Realisation of genomic selection in the honey bee

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

eingereicht an der Lebenswissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

Dipl.-Math. M.A. Richard Bernstein

Abstract

This thesis presents results from genomic breeding value estimation for six economically important traits in the honey bee, and evaluates several strategies to optimize the genetic gain using genomic selection.

Genomic selection is a routine practice for several important livestock species but not yet in honey bees, due to the peculiarities of this species. Chapter 1 introduces the basic concepts of genomic selection and the best linear unbiased prediction method to calculate pedigree-based breeding values honey bees. The model uses the pedigree information via the numerator relationship matrix. For honey bees, a specialized genetic relationship matrix is required, since their mating biology involves uncertain paternity, diploid queens, and haploid drones. Chapter 2 presents a novel algorithm for the efficient computation of the inverse of the numerator relationship matrix and the coefficients of inbreeding on large data sets. In Chapter 3, the method to calculate the inverse numerator relationship matrix was included into methods to estimate genomic and pedigree-based breeding values. The accuracy and bias of the estimated breeding values were evaluated in a simulation study where various sizes of the reference population were considered. Subsequently, the genetic gain in the initial cycle of breeding programs was evaluated for several breeding schemes employing genomic or pedigree-based selection. A considerably higher genetic gain than with pedigree-based selection was achieved with genomic preselection, for which queens were genotyped early in life, and only the candidates of high genomic breeding value were admitted for mating or phenotyping.

In Chapter 4, pedigree-based and genomic breeding values were predicted on a data set of about 3000 genotyped queens for six economically relevant traits. Three traits showed significantly higher prediction accuracy with genomic compared to pedigree-based methods, and the differences between all the six traits could be explained mainly from their genetic parameters and the limited size of the reference population. The results show that genomic selection can be applied in honey bees, and the thesis provides appropriate breeding schemes and mathematical methods for its implementation.

Chapters 2, 3, and 4 have previously been published in, or submitted to peer-reviewed journals; the articles are referenced at the beginning of the respective chapters. In comparison with the

published or submitted versions, the layouts were adjusted to achieve a uniform appearance. This includes the formula notation and the display and numbering of tables and figures. The few cases where content was changed are clearly marked with explaining footnotes.

Zusammenfassung

Diese Arbeit präsentiert Ergebnisse der genomischen Zuchtwertschätzung für sechs wirtschaftlich bedeutende Merkmale bei der Honigbiene und untersucht verschiedene Strategien, um den optimalen Zuchtfortschritt durch genomische Selektion zu erreichen.

Genomische Selektion ist ein Routine-Verfahren bei verschiedenen Nutztierarten, aber noch nicht bei der Honigbiene wegen der Besonderheiten dieser Spezies. Kapitel 1 erläutert grundlegende Begriffe der genomischen Selektion und die Methode der besten, linearen, unverzerrten Vorhersage der Stammbaum-basierten Zuchtwerte bei der Honigbiene. Das Modell verwendet Stammbauminformationen in Form der genetischen Verwandtschaftsmatrix. Für die Honigbiene ist eine spezielle genetische Verwandtschaftsmatrix erforderlich, da die Paarungsbiologie dieser Spezies ungesicherte Vaterschaft, diploide Königinnen und haploide Drohnen umfasst. Kapitel 2 präsentiert einen neu-entwickelten Algorithmus zur effizienten Berechnung der Inversen der genetischen Verwandtschaftsmatrix und der Inzuchtkoeffizienten auf großen Datensätzen. In Kapitel 3 wird die Methode zur Berechnung der Inversen der genetischen Verwandtschaftsmatrix in Methoden zur Voraussage von genomischen und Stammbaum-basierten Zuchtwerten integriert. Die Genauigkeit und die Verzerrung der geschätzten Zuchtwerte wurden ausgewertet in einer Simulationsstudie, bei der verschiedene Größen der Referenzpopulation berücksichtigt werden. Anschließend wurde der Zuchtfortschritt im ersten Durchlauf von Zuchtprogrammen ausgewertet, die Zuchtschemata mit genomischer oder Stammbaum-basierter Selektion nutzten. Ein erheblich größerer Zuchtfortschritt als bei Stammbaum-basierter Selektion wurde mit genomischer Vorselektion erzielt, für die junge Königinnen genotypisiert wurden, und nur die Kandidaten mit den höchsten genomischen Zuchtwerten zur Anpaarung oder Leistungsprüfung zugelassen wurden.

In Kapitel 4 wurden Stammbaum-basierte und genomische Zuchtwerte für einen Datensatz von ungefähr 3000 genotypisierten Königinnen für sechs wirtschaftlich bedeutende Merkmale vorhergesagt. Drei Merkmale zeigten eine signifikant höhere Vorhersagegenauigkeit bei genomischer Zuchtwertschätzung gegenüber Stammbaum-basierten Verfahren und die Unterschiede zwischen allen sechs Merkmalen konnten im Wesentlichen aus den genetischen Parametern der Merkmale und der begrenzten Größe der Referenzpopulation erklärt werden. Die Ergebnisse zeigen, dass sich die genomische Selektion für die Honigbiene eignet, und die

Arbeit stellt angemessene Zuchtschemata und mathematische Methoden zu ihrer Umsetzung vor.

Kapitel 2, 3, und 4 sind bereits in referierten Fachzeitschriften veröffentlicht bzw. eingereicht worden. Die entsprechenden Literaturangaben finden sich am Anfang des jeweiligen Kapitels. Das Layout wurde gegenüber der veröffentlichen bzw. eingereichten Versionen angepasst, um ein einheitliches Erscheinungsbild zu gewährleisten. Angepasst wurden die Notation in Formeln und die Darstellung und Nummerierung von Abbildungen und Tabellen. Die wenigen Stellen, an denen inhaltliche Änderungen durchgeführt wurden, sind mit erläuternden Fußnoten gekennzeichnet.

Table of Contents

Acknowledgement

First, I would like to thank all my colleagues at the Institute for Bee Research Hohen Neuendorf. Most notably, I would like to thank my supervisor Prof. Bienefeld, whose continuous drive to improve bee breeding made my thesis possible, and who gave me great advice and support. Further, I profited immensely from the experience, patience, and brilliant ideas of Andreas Hoppe, Anja Strauß, Julia Jones, Gracie Zhipei Du, and Manuel Du. I cordially thank them for all their support.

Lastly, I wish to thank Vivian for being herself and staying by my side during the time of this thesis.

1 Introduction

1.1 Genomic selection

Traditionally, breeders record phenotypes, collect genetic data in the form of pedigrees, and select the parents of future generations accordingly to achieve genetic gain. This thesis focusses on quantitative traits, which include production traits such as honey yield in the honey bee, and the size, weight or predisposition to disease of breeding stock in general. The genetic loci influencing these traits, so-called quantitative trait loci (**QTL**s), can be genes or other loci of biological function. To genotype animals for QTLs efficiently, genetic markers are used which are inherited together with the QTLs, due to linkage disequilibrium. Genetic markers are mutations, and can be void of biological functions. From 1990 onward, advances in molecular genetics promised that genomic marker data could be integrated into breeding value estimations to achieve higher genetic gain (Fernando and Grossman 1989). Single-nucleotide polymorphisms (**SNP**s) proved to be useful as genetic markers, because they are abundantly present in the genome and can be automatically and cost-efficiently detected using SNP chips. After Meuwissen et al. (2001) described how to include genome-wide dense SNP data into breeding value estimation, genomic selection was implemented for several species, e.g. dairy cattle (Boichard et al. 2016), pigs (Samorè and Fontanesi 2016), Atlantic salmon and rainbow trout (Boudry et al. 2021).

1.1.1 Breeding schemes

The strategies for genomic selection vary across different species. In dairy cattle, pedigree-based selection relied on progeny-proven bulls but with genomic selection, young bulls could be selected based on their genotype. Using genomic breeding values shortened the generation interval and increased genetic gain sufficiently to outweigh the higher costs for genotyping (Doublet et al. 2019; García-Ruiz et al. 2016). By contrast, in pigs (Lillehammer et al. 2011), sheep (Lillehammer et al. 2020), or Atlantic salmon (Verbyla et al. 2018) the generation intervals could not be shortened significantly, instead the higher accuracy of genomic compared to pedigree-based selection convinced breeders to implement genomic selection.

1.1.2 Reference population

To guarantee a high accuracy of genomic breeding values, a large reference population is required, i.e. animals which are genotyped and phenotyped. With this data, the SNP chip can be calibrated by training the model with SNP genotypes and corresponding phenotypes. The effect of a SNP on a trait depends on the QTLs which are inherited together with the SNP, due to linkage disequilibrium. Over several generations of breeding, the linkage between SNPs and QTLs changes due to recombination events. Therefore, reference populations must be regularly updated to include relatives of the current selection candidates (Habier et al. 2007).

1.2 Mathematical models for breeding value estimation

Breeding value estimation provides a basis for breeding decisions by ranking the candidate animals according to their genetic quality. A plausible model for breeding value estimation should be based on phenotypes but account for the fact that even animals of high genetic value cannot achieve top performances in unfavourable environments. The model should rely on genetic data, e.g., the pedigree or genomic marker data, to provide mostly similar breeding values for relatives, but leave room for extreme differences even between breeding values of close relatives in exceptional circumstances. Such exceptional circumstances include, but are not limited to extreme differences in the phenotype which cannot be explained from different environments. The animal model (Mrode 2005) fulfils these conditions by decomposing the phenotype of animal i as

$$y_i = b_i + a_i + e_i,$$

where y_i and a_i are the phenotype and the breeding value of animal i, respectively, while b_i is an estimation of the effect of the herd, the year and the season of i on the phenotype of i, and e_i is the residual difference between the estimated effects and the phenotype. The pedigree information enters the model in the form of the numerator relationship matrix, $\boldsymbol{A}$, which is used to calculate the covariance matrix of the breeding values. Pedigree-based best linear unbiased prediction (**PBLUP**) is a mathematical method to estimate the breeding values according to the animal model (Henderson 1975; 1988a).

1.2.1 Calculation of the numerator relationship matrix

The coefficient of inbreeding, F_i, of animal i is a number between 0 and 1, which is closely connected to the pedigree-based relationship, A_{sd}, between the parents of i, denoted by s and

d, via $F_i = 0.5\, A_{sd}$ (Mrode 2005). This reflects the fact that closely related parents lead to inbred offspring. The diagonal entry, A_{ii}, of animal i is calculated as $A_{ii} = 1 + F_i$. Under the provision that animal j is no offspring of animal i, their pedigree-based relationship, A_{ij}, is calculated from the relationships of j to s and j to d, the sire and dam of i, respectively, via $A_{ij} = 0.5\,(A_{js} + A_{jd})$. The formula for A_{ij} reflects the fact that offspring inherit about half of their genome from each parent. E. g., if i and j are siblings with the same non-inbred and unrelated combination of sire and dam, then their relationship equals $A_{ij} = 0.5$, since the relationships of j to s and d are equal to 0.5.

1.2.2 Genomic estimation of breeding values

The mathematical method genomic BLUP (**GBLUP**) uses a genomic relationship matrix in the place of the pedigree-based relationship matrix to estimate breeding values for the genotyped animals (VanRaden 2008). In early versions of GBLUP, phenotypes of non-genotyped animals could not be used, and practical applications required specific adaptations. E.g., in early genomic selection of dairy cattle, only bulls were genotyped and only cows were phenotyped (VanRaden et al. 2009). Therefore, the estimation of breeding values required additional steps to equip bulls with pseudo-phenotypes. The combination of calculating pseudo-phenotypes and applying GBLUP has retrospectively been called "multi-step procedure", because it was succeeded by the method single-step GBLUP (**ssGBLUP**) (Christensen and Lund 2010; Legarra et al. 2009). The method ssGBLUP combines pedigree and genomic information in a combined numerator relationship matrix, performs better than the multi-step procedure, and has become standard methodology for breeding value estimation in several species (Legarra et al. 2014b).

1.3 The honey bee

1.3.1 Quantitative traits

Traditionally, beekeepers generate profit from the production of honey, beeswax, pollen, propolis or royal jelly. Other traditional aims of breeding are calmness during inspection, reduced stinging towards the beekeeper, and reduced swarming drive which refers to a behaviour of bees, where half of the colony relocates its nest (Petersen et al. 2020; Ruttner 1988; Uzunov et al. 2017). Over the last few decades, breeding against the parasitic mite *Varroa destructor* has become a priority, because the parasite contributes to high colony losses

(Genersch et al. 2010; Guichard et al. 2020a; Traynor et al. 2016). Furthermore, climate change has become a threat especially to honey bees, because their hives must be located in the open air to allow foraging (Flores et al. 2019; Le Conte and Navajas 2008; Vercelli et al. 2021). Genomic selection could be an effective tool to increase disease resistance and face climate change without sacrificing genetic progress in traditional aims of breeding.

1.3.2 Genome

The honey bee genome poses mixed conditions for genomic selection. On the one hand, at about 250 Mbp (mega base pairs) (Wallberg et al. 2019) the honey bee genome is short compared to most other livestock species (see Stapley et al. 2017) for a review on genome size and recombination rate across a wide range of species). Therefore, a rather small number of SNPs can densely cover the genome. On the other hand, the recombination rate is extremely high with estimates ranging from 19 cM/Mbp (centi-Morgan per mega base pairs) (Beye et al. 2006) to 37 cM/Mbp (Liu et al. 2015). Therefore, a SNP is unlikely to be linked to many QTLs, and the effects of SNPs on quantitative traits can change quickly from generation to generation due to recombination events. This suggests that large number of SNPs and a large reference population are required to guarantee a high accuracy of genomic breeding values.

1.3.3 Social behaviour

A fully developed hive consists of a single queen and several thousand workers (Koeniger et al. 2015). The main role of the queen is egg-laying which influences the number of workers, but she also uses pheromones to influence worker behaviour. All non-reproductive tasks in the hive, including cleaning, foraging and nursing offspring are performed by worker bees. During summer, the hive additionally contains a few thousand drones whose only function is reproduction. A virgin queen undertakes nuptial flights to mate with several drones. She collects the sperm in her spermatheca and uses it later to lay diploid eggs, which develop into workers or queens, depending on nutrition. She can also lay haploid eggs from which drones develop.

1.4   Breeding value estimation adapted to the honey bee

PBLUP has led to considerable genetic improvement in honey bees (Bienefeld et al. 2008; Hoppe et al. 2020) which was only possible after two main problems were solved.

1.4.1 First obstacle: mating biology

To achieve lasting genetic gain, the genetic quality of the drones mating a queen must be controlled (Plate et al. 2019b). This is usually realised by selecting a single queen of high breeding value and raising a number of her daughter-queens. These daughters are then mated and later placed as drone producing queens (**DPQ**s) on mating stations, often in islands or valleys, where other mated queens are not permitted to produce drones. Virgin queens, which are designated for breeding, are brought to these mating stations to mate with drones of high genetic quality. Since drones cannot be tracked, the exact number of drones a queen mated, and which DPQ the drones originate from, are unknown. Specialized methods have been developed to account for the mating biology of honey bees and the specialized mating system to correctly calculate the numerator relationship matrix A (Bienefeld et al. 2007; Brascamp and Bijma 2014).

1.4.2 Second obstacle: modelling a hive

For economically important traits in honey bees, a single phenotype is attributed to the hive which comprises various individuals, and the contributions of the queen and the workers must be distinguished. In honey yield for example, workers forage but the queen must lay new eggs to maintain a high number of workers. This situation is similar to maternally affected traits in other species, e.g., weaning weight in pigs, where the dam is tasked with nurturing her piglets which are phenotyped for weight gain. For the breeding value estimation, a direct genetic effect is modelled for each piglet, and a maternal effect for the sow. In honey bees, the phenotype of a colony c is attributed to a maternal genetic effect of the queen q, and a single direct effect for all workers in the colony comprised into a so-called worker group w according to

$$y_c = b_c + a_w + m_q + e_c,$$

where y_c is the phenotype of the colony, b_c is an estimation of the effect of the colony's apiary and year, a_w is the direct effect of the worker group, m_q is the maternal effect of the queen, and e_c is the residual.

1.5 State of the art and objectives

This thesis aimed to optimize the use of genome-wide dense marker (i.e. SNP) data for the breeding value estimation for large populations of honey bees. A simulation study previously showed that ssGBLUP is applicable to honey bees (Gupta et al. 2013), but recently the underlying model of genetic relationships in honey bees was revised, and especially the calculation of the numerator relationship matrix became more complex (Brascamp and Bijma 2014). The thesis developed new methods to combine ssGBLUP with the most recent numerator relationship matrix and enable the computation of pedigree-based and genomic breeding values in large populations of honey bees. The accuracy and the quality of the resulting genomic breeding values were evaluated by comparing them to pedigree-based breeding values. During the time of this thesis, simulation software, which considers the peculiarities of the honey bee, was developed with my participation, especially in the efficient computation of breeding values (Plate et al. 2019a). The simulation software was used for the initial validation of the methods developed in this thesis. Breeding schemes for genomic selection in honey bees were suggested in a small-scale simulation study (Brascamp et al. 2018). The thesis analysed the breeding schemes with different sizes of the reference population to determine the optimal use of genomic data for increased genetic gain. During the time of this thesis, a 100k-SNP-chip was developed and about 3000 queens were genotyped and phenotyped with my participation especially in the curation and analysis of the data (Jones et al. 2020). The reference population was gathered to enable the genomic selection of honey bees, and was used in this thesis to validate genomic prediction in the honey bee. This thesis aims to initiate genomic selection in the honey bee.

2 Computing inbreeding coefficients and the inverse numerator relationship matrix in large populations of honey bees

This is the accepted version of the following article:

Bernstein R, Plate M, Hoppe A, Bienefeld K. 2018. Computing inbreeding coefficients and the inverse numerator relationship matrix in large populations of honey bees. Journal of Animal Breeding and Genetics. 135:323-332. DOI: 10.1111/jbg.12347.

, which has been published in final form at [

https://onlinelibrary.wiley.com/doi/abs/10.1111/jbg.12347

]. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy [

https://authorservices.wiley.com/author-resources/Journal-Authors/licensing-open-access/open-access/self-archiving.html

].

The publication included four appendices which are included in this thesis as:

- Appendix A: Variance of Mendelian sampling terms
- Appendix B: Correction of two formulas in Brascamp and Bijma (2014)
- Appendix C: Inversion of D
- Appendix D: Storage and calculation of D

ABSTRACT

The inbreeding coefficients are considered in breeding decisions and the inverse numerator relationship matrix A^{-1}, is a prerequisite for breeding value estimation. Polyandry and haploid males are among the specifics of relationships between honey bees. Brascamp and Bijma (2014) averaged out the manifold possible relationships among honey bees that appear to have the same parents in a pedigree and assigned a single entry in A to animals that behave as a unit, e.g. the workers of a hive. Their methods of calculation connected full-sibs in the variance matrix of the Mendelian sampling terms D, via nonzero off-diagonal elements. This impedes the inversion of A and the closely connected calculation of inbreeding coefficients, because efficient algorithms for this task take D to be a diagonal matrix. Memory limitations necessitate their use for large data sets. We adapted the quickest of them to the block diagonal matrix D, that is postulated for the honey bee. To our knowledge, the presented algorithm is the first one that facilitates the method of Brascamp and Bijma (2014) on large data sets.

2.1   INTRODUCTION

The peculiarities of the honey bee require a special approach in the calculation of relationships and inbreeding coefficients. While the maternal side is in accordance with diploid farm animals, the paternal side contains systematic unknown parenthood and haploid males. In the most frequently used form of mating the resulting relationship between siblings is on average smaller than in diploid species. Therefore, Bienefeld et al. (2007) employed a reduced relationship of daughters to their sire, i.e. a reduced paternal path coefficient, in their calculation of the numerator relationship matrix, A. This method yielded the correct relationship between sisters, but underestimated the relationship to the sire and consequently the inbreeding. Brascamp and Bijma (2014) presented a new approach referred to here as **method BB**, which avoids these shortcomings. Both matrices were compared for their performance in the BLUP breeding value estimation by Brascamp and Bijma (2014), and method BB performed slightly better.

2.1.1   Literature on the inversion of A

Until now, method BB has not been applied to large data sets, because methods to invert A efficiently have not been developed yet. Henderson (1976) presented an efficient way to do this for diploid species via the following decomposition

$$A = TDT',\qquad\qquad(2.1)$$

where T follows directly from the pedigree, and D is a diagonal matrix. The ith diagonal entry of D is the variance of the Mendlian sampling term of individual i. Inbreeding of the parents of i reduces this variance, and matrix D can be calculated from the coefficients of inbreeding. Conversely, the diagonal entries of A minus 1 equal the coefficients of inbreeding, so the inbreeding coefficients can in turn be calculated from D via (2.1).

Quaas (1976) proposed a way of calculating A^{-1}, which used this interrelationship to reduce the required memory. Golden et al. (1991) improved the algorithm of Quaas (1976) to make it more efficient on large data sets by using a branching stack. Meuwissen and Luo (1992) developed the algorithm of Quaas (1976) in a different direction and outperformed Golden et al. (1991). Sargolzaei et al. (2005) presented an algorithm based on one by Colleau (2002), which minimized the computation time as well as memory requirements more than Meuwissen and Luo (1992). While Sargolzaei et al. (2005) aimed at calculating inbreeding coefficients,

their use of D makes it easy to obtain A^{-1} as well. We will refer to their method as **method SIC**.

The aim of this paper is to present a method for the rapid inversion of the relationship matrix of BB that also computes inbreeding coefficients in large populations of honey bees.

2.1.2   Special versions of A inbreeding diploid animals

Inverses of the numerator relationship matrix that account for unknown parenthood only or haplodiploidy only have already been presented. Henderson (1988b) proposed an average numerator relationship matrix to account for multiple sires in a herd. Bayesian approaches can be found in Famula (1992), where the formulas from Henderson (1988b) have been inserted into an algorithm for the calculation of A^{-1} from D.

Haplodiploidy can be dealt with by a gametic relationship matrix, which relates gametes rather than animals. This matrix was first described by Smith and Allaire (1985) who remarked that the methods of Henderson (1976) can be used for inversion. Schaeffer et al. (1989) gave a detailed account of this calculation and highlighted its uses for modeling X-chromosomal inheritance. Fernando and Grossman (1990) showed how to construct and invert a relationship matrix that relates animals, and can be used under haplodiploid inheritance.

2.1.3   Diagonality of D

In honey bees, the paternal descent can only be ascribed to a mixture of gametes from related sires. Therefore, D is a block diagonal matrix (Bienefeld et al. 1989). Bienefeld et al. (2007) reduced D to a diagonal matrix for reasons of efficiency, and were able to readily apply methods like SIC. In contrast, method BB includes off-diagonal elements between daughters of the same queen. This contrasts with the averaged relationship matrix of Henderson (1988b) as well as the relationship matrices accounting for haploidy (Fernando and Grossman 1990; Schaeffer et al. 1989). However, off-diagonal elements were used by Oikawa and Yasuda (2009), who when modeling the genome of cloned cattle, supposed partial genetic identity among animals. The main reason for the off-diagonal elements in honey bees is the omission of drones from the pedigree.

2.1.4 Reproduction of the honey bee

A honey bee queen mates only once in her life, before laying her first egg. Hence, all daughters and sons of a queen have the same possible sires. While she mates with 12 drones on average with differences across subspecies (Tarpy and Nielsen 2002), the precise number also depends on the mating location (Neumann et al. 1999a). Because a drone dies after mating, it can only mate with one queen. In contrast to the diploid queens and workers, drones are haploid. Daughters of the same drone share all of their paternal DNA, which makes them very closely related with an additive relationship of 0.75 without prior inbreeding, and are called supersisters.

In the most common form of controlled mating in Germany, virgin queens are brought to mating stations, where drones of several drone-producing queens (**DPQ**s) are present. To prevent foreign drones from intervening, mating stations are located in remote areas such as islands or valleys. The DPQs have a common dam, thus, there are three possible ways how the daughters of a queen can be related: they may be daughters of the same drone (supersisters, relationship 0.75), of different drones that share a common dam (full-sibs in the usual sense, 0.5), or of drones from different dams (half-sibs in the usual sense, 0.25, if the dams were unrelated).

2.2   MATERIALS AND METHODS

2.2.1   Method BB

Three types of individuals appear in the numerator relationship matrix, these are: (i) dams, (ii) sires, and (iii) worker groups. We will only refer to the queens that are not DPQs as dams. All DPQs at a mating station are a single sire (compare Figure 2.1), and all workers in a colony make up a single worker group. We will call the individuals of type (i-iii) breeding units to distinguish them from single bees.
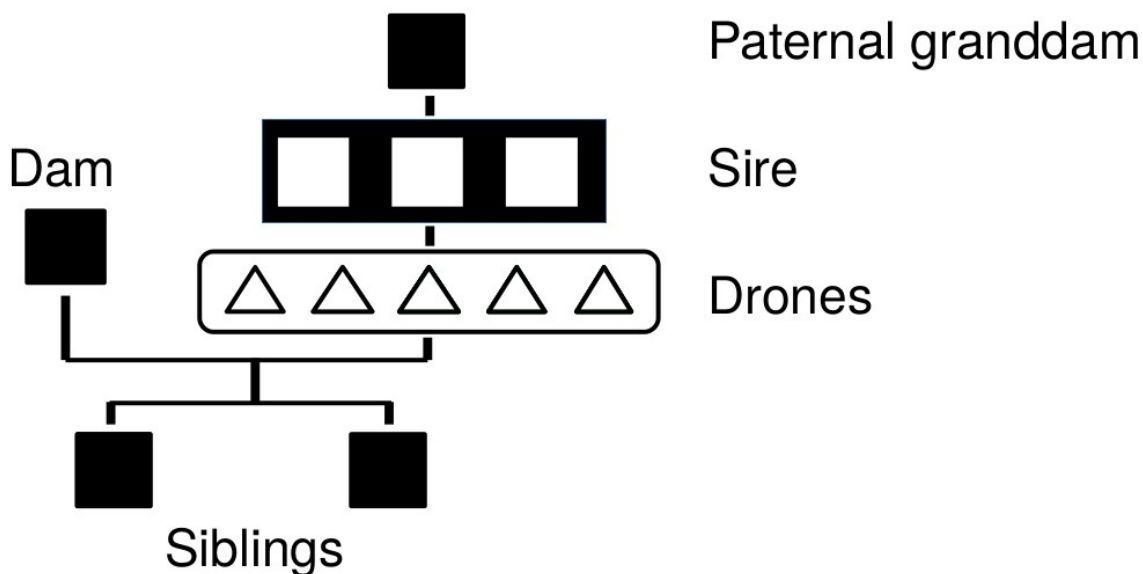
Figure 2.1 *The diagram shows two full-sibs and their parents. Solid symbols (■) represent individuals that are listed in the pedigree and the numerator relationship matrix, while hollow symbols (□) represent those that are not listed. Squares represent queens, while triangles represent drones, which are not listed. Multiple drone producing queens (DPQs) are recorded as a single sire in the pedigree and a single entry in the relationship matrix, therefore they are drawn into a solid rectangle.*

Method BB calculates A^{-1} using Henderson's decomposition (Henderson 1976) via

$$A^{-1} = (T^{-1})'D^{-1}T^{-1}, \tag{2.2}$$

where T^{-1} "is a lower triangular matrix with all 1's in the diagonal, and with the only nonzero elements to the left of the diagonal of the ith row being -.5 for columns corresponding to known parents" (Henderson, 1976, p. 74). While matrix D is a diagonal matrix under diploid inheritance. The elements of D are the within family segregation variances, also known as the variances of the Mendelian sampling terms.

