
# Integrating high-density marker information into the genetic evaluation of the honey bee

Dissertation

Submitted in partial fulfilment of the requirements for the

Doctor of Natural Sciences (Dr. rer. nat.)

to the Department of Mathematics and Informatics

of the Freie Universität Berlin

Pooja Gupta

Berlin, 2012

Reviewers: Prof. Dr. Christof Schütte

Prof. Dr. Kaspar Bienefeld

Prof. Dr. Norbert Reinsch

Date of defence: 28.02.2013

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, den 25.10.2012

Pooja Gupta

# Bibliographic information

Parts of this work have either been published or are a part of submitted manuscript.

Chapter 2:

- Gupta P, Conrad T, Spötter A, Reinsch N, Bienefeld K (2012). Simulating a base population in the honey bee for molecular genetic studies. *Genetics Selection Evolution* **44**: 14.

  DOI: 10.1186/1297-9686-44-14.

- Gupta P, Reinsch N, Spötter A, Conrad T, Bienefeld K (2012). Accuracy of the unified approach in composite traits - illustrated by a simulation study in the honey bee (*Apis mellifera*). Submitted.

Chapter 3:

- Gupta P, Reinsch N, Spötter A, Conrad T, Bienefeld K (2012). Accuracy of the unified approach in composite traits - illustrated by a simulation study in the honey bee (*Apis mellifera*). Submitted.

Chapter 4:

- Spötter A, Gupta P, Nürnberg G, Reinsch N, Bienefeld K (2012). Development of a 44K SNP assay focussing on the analysis of a varroa-specific defence behaviour in honey bees (*Apis mellifera carnica*). *Molecular Ecology Resources* **12**: 323–332.

  The definitive version is available at www.blackwell-synergy.com

  DOI: 10.1111/j.1755-0998.2011.03106.x.

Chapter 5:


- Gupta P, Conrad T, Spötter A, Reinsch N, Bienefeld K (2012). Simulating a base population in the honey bee for molecular genetic studies. *Genetics Selection Evolution* **44**: 14.

  DOI: 10.1186/1297-9686-44-14.

- Gupta P, Reinsch N, Spötter A, Conrad T, Bienefeld K (2012). Accuracy of the unified approach in composite traits - illustrated by a simulation study in the honey bee (*Apis mellifera*). Submitted.

- Spötter A, Gupta P, Nürnberg G, Reinsch N, Bienefeld K (2012). Development of a 44K SNP assay focussing on the analysis of a varroa-specific defence behaviour in honey bees (*Apis mellifera carnica*). *Molecular Ecology Resources* **12**: 323–332.

  The definitive version is available at www.blackwell-synergy.com

  DOI: 10.1111/j.1755-0998.2011.03106.x.

# Acknowledgement

I would like to express my gratitude to Prof. Dr. Kaspar Bienefeld for providing me the opportunity to work on this thesis. Without his scientific guidance, it would not have been possible for me to complete this project. I would also like to thank him for introducing me to his work group and for helping me to establish collaborations during my Ph.D. His unrelenting encouragement and support was always a source of motivation for me.

I would like to thank Dr. Tim Conrad for giving me the opportunity to work in his group and providing me with the great infrastructural support. I am obliged to him for his constant supervision and support during my Ph.D. I am also extremely grateful to him for the financial support during the final months of my Ph.D.

Furthermore, I would sincerely like to thank Prof. Dr. Norbert Reinsch for guiding me through this project. His knowledge in the field of genetics is awe-inspiring and I learned a lot from him. I would like to thank him for inviting me at his institute for scientific discussions that helped me to gain a better understanding of this field.

I would also like to thank Dr. Andreas Spötter for helping me so much during my initial phase in Germany. It was an enriching experience to learn from him about the scientific research in the area of molecular genetics.

I would like to thank Prof. Dr. Christof Schütte for providing me the opportunity to be a part of the Biocomputing group. This allowed me to work in a great environment, meet a lot of nice people and gain many friends.

Last but not the least, I would like to thank my mother, father and brother for believing so much in me. This journey would not have been possible without their support. A special thank to Maxi who always stood by me. Additionally, I would like to thank all my friends for their encouragement.

