used in this study as well as the list of reference individuals and individuals (male offspring) used for validation can be found in the Tables S2–S4, together with their accession names. The parameter files for simulations as well as the pool sequencing experiment data for the 34 colonies used for validation can be found on Data INRAE, https://doi.org/10.57745/mh1wfp. The external data set used for validation can be found in (Liu et al., 2015).

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://doi.org/10.57745/mh1wfp.

ORCID

Sonia E. Eynard https://orcid.org/0000-0002-8609-5869
Alain Vignal https://orcid.org/0000-0002-6797-2125
Kamila Canale-Tabet https://orcid.org/0000-0002-5760-7773
Yves Le Conte https://orcid.org/0000-0002-8466-5370
Fanny Mondet https://orcid.org/0000-0002-7737-0101
Bertrand Servin https://orcid.org/0000-0001-5141-0913

REFERENCES

Alexander, D., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109

Arca, M., Gouesnard, B., Mary-Huard, T., Le Paslier, M., Bauland, C., Combes, V., Madur, D., Charcosset, A., & Nicolas, S. (2020). Genome-wide snp genotyping of dna pools identifies untapped landraces and genomic regions that could enrich the maize breeding pool. *bioRxiv*. https://doi.org/10.1101/2020.09.30.321018

Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of dna pools. *Bioinformatics*, *26*(12), i318–i324. https://doi.org/10.1093/bioinformatics/btq214

Barron, A., Oldroyd, B., & Ratnieks, F. (2001). Worker reproduction in honey-bees (apis) and the anarchic syndrome: A review. *Behavioral Ecology and Sociobiology*, *50*, 199–208. https://doi.org/10.1007/s002650100362

Bastin, F., Savarit, F., Lafon, G., & Sandoz, J. (2017). Age-specific olfactory attraction between western honey bee drones (*Apis mellifera*) and its chemical basis. *PLoS One*, *12*, e0185949. https://doi.org/10.1371/journal.pone.0185949

Brascamp, E., & Bijma, P. (2014). Methods to estimate breeding values in honey bees. *Genetics Selection Evolution*, *46*(1), 53. https://doi.org/10.1186/s12711-014-0053-9

Bubniĉ, J., Mole, K., Prešern, J., & Moškrič, A. (2020). Non-destructive genotyping of honeybee queens to support selection and breeding. *Insects*, *11*(12), 896. https://doi.org/10.3390/insects11120896

Chang, C., Chow, C., Tellier, L., Vattikuti, S., Purcell, S., & Lee, J. (2015). Second-generation plink: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Estoup, A., Solignac, M., & Cornuet, J. (1994). Precise assessment of the number of patrilines and of genetic relatedness in honeybee

colonies. *Proceedings of the Biological Sciences*, *258*(1351), 1–7. https://doi.org/10.1098/rspb.1994.0133

Gautier, M., Vitalis, R., Flori, L., & Estoup, A. (2022). f-Statistics estimation and admixture graph construction with Pool-Seq or allele count data using the package poolfstat. *Molecular Ecology Resources*, *22*, 1394-1416. https://doi.org/10.1111/1755-0998.13557

Gregory, P., & Rinderer, T. (2004). Non-destructive sources of DNA used to genotype honey bee (*Apis mellifera*) queens. *Entomologia Experimentalis et Applicata*, *111*, 173–177. https://doi.org/10.1111/j.0013-8703.2004.00164.x

Hivert, V., Leblois, R., Petit, E., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics*, *210*(1), 315–330. https://doi.org/10.1534/genetics.118.300900

Johnston, S., Lindqvist, M., Niemelä, E., Orell, P., Erkinaro, J., Kent, M., Lien, S., Vähä, J. P., Vasemägi, A., & Primmer, C. (2013). Fish scales and snp chips: Snp genotyping and allele frequency estimation in individual and pooled dna from historical samples of atlantic salmon (*Salmo salar*). *BMC Genomics*, *14*(1), 1471–2164. https://doi.org/10.1186/1471-2164-14-439

Jones, J., Du, Z., Bernstein, R., Meyer, M., Hoppe, A., Schilling, E., Ableitner, M., Juling, K., Dick, R., Strauss, A., & Bienefeld, K. (2020). Tool for genomic selection and breeding to evolutionary adaptation: Development of a 100k single nucleotide polymorphism array for the honey bee. *Ecology and Evolution*, *10*(13), 6246–6256. https://doi.org/10.1002/ece3.6357

Kofler, R., Pandey, R., & Schlotterer, C. (2011). Popoolation2: Identifying differentiation between populations using sequencing of pooled dna samples (pool-seq). *Bioinformatics*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwamem *arXiv preprint arXiv:1303.3997*.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liu, H., Zhang, X., Huang, J., Chen, J., Tian, D., Hurst, L., & Yang, S. (2015). Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. *Genome Biology*, *16*(1), 15. https://doi.org/10.1186/s13059-014-0566-0

Micheletti, S., & Narum, S. (2018). Utility of pooled sequencing for association mapping in nonmodel organisms. *Molecular Ecology Resources*, *18*(4), 825–837. https://doi.org/10.1111/1755-0998.12784

Oxley, P., & Oldroyd, B. (2010). The genetic architecture of honeybee breeding, *39*, 83–118.

Palmer, K., & Oldroyd, B. (2000). Evolution of multiple mating in the genus apis. *Apidologie*, *31*(2), 235–248. https://doi.org/10.1051/apido:2000119

Petersen, G., Fennessy, P., Van Stijn, T., Clarke, S., Dodds, K., & Dearden, P. (2020). Genotyping-by-sequencing of pooled drone dna for the management of living honeybee (*Apis mellifera*) queens in commercial beekeeping operations in New Zealand. *Apidologie*, *51*, 545–556. https://doi.org/10.1007/s13592-020-00741-w

Pritchard, J., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M., & Sham, P. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Saelao, P., Simone-Finstrom, M., Avalos, A., Bilodeau, L., Danka, R., de Guzman, L., Rinkevich, F., & Tokarz, P. (2020). Genome-wide patterns of differentiation within and among US commercial honey bee stocks. *BMC Genomics*, *21*(1), 704. https://doi.org/10.1186/s12864-020-07111-x

Schlotterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nature Reviews. Genetics*, *15*, 749. https://doi.org/10.1038/nrg3803

Speed, D., Holmes, J., & Balding, D. (2020). Evaluating and improving heritability models using summary statistics. *Nature Genetics*, *52*(4), 458–462. https://doi.org/10.1038/s41588-020-0600-y

Su, S., Albert, S., Zhang, S., Maier, S., Chen, S., Du, H., & Tautz, J. (2007). Non-destructive genotyping and genetic variation of fanning in a honey bee colony. *Journal of Insect Physiology*, *53*(5), 411–417. https://doi.org/10.1016/j.jinsphys.2007.01.002

Tarpy, D., & Nielsen, D. (2002). Sampling error, effective paternity, and estimating the genetic structure of honey bee colonies (Hymenoptera: Apidae). *Annals of the Entomological Society of America*, *95*(4), 513–528. https://doi.org/10.1603/0013-8746(2002)095[0513:SEEPAE]2.0.CO;2

Tarpy, D., Nielsen, R., & Nielsen, D. (2004). A scientific note on the revised estimates of effective paternity frequency in apis. *Insectes Sociaux*, *51*(2), 203–204. https://doi.org/10.1007/s00040-004-0734-4

Toth, A., & Zayed, A. (2021). The honey bee genome–what has it been good for? *Apidologie*, *52*, 45–62. https://doi.org/10.1007/s13592-020-00829-3

Uzunov, A., Brascamp, E., & Buchler, R. (2017). The basic concept of honey bee breeding programs. *Bee World*, *94*(3), 84–87. https://doi.org/10.1080/0005772X.2017.1345427

Wallberg, A., Bunikis, I., Pettersson, O., Mosbech, M., Childers, A., Evans, J., Mikheyev, A., Robertson, H., Robinson, G., & Webster, M. (2019). A hybrid de novo genome assembly of the honeybee, *Apis*, with chromosome-length scaffolds. *BMC Genomics*, *20*(1), 275. https://doi.org/10.1186/s12864-019-5642-0

Wallberg, A., Schöning, C., Webster, M., & Hasselmann, M. (2017). Two extended haplo-type blocks are associated with adaptation to high altitude habitats in east African honey bees. *PLoS Genetics*, *13*(5), e1006792. https://doi.org/10.1371/journal.pgen.1006792

Wragg, D., Eynard, S., Basso, B., Canale-Tabet, K., Labarthe, E., Bouchez, O., Bienefeld, K., Bien'kowska, M., Costa, C., Gregorc, A., Kryger, P., Parejo, M., Pinto, M., Bidanel, J., Servin, B., Le Conte, Y., & Vignal, A. (2022). Complex population structure and haplotype patterns in western europe honey bee from sequencing a large panel of haploid drones. *Molecular Ecology Resources*, 1-19. https://doi.org/10.1111/1755-0998.13665

Wragg, D., Marti-Marimon, M., Basso, B., Bidanel, J., Labarthe, E., Bouchez, O., Le Conte, Y., & Vignal, A. (2016). Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Scientific Reports*, *6*, 27168.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Eynard, S. E., Vignal, A., Basso, B., Canale-Tabet, K., Le Conte, Y., Decourtye, A., Genestout, L., Labarthe, E., Mondet, F., & Servin, B. (2022). Reconstructing queen genotypes by pool sequencing colonies in eusocial insects: Statistical Methods and their application to honeybee. *Molecular Ecology Resources*, *22*, 3035–3048. https://doi.org/10.1111/1755-0998.13685