



HAL
open science

From group to individual - Genotyping by pool sequencing eusocial colonies

Sonia Eynard, Alain Vignal, Benjamin B. Basso, Yves Le Conte, Axel Decourtye,
Lucie Genestout, Emmanuelle Labarthe, Fanny Mondet, Kamila Tabet, Bertrand
Servin

► **To cite this version:**

Sonia Eynard, Alain Vignal, Benjamin B. Basso, Yves Le Conte, Axel Decourtye, et al.. From group to individual
- Genotyping by pool sequencing eusocial colonies. 2021. <hal-03482633>

HAL Id: hal-03482633

<https://hal.inrae.fr/hal-03482633v1>

Preprint submitted on 21 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Copyright - All rights reserved

From group to individual - Genotyping by pool sequencing eusocial colonies

Sonia E Eynard^{1,*}, Alain Vignal¹, Benjamin Basso^{2,3}, Yves Le Conte², Axel Decourtye³,
Lucie Genestout⁴, Emmanuelle Labarthe¹, Fanny Mondet², Kamila Tabet¹, Bertrand
Servin¹

1 GenPhySE, Université de Toulouse, INRAE, INP, ENVT, Castanet-Tolosan, 31320,
France

2 INRAE, Abeilles et Environnement, Avignon, 84914, France

3 ITSAP, Avignon, 84914, France

4 LABOGENA DNA, Jouy-en-Josas, 78353, France

* sonia.eynard@inrae.fr

Abstract

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

pool sequencing; eusocial insects; *Apis mellifera*; genotype

1 Introduction

2 Eusocial organisms such as bees, ants or wasps live in large colonies produced by a single
3 individual (the queen) and have a specific mating system in which the queen is mated
4 to a cohort of males. In the case of the honeybee, *Apis mellifera*, a colony is typically
5 composed of a single queen, a large number (up to tenths of thousands) of workers and
6 a few males. The queen is usually the only reproducing individual and all individuals
7 present in the colony are its offspring. In the wild, after mating with a cohort of 10 to 20
8 males the virgin queen will return to the colony and maintain its population, throughout
9 her life, by continuously laying eggs. Fertilised eggs will produce diploid worker females,
10 while unfertilised eggs will produce haploid males. Males are therefore a direct sample
11 of the queen genome and can be considered as flying gametes. The mosaic composition
12 of a colony makes standard genomics analysis complex especially when making breeding
13 decisions (Brascamp and Bijma, 2014; Uzunov, Brascamp, and Büchler, 2017). In eu-
14 social populations, each worker performs individual tasks participating in the collective
15 phenotype of the colony. However, although the phenotype of the colony is collective, the
16 queen contributes to more than half of the genetics of the colony (through diploid female
17 and haploid male offspring) that will be passed on to next generations. Thus, the queen's
18 genotype itself is an essential piece of information for genetic analysis aimed at studying
19 the evolution of populations or performing selective breeding. Even though the field of
20 insect genomics has boomed in the past decades there still is a need to expand traditional
21 approaches of population genetics for this specific kind of organisms (Toth and Zayed,
22 2021). However, contrary to large animal species, sampling the queen for genotyping is
23 impossible without threatening its integrity and is therefore rarely performed in routine
24 beekeeping practices. One possible approach to overcome these problems is to perform
25 individual or pool genotyping (Petersen et al., 2020) of a set of males. However this im-
26 plies an increased manipulation effort to sample the individual males or sequencing cost
27 as multiple genotyping experiments are required to infer the genotype of a single queen.

28 Advances in sequencing technologies have brought new opportunities to develop tools

29 for genomics and genetics. Amongst these, parallel sequencing allows for counting of se-
30 quencing reads at all positions on the genome which thus permitted the development of
31 pool sequencing for allele frequencies estimations (Schlotterer, Tobler, Kofler, and Nolte,
32 2014). By combining DNA from multiple individuals into a unique sequencing experiment,
33 pool sequencing allows for cheap and fast data acquisition, especially for non-model or-
34 ganisms for which resources are limited. However pool sequencing outcomes, allele counts
35 in the pool instead of genotypes, are more difficult to use in practice and require specific
36 programs and software to perform SNP calling, mainstream population genetics analysis,
37 association testing (Kofler, Pandey, and Schlötterer, 2011; Bansal, 2010; Purcell et al.,
38 2007; Chang et al., 2015; Zhou and Stephens, 2012; Speed, Holmes, and Balding, 2020)
39 and more. Additionally traditional pool sequencing is performed on a group of unre-
40 lated individuals representing a population often linked by an environmental factor (e.g
41 a population in a specific location, a genetic type ...).

42 In this study, we propose a new application of pool sequencing to multiple individuals
43 from a single colony in the context of eusocial insects. Hence, contrastingly to standard
44 pool experiment, representing a population of individuals, pool experiment on colonies can
45 be seen as sequencing of a meta-individual. Using this specificity we introduce dedicated
46 statistical methods to estimate the genetic ancestry of the queen and reconstruct its
47 genotype from pool sequencing of workers. The acquisition of genotype data will on the
48 one hand provide information on the queen that can further benefit breeding decisions
49 and will on the other hand allow the use of standard programs and software for population
50 genetics analysis such as admixture or association studies. Two models are proposed and
51 evaluated: the first model estimates the genetic ancestry of the queen, based on single
52 colony data but assuming information on the allele frequencies of markers in reference
53 populations and the second model exploits information available across multiple colonies
54 to reconstruct the queen genotype. Performances of the models are evaluated through
55 simulations including some based on real data from a diversity panel in *Apis mellifera*
56 (Wragg et al., 2021). Using these simulations we show that the genetic ancestry of the

57 queen estimated from the pool sequencing data matches results from standard population
58 genetics methods results on genotype data and that the genotype of the queen can be
59 reconstructed with an error rate limited to a few percent. To evaluate the interest of pool
60 sequencing compared to individual genotyping, we applied our genotype reconstruction
61 models to real data in this species from a field experiment where both pool sequences of
62 workers and individual sequences of male offspring from the same colony were available.
63 We showed that inference of the genetic ancestry and the genotype of the queen based on
64 pool sequencing data matches results obtained from individual data on male offspring.

65 Models introduced in this study can be used sequentially to first estimate the genetic
66 ancestries of a population of colonies, then use this information to cluster the dataset
67 into homogeneous populations and finally infer genotypes of colonies by considering them
68 jointly within these homogeneous clusters. Finally we discuss the interpretation of the
69 results obtained with the models proposed, their applicability and possible extensions.

70 **Materials and Methods**

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

73 **Models**

74 We consider data coming from colony pool sequencing experiments. For each colony,
75 whole genome sequencing is assumed to be performed on DNA extracted from the mix of
76 a large number of worker bees. For a colony c , the raw data consist of the reference allele
77 counts and sequencing depths at a fixed set of L biallelic loci. At a locus l , with observed
78 reference allele count x_l^c and sequencing depth d_l^c , we have:

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

79 where g_l^c is the (unknown) queen genotype expressed as the frequency of the reference

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

87 **Homogeneous Population Model** In this approach, we add to model (1) the hypoth-
 88 esis that queens and males of all colonies come from the same random mating population.
 89 Under this hypothesis, (i) the allele frequency at a given locus is the same for all colonies
 90 and (ii) genotypes at a locus are sampled according to this frequency, so we have for a
 91 locus l :

$$\forall c, f_l^c = f_l$$

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

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

$$P(x_l^c | d_l^c, f_l) = \sum_{G \in \{0, 0.5, 1\}} P(x_l^c | d_l^c, f_l, g_l^c) P(g_l^c = G | f_l)$$

$$\mathcal{L}(f_l; \mathbf{x}_l, \mathbf{d}_l) = \prod_c P(x_l^c | d_l^c, f_l) \quad (3)$$

93 where \mathbf{x}_l is the vector of reference allele counts in all colonies and \mathbf{d}_l the correspond-
 94 ing vector of sequencing depths. The likelihood (3) is maximized numerically for f_l
 95 on $[0, 1]$. The maximizing value (called the Maximum Likelihood Estimate, MLE) \hat{f}_l
 96 can be used for inference on \mathbf{g}_l based on the posterior distribution $P(\mathbf{g}_l | \mathbf{x}_l, \mathbf{d}_l, \hat{f}_l) \propto$
 97 $P(\mathbf{x}_l | \mathbf{d}_l, \mathbf{g}_l, \hat{f}_l) P(\mathbf{g}_l | \hat{f}_l)$.

98 This homogeneous population model (HP) should only be applied when the set of
 99 colonies have a similar genetic background. We therefore developed another approach,
 100 the admixture model, aimed at estimating the genetic ancestry of a single colony from
 101 pool sequencing data.

102 **Admixture Model** The objective of this model is to describe the “genetic background”,
 103 the subspecies, of a colony. To do so, we will adopt the widely used modeling framework in-
 104 troduced by Pritchard, Stephens, and Donnelly (2000) and define the genetic background
 105 of a colony as *the proportions of the queen genome that come from a set of pre-defined*
 106 *reference populations* (in our applications below, the reference populations considered are
 107 *Apis mellifera mellifera*, *Apis mellifera ligustica* & *Apis mellifera carnica* and *Apis mellifera caucasica*, the
 108 three main populations found in Western Europe (Wragg et al., 2021)). We will do that
 109 in a supervised manner so we will assume that we are provided with allele frequencies in
 110 a set of K reference populations at the L loci : this takes the form of an $L \times K$ matrix \mathbf{F}
 111 where F_{lk} is the frequency of the reference allele at locus l in population k . Here we are
 112 interested in inferring \mathbf{q} , the K -vector of admixture proportions for the queen: q_k is the
 113 proportion of alleles over all loci that come from population k . Dropping the c index as
 114 the model is fitted for each colony independently, the likelihood for \mathbf{q} is:

$$P(x_l|d_l, \mathbf{F}_l, \mathbf{q}) = \sum_g \int_0^1 P(x_l|d_l, g_l, f_l)P(g_l|\mathbf{q}, \mathbf{F})P(f_l|\mathbf{F})df_l \quad (4)$$

$$\mathcal{L}(\mathbf{q}; \mathbf{x}, \mathbf{d}) = \prod_l P(x_l|d_l, \mathbf{F}, \mathbf{q})$$

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

119 **Priors** To specify the prior $P(g_l|\mathbf{q}, \mathbf{F})$, we use the classical approach of introducing
 120 latent variables $\mathbf{Z}_l = (z_l^1, z_l^2)$ at each locus l that denotes the origins (in terms of reference

121 populations) of the two alleles carried by the queen. Then we can write:

$$P(g_l|\mathbf{q}, \mathbf{F}) = \sum_{\mathbf{Z}_l} P(g_l|\mathbf{Z}_l, \mathbf{F})P(\mathbf{Z}_l|\mathbf{q}) \quad (5)$$

122 where $P(g_l|\mathbf{Z}_l, \mathbf{F})$ is the probability of the queen genotype given the origins of the two
 123 alleles, which is a function of the allele frequencies in the K reference populations (*e.g.*
 124 $P(g_l = 0.5|\mathbf{Z}_l = (2, 2), \mathbf{F}) = 2F_{2l}(1 - F_{2l})$), and $P(\mathbf{Z}_l|\mathbf{q})$ is the probability of the pair of
 125 origins that depends on the admixture proportions \mathbf{q} (*e.g.* $P(\mathbf{Z}_l = (0, 0)) = q_0^2$).

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

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

128 This prior is informative if all reference populations have similar allele frequencies
 129 and more diffuse if allele frequencies in reference populations differ greatly. Finally, the
 130 estimation of the vector \mathbf{q} is performed using an EM algorithm. Note that this is similar
 131 to the supervised version of the estimation procedure of the Pritchard et al. (2000) model
 132 as the matrix of allele frequencies \mathbf{F} is considered known a priori.