# Abstract

Over the past years, a decline of the most commonly domesticated European honey bee (*Apis mellifera*) populations has been reported, mainly caused by infestation with an ecto-parasitic mite *Varroa destructor*. Selective breeding of genetically superior bees can help to establish resistant lines which will prevent losses due to the parasite. It will also improve several other economically important quantitative traits such as honey yield, swarming tendency and calmness. However, this requires a robust breeding program and the implementation of genetic evaluation to predict the 'breeding values' for selecting genetically superior individuals based on information on phenotype, pedigree and genotype.

This thesis describes a method to integrate high-density single nucleotide polymorphism (SNP) data for genetic evaluation in the honey bee using the 'unified approach'. In order to assess the potential of this approach and its applicability to the honey bee population, a simulation study was conducted. A framework for simulating a honey bee population was developed by modelling the reproductive and genetic biology of the honey bee such as high genomic recombination rates, haplo-diploid sex determination, polyandry, uncertain paternity and negative correlation between maternal and direct effects. This provided genomic, pedigree and phenotypic datasets required for implementing the unified approach. The linear mixed model equations were solved to obtain the 'best linear unbiased predictions' of breeding values based on the unified approach. The influence of maternal effects, negative correlation between maternal and direct effects, uncertain paternity and different magnitudes of maternal and direct heritabilities were also addressed, thus making this study of interest for research in other livestock species as well. In addition, a 44k SNP assay was designed for the purpose of genome-wide association studies and marker based selection strategies.

This is the first study that gives background knowledge about the simulation and modelling of genomic and pedigree datasets in honey bees for genetic evaluation, thus, providing an important framework for future studies. The unified approach is a progressive step for the genetic evaluation in honey bees. It is expected that the study will give directions to further research in

the honey bee as well as other species concerning genetic evaluation based on high-density molecular marker data.

# Contents

## List of Figures

# List of Tables

# Introduction

The honey bee (*Apis* sp.) is an important species that serves as a major pollinator of wild plants and agricultural crops. It contributes significantly to the agricultural economy, as a large fraction (Roubik, 1995) of the agricultural crops is being pollinated by honey bees. Furthermore, the honey bee exhibits high degree of 'eusociality', and owing to this characteristic it has been recognized as a model organism for understanding and studying the dynamics of social interaction (Oldroyd and Thompson, 2007). An important milestone for research in fields associated with the honey bee was the sequencing of the genome of *Apis mellifera* (The Honeybee Genome Sequencing Consortium, 2006). The unique properties associated with its genome (The Honeybee Genome Sequencing Consortium, 2006) such as high A+T content, high CpG content, fewer genes for innate immunity, absence of transposons, slower rate of evolution, similarity to vertebrates for circadian rhythm, RNA interference and DNA methylation genes show that it is a promising model organism for elucidating key biological processes. For example, honey bees possess a rather simple nerve system but still show repertoire of cognitive behaviour, thus, can serves as a model organism for investigating the fundamentals of learning behaviour, memory consolidation, circadian rhythms etc. It can help to examine how the insulin/insulin-like growth factor signalling pathway could have been modified to extend lifespan without negatively affecting reproductive capabilities (The Honeybee Genome Sequencing Consortium, 2006). Apart from the area of biological sciences, the swarming behaviour of honey bees has also inspired researchers in the field artificial intelligence (Karaboga, 2005; Karaboga and Akay, 2009).

Over the past years, a decline of the most commonly domesticated European honey bee (*Apis mellifera*) population has been reported (Brown and Paxton, 2009; De la Rúa *et al*., 2009; Neumann and Carreck, 2010), mainly caused by infestation with an ecto-parasitic mite (*Varroa destructor*). It has been observed that honey bees show resistance to *Varroa* mite (Boecking and Drescher, 1992; Spivak, 1996) through the manifestation of 'hygienic behaviour'. Hygienic behaviour can be defined as the ability of worker bees to detect and remove infected broods before the pathogen reaches the stage of infection (Spivak and Gilliam, 1998). Several studies suggest that the hygienic behaviour is a heritable trait (Harbo and Harris, 1999; Boecking *et al*.,