Matrix D is a block diagonal matrix in honey bees, two breeding units of the relationship matrix belong to the same block in D if, and only if, they have the same dam. We will call the breeding units in a block a *full-sib group* as a technical term in the following text. Within a full-sib group

22

g, there are up to three types of diagonal elements corresponding to queens, sires, and the worker group, whereas the off-diagonal elements are mutually equal. The diagonal elements of $\boldsymbol{D}$ for a queen are given by

$$D_{ii} = D_g^Q, \tag{2.3}$$
<div align="right">(BB 10)</div>

where the superscript Q denotes queens. We link our equations to those in (Brascamp and Bijma 2014) by providing their equation number below ours with the prefix BB. The off-diagonal elements of $\boldsymbol{D}$ are equal to

$$D_{ij} = D_g^{FS}, \tag{2.4}$$
<div align="right">(BB 20)</div>

if the corresponding breeding units belong to the same full-sib group, g, and zero otherwise. The superscript FS denotes full-sibs. This is the covariance of the Mendelian sampling term between two bees that derive from the same dam and sire combination. A sire contains multiple bees, therefore, its diagonal entry in $\boldsymbol{D}$ is given by

$$D_{ii} = \frac{1}{n} D_g^Q + \frac{n-1}{n} D_g^{FS}, \tag{2.5}$$
<div align="right">(BB 17)</div>

where the sire contains n DPQs, and belongs to full-sib group g. In the diagonal entry D for a worker group, the variance of the Mendelian sampling term for a single bee vanishes, because there is a large number of workers in a hive (during summer between 30,000 and 50,000). The diagonal entry of $\boldsymbol{D}$ for a worker group belonging to full-sib group g is equal to

$$D_{ii} = D_g^{FS}. \tag{2.6}$$
<div align="right">(BB 19)</div>

The values of D_g^Q and D_g^{FS} for a full-sib group, g, can be obtained as follows: if the parents are known, then the variance term for a single bee is given by

$$D_g^Q = 1 - \frac{1 + F_d}{4} - \frac{1}{n_S} \frac{1 + F_s}{4} - \frac{n_S - 1}{n_S} \frac{F_s^{FS}}{2}, \tag{2.7}$$
<div align="right">(BB 13)</div>

where F_d is the inbreeding coefficient of the full-sib group's dam, F_s that of the group's sire, n_S the sire's number of DPQs, and F_s^{FS} the coefficient of coancestry between two of these DPQs,

which is half of their additive genetic relationship, called a_{SS} by Brascamp and Bijma (2014). The covariance term between two bees is equal to

$$D_g^{FS} = p_1 \frac{1 - F_s}{4} + \left(p_2 - \frac{1}{n_S}\right) \frac{1 + F_s - 2F_s^{FS}}{4}, \qquad \begin{array}{l} (2.8) \\ (\text{BB } 21) \end{array}$$

where p_1 denotes the probability that two bees belonging to g descend from the same drone, and p_2 the probability that they descend from the same DPQ (including the case where they descend from the same drone, so $p_2 \geq p_1$). Brascamp and Bijma (2014) assumed that the number of offspring per drone and the number of drones per DPQ follow a Poisson distribution. They arrived at

$$p_1 = \frac{1}{n_S} \text{ (BB 22a)} \quad \text{and} \quad p_2 = \frac{1}{n_D} + \frac{1}{n_S} \text{ (BB 23a)}, \qquad (2.9)$$

where n_D ist the number of drones with which the dam of g mated, and n_S the number of DPQs belonging to the sire of g. For cases where one or both parents are unknown, see Appendix A. Appendix B provides proofs for two formulas of Appendix A. The other formulas have already been proven by Brascamp and Bijma (2014).

We will obtain the coefficients of inbreeding and coancestry via SIC except for the coefficients of coancestry between the DPQs of a sire. The latter are given by

$$F_i^{FS} = \frac{1 + F_d}{8} + \frac{1}{2} F_i + \frac{1}{4} \left[p_1 + (p_2 - p_1) \frac{1 + F_s}{2} + (1 - p_2) F_s^{FS} \right], \qquad \begin{array}{l} (2.10) \\ (\text{BB } 24) \end{array}$$

where F_i is the inbreeding coefficient of the sire, F_d and F_s those of its parents, and F_s^{FS} the coefficient of coancestry of the sire in the former generation. The term in square brackets relates to the older sire, and is split into three addends: the first addend relates to the case where the DPQs descend from the same drone, the second to the one where they descend from different drones but the same DPQ, and the third relates to the case where they descend from different DPQs.

2.2.2   Method SIC

Method SIC determines the order in which the elements of D are calculated (Sargolzaei et al. 2005). It passes through all the sires, and calculates each sire's relationship to its mates. Let us

consider a particular sire s, and two different submatrices of A. Submatrix $A^{(1)}$ involves s and his ancestors while $A^{(2)}$ involves the mates of s and their ancestors, in addition to the breeding units in $A^{(1)}$. The two sets were introduced to better handle overlapping generations and are used for different tasks. We have

$$A^{(1)}x^{(1)} = y^{(1)}, \tag{2.11}$$

$$A^{(2)}x^{(2)} = y^{(2)}, \tag{2.12}$$

where $x^{(1)}$ and $x^{(2)}$ contain 1 at the position of s and 0 otherwise. Vectors $y^{(1)}$ and $y^{(2)}$ contain the relationships of s to the breeding units involved in $A^{(1)}$ and $A^{(2)}$ respectively. We aim to calculate $y^{(2)}$ as it contains the relationships between s and its mates, from which the inbreeding coefficients of the direct offspring of s can be calculated. These inbreeding coefficients are needed to calculate D.

In analogy to (2.1), the matrices $A^{(1)}$ and $A^{(2)}$ can be decomposed into $D^{(1)}$, $T^{(1)}$, and $D^{(2)}$, $T^{(2)}$. The idea behind SIC is to separate the multiplication of $A^{(2)}$ in (2.12) into three simple multiplications by $(T^{(2)})'$, $D^{(2)}$ and $T^{(2)}$, respectively. We start with the multiplication by $(T^{(2)})'$, consider two vectors $z^{(1)}$ and $z^{(2)}$ defined by

$$z^{(1)} = (T^{(1)})'x^{(1)}, \tag{2.13}$$

$$z^{(2)} = (T^{(2)})'x^{(2)}. \tag{2.14}$$

It can be shown that the vectors $z^{(1)}$ and $z^{(2)}$ have the same values for s and its ancestors, and that all the other entries are zero. Thus, the result of the multiplication by $(T^{(2)})'$ can even be obtained from a multiplication by $(T^{(1)})'$ which is easier. We obtain $z^{(1)}$ by solving

$$((T^{(1)})')^{-1}z^{(1)} = x^{(1)}, \tag{2.15}$$

which is feasible because $((T^{(1)})')^{-1}$ is upper triangular, via tracing back the pedigree. This is done by calculating $z_i^{(1)}$ from the younger breeding units up to the older breeding units according to

$$z_i^{(1)} = x_i^{(1)} + \frac{1}{2}\sum_{j \in O(i)} z_j^{(1)}, \tag{2.16}$$

where for every breeding unit i the set $O(i)$ includes all breeding units, that are direct offspring of i. This leaves us with the multiplications of $\boldsymbol{D}^{(2)}$ and $\boldsymbol{T}^{(2)}$ to $\boldsymbol{z}^{(2)}$. Vector $\boldsymbol{D}^{(2)}\boldsymbol{z}^{(2)}$ can be derived from $\boldsymbol{D}^{(1)}\boldsymbol{z}^{(1)}$. We solve

$$(\boldsymbol{T}^{(2)})^{-1}\boldsymbol{y}^{(2)} = \boldsymbol{D}^{(2)}\boldsymbol{z}^{(2)}, \tag{2.17}$$

by tracing forth the pedigree, which is feasible because $(\boldsymbol{T}^{(2)})^{-1}$ is lower triangular, and obtain $\boldsymbol{y}^{(2)}$ from which the inbreeding coefficients for the direct offspring of s are calculated.

Before we work our way through the pedigree sire by sire, we need to order the breeding units according to their sires. Method SIC has the following steps:

1. order the breeding units according to descendence;
2. reduce the pedigree by cutting out breeding units with no progeny;
3. sort the breeding units according to the number of their sire, putting breeding units with an unknown father at the start;
4. let i go through the breeding units
    4.1. put $F_i = 0$, if the sire of i is unknown;
    4.2. calculate the elements of $\boldsymbol{D}$ up to and including the sire of i, otherwise;
    4.3. obtain F for the progeny of the current sire.

2.2.3   Adaptation of SIC to the honey bee

Our aim is to use method SIC to calculate the inbreeding coefficients. The formulas from the previous subsection are valid for method BB. However, the new structure of the matrix of Mendelian sampling terms, $\boldsymbol{D}$, modifies how the pedigree is traced forth.

Consider a particular sire, s, whose relationships to his mates we want to calculate. Let $\boldsymbol{z}^{(1)}$ be the vector resulting from tracing back the ancestors of s via (2.16). Add zeros to $\boldsymbol{z}^{(1)}$ - so that it matches the length of the result of tracing back the ancestors of s and its mates. This gives us $\boldsymbol{z}^{(2)}$.

Let $\boldsymbol{A}^{(2)}$ be the numerator relationship matrix involving s, the mates of s and their ancestors. We aim to calculate the row in $\boldsymbol{A}^{(2)}$ that relates to s, and call it $\boldsymbol{y}^{(2)}$. We calculate the entry

relating to the breeding unit i from entries in $\boldsymbol{y}^{(2)}$, $\boldsymbol{z}^{(2)}$, and $\boldsymbol{D}$ relating to breeding units that are older than or full-sibs of s by

$$y_i^{(2)} = D_{ii} z_i^{(2)} + \sum_{j \in g} D_g^{FS} z_j^{(2)} + \frac{1}{2} y_d^{(2)} + \frac{1}{2} y_f^{(2)} , \qquad (2.18)$$

where D_g^{FS} is the off-diagonal element of $\boldsymbol{D}$ for the full-sib group that i belongs to, $y_d^{(2)}$ and $y_f^{(2)}$ are the elements of $\boldsymbol{y}^{(2)}$ for the parents of i. Equation (2.18) is the key result of this paper.

2.2.4   Combined algorithm

We use method BB to calculate matrix $\boldsymbol{D}$ inside of method SIC. To have $\boldsymbol{D}$ in the convenient shape of a block diagonal matrix, the breeding units are sorted in such a way that full-sibs follow each other directly. See Appendix D for details on storing $\boldsymbol{D}$. The inverse of the numerator relationship matrix is calculated according to the following steps:

1. carry out the steps 1.- 4. From subsection 2.2.2, "method SIC";
2. calculate the remaining elements of $\boldsymbol{D}$;
3. invert $\boldsymbol{D}$;
4. calculate $\boldsymbol{A}^{-1}$.

The previous subsection implies, that the combined algorithm yields the correct result. The algorithm was implemented by modifying the code provided in the appendix of (Sargolzaei et al. 2005). The sizes of the blocks to be inverted in step 3 are limited by the number of daughter queens a queen can produce (max. 224 in the BeeBreed data set). Therefore, numerical methods can be used. See Appendix C for an analytical approach using a result of Sherman and Morrison (1950).

2.2.5   Simulation studies

To evaluate the algorithm's running time, artificial pedigrees were created. The program simulated populations over 10, 20, and 40 years, in each year there were either 3000 dams and 100 sires, or 2950 dams and 200 sires. A worker group was included for each queen that was not a DPQ. The number of dams in a single full-sib group was fixed at either 4 or 8 per year. In one scenario, a maximum of 2 dams could be selected from a full-sib group, in the other scenario

all could be selected. A single trait model was simulated, and breeding values were estimated as suggested by Brascamp and Bijma (2014). The additive genetic variances were set to 1 for the maternal effects, to 2 for the direct effects, and to -0.5 for their covariance.

The average generation numbers of the animals in the last year were calculated as described by Meuwissen and Luo (1992). Computations were carried out on a i7-4790 Intel, 3.6 gigahertz processor under Ubuntu 16.04 LTS, 64 bit (15.6 gigabyte RAM).

2.3 RESULTS AND DISCUSSION

2.3.1 Properties of the implementation

Table 2.1 presents the computing properties of the combined algorithm. The time required to calculate the inbreeding coefficients was measured. The memory required was nearly the same for both the calculation of the inbreeding coefficients and the inversion of A. It increased linearly with the size of the input data.

The computation of the inbreeding coefficients took longer when more sires per generation were involved, as the pedigree had to be traced twice for each sire. This was also true for the algorithm of Sargolzaei et al. (2005).

The time required to compute A^{-1} was almost exclusively influenced by the size of the data set. This method was tested with 240,000 breeding units on a desktop computer, which was completed in several seconds. As the computation time increased by a factor of 4 when the number of breeding units was doubled, a data set consisting of a million breeding units would be finished within 45 minutes. The program is easily transferrable to a server environment, which enables even faster computation.

When the full-sib groups were larger, the computation of the inbreeding coefficients took longer. The reason might, be that the tracing forth of the pedigree picked up more breeding units because selected dams had more offspring. The time required for tracing back the pedigree was not altered.

Table 2.1 *Time and memory (in Megabyte) required to calculate the coefficients of inbreeding were measured to compare the algorithms of Sargolzaei et al. (2005) (ISa) and Meuwissen and Luo (1992) (IMe). Time and memory required for the inversion of the numerator relationship matrix were only measured for the algorithm of Sargolzaei et al. (2005) (AiSa). Avg. stands for average.*

Eligible full-sibs[†]	Family size[‡]	Years evaluated	No. of breeding units[§]	Avg. no. of generations[¶]	Avg. In-breeding	Computing time (s)			Memory requirements (MB)		
						ISa	IMe	AiSa	ISa	IMe	AiSa
100 sires per year and 3000 dams per year											
2	4	10	60,800	3.3	0.0013	0.6	1.4	7.3	16.3	15.2	16.3
2	4	20	121,800	7.7	0.0040	1.7	6.8	30.2	30.5	28.5	30.7
2	4	40	243,800	14.9	0.0103	5.9	122.4	140.8	59.0	55.1	59.4
2	8	10	60,800	3.1	0.0012	0.7	1.4	7.2	16.2	15.2	16.2
2	8	20	121,800	6.7	0.0036	1.9	6.6	29.8	30.6	28.3	30.5
2	8	40	243,800	13.9	0.0084	8.2	97.3	144.4	59.1	55.2	59.0
4	4	10	60,800	3.4	0.0017	0.6	1.4	7.2	16.4	15.2	16.4
4	4	20	121,800	7.3	0.0039	1.5	6.8	30.2	30.6	28.5	30.5
4	4	40	243,800	15.2	0.0104	5.7	128.0	149.3	59.4	55.2	59.0
8	8	10	60,800	3.4	0.002	0.7	1.4	7.3	16.3	15.2	16.2
8	8	20	121,800	7.3	0.0042	2.1	6.8	29.6	30.5	28.2	30.5
8	8	40	243,800	15.1	0.0114	7.3	112.2	144.8	59.2	54.9	59.2
200 sires per year and 2950 dams per year											
2	4	10	60,600	3.3	0.0008	1.0	1.4	7.6	16.4	15.1	16.3
2	4	20	121,600	7.1	0.0023	3.1	6.5	31.8	30.6	28.4	30.7
2	4	40	243,600	14.7	0.0064	12.0	200.1	153.0	59.1	55.2	59.1
2	8	10	60,600	3.1	0.0010	1.1	1.4	7.8	16.3	15.1	16.2
2	8	20	121,600	6.6	0.0022	4.3	6.6	31.9	30.6	28.3	30.6
2	8	40	243,600	13.8	0.0058	15.8	143.2	160.0	59.2	54.9	59.2
4	4	10	60,600	3.4	0.0090	0.9	1.4	7.6	16.2	15.0	16.2
4	4	20	121,600	7.3	0.0029	2.9	6.6	32.0	30.7	28.5	30.6
4	4	40	243,600	15.1	0.0066	12.0	214.9	158.6	59.1	55.3	59.1
8	8	10	60,600	3.4	0.0013	1.1	1.4	7.7	16.2	15.3	16.2
8	8	20	121,600	7.3	0.0029	4.0	6.6	31.6	30.6	28.4	30.5
8	8	40	243,600	15.1	0.0079	15.0	173.4	152.7	59.0	55.0	59.0

The maximum number of members of the same full-sib group that were used as dams is shown under "max. siblings in breeding," and did not have a significant impact on computation time or memory.

An adaptation of the approach of Meuwissen and Luo (1992) to the honey bee was also implemented, but the time required to compute the inbreeding coefficients on pedigrees with more than 240,000 breeding units ranged between 100 and 200 seconds. This is in accordance with the findings of Sargolzaei et al. (2005), therefore, we did not investigate this method any further.

Comparison to Bienefeld et al. (2007)

The BeeBreed data set was used to compare method BB to the method of Bienefeld et al. (2007) which we will abbreviate by **method BER.** The pedigree contains 202,602 dams. Only the inbreeding coefficients of queens were evaluated. The full-sib group size was 9.25 on average with a maximum of 224.

BER yields in some short pedigrees higher inbreeding coefficients because the relationship of a breeding unit to its sire without prior inbreeding is higher in BER (0.367) than in BB (0.204). However, BB tends to yield higher values especially on long pedigrees. The results of BB are more plausible, because the reduced paternal path coefficient reduces not only the relationship of a breeding unit to its sire but to all the ancestors of the sire too.

E.g., the relationship of a queen to its paternal granddam is 0.25 in BB without prior inbreeding, which is plausible because the granddam passes a quarter of its genome on to the queen. The same relationship is put to 0.184 in BER to achieve a plausible relationship among full-sibs.
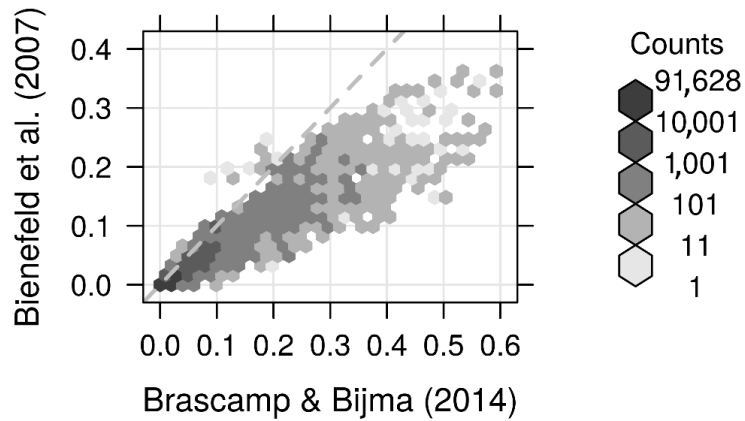
Figure 2.2 *Inbreeding coefficients were calculated according to BB and BER. The colours code how often a combination occurred. The highest inbreeding coefficient was 0.593 by method BB and 0.367 according to BER. The average inbreeding in 2016 was 0.0419 according to BB and 0.023 according to BER.*
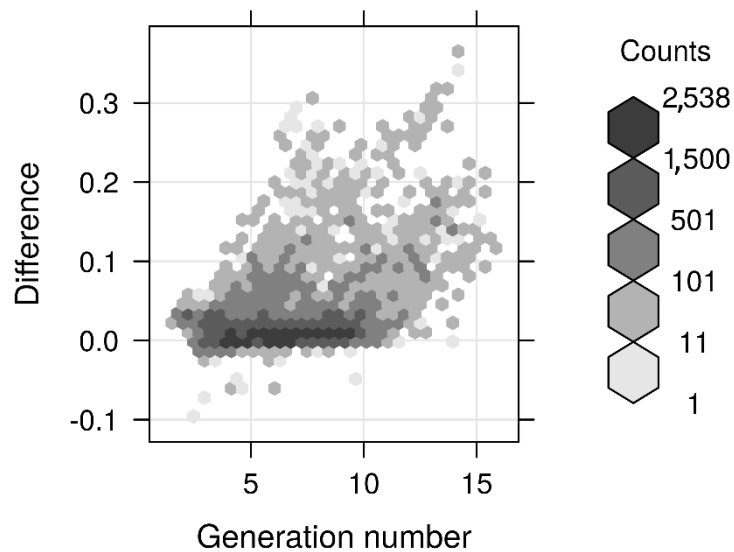
Figure 2.3 *The inbreeding coefficients calculated by BER were subtracted from those calculated by BB for animals that did have a positive inbreeding coefficient. Inbreeding coefficients according to BB were at maximum 0.365 higher than those of BER, while the results of BER were at maximum 0.096 higher than those of BB.*

2.3.2 Other kinds of mating

Various mating situations demand further development of method BB. On most mating stations the presence of foreign drones is possible, although it is usually tested before use (Jensen et al. 2005; Neumann et al. 1999b). Foreign drones reduce the relationship of a sire to its putative daughters, and so does the paternal path coefficient in BER. Thus, BER might fit these insecure mating stations. In BB, a ratio of foreign drones could be taken into account by the introduction of a correction term.

Artificial insemination can already be accounted for by BB, if the drones are taken from DPQs whose performance is not recorded. This excludes single drone insemination suggested by Harbo and Harris (1999) for furthering hygienic behavior in workers, but this situation is less complex than when DPQs are present. When mixed semen from multiple recorded queens is used, as Spivak and Reuter (2001) did, formulas need further adaptation.

On Varroa tolerance mating stations (Büchler et al. 2010), the situation is more complex than it is sketched in Figure 2.1 because the DPQs have various possible dams. To account for this, new terms would have to be added to the formulas presented here. Furthermore, the reliability of the estimated relationships would be considerably reduced as more possibilities become involved.

2.4 CONCLUSION

Method BB promised more accurate relationships, and consequently more precise breeding values and improved selection. Therefore, we faced a challenge to develop a memory efficient algorithm. To our knowledge, the presented algorithm is the first one that facilitates method BB on large data sets. Method SIC from Sargolzaei et al. (2005) was adapted to the peculiarities of the honey bee. Now the off-diagonal elements of D are taken into account when the pedigree is traced forth. The running time was influenced by the proportional number of sires but not the size of memory required. Theoretical advantages of method BB can now be verified in large simulations, or breeding programs.

3 Simulation studies to optimize genomic selection in honey bees

This paper has originally been published as:

The publication included eight appendices which are included in this thesis as:

- Appendix E: Merging DPQs for ssGBLUP$_{DPQ+BQ}$

- Appendix F: Accuracy for replacement queens

- Appendix G: Comparison of simulated and estimated genetic gain (R_{PB}) from year 8 to year 9 in the simulated breeding population relying on PBLUP

- Appendix H: Accuracies of breeding values when all BQs from years 4 to 9 were genotyped

- Appendix I: Genetic gain, R_{GS}, in the initial selection cycle of different breeding schemes applying CBS and GPS; including the additional file "Appendix_I.xlsx"

- Appendix J: Standard deviations of true breeding values

- Appendix K: Regression coefficients of true on estimated breeding values, b_1, when all queens from years 4-9 were genotyped

- Appendix L: Genetic in the first generation of a genomic breeding program applying a shorter generation interval

ABSTRACT

With the completion of a single nucleotide polymorphism (SNP) chip for honey bees, the technical basis of genomic selection is laid. However, for its application in practice, methods to estimate genomic breeding values need to be adapted to the specificities of the genetics and breeding infrastructure of this species. Drone-producing queens (DPQ) are used for mating control, and usually, they head non-phenotyped colonies that will be placed on mating stations. Breeding queens (BQ) head colonies that are intended to be phenotyped and used to produce new queens. Our aim was to evaluate different breeding program designs for the initiation of genomic selection in honey bees.

Stochastic simulations were conducted to evaluate the quality of the estimated breeding values. We developed a variation of the genomic relationship matrix to include genotypes of DPQ and tested different sizes of the reference population. The results were used to estimate genetic gain in the initial selection cycle of a genomic breeding program. This program was run over six years, and different numbers of genotyped queens per year were considered. Resources could be allocated to increase the reference population, or to perform genomic preselection of BQ and/or DPQ.

Including the genotypes of 5000 phenotyped BQ increased the accuracy of predictions of breeding values by up to 173%, depending on the size of the reference population and the trait considered. To initiate a breeding program, genotyping a minimum number of 1000 queens per year is required. In this case, genetic gain was highest when genomic preselection of DPQ was coupled with the genotyping of 10 to 20% of the phenotyped BQ. For maximum genetic gain per used genotype, more than 2500 genotyped queens per year and preselection of all BQ and DPQ are required.

This study shows that the first priority in a breeding program is to genotype phenotyped BQ to obtain a sufficiently large reference population, which allows successful genomic preselection of queens. To maximize genetic gain, DPQ should be preselected, and their genotypes included in the genomic relationship matrix. We suggest, that the developed methods for genomic prediction are suitable for implementation in genomic honey bee breeding programs.

3.1 BACKGROUND

Currently, genomic selection is applied in various livestock species (Fulton 2012; Samorè and Fontanesi 2016; Wiggans et al. 2017) but not in honey bees. Although honey bees contribute to agriculture as a key pollinator (Gallai et al. 2009), adaptation of modern breeding methods to apiculture is comparatively slow. Systematic collection of performance and pedigree data on honey bees in Germany started in the 1950s (Bienefeld and Pirchner 1990) and the estimation of best linear unbiased prediction (BLUP) breeding values began in 1994 (Bienefeld et al. 2007; Hoppe et al. 2020), but to date honey bee breeding programs do not use genomic marker data (Uzunov et al. 2017). Recently, cost-efficient methods for the collection of genomic data in honey bees have become available in the form of a high-density 100K single nucleotide polymorphisms (SNP) chip (Jones et al. 2020).

Progress in other livestock is faster than in honey bees, which is due to the specific reproduction characteristics of honey bees. Queens can only mate during the first weeks of their life, and at this time, the unfertilized queens undertake nuptial flights during which they mate with several drones (Tarpy and Nielsen 2002), and store the drones' sperm in their spermatheca for subsequent use to fertilize eggs. Under normal circumstances, workers do not reproduce (Koeniger et al. 2015). Queens and workers hatch from fertilized eggs, while drones hatch from unfertilized eggs. The consequence of the honey bee's reproductive biology is uncertain paternity. To alleviate the resulting practical problems, controlled mating is applied in various honey bee populations (Brascamp et al. 2016; Guichard et al. 2020b; Hoppe et al. 2020), where unfertilized queens are brought to mating stations to mate with drones. Mating stations are located in isolated areas, which in practice are often islands or valleys, where only drones from selected colonies headed by drone-producing queens (DPQ) are available. Typically, all DPQ on a mating station share a single dam, which restricts their genetic diversity. We refer to a group of DPQ on a mating station as a pseudo-father (Bienefeld et al. 1989). In practice (Hoppe et al. 2020), DPQ are at least one year old when they are deployed on mating stations.

Performance testing for relevant traits, such as honey yield, gentleness, or disease resistance, is only possible when the colony is fully developed. A colony contains up to 50,000 workers during spring, which are usually all offspring of the same queen and considered as a single worker group. The workers perform a wide range of tasks, such as foraging, cleaning, and feeding the queen, drones, and larvae. Collecting data on all the relevant traits is usually

completed in the year after the queen hatches. All queens that were or will be performance-tested, and queens which are or were candidates for a preselection step before phenotyping, are referred to as breeding queens (BQ). We call the selection of phenotyped queens 'colony-based selection' (CBS), since phenotyping requires a colony. Figure 3.1 shows the set-up of classic CBS, demonstrating that DPQ and unfertilized BQ are promising candidates for genomic selection. For BQ, genomic selection could be efficient before fertilization, which would save the costs of phenotyping. We call this step 'genomic preselection' (GPS) because it is applied before mating during the life of a BQ, and before deployment on mating stations for DPQ. In schemes with controlled mating, DPQ are usually not performance-tested, and only BQ can be selected as dams of BQ or as dams of DPQ. Although some populations are built without controlled mating (Andonov et al. 2019; Maucourt et al. 2020), in our work, we considered only populations with controlled mating, because it increases genetic gain (Plate et al. 2019b).

The use of genomic selection in honey bee breeding programs is expected to enhance genetic gain. Gupta et al. (2013) reported considerably improved prediction accuracies using a single-step approach (Christensen and Lund 2010; Legarra et al. 2009) and a genetic model that was honey-bee-specific, but the underlying theory was considerably improved in a later study (Brascamp and Bijma 2014). E.g., because the queen and her worker group contribute to the phenotype of a colony, both the direct effect of the worker group and the maternal effect of the queen should be linked to the phenotype. In (Gupta et al. 2013), phenotypes were linked to the genetic effects of the queens, but not of the worker groups. To address this issue and provide realistic estimates of the accuracies of maternal and direct effects, we used recently published software that accommodates these effects (Plate et al. 2019a).

To estimate genomic breeding values in honey bees (Gupta et al. 2013), the genotypes of phenotyped BQ are required. Each DPQ has a large number of offspring but is not performance-tested in the scenarios under consideration, a situation which is similar to that of sires in mammalian species, for which it is usually more useful to genotype the males than the females (Howard et al. 2018). Therefore, we examined whether the genotypes of DPQ can be used together with the genotypes of BQ to increase the prediction accuracy for DPQ and BQ.
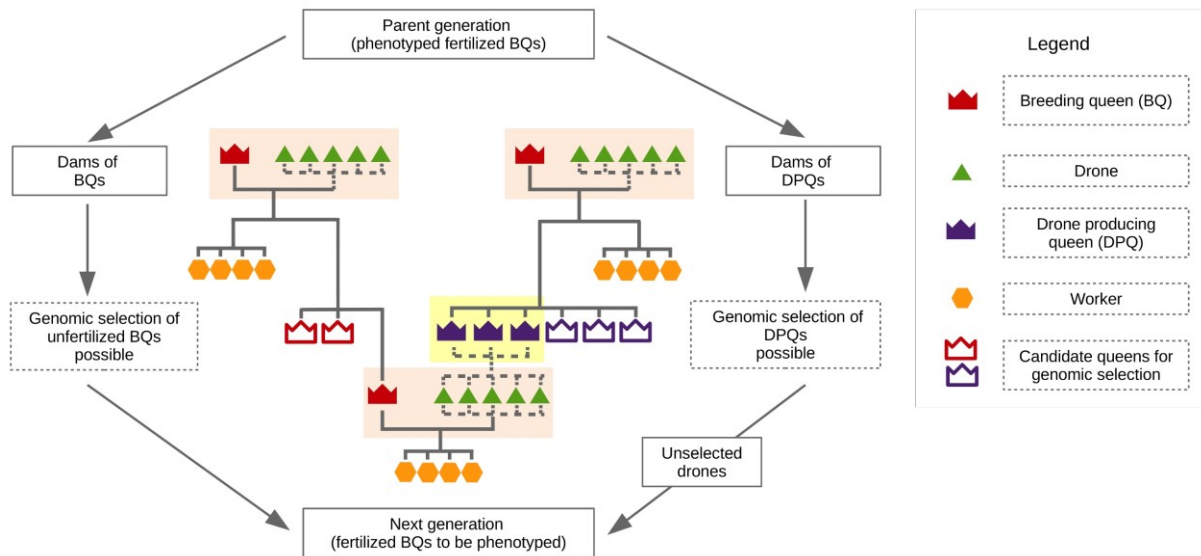
Figure 3.1 Population structure of a honey bee population under controlled mating. *The parent generation consists of phenotyped colonies headed by fertilized queens (highlighted in beige). Each fertilized queen mates with several drones and produces a worker group. Dams of breeding queens (BQ) and drone producing queens (DPQ) are selected from the phenotyped colonies. With genomic selection, a larger number of queens is reared per dam, and the daughters with the highest genomic estimated breeding value are kept. No preselection among queens is applied in the pedigree-based breeding scheme. Sister DPQ are deployed together on a mating station and form a pseudo-father (highlighted in yellow). BQ are brought to mating stations to mate drones. This enables the fertilized BQ to produce a worker group and the colony is later phenotyped.*

A sufficiently large reference population must be gathered and maintained to ensure accurate genomic prediction of breeding values (Sonesson and Meuwissen 2009). Brascamp et al. (2018) addressed the design of breeding programs in honey bees, and their results suggested that genomic selection applied sequentially to several generations of virgin queens results, by far, in the highest genetic gain. However, the authors did not investigate what size of the reference population was most appropriate Brascamp et al. (2018). In our study, we assumed that only a limited number of queens could be genotyped and we analyzed the trade-off between primarily genotyping queens with a phenotype to enlarge the reference population, or investing in preselection of BQ and/or DPQ. The optimum values in this trade-off depend on the costs of genotyping the queens and the profit from the greater genetic gain from genomic selection. It

is difficult to assess the monetary value of some traits that are under selection in honey bees, such as gentleness or disease resistance. Therefore, we did not attempt an economic calculation, but focused on response to selection.