133 Simulations

134 To evaluate the performance of the two models, we simulated data as obtained from a
 135 pool sequencing experiment. We assume these data come in the form of the reference
 136 allele counts x_l^c and sequencing depths d_l^c at each locus l , knowing the queen genotypes g_l^c
 137 and allele frequencies in the inseminating drones f_l^c . To further condition our simulations
 138 on what can be expected from real data, we exploited information available in a reference
 139 population of *Apis mellifera* (Wragg et al., 2021). This data consists of 628 European sam-
 140 ples of haploid drones (Supplementary Table ST2) with genotypes available at 6,914,704
 141 Single Nucleotide Polymorphisms (SNPs). Wragg et al. (2021) showed that this panel is
 142 structured into three main genetic background for which unadmixed (reference) individ-

143 uals can be identified, with a threshold of 99% of their genetic background being from a
 144 unique type: the **M** background (*Apis mellifera mellifera*) with 85 reference individuals,
 145 the **L** background (*Apis mellifera ligustica* & *carstica*) with 44 reference individuals and
 146 the **C** background (*Apis mellifera caucasica*) with 16 reference individuals (Supplementary
 147 Table ST3). In the simulations described below, the reference panel information used was
 148 either the allele frequencies in the three main backgrounds ($\mathbf{F} = (F_{lp}) \in [0, 1]^{L \times 3}$, where
 149 the columns contain the allele frequencies of all L markers in genetic backgrounds L, M
 150 and C in this order) and/or the genotypes of the reference individuals.

151 **Independent markers** To evaluate the performance of the models proposed, a first
 152 set of simulations was performed on 1000 independent SNPs chosen to be common and
 153 ancestry informative with respect to the L, M and C genetic backgrounds. To this goal,
 154 the 1000 SNPs were randomly sampled from the 722,170 SNPs out of the 6,914,704 that
 155 had a minor allele frequency (MAF) ≥ 0.1 and a variance across genetic backgrounds
 156 ≥ 0.1 . For this first set of simulations, only the allele frequencies in the reference panel
 157 at the 1000 SNPs were used.

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

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

161 Different values were considered for the α parameters to consider different levels of
 162 admixed ancestries for the colony (Table 1). Simulated genetic ancestries are represented
 163 in Figure S1.

164 Second, the allele frequencies of each SNP l in the cohort of inseminating drones was

165 simulated as:

$$f_l^c \sim \frac{1}{n_d} \text{Binomial}(n_d, \mathbf{F}_{l,\bullet} \mathbf{q}_d^c) \quad (8)$$

166 where $\mathbf{F}_{l,\bullet}$ is the l -th line of the \mathbf{F} matrix and n_d is the number of inseminating drones,
 167 here fixed at 15 (Tarpy and Nielsen, 2002; Tarpy, Nielsen, and Nielsen, 2004).

168 Third, the genotype of the queen at a SNP l was simulated by first drawing the
 169 population of origin of each of the two allele of the queen ($\mathbf{Z}_l = (z_l^1, z_l^2)$) from a multino-
 170 mial distribution with parameter \mathbf{q}_q^c . The genotype of the queen was finally obtained as
 171 $g_l^c = \frac{a_{i1}^c + a_{i2}^c}{2}$ where :

$$\begin{cases} a_{i1}^c \sim \text{Bernoulli}(F_{l,z_l^1}) \\ a_{i2}^c \sim \text{Bernoulli}(F_{l,z_l^2}) \end{cases} \quad (9)$$

172 Finally, pool sequencing data was simulated as

$$x_l^c \sim \text{Binomial}(d_c, \frac{g_l^c + f_l^c}{2}) \quad (10)$$

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

175 **Linked markers** Pool sequencing experiments provide information on a large number
 176 of markers distributed throughout the genome. In order to evaluate the performance of
 177 the models in realistic conditions for the distribution of allele frequencies and the genetic
 178 structure, a second set of simulations was performed using individual genotypes of 628
 179 individuals from the diversity panel previously described in Wragg et al. (2021) and used
 180 beforehand to define reference genetic backgrounds. First, individuals were clustered
 181 into seven groups, of all potential combinations of admixture between the three genetic
 182 backgrounds, using hard thresholds on their initial vectors of genetic ancestry estimated
 183 with ADMIXTURE (Alexander, Novembre, and Lange, 2009) (Figure S2). Then, each

184 colony was simulated by sampling haploid genotypes of 17 individuals two of which were
185 united to create the genotype of the queen (replacing step (9) above) and the remaining
186 15 were used as inseminating drones under different scenarios of admixture between the
187 three populations, replacing step (8). Then pool sequencing data x_i^c was simulated as
188 in (10). The simulated scenarios are the same as for independent markers, despite that
189 only 20 colonies are simulated per scenario because of sampling limitation due to the
190 restricted number of individuals to select from. As an example, when the queen of the
191 colony is L genetic background and the inseminating drones are LMC genetic background
192 the two individuals to make the queen were sampled from the group of 'pure' L and the
193 15 inseminating drones were sampled from all the possible groups, as their combination
194 will create a mixture of genetic backgrounds.

195 **Evaluation of statistical models**

196 **Genetic ancestry** For each colony and for each set of simulations, the queen genetic
197 ancestry q^c was estimated using the Admixture model (AM). For independent marker sim-
198 ulations, the estimates were compared to the true simulated value, while for linked marker
199 simulations they were compared to the estimates obtained by running ADMIXTURE on
200 the queen genotype. All simulated colonies were analysed jointly with AM and thereafter
201 clustered into seven groups based on their ancestry vectors. Hence, each cluster was a
202 group of colonies with homogeneous genetic ancestry.

203 **Genotype reconstruction** The HP model was used to reconstruct the queen geno-
204 type, within each of the ancestry clusters described above, in the linked marker simula-
205 tions. Criteria for evaluating the model were :

- 206 • the genotyping error rate measured as the proportion of errors in the reconstructed
207 genotypes among all markers. We measured the genotyping error rate for different
208 calling probability thresholds (see Results).

209 • the calibration of the posterior genotype probabilities. For each locus and each
210 simulated colony, the HP model provides the posterior probabilities of the three
211 possible genotypes. Because in the simulations the true genotype is known, we
212 can evaluate in which proportion of the simulations (π) a genotype with posterior
213 probability P is the true genotype. If the model is perfectly calibrated $\pi = P$.
214 Hence, the calibration of the model was measured as

$$AUC = \int_0^1 |P - \pi| \quad (11)$$

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

216 **Validation on experimental data**

217 **Dataset** In order to evaluate the performance of the genotyping by pool sequencing
218 approach, we produced a new dataset where colonies were both pool sequenced and in-
219 dividual drones were sampled. Thirty four colonies, present at an experimental apiary
220 and representing the diversity of French honeybee populations, were sampled in 2016. For
221 each colony between approximately 300 and 500 worker bees were collected and pooled for
222 sequencing purposes. DNA extraction was performed from a blended solution of all the
223 workers of the colony with 4 m urea, 10 mm Tris-HCl pH 8, 300 mm NaCl, 10 mm EDTA.
224 The elution was centrifuged for 15 min at 3500 g, and 200 μ l of supernatant was preserved
225 with 0.5 mg proteinase K and 15 μ l of DTT 1 m for incubation overnight at 56 °C. After
226 manual DNA extraction and DNA Mini Kit (Qiagen) a volume of 100 μ l was used to per-
227 form pair-end sequencing on the Illumina™ HiSeq 3000 or NovaSeq 6000 platform with
228 the aim to obtain approximately 30 \times raw sequencing data per sample. Raw reads were
229 then aligned to the honeybee reference genome Amel HAV3.1, Genebank assembly acces-
230 sion GCA_003254395.2 (Wallberg et al., 2019), using BWA-MEM (v0.7.15; (Li, 2013)).
231 For pool sequenced experiments the resulting BAM files were converted into pileup files
232 using Samtools mpileup (Li and Durbin, 2009) with the parameters: -C 50 coefficient of

233 50 for downgrading mapping quality for reads with excessive mismatches, -q 20 minimum
234 mapping quality of 20 for an alignment, -Q 20 and minimum base quality of 20, following
235 standard protocols. This procedure was applied exclusively to the 6,914,704 Single Nu-
236 cleotide Polymorphisms (SNPs) identified in Wragg et al. (2021) as polymorphic in the
237 European honeybee population. The pileup files were interpreted by the PoPoolation2
238 utility mpileup2sync (Kofler et al., 2011) for the Sanger Fastq format, with a minimum
239 quality of 20 and were finally converted to allele counts and sequencing depth files using
240 a custom-made script. In addition, for each of these 34 colonies 4 male offspring of the
241 queen, genetically equivalent to queen gametes, were individually sequenced as in Wragg
242 et al. (2021) (Supplementary Table ST4). In order to reduce computation time this anal-
243 ysis was performed on a subset of about 50000 markers. These markers were selected
244 following the criteria: 1) maximum of two polymorphic sites within a 100 base pair win-
245 dow, 2) only one representative marker per linkage disequilibrium block with r^2 higher
246 than 0.8, 3) variance between allele frequencies in the different genetic backgrounds higher
247 than zero, to allow for population identification and 4) sampled so that the minor allele
248 frequency follows a uniform distribution. This selection led to exactly 48 334 markers in
249 the experimental dataset.

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

- 253 1. By averaging the genetic ancestry vectors of the four males as estimated by AD-
254 MIXTURE.
- 255 2. By first reconstructing the queen genotype from the male offspring data (see below)
256 and then analysing the resulting genotype with ADMIXTURE.

257 **Genotype reconstruction** For pool sequencing data, queen genotypes were recon-
258 structed using HP, considering the 34 colonies jointly. For the male offspring data, queen

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

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

266 To compute the genotype posterior probability when r_l reference alleles are observed
 267 at a locus, we specify a uniform prior on the three possible genotypes, so that $P(g_l =$
 268 $x|R = r_l) = P(R = r_l|g_l = x)/\sum_{x' \in [0,0.5,1]} P(R = r_l|g_l = x')$. For our application, we
 269 fixed $\epsilon = 10^{-3}$ and n_g is four as described above. Because we have only four drones
 270 per colony in this dataset, there is still some uncertainty in the genotype of the queen.
 271 For example the highest posterior probability achievable for a genotype with $n_g = 4$ is
 272 ≈ 0.94 . This has to be taken into account when comparing the genotypes reconstructed
 273 from the offspring data and from the pool sequencing data: the concordance between the
 274 two approaches has to be measured with respect to what is expected between the true
 275 genotype of the queen and the one reconstructed from noisy data (either offspring or pool
 276 sequencing). Unfortunately we do not know the true genotype of the queen in our dataset
 277 but we can measure the concordance between the genotype reconstructed with four male
 278 offspring to the true genotype of the queen using data from Liu et al. (2015). In this
 279 dataset, genotypes of 13 to 15 offspring are available for three colonies. With that many
 280 offspring the genotype of the queen can be reconstructed with certainty and be compared
 281 to the one obtained by downsampling the data to four offspring per colony. Therefore,
 282 for each of the three colonies in Liu et al. (2015), we called the offspring genotypes at the

283 set of markers present in the diversity panel, reconstructed the queen genotype using (i)
284 all offspring ($n_g = 15$ or 13) and (ii) a 100 randomly downsampled datasets consisting of
285 four offspring only.

286 **Results**

287 In this study we developed statistical models to estimate genetic ancestry and queen
288 genotypes from pool sequencing data from workers of the colony. Simulations, from inde-
289 pendent and linked markers, were performed to evaluate the performance of our models in
290 terms of queen genetic ancestry inference and genotype reconstruction. The scenarios are
291 described in Figure S1. Moreover, these models were applied to an experimental dataset
292 composed of both pool sequenced data and individual male offspring of the queen. In fact
293 male offspring of the queen, haploid individuals coming from unfertilised queen gametes,
294 are direct sampling of the queen genetics and their use is often suggested in literature as
295 a proxy for queen information.

296 **Validation on simulations**

297 **Genetic ancestry** For independent markers, correlations between simulated genetic
298 ancestries and estimated genetic ancestries using the Admixture Model (AM) ranged
299 between 0.88 and 0.9 depending on the genetic background and for linked markers corre-
300 lations between genetic ancestries estimated using ADMIXTURE (Alexander et al., 2009)
301 on the queen genotypes simulated from real data and estimated by AM ranged between
302 0.93 and 0.95 depending on the genetic background (Figure 1). In addition to the 15
303 scenarios listed we also estimated genetic ancestries by AM on scenarios in which queen
304 and drones had divergent ancestries (Supplementary table ST1). We observed that shift-
305 ing from the initial hypothesis that queen and drones come from the same origin led to
306 highly biased genetic ancestry estimations with AM (Figure S3). It should be noted that
307 the statistical model under AM is based on the assumption that markers are indepen-

308 dent. To match this assumption a subset of 1000 markers, rather than the whole genome,
309 was used to estimate genetic ancestry for simulations with linked markers. These results
310 show that AM outputs accurate genetic ancestry estimates and show inference with high
311 agreement to standard population genetics models such as ADMIXTURE, under the as-
312 sumption that queen and drones are of the same origin. Moreover, the observed results
313 confirm that using only a subset of ancestry informative markers, here 1000 from the
314 whole genome, is sufficient to accurately estimate genetic ancestries using AM.