2000; Lapidge *et al*., 2002). Thus, through selective breeding, it would be possible to exploit the heritability to establish resistant lines in the honey bee as well as to improve several important quantitative traits such as honey yield, swarming tendency and calmness. Selective breeding requires a robust breeding program and the implementation of genetic evaluation. Until now, a method based on the pedigree and phenotypic information has been used for genetic evaluation in the honey bee (Bienefeld *et al*., 2007). In other species, this approach has been mostly succeeded by marker based methodologies that employ high-density marker information across the genome such as the single nucleotide polymorphism[1] (SNP) for genetic evaluation. Marker based selection has been widely tested in several species either with simulated datasets (Sonesson and Meuwissen, 2009; Christensen and Lund, 2010) or with real datasets (Dekkers and Hospital, 2002; de Roos *et al*., 2007; Legarra *et al*., 2008; Aguilar *et al*., 2010) but, to date, not in honey bees. Therefore, one of the main objectives of this study was to integrate SNP information into the genetic evaluation of the honey bee.

Over the last decade, a high-density marker based methodology known as 'genomic selection' has replaced most other genetic evaluation methodologies in the livestock species. It is now established as a 'state of the art' method for genetic evaluation that has resulted in a significant advancement in the field of animal breeding. In order to implement genetic evaluation based on the genomic selection strategy, a multi-step procedure was proposed for many livestock species, for example, in the US dairy cattle (VanRaden, 2009) and pigs (Ostersen *et al*., 2011). However, this multi-step procedure has certain disadvantages with respect to the honey bee. In comparison to cattle, it is complicated to define the daughter yield deviation[2] in the honey bee because of its complex population structure as well as due to the influence of maternal effects on traits. In addition, due to economical and technical constraints, it may not be possible to genotype all animals in the population. Thus, instead of a multi-step procedure, this thesis describes the implementation of a single-step 'Unified Approach' for the integration of molecular marker information (i.e. SNP) for genetic evaluation in the honey bee. The unified approach, first proposed by Legarra *et al*. (2009) and Christensen and Lund (2010), combines full pedigree and

[1] A single nucleotide polymorphism is a difference in the DNA sequence among individuals resulting from the variation of a single nucleotide i.e. A, G, C or T.

[2] Daughter yield deviation is defined as a weighted average of the yield deviation of all progeny of a sire corrected for fixed effects and the breeding values of the mates of the sire (Mrode, 2005).

genomic information from both genotyped and ungenotyped individuals. The advantage of this procedure over the multi-step approach is that it gives a more accurate estimate of the breeding values for ungenotyped animals (Aguilar *et al.*, 2010; Christensen and Lund, 2010) and is resistant to selection bias (Vitezica *et al.*, 2011). Moreover, it is simpler to implement as compared to the multi-step approach and provides an easy extension to a multi-trait model with maternal effects in honey bees.

This thesis illustrates the implementation of the unified approach in honey bees using simulations. A dataset was generated by modelling and simulating the reproductive and genetic biology characteristics of the honey bee. Comparative analyses of the unified and the traditional pedigree based genetic evaluation approaches was performed on the simulated population. Additionally, since future studies require a real genotyping dataset for implementing the unified approach, a 44k SNP assay was developed during the study. To the best of the knowledge, this is the first study that assesses the impact of both negative correlation between maternal and direct effects and uncertain paternity on marker based genetic evaluation. Thus, the study is of great interest for research concerning genetic evaluation in other species as well.

## Objectives

The primary objective of this thesis was to integrate high-density marker information (i.e. SNP) into the genetic evaluation of the honey bee. This integration of marker information into genetic evaluation of the honey bee was achieved by completing the following milestones in the study:

1. Developing a theoretical framework for modelling the reproductive and genetic biology of the honey bee.

2. Developing a software program for simulating a honey bee population and for analyzing the generated genomic, pedigree and phenotypic datasets.

3. Implementing the unified approach of genetic evaluation in the honey bee to investigate (a) the accuracy of the estimated breeding values and (b) the applicability of this approach.

4. Designing a 44k SNP assay (Spötter *et al*., 2012) that will be used for the purpose of genome-wide association studies and marker based selection strategies.