The aim of our simulation study was to provide insights into the optimization of genomic selection for honey bee breeding programs by using stochastic simulations to generate a breeding population. Genomic breeding values were estimated for the population using either the genotypes of BQ only, or the genotypes of both DPQ and BQ. The obtained accuracies of genomic predictions allowed us to compare the quality of the different analyses. Genetic gain was predicted by using a deterministic model for breeding schemes. We implemented GPS of BQ and/or DPQ with different selection intensities. Furthermore, the breeding schemes under consideration covered different budgets for genotyping queens, and different sizes of the reference population.

3.2   METHODS

Our study can be divided into two steps, each using very different methods. In the first step, honey bee populations were stochastically simulated using the program BeeSim (Plate et al. 2019a). The breeding value estimation for the last generation was subsequently repeated with pedigree-based and genomic methods to evaluate the quality of resulting the breeding values. In the second step, we used the accuracies obtained from the stochastic simulation as input to predict genetic gain, for which we used a deterministic model, since the BeeSim program does not accommodate a preselection step.

3.2.1   Model and selection criteria

The phenotypes of economically-relevant traits of a honey bee colony are influenced by the queen and her workers. While all non-reproductive tasks are performed by the workers, the egg-laying rate of the queen is one example of her essential qualities, since it is crucial for the number of workers. Therefore, in honey bees, the genetic model for most traits includes direct effects due to the contribution of the workers and maternal effects due to the contribution of the queen and the phenotype, y, of a colony, C, is modeled as follows:

$$y = \overline{a_W} + m_Q + e, \tag{3.1}$$

where $\overline{a_W}$ is the average of the direct effects of the workers in C, and m_Q the maternal effect of the queen in C, and e is a non-heritable residual.

For a queen, Q, the selection criterion in CBS is equal to the sum of the estimated breeding values (EBV) of the maternal and direct effects of Q's worker group, and in GPS it is equal to the sum of the EBV for the direct and maternal effects of Q. This choice is motivated by the reproductive biology of honey bees. The maternal and direct effects of Q do not account for the quality of the drones she mated with. By contrast, the maternal and direct effects of Q's worker group reflect the genetic quality of Q and of the drones with which she mated. Therefore, the selection criterion for fertilized BQ is equal to the sum of the EBV for the direct and maternal effects of their worker groups (see (Brascamp and Bijma 2014; Brascamp et al. 2016) for more details justifying the selection criterion in CBS). In GPS, the sum of the EBV for the direct and maternal effects of a BQ serves as the selection criterion for unfertilized BQ. In GPS and CBS, the selection criterion for DPQ is equal to the sum of the EBV for the maternal and direct effects of the DPQ, because DPQ are selected for their drones that hatch from unfertilized eggs.

3.2.2 Scenarios for breeding schemes

The number of BQ per year was set to 1000. BQ were mated in the year they were born and tested in the next year, when their colonies were fully developed. At the age of one year, DPQ were deployed on mating stations. At each mating station, 8 daughters of a single dam were placed. Figure 3.2 (a) illustrates the classic CBS, in which the top 200 of the 2-year-old BQ were selected as dams of BQ (selection intensity $i_{BQ}^{CBS} = 1.40$) and the top 50 were selected as dams of DPQ (selection intensity $i_{DPQ}^{CBS} = 2.06$). For GPS, specific dams of BQ or dams of DPQ were assumed to produce more offspring than required in the classic CBS to allow genomic preselection. The candidate BQ and candidate DPQ were genotyped, and 5 BQ and 8 DPQ were kept for each dam of BQ and dam of DPQ, respectively.

The initial selection cycle in a genomic breeding program using GPS, as illustrated in Figure 3.2 (b), was examined for different numbers of genotyped queens per year, n_{gpy}, ranging from 0 to 4000 in steps of 500. In the first step of the genomic breeding program, phenotyped colonies were selected based on genomic EBV using phenotypic information, as appropriate, and a

relationship matrix based on genomic and pedigree relationships. In the second step, unfertilized queens or DPQ were preselected based on their genomic estimated breeding value.
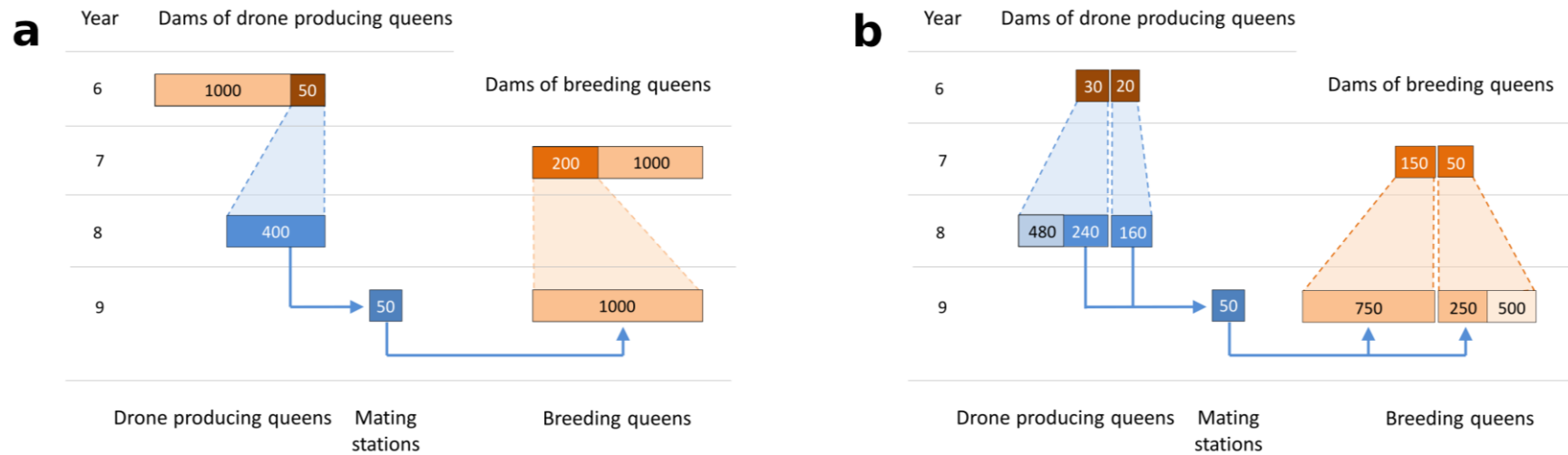
Figure 3.2 Pathway-model of pedigree-based selection (a) and genomic preselection of queens (b) in year 9 as an example.

(a) In year 8, the top 50 of the 2-year-old breeding queens (BQ) were selected as dams of drone-producing queens (DPQ; selection intensity $i_{DPQ}^{CBS} = 2.06$). 400 DPQ were reared in year 8 and deployed on 50 mating stations in year 9. In the same year, the top 200 of the 2-year-old BQ were selected as dams of BQ (selection intensity $i_{BQ}^{CBS} = 1.40$), and 1000 BQ were reared from them. The new BQ were mated on the 50 mating stations.   (b) From the 50 dams of DPQ, $N_{DPQ}^{GPS} = 30$ were chosen for preselection based on genomic estimated breeding values (EBV, $p_{DPQ} = 0.6$), and each produced $n_{DPQ}^{GPS} = 16$ candidate DPQ. The 240 candidate DPQ with the highest genomic EBV were selected as DPQ. Each group of eight DPQ was deployed on a separate mating station. From the 20 dams of DPQ not chosen for preselection, 160 DPQ were reared and groups of 8 sister DPQ were deployed on mating stations. From the 200 dams of BQ, $N_{BQ}^{GPS} = 50$ were chosen for preselection based on genomic EBV and each produced $n_{BQ}^{GPS} = 10$ daughters. The 250 candidate BQ with the highest genomic EBV were selected to be mated and later phenotyped. This left $N_{rest} = 20$ open slots to genotype more phenotyped BQ. Consequently, the proportion of BQ in the reference population per year was $p_{ref} = 0.27$.

Among the 50 selected dams of DPQ, N_{DPQ}^{GPS} (ranging from 0 to 50 in steps of 1) dams were chosen for GPS. From a dam of DPQ chosen for GPS, n_{DPQ}^{GPS} (ranging from 9 to 64 in steps of 1) daughter queens were reared and genotyped. The top $8N_{DPQ}^{GPS}$ of all candidate DPQ were deployed on a mating station (with a selection intensity i_{DPQ}^{GPS} that ranged from 0.21 to 1.65). The proportion of DPQ selected by GPS compared to all DPQ deployed on mating stations (p_{DPQ}) is given by $p_{DPQ} = 8N_{DPQ}^{GPS}/400$, which ranged from 0 to 1 in steps of 0.02. Of the 200 selected dams of BQ, N_{BQ}^{GPS} (ranging from 0 to 200 in steps of 1) dams were chosen for GPS. From a dam of BQ chosen for preselection, n_{BQ}^{GPS} (ranging from 6 to 32 in steps of 1) daughter queens were reared and genotyped. The top $5N_{BQ}^{GPS}$ candidate BQ were kept for phenotyping (with a selection intensity i_{BQ}^{GPS} that ranged from 0.30 to 1.53). The proportion of BQ selected by GPS compared to all phenotyped BQ (p_{BQ}) is given by $p_{BQ} = 5N_{BQ}^{GPS}/1000$, which ranged from 0 to 1 in steps of 0.005.

With a total genotyping capacity of 1000, the number of remaining animals to genotype is $N_{rest} = 1000 - n_{BQ}^{GPS}N_{BQ}^{GPS} - n_{DPQ}^{GPS}N_{DPQ}^{GPS}$. We assumed that these N_{rest} genotypes were allocated to phenotyped BQ and that this was also the case in previous years. To obtain the proportion of BQ in the reference population per year, p_{ref}, the BQ selected by GPS were added, thus $p_{ref} = (N_{rest} + 5N_{BQ}^{GPS})/1000$. Combinations of parameter values with $n_{gpy} < n_{BQ}^{GPS}N_{BQ}^{GPS} + n_{DPQ}^{GPS}N_{DPQ}^{GPS}$ were not evaluated.

In total, 14 million scenarios were evaluated. The parameters of the evaluated breeding schemes are shown in Table 3.1.

3.2.3 Scenarios for breeding value estimation

To evaluate the quality of the EBV and estimate genetic gain, a breeding population under classic CBS was simulated over 10 years (Fig.2a). BQ were randomly assigned to mating stations, but BQ that shared a common dam were assigned to the same mating station. Each BQ was mated with 12 drones. The dam of each drone was randomly sampled from the 8 DPQ deployed on the mating station. Pedigree was recorded for BQ. In the first years, BQ were too young to be selected. The DPQ on mating stations in years 0 to 2 were unrelated and not

recorded in the pedigree. Dams of BQ were selected from year 2 onwards and dams of DPQ were selected from year 3 onwards.

Table 3.1 *Parameters evaluated for breeding schemes with genomic preselection (GPS).*

Number of genotyped queens per year (n_{gpy})	0 to 4000 in steps of 500
Number of dams of drone producing queens (DPQ) chosen for GPS (N_{DPQ}^{GPS})	0 to 50 in steps of 1
Number of dams of breeding queens (BQ) chosen for GPS (N_{BQ}^{GPS})	0 to 200 in steps of 1
Number of candidate DPQ per dam for GPS (n_{DPQ}^{GPS})	9 to 64 in steps of 1
Number of candidate BQ per dam for GPS (n_{BQ}^{GPS})	6 to 32 in steps of 1
Proportion of DPQ selected by GPS compared to all DPQ deployed on mating stations (p_{DPQ})	0 to 1 in steps of 0.02
Proportion of BQ selected by GPS compared to all phenotyped BQ (p_{BQ})	0 to 1 in steps of 0.005
Selection intensity of GPS on DPQ (i_{DPQ}^{GPS})	0.21 to 1.65
Selection intensity of GPS on BQ (i_{BQ}^{GPS})	0.30 to 1.53

To build the simulated population, breeding values were estimated via pedigree-based BLUP (PBLUB), i.e. BLUP without marker data. When the simulation of the population over 10 years was complete, breeding values were estimated with PBLUP and single-step genomic BLUP. Two versions of the genomic relationship matrix, $\mathbf{G}_{BQ}$ and $\mathbf{G}_{DPQ+BQ}$, were used, leading to two analyses, i.e. ssGBLUP$_{BQ}$ and ssGBLUP$_{DPQ+BQ}$, respectively. Matrix $\mathbf{G}_{BQ}$ included only the genotypes of BQ and $\mathbf{G}_{DPQ+BQ}$ included the genotypes of DPQ and BQ.

The final pedigree contained 10,000 BQ, with 10,000 worker groups, and 2800 DPQ on 350 mating stations, since we did not include mating stations from years 0 to 2 in the pedigree. Worker groups from year 8 represented phenotyped colonies. BQ from year 9 represented unphenotyped queens, but had worker groups. Since unphenotyped worker groups do not affect the EBV of the other individuals, the accuracies of the EBV of queens from year 9 represented those of unfertilized queens. Older BQ were included in the reference population to increase the accuracy of genomic EBV.

Proportions (p_{ref}) of 5, 10, 20, 30, 50 and 100% of BQ from years 4 to 7 were randomly chosen for genotyping. Separately, p_{ref} of the queens from years 8 and 9 were randomly sampled for genotyping. Consequently, the data set contained 300, 600, 1200, 1800, 3000, or 6000, respectively, genotyped BQ. The reference population included 250, 500, 1000, 1500, 2500, or 5000, respectively, phenotyped BQ from years 4 to 8. For $\mathbf{G}_{DPQ+BQ}$, all 2400 DPQ from years 4 to 9 were included.

3.2.4 Genetic parameters

Two quantitative traits that were affected by direct (worker group) and maternal (queen) genetic effects were simulated, with parameters as specified in Table 3.2. For the first trait (MOD), a moderate negative correlation between direct and maternal effects was assumed. The second trait had a higher negative genetic correlation (HGC). The chosen genetic parameters roughly represented the estimates for economically important traits such as honey yield or gentleness (Bienefeld and Pirchner 1991; Brascamp et al. 2016; Hoppe et al. 2020). Table 3.2 summarises all genetic parameters.

The genetic variances and heritabilities of the direct and maternal effects take the specificities of the honey bee into account. The phenotypic variance was calculated (formula (2) in Brascamp and Bijma 2019) as:

$$\sigma_{ph}^2 = A_{ii}\sigma_a^2 + \sigma_m^2 + \sigma_{am} + \sigma_e^2, \tag{3.2}$$

where σ_a^2 and σ_m^2 are the additive genetic variances of the direct and maternal effects, σ_{am}^2 is the covariance between direct and maternal effects, σ_e^2 is the residual variance, and A_{ii} is the average relationship between two workers of the same colony. We used $A_{ii} = 0.32$, which was calculated under the assumption that the queen is not inbred and is mated to unrelated DPQ. The heritabilities of the direct and maternal effects, h_a^2 and h_m^2 were calculated according to formulas (6b) and (6c) in (Brascamp and Bijma 2019), respectively. The genetic variance for queens, σ_Q^2, is given by:

$$\sigma_Q^2 = \text{Var}(\, a_Q + m_Q \,) = \sigma_a^2 + \sigma_m^2 + 2\sigma_{am}, \tag{3.3}$$

where a_Q and m_Q are the direct and maternal effects of queens, respectively.

The heritability of the sum of the queens' maternal and direct effects, which was the selection objective for queens in GPS, h_{GPS}^2, was calculated according to formula (7a) in (Brascamp and Bijma 2019). The genetic variance of the worker groups' direct and maternal effects, σ_W^2, is given by:

$$\sigma_W^2 = \text{Var}(\overline{a_W + m_W}), \tag{3.4}$$

where a_W and m_W are the direct and maternal effects of single workers from the same worker group, respectively. The genetic effects a_W and m_W have variances σ_a^2 and σ_m^2, respectively, and covariance σ_{am}. Because the number workers within a worker group, n_W, is very large, σ_W^2 is given by:

$$\sigma_W^2 = \frac{1 + n_W A_{ii}}{n_W} \sigma_Q^2 = A_{ii} \sigma_Q^2. \tag{3.5}$$

The genetic variance in worker groups equals the variance of the selection criterion in CBS. Therefore, the heritability of the selection criterion in CBS (called accessible heritability in Hoppe et al. 2020) is equal to:

$$h_{CBS}^2 = \frac{\sigma_W^2}{\sigma_{ph}^2}. \tag{3.6}$$

Table 3.2 Simulated variance and covariance components and genetic parameters derived from these (co)variances. We simulated settings with a moderate negative genetic correlation (MOD), or a high negative genetic correlation (HGC). The last 11 columns show the additive genetic variances of the direct (σ_a^2) and maternal effects (σ_m^2), their covariance (σ_{am}), the residual variance (σ_e^2), the heritabilities of the direct effects (h_a^2), maternal effects (h_m^2), the genetic correlation (r_G), the genetic variance of the worker groups (σ_W^2), the genetic variance of the queens (σ_Q^2), and the heritabilities of the sum of the maternal and direct effects of the worker groups (h_{CBS}^2) and queens (h_{GPS}^2), respectively.

Trait	σ_a^2	σ_m^2	σ_{am}	σ_e^2	h_a^2	h_m^2	r_G	σ_W^2	σ_Q^2	h_{CBS}^2	h_{GPS}^2
MOD	2	1	-0.5	1	0.299	0.467	-0.354	0.64	2	0.299	0.935
HGC	2	1	-1	1	0.390	0.610	-0.707	0.32	1	0.195	0.610

3.2.5 Simulation of the breeding population

We simulated a genome of 16 chromosomes with a recombination rate of 19 cM/Mb (Beye et al. 2006), with lengths based on the reference genome Amel_4.5 (INSDC assembly GCA_000002195.1) used by Jones et al. (2020). The level of linkage disequilibrium (LD) aimed for in the simulated genome was based on the genotypes for Supplementary table 2 of (Jones et al. 2020), with 44,113 SNPs remaining after quality control, for which the average LD between neighbouring loci was $r^2 = 0.215$.

To achieve this, a historical population of 50 queens per year was simulated, spanning 20,000 years, with a mutation rate of 0.0005 per locus (see (Gupta et al. 2012) for more details on the impact of the parameters). All loci were bi-allelic. There were no mating stations and queens mated with 12 drones. The dam of each drone was randomly sampled from all queens in the population. The allele frequencies in the final generation followed a U-shaped distribution. Loci with an allele frequency lower than 0.05 were discarded, which decreased the number of SNPs from an initial 100,000 to 48,419, with an average LD of $r^2 = 0.217$ between neighbouring loci. After creation of the LD, the population size was increased to 2400 queens per year and random mating was continued for six years.

A breeding population was simulated from years 0 to 9 after the historical generations, using the BeeSim program (Plate et al. 2019a), as shown in Figure 3.2 (a). The mutation rate was set to 0. Among the remaining 48,419 loci, 1000 were randomly chosen as QTL. Each QTL was assigned direct and maternal additive allele effects. Preliminary allele effects were drawn from the following distribution, as used by (Esfandyari et al. 2017; Plate et al. 2019a):

$$0.95 \cdot L(\mathbf{0}, \mathbf{V}_a) + 0.05 \cdot N(\mathbf{0}, \mathbf{V}_a) \text{ with } \mathbf{V}_a = \begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix}, \qquad (3.7)$$

where L and N denote 2-dimensional Laplace and normal distributions, respectively, and $\mathbf{V}_a$ is the additive genetic covariance matrix of the direct and maternal effects, as specified in Table 3.2. The preliminary allele effects were adjusted as described by (Plate et al. 2019a) to ensure that the average true breeding values (TBV) and the additive genetic variance in the base population were equal to $\mathbf{0}$ and $\mathbf{V}_a$, respectively. See (Plate et al. 2019a) for a detailed description of the modelling of the worker groups, phenotypes, and TBV.

After the breeding population was simulated, breeding values were estimated using PBLUP, ssGBLUP$_{BQ}$, and ssGBLUP$_{DPQ+BQ}$. One hundred replicates were simulated, starting from the same historical base population. For each replicate, new QTL were randomly chosen from the available SNPs.

3.2.6 Estimation of breeding values

For PBLUP, the following mixed linear model was used:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_a\mathbf{a} + \mathbf{Z}_m\mathbf{m} + \mathbf{e}, \tag{3.8}$$

where $\mathbf{y}$ is a vector of observations; $\mathbf{b}$ is a vector of the fixed effects (year); $\mathbf{a}$ is a vector of the direct effects of animals or groups; $\mathbf{m}$ is a vector of the maternal effects of animals or groups; $\mathbf{e}$ is a vector of residuals; and $\mathbf{X}$, $\mathbf{Z}_a$, and $\mathbf{Z}_m$ are known incidence matrices for $\mathbf{b}$, $\mathbf{a}$, and $\mathbf{m}$, respectively. The expected values of $\mathbf{a}$, $\mathbf{m}$, and $\mathbf{e}$ were assumed to be equal to $\mathbf{0}$, with the following variances:

$$\text{Var}\begin{pmatrix} \mathbf{a} \\ \mathbf{m} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \sigma_a^2\mathbf{A} & \sigma_{am}\mathbf{A} & \mathbf{0} \\ \sigma_{am}\mathbf{A} & \sigma_m^2\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2\mathbf{I} \end{pmatrix}, \tag{3.9}$$

where $\mathbf{A}$ is the honey bee specific numerator relationship matrix derived from pedigree (Bernstein et al. 2018), $\mathbf{I}$ is an identity matrix, and σ_a^2, σ_m^2, σ_{am} and σ_e^2 are the additive genetic variances of direct and maternal effects, their covariance and the residual variance, respectively.

In the ssGBLUP$_{BQ}$ analysis, pedigree information and genomic information were combined. The number of bi-allelic loci and the number of genotyped BQ are denoted m_g and n_g, respectively. The model and the variances were the same as for PBLUP, except that the numerator relationship matrix $\mathbf{A}$ was replaced by $\mathbf{H}$, which was constructed from $\mathbf{A}$ and the genomic relationship matrix, $\mathbf{G}_{BQ}$, following (Aguilar et al. 2010; Christensen and Lund 2010).

The genomic relationship matrix, (VanRaden 2008, method 1) was constructed as:

$$\mathbf{G}_{BQ} = \frac{\mathbf{ZZ}^T}{2\sum_i p_i(1-p_i)}, \tag{3.10}$$

where the $n_g \times m_g$ matrix $\mathbf{Z}$ is given by $\mathbf{Z} = \mathbf{M} - \mathbf{P}$, where matrix $\mathbf{M}$ contains the marker information of all genotyped BQ given as 0, 1, 2, and column j of matrix $\mathbf{P}$ is defined by $P_{ij} = 2p_i$, where p_i is the allele frequency at locus i. Matrix $\mathbf{G}_{BQ}$ was adjusted to $\mathbf{A}$ by adjusting the means of diagonal and off-diagonal entries, as described by (Christensen et al. 2012). To obtain an invertible genomic relationship matrix, a weighted genomic relationship matrix, $\mathbf{G}_{BQ,w}$, was constructed as follows:

$$\mathbf{G}_{BQ,w} = 0.95\mathbf{G}_{BQ} + 0.05\mathbf{A}_{BQ,g}, \tag{3.11}$$

where $\mathbf{A}_{BQ,g}$ is the submatrix of $\mathbf{A}$ relating to the genotyped animals. Finally, the inverse of $\mathbf{H}_{BQ}$ was computed following (Aguilar et al. 2010; Christensen and Lund 2010) as:

$$\mathbf{H}_{BQ}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{BQ,w}^{-1} - \mathbf{A}_{BQ,g}^{-1} \end{pmatrix}. \tag{3.12}$$

For ssGBLUP$_{\text{DPQ+BQ}}$, the ssGBLUP$_{\text{BQ}}$ analysis was supplied with genotypes of DPQ. The honey bee specific pedigree relationship matrix (Bernstein et al. 2018; Brascamp and Bijma 2014) does not consider individual DPQ. Instead, they are merged into pseudo-fathers. To have an equal structure in $\mathbf{A}$ and $\mathbf{G}_{BQ+DPQ}$, we merged entries as described below, such that every pseudo-father is represented by a single element that combines the genotypes of the DPQ which comprise it. In the pedigree-based relationship matrix (formula (17) in Brascamp and Bijma 2014 and Equation (2.5) here), the diagonal entry of a pseudo-father is calculated as:

$$A_{pp} = \frac{1}{n_S}(1 + F_d) + \frac{n_S - 1}{n_S}\tilde{A}_{db}, \tag{3.13}$$

where A_{pp} represents the diagonal entry of pseudo-father p, comprising n_S DPQ, F_d represents the coefficient of inbreeding of p, and $\tilde{A}_{db}$ represents the additive genetic relationship of two DPQ contained in p. In the pedigree-based relationship matrix, all DPQ of p have the same coefficients of relationship. Let $\tilde{\mathbf{A}}$ be a honey-bee specific relationship matrix, where all pseudo-fathers are replaced by the groups of DPQ they represent. Then, we have:

$$A_{pp} = \frac{1}{n_S^2}\sum_d \tilde{A}_{dd} + \frac{1}{n_S^2}\sum_{d,b}\tilde{A}_{db}, \tag{3.14}$$

where $\tilde{A}_{dd}$ represents the diagonal entry of a DPQ of p, and $\tilde{A}_{db}$ represents the off-diagonal entry of two DPQ, d and b, of p.

The genomic relationship matrix with entries for individual DPQ, $\tilde{\mathbf{G}}_{DPQ+BQ}$, was calculated according to Equation (3.10) by including the genotypes of individual DPQ. The following conversion of Equation (3.14) to the genomic relationship matrix was used:

$$G_{DPQ+BQ,pp} = \frac{1}{n_S{}^2}\sum_d \tilde{G}_{DPQ+BQ,dd} + \frac{1}{n_S{}^2} \sum_{d,b} \tilde{G}_{DPQ+BQ,db}, \qquad (3.15)$$

where $G_{DPQ+BQ,pp}$ represents the diagonal entry of p, $\tilde{G}_{DPQ+BQ,dd}$ represents the diagonal entry of a DPQ of p, and $\tilde{G}_{DPQ+BQ,db}$ represents the off-diagonal entry of two DPQ of p. The genomic relationships of p to all other animals were calculated as the mean of the relationships of the DPQ of p to these animals. Exchanging $\tilde{\mathbf{G}}_{DPQ+BQ}$ for $\mathbf{G}_{DPQ+BQ}$ in ssGBLUP$_{DPQ+BQ}$ does not change the EBV of BQ or worker groups [see Appendix E]. Matrix $\mathbf{G}_{DPQ+BQ}$ was then adjusted to the submatrix of $\mathbf{A}$ that contains the same animals, analogous to Equation (3.11). Finally, $\mathbf{H}_{DPQ+BQ}^{-1}$ was obtained analogous to Equation (3.12).

Programs from the BLUPf90 family (Aguilar et al. 2014; Misztal et al. 2002) were used to calculate the genomic relationship matrix and to perform estimation of breeding values, using the variance components in Table 3.2. Equations (3.11), (3.12), and (3.15), and adjustment of the genomic relationship matrix to the pedigree relationship matrix were implemented in R (R Development Core Team 2020). The pedigree relationship matrix and its submatrix relating to genotyped animals were calculated in a C-program according to (Bernstein et al. 2018).

3.2.7 Evaluation of the breeding value estimates

EBV were evaluated for prediction accuracy and bias. The accuracy was calculated as the correlation coefficient between (simulated) TBV and EBV. Bias was evaluated based on deviations of the regression coefficient of TBV on EBV, b_1, from 1.

The accuracies for phenotyped colonies, ρ_{pW}, and unfertilized queens, ρ_{uQ}, were represented by the accuracy for worker groups from year 8 and queens from year 9, respectively. The accuracies for ssGBLUP$_{BQ}$ and ssGBLUP$_{DPQ+BQ}$ were calculated for the genotyped BQ. The

variance of TBV for phenotyped colonies, σ_{pW}^2, was calculated as the variance of the TBV of worker groups in year 8. The variance of the TBV of the BQ that head phenotyped colonies, σ_{pQ}, was calculated as the variance of the TBV of the BQ in year 8. The variance of the TBV of the unfertilized queens or DPQ, σ_{uQ}^2, was calculated as the variance of the TBV of the BQ in year 9.

Accuracies for queens and worker groups cannot be directly compared, since different genetic variances must be used to estimate genetic gain from them. Therefore, we rescaled ρ_{pW} to an accuracy of queens, ρ_{pR} (R for replacement queen) as:

$$\rho_{pR} = \frac{\sigma_{pW}}{\sigma_{pQ}} \rho_{pW} \tag{3.16}$$

This can interpreted as an accuracy of fictional queens. If a daughter was reared from each of the colonies in year 8, then the correlation between the daughters' EBV and the daughters' TBV would be ρ_{pR}. Equation (3.16) was derived from formulas of Brascamp and Bijma (2019) and [see Appendix F].

3.2.8 Genetic gain in different breeding schemes

Table 3.3 shows the notation used in the equations to estimate genetic gain in the sum of direct and maternal effects (SDME), based on the following basic formula for expected genetic gain (Falconer and Mackay 1996):

$$R = i\rho\sigma, \tag{3.17}$$

where R is response to selection in SDME, i is intensity of selection, ρ is the accuracy of selection, i.e. the correlation between the true and estimated value of the SDME for selection candidates, and σ is the standard deviation of the TBV for SDME among selection candidates. The generation interval was not considered in the calculations since it was two years for BQ and three years for DPQ for all breeding schemes.

The average TBV of a generation of colonies is given by the average TBV of the selected BQ and DPQ in the parental generation. The average TBV of a queen reared from a colony is equal

to the breeding value of the colony's worker group (Brascamp and Bijma 2014). Consequently, the response to CBS, R_{CBS}, is given by:

$$R_{CBS} = \frac{i_{DPQ}^{CBS}}{2} \rho_{pW} \sigma_{pW} + \frac{i_{BQ}^{CBS}}{2} \rho_{pW} \sigma_{pW}.$$

(3.18)

Alternatively, R_{CBS} can be calculated from Equation (3.18) by replacing ρ_{pW} and σ_{pW} by ρ_{pR} and σ_{pQ}, respectively. Accuracies for ssGBLUP are chosen according to the size of the reference population given by 5000 p_{ref}. Table 3.3 shows the notation used in the formulas for to calculate genetic gain.

Based on the average TBV of colonies when BQ and DPQ were preselected and the dams of BQ and dams of DPQ have a TBV of 0, response to GPS, R_{GPS}, was predicted as:

$$R_{GPS} = \frac{i_{DPQ}^{GPS} p_{DPQ}}{2} \rho_{uQ} \sigma_{uQ} + \frac{i_{BQ}^{GPS} p_{BQ}}{2} \rho_{uQ} \sigma_{uQ},$$

(3.19)

Table 3.3 *Notation key for symbols in the estimation of genetic gain.*

n_{gpy}	Number of genotyped queens per year
$i_{DPQ}^{CBS}, i_{BQ}^{CBS}$	Selection intensity for dams of drone producing queens (DPQ), and dams of breeding queens (BQ), respectively
$i_{DPQ}^{GPS}, i_{BQ}^{GPS}$	Selection intensity for DPQ, and BQ, respectively
p_{DPQ}	Proportion of preselected DPQ compared to all DPQ deployed on mating stations
p_{BQ}	Proportion of preselected BQ compared to all phenotyped BQ
p_{ref}	Proportion of BQ in the reference population
$\sigma_{uQ}, \sigma_{pQ}, \sigma_{pW}$	Standard deviation of the true breeding values for the sum of maternal and direct effects of unfertilized queens (BQ from year 9), queens heading phenotyped colonies (BQ from year 8), and phenotyped colonies (worker groups from year 8), respectively
$\rho_{uQ}, \rho_{pW}, \rho_{pR}$	Prediction accuracy for the sum of maternal and direct effects of unfertilized queens (queens from year 9), phenotyped colonies (worker groups from year 8), and replacement queens from phenotyped colonies, respectively
R_{PB}	Response to selection in a single generation in classical (pedigree-based) selection program
R_{GS}	Response to selection in the initial selection cycle of a genomic selection program

Response to selection in a genomic breeding program with CBS and GPS is given by $R_{GS} = R_{CBS} + R_{GPS}$, and was predicted as:

$$R_{CBS+GPS} = \frac{i_{DPQ}^{CBS}}{2}\rho_{pW}\sigma_{pW} + \frac{i_{BQ}^{CBS}}{2}\rho_{pW}\sigma_{pW} + \frac{i_{DPQ}^{GPS}p_{DPQ}}{2}\rho_{uQ}\sigma_{uQ} + \frac{i_{BQ}^{GPS}p_{BQ}}{2}\rho_{uQ}\sigma_{uQ}. \qquad (3.20)$$

Response to pedigree-based selection, R_{PB} (PB for pedigree-based) was predicted based on Equation (3.18), which correctly predicted the average genetic gain in the stochastically simulated breeding population [see Appendix G]. Equations (3.18), (3.19), and (3.20) were used to predict R_{GS} in the scenarios described in Table 3.1. For ρ_{uQ} and σ_{uQ}, we used values obtained for the BQ from year 9 in the simulated breeding population. For ρ_{pW} and σ_{pW}, we used values obtained for the worker groups from year 8 in the simulated breeding population. However, for this analysis, ρ_{pW} was calculated by including worker groups of non-genotyped queens. The values of the accuracies ρ_{uQ} and ρ_{pW} were chosen for each scenario according to p_{ref}. For values of p_{ref} that were not explicitly simulated, the accuracy was obtained by linear interpolation. For $p_{ref} = 0$ and $p_{ref} > 0$, the accuracies of PBLUP and ssGBLUP were used, respectively. When all DPQ were preselected, the formulas were also evaluated using the accuracy of ssGBLUP$_{DPQ+BQ}$.

To evaluate the impact of increasing the annual budget for genotyping queens, n_{gpy}, we calculated the increase in genetic gain (IGG) by adding genotypes on 500 queens. For a budget $n_{gpy} = x$, the increase in genetic gain by adding 500 genotypes, IGG(x), was defined as:

$$\text{IGG}(x) = R_{GS}(x + 500) - R_{GS}(x), \qquad (3.21)$$

where the values of $R_{GS}(x)$ and $R_{GS}(x + 500)$ were taken from the scenarios with the highest genetic gain for budgets $n_{gpy} = x$ and $n_{gpy} = x + 500$, respectively. We will focus on budgets $n_{gpy} = x$ which satisfy IGG$(x) <$ IGG$(x - 500)$. Such points x provide potential values of the minimal budget required to initiate a breeding program, because using $n_{gpy} = x + 500$ genotypes instead of $n_{gpy} = x$ might add too little genetic gain to justify the investment. Note that these measures do not the monetary value of genetic gain into account.

3.3 RESULTS

3.3.1 Prediction accuracy for the sum of direct and maternal effects

Accuracy was measured as the correlation between EBV and TBV. The results obtained with ssGBLUP and PBLUP are shown in Figure 3.3, and those obtained with 5000 BQ in the reference population are in Appendix H. The largest improvements over PBLUP were observed for unfertilized queens, which is desirable for GPS.

We focused on ssGBLUP$_{DPQ+BQ}$, since the accuracy with ssGBLUP$_{BQ}$ was just slighlty lower with 5000 BQ in the reference population. The accuracies for unfertilized queens with ssGBLUP$_{DPQ+BQ}$ were higher than those with PBLUP by 160.1 (MOD) and 126.2% (HGC). The accuracies for replacement queens from phenotyped colonies with ssGBLUP$_{DPQ+BQ}$ were higher than those with PBLUP by 9.9 (MOD) and 12.5% (HGC).
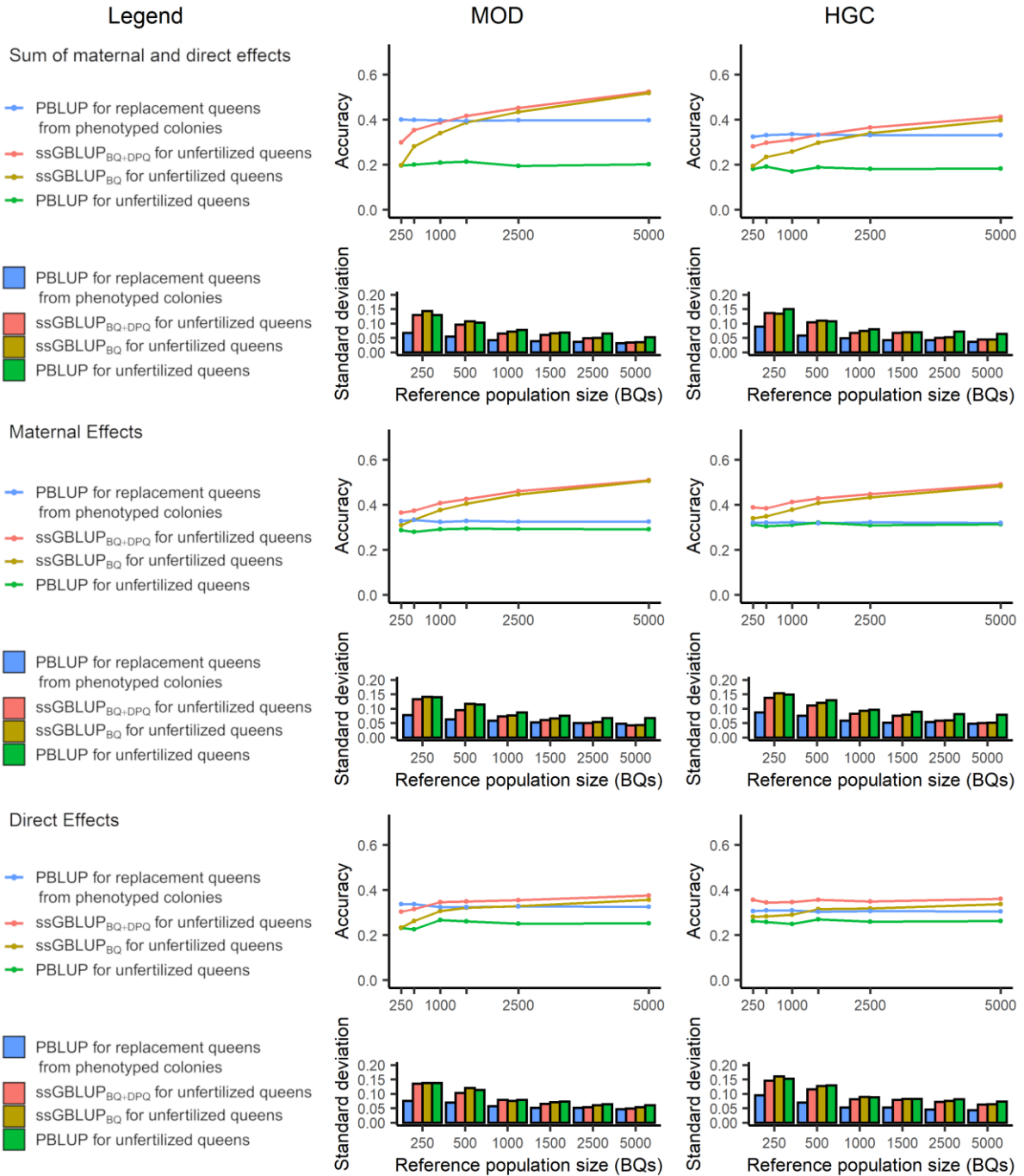
Figure 3.3 Accuracies of genomic estimated breeding values. Accuracies of estimated breeding values of unfertilized queens (from year 9) and of replacement queens from phenotyped colonies (from year 8), and the standard deviations of these accuracies in 100 replicates. The standard deviations of the true breeding values used to calculate the accuracy for replacement queens σ_{pW} (worker groups from year 8) and σ_{pQ} (queens from year 8) are shown in Appendix J.

The differences in the accuracies of ssGBLUP$_{DPQ+BQ}$ and ssGBLUP$_{BQ}$ were larger when fewer BQ were included in the reference population. We focused on the accuracies for unfertilized queens. For a reference population size of 250 BQ, accuracies obtained with ssGBLUP$_{BQ}$ differed from those obtained with PBLUP by -2.1 (MOD) and 6.9% (HGC). Addition of the genotypes of all 2800 DPQ increased the accuracy considerably. With 250 BQ in the reference population, accuracies obtained with ssGBLUP$_{DPQ+BQ}$ were higher than those obtained with PBLUP by 47.9 (MOD) and 54.4% (HGC). However, genotyping phenotyped BQ instead of DPQ yields more accurate genomic EBV for a smaller number of genotypes. E.g., with a reference population of 1500 BQ, accuracies obtained with ssGBLUP$_{BQ}$ were higher cthan those with PBLUP by 91.7 (MOD) and 63.2% (HGC).

3.3.2 Prediction accuracy of maternal and direct effects

The results for the accuracies for estimates of direct and maternal effects with ssGBLUP and PBLUP are shown in Figure 3.3. We focused on unfertilized queens and on a reference population size of 5000 BQ. Increases in the accuracies of estimates of maternal effects were larger than those of direct effects. The accuracies for maternal effects with ssGBLUP$_{BQ}$ were 73.6 (MOD) and 53.8% (HGC) higher than those with PBLUP, while the accuracies for direct effects with ssGBLUP$_{BQ}$ were 41.4 (MOD) and 28.6% (HGC) higher than with PBLUP. With ssGBLUP$_{DPQ+BQ}$, the accuracies for direct effects were higher than those with ssGBLUP$_{BQ}$. For maternal effects, the accuracy with ssGBLUP$_{DPQ+BQ}$ was almost equal to that with ssGBLUP$_{BQ.}$

3.3.3 Genetic gain

Figure 3.4 shows the genetic gain, R_{GS}, for different breeding schemes and different numbers of genotyped queens per year, n_{gpy}. The configurations of these scenarios are shown in Appendix I. The standard deviations of the TBV used to calculate R_{GS} are shown in Appendix J. We focused on the optimal breeding schemes with the configurations shown in Table 3.4.

Up to $n_{gpy} = 500$, increases in R_{GS} were small since the accuracy of EBV based on ssGBLUP was still low (Table 3.4). Genetic gain, R_{GS}, increased strongly as n_{gpy} increased from 500 to 1000. The increase in genetic gain (IGG) diminished slightly from $n_{gpy} = 1000$, onwards. Therefore, we suggest that $n_{gpy} = 1000$ is the minimal budget required to initiate a breeding program.

55
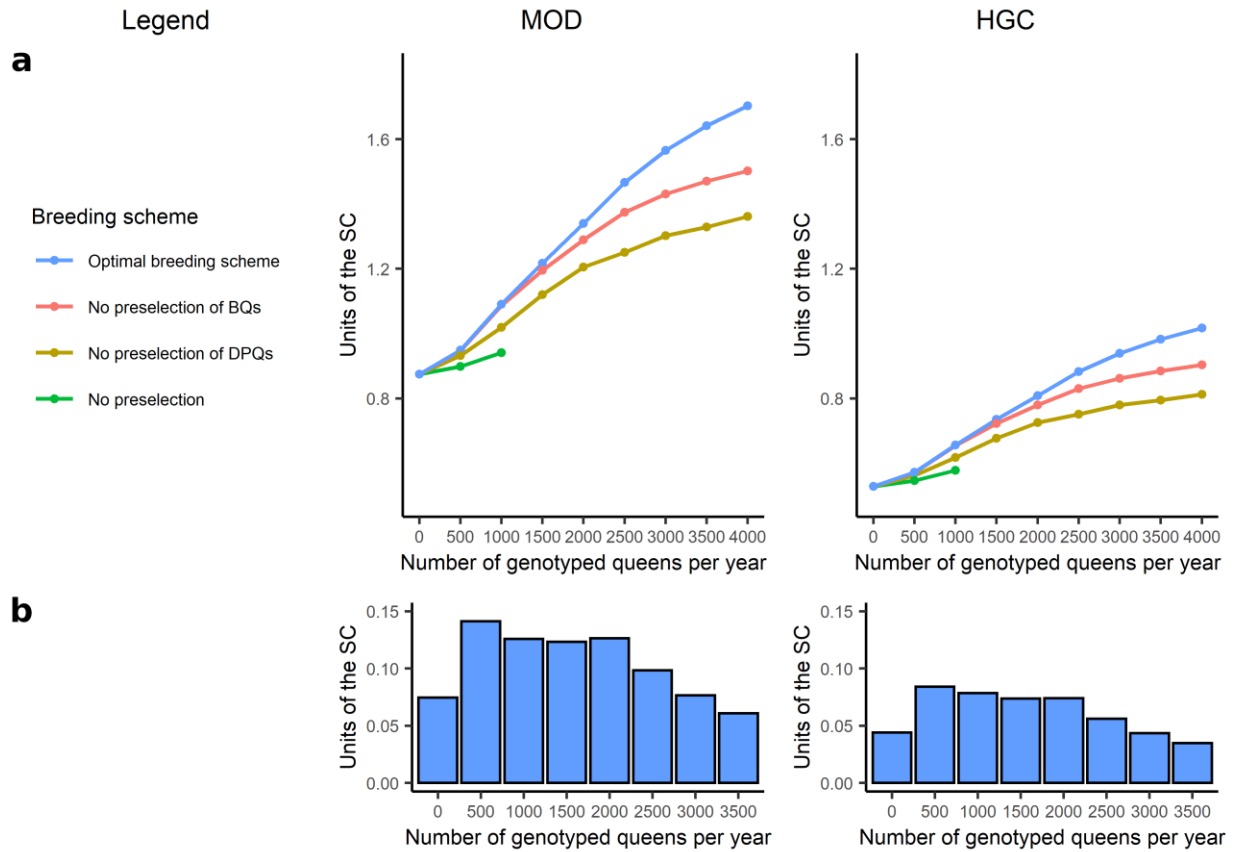
Figure 3.4 Predicted genetic gain for different breeding schemes (a) and increase in genetic gain by adding genotypes of 500 queens (b).

(a) Genetic gain, R_{GS}, was calculated according to Equation (3.20). For strategies without preselection of BQ, or without preselection of DPQ, the scheme with the highest genetic gain is shown. The value for zero genotyped queens represents the gain with pedigree-based selection, R_{PB}, predicted according to Equation (3.18).

(b) The increase in genetic gain by adding 500 genotypes of queens (IGG) is shown for the optimal breeding scheme in (a). Genetic gain and IGG are given in the units of the selection criterion.

Table 3.4 *Genetic gain, R_{GS}, in the initial selection cycle of different breeding schemes with colony-based (CBS) and genomic preselection (GPS). Breeding queens and drone producing queens are referred to as BQ and DPQ, respectively; Settings with a moderate negative genetic correlation (MOD), or a high negative genetic correlation (HGC) were simulated; Pedigree-based BLUP and Single-Step-genomic-BLUP are referred to as PBLUP and ssGBLUP, respectively. The relationship matrix for ssGBLUP contained either exclusively BQ (ssGBLUP$_{BQ}$) or BQ and DPQ (ssGBLUP$_{DPQ+BQ}$); Genetic gain is given in the units of the selection criterion; The standard deviations of the true breeding values σ_{pW} (worker groups from year 8) and σ_{uQ} (queens from year 9) are in Appendix J. Genetic gain, R_{GS}, was calculated according to Equations (3.18), (3.19) , and (3.20).*

Number of genotyped queens per year (n_{gpy})	Proportion of DPQ selected by GPS compared to all DPQ deployed on mating stations (p_{DPQ})	Selection intensity of GPS on DPQ (i^{GPS}_{DPQ})	Proportion of BQ selected by GPS compared to all phenotyped BQ (p_{BQ})	Selection intensity of GPS on BQ (i^{GPS}_{BQ})	Number of BQ in the reference population (5 years in total)	Proportion of BQ in the reference population (p_{ref})	Breeding value estimation method	Genetic gain R_{GS}	
								MOD	HGC
0	0	0	0	0	0	0	PBLUP	0.8751	0.5286
500	0.4	0.7979	0.15	0.2998	750	0.15	ssGBLUP$_{BQ}$	0.9497	0.5715
500	0.46	0.7979	0.11	0.2998	550	0.11	ssGBLUP$_{BQ}$	0.9492	0.5725
1000	1	0.7454	0.205	0.2998	1045	0.209	ssGBLUP$_{DPQ+BQ}$	1.0909	0.6559
1000	1	0.8454	0.125	0.2998	625	0.125	ssGBLUP$_{DPQ+BQ}$	1.0878	0.6567
1500	1	0.8454	0.46	0.4759	2330	0.466	ssGBLUP$_{DPQ+BQ}$	1.2169	0.7344
1500	1	0.8889	0.5	0.2998	2500	0.5	ssGBLUP$_{DPQ+BQ}$	1.2167	0.7353
2000	1	0.8889	0.785	0.4759	3930	0.786	ssGBLUP$_{DPQ+BQ}$	1.3403	0.8091
2500	1	1.0324	1	0.4759	5000	1	ssGBLUP$_{DPQ+BQ}$	1.4668	0.8831
3000	1	1.0908	1	0.7111	5000	1	ssGBLUP$_{DPQ+BQ}$	1.5652	0.9391
3500	1	1.2322	1	0.7979	5000	1	ssGBLUP$_{DPQ+BQ}$	1.6416	0.9828
4000	1	1.3401	1	0.872	5000	1	ssGBLUP$_{DPQ+BQ}$	1.7026	1.0175

At $n_{gpy} = 1000$, the optimum scenarios for MOD and HGC increased genetic gain by around 24% compared to pedigree-based selection (Table 3.4). For both these scenarios, BQ were preselected at the lowest possible non-zero selection intensity, and scenarios without preselection of BQ achieved very similar results [see Appendix I]. These increases were only possible with ssGBLUP$_{\text{DPQ+BQ}}$, as the optimal scenarios with ssGBLUP$_{\text{BQ}}$ increased genetic gain by 18.7 to 20.8% compared to pedigree-based selection [see Appendix I]. The optimum scenarios and the optimal scenarios without preselection of BQ were those for which 10 to 20% of the phenotyped BQ were genotyped per year and around 50% of DPQ were preselected.

The IGG remained high up to $n_{gpy} = 2500$ (Figure 3.4). Between $n_{gpy} = 1000$ and $n_{gpy} = 2500$, BQ were increasingly preselected (Table 3.4). At $n_{gpy} = 2500$, all BQ and DPQ of a given year were selected by GPS, with a higher selection intensity for DPQ. This increased genetic gain by around 67.5% compared to pedigree-based selection for both MOD and HGC. For n_{gpy} larger than 2500, p_{DPQ} and p_{BQ} were at their maximum value in the optimal scenarios, and only the selection intensities i_{DPQ}^{GPS} and i_{BQ}^{GPS} could increase. Consequently, IGG diminished strongly, which implies that n_{gpy} much larger than 2500 yield little additional genetic gain per genotype. Since IGG were constantly high between $n_{gpy} = 1000$ and $n_{gpy} = 2500$, an n_{gpy} of 2500 or larger is required to maximize the total amount of genetic gain with genomic over pedigree-based selection divided by the total number of genotypes used.

3.3.4 Bias of the estimated breeding values

Results for the coefficient of regression of TBV on EBV for PBLUP and ssGBLUP are shown in Figure 3.5 and Appendix K. Pedigree-based methods showed very little bias ($b_1 = 0.97$ to 1.02) for both maternal and direct effects, and the SDME and across all groups of animals considered. Worker groups from year 8 showed nearly no bias. Non-phenotyped queens showed a maximum bias of $b_1 = 1.11$ for ssGBLUP$_{\text{BQ}}$ and ssGBLUP$_{\text{DPQ+BQ}}$ (HGC). Worker groups in years 4 to 7 showed a greater bias than $b_1 = 1.15$ with ssGBLUP for HGC. This might be due to the mating stations using unrelated DPQ in years 0 to 2. We did not investigate the cause of this bias in more detail since the bias of GEBV of the candidates in years 8 and 9 was within the acceptable range.
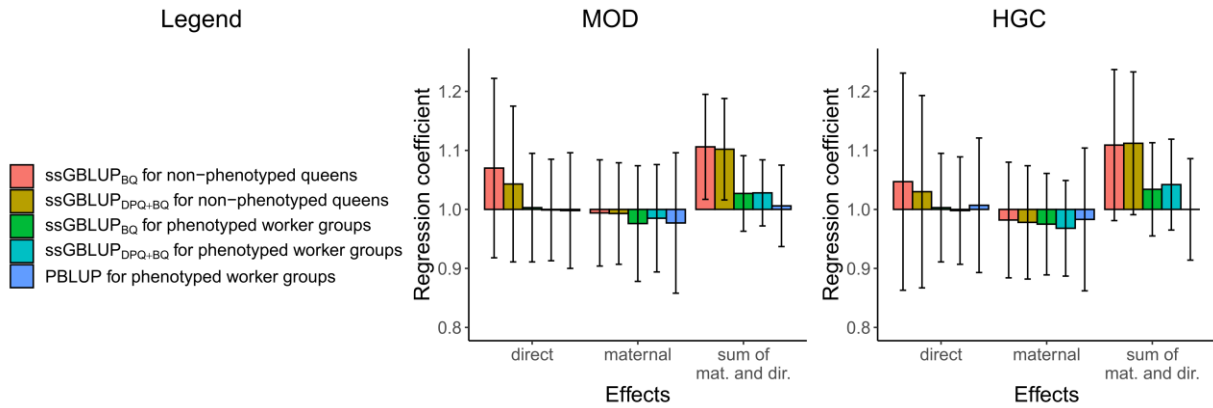
Figure 3.5 Regression coefficients of true on estimated breeding values. Regression coefficients of true on estimated breeding values for unphenotyped queens (from year 9) and for phenotyped worker groups (from year 8) with a reference population of 5000 BQ. The sum of maternal and direct effects was considered as the selection criterion.

The bias for estimates was often smaller for maternal effects ($b_1 = 0.98$ to 1.00 for non-phenotyped queens) than for direct effects ($b_1 = 1.01$ to 1.07 for non-phenotyped queens), while the largest bias was found for the SDME ($b_1 = 1.02$ to 1.11 for non-phenotyped queens).

3.4   DISCUSSION

We used a honey bee specific relationship matrix to perform genetic evaluations with PBLUP, ssGBLUP$_{BQ}$, and ssGBLUP$_{DPQ+BQ}$ and investigated the prediction accuracy and bias of overall, direct, and maternal EBV for queens and worker groups, as well as their change with the number of genotyped queens. We used these statistics to estimate genetic gain in different breeding programs.

3.4.1   Estimation of breeding values

We compared the quality of the EBV for different reference population sizes by genotyping different proportions of phenotyped BQ (see Figure 3.3). Rather high increases in prediction accuracy in honey bees could be expected, since controlled mating in honey bees leaves uncertain relationships, and simulations in other species (Tonussi et al. 2017) show that uncertain relationships reduce the accuracy of PBLUP EBV more than that of ssGBLUP EBV. Our accuracies for the SDME, i.e. 0.18 and 0.2 for PBLUP and 0.39 and 0.52 for ssGBLUP$_{BQ}$, are similar to results for young bulls with 25% unknown sires in beef cattle (Tonussi et al.

2017). For the two traits considered in that study, the accuracies of PBLUP were 0.14 and 0.16 and those of ssGBLUP were 0.41 and 0.65, respectively. In real datasets, pedigree errors can be corrected from genomic data, which may yield higher improvements.

We found a bias of up to $b_1 = 1.11$ for the EBV of non-phenotyped queens with ssGBLUP (Figure 3.5 and Appendix K, which is low, since values between 0.85 and 1.15 appear to be acceptable in practice (Tsuruta et al. 2011). Uncertain relationships can explain our deviations of b_1 from 1, as even higher deviations were reported by (Tonussi et al. 2017) for young males with 25% unknown sires. The controlled mating used in our study probably reduced the bias of EBV.

3.4.2   Accuracy with genotyped DPQ

We examined the effect of including the genotypes of DPQ into the genomic relationship matrix. This proved useful, when a small number of phenotyped BQ and a large number of DPQ were genotyped. This situation is likely to occur in practice when genomic preselection is applied to DPQ (see Table 3.4). However, a considerable increase in prediction accuracy was only possible when the number of genotyped DPQ was much larger than the number of genotyped BQ (see Figure 3.3).

Genotyping sires increases the prediction accuracies of their offspring, because a genotyped sire adjusts its offspring's relationships in ssGBLUP (Christensen and Lund 2010). However, genotyping all DPQ and using these data, as described here, proved not as effective as adding more BQ to the reference population. E.g., genotyping 2800 DPQ and 250 phenotyped BQ, yielded a lower prediction accuracy than genotyping 1500 phenotyped BQ.

Specificities of the honey bee can explain our results for prediction accuracy. When pseudo-fathers are modelled as groups of DPQ, the relationship of a daughter with its pseudo-father is lower (= 0.203 with the parameters in this study) than with its dam (= 0.5), regardless of whether it is a queen or worker group. The reason is that the daughter is related by 0.5 to exactly one of the DPQ which are daughters of the pseudo-father, but it is uncertain which DPQ from the mating station it is. The relationship of a daughter to its pseudo-father is its average relationship to all DPQ from the mating station.

For ssGBLUP$_{DPQ+BQ}$ with non-phenotyped DPQ, separating pseudo-fathers into individual DPQ in the relationship matrix and identifying sires from genomic data does not increase the accuracy of genomic EBV of BQ and their worker groups [see Appendix E]. This is corroborated by results from Maiorano et al. (2019), who simulated a pig population in which mixed semen was used and showed that using the genomic relationship matrix in single-step genomic BLUP obviated the need for identifying the sires from genomic relationships.

However, for phenotyped DPQ, splitting pseudo-fathers into individual DPQ in the relationship matrix looks more promising. The main reason why DPQ contributed less to prediction accuracy than BQ is probably due to the fact that DPQ were not phenotyped, although they sometimes are in practice. In this case, their genotypes are more important than those of the BQ, since they have a strong impact on genetic gain. Individual DPQ should then be incorporated into the genomic relationship matrix and the algorithm of Bernstein et al. (2018) to estimate their breeding values.

3.4.3   Optimal breeding scheme

We compared genetic gain from different breeding schemes that use GPS for several numbers of genotyped queens per year (Table 3.4). Our deterministic model correctly predicted genetic gain for an initial CBS-step followed by a GPS-step. Our model would, however, overestimate the genetic gain for a CBS-step following a GPS-step, since GPS reduces the genetic variance in ways that are rather specific to honey bees. Deterministic models of genomic preselection (Pryce et al. 2010; Shumbusho et al. 2013) are usually models of 2-stage selection of the same animals (Saxton 1983). However, in honey bees, the candidates of GPS are DPQ and/or unfertilized BQ. Subsequently, the drone offspring of the DPQ mate with the BQ, which results in fertilized BQ that are the candidates for CBS. Separating the genetic variance contributed by the new generation of drones from the genetic variance of the BQ was beyond the scope of our study (see Du et al. 2021a for the situation in CBS). However, we expect that would not change the main conclusions.

We suggest that a budget to genotype 1000 queens per year should be the minimal target to initiate a genomic selection breeding program, since the IGG decreased slightly from there on. The optimal configuration for the use of this budget involves GPS of all DPQ at a proportion of 1:2 and genotyping between 10 and 20% of the phenotyped BQ. Smaller reference

populations should be avoided, since the optimal breeding schemes for $n_{gpy} = 500$ also relied on genotyping more than 10% of the phenotyped queens. However, pure GPS of DPQ can be suboptimal when the number of DPQ is greater and the number of BQ is smaller than we assumed.

The IGG decreased strongly for n_{gpy} larger than 2500, when all BQ and DPQ were preselected based on genomic EBV. This suggests that an n_{gpy} of 2500 or larger optimizes genetic gain per used genotype. The economic optimum depends on the costs of genotyping and the monetary benefit of increased genetic gain.

Extra genetic gain from genomic selection compared to pedigree-based selection resulted from an increase in prediction accuracy for non-phenotyped queens and a larger number of candidates. Similar results were found for maternal traits in sheep (Lillehammer et al. 2020), pigs (Lillehammer et al. 2011), and Atlantic salmon (Verbyla et al. 2018). However, genomic selection is often most efficient when young animals are selected based on genomic EBV and the generation interval is shortened (Goddard and Hayes 2007).

We considered a generation interval of two years for BQ and three years for DPQ. However, queens can be reared from very young queens shortly after mating and a colony headed by a one-year old queen produces a sufficient number of drones to fertilize hundreds of queens. Therefore, the generation interval could be shortened by at least one year. We consider such a scheme in Appendix L and showed that it can improve genetic gain considerably. Brascamp et al. (2018) considered an even more refined scheme in which several generations of queens are reared during a single summer. Besides possible issues with its practical implementation, further simulation studies are required before such a scheme can be recommended. Phenotyping would lag behind, as phenotyping is done in the second summer of a colony's life, which leads to a lower accuracy of GEBV (Habier et al. 2007). Furthermore, the intensity of selection should be carefully considered. Shortening the generation interval also requires changes to the structure of a breeding program, which can substantially increase the rate of inbreeding per generation (de Roos et al. 2011; Doublet et al. 2019; Lillehammer et al. 2020). We suggested GPS as an additional selection step, which is unlikely to alter the rate of inbreeding significantly.

In our simulations, the queens to be genotyped were randomly chosen. However, genotyping a small proportion of all selection candidates can yield most of the benefits from genomic selection, especially when animals to be genotyped are preselected (Granleese et al. 2019; Henryon et al. 2012; Howard et al. 2018). In honey bees, the dams of DPQ and dams of BQ that produce candidates for GPS could be selected based on pedigree-based EBV (Henryon et al. 2012). However, genotyping queens with contrasting phenotypes should be considered to maintain prediction accuracy (Chu et al. 2020; Gowane et al. 2019).

We considered only the beginning of genomic selection in the breeding schemes investigated. Genetic gain may increase in future generations because the reference population grows yearly. However, the early animals will become less useful over time, since their relationship to the selection candidates will decrease (Habier et al. 2007). In addition, the Bulmer effect will decrease response to selection (Van Grevenhof et al. 2012).

3.4.4 Maternal and direct effects

A previous study (Gupta et al. 2013) showed that prediction accuracy for non-phenotyped queens is considerably higher with genomic methods than with pedigree-based methods. We found considerably higher gains in accuracies for estimates of maternal effects than for estimates of direct effects with ssGBLUP compared to PBLUP (see Figure 3.3 and Appendix H). To the best of our knowledge, this result stands out among those of other simulation studies in the literature (Gupta et al. 2013; Lourenco et al. 2013; Maiorano et al. 2019).

For honey bees, Gupta et al. (2013) reported gains in accuracy of estimates of maternal effects with ssGBLUP over PBLUP that were similar to the gains in accuracy for estimates of direct effects, which is due to the fact that phenotypes were directly associated with the genetic effects of queens, as mentioned previously. Several other studies have reported simulated accuracies of estimates of maternal and direct effects in other species. For example, Lourenco et al. (2013) compared estimates of breeding values in beef cattle based on ssGBLUP and Bayesian methods. However, the authors used PBLUP as a reference and reported a higher increase in accuracy by switching from PBLUP to ssGBLUP for direct than for maternal effects. In another study, Maiorano et al. (2019) investigated how the use of pooled semen in pigs affected the accuracies of EBV from ssGBLUP and reported considerably higher gains in accuracies for estimates of direct effects than for maternal effects. This can be explained by major species-specific

differences. For almost all agricultural species, the maternal effect of a dam is expressed in multiple phenotypic records, one for each offspring. In contrast, the maternal effect of a honey bee queen is only expressed in the single phenotypic record of her worker group. Thus, the maternal effect of a dam in other agricultural species is expressed in more phenotypes than the maternal effect of a honey bee queen. Consequently, the impact of using genotyping data on maternal effects is greater for honey bee queens than it is for other agricultural species.

In our study, the increase in accuracy by switching from PBLUP to ssGBLUP was higher for MOD than for HGC in unphenotyped queens, which is explained by the high negative genetic correlation, r_G, for HGC, which is the only difference between MOD and HGC. However, the relative increase in accuracy of EBV from PBLUP to ssGBLUP was smaller for MOD than for HGC in phenotyped queens. A study in beef cattle found no significant difference in the increase in accuracy of EBV from PBLUP compared to ssGBLUP with $r_G = 0$ and $r_G = -.3$ (Lourenco et al. 2013). The higher negative values of r_G that we considered, probably amplified the differences in the increase in accuracy of EBV from PBLUP compared to ssGBLUP between the maternal and direct effects. However, parameter estimates in honey bees can yield even lower values for r_G than we assumed in our study (Brascamp et al. 2016). The high negative value for r_G in HGC reduced genetic gain compared to MOD, but the optimal breeding schemes for MOD and HGC were very similar, for each n_{gpy} (see Table 3.4).

3.5   CONCLUSIONS

We used a honey bee specific relationship matrix in simulation studies to evaluate methods of breeding value prediction and the design of genomic breeding programs for honey bees and we found that ssGBLUP outperformed PBLUP. Adding the genotypes of DPQ was found to increase the accuracy considerably if the reference population is small. Prediction accuracies of EBV and standard deviations of TBV were used in a deterministic model to predict genetic gain from one round of selection. The model correctly predicted genetic gain for an initial CBS-step followed by a GPS-step. To initiate a breeding program, genotyping a minimum number of 1000 queens per year is required. With 1000 genotypes, genotyping phenotyped BQ and preselection of DPQ based on GEBV achieved the highest genetic gain. Genotyping at least 2500 queens per year and applying GPS to all BQ and all DPQ is required to maximize genetic gain per used genotype. However, economic aspects, e.g., the costs of genotyping, and the

monetary benefits from increased genetic gain should be included in such considerations. We suggest that the methods ssGBLUP$_{BQ}$ and ssGBLUP$_{DPQ+BQ}$ are suitable for implementation in a genomic honey bee breeding program.

4   First large-scale genomic prediction in the honey bee

This paper has been submitted to Heredity as:

The preprint is archived on ArXiv at:

The genotypes used for this study are available in (Jones et al. 2020). The phenotype data of this study belongs to several breeding associations and is unavailable due to legal reasons.

ABSTRACT

Genomic selection has increased genetic gain in several livestock species, but due to the complicated genetics and reproduction biology not yet in honey bees. Recently, 2970 queens were genotyped to gather a reference population. For the application of genomic selection in honey bees, this study analyses the predictive ability and bias of pedigree-based and genomic breeding values for honey yield, three workability traits and two traits for resistance against the parasite *Varroa destructor*. For breeding value estimation, we use a honey bee-specific model with maternal and direct effects, to account for the contributions of the workers and the queen of a colony to the phenotypes. We conducted a validation for the last generation and a five-fold cross-validation. In the validation for the last generation, the predictive ability of pedigree-based estimated breeding values was 0.06 for honey yield, and ranged from 0.2 to 0.41 for the workability traits. The inclusion of genomic marker data improved these predictive abilities to 0.11 for honey yield, and a range from 0.22 to 0.44 for the workability traits. The inclusion of genomic data did not improve the predictive ability for the disease related traits. Traits with high heritability for maternal effects compared to the heritability for direct effects showed the most promising results. Across all traits, the bias with genomic methods was close to the bias with pedigree-based BLUP. The results show that genomic selection can successfully be applied to honey bees.

4.1 INTRODUCTION

Genomic selection (Meuwissen et al. 2001) incorporates genome-wide marker data into breeding value estimation. Compared to pedigree-based breeding values, the use of genomic data can increase the predictive ability of estimated breeding values, or enable the selection of animals before they are phenotyped. Both strategies have been realised to increase the genetic gain in several livestock species (Doublet et al. 2019; Fulton 2012; Samorè and Fontanesi 2016). Honey bee breeders, by contrast, employ phenotypic selection (De la Mora et al. 2020; Maucourt et al. 2020) or pedigree-based breeding value estimation (Bienefeld et al. 2007; Brascamp et al. 2016; Hoppe et al. 2020). Recently, a high density SNP chip was developed and genotypes of phenotyped queens are now available to validate genomic prediction (Jones et al. 2020).

Pedigree-based best linear unbiased prediction (**PBLUP**) of breeding values began in 1994 for the population registered on Beebreed. The estimated breeding values enabled hundreds of mostly Central European bee breeders to improve the quality of their stock (Hoppe et al. 2020). To ensure the quality of the estimated breeding values, the program relies on a specialized infrastructure for mating control and an adapted genetic model to account for the peculiarities of the honey bee (Bienefeld et al. 2007; Brascamp and Bijma 2014).

While the up to 50,000 workers in a hive do not reproduce under normal circumstances, they perform all other tasks in the hive, such as foraging, brood care, and cleaning (Koeniger et al. 2015). A queen mates with six to twenty drones during flight (Tarpy and Nielsen 2002). Therefore, the daughters of the queen will belong to various patrilines, and uncertain paternity is a challenge for the breeding value estimation. However, lasting breeding success in honey bees requires adequate measures for mating control (Plate et al. 2019b). Therefore, mating stations are maintained with several drone producing queens (**DPQ**), which are usually unselected daughters of s single dam of high genetic quality.

The phenotypes of honey bee colonies for economically relevant traits result from the collaboration of worker groups and queens. In honey yield for example, the workers of a colony perform foraging and storing, but the queen affects the number of workers via her egg-laying rate, and influences the behavior of the workers via pheromones. Therefore, the genetic model

for the traits includes direct and maternal effects for the contribution of workers and queens, respectively.

In commercial honey bee breeding programs, the demands of beekeepers lead to selection traits which differ significantly in terms of methodology and effort for recording and mathematical modelling. Typical aims include increased honey yield, better workability for the beekeeper, and more disease resistance (Petersen et al. 2020; Uzunov et al. 2017). Especially resistance against *Varroa destructor* is targeted, since this parasitic mite contributes to severe colony losses in numerous countries (Genersch et al. 2010; Guichard et al. 2020a; Traynor et al. 2016).

Genomic breeding value estimation in honey bees has been tried in simulation studies, and single step genomic BLUP (**ssGBLUP**) appeared as an efficient solution (Bernstein et al. 2021; Gupta et al. 2013) to combine pedigree information with genomic information. The simulations showed that ssGBLUP can increase the accuracy of genomic breeding values considerably and enables high genetic gains, if the infrastructure is appropriately adapted. Augmenting ssGBLUP with trait-specific weights leads to weighted ssGBLUP (**WssGBLUP**) (Wang et al. 2012), which can increase the prediction accuracy further, as results from other species have shown (Lourenco et al. 2014; Teissier et al. 2019; Vallejo et al. 2019).

To our knowledge, only simulated results on genomic estimated breeding values in honey bees have been published until now. In this study, we firstly report the predictive abilities and the bias of PBLUP, ssGBLUP, and WssGBLUP for a number of key traits of economic importance in a large breeding population of honey bees.

4.2 MATERIALS AND METHODS

4.2.1 Data

Pedigree and performance data from the *Apis mellifera carnica* population were used, since the genotyped queens belonged to this subspecies, which is native and widespread in Central Europe (Lodesani and Costa 2003; Ruttner 1988; Wallberg et al. 2014). The data was downloaded from BeeBreed on the 14th of February 2021, totalling 201,304 valid performance tests and pedigree data of 234,519 queens. The dataset was reduced and refined to the queens and performance tests relevant to compare classical and genomic selection, as follows. Performance tested queens on apiaries from test year 2010 on formed the set of potential

phenotypes. Queens with a valid phenotype whose genotypes passed the quality control (see below) defined the seed. In an iterative process, phenotypes were added by (1) the completion of testing apiaries, (2) the completion of sister groups, and (3) the closing of pedigree gaps, until no further phenotypes could be added. Finally, the full ancestry of all resulting queens was added without phenotype. The final enriched dataset contained 36,509 phenotypes in a pedigree of 44,183 queens and 4512 sires composed of DPQ. Table 4.1 lists the countries of origin for all colonies.

Table 4.1 *Number of phenotyped and genotyped queens included in the data set by country.*

Country	Phenotyped queens	Genotyped queens after quality control
Germany	24,019	1982
Austria	9 618	372
Italy	796	1
Switzerland	619	17
Ukraine	467	0
Belgium	368	4
Netherlands	275	11
Sweden	133	0
France	117	0
Croatia	91	2

The phenotypes covered honey yield, gentleness, calmness, swarming drive, hygienic behavior, and *Varroa* infestation development (**VID**). Honey yield was measured in kg, and the values were corrected for outliers as described in (Hoppe et al. 2020). Gentleness, calmness, and swarming tendency were recorded as marks from 1 to 4 where 4 is best. Records for these traits were discarded if all colonies on an apiary received the same mark. For hygienic behavior, larvae were artificially killed with a pin and the percentage of cleared cells was recorded (Büchler et al. 2013). VID indicates the resistance of a colony against *Varroa*, based on the

change of the level of *Varroa* infestation from early spring to late summer (see Hoppe et al. 2020 for the calculation of VID). For a measurement of *Varroa* infestation, a bee sample is taken from the hive, and the number of mites per 10g bees is determined (Büchler et al. 2013). Table 4.2 shows the descriptive statistics of the phenotypes available for each trait.

***Table 4.2** Descriptive statistics for honey yield, gentleness, calmness, swarming drive, hygienic behavior, and Varroa infestation development (VID). Honey yield is given in kg. Marks from 1 to 4 were recorded for gentleness, calmness, and swarming drive. Hygiene is given as the percentage of cleared cells. VID is a Varroa resistance score and higher values indicate more resistance.*

Trait	Number of records	Number of genotyped queens with record	Average size of apiaries with a genotyped queen (SD)	Mean	SD	Min.	Max.
Honey yield	35,888	2,046	13.62 (8.33)	40.71	22.84	0	199.8
Gentleness	35,187	2,013	13.80 (8.48)	3.52	0.48	1	4
Calmness	34,652	2,016	13.76 (8.50)	3.49	0.48	1	4
Swarming drive	26,937	1,549	14.57 (8.88)	3.55	0.76	1	4
Hygienic behavior	23,924	1,781	13.36 (7.86)	62.26	23.13	0	100
VID	24,650	1,787	13.48 (7.85)	-1.55	2.38	-77.12	6.93

The 100-K-SNP chip (Jones et al. 2020) was used to genotype 2970 queens which were registered on BeeBreed and born between 2009 and 2017. Markers which were called in less than 90% of the samples, had minor allele frequency below 1%, or showed significant deviations from Hardy-Weinberg-equilibrium after Bonferroni-correction (chi-square p-value $< 0.05 \times 10^{-5}$) were removed. This left 63,240 markers for further analysis. 312 queens were

removed because less than 90% of all the valid markers were called in their samples. After comparisons of daughter and parent based on the number of opposing homozygotes, 207 queens were removed. Subsequently, 62 samples were removed based on comparison of genomic and classic relationship matrix (Calus et al. 2011). This left 2389 genotyped queens for further analysis.

4.2.2 Model and genetic parameters

The complex collaboration between the workers and the queen of a colony must be reflected in the model, and carefully analyzed in the calculation of genetic parameters (Brascamp and Bijma 2019). The phenotype, y, of a colony is modeled as

$$y = a_W + m_Q + e, \tag{4.1}$$

where a_W is the direct effect of the worker group in the colony, and m_Q the maternal effect of the queen in the colony, while e is a non-heritable residual.

The phenotypic variance was calculated according to formula (2) in (Brascamp and Bijma 2019) as

$$\sigma_{ph}^2 = A_{base}\sigma_a^2 + \sigma_m^2 + \sigma_{am} + \sigma_e^2, \tag{4.2}$$

where σ_a^2 and σ_m^2 are the additive genetic variances of direct and maternal effects, σ_{am} is the covariance between direct and maternal effects, σ_e^2 is the residual variance, and A_{base} is the average relationship between two workers of the same colony in the base population. The variance components were estimated via AIREML with the complete phenotypic information, using the model for PBLUP (see below). We used $A_{base} = 0.40$ (Brascamp and Bijma 2019), because even the oldest queens in our pedigree came from populations with established mating control (Armbruster 1919). The heritabilities of direct and maternal effects, and the total heritability, h_a^2, h_m^2, and h_T^2 were calculated according to formulas (6b), (6c), and (7c) in (Brascamp and Bijma 2019) , respectively, as

$$h_a^2 = A_{base}\,\sigma_a^2/\sigma_{ph}^2 \;,\; h_m^2 = \sigma_m^2/\sigma_{ph}^2 \text{ and } h_T^2 = \frac{\sigma_a^2+\sigma_m^2+2\sigma_{am}}{\sigma_{ph}^2}. \tag{4.3}$$

However, because the Beebreed dataset relies on colony based selection (CBS), we calculate the heritability of the corresponding selection criterion, h_{CBS}^2 (derived as formula (6) in Bernstein et al. 2021; called accessible heritability in Hoppe et al. 2020), as

$$h_{CBS}^2 = A_{base} h_T^2. \tag{4.4}$$

4.2.3 Breeding value estimation

The following mixed linear model was used for PBLUP.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_a \mathbf{a} + \mathbf{Z}_m \mathbf{m} + \mathbf{e}, \tag{4.5}$$

where $\mathbf{y}$ is a vector of observations on colonies; $\mathbf{b}$ a vector of fixed effects (year and apiary); $\mathbf{a}$ a vector of direct effects of queens, worker groups or sires; $\mathbf{m}$ a vector of maternal effects of queens, worker groups or sires; $\mathbf{e}$ a vector of residuals; and $\mathbf{X}$, $\mathbf{Z}_a$, and $\mathbf{Z}_m$ are known incidence matrices for $\mathbf{b}$, $\mathbf{a}$, and $\mathbf{m}$, respectively. The expected values of $\mathbf{a}$, $\mathbf{m}$, and $\mathbf{e}$ were assumed to equal $\mathbf{0}$, with the following variances:

$$\mathrm{Var}\begin{pmatrix} \mathbf{a} \\ \mathbf{m} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \sigma_a^2 \mathbf{A} & \sigma_{am} \mathbf{A} & \mathbf{0} \\ \sigma_{am} \mathbf{A} & \sigma_m^2 \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \mathbf{I} \end{pmatrix}, \tag{4.6}$$

where $\mathbf{A}$ is the honey bee specific numerator relationship matrix derived from pedigree (Brascamp and Bijma 2014), $\mathbf{I}$ is an identity matrix, and σ_a^2, σ_m^2, σ_{am} and σ_e^2 are the additive genetic variance of worker and queen effects, their covariance and the residual variance, respectively.

The model equation and variances for ssGBLUP were the same as for PBLUP, except for the fact that matrix $\mathbf{H}$ replaced matrix $\mathbf{A}$. Matrix $\mathbf{H}$ was constructed from the numerator relationship matrix $\mathbf{A}$ which is calculated from pedigree information, and the marker information in the following steps (Aguilar et al. 2010; Christensen and Lund 2010). The genomic relationship matrix , $\mathbf{G}$, (VanRaden 2008, method 1) was constructed by the following equation.

$$\mathbf{G} = \frac{\mathbf{ZZ}^T}{2\sum_i p_i(1-p_i)}, \tag{4.7}$$

where p_i is the allele frequency of the SNP at locus i; $\mathbf{Z} = \mathbf{M} - \mathbf{P}$ with $\mathbf{M}$ containing the marker information of all genotyped queens given as 0, 1, 2, and matrix $\mathbf{P}$ defined column-wise by $P_{ij} = 2p_i$ for all j. Matrix $\mathbf{G}$ was adjusted to $\mathbf{A}$ by adjusting the means of diagonal and off-diagonal elements as described by (Christensen et al. 2012). To have an invertible genomic relationship matrix, we used the weighted genomic relationship matrix, $\mathbf{G}_w$, given by the following equation.

$$\mathbf{G}_w = 0.95\mathbf{G} + 0.05\mathbf{A}_g, \tag{4.8}$$

where $\mathbf{A}_g$ is the submatrix of $\mathbf{A}$ relating to the genotyped animals. Finally, the inverse of $\mathbf{H}$ was computed according to the following formula.

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_w^{-1} - \mathbf{A}_g^{-1} \end{pmatrix}. \tag{4.9}$$

The model equation and variances for WssGBLUP were the same as for ssGBLUP, except for the fact that matrix $\mathbf{G}^*$ replaced matrix $\mathbf{G}$. Matrix $\mathbf{G}^*$ was constructed from the vectors of direct and maternal additive genetic effects, $\mathbf{a}$ and $\mathbf{m}$, and the genomic relationship matrix $\mathbf{G}_w$, which were obtained from ssGBLUP. The vectors of the direct and maternal SNP effects, $\mathbf{u}$ and $\mathbf{v}$, were estimated by:

$$\mathbf{u} = \lambda\mathbf{M}^T\mathbf{G}_w^{-1}\mathbf{a},$$
$$\mathbf{v} = \lambda\mathbf{M}^T\mathbf{G}_w^{-1}\mathbf{m}, \tag{4.10}$$

with $\lambda = \frac{1}{2\sum_i p_i(1-p_i)}$, where p_i and $\mathbf{M}$ have the same value as in ssGBLUP. SNP weights $\mathbf{d}$ were calculated using the average of the direct and maternal SNP effects, deviating from the original algorithm which considered only single-trait models (Wang et al. 2012) as follows.

$$d_i = \left(\frac{u_i + v_i}{2}\right)^2 2p_i(1-p_i). \tag{4.11}$$

Diagonal matrix $\mathbf{D}$ was defined by $D_{ii} = d_i / \overline{\mathbf{d}}$, where $\overline{\mathbf{d}}$ is the average of $\mathbf{d}$. The trait-specific matrix $\mathbf{G}^*$ was calculated by the following formula.

$$\mathbf{G}^* = \frac{\mathbf{ZDZ}^T}{2\sum_i p_i(1-p_i)}, \tag{4.12}$$

where $\mathbf{Z}$ is the same matrix as in ssGBLUP.

Programs from the BLUPf90 software (Misztal et al. 2002) were used to estimate the genetic parameters, predict breeding values and calculate relationship matrices $\mathbf{G}$ and $\mathbf{G}^*$. To account for the specifics of honey bees, PInCo (Bernstein et al. 2018) was used to calculate the pedigree-based relationship matrices. Equations (4.8) to (4.11) were implemented in R (R Development Core Team 2020).

4.2.4   Validation

We performed two types of cross-validation. In the generation validation, estimated breeding values (**EBV**) were predicted using PBLUP, ssGBLUP and WssGBLUP (1) without the phenotypes of all queens born in 2017 or later, and (2) without the phenotypes of queens born in 2016 or later. The EBV of the 265 genotyped queens born in 2017 from scenario 1 were pooled with EBV of the 994 genotyped queens born in 2016 from scenario 2.

In the five-fold cross-validation, only apiaries with at least five performance tested queens were included to ensure reliable estimates of fixed effects. This left 1281 genotyped queens for the validation. The 1281 queens were split into five partitions, where apiaries were evenly distributed onto the partitions. For each partition, EBV were estimated using PBLUP, ssGBLUP and WssGBLUP without the phenotypes of the animals on this partition, and the results from all partitions were pooled. The procedure was repeated six times.

To assess the predictive ability of PBLUP, ssGBLUP and WssGBLUP, *predicted phenotypes* were correlated to *phenotypes corrected for fixed effects*, where the predicted phenotype of a colony is the sum of the direct effect of its worker group and the maternal effect of its queen. For each method to predict EBV, the phenotypes corrected for fixed effects were calculated using fixed effects from the same method. In the generation-validation, PBLUP, ssGBLUP and WssGBLUP were run on the complete data set to obtain appropriate fixed effects. In the five-

fold cross-validation, the fixed effects for the correction of the phenotypes were taken from the same run of the same partition as the predicted phenotypes.

A bootstrap procedure was used to test whether the predictive abilities of WssGBLUP and ssGLUP were significantly higher than the predictive ability of PBLUP. 10,000 bootstrap sample vectors were constructed by sampling validation queens with replacement, and the predictive ability with PBLUP, ssGLUP, and WssGBLUP was calculated for each vector. Two methods were considered significantly different, if the same method had higher predictive ability in 97.5% of all sample vectors (p-value of 0.05 in a two-sided test). Similar bootstrapping methods were used in other studies (Iversen et al. 2019; Legarra et al. 2008).

The regression coefficient of *realized total EBV of queens* on *predicted total EBV of queens*, b_1, was used as a measure of bias. Values of $b_1 < 1$ and $b_1 > 1$ indicate inflation and deflation of the predicted breeding values compared to the realized total EBV, respectively. For both the generation validation and the five-fold cross-validation, the realized total EBV of a queen and its predicted total EBV were the sum of its direct and maternal effect calculated using the complete data set, and restricted data, respectively. The predicted total EBV of PBLUP, ssGBLUP, and WssGBLUP were compared to realised total EBV from the same method.

4.3   RESULTS

4.3.1   Genetics parameters

Estimates of the genetic parameters are shown in Table 4.3. The total heritability was very high for gentleness and calmness, medium for hygienic behavior, honey yield and swarming drive, low for VID. All traits showed considerable negative genetic correlations between maternal and direct effects. The heritability for direct effects was considerably larger than the heritability for maternal effects in gentleness, calmness, and hygienic behavior, but equal to or smaller than the heritability for maternal effects for all other traits.

Table 4.3 *Estimated variance and covariance components, genetic parameters derived from these (co)variances. The approximate standard deviations are given in brackets. VID, Varroa infestation development; the last nine columns show the additive genetic variances of direct (σ_a^2) and maternal effects (σ_m^2), their covariance (σ_{am}), the residual variance (σ_e^2), the heritabilities of direct effects (h_a^2), maternal effects (h_m^2), the genetic correlation (r_G), the total heritability (h_T^2), and the heritability for the selection criterion in colony based selection (h_{CBS}^2).*

Trait	σ_a^2	σ_m^2	σ_{am}	σ_e^2	h_a^2	h_m^2	r_G	h_T^2	h_{CBS}^2
Honey yield	27.235 (5.406)	13.841 (3.015)	-6.124 (3.372)	61.549 (1.476)	0.136 (0.027)	0.173 (0.037)	-0.315 (0.142)	0.360 (0.047)	0.144 (0.019)
Gentleness	0.134 (0.016)	0.027 (0.006)	-0.020 (0.008)	0.062 (0.003)	0.435 (0.053)	0.221 (0.049)	-0.325 (0.096)	0.991 (0.083)	0.396 (0.033)
Calmness	0.103 (0.013)	0.019 (0.005)	-0.013 (0.006)	0.059 (0.003)	0.387 (0.048)	0.181 (0.043)	-0.289 (0.105)	0.907 (0.076)	0.363 (0.03)
Swarming drive	0.143 (0.033)	0.054 (0.017)	-0.008 (0.019)	0.356 (0.010)	0.124 (0.029)	0.117 (0.037)	-0.087 (0.249)	0.394 (0.056)	0.158 (0.022)
Hygienic behavior	111.61 (22.027)	25.571 (8.508)	-5.454 (10.688)	174.73 (5.688)	0.186 (0.037)	0.107 (0.035)	-0.102 (1.645)	0.527 (0.064)	0.211 (0.026)
VID	0.159 (0.045)	0.068 (0.024)	-0.028 (0.027)	0.619 (0.014)	0.088 (0.025)	0.095 (0.033)	-0.270 (0.275)	0.236 (0.044)	0.095 (0.018)

4.3.2 Accuracy of breeding values

The predictive abilities of the methods under investigation in the generation validation are shown in Figure 4.1. Compared to PBLUP, the predictive ability was improved with WssGBLUP for honey yield (94%), swarming drive (7%), gentleness (6%), calmness (5%), and VID (20%), and with ssGBLUP, improvements were observed for honey yield (48%), VID (41%), and gentleness (6%). The improvement with WssGBLUP over PBLUP for honey yield was statistically significant. No improvement was observed for hygienic behavior, and ssGBLUP did not yield a higher accuracy than PBLUP for calmness and swarming drive.

Figure 4.1 Predictive abilities of pedigree-based BLUP (PBLUP), single step genomic BLUP (ssGBLUP) and weighted ssGBLUP (WssGBLUP) in the generation validation.

The accuracies of the methods under investigation in the five-fold cross-validation are shown in Figure 4.2. Improvements over PBLUP were achieved for swarming drive (20%), honey yield (15%), calmness (2%), and gentleness (3%) with WssGBLUP. Improvement over PBLUP with ssGBLUP were achieved for honey yield (10%), and swarming drive (3%). The improvements with WssGBLUP over PBLUP were statistically significant for calmness and swarming drive. No improvement was observed for hygienic behavior and VID.

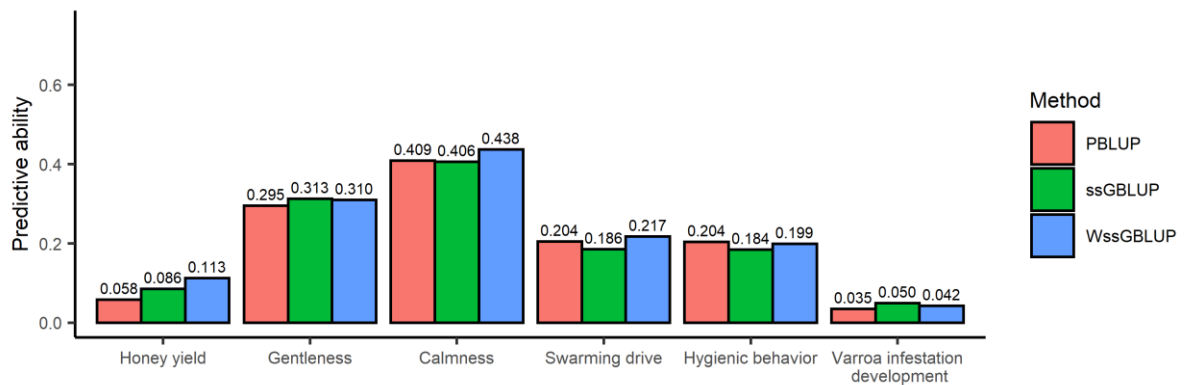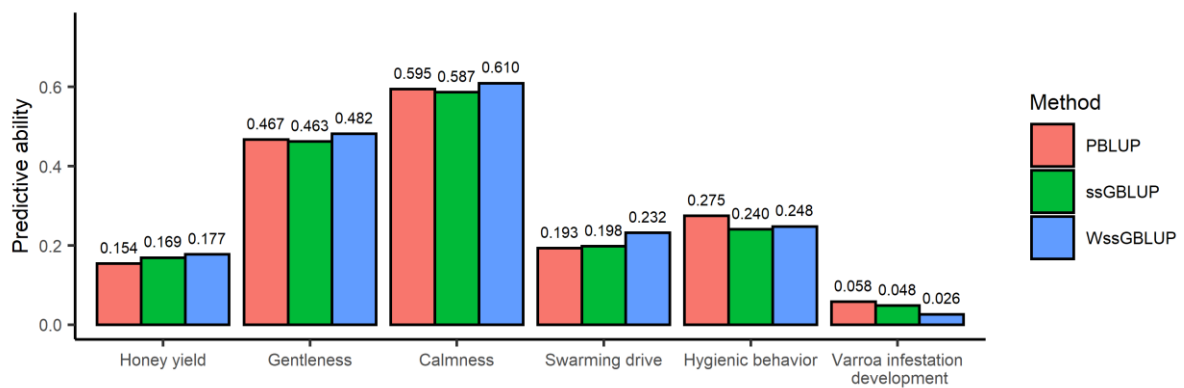*Figure 4.2 Mean Predictive abilities of pedigree-based BLUP (PBLUP), single step genomic BLUP (ssGBLUP) and weighted ssGBLUP (WssGBLUP) in the five-fold cross-validation, calculated across the six repetitions. The standard deviations over the six repetitions are not shown, as they were smaller than 0.007.*

Overall, both validations showed similar results, although the predictive ability was higher in the five-fold cross-validation, and the increases in predictive ability with ssGBLUP and WssGBLUP over PBLUP were higher in the generation validation.

4.3.3 Bias of breeding values

Bias was calculated as the regression coefficient b_1 of total realised EBV on total predicted EBV. The results for EBV from PBLUP, ssGBLUP and WssGBLUP in the generation validation are shown in Figure 4.3. The regression coefficient b_1 deviated the most from 1 for honey yield and VID with -0.33, and -0.28, respectively, for PBLUP. The results for WssGBLUP, and ssGBLUP are rather similar to each other, and more stable across the traits than with PBLUP. With WssGBLUP, the regression coefficient b_1 deviated from 1 by a number between -0.13 and -0.26 across all traits. The results for all three methods show inflated EBV estimates.

Figure 4.3 Regression coefficient b_1 of pedigree-based BLUP (PBLUP), single step genomic BLUP (ssGBLUP) and weighted ssGBLUP (WssGBLUP) in the generation validation.

The results for EBV from PBLUP, ssGBLUP and WssGBLUP in the five-fold cross-validation are shown in Figure 4.4. For PBLUP, the regression coefficient b_1 deviated from 1 by a number between -0.20 and -0.36. The deviations of b_1 from 1 for ssGBLUP were slightly higher, and WssGBLUP showed the highest bias for all traits. All three methods showed inflated EBV estimates.
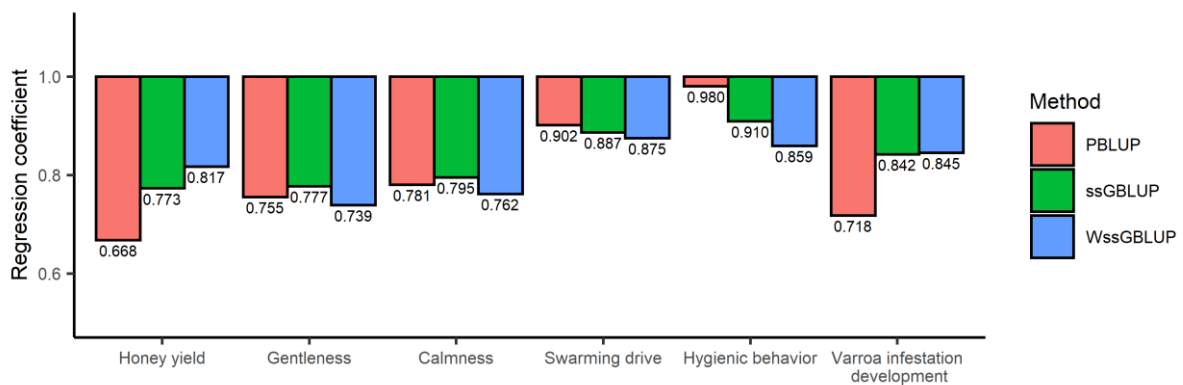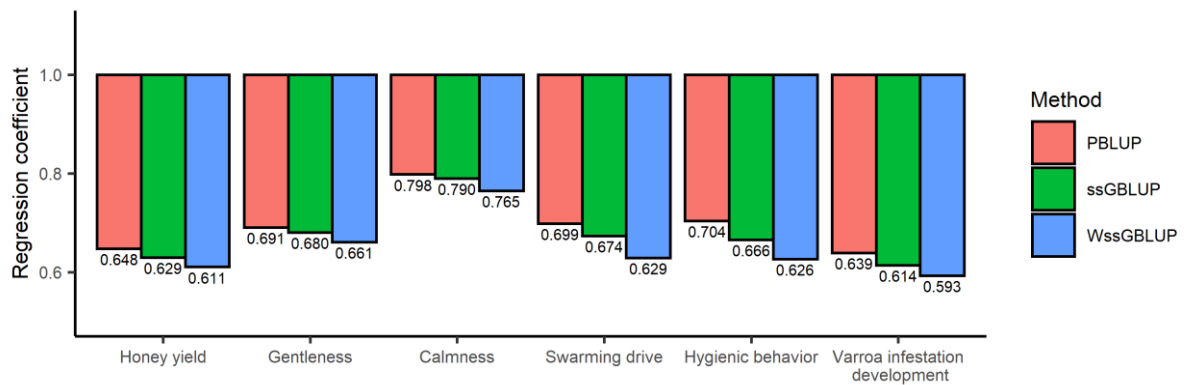
Figure 4.4 Mean Regression coefficient b_1 of pedigree-based BLUP (PBLUP), single step genomic BLUP (ssGBLUP) and weighted ssGBLUP (WssGBLUP) in the five-fold cross-validation, calculated across the six repetitions. The standard deviations over the six repetitions are not shown, as they were smaller than 0.004.

4.4   DISCUSSION

4.4.1   Genetic parameters and predictive ability

The estimated genetic parameters (Table 4.3) are in line with the results for the multiple trait models of the complete Beebreed data set (Hoppe et al. 2020). The results on the predictive abilities in the generation validation (Figure 4.1) and in the five-fold cross-validation (Figure 4.2) show improvements with WssGBLUP over PBLUP for honey yield, gentleness, calmness and swarming drive. These results were within the range reported for data sets of similar size in dairy goats (Legarra et al. 2014a), or for traits affected by maternal effects in beef cattle (Lourenco et al. 2015) or pigs (Putz et al. 2018).

The results on the difference in predictive ability between WssGBLUP and PBLUP can be explained with the results on the heritabilities (Table 4.3). Because simulation studies in honey bees showed greater increases in predictive ability with ssGBLUP over PBLUP for maternal effects than for direct effects (Bernstein et al. 2021), traits with a higher heritability for maternal effects than for direct effects can be expected to show higher increases than other traits in predictive ability with WssGBLUP and ssGBLUP over PBLUP. Honey yield and swarming

drive showed the highest improvements in predictive ability with WssGBLUP over PBLUP, and the heritability for maternal effects is equal to or greater than the heritability for direct effects in both traits. Although the heritability for maternal effects is also equal to the heritability for direct effects in VID, the results for this trait are rather ambiguous. This is due to the low total heritability for this trait, because simulation studies in honey bees and other species show that traits with low heritability also have low accuracy of pedigree-based and genomic EBV (Gowane et al. 2019; Gupta et al. 2013). This result stands out from other species where maternal effects are modelled, as in beef cattle (Lourenco et al. 2018) and simulation studies for beef cattle and pigs (Lourenco et al. 2013; Putz et al. 2018), the predictive ability for direct effects showed higher increases in predictive ability with ssGBLUP over PBLUP than the predictive ability for maternal effects.

An analogy between direct effects in honey bees and traits in other species can explain, why traits with a higher heritability for direct effects than for maternal effects (e.g. hygienic behavior) showed less increase in predictive ability from PBLUP to WssGBLUP than other traits of similar total heritability (e.g. swarming drive). For our study, only queens were genotyped, since they are the candidates of selection. The primary selection candidates in dairy cattle or pigs are males. For traits for which only females are phenotyped, simulations showed that male selection candidates should be genotyped to achieve high genetic gain, but including additional genotypes of females increases the accuracy of prediction considerably (Buch et al. 2012; Lillehammer et al. 2011; Plieschke et al. 2016). The situation of the direct effects of queens and workers in our study is similar to the situation of males and females in dairy cattle and specific traits in pigs, respectively, since direct effects quantify the impact of workers on the phenotypes of colonies. The results on dairy cattle and pigs suggest that genotyping queens has the highest priority, but including genotypes of workers might increase the accuracy of genomic prediction, especially for direct effects. To our knowledge, there are no studies in honey bees on this subject. However, since the workers of a hive belong to different patrilines, the price for genotyping enough workers to represent the workers as a whole is probably too high in practice.

The results for the *Varroa* resistance related traits were also by problems in gathering data. The number of genotyped queens with phenotype for both traits was about 200 queens lower than for honey yield, gentleness, and calmness. Furthermore, the number of phenotyped queens on

apiaries with a genotyped queen (Table 4.2) was low for the *Varroa* related traits, which might have led to less accurate fixed effects. However, *Varroa* specific hygienic behavior is the subject of ongoing research (Conlon et al. 2019; Farajzadeh et al. ; Mondet et al. 2020). The discovery of new quantitative trait loci (QTL) which are then covered by causative SNPs on a new chip can increase predictive ability for the *Varroa* related traits considerably.

The predictive ability of ssGBLUP was slightly lower than the predictive ability of WssGBLUP for most traits. This result is common in studies for several other agricultural species using WssGBLUP (e.g. Lu et al. 2020; Teissier et al. 2019; Wang et al. 2014). In simulation studies (Lourenco et al. 2017; Wang et al. 2012), WssGBLUP had higher predictive ability than ssGBLUP when the trait was controlled by few QTL, and both methods showed equal predictive ability when the trait was polygenic. As the predictive ability for VID was higher with ssGBLUP than with WssGBLUP in both validations, the genetic architecture of the trait appears to be highly polygenic. However, this is a preliminary conclusion, as VID has the lowest heritability of the traits we considered, due to the many factors that affect it (see Guichard et al. 2020a for a review).

The predictive abilities in the five-fold cross-validation were for the majority of the traits higher than in the generation validation. This is due to the fact that in the five-fold cross-validation, sibling groups are evenly distributed across the partitions, while the phenotypes of whole sibling groups might be removed for the calculation of EBVs in the generation-validation. Therefore, the five-fold cross-validation is a validation within sibling groups, while the generation validation is similar to a validation across sibling groups. Studies in other species found that validations within sibling groups show higher predictive abilities than validations across siblings groups (Gao et al. 2019; Kjetså et al. 2020; Legarra et al. 2008). The standard deviations of the predictive abilities in the five-fold cross-validation were extremely small in our study, but the predictive abilities for individual partitions showed large differences.

According to a simulation study in honey bees (Bernstein et al. 2021), the size of the reference population in our study is close to the minimal size which should be available to initiate a breeding program. We expect the reference population to grow in the future, when breeders start to apply genomic selection.

The larger reference population is likely to obviate the need to run WssGBLUP instead of ssGBLUP, since a simulation study showed that WssGBLUP and ssGBLUP yield the same results for large reference sets (Lourenco et al. 2017). The larger reference population will also result in an increase of the predictive ability of genomic methods, as results from other species demonstrate (Daetwyler et al. 2012; Lourenco et al. 2015; Mehrban et al. 2017; Moser et al. 2009).

4.4.2 Bias of the estimated breeding values

The regression coefficients, b_1 of realised total EBV of queens on predicted total breeding values of queens are shown in Figure 4.3 for the generation validation. The results for WssGBLUP deviate slightly further than -0.15 from 1 for honey yield, gentleness, and calmness. As 0.15 is seen as the maximum deviation from 1 for acceptable EBV (Tsuruta et al. 2011), our results for these traits are biased. The results of the five-fold cross-validation (Figure 4.4) confirm that the estimates show high inflation. The difference between the validations is due to the different ways of accounting for fixed effects. In the generation validation, the same value of the fixed effect was modelled for all colonies on the same apiary in a single year, while in the five-fold cross-validation different values of the fixed effect for colonies on the same apiary in a single year were used, depending on the partition. This difference was exacerbated by the small apiaries in honey bees, which lead to a higher variance of the predicted EBV in the five-fold cross-validation than in the generation validation, and consequently lower regression coefficients. Results from routine evaluations are expected to be closer to the generation validation, where the bias was acceptable for swarming drive, hygienic behavior, and VID, and close to acceptable for the remaining traits.

The bias with genomic methods can be reduced by e.g. increasing the share of the classic relationship matrix $\mathbf{A}_g$ in Equation (4.8) (McMillan and Swan 2017; Misztal et al. 2017). However, considerable bias was neither observed in simulations for honey bees (Bernstein et al. 2021) for PBLUB and ssGBLUP, nor the Austrian data set (Brascamp et al. 2016) with PBLUP. Since, our data set is a small outtake of a very large population, we expect that PBLUP and genomic breeding values will show less bias as the data set increases.

4.4.3 Practical application of genomic selection in the honey bee

Gathering genomic data from honey bees for genomic selection requires special considerations, due to their small body size, and their genetic diversity within a hive. Non-lethal ways to genotype queens are available for genomic selection (Jones et al. 2020). The exuviae which queens leave behind after hatching offer a non-lethal option to genotype virgin queens, but just one exuvia is available for each queen, and exuviae showed low DNA quality in several cases. Alternatively, drones can be gathered from a hive to genotype the queen, since drones are haploid offspring. This method is a viable option to genotype queens before the colony is phenotyped.

The availability of genomic breeding values offers new possibilities in breeding schemes for honey bees. Queens can be genomically preselected by genotyping candidate queens, and keeping only the ones with the highest genomic breeding value for phenotyping or deployment as DPQ on mating stations. A simulation study suggests that genomic preselection can increase the genetic gain per year considerably, if a budget to genotype at least 1000 queens per year is available[1] (Bernstein et al. 2021). Another simulation study suggests, that if the structure of the breeding program is changed so that several generations of queens are genomically preselected in a single summer, the generation interval will at least be halved, enabling even higher genetic gains (Brascamp et al. 2018).

The routine evaluations of queens on a SNP chip can also benefit beekeepers by checking the subspecies of the queens and monitoring the genetic diversity within the population. Conservation efforts have been initiated for several subspecies in Europe (De la Rúa et al. 2009; Janczyk et al. 2021; Pinto et al. 2014), at least in part because locally adapted ecotypes have a higher survivability (Büchler et al. 2014). While microsatellites offer a cheap and widely used way to monitor genetic diversity in honey bees (Meixner et al. 2013), a sufficiently high number of SNPs is more accurate when the SNPs are properly validated (Muñoz et al. 2017).

4.5   CONCLUSIONS

WssGBLUP offers significantly greater predictive ability than PBLUP for honey yield, calmness, and swarming drive. For gentleness, the predictive ability of WssGBLUP was greater

---

[1] The submitted version of the manuscript inaccurately stated that all 1000 genotyped queens had to be phenotyped.

than the predictive ability of PBLUP to a similar degree as for calmness, but the difference remained below the threshold for significance. Across all traits, the bias with WssGBLUP and ssGBLUP was close to the bias with PBLUP. However, ssGBLUP offers too little improvement over PBLUP to be recommended based on the current data set for *Varroa* resistance traits, which is likely due to the size of the reference population. A larger reference population or the discovery of new causative SNPs for *Varroa* resistance are required to increase the predictive ability of genomic methods for hygienic behavior and VID. The results suggest that genomic selection can be successfully applied to honey bees.

5   Conclusion and outlook

This thesis realised the first genomic prediction on a large population of honey bees, and analysed the genomic breeding values in simulations to compare several strategies to optimally use genomic selection in the honey bee. To ensure the highest possible quality for the estimated pedigree-based and genomic breeding values, this thesis presented a fast, memory-efficient algorithm to facilitate the calculation of the honey bee-specific inverse numerator relationship matrix on a large data set, according to the most recent model (Brascamp and Bijma 2014). The algorithm proved to be useful for this thesis (Sections 3.2.6 and 4.2.3), as well as in simulation studies (Du et al. 2021a; 2021b; Plate et al. 2019a), and routine breeding value estimations for European beekeepers (Hoppe et al. 2020).

For the estimation of genomic breeding values, method ssGBLUP was used in this thesis in a simulation study as well as on a real data set, where WssGBLUP was additionally implemented. Both studies validate that the algorithms for genomic prediction offer breeding values of higher quality than pedigree-based prediction. For some traits in the real data set, however, PBLUP had a higher predictive ability than ssGBLUP and WssGBLUP (Section 4.3.2), which is largely due to the limited reference population size, as the simulated data shows (Sections 4.4.1 and 3.3.2). The simulations predict that expanding the proportion of genotyped queens heading phenotyped colonies would yield considerably more accurate genomic breeding values in the real data set. However, differences between the traits must be expected due to their different genetic parameters. Traits which are more affected by the direct effects of the workers than by the maternal effects of the queens, especially the *Varroa* traits, require larger reference populations than traits which are more affected by maternal effects than by direct effects, such as honey yield or swarming drive.

After a reference population has been established, genotyping a minimum number of 1000 queens per year is required at the start of a genomic breeding program in honey bees (Chapter 3). While only breeding queens were genotyped in the real data set, ssGBLUP was adjusted to include the genotypes of DPQs for the simulation study. The genetic gain with several breeding schemes was estimated and compared to the required budget for genotyping queens. Genomic preselection offered significantly higher genetic gain than pedigree-based selection, and can

already be realised since reliable methods for non-lethal genotyping of queens are available (Jones et al. 2020).

To summarise, this thesis provides new insights into the efficient calculation of breeding values in large populations of honey bees which enables complex simulation studies for researchers, and more accurate selection decisions for breeders. The use of genomic data will accelerate genetic gain in economically important traits in honey bees. The thesis described methods for genomic prediction comprehensively, expanded them, and analysed realistic breeding schemes for genomic selection. The results provide guidelines for bee breeders to achieve optimal genetic gains with genomic selection.

The method for breeding value estimation used in this thesis is open to further improvements. To allow more accurate breeding decisions for populations with less common types of mating, e.g. single drone insemination, the software for the calculation of the inverse numerator relationship matrix and coefficients of inbreeding requires further adaptation. Subsequently, simulation studies for genomic selection could optimise nucleus breeding schemes employing single drone insemination along with traditional mating control. Because single drone insemination offers very precise control over the genetic quality of the offspring, breeders can aim for higher genetic gain.

The changes of genetic variance due to genomic selection, and the cost-efficient maintenance of the reference populations over several generations are important concerns for breeders. The software for simulating breeding populations requires further adjustment to investigate these questions, and develop breeding schemes employing shorter generation intervals in honey bees. Such breeding schemes promise even higher genetic gain than GPS, but could also lead to considerable loss of genetic diversity (Appendix L).

Worldwide, there are several subspecies of the honey bee (*Apis mellifera*) and they should be maintained, since native subspecies are usually better adapted to the local environment in a particular area than imported subspecies (Büchler et al. 2014). However, beekeepers often consider outbred subspecies, e.g. *Apis m. carnica* or *Apis m. ligustica*, as economically more viable and the import of foreign queens threatens to hybridize or replace native subspecies (De la Rúa et al. 2009; Theisen-Jones and Bienefeld 2016). While some native populations, e.g. *Apis m. ruttneri* on Malta, are too small for directed selection (Plate et al. 2020), genomic

selection can contribute to the adaptation of native populations of appropriate size to the requirements of beekeepers. Chapter 4 focused on a data set from the Central European *Carnica* population, and the simulated level of linkage disequilibrium in Chapter 3 was taken from the real data set. The level of linkage disequilibrium in the *Carnica* population is rather high compared to other subspecies, although different values have been reported (Supplementary Figures 4 and 1 in Wallberg et al. 2014 and Parejo et al. 2016, respectively). Therefore, genomic selection in other subspecies, is expected to require larger reference populations.

To discover the secrets of honey bee social behaviour and biology and to improve health, production and workability, a better understanding of the honey bee genome is required. Genome-wide association studies establish links between differences in these traits and genetic mutations to broaden our knowledge of QTLs. The genotyped queens used for the genomic prediction in a real data set in this thesis provide promising material for association studies.

References

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci. 93:743-752.

Aguilar I, Misztal I, Tsuruta S, Legarra A, Wang H. 2014. PREGSF90 – POSTGSF90: computational tools for the implementation of single-step genomic selection and genome-wide association with ungenotyped individuals in BLUPF90 programs. Paper presented at: 10th World Congress of Genetics Applied to Livestock Production. Vancouver.

Andonov S, Costa C, Uzunov A, Bergomi P, Lourenco D, Misztal I. 2019. Modeling honey yield, defensive and swarming behaviors of Italian honey bees (Apis mellifera ligustica) using linear-threshold approaches. BMC Genet. 20:78.

Armbruster L. 1919. Bienenzüchtungskunde. Leipzig, Germany: Verlag von Theodor Fisher.

Bernstein R, Du M, Hoppe A, Bienefeld K. 2021. Simulation studies to optimize genomic selection in honey bees. Genet Sel Evol. 53:64.

Bernstein R, Plate M, Hoppe A, Bienefeld K. 2018. Computing inbreeding coefficients and the inverse numerator relationship matrix in large populations of honey bees. J Anim Breed Genet. 135:323-332.

Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O et al. 2006. Exceptionally high levels of recombination across the honey bee genome. Genome Research. 16:1339-1344.

Bienefeld K, Ehrhardt K, Reinhardt F. 2007. Genetic evaluation in the honey bee considering queen and worker effects-a BLUP-animal model approach. Apidologie. 38:77-85.

Bienefeld K, Ehrhardt K, Reinhardt F. 2008. Bee breeding around the world: Noticeable success in honey bee selection after the introduction of genetic evaluation using BLUP. American bee journal. 148:739-742.

Bienefeld K, Pirchner F. 1990. Heritabilities for several colony traits in the honeybee (Apis mellifera carnica). Apidologie. 21:175-183.

Bienefeld K, Pirchner F. 1991. Genetic correlations among several colony characters in the honey bee (Hymenoptera: Apidae) taking queen and worker effects into account. Ann Entomol Soc Am. 84:324-331.

Bienefeld K, Reinhardt F, Pirchner F. 1989. Inbreeding effects of queen and workers on colony traits in the honey bee. Apidologie. 20:439-450.

Boichard D, Ducrocq V, Croiseau P, Fritz S. 2016. Genomic selection in domestic animals: Principles, applications and perspectives. Comptes Rendus Biologies. 339:274-277.

Boudry P, Allal F, Aslam ML, Bargelloni L, Bean TP, Brard-Fudulea S, Brieuc MSO, Calboli FCF, Gilbey J, Haffray P et al. 2021. Current status and potential of genomic selection to improve selective breeding in the main aquaculture species of International Council for the Exploration of the Sea (ICES) member countries. Aquaculture Reports. 20:100700.

Brascamp EW, Bijma P. 2014. Methods to estimate breeding values in honey bees. Genet Sel Evol. 46:53.

Brascamp EW, Bijma P. 2019. A note on genetic parameters and accuracy of estimated breeding values in honey bees. Genet Sel Evol. 51:71.

Brascamp EW, Wanders THW, Wientjes YCJ, Bijma P. 2018. Prospects for genomic selection in honey-bee breeding. Paper presented at: 11th World Congress on Genetics Applied to Livestock Production. Auckland, New Zealand.

Brascamp EW, Willam A, Boigenzahn C, Bijma P, Veerkamp RF. 2016. Heritabilities and genetic correlations for honey yield, gentleness, calmness and swarming behaviour in Austrian honey bees. Apidologie. 47:739-748.

Buch LH, Kargo M, Berg P, Lassen J, Sørensen AC. 2012. The value of cows in reference populations for genomic selection of new functional traits. Animal. 6:880-886.

Büchler R, Andonov S, Bienefeld K, Costa C, Hatjina F, Kezic N, Kryger P, Spivak M, Uzunov A, Wilde J. 2013. Standard methods for rearing and selection of Apis mellifera queens. J Apic Res. 52:1-30.

Büchler R, Berg S, Le Conte Y. 2010. Breeding for resistance to Varroa destructor in Europe. Apidologie. 41:393-408.

Büchler R, Costa C, Hatjina F, Andonov S, Meixner MD, Le Conte Y, Uzunov A, Berg S, Bienkowska M, Bouga M et al. 2014. The influence of genetic origin and its interaction with environmental effects on the survival of Apis mellifera L. colonies in Europe. J Apic Res. 53:205-214.

Calus MPL, Mulder HA, Bastiaansen JWM. 2011. Identification of Mendelian inconsistencies between SNP and pedigree information of sibs. Genet Sel Evol. 43:34.

Christensen OF, Lund MS. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 42:2.

Christensen OF, Madsen P, Nielsen B, Ostersen T, Su G. 2012. Single-step methods for genomic evaluation in pigs. Animal. 6:1565-1571.

Chu TT, Sørensen AC, Lund MS, Meier K, Nielsen T, Su G. 2020. Phenotypically Selective Genotyping Realizes More Genetic Gains in a Rainbow Trout Breeding Program in the Presence of Genotype-by-Environment Interactions. Front Genet. 11:866.

Colleau J-J. 2002. An indirect approach to the extensive calculation of relationship coefficients. Genet Sel Evol. 34:409.

Conlon BH, Aurori A, Giurgiu A-I, Kefuss J, Dezmirean DS, Moritz RFA, Routtu J. 2019. A gene for resistance to the Varroa mite (Acari) in honey bee (Apis mellifera) pupae. Mol Ecol. 28:2958-2966.

Daetwyler HD, Swan AA, van der Werf JHJ, Hayes BJ. 2012. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. Genet Sel Evol. 44:33.

De la Mora A, Emsen B, Morfin N, Borges D, Eccles L, Kelly PG, Goodwin PH, Guzman-Novoa E. 2020. Selective Breeding for Low and High Varroa destructor Growth in Honey Bee (Apis mellifera) Colonies: Initial Results of Two Generations. Insects. 11:864.

De la Rúa P, Jaffé R, Dall'Olio R, Muñoz I, Serrano J. 2009. Biodiversity, conservation and current threats to European honeybees. Apidologie. 40:263-284.

de Roos APW, Schrooten C, Veerkamp RF, van Arendonk JAM. 2011. Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. J Dairy Sci. 94:1559-1567.

Doublet A-C, Croiseau P, Fritz S, Michenet A, Hozé C, Danchin-Burge C, Laloë D, Restoux G. 2019. The impact of genomic selection on genetic diversity and genetic gain in three French dairy cattle breeds. Genet Sel Evol. 51:52.

Du M, Bernstein R, Hoppe A, Bienefeld K. 2021a. Short-term effects of controlled mating and selection on the genetic variance of honeybee populations. Heredity. 126:733-747.

Du M, Bernstein R, Hoppe A, Bienefeld K. 2021b. A theoretical derivation of response to selection with and without controlled mating in honeybees. Genet Sel Evol. 53:17.

Esfandyari H, Henryon M, Berg P, Thomasen JR, Bijma P, Sørensen AC. 2017. Response to Selection in Finite Locus Models with Nonadditive Effects. J Hered. 108:318-327.

Falconer DS, Mackay TFC. 1996. Introduction to Quantitative Genetics. Longman: London.

Famula TR. 1992. Simple and rapid inversion of additive relationship matrices incorporating parental uncertainty. J Anim Sci. 70:1045-1048.

Farajzadeh L, Wegener J, Momeni J, Nielsen R, Zautke F, Thomsen B, Larsen K, Bendixen C, Bienefeld K. Detection of genes underlying individual hygienic behaviour towards Varroa parasitized brood in honey bees using a pool-sequencing approach. in prep.

Fernando RL, Grossman M. 1989. Marker assisted selection using best linear unbiased prediction. Genet Sel Evol. 21:467.

Fernando RL, Grossman M. 1990. Genetic evaluation with autosomal and X-chromosomal inheritance. Theoretical and Applied Genetics. 80:75-80.

Flores JM, Gil-Lebrero S, Gámiz V, Rodríguez MI, Ortiz MA, Quiles FJ. 2019. Effect of the climate change on honey bee colonies in a temperate Mediterranean zone assessed through remote hive weight monitoring system in conjunction with exhaustive colonies assessment. Science of The Total Environment. 653:1111-1119.

Fulton JE. 2012. Genomic selection for poultry breeding. Anim Front. 2:30-36.

Gallai N, Salles J-M, Settele J, Vaissière BE. 2009. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Ecological Economics. 68:810-821.

Gao N, Teng J, Pan R, Li X, Ye S, Li J, Zhang H, Zhang X, Zhang Z. 2019. Accuracy of whole genome prediction with single-step GBLUP in a Chinese yellow-feathered chicken population. Livest Sci. 230:103817.

García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, Van Tassell CP. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proceedings of the National Academy of Sciences. 113:E3995.

Genersch E, von der Ohe W, Kaatz H, Schroeder A, Otten C, Büchler R, Berg S, Ritter W, Mühlen W, Gisder S et al. 2010. The German bee monitoring project: a long term study to understand periodically high winter losses of honey bee colonies. Apidologie. 41:332-352.

Goddard ME, Hayes BJ. 2007. Genomic selection. J Anim Breed Genet. 124:323-330.

Golden BL, Brinks JS, Bourdon RM. 1991. A performance programmed method for computing inbreeding coefficients from large data sets for use in mixed-model analyses. J Anim Sci. 69:3564-3573.

Gowane GR, Lee SH, Clark S, Moghaddar N, Al-Mamun HA, van der Werf JHJ. 2019. Effect of selection and selective genotyping for creation of reference on bias and accuracy of genomic prediction. J Anim Breed Genet. 136:390-407.

Granleese T, Clark SA, van der Werf JHJ. 2019. Genotyping strategies of selection candidates in livestock breeding programmes. J Anim Breed Genet. 136:91-101.

Guichard M, Dietemann V, Neuditschko M, Dainat B. 2020a. Advances and perspectives in selecting resistance traits against the parasitic mite Varroa destructor in honey bees. Genet Sel Evol. 52:71.

Guichard M, Neuditschko M, Soland G, Fried P, Grandjean M, Gerster S, Dainat B, Bijma P, Brascamp EW. 2020b. Estimates of genetic parameters for production, behaviour, and health traits in two Swiss honey bee populations. Apidologie. 51:876-891.

Gupta P, Conrad T, Spötter A, Reinsch N, Bienefeld K. 2012. Simulating a base population in honey bee for molecular genetic studies. Genet Sel Evol. 44:14.

Gupta P, Reinsch N, Spötter A, Conrad T, Bienefeld K. 2013. Accuracy of the unified approach in maternally influenced traits--illustrated by a simulation study in the honey bee (Apis mellifera). BMC Genet. 14:36.

Habier D, Fernando RL, Dekkers JCM. 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics. 177:2389-2397.

Harbo JR, Harris JW. 1999. Selecting honey bees for resistance to Varroa jacobsoni. Apidologie. 30:183-196.

Henderson CR. 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics. 31:423-447.

Henderson CR. 1976. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics. 32:69-83.

Henderson CR. 1988a. Theoretical Basis and Computational Methods for a Number of Different Animal Models. J Dairy Sci. 71:1-16.

Henderson CR. 1988b. Use of an Average Numerator Relationship Matrix for Multiple-Sire Joining. J Anim Sci. 66:1614-1621.

Henryon M, Berg P, Ostersen T, Nielsen B, Sorensen AC. 2012. Most of the benefits from genomic selection can be realized by genotyping a small proportion of available selection candidates. J Anim Sci. 90:4681-4689.

Hoppe A, Du M, Bernstein R, Tiesler F-K, Kärcher M, Bienefeld K. 2020. Substantial Genetic Progress in the International Apis mellifera carnica Population Since the Implementation of Genetic Evaluation. Insects. 11:768.

Howard JT, Rathje TA, Bruns CE, Wilson-Wells DF, Kachman SD, Spangler ML. 2018. The impact of selective genotyping on the response to selection using single-step genomic best linear unbiased prediction. J Anim Sci. 96:4532-4542.

Iversen MW, Nordbø Ø, Gjerlaug-Enger E, Grindflek E, Lopes MS, Meuwissen T. 2019. Effects of heterozygosity on performance of purebred and crossbred pigs. Genet Sel Evol. 51:8.

Janczyk A, Meixner MD, Tofilski A. 2021. Morphometric identification of the endemic Maltese honey bee (Apis mellifera ruttneri). J Apic Res. 60:157-164.

Jensen AB, Palmer KA, Chaline N, Raine NE, Tofilski A, Martin SJ, Pedersen BV, Boomsma JJ, Ratnieks FLW. 2005. Quantifying honey bee mating range and isolation in semi-isolated valleysby DNA microsatellite paternity analysis. Conservation Genetics. 6:527-537.

Jones JC, Du ZG, Bernstein R, Meyer M, Hoppe A, Schilling E, Ableitner M, Juling K, Dick R, Strauss AS et al. 2020. Tool for genomic selection and breeding to evolutionary adaptation: Development of a 100K single nucleotide polymorphism array for the honey bee. Ecol Evol. 10:6246-6256.

Kjetså MH, Ødegård J, Meuwissen THE. 2020. Accuracy of genomic prediction of host resistance to salmon lice in Atlantic salmon (Salmo salar) using imputed high-density genotypes. Aquaculture. 526:735415.

Koeniger G, Koeniger N, Ellis J, Connor L. 2015. Mating biology of honey bees (Apis mellifera). Kalamazoo: Wicwas Press.

Le Conte Y, Navajas M. 2008. Climate change: impact on honey bee populations and diseases. Rev Sci Tech. 27:485-497, 499-510.

Legarra A, Aguilar I, Misztal I. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 92:4656-4663.

Legarra A, Baloche G, Barillet F, Astruc JM, Soulas C, Aguerre X, Arrese F, Mintegi L, Lasarte M, Maeztu F et al. 2014a. Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J Dairy Sci. 97:3200-3212.

Legarra A, Christensen OF, Aguilar I, Misztal I. 2014b. Single Step, a general approach for genomic selection. Livest Sci. 166:54-65.

Legarra A, Robert-Granié C, Manfredi E, Elsen J-M. 2008. Performance of Genomic Selection in Mice. Genetics. 180:611-618.

Lillehammer M, Meuwissen TH, Sonesson AK. 2011. Genomic selection for maternal traits in pigs. J Anim Sci. 89:3908-3916.

Lillehammer M, Sonesson AK, Klemetsdal G, Blichfeldt T, Meuwissen THE. 2020. Genomic selection strategies to improve maternal traits in Norwegian White Sheep. J Anim Breed Genet. 134:384-394.

Liu H, Zhang X, Huang J, Chen J-Q, Tian D, Hurst LD, Yang S. 2015. Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. Genome Biology. 16:15.

Lodesani M, Costa C. 2003. Bee breeding and genetics in Europe. Bee World. 84:69-85.

Lourenco DAL, Fragomeni BO, Bradford HL, Menezes IR, Ferraz JBS, Aguilar I, Tsuruta S, Misztal I. 2017. Implications of SNP weighting on single-step genomic predictions for different reference population sizes. J Anim Breed Genet. 134:463-471.

Lourenco DAL, Misztal I, Tsuruta S, Aguilar I, Ezra E, Ron M, Shirak A, Weller JI. 2014. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. J Dairy Sci. 97:1742-1752.

Lourenco DAL, Misztal I, Wang H, Aguilar I, Tsuruta S, Bertrand JK. 2013. Prediction accuracy for a simulated maternally affected trait of beef cattle using different genomic evaluation models. J Anim Sci. 91:4090-4098.

Lourenco DAL, Tsuruta S, Fragomeni BO, Masuda Y, Aguilar I, Legarra A, Bertrand JK, Amen TS, Wang L, Moser DW et al. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J Anim Sci. 93:2653-2662.

Lourenco DAL, Tsuruta S, Fragomeni BO, Masuda Y, Aguilar I, Legarra A, Miller S, Moser D, Misztal I. 2018. Single-step genomic BLUP for national beef cattle evaluation in US: from initial developments to final implementation. Paper presented at: 11th World Congress on Genetics Applied to Livestock Production. Auckland, New Zealand.

Lu S, Liu Y, Yu X, Li Y, Yang Y, Wei M, Zhou Q, Wang J, Zhang Y, Zheng W et al. 2020. Prediction of genomic breeding values based on pre-selected SNPs using ssGBLUP, WssGBLUP and BayesB for Edwardsiellosis resistance in Japanese flounder. Genet Sel Evol. 52:49.

Maiorano AM, Assen A, Bijma P, Chen CY, Silva J, Herring WO, Tsuruta S, Misztal I, Lourenco DAL. 2019. Improving accuracy of direct and maternal genetic effects in genomic evaluations using pooled boar semen: a simulation study1. J Anim Sci. 97:3237-3245.

Maucourt S, Fortin F, Robert C, Giovenazzo P. 2020. Genetic Parameters of Honey Bee Colonies Traits in a Canadian Selection Program. Insects. 11:587.

McMillan AJ, Swan AA. 2017. Weighting of genomic and pedigree relationships in single step evaluation of carcass traits in Australian sheep. Paper presented at: 22nd conference of the Association for the Advancement of Animal Breeding and Genetics. Townsville, Queensland, Australia.

Mehrban H, Lee DH, Moradi MH, IlCho C, Naserkheil M, Ibáñez-Escriche N. 2017. Predictive performance of genomic selection methods for carcass traits in Hanwoo beef cattle: impacts of the genetic architecture. Genet Sel Evol. 49:1.

Meixner MD, Pinto MA, Bouga M, Kryger P, Ivanova E, Fuchs S. 2013. Standard methods for characterising subspecies and ecotypes of Apis mellifera. J Apic Res. 52:1-28.

Meuwissen TH, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-1829.

Meuwissen THE, Luo Z. 1992. Computing inbreeding coefficients in large populations. Genet Sel Evol. 24:305.

Misztal I, Bradford HL, Lourenco DAL, Tsuruta S, Masuda Y, Legarra A, Lawlor TJ. 2017. Studies on Inflation of GEBV in Single-Step GBLUP for Type. Paper presented at: 2017 Interbull Meeting. Tallinn, Estonia.

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH. 2002. BLUPF90 and related programs (BGF90). Paper presented at: 7th World Congress on Genetics Applied to Livestock Production. Montpellier, France.

Mondet F, Beaurepaire A, McAfee A, Locke B, Alaux C, Blanchard S, Danka B, Le Conte Y. 2020. Honey bee survival mechanisms against the parasite Varroa destructor: a systematic review of phenotypic and genomic research efforts. Int J Parasitol. 50:433-447.

Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol. 41:56.

Mrode RA. 2005. Linear Models for the Prediction of Animal Breeding Values. Wallingford, UK: CABI Publishing.

Muñoz I, Henriques D, Jara L, Johnston JS, Chávez-Galarza J, De La Rúa P, Pinto MA. 2017. SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered dark European honeybee (Apis mellifera mellifera). Mol Ecol Resour. 17:783-795.

Neumann P, Moritz RFA, van Praagh J. 1999a. Queen mating frequency in different types of honey bee mating apiaries. J Apic Res. 38:11-18.

Neumann P, van Praagh JP, Moritz RFA, Dustmann JH. 1999b. Testing reliability of a potential island mating apiary using DNA microsatellites. Apidologie. 30:257-276.

Oikawa T, Yasuda K. 2009. Inclusion of genetically identical animals to a numerator relationship matrix and modification of its inverse. Genet Sel Evol. 41:25.

Parejo M, Wragg D, Gauthier L, Vignal A, Neumann P, Neuditschko M. 2016. Using Whole-Genome Sequence Information to Foster Conservation Efforts for the European Dark Honey Bee, Apis mellifera mellifera. Frontiers in Ecology and Evolution. 4:140.

Petersen GEL, Fennessy PF, Amer PR, Dearden PK. 2020. Designing and implementing a genetic improvement program in commercial beekeeping operations. J Apic Res.1-10.

Pinto MA, Henriques D, Chávez-Galarza J, Kryger P, Garnery L, van der Zee R, Dahle B, Soland-Reckeweg G, de la Rúa P, Dall' Olio R et al. 2014. Genetic integrity of the Dark European honey bee (Apis mellifera mellifera) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. J Apic Res. 53:269-278.

Plate M, Bernstein R, Hoppe A, Bienefeld K. 2019a. Comparison of infinitesimal and finite locus models for long-term breeding simulations with direct and maternal effects at the example of honeybees. PLOS ONE. 14:e0213270.

Plate M, Bernstein R, Hoppe A, Bienefeld K. 2019b. The importance of controlled mating in honeybee breeding. Genet Sel Evol. 51:74.

Plate M, Bernstein R, Hoppe A, Bienefeld K. 2020. Long-Term Evaluation of Breeding Scheme Alternatives for Endangered Honeybee Subspecies. Insects. 11.

Plieschke L, Edel C, Pimentel ECG, Emmerling R, Bennewitz J, Götz K-U. 2016. Systematic genotyping of groups of cows to improve genomic estimated breeding values of selection candidates. Genet Sel Evol. 48:73.

Pryce JE, Goddard ME, Raadsma HW, Hayes BJ. 2010. Deterministic models of breeding scheme designs that incorporate genomic selection. J Dairy Sci. 93:5455-5466.

Putz AM, Tiezzi F, Maltecca C, Gray KA, Knauer MT. 2018. A comparison of accuracy validation methods for genomic and pedigree-based predictions of swine litter size traits using Large White and simulated data. J Anim Breed Genet. 135:5-13.

Quaas RL. 1976. Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix. Biometrics. 32:949-953.

R Development Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Ruttner F. 1988. Biogeography and taxonomy of honeybees. Berlin, Germany: Springer-Verlag.

Samorè AB, Fontanesi L. 2016. Genomic selection in pigs: state of the art and perspectives. Ital J Anim Sci. 15:211-232.

Sargolzaei M, Iwaisaki H, Colleau JJ. 2005. A fast algorithm for computing inbreeding coefficients in large populations. J Anim Breed Genet. 122:325-331.

Saxton AM. 1983. A comparison of exact and sequential methods in multi-stage index selection. Theoretical and Applied Genetics. 66:23-28.

Schaeffer LR, Kennedy BW, Gibson JP. 1989. The Inverse of the Gametic Relationship Matrix. J Dairy Sci. 72:1266-1272.

Sherman J, Morrison WJ. 1950. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. The Annals of Mathematical Statistics. 21:124-127.

Shumbusho F, Raoul J, Astruc JM, Palhiere I, Elsen JM. 2013. Potential benefits of genomic selection on genetic gain of small ruminant breeding programs1. J Anim Sci. 91:3644-3657.

Smith SP, Allaire FR. 1985. Efficient selection rules to increase non-linear merit: application in mate selection. Génétique, sélection, évolution. 17:387.

Sonesson AK, Meuwissen THE. 2009. Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol. 41:37.

Spivak M, Reuter GS. 2001. Varroa destructor Infestation in Untreated Honey Bee (Hymenoptera: Apidae) Colonies Selected for Hygienic Behavior. Journal of Economic Entomology. 94:326-331.

Stapley J, Feulner PGD, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc Lond, B, Biol Sci. 372:20160455.

Tarpy DR, Nielsen DI. 2002. Sampling Error, Effective Paternity, and Estimating the Genetic Structure of Honey Bee Colonies (Hymenoptera: Apidae). Ann Entomol Soc Am. 95:513-528.

Teissier M, Larroque H, Robert-Granie C. 2019. Accuracy of genomic evaluation with weighted single-step genomic best linear unbiased prediction for milk production traits, udder type traits, and somatic cell scores in French dairy goats. J Dairy Sci. 102:3142-3154.

Theisen-Jones H, Bienefeld K. 2016. The Asian Honey Bee (Apis cerana) is Significantly in Decline. Bee World. 93:90-97.

Tonussi RL, Silva RMdO, Magalhães AFB, Espigolan R, Peripolli E, Olivieri BF, Feitosa FLB, Lemos MVA, Berton MP, Chiaia HLJ et al. 2017. Application of single step genomic BLUP under different uncertain paternity scenarios using simulated data. PLOS ONE. 12:e0181752.

Traynor KS, Rennich K, Forsgren E, Rose R, Pettis J, Kunkel G, Madella S, Evans J, Lopez D, vanEngelsdorp D. 2016. Multiyear survey targeting disease incidence in US honey bees. Apidologie. 47:325-347.

Tsuruta S, Misztal I, Aguilar I, Lawlor TJ. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. J Dairy Sci. 94:4198-4204.

Uzunov A, Brascamp P, Büchler R. 2017. The basic concept of honey bee breeding programs. Bee World. 94:84-87.

Vallejo RL, Cheng H, Fragomeni BO, Shewbridge KL, Gao G, MacMillan JR, Towner R, Palti Y. 2019. Genome-wide association analysis and accuracy of genome-enabled breeding value predictions for resistance to infectious hematopoietic necrosis virus in a commercial rainbow trout breeding population. Genet Sel Evol. 51:47.

Van Grevenhof EM, Van Arendonk JAM, Bijma P. 2012. Response to genomic selection: The Bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. Genet Sel Evol. 44:26.

VanRaden PM. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414-4423.

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 92:16-24.

Verbyla K, Kube P, Evans B. 2018. Implementation of Genomic Selection in Tasmanian Atlantic Salmon. Paper presented at: 11th World Congress on Genetics Applied to Livestock Production. Auckland, New Zealand.

Vercelli M, Novelli S, Ferrazzi P, Lentini G, Ferracini C. 2021. A Qualitative Analysis of Beekeepers' Perceptions and Farm Management Adaptations to the Impact of Climate Change on Honey Bees. Insects. 12.

Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, Evans JD, Mikheyev AS, Robertson HM, Robinson GE, Webster MT. 2019. A hybrid de novo genome assembly of the honeybee, Apis mellifera, with chromosome-length scaffolds. BMC Genom. 20:275.

Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, Simões ZLP, Allsopp MH, Kandemir I, De la Rúa P et al. 2014. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera. Nat Genet. 46:1081-1088.

Wang H, Misztal I, Aguilar I, Legarra A, Fernando RL, Vitezica Z, Okimoto R, Wing T, Hawken R, Muir WM. 2014. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. Front Genet. 5:134.

Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res (Camb). 94:73-83.

Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. 2017. Genomic Selection in Dairy Cattle: The USDA Experience. Annu Rev Anim Biosci. 5:309-327.

Appendix A: Variance of Mendelian sampling terms

In this Appendix, we state the formulas for the variances of the Mendelian sampling terms for single queens, D_i^Q, or between full-sibs, D^{FS}, in the cases where at least one parent is unknown. Note that both terms are needed to calculate the diagonal entry of sire in $\boldsymbol{D}$.

Consider a breeding unit i. Let F_s and F_d be the inbreeding coefficients belonging to the sire and the dam of i, respectively. The coefficient of coancestry between two DPQs in s is F_s^{FS}. Their number is given by n_s. The probability that two full-sibs descend from the same drone is p_1. Their chance of descending from the same DPQ, including p_1, is p_2.

1. If the dam of i is known, but not the sire, then

$$D_i^Q = \frac{3 - F_d}{4},\qquad\qquad\qquad\text{(A.1)}$$

$$D^{FS} = \frac{1}{2}\left[p_1 + \frac{p_2 - p_1}{2}\right].\qquad\qquad\text{(A.2)}$$

2. If the sire of i is known, but not the dam, then

$$D_i^Q = 1 - \frac{1}{n_s}\frac{1 + F_s}{4} - \frac{n_s - 1}{n_s}\frac{F_s^{FS}}{2},\qquad\qquad\text{(A.3)}$$

$$D_i^{FS} = \frac{1}{4}\left[1 + p_1(1 - F_s) + \left(p_2 - \frac{1}{n_s}\right)(1 + F_s - 2F_s^{FS})\right].\qquad\text{(A.4)}$$

3. If the parents of i are unknown, then

$$D_i^Q = 1,\qquad\qquad\qquad\text{(A.5)}$$

$$D_i^{FS} = \frac{1}{n_s} + \frac{n_s - 1}{4n_s}[1 + p_1 + p_2],\qquad\qquad\text{(A.6a)}$$

where we assumed that the full-sibs or DPQs have a common dam. However, when they are unrelated, the following should be used

$$D_i^{FS} = 0.\qquad\qquad\qquad\text{(A.6b)}$$

Appendix B: Correction of two formulas in Brascamp and Bijma (2014)

Here, we derive the formulas (A.4) and (A.2) for the variance of the Mendelian sampling terms in two special cases. Brascamp and Bijma (2014) provided proofs of the remaining formulas. We will use the formula

$$a_i = \frac{1}{2} a_d + \frac{1}{2} a_s + \delta_i, \tag{B.1}$$

which relates the breeding value, a_i, of animal i to the breeding values, a_d and a_s, of its dam, d, and its sire, s, leaving the Mendelian sampling term, δ_i, as a residual. We also require the coefficient of coancestry, F_{is}, of a breeding unit, i, to its sire, s. It is given by

$$F_{is} = \frac{1}{2} \left[\frac{1}{n_s} \frac{1 + F_s}{2} + \frac{n_s - 1}{n_s} F_s^{FS} + F_{ds} \right], \tag{B.2}$$

where n_s the number of DPQs in s. We will use A_{ij} and D_{ij} to denote entries in $\boldsymbol{A}$ and $\boldsymbol{D}$, respectively.

1. A Sire with an Unknown Dam.

Let i be a sire, let s be its sire, and let its dam be unknown. Equation (B.1) takes the form

$$a_i = \frac{1}{2} a_s + \delta_i. \tag{B.3}$$

We take the variance, divide by σ_A^2, and assume that δ_i is independent from a_s.

$$D_{ii} = A_{ii} - \frac{1}{4} A_{ss}. \tag{B.4}$$

We replace the relationships by inbreeding and coefficients of coancestry. Let s contain n_s DPQs.

$$D_{ii} = \frac{1}{n_S} + \frac{n_S - 1}{n_S} 2F_i^{FS} - \frac{1}{4} \frac{1 + F_s}{n_S} - \frac{1}{2} \frac{n_S - 1}{n_S} F_s^{FS} \quad \text{(B.5)}$$

We can use (B.2). The coefficient of coancestry between the DPQs in i, whose dam is unknown follows from (2.10). From this we get

$$D_{ii} = \frac{1}{n_S} - \frac{1}{4} \frac{1 + F_s}{n_S} - \frac{1}{2} \frac{n_S - 1}{n_S} F_s^{FS}$$

$$+ \frac{1}{4} \frac{n_S - 1}{n_S} [1 + 2p_1 + (p_2 - p_1)(1 + F_s) + (1 - p_2) 2F_s^{FS}] . \quad \text{(B.6)}$$

We split the terms into D^Q, which is given by (A.3), and D_i^{FS} according to (2.5). We have

$$D^{FS} = \frac{1}{4} \left[1 + p_1(1 - F_s) + \left(p_2 - \frac{1}{n_S} \right)(1 + F_s - 2F_s^{FS}) \right]. \quad \text{(A.4)}$$

This contradicts the result in case 5 in Appendix 1 of Brascamp and Bijma (2014).

2. Full-sibs with an Unknown Sire.

Let i and j be full-sibs, d their dam, and suppose we do not know their sire. We leave the term relating to the sire out in (B.1), and consider the covariance of their breeding values, a_i and a_j. We treat the Mendelian samplings as independent from the breeding values. After rearranging, we have

$$D_{ij} = A_{ij} - \frac{1}{4} A_{dd} . \quad \text{(B.7)}$$

We replace the relationships by coefficients of inbreeding and coancestry, arriving at

$$D_{ij} = 2F_{ij} - \frac{1}{4}(1 + F_d). \quad \text{(B.8)}$$

The relationship between full-sibs with unknown father is given by

$$F_{ij} = \frac{1}{2} F_{id} + \frac{1}{4} \left[p_1 + \frac{p_2 - p_1}{2} \right]. \quad \text{(B.9)}$$

We replace F_{id}, and arrive at

$$D_{ij} = \frac{1}{2}\left[p_1 + \frac{p_2 - p_1}{2}\right]. \tag{A.2}$$

This contradicts the result in case 6 in Appendix 1 of Brascamp and Bijma (2014).

Appendix C: Inversion of D

The blocks of D are the full-sib groups in the population. Here, we consider a full-sib group and its corresponding submatrix, K. Matrix D is a block diagonal matrix. Therefore, every block can be inverted separately. The value of the ith diagonal entry in K is determined by the type breeding unit that i has, and the number of bees contained in i, if i is a sire. For a queen, the diagonal entry is given by

$$K_{ii} = K_Q. \tag{C.1}$$

The diagonal entry of a sire containing s bees is given by

$$K_{ii} = K_{n_S(s)} . \tag{C.2}$$

The diagonal entry of a worker group is given by

$$K_{ii} = K_W . \tag{C.3}$$

The off-diagonal entries of K are all equal to K_W.

K^{-1} is a symmetric matrix. The value of entry K_{ij}^{-1} in K^{-1} depends on the types of the breeding units i and j. Let n_{BQ} be the number of (breeding) queens in K excluding all DPQs. Different sires may contain different numbers of DPQs. For every non-negative integer r, let $m_{S(r)}$ be the number of sires in K that contain exactly r bees. If there is a worker group in K, then the solutions are given by the following equations. Let i be a queen, then

$$K_{ii}^{-1} = \frac{1}{K_Q - K_W}. \tag{C.4}$$

Let i be a sire containing r DPQs, then

$$K_{ii}^{-1} = \frac{1}{K_{n_S(r)} - K_W}. \tag{C.5}$$

Let i be a worker group, then

$$K_{ii}^{-1} = \frac{1}{K_W} + \frac{n_{BQ}}{K_Q - K_W} + \sum_r \frac{m_{S(r)}}{K_{n_{S(r)}} - K_W} \,. \qquad (C.6)$$

There can only be one worker in a full-sib group. Let i be this worker group, and j a queen, then

$$K_{ij}^{-1} = \frac{1}{K_W - K_Q} \,. \qquad (C.7)$$

Let i again be the worker, and j a sire, then

$$K_{ij}^{-1} = \frac{1}{K_W - K_{n_{S(r)}}} \,. \qquad (C.8)$$

In the remaining cases, it holds that

$$K_{ij}^{-1} = 0 \,. \qquad (C.9)$$

To check these formulas, multiply an arbitrary row of $\boldsymbol{K}$ to an arbitrary row of $\boldsymbol{K}^{-1}$.

If no worker group is present, the formulas change. They can be proven via the Sherman Morrison formula (Sherman and Morrison 1950):

Let $\boldsymbol{W}$ be an invertible $n \times n$ - matrix, u and v column vectors of length n. If $1 + v^T \boldsymbol{W} u \neq 0$ is satisfied, then

$$(\boldsymbol{W}^T + uv^T)^{-1} = \boldsymbol{W}^{-1} - \frac{\boldsymbol{W}^{-1} uv^T \boldsymbol{W}^T}{1 + v^T \boldsymbol{W} u} \,. \qquad (C.10)$$

Let $\boldsymbol{K}$ contain n breeding units. The off-diagonal entry is still denoted by K_W. Putting all together gives

$$K_{ii}^{-1} = \frac{1}{K_{ii} - K_W} - \frac{K_W}{\left(1 + \sum_{i=1}^n \frac{K_W}{K_{ii} - K_W}\right)(K_{ii} - K_W)^2} \,, \qquad (C.11)$$

and for $i \neq j$ we have

$$K_{ij}^{-1} = -\frac{K_W}{\left(1 + \sum_{i=1}^{n} \frac{K_W}{K_{ii} - K_W}\right)(K_{ii} - K_W)\left(K_{jj} - K_W\right)}. \tag{C.12}$$

Appendix D: Storage and calculation of D

In this Appendix, we discuss how the algorithm creates and handles the matrix of Mendelian samplings, $\boldsymbol{D}$. It is a sparse matrix. The non-zero elements are the off-diagonal that connect full-sibs and the diagonal elements.

The diagonal entries of $\boldsymbol{D}$ are stored in a vector, where D_{ii} is stored at the ith position. The off-diagonal entries of $\boldsymbol{D}$ are stored in a separate vector. For every full-sib group, K, just one off-diagonal entry, D_K^{FS}, is stored, because all off-diagonal entries between full-sibs of the same group have the same value.

When the breeding units are ordered so that full-sibs follow each other directly, then $\boldsymbol{D}$ is a block diagonal matrix. The following matrix is an example of 8 breeding units, where 3 to 5 belong to full-sib group 1 while 7 and 8 belong to group 2.

$$
\begin{pmatrix}
D_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & D_{22} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & D_{33} & D_1^{FS} & D_1^{FS} & D_1^{FS} & 0 & 0 \\
0 & 0 & D_1^{FS} & D_{44} & D_1^{FS} & D_1^{FS} & 0 & 0 \\
0 & 0 & D_1^{FS} & D_1^{FS} & D_{55} & D_1^{FS} & 0 & 0 \\
0 & 0 & D_1^{FS} & D_1^{FS} & D_1^{FS} & D_{66} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & D_{77} & D_2^{FS} \\
0 & 0 & 0 & 0 & 0 & 0 & D_2^{FS} & D_{88}
\end{pmatrix}
$$

In the calculation, it is useful to calculate all entries for a full-sib group, K, when its first member is reached. The values of D_K^Q and D_K^{FS} are computed after a case distinction based on which parents are known. This gives us the off-diagonal entry directly, which is also the entry of a worker, and the entry for queens, while the diagonal entries of sires are obtained via (2.5).

Appendix E: Merging DPQs for ssGBLUP$_{\text{DPQ+BQ}}$

In this appendix, we show that merging DPQs of a pseudo-father for ssGBLUP leaves the breeding values of all other animals unchanged. DPQs are not phenotyped. Therefore, it is enough to show, that the relationships among the other animals do not change. To do this, we show:

$$\mathbf{H}_{DPQ+BQ} = \mathbf{B}\widetilde{\mathbf{H}}_{DPQ+BQ}\mathbf{B}^T \qquad (E.1)$$

where $\widetilde{\mathbf{H}}_{DPQ+BQ}$ ($n \times n$) is the relationship matrix for ssGBLUP. We drop the subscript DPQ+BQ for this Appendix. $\widetilde{\mathbf{H}}$ was calculated from the $g \times g$ matrix $\widetilde{\mathbf{G}}$, and $\mathbf{B}$ is a $n \times n$ matrix that merges the rows and columns of a group of DPQs on a mating station. If one PF can be merged, then all PFs can be merged. We consider a single pseudo-father, p, in the top left corner.

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}^T & \mathbf{0}_{1\times(n-8)} \\ \mathbf{0}_{(n-1)\times 8} & \mathbf{I}_{n-8} \end{pmatrix} \qquad (E.2)$$

where vector $\mathbf{b}$ has only $\frac{1}{8}$ as entry with 8 being the size of the block to merge. Note that $\mathbf{G} = \mathbf{B}\widetilde{\mathbf{G}}\mathbf{B}^T$. Christensen and Lund (2010) define $\mathbf{H}$ by $\mathbf{H} = 0.95\mathbf{H}^* + 0.05\mathbf{A}$. where $\mathbf{H}^*$ is defined by (4) in (Christensen and Lund 2010).

$$\mathbf{H}^* = \begin{pmatrix} \mathbf{G} & \mathbf{G}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \\ \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G} & \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} + \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \end{pmatrix} \qquad (E.3)$$

where $\mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{21},$ and $\mathbf{A}_{22}$ are the submatrices of $\mathbf{A}$ for the genotyped animals, relating the genotyped animals to the non-genotyped animals, relating the non-genotyped animals to the genotyped animals, and for the non-genotyped animals, respectively. The analogon holds for $\widetilde{\mathbf{H}}^*$ with $\widetilde{\mathbf{G}}$ and $\widetilde{\mathbf{A}}$. We have $\mathbf{B}\widetilde{\mathbf{H}}\mathbf{B}^T = 0.95\mathbf{B}\widetilde{\mathbf{H}}^*\mathbf{B}^T + 0.05\mathbf{B}\widetilde{\mathbf{A}}\mathbf{B}^T$. Obviously, $\mathbf{B}\widetilde{\mathbf{A}}\mathbf{B}^T = \mathbf{A}$ holds true. We need to show

$$\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G} = \widetilde{\mathbf{A}}_{21}\widetilde{\mathbf{A}}_{11}^{-1}\widetilde{\mathbf{G}}\mathbf{B}^T \qquad (E.4)$$

and

$$\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} = \widetilde{\mathbf{A}}_{21}\widetilde{\mathbf{A}}_{11}^{-1}\widetilde{\mathbf{G}}\widetilde{\mathbf{A}}_{11}^{-1}\widetilde{\mathbf{A}}_{12} \qquad (E.5)$$

and

$$A_{21}A_{11}^{-1}G = B\widetilde{G}\widetilde{A}_{11}^{-1}\widetilde{A}_{21} \tag{E.6}$$

The proofs are very similar, so we only show (E.4). We define submatrices of $\mathbf{A}$.

$$\mathbf{A} = \begin{pmatrix} d & \mathbf{y}^T & \mathbf{x}^T \\ \mathbf{y} & \mathbf{X} & \mathbf{Z}^T \\ \mathbf{x} & \mathbf{Z} & \mathbf{A}_{22} \end{pmatrix} \tag{E.7}$$

where d is diagonal entry of p, $\mathbf{y}$ are the relationships of the other genotyped animals to p, and $\mathbf{x}$ are the relationships of the non-genotyped animals to p, matrix $\mathbf{X}$ holds the relationship of the other genotyped animals, and $\mathbf{Z}$ the relationships of the other genotyped animals to the non-genotyped animals. Matrix $\widetilde{\mathbf{A}}$ has similar submatrices as $\mathbf{A}$.

$$\widetilde{\mathbf{A}} = \begin{pmatrix} \widetilde{\mathbf{D}} & \mathbf{1}\mathbf{y}^T & \mathbf{1}\mathbf{x}^T \\ \mathbf{y}\mathbf{1}^T & \mathbf{X} & \mathbf{Z}^T \\ \mathbf{x}\mathbf{1}^T & \mathbf{Z} & \mathbf{A}_{22} \end{pmatrix} \tag{E.8}$$

where $\widetilde{\mathbf{D}}$ is the relationship matrix of the DPQs within p. As all DPQs have the same pedigree relationships to all other animals, the corresponding submatrices can be written using vectors of ones, called $\mathbf{1}$. Note that $d = \mathbf{b}^T\widetilde{\mathbf{D}}\mathbf{b}$ holds. To prove (E.4), we focus on the following submatrices of $\mathbf{A}$.

$$\mathbf{A}_{11} = \begin{pmatrix} d & \mathbf{y}^T \\ \mathbf{y} & \mathbf{X} \end{pmatrix} \text{ and } \mathbf{A}_{21} = \begin{pmatrix} \mathbf{x} & \mathbf{Z} \end{pmatrix} \tag{E.9}$$

Matrix $\mathbf{G}$ has similar submatrices as $\mathbf{A}_{11}$.

$$\mathbf{G} = \begin{pmatrix} u & \mathbf{b}^T W \\ W\mathbf{b} & V \end{pmatrix} \text{ and } \widetilde{\mathbf{G}} = \begin{pmatrix} \widetilde{U} & W \\ W & V \end{pmatrix} \tag{E.10}$$

where $u = \mathbf{b}^T\widetilde{U}\mathbf{b}$ and $\widetilde{U}$ relate to p. We use block-wise inversion to calculate $\mathbf{A}_{11}^{-1}$. Putting $k = d - \mathbf{y}^T\mathbf{X}^{-1}\mathbf{y}$, we arrive at:

$$\mathbf{A}_{11}^{-1} = \begin{pmatrix} k^{-1} & -k^{-1}\mathbf{y}^T\mathbf{X}^{-1} \\ -\mathbf{X}^{-1}\mathbf{y}k^{-1} & \mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{y}k^{-1}\mathbf{y}^T\mathbf{X}^{-1} \end{pmatrix} \tag{E.11}$$

To obtain $\widetilde{\mathbf{A}}_{11}^{-1}$, we put $\mathbf{K} = \widetilde{\mathbf{D}} - \mathbf{1}y^T\mathbf{X}^{-1}y\mathbf{1}^T$ which yields:

$$\widetilde{\mathbf{A}}_{11}^{-1} = \begin{pmatrix} \mathbf{K}^{-1} & -\mathbf{K}^{-1}\mathbf{1}y^T\mathbf{X}^{-1} \\ -\mathbf{X}^{-1}y\mathbf{1}^T\mathbf{K}^{-1} & \mathbf{X}^{-1} + \mathbf{X}^{-1}y\mathbf{1}^T\mathbf{K}^{-1}\mathbf{1}y^T\mathbf{X}^{-1} \end{pmatrix} \tag{E.12}$$

As all DPQs in p have the same pedigree inbreeding and pedigree relationship to each other, all diagonal entries of $\widetilde{\mathbf{D}}$ have the same value, l_1, and all off-diagonal entries have the same value, l_2. The Sherman-Morrison formula shows that there are scalars, x and y, such that $\mathbf{K} = x\mathbf{I} + y\mathbf{1}\mathbf{1}^T$ and $\mathbf{K}^{-1} = x^{-1}\mathbf{I} - \frac{x^{-2}y\mathbf{1}\mathbf{1}^T}{1+x^{-1}8y}$ hold true. $k = \frac{1}{8}x + y$ is the average of $\mathbf{K}$, and the average of $\mathbf{K}^{-1}$ is $\frac{1}{8}x^{-1} - \frac{x^{-2}y}{1+x^{-1}8y} = \frac{1}{64}k^{-1}$. I. e., the sum of the entries in $\mathbf{K}$ is given by:

$$\mathbf{1}^T\mathbf{K}^{-1}\mathbf{1} = k^{-1} \tag{E.13}$$

Due to the simple structure of $\mathbf{K}$ we also have:

$$\mathbf{1}^T\mathbf{K}^{-1} = \frac{1}{8}k^{-1}\mathbf{1}^T \tag{E.14}$$

We consider $\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}$ in two parts. Vector $[\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}]_1$ is the column relating p to the non-genotyped animals.

$$[\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}]_1 = (x - \mathbf{Z}\mathbf{X}^{-1}y)k^{-1}u + (-xk^{-1}y^T\mathbf{X}^{-1} + \mathbf{Z}\mathbf{X}^{-1} \\ + \mathbf{Z}\mathbf{X}^{-1}yk^{-1}y^T\mathbf{X}^{-1})\mathbf{W}b \tag{E.15}$$

The first column of $\widetilde{\mathbf{A}}_{21}\widetilde{\mathbf{A}}_{11}^{-1}\widetilde{\mathbf{G}}\mathbf{B}^T$ is given by:

$$[\widetilde{\mathbf{A}}_{21}\widetilde{\mathbf{A}}_{11}^{-1}\widetilde{\mathbf{G}}\mathbf{B}^T]_1 \\ = (x - \mathbf{Z}\mathbf{X}^{-1}y)\mathbf{1}^T\mathbf{K}^{-1}\widetilde{\mathbf{U}}b + (-x\mathbf{1}^T\mathbf{K}^{-1}\mathbf{1}y^T\mathbf{X}^{-1} \\ + \mathbf{Z}\mathbf{X}^{-1} + \mathbf{Z}\mathbf{X}^{-1}y\mathbf{1}^T\mathbf{K}^{-1}\mathbf{1}y^T\mathbf{X}^{-1})\mathbf{W}b \tag{E.16}$$

The matrices in (E.15) and (E.16) are equal, because of (E.13) and the following which is implied by (E.14).

$$\mathbf{1}^T\mathbf{K}^{-1}\widetilde{\mathbf{U}}b = \frac{1}{8}k^{-1}\mathbf{1}^T\widetilde{\mathbf{U}}b = k^{-1}b^T\widetilde{\mathbf{U}}b = k^{-1}u \tag{E.17}$$

The second part of $\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}$ is matrix $[\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}]_2$ which relates the other genotyped animals to the non-genotyped animals.

$$[\mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{G}]_2 = (\mathbf{x} - \mathbf{ZX}^{-1}\mathbf{y})k^{-1}\mathbf{b}^T\mathbf{W} + (-\mathbf{x}k^{-1}\mathbf{y}^T\mathbf{X}^{-1} + \mathbf{ZX}^{-1}$$
$$+ \mathbf{ZX}^{-1}\mathbf{y}k^{-1}\mathbf{y}^T\mathbf{X}^{-1})\mathbf{V} \tag{E.18}$$

The corresponding submatrix of $\widetilde{\mathbf{A}}_{21}\widetilde{\mathbf{A}}_{11}^{-1}\widetilde{\mathbf{G}}\mathbf{B}^T$ is given by:

$$[\widetilde{\mathbf{A}}_{21}\widetilde{\mathbf{A}}_{11}^{-1}\widetilde{\mathbf{G}}\mathbf{B}^T]_2$$
$$= (\mathbf{x} - \mathbf{ZX}^{-1}\mathbf{y})\mathbf{1}^T\mathbf{K}^{-1}\mathbf{Wb} + (-\mathbf{x}\mathbf{1}^T\mathbf{K}^{-1}\mathbf{1}\mathbf{y}^T\mathbf{X}^{-1} \tag{E.19}$$
$$+ \mathbf{ZX}^{-1} + \mathbf{ZX}^{-1}\mathbf{y}\mathbf{1}^T\mathbf{K}^{-1}\mathbf{1}\mathbf{y}^T\mathbf{X}^{-1})\mathbf{V}$$

The matrices in (E.18) and (E.19) are equal, because of (E.13) and the following which is implied by (E.14).

$$\mathbf{1}^T\mathbf{K}^{-1}\mathbf{Wb} = k^{-1}\mathbf{b}^T\mathbf{W} \tag{E.20}$$

as $\mathbf{W}$ is symmetric.

QED

Appendix F: Accuracy for replacement queens

In this appendix, we derive our formula (3.16) for the accuracy of the replacement queens in year 8.

$$\rho_{pR} = \frac{\sigma_{pW}}{\sigma_{pQ}} \rho_{pW} \tag{3.16}$$

This is close to the accuracy of replacement queens developed by Brascamp and Bijma (2019). Formula (10) in (Brascamp and Bijma 2019) states

$$\rho_{pR} = \sqrt{A_{ii}} \rho_{pW} \tag{F.1}$$

where A_{ii} is the diagonal entry of the numerator relationship matrix for a particular worker group. An unnumbered formula before formula (10) in (Brascamp and Bijma 2019)

$$\widetilde{\sigma_{pW}} \sqrt{A_{ii}} = \sigma_{pW} \tag{F.2}$$

where $\widetilde{\sigma_{pW}}$ is the standard deviation of the true breeding value of a single worker from year 8, which equals

$$\widetilde{\sigma_{pW}} = \sigma_{pR} \tag{F.3}$$

where σ_{pR} is the standard deviation of the true breeding value of the corresponding replacement queen. Because these queens were not simulated, we approximate:

$$\sigma_{pR} = \sigma_{pQ} \tag{F.4}$$

This is reasonable, because the replacement queens are the unselected offspring of queens from year 8. In summary we have

$$\sqrt{A_{ii}} = \frac{\sigma_{pW}}{\sigma_{pQ}} \tag{F.5}$$

Appendix G: Comparison of simulated and estimated genetic gain (R_{PB}) from year 8 to year 9 in the simulated breeding population relying on PBLUP

Parameter setting abbreviated	ρ_{pW} with PBLUP	σ_{pW}	Estimated R_{PB} per year	Average TBV of worker groups in year 8	Average TBV of worker groups in year 9	Simulated R_{PB} per year	Difference between estimated and simulated R_{PB}
MOD	0.6501 (0.035)	0.7777 (0.0335)	0.3506 (0.0315)	2.5768 (0.1422)	2.9352 (0.1433)	0.3584 (0.1048)	-0.0078 (0.1103)
HGC	0.5376 (0.0448)	0.5679 (0.0244)	0.2118 (0.0240)	1.4705 (0.1134)	1.6901 (0.1152)	0.2196 (0.0871)	-0.0077 (0.0892)

Genetic gain is given in the units of the selection criterion. The simulated values show the difference of the average breeding values between worker groups of years 9 and worker groups of year 8. For the estimated values, we used formula (3.18), and divided the result by 2.5 to account for the generation interval. The accuracy for phenotyped worker groups with PBLUB, ρ_{pW}, and the standard deviation of the TBV, σ_{pW}, were taken from the worker groups in year 8. The difference between the estimated and simulated genetic gain was calculated for each replicate and the average and the standard deviation among all replicates is given. The difference is small on average and the standard deviation of the difference is slightly greater than the standard deviation of the simulated genetic gain. Accuracies of breeding values when all BQ from years 4-9 were genotyped

Appendix H: Accuracies of breeding values when all BQs from years 4 to 9 were genotyped

parameter setting	effect	method	year 9		year 8		years 4 to 7	
			queens	workers	queens	workers	queens	workers
MOD	maternal	PBLUP	0.291 (0.068)	0.386 (0.08)	0.551 (0.039)	0.519 (0.061)	0.572 (0.023)	0.577 (0.033)
		ssGBLUP$_{BQ}$	0.506 (0.044)	0.522 (0.057)	0.626 (0.038)	0.592 (0.056)	0.643 (0.024)	0.637 (0.031)
	direct	PBLUP	0.252 (0.06)	0.274 (0.078)	0.311 (0.044)	0.522 (0.056)	0.4 (0.036)	0.608 (0.034)
		ssGBLUP$_{BQ}$	0.357 (0.053)	0.354 (0.074)	0.38 (0.046)	0.553 (0.055)	0.447 (0.037)	0.631 (0.032)
	sum of dir. and mat. eff.	PBLUP	0.202 (0.053)	0.26 (0.061)	0.588 (0.028)	0.65 (0.035)	0.657 (0.018)	0.765 (0.018)
		ssGBLUP$_{BQ}$	0.518 (0.035)	0.486 (0.043)	0.677 (0.025)	0.699 (0.032)	0.716 (0.016)	0.793 (0.016)
HGC	maternal	PBLUP	0.314 (0.079)	0.394 (0.074)	0.491 (0.049)	0.511 (0.064)	0.502 (0.027)	0.536 (0.035)
		ssGBLUP$_{BQ}$	0.483 (0.051)	0.504 (0.056)	0.581 (0.043)	0.59 (0.052)	0.59 (0.026)	0.608 (0.031)
	direct	PBLUP	0.262 (0.073)	0.293 (0.082)	0.275 (0.062)	0.489 (0.054)	0.355 (0.042)	0.566 (0.042)
		ssGBLUP$_{BQ}$	0.337 (0.064)	0.352 (0.074)	0.344 (0.061)	0.526 (0.051)	0.399 (0.045)	0.591 (0.04)
	sum of dir. and mat. eff.	PBLUP	0.182 (0.064)	0.233 (0.073)	0.413 (0.032)	0.538 (0.045)	0.503 (0.024)	0.664 (0.028)
		ssGBLUP$_{BQ}$	0.398 (0.044)	0.386 (0.059)	0.509 (0.039)	0.589 (0.044)	0.546 (0.025)	0.685 (0.025)

Queens from year 9 were not phenotyped; queens from year 8 were phenotyped, but none of them were dams of queens.

Appendix I: Genetic gain, R_{GS}, in the initial selection cycle of different breeding schemes applying CBS and GPS

The table in the additional file "Appendix_I.xlsx" presents the configuration of the breeding schemes shown in Figure 3.4, as well as 9 reruns of the optimal breeding scheme with ssGBLUP$_{\text{BQ}}$, and 1765 other runs chosen at even spacing to represent the remaining schemes. Equations (3.18), (3.19) , and (3.20) were used to calculate R_{GS} in the scenarios described in Table 3.1 for different ratios of preselected BQ and DPQ, with the prediction accuracy adjusted to the proportion on preselected BQ in all scenarios. Genetic gain is given in the units of the selection criterion. The parameter setting MOD was used. The standard deviations of the true breeding values σ_{pW} (worker groups from year 8) and σ_{uQ} (queens from year 9) are shown in Appendix J. The 1765 remaining schemes were picked by the numbers of dams of DPQ chosen for GPS, (N_{DPQ}^{GPS}), dams of BQ chosen for GPS (N_{BQ}^{GPS}), candidate DPQ per dam for GPS (n_{DPQ}^{GPS}), and candidate BQ per dam for GPS (n_{BQ}^{GPS}) with the step widths 10, 20, 8, and 8, respectively.

Appendix J: Standard deviations of true breeding values

parameter setting	effect	year 9		year 8		years 4 to 7	
		queens	workers	queens	workers	queens	workers
MOD	maternal	0.977 (0.041)	0.613 (0.032)	0.963 (0.038)	0.602 (0.034)	0.987 (0.022)	0.619 (0.019)
	direct	1.346 (0.05)	0.831 (0.042)	1.342 (0.042)	0.833 (0.035)	1.402 (0.031)	0.901 (0.03)
	sum of dir. and mat. eff.	1.277 (0.044)	0.773 (0.032)	1.274 (0.039)	0.778 (0.033)	1.376 (0.034)	0.907 (0.031)
HGC	maternal	0.991 (0.042)	0.618 (0.029)	0.983 (0.041)	0.612 (0.03)	0.997 (0.026)	0.621 (0.02)
	direct	1.368 (0.059)	0.847 (0.039)	1.364 (0.053)	0.844 (0.038)	1.421 (0.037)	0.907 (0.034)
	sum of dir. and mat. eff.	0.926 (0.035)	0.567 (0.025)	0.925 (0.035)	0.568 (0.024)	0.985 (0.025)	0.642 (0.023)

Queens from year 9 were not phenotyped; queens from year 8 were phenotyped, but none of them were dams of queens.

Appendix K: Regression coefficients of true on estimated breeding values, b_1, when all queens from years 4-9 were genotyped

parameter setting	Effect	method	year 9		year 8		years 4 to 7	
			queens	workers	queens	workers	queens	workers
MOD	maternal	PBLUP	0.995 (0.213)	1.009 (0.208)	0.978 (0.073)	0.977 (0.119)	1.001 (0.043)	0.996 (0.054)
		ssGBLUP$_{BQ}$	0.994 (0.09)	1 (0.118)	0.98 (0.068)	0.976 (0.098)	1.012 (0.039)	1.005 (0.05)
	direct	PBLUP	1.009 (0.223)	0.985 (0.283)	0.988 (0.149)	0.998 (0.098)	0.995 (0.079)	0.996 (0.059)
		ssGBLUP$_{BQ}$	1.07 (0.152)	1.031 (0.221)	1.013 (0.131)	1.003 (0.092)	1.118 (0.076)	1.079 (0.056)
	sum of dir. and mat. eff.	PBLUP	1.019 (0.257)	0.988 (0.252)	1.009 (0.053)	1.006 (0.069)	1.007 (0.037)	1.005 (0.041)
		ssGBLUP$_{BQ}$	1.106 (0.089)	1.087 (0.118)	1.04 (0.05)	1.027 (0.064)	1.136 (0.033)	1.143 (0.04)
HGC	maternal	PBLUP	0.998 (0.24)	0.997 (0.179)	0.983 (0.093)	0.983 (0.121)	1.004 (0.052)	1.006 (0.057)
		ssGBLUP$_{BQ}$	0.982 (0.098)	0.976 (0.117)	0.984 (0.08)	0.975 (0.086)	0.99 (0.046)	0.979 (0.044)
	Direct	PBLUP	1.008 (0.268)	1.014 (0.275)	1.014 (0.209)	1.007 (0.114)	1.004 (0.082)	1.007 (0.066)
		ssGBLUP$_{BQ}$	1.047 (0.184)	1.037 (0.212)	1.022 (0.16)	1.003 (0.092)	1.069 (0.09)	1.078 (0.062)
	sum of dir. and mat. eff.	PBLUP	1.021 (0.334)	1.016 (0.32)	1.002 (0.091)	1 (0.086)	0.996 (0.048)	0.996 (0.053)
		ssGBLUP$_{BQ}$	1.109 (0.128)	1.101 (0.18)	1.053 (0.094)	1.034 (0.079)	1.151 (0.053)	1.185 (0.057)

Queens from year 9 were not phenotyped; queens from year 8 were phenotyped, but none of them were dams of queens.

Appendix L: Genetic in the first generation of a genomic breeding program applying a shorter generation interval

In this appendix, we estimate genetic gain in for a breeding scheme with a shorter generation interval. We assume a parent generation of 1000 BQs per year in the population, with unphenotyped parents. Furthermore, a budget for 1000 genotyped queens per year is assumed. Furthermore, we assume the lag of phenotyping reduces the accuracy to $\frac{3}{4}$ of the calculated values. This is based on results from (Habier et al. 2007), where the predicted accuracy due to LD was estimated. An additional study would be required to determine this value in honey bees.

The 1000 parent BQs are genotyped as they hatch. The top 5% and 20% are selected as dam of DPQs ($i_{DPQ}^{CBS} = 2.06$) and as dams of BQs ($i_{BQ}^{CBS} = 1.40$), respectively, as we assumed for CBS. DPQs are reared in the same year as the BQs hatch. A year later, the daughter BQs are reared. To calculate the genetic gain with a shorter generation interval, R^{SGI}, we modify Equation (3.19) to selection of unfertilized queens.

$$R^{SGI} = \left(\frac{i_{DPQ}^{CBS} + i_{BQ}^{CBS}}{2}\right) \left(\frac{3}{4}\rho_{uQ}\right) \sigma_{uQ} \qquad \text{(L.1)}$$

where $\sigma_{uQ} = 1.28$ is the standard deviation of TBV as for Table 3.4; and ρ_{uQ} is the accuracy for unfertilized queens. We consider the trait MOD. We use $\rho_{uQ} = 0.52$, which is the accuracy of ssGBLUP$_{BQ}$ with 5000 BQs in the reference population (Figure 3.3). This yields $R^{SGI} = 0.86$. The gain in the optimal scheme of Table 3.4 was $R_{GS} = 1.09$ for 1000 genotyped queens per year. The generation interval for R^{SGI} is 1 year, while the generation interval for R_{GS} is 2.5 years. The expected yearly gain with R_{GS} is 0.44. The expected gain with a shortened generation interval is therefore 95% higher compared to the optimal scheme for 1000 genotyped queens per year in Table 3.4.

Statement of authorship

I hereby declare that I completed the doctoral thesis independently based on the stated resources and aids. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Selbstständigkeitserklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsbearbeiterinnen/ Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

_____   _____

Datum   Unterschrift