315 **Genotype reconstruction** One major assumption of the Homogeneous Population
316 Model (HP) is that colonies within the population are of homogeneous genetic ancestries.
317 Therefore, using simulations for linked markers across the whole genome, we compared
318 and clustered all the simulated colonies based on their genetic ancestries estimated by
319 AM. In our study we assume that colonies come from a mixture of three main genetic
320 backgrounds (as described in Wragg et al. (2021)), we thus clustered our simulated colonies
321 in seven groups from pure to hybrid genetic types (Figure 2).

322 Thereafter, to evaluate queen genotype reconstruction performance we implemented
323 the Homogeneous Population Model (HP) on our seven groups of homogeneous colonies
324 for linked markers. As the HP model does not make the assumption of independence
325 of markers the inference could be performed on the whole genome, approximately 7 mil-
326 lion markers. Across all simulations and all scenarios, we observed a good correlation
327 between the rate of genotype agreement between simulated and estimated genotypes and
328 the associated estimated genotype probability. In other terms genotypes inferred with a
329 high probability are often correctly predicted by HP whereas genotypes inferred with a
330 low probability are often wrongly predicted by HP, making genotypes with a probability
331 close to 0.5 the hardest to infer precisely. The calibration of the HP model for geno-
332 type reconstruction, measured as the Area under the Curve between agreement rates and
333 probabilities was 0.055 (Figure 3A), when AUC ranged between 0, for perfect correlation
334 and 0.5 for completely imperfect correlation. A large proportion of the markers have

335 probabilities close to zero or to one, making the genotypes drawn for these markers close
336 to certain (Figure 3A). As expected we observed that the genotyping error rate decreases
337 slightly when the best genotype probability threshold increases meaning that filtering for
338 markers with higher best genotype probability leads to more accurate genotype recon-
339 struction. However such filtering is accompanied with a small reduction in genotype call
340 rate. For example if no filtering on best genotype probability is applied, 100% of the
341 genome will be reconstructed with an average genotyping error rate of 4%, if filtering
342 for markers with best genotype probabilities above 0.9 is applied about 95% of the whole
343 genome will be reconstructed with an average genotyping error rate as little as 2% (Figure
344 3B). Additionally we observed that the genotyping error rate increased when the MAF
345 threshold increased meaning that filtering on MAF might cause an increase in genotyping
346 error, accompanied by a drastic reduction in genotype call rate (Figure 3C). Minor Allele
347 Frequency and best genotype probability are highly linked as markers with low MAF
348 tend to be easier to infer with high probability. In our simulation a large proportion,
349 more than 50%, of the whole genome is composed by markers with MAF below 0.05.
350 Yet applying a filter on best genotype probability does not seem to highly impact the
351 distribution of MAF on the whole genome (Figure S4). Rather than filtering on MAF we
352 suggest to filter on best genotype probability, for example equal to or greater than 0.95.
353 Indeed, such filtering will improve the queen genotype reconstruction accuracy without
354 heavily impacting the allele frequency distribution of the markers genotyped on the whole
355 genome. In fact, we observed that genotyping error, on the whole genome and without
356 filtering, is on average about 3% (Figure 3D). After applying a filter on best genotype
357 probability equal to or greater than 0.95 genotyping error becomes on average as low as
358 about 2%.

359 These results show average estimates across all simulation scenarios and colonies after
360 grouping based on genetic ancestry. Detailed results for calibration and genotyping error
361 are presented Figure S5.

362 To conclude, using simulations we confirm that the statistical model AM performs

363 similarly to ADMIXTURE leading to highly accurate genetic ancestry inference. A small
364 set of markers, as low as 1000 in our example where genetic background differentiation is
365 strong, seems sufficient to accurately estimate genetic ancestry with AM. Using simulation
366 of linked markers across the whole genome we confirmed that HP reconstructed queen
367 genotypes with high accuracy. Furthermore, we inferred the impact of MAF and best
368 genotype probability thresholds on the genotype call rate and the associated genotyping
369 error rates, giving the advice to filter on best genotype probability equal to or greater
370 than 0.95 to reduce genotyping error, without drastic loss of predicted markers and while
371 preserving allele frequency distribution across the genome.

372 **Application on experimental data**

373 To further evaluate the performance of the AM and HP models, we analyzed real data
374 on honeybee colonies for which 4 drones were individually sequenced (see Materials and
375 methods).

376 **Genetic ancestry** For each colony, the genetic ancestry of the queen was estimated
377 either from the group of male offspring or from the pool sequences of workers. Genetic
378 ancestry from worker pool sequence were estimated using the Admixture Model (AM).
379 For male offspring, it was estimated with ADMIXTURE (Alexander et al., 2009) either
380 using the male offspring directly (admix_males) or from the genotype of the queen recon-
381 structed using male offspring (admix_proba), as described in the Material and Methods
382 section. Using male offspring data directly (admix_males) or through queen genotype
383 reconstruction (admix_proba) genetic ancestry from ADMIXTURE were virtually equal
384 with a Mean Squared Difference (MSD) of 1.4×10^{-3} (standard deviation 1.1×10^{-3}).
385 Comparing estimates based on male offspring versus worker pool sequence (AM) MSD
386 were slightly higher with 0.024 and 0.026 with standard errors of 0.025 and 0.021 for ad-
387 mix_males and admix_proba respectively (Table 2). Out of the 34 experimental colonies
388 most of the genetic ancestry estimated using queen reconstructed genotypes from worker

389 pool sequencing data, male offspring or using individual sequencing of male offspring gave
390 nearly identical q vectors (Figure S6).

391 **Genotype reconstruction** To validate queen genotype reconstruction from worker
392 pool sequence on our experimental dataset we used publicly available data from Liu et al.
393 (2015) on three colonies for which both queen and 13 to 15 male offspring were individually
394 sequenced were used. Of the 50000 selected markers only 14988 were available, as polymor-
395 phic SNPs, on the dataset from Liu et al. (2015). This reduction in the number of markers
396 available for the analysis can be explained as the population used for SNP calling was
397 composed of fewer individuals from a unique and uniform origin in the dataset from Liu
398 et al. (2015). We compared queen genotypes reconstructed from worker pool sequence and
399 queen genotype reconstructed on probabilities from four male offspring (pool/offspring)
400 on the experimental dataset, genotypes from individually sequenced queens and queen
401 genotype reconstructed on probabilities from four male offspring (queen/offspring) and
402 pairs of queen genotype reconstructed on probabilities from four independent male off-
403 spring (offspring/offspring) on the dataset from Liu et al. (2015). Genotype concordance
404 was on average 0.94 (standard deviation 0.03), 0.96 (standard deviation 0.01) and 0.92
405 (standard deviation 0.01) for pool/offspring, queen/offspring and offspring/offspring re-
406 spectively (Figure 4). The highest concordance is observed between the actual queen
407 genotypes and its reconstruction from four male offspring; however queen genotype re-
408 construction from pool and from male offspring seem to present similar concordance than
409 when pairs of independent male offspring are compared. The few colonies showing more
410 discrepancy between genetic ancestry estimates always showed a genetic ancestry from
411 worker pool sequence mostly divergent from the estimates based on males, despite having
412 high concordance between genotype reconstruction. This can be either due to limitations
413 in AM when it comes to disentangling queen genotype from cohort of inseminating drones
414 in the worker pool sequencing data, to the fact that sampling only four male offspring is
415 not sufficient to accurately represent the queen genetic ancestry, because of genetic con-

416 tradition between the queen that produced the male offsprings and the one that produced
417 the workers or to a bias in the markers used for AM. However, this validation confirms
418 that queen genotype reconstructed using worker pool sequencing data performs as well
419 as individually sequencing multiple male offspring. Additionally we showed, on the data
420 from Liu et al. (2015), that increasing the number of male offspring individually sequenced
421 to six, eight or even ten improved the genotype concordance quite substantially (Figure
422 S7) with eight and ten male offspring showing a concordance between reconstructed and
423 real genotype close to one.

424 To summarise, the difference between genetic ancestry estimated from male offspring
425 or worker pool sequencing data, using AM, were small. Queen genotype reconstruction
426 from worker pool sequencing data was in agreement with queen genotype reconstructed
427 from male offspring. This value was slightly lower than when comparing queen recon-
428 structed genotypes from male offspring with the real queen genotype and slightly higher
429 than when comparing queen reconstructed genotypes from different sets of male offspring
430 of the same queen. HP on worker pool sequencing data is an accurate alternative to
431 individually sequence a limited number of male offspring of the queen when one wants to
432 access the queen genotype.

433 **Discussion**

434 The past decade has seen the growth of the molecular genomics era with the development
435 of new sequencing platforms and technologies, one of them being pool sequencing. This
436 technology allows for the combination of multiple individuals in one sequencing experi-
437 ment, reducing drastically preparation and sequencing costs and therefore making high
438 depth sequencing available for a wide variety of samples. Traditionally pool sequencing
439 is used to perform analysis on multiple individuals from a population. Additionally pool
440 sequencing might be of interest when group level information is desired as for example in
441 the context of eusocial organisms. In such cases the pool will represent a meta-individual

442 of the colony rather than a population. One pitfall of using such sequencing method it
443 that the outcome of pool sequencing comes in the form of allele read counts and sequenc-
444 ing depths rather than diploid genotype observations making it a non standard format
445 for downstream analysis.

446

447 So far only a few programs, for example Popoolation (Kofler et al., 2011) and CRISP
448 (Bansal, 2010) for SNP calling, Plink (Purcell et al., 2007; Chang et al., 2015) and the
449 R package poolfstat (Hivert, Leblois, Petit, Gautier, and Vitalis, 2018; Gautier, Vitalis,
450 Flori, and Estoup, 2021) for population genetics or GEMMA (Zhou and Stephens, 2012)
451 and LDAK (Speed et al., 2020) for association study handle non genotype data. However,
452 when considering eusocial insects from the same colony as a pool we might break under-
453 lying assumptions made by these models. In fact, eusocial insects present characteristics
454 deviating from what could be expected in a standard population used for pool sequencing
455 experiments. First, in hymenopterans, reproductive systems are often polyandric, leading
456 to non standard genetic relationships across individuals in the colony. Second, traits of
457 interest are likely to be measured at the colony level. Therefore, in order to avoid compu-
458 tational limitations and biases that could be brought by the use of pool sequencing with
459 unadapted models one may want to infer individual genetic information (e.g. ancestry
460 and genotypes) from a pool from the group. In honeybee, for instance, a colony can be
461 considered as a polyploid organism (with two major chromosomes, coming from the queen
462 and being present in the whole population, and about 15, the number of inseminating
463 drones, minor chromosomes) constituted of haploid male offspring of the queen that can
464 be described as ‘flying gametes’ as they come from queen unfertilised eggs and diploid
465 female offspring of the queen, worker bees, descendant from the mating of a queen with
466 a cohort of about 15 inseminating drones. Genetic relationships between colony inmates
467 is more complex than in other animal species as they range between 1 to 0.25 depending
468 on the patriline from which the individual belongs (Oxley and Oldroyd, 2010). The hon-
469 eybee queen carries the largest part of the genetic information of the colony and is the

470 producing organ of the next generation making it a favored pathway for breeding selection.
471 In addition, the honeybee populations used by breeders and beekeepers are often highly
472 structured with vast differences between genetically pure and highly admixed colonies.
473 The honeybee population has been influenced by domestication and selection performed
474 by beekeepers often on traits measured at the colony level making the use of pool highly
475 relevant. These features make the use of *Apis mellifera* as a model organism, to develop
476 statistical models to use pool sequencing data, greatly relevant. Moreover we also benefit
477 from the available knowledge on the organism compared to other eusocial insects. For
478 example we can exploit the diversity panels, such as built in Wragg et al. (2021), as priors
479 in our models to facilitate inference. In this context the developed methods are expected
480 to be easily applicable to organisms with lower level of population stratification, as can
481 be for some other eusocial insects.

482

483 Here we present two statistical models to infer queen information from pool exper-
484 iment data. First, the Admixture Model (AM) allows to infer queen genetic ancestry
485 from worker pool sequencing data knowing expected allele frequencies in a reference pop-
486 ulations with high correlation between predicted and expected ancestry (about 0.9) and
487 computational efficiency as it can be run rapidly for each colony independently, thus
488 parallelisable, on a small subset of markers. Second, the Homogeneous Population Model
489 (HP) allows for an accurate queen genotype reconstruction with as little as 2% genotyping
490 error. This model takes advantage of the information from other colonies of the group to
491 complete genotype reconstruction, making the assumption that colonies within a group
492 are of homogeneous genetic ancestry. Within the context of population genetics study,
493 when genetic ancestry is unknown prior to the analysis and knowing the results of this
494 study we suggest to first infer genetic ancestry using AM for all the colony DNA pools of
495 interest, then group them based on similarities in their ancestries and perform genotype
496 reconstruction on these groups separately with HP. Therefore, we propose to use our sta-
497 tistical models sequentially to reach highly accurate genotype reconstruction. To date a

508 common way to infer honeybee queen genotype without manipulating and sacrificing this
509 queen is to perform pool sequencing on multiple honeybee queen male offspring (Petersen
500 et al., 2020). For this purpose Jones et al. (2020) suggests, using theoretical estimations,
501 to sequence at least 30 individuals. This procedure requires to be able to identify and
502 sample enough male offspring from the colony, which is not always easy depending on the
503 season, the colony and the time available for sampling. An alternative is to individually
504 sequence multiple honeybee queen male offspring, in such case, the number of individ-
505 ual sequences is the limiting factor to an accurate queen genotype reconstruction with
506 at least eight to ten individuals needed to accurately deduce queen genome phase, that
507 we cannot obtain from a pool experiment, and to lower the risk of incorrect genotype
508 reconstruction (Figure S7). Using real data we saw that our statistical models, based on
509 pool sequence experiments, reconstructed queen genotypes at least as well as using four
510 individual male offspring sequences. Queen genotype reconstruction from pool sequenc-
511 ing data from workers of the colony appears to be a relevant alternative, cheaper as only
512 one sequencing procedure needs to be performed. Simulations, of independent and linked
513 markers, and the experimental field dataset concluded that we could estimate honeybee
514 queen genetic ancestry and genotype accurately and efficiently using our methods.

515

516 Despite the efficiency of the statistical models described in this study some limitations
517 have been identified and further improvements can be conducted. One crucial assumption
518 of our model is that honeybee queens and inseminating drones have similar genetic an-
519 cestry, which is often true when natural breeding is conducted. However this assumption
520 might be broken when conducting queen artificial insemination for breeding purposes,
521 in extremely controlled breeding environments or even when the breeding environment
522 is 'polluted' by unexpected genetics. In fact, when queen and inseminating drones have
523 highly divergent ancestries our models will estimate biased genetic ancestry and queen
524 genotypes (Figure S3). Additional external information is necessary to account for het-
525 erogeneity in the origin of breeding parents of the pool. One way to do so would be

526 by implementing a two step reconstruction algorithm focusing first on the inseminat-
527 ing drones allele frequencies, for example using information on the breeding practices or
528 sampling drones from the environment as a representation of the mating cohort. Once
529 information on the mating cohort is available it can be easily implemented in our model
530 by adapting the prior in the equation (6). In this study we performed simulations of pool
531 experiments with a sequencing depth of 30x. In practice, and especially in the context
532 of non-model organisms, such sequencing depth might be difficult to reach either due to
533 sequencing cost or to genetic material availability. Therefore, we also tested the simula-
534 tions with a depth of 10 or 100. We compared our results in terms of genotyping error
535 rate and genotype call rate on the genome after filtering for best genotype probability.
536 In Figure S8 we can see that increasing sequencing depth from 10 to 30 improved the
537 accuracy of genotype inference and the genotype call rate. At high sequencing depth,
538 100, we observed higher genotyping error rate overall and limited improvement in the
539 fraction of markers inferred with certainty. It is likely that some level of heterogeneity
540 within the groups used to reconstruct queen genotype led to wrong decisions at higher
541 sequencing depth. Increasing sequencing depth seems to cause higher sensitivity to the
542 hypothesis of homogeneous population by the statistical HP model. One option to reduce
543 this impact would be by grouping colonies based on their genetic ancestries to a more
544 refined scale. Indeed, further developments in the HP model could allow one to take into
545 account a level of heterogeneity in the population to reduce the sensitivity of the model
546 to the homogeneity assumption.

547

548 We observed that HP performed better, had a lower genotyping error rate, if inferred
549 genotypes along the genome were filtered based on their certainty, measured as a proba-
550 bility. In our simulations such filtering did not affect the allele frequency distribution and
551 reduced only slightly the number of inferred markers along the genome while reducing
552 genotyping error rate (Figure S4). An imputation step would contribute to the improve-
553 ment of genome reconstruction completeness. Also taking into consideration Linkage

554 Disequilibrium (LD) along the genome to refine the genotypes inferred by HP could be
555 adapted in our statistical model. Such development would benefit from identification of
556 haplotype blocks in the honeybee genome (Saelao et al., 2020; Wallberg, Schöning, Web-
557 ster, and Hasselmann, 2017; Wragg et al., 2016; Wragg et al., 2021) tagging the different
558 *Apis mellifera* populations. An efficient strategy would be to reconstruct queen genotypes
559 with HP, filter on genotype probability to retain only markers from which reconstruction
560 is satisfying and then apply an imputation step taking into account known haplotype
561 blocks and LD between markers.

562

563 To conclude, colony pool sequencing data can be used to infer queen genetic ances-
564 try when knowing allele frequencies in reference populations present in the environment.
565 Moreover, using pool sequencing data across multiple colonies of homogeneous genetic an-
566 cestry in which queen and inseminating drones come from a similar origin, it is possible to
567 reconstruct honeybee queen genotypes accurately. Such genotypes are valuable for exam-
568 ple to run population genetics analysis and association studies with mainstream models
569 currently available and genetic ancestry estimates can be useful for selective breeding
570 purposes. Additional developments to take into consideration some level of heterogeneity,
571 discrepancy of origins between queen and inseminating drone cohort and linkage dise-
572 quilibrium along the genome will help further increase genotype reconstruction accuracy.
573 The statistical models described in the study have been designed within the context of
574 eusocial hymenoptera but tested solely on *Apis mellifera*. Such models could be tested
575 within the framework of studies on other eusocial species with multiple mating of a single
576 queen (Micheletti and Narum, 2018) and with known genetic diversity panels to estimate
577 priors for allelic frequencies.

578 **Data accessibility statement**

579 Scripts developed to perform the simulation are available at xxxxx for download. The
580 vcf file containing the filtered SNPs and the complete diversity panel can be found in
581 Wragg et al. (2021). The list of 628 individuals used in this study as well as the list of
582 reference individuals and individuals (male offsprings) used for validation can be found in
583 the Supplementary Table S1, together with their accession names. The pool sequencing
584 experiment data for the 34 colonies used for validation can be found at xxx. The external
585 data set used for validation can be found in Liu et al. (2015).

586 **Competing interests**

587 The authors declare that they have no competing interests.

588 **Author's contributions**

589 AV, BS, FM, BB, YLC and AD designed the data collection. FM, BB and YLC performed
590 the data collection. KT, and EL performed the laboratory preparation of the samples,
591 DNA extraction, library preparation and sequencing. SEE, BS and AV designed the
592 study. BS developed the methods and wrote the models. SEE designed and performed
593 the simulations and model comparisons. SEE, BS and AV interpreted the results. FM,
594 YLC, LG, and AD contributed to the discussion. SEE, BS and AV drafted and reviewed
595 the manuscript. All authors have read and approved the manuscript.

596 **Acknowledgements**

597 This study was performed with the support of the ITSAP team for the maintenance
598 of the honeybee colonies and the data collection, the sequencing platform GeT-PlaGe,
599 Toulouse (France), a partner of the National Infrastructure France Génomique, thanks to

600 support by the Commissariat aux Grands Investissements (ANR-10-INBS-0009), for the
601 sequencing and especially Olivier Bouchez. Bioinformatics analyses were performed on the
602 computing facility Genotoul. This research was funded by the Ministère de l'Agriculture
603 de l'Agroalimentaire et de la Forêt within the framework of MOSAR RT 2015-776 project
604 and the Ministère de l'Agriculture de l'Agroalimentaire et de la Forêt and Investissement
605 d'avenir for BeeStrong PIA P3A project. Thanks to Claude Chevalet for the initial
606 discussions on the idea, the members of the BeeStrong project, Florence Phocas and
607 François Guillaume, for their contributions to the discussion during the development of
608 this study.

References

- Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. doi:10.1101/gr.094052.109
- Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of dna pools. *Bioinformatics (Oxford, England)*, *26*(12), i318–i324. doi:10.1093/bioinformatics/btq214
- Brascamp, E. W. & Bijma, P. (2014). Methods to estimate breeding values in honey bees. *Genetics Selection Evolution*, *46*(1), 53. doi:10.1186/s12711-014-0053-9
- Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation plink: Rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1). doi:10.1186/s13742-015-0047-8
- Gautier, M., Vitalis, R., Flori, L., & Estoup, A. (2021). F-statistics estimation and admixture graph construction with pool-seq or allele count data using the r package poolstat. *submitted*. doi:10.1101/2021.05.28.445945
- Hivert, V., Leblois, R., Petit, E., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics*, *210*(1), 315–330. doi:10.1534/genetics.118.300900. eprint: <https://www.genetics.org/content/210/1/315.full.pdf>
- Jones, J. C., Du, Z. G., Bernstein, R., Meyer, M., Hoppe, A., Schilling, E., ... 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. doi:<https://doi.org/10.1002/ece3.6357>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.6357>
- Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). Popoolation2: Identifying differentiation between populations using sequencing of pooled dna samples (pool-seq). *Bioinformatics*, *27*(24), 3435–3436. doi:10.1093/bioinformatics/btr589
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *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. doi:10.1093/bioinformatics/btp324
- Liu, H., Zhang, X., Huang, J., Chen, J. Q., Tian, D., Hurst, L. D., & Yang, S. (2015). Causes and consequences of crossing-over evidenced via a high-resolution recombination landscape of the honey bee. *16*(1), 15. Retrieved from <https://doi.org/10.1186/s13059-014-0566-0>
- Micheletti, S. J. & Narum, S. R. (2018). Utility of pooled sequencing for association mapping in nonmodel organisms. *Molecular Ecology Resources*, *18*(4), 825–837. doi:<https://doi.org/10.1111/1755-0998.12784>
- Oxley, P. R. & Oldroyd, B. P. (2010). The genetic architecture of honeybee breeding. *39*, 83–118.
- Petersen, G. E. L., Fennessy, P. F., Van Stijn, T. C., Clarke, S. M., Dodds, K. G., & Dearden, P. K. (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*. doi:10.1007/s13592-020-00741-w

- 652 Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population struc-
653 ture using multilocus genotype data. *Genetics*, *155*(2), 945–959. Retrieved from
654 %3CGo%20to%20ISI%3E://WOS:000087475100039
- 655 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ...
656 Sham, P. C. (2007). Plink: A tool set for whole-genome association and population-
657 based linkage analyses. *American journal of human genetics*, *81*(3), 559–575. doi:10.
658 1086/519795
- 659 Saelao, P., Simone-Finstrom, M., Avalos, A., Bilodeau, L., Danka, R., de Guzman, L., ...
660 Tokarz, P. (2020). Genome-wide patterns of differentiation within and among u.s.
661 commercial honey bee stocks. *BMC Genomics*, *21*(1), 704. doi:10.1186/s12864-020-
662 07111-x
- 663 Schlotterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals
664 — mining genome-wide polymorphism data without big funding. *15*, 749. Retrieved
665 from <http://10.1038/nrg3803>
- 666 Speed, D., Holmes, J., & Balding, D. J. (2020). Evaluating and improving heritability
667 models using summary statistics. *Nature Genetics*, *52*(4), 458–462. doi:10.1038/
668 s41588-020-0600-y
- 669 Tarpy, D. R. & Nielsen, D. I. (2002). Sampling error, effective paternity, and estimating
670 the genetic structure of honey bee colonies (hymenoptera: Apidae). *Annals of the*
671 *Entomological Society of America*, *95*(4), 513–528. doi:10.1603/0013-8746(2002)
672 095[0513:SEEPAE]2.0.CO;2
- 673 Tarpy, D. R., Nielsen, R., & Nielsen, D. I. (2004). A scientific note on the revised estimates
674 of effective paternity frequency in apis. *Insectes Sociaux*, *51*(2), 203–204. doi:10.
675 1007/s00040-004-0734-4
- 676 Toth, A. L. & Zayed, A. (2021). The honey bee genome— what has it been good for?
677 *Apidologie*. doi:10.1007/s13592-020-00829-3
- 678 Uzunov, A., Brascamp, E. W., & Büchler, R. (2017). The basic concept of honey bee
679 breeding programs. *Bee World*, *94*(3), 84–87. doi:10.1080/0005772X.2017.1345427
- 680 Wallberg, A., Schöning, C., Webster, M. T., & Hasselmann, M. (2017). Two extended
681 haplotype blocks are associated with adaptation to high altitude habitats in east
682 african honey bees. *PLOS Genetics*, *13*(5), e1006792. doi:10.1371/journal.pgen.
683 1006792
- 684 Wallberg, A., Bunikis, I., Pettersson, O. V., Mosbech, M. B., Childers, A. K., Evans, J. D.,
685 ... Webster, M. T. (2019). A hybrid de novo genome assembly of the honeybee, apis
686 mellifera, with chromosome-length scaffolds. *BMC Genomics*, *20*(1), 275. doi:10.
687 1186/s12864-019-5642-0
- 688 Wragg, D., Marti-Marimon, M., Basso, B., Bidanel, J. P., Labarthe, E., Bouchez, O., ...
689 Vignal, A. (2016). Whole-genome resequencing of honeybee drones to detect ge-
690 nomic selection in a population managed for royal jelly. *6*, 27168. Retrieved from
691 <https://www.nature.com/articles/srep27168#supplementary-information>
- 692 Wragg, D., Eynard, S. E., Basso, B., Canale-Tabet, K., Labarthe, E., Bouchez, O., ...
693 Vignal, A. (2021). Complex population structure and haplotype patterns in west-
694 ern europe honey bee from sequencing a large panel of haploid drones. *bioRxiv*,
695 2021.09.20.460798. doi:10.1101/2021.09.20.460798
- 696 Zhou, X. & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for associ-
697 ation studies. *Nature Genetics*, *44*(7), 821–824. doi:10.1038/ng.2310

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)

Description of the simulations for colony and population size, composition and genetic ancestries. For each of the 15 scenarios designed for simulations we present the number of simulated colonies, the queen's genetic ancestry in term of genetic backgrounds *Apis m. caucasia* C, *Apis m. ligustica* & *carstica* L and *Apis m. mellifera* M, the associated Dirichlet alpha vectors, and the same information for the inseminating drones.

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	0.020
max	4.19E-03	0.085	0.082
sd	1.16E-03	0.025	0.021

Genetic ancestry Mean Squared Differences for different data and models on experimental colonies. Minimum, average, median, maximum and standard deviations are calculated for each combination.

699 **Figures**

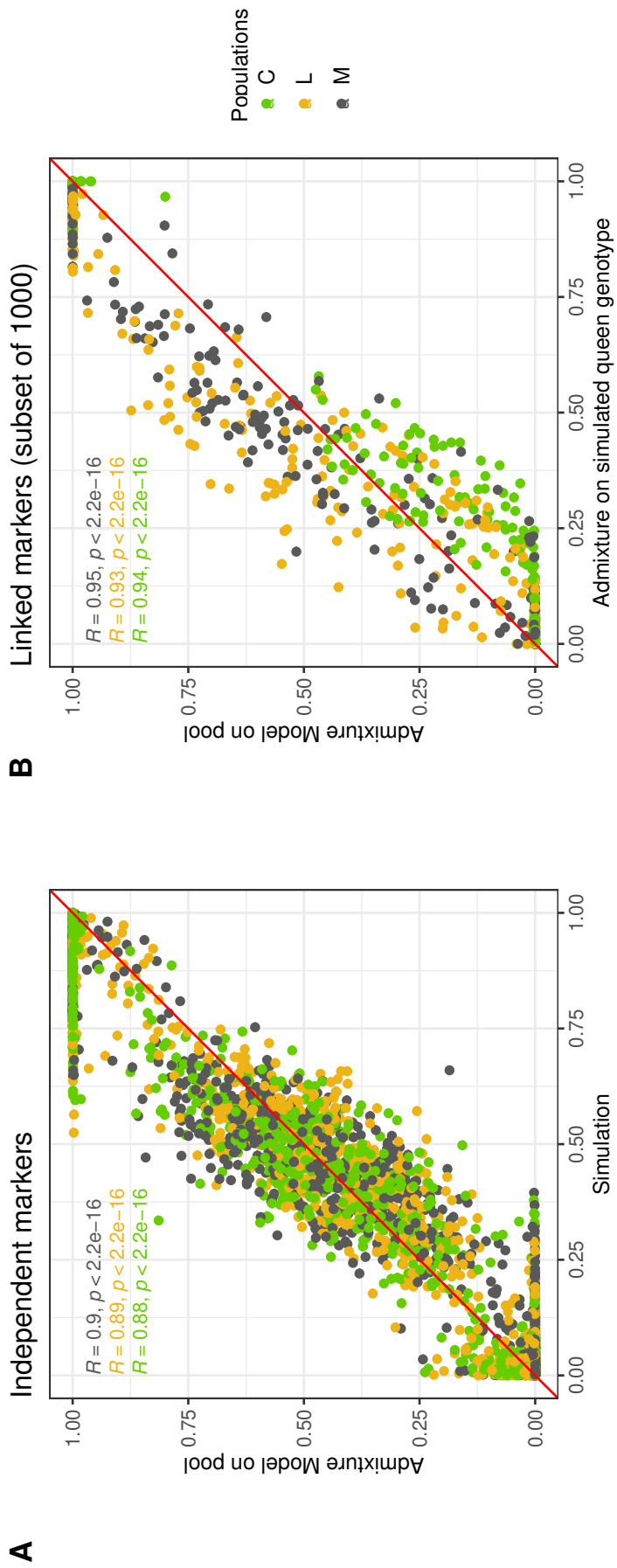


Figure 1: Genetic ancestry comparison Regression of the genetic ancestry vectors estimated by the Admixture Model against simulated. Genetic ancestries estimated with AM against simulated for independent markers, for each scenario, each colony, for each genetic background ($15 * 100 * 3$) (A) or estimated with AM against by ADMIXTURE for simulations for linked markers (subset of 1000), for each scenario, each colony, for each genetic background ($15 * 20 * 3$, the number of simulated colonies is lower due to limitation in the number of individuals to sample from in the real dataset) (B). The red line represents the regression with intercept 0 and slope 1, meaning perfect agreement between the two estimates. Values for spearman rank correlations between ancestry vectors are shown in the top left corner for each of the three genetic backgrounds in green *Apis m. caucasica* C, in yellow *Apis m. ligustica* L and in grey *Apis m. mellifera* M.

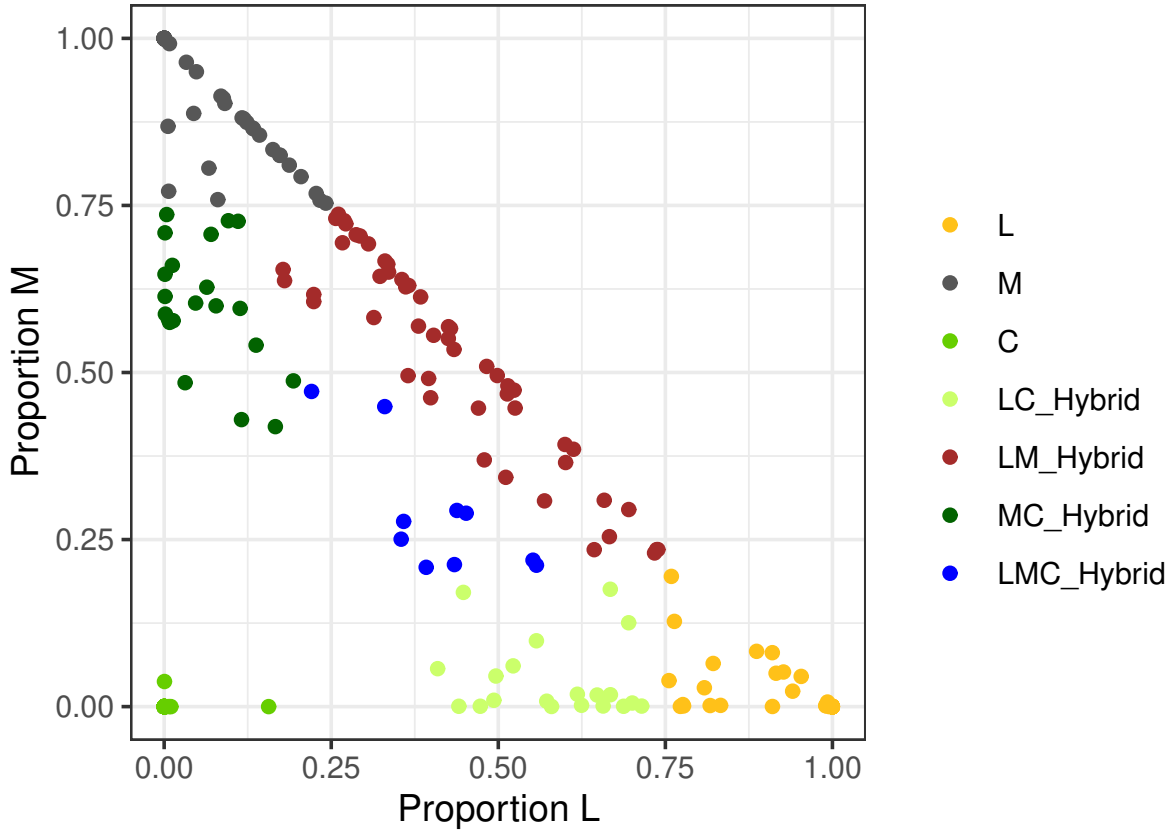


Figure 2: **Genetic ancestries for the simulated colonies as estimated by the Admixture Model** Two dimensions plot of genetic ancestries estimated by AM for colonies simulated for linked markers. X and y axis give the genetic ancestry values in two of the three populations of honeybee in our dataset, for all the colonies in all scenarios (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* 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.

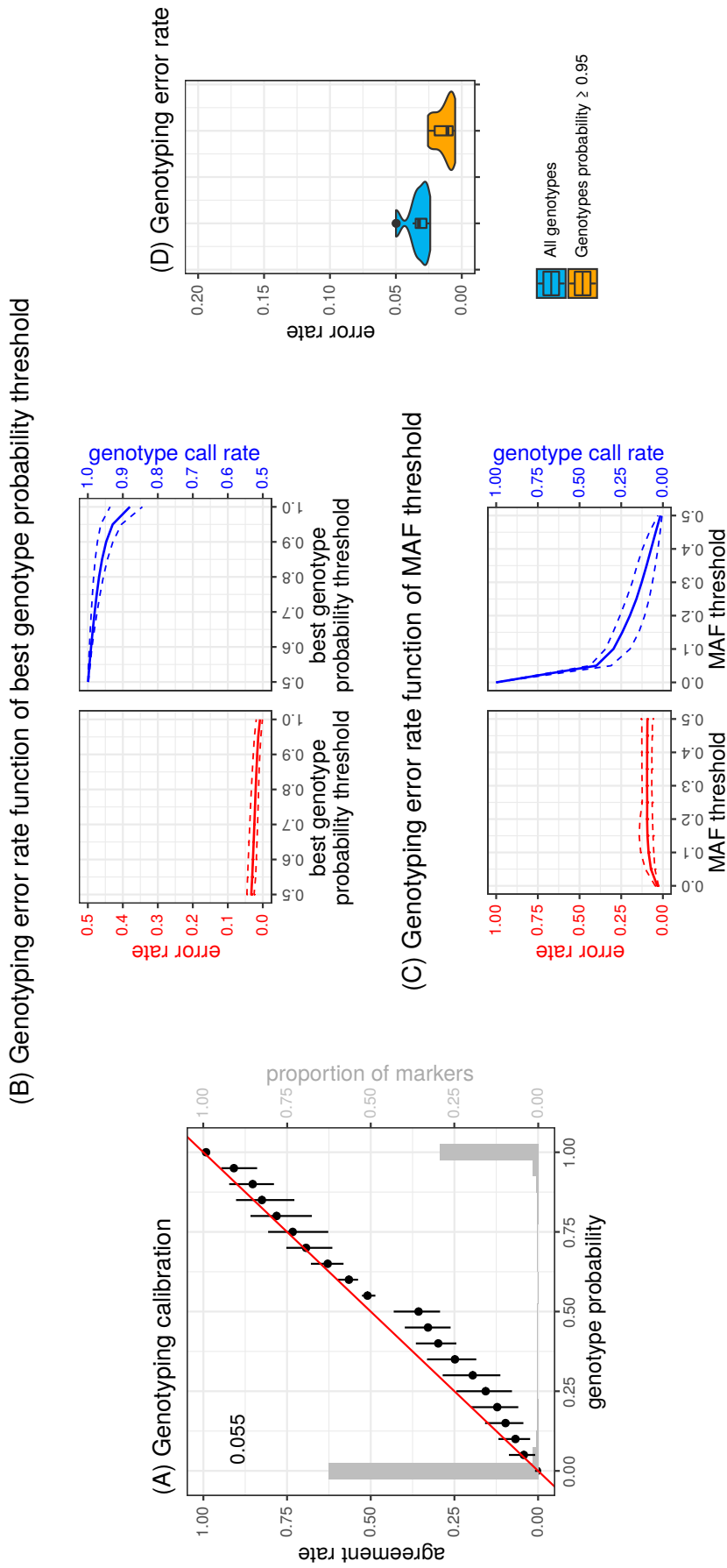


Figure 3: Queen genotype reconstruction For linked marker simulations on the whole genome, values averaged across all colonies and scenarios. A) Genotyping calibration, each point represents the genotype agreement rate per genotype probability value with bars representing the quantiles 95% to 5%. The 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 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 quantiles 95% and 5% for genotype call rate. D) Violin plot of the genotyping error, for all markers or filtering on best genotypes probability equal to or greater than 0.95.

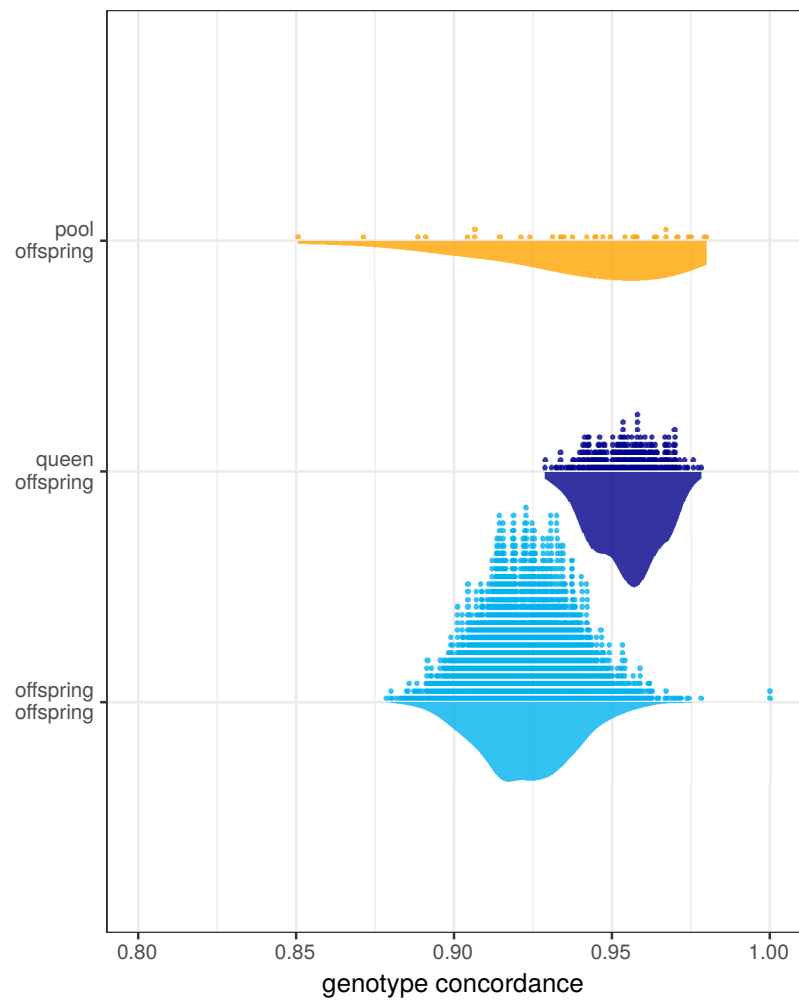


Figure 4: Concordance between queen genotype reconstruction based on different data Concordance between reconstructed genotypes from different data types. The densities, bottom, represent the concordance, only for markers after filtering for best genotype probability equal to or greater than 0.94, between i) queen genotype reconstructed from pool sequencing data using HP and queen genotype reconstructed from genotype probabilities (pool/offspring), based on four male offspring for experimental colonies, in orange ii) queen genotype 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 genotype 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.

700 **Supplementary Figures**

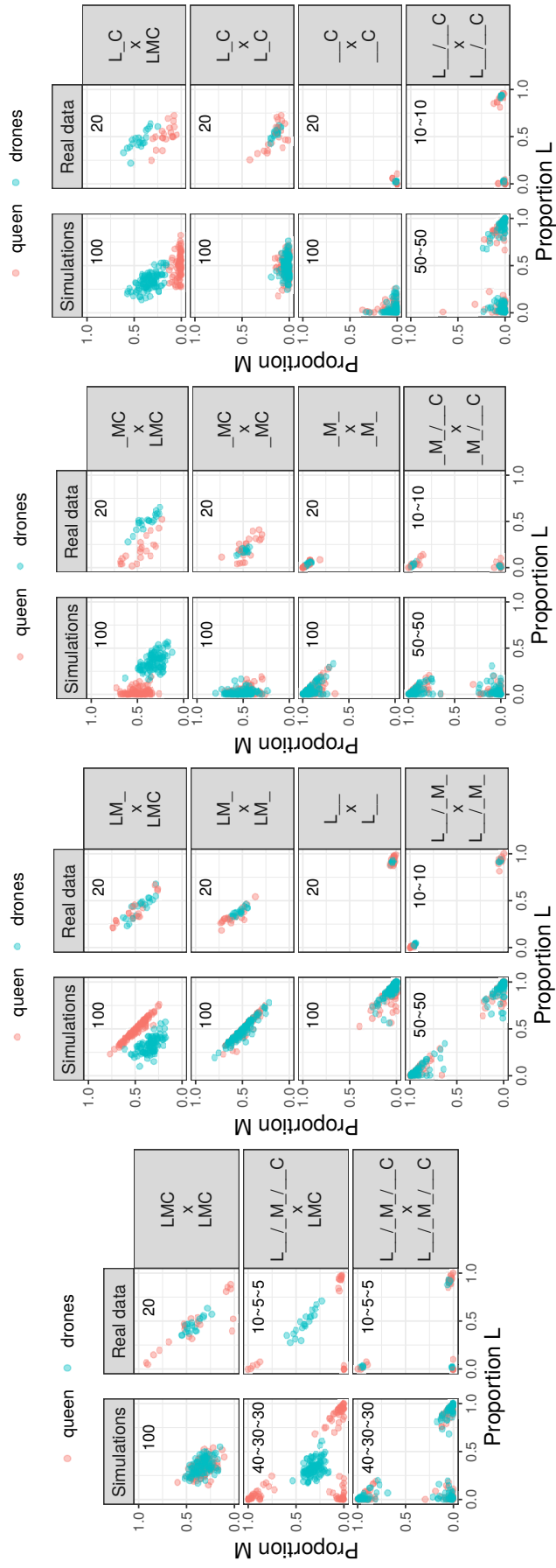


Figure S1: Simulated genetic ancestries for queen and drones. Two dimensions plot of genetic ancestries simulated for queens (pink) and drones (blue) for all scenarios.

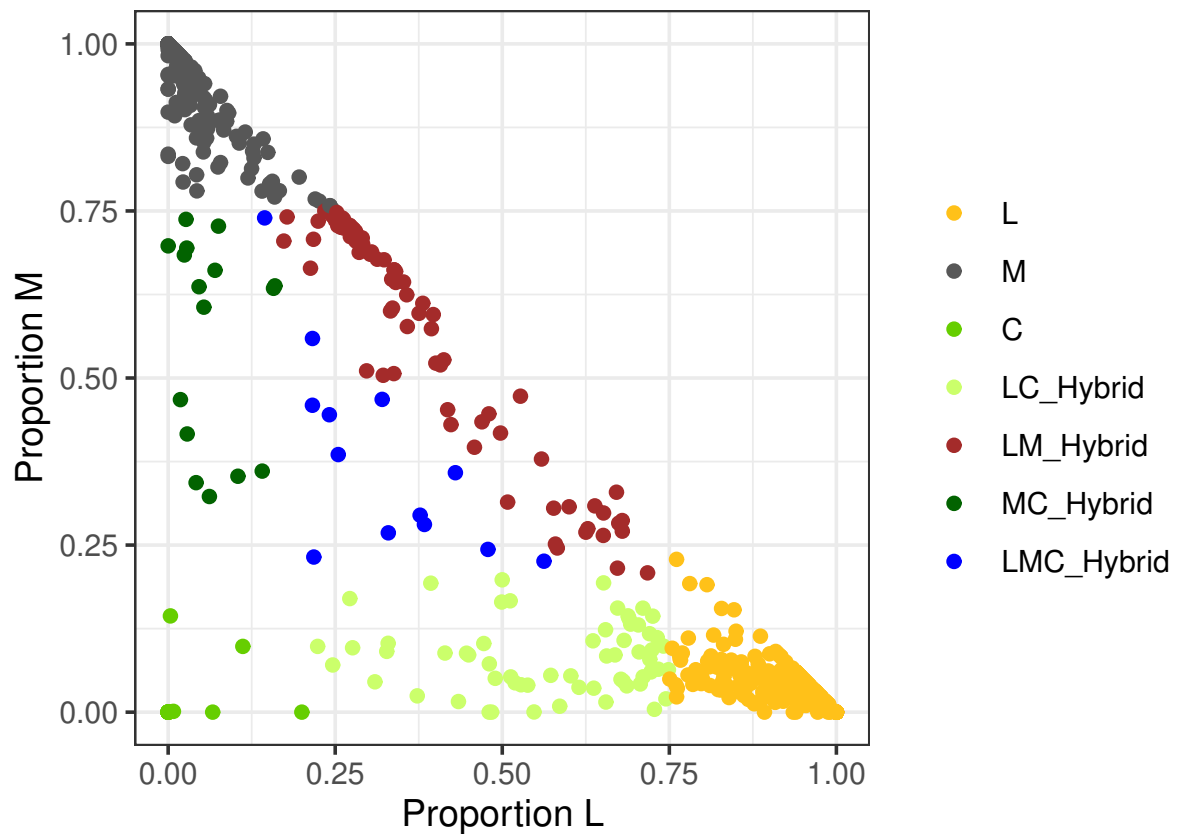


Figure S2: **Genetic ancestries of 628 male individuals from the diversity panel of Wragg et al. (2021)** Two dimensions plot of genetic ancestries for the individuals from the diversity panel. Individuals can be grouped by genetic ancestry. Here we decided on seven groups, each in a different colour, 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* 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.

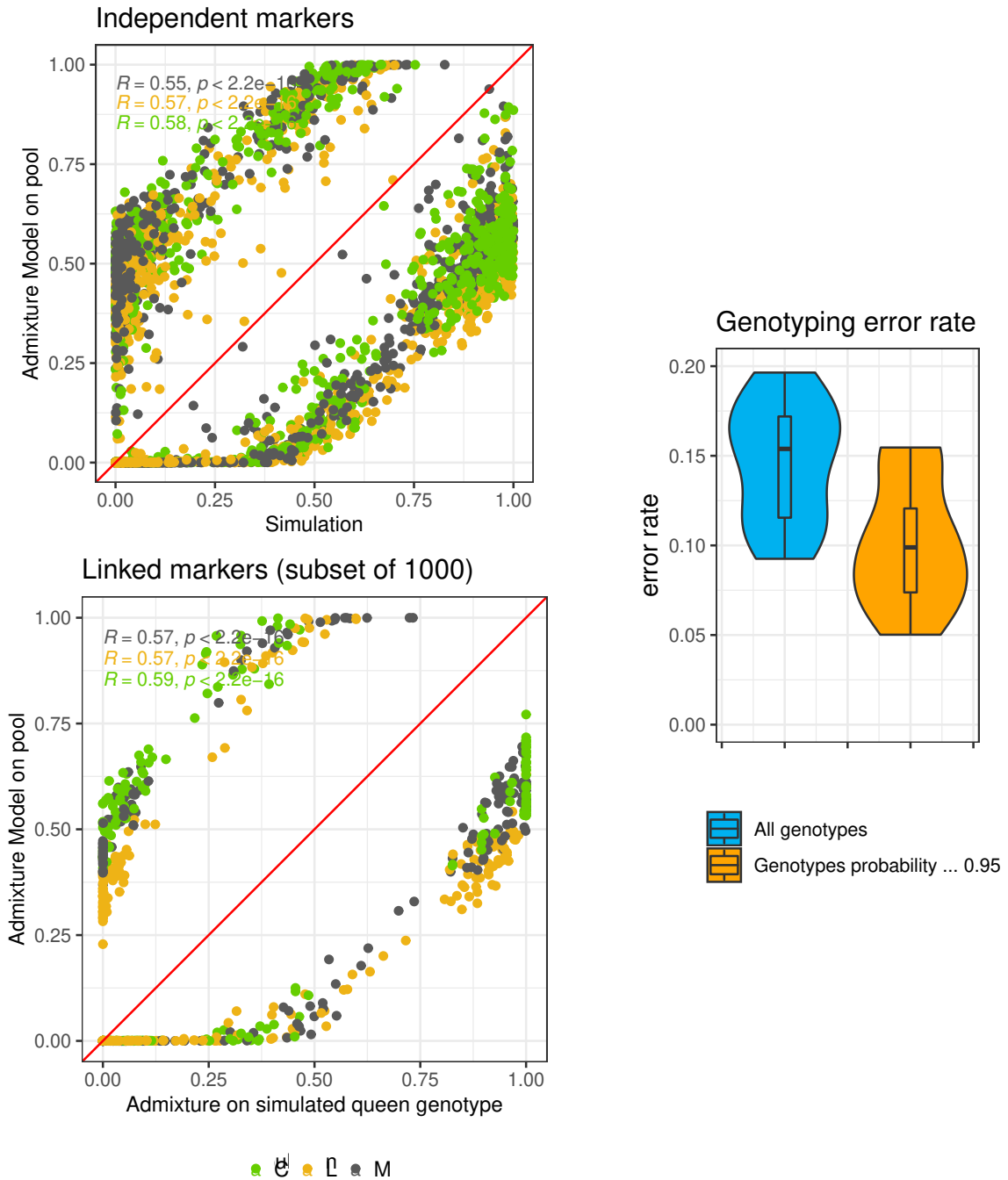


Figure S3: Genetic composition and genotyping error when queen and drones come from different ancestries Detailed information are available in Supplementary Table ST1

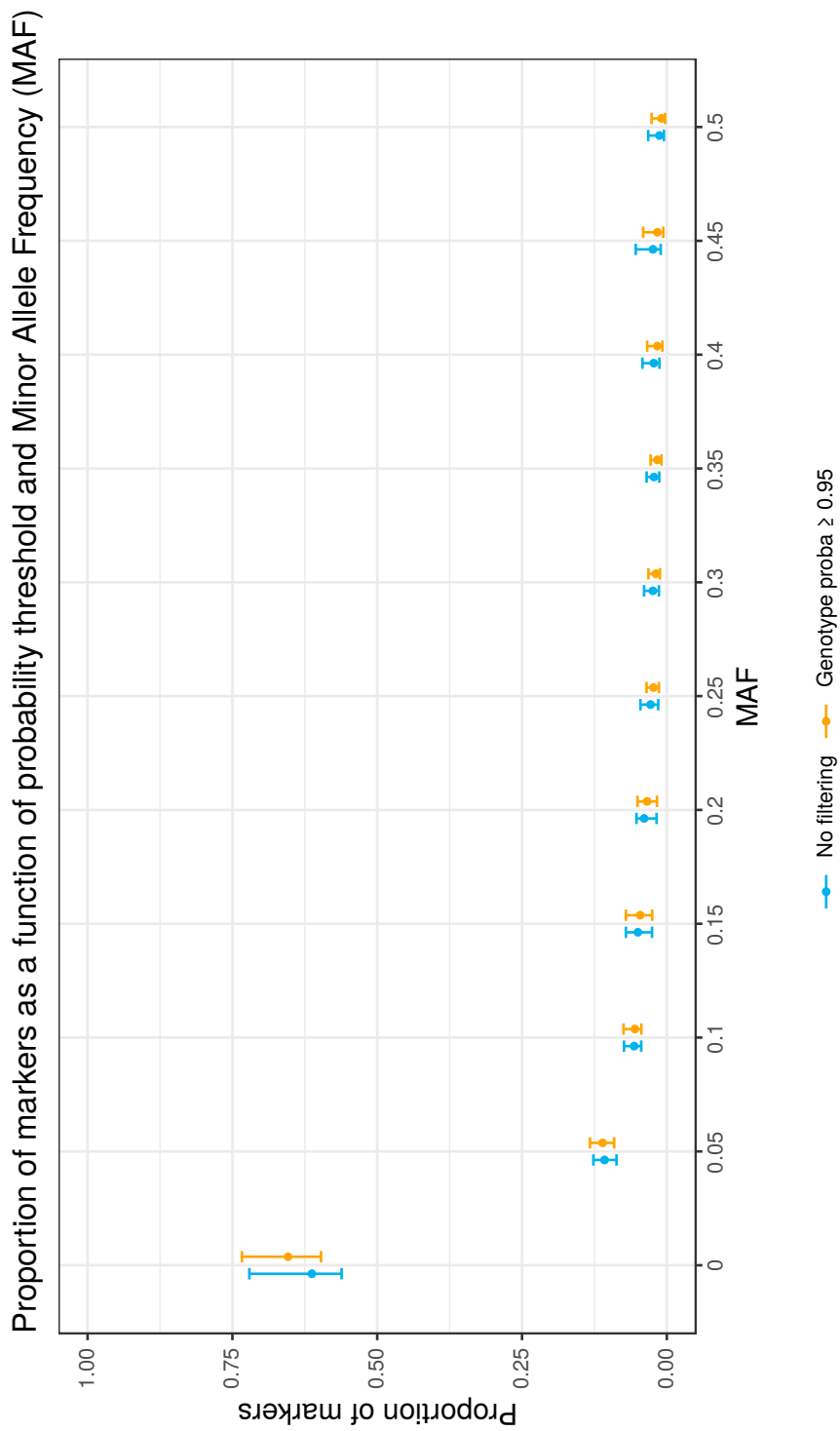


Figure S4: **Proportion of markers in each MAF categories under best genotype probability thresholds** Representation for each MAF category of the mean, with interval, proportion of markers. For MAF 0 to 0.5 we represented the mean and the quantiles 95% and 5% proportion of markers across all simulations. In blue without filtering on best genotype probability, in orange after filtering for markers with best genotype probability equal to or greater than 0.95 on the whole genome or only on these filtered markers.

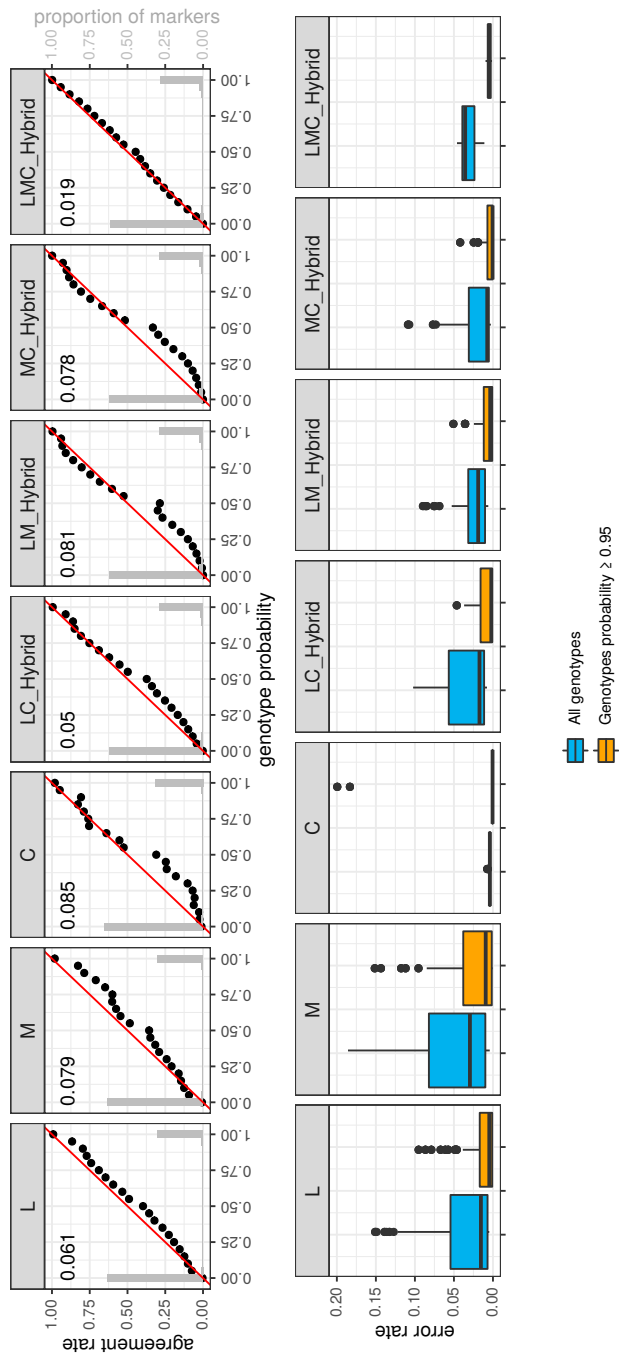


Figure S5: Queen genotype reconstruction for each of the seven groups Detailed genotype calibration and genotyping error rate for each group. The first row represents the genotype calibration, with AUC and genotype probability distribution, for each of the groups tested when performing queen genotype reconstruction using simulations from real data on the whole genome are clustering on genetic ancestries estimated with AM. The second row represents the genotyping error rate for each of the scenarios tested when performing queen genotype reconstruction using simulations from real data on the whole genome or after filtering on best genotype probability equal to or greater than 0.95.

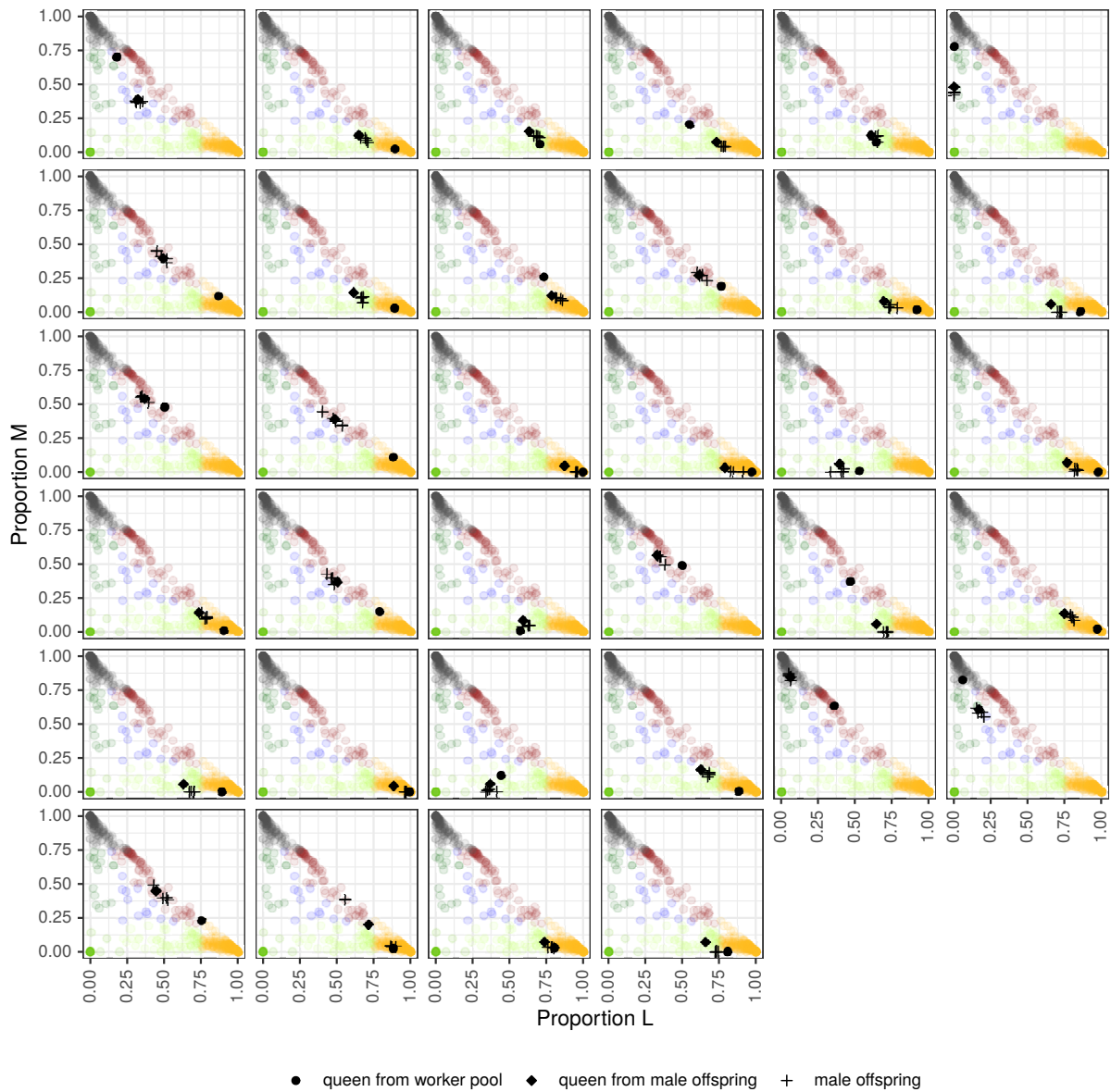


Figure S6: Genetic ancestries on experimental colonies estimated from different models and data on experimental colonies Two dimensions plot of genetic ancestries for the different estimates on the experimental colonies. For the 34 experimental colonies, drones offspring of the queen (crosses), queen reconstructed from these drones (diamonds) and queen reconstructed from the pool experiment (circles) projected on top of the individuals from the diversity panel (628 from Wragg et al. (2021)), representing genetic ancestries in two dimensions.

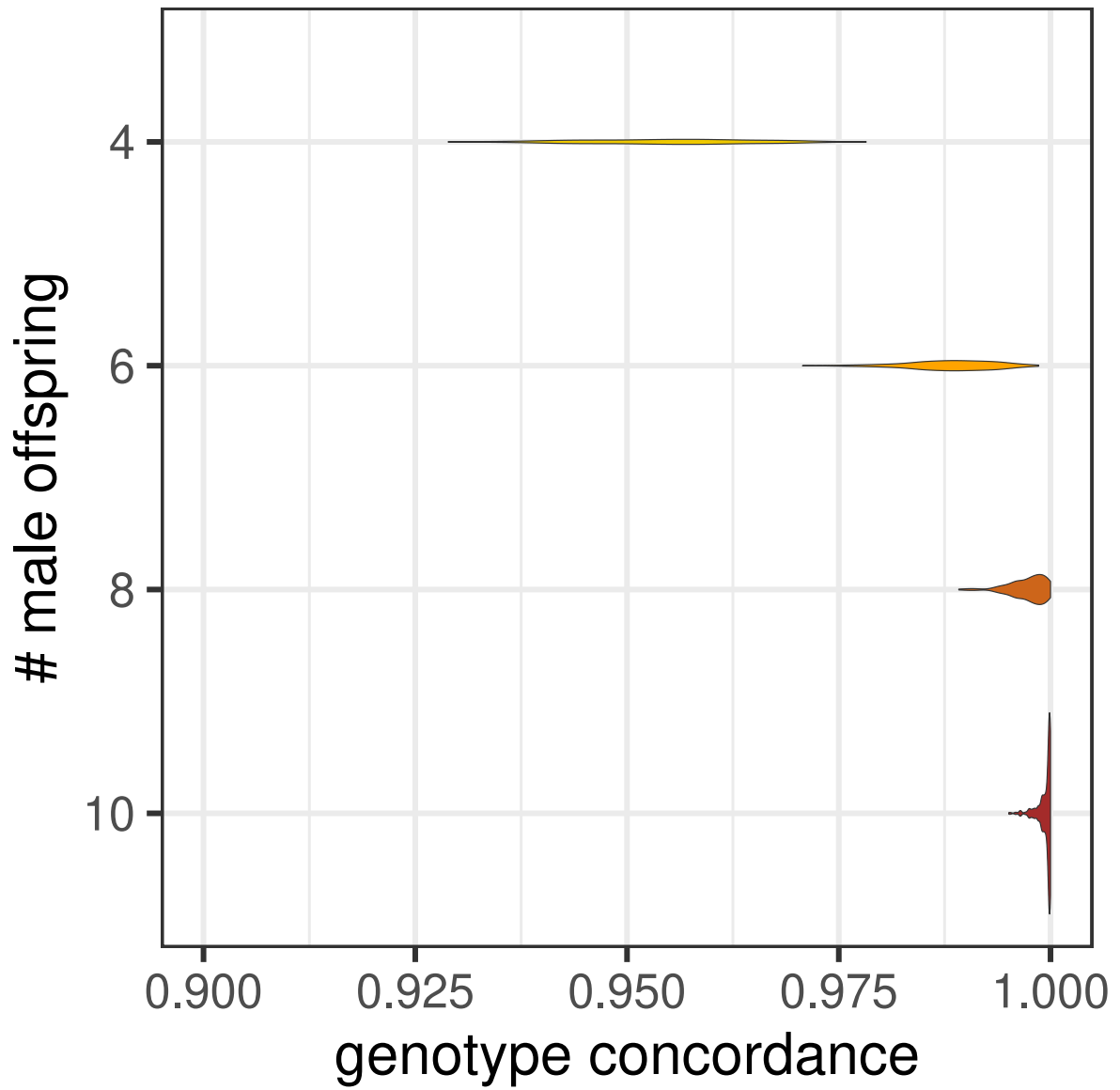


Figure S7: textbfConcordance between real and reconstructed queen genotypes as a function of the number of male offspring available

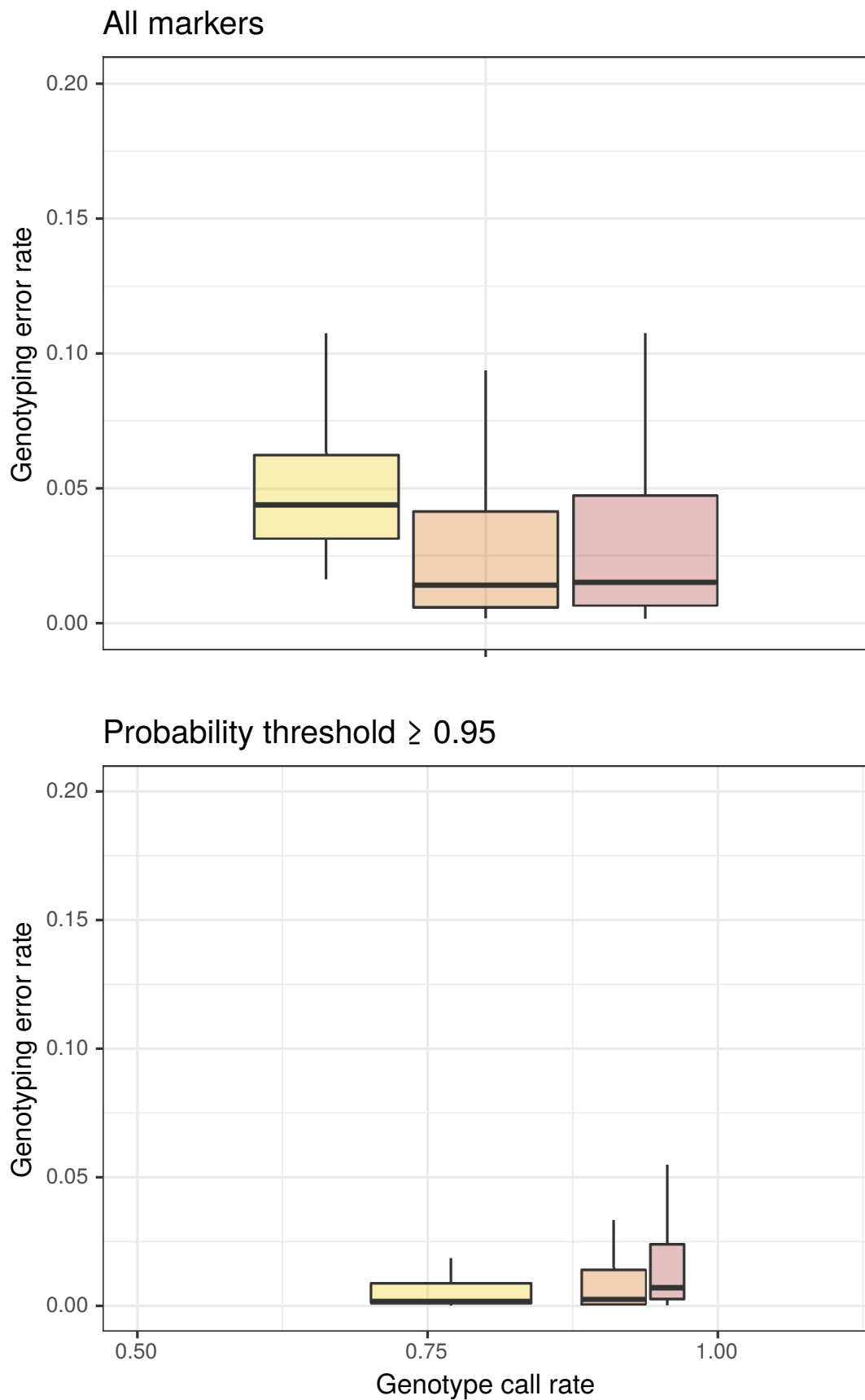
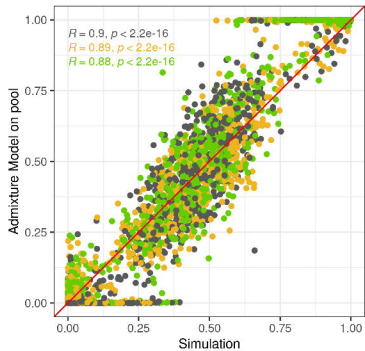
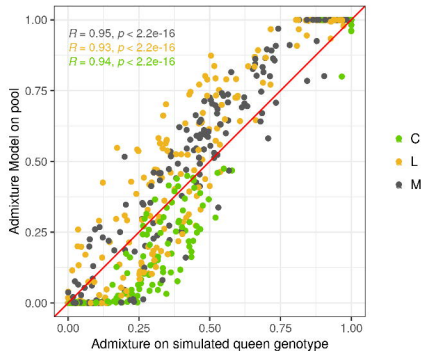
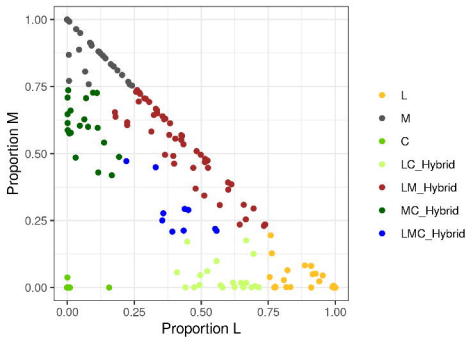
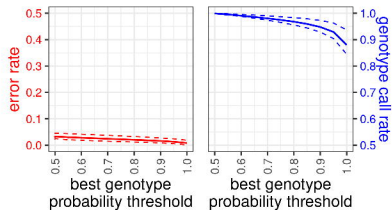


Figure S8: **Genotyping error rates for different sequencing depths** Impact of pool sequencing depth on genotyping error rate. Genotyping error across each colony simulated for linked markers across the whole genome after genotype reconstruction within groups of homogeneous genetic ancestries based on estimations from AM for depth 10 (yellow), 30 (orange) and 100 (brown). The top panel is for all markers on the genome, the bottom panel is for markers with best genotype probability higher or equal than 0.95, the x axis represents the genotype call rate.

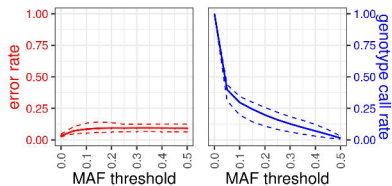
A**Independent markers****B****Linked markers (subset of 1000)**



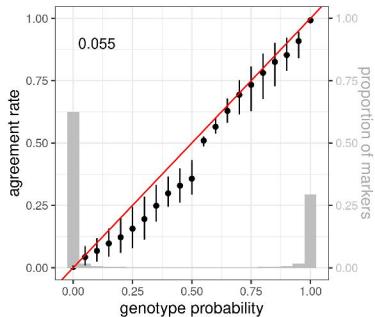
(B) Genotyping error rate function of best genotype probability threshold



(C) Genotyping error rate function of MAF threshold



(A) Genotyping calibration



(D) Genotyping error rate