**Survey of contents**

Chapter 1: To enable a comprehensive understanding of the thesis, Chapter 1 introduces the essential aspects of the honey bee's biology and explains the principles integral to genetic evaluation. It summarizes honey bee's taxonomic classification, reproductive biology, genome characteristics and addresses a key issue related to threat due to *Varroa* infestation. Furthermore, it explains the fundamental concepts associated with genetic evaluation such as quantitative traits, breeding values and the methodology of genetic evaluation.

Chapter 2: This chapter describes the generation of a dataset for the honey bee population. It deals with the important aspects of population modelling and simulation. In addition, it also describes the simulation of true breeding values, phenotypic values, a negative correlation between maternal and direct effects, heritability, genetic and residual variances. Ideas that could be exploited in future to derive genotyping information for ungenotyped animals are presented at the end of the chapter.

Chapter 3: This chapter gives a chronological review of the progress in genetic evaluation in honey bees along with an introduction to the best linear unbiased prediction methodology, mixed model equations and maternal effects. It presents the implementation of the following genetic evaluation methods: (1) traditional approach using pedigree and phenotypic data (PED_BLUP) (2) the advanced unified approach using marker, pedigree and phenotypic data (UNI_BLUP). In addition, ideas for improving the method of construction of the numerator relationship matrix in the honey bee are discussed.

Chapter 4: This chapter outlines the procedure of development and a preliminary test of a 44k SNP assay for the honey bee required for future association studies and genetic evaluation.

Chapter 5: This chapter presents the results and discussion of this thesis. It describes the validation results for the base population simulation software program. Furthermore, for the comparative analyses of the unified and pedigree based approaches, results are reported regarding the accuracy of the overall, direct and maternal estimated breeding values under

varying heritability of trait and correlation between maternal and direct effects. The discussion section describes the effect of the unified approach on the accuracy of estimating breeding values with regard to the influence of heritability of the trait and the genetic correlation between maternal and direct effects. Towards the end of this chapter, the results and discussion about the 44k SNP assay are presented.

Chapter 6: This chapter gives the conclusion and presents the possible future work which can be undertaken in the direction of marker based genetic evaluation.

# Chapter 1. Background

To enable a comprehensive understanding the thesis, this chapter introduces the essential aspects of the honey bee's biology and explains the principles integral to genetic evaluation. This chapter has been divided into two sections. The first section summarizes honey bee's taxonomic classification, reproductive biology, genome characteristics and addresses a key issue related to threat due to *Varroa* infestation. The second section introduces fundamental concepts associated with genetic evaluation such as quantitative traits, breeding values and the methodology of genetic evaluation.

## 1.1. Overview of the honey bee biology

### 1.1.1. Taxonomy

The honey bee (*Apis* sp.) belongs to a large family of bees, Apidae, which are characterized by the presence of a pollen basket. Taxonomically, this family is classified under the phylum Arthropoda, class Insecta and order Hymenoptera. The genus *Apis* evolved in tropical Eurasia (Ruttner, 1988; The Honeybee Genome Sequencing Consortium, 2006), migrating to northern and western regions, eventually reaching Europe by the end of the Pleistocene epoch, 10,000 years ago (The Honeybee Genome Sequencing Consortium, 2006). Several species and sub-species of the honey bee, distributed in different regions of the world (Figure 1.1), have been recognized. One of the most commonly domesticated species is the European honey bee, *Apis mellifera*. On the basis of morphometric, mtDNA and microsatellite studies, the subspecies within *Apis mellifera* have been grouped into three main evolutionary branches (Franck *et al*., 1998), the African subspecies (branch A), northern Mediterranean subspecies (branch C) and western European subspecies (branch M).

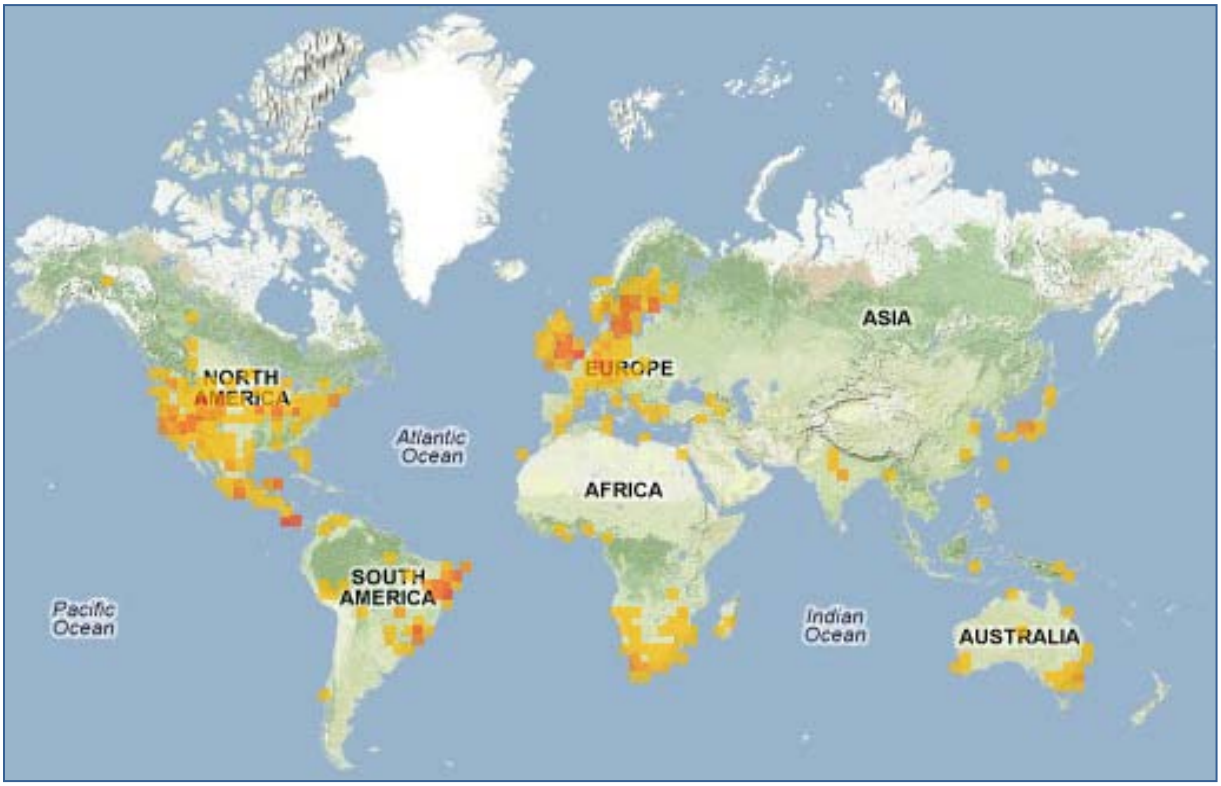


Figure 1.1. Global distribution of the *Apis* sp.

(Source: Encyclopedia of life http://eol.org/pages/104135/maps)

**1.1.2. Reproductive biology**

The honey bee exhibits a high degree of social organization. It has a complex colony structure consisting of individuals belonging to three different castes i.e. queen, workers and drones, each specialized to perform distinctive functions. An important feature of the honey bee is its unique reproductive biology. Sex determination in the honey bee is neither determined by the presence or absence of a sex chromosome nor the ploidy of the organism. Interestingly, a sex determination locus harbouring the *complementary sex determiner gene* (*csd* gene) (Beye *et al*., 2003) is responsible for the sex in honey bees. The *csd* gene encodes a potential splicing factor that exists in at least 15 allelic variants (Hasselmann and Beye, 2004). Thus, honey bees possess a 'Complementary sex determination' system; individuals which are hemizygous or homozygous for this gene develop into male drones whereas heterozygous individuals develop into female queens and workers. The fertilized homozygous males are sterile and are eaten by workers shortly after emerging (Woyke, 1963). The caste differentiation between a queen and a worker results from the selective feeding of the royal jelly protein. This selective feeding leads to

differential DNA methylation which causes different reproductive and behavioural statuses in workers and queens (Kucharski *et al*., 2008). Thus, the genes encoding the major royal proteins play a key role in the establishment of this highly evolved social structure (The Honeybee Genome Sequencing Consortium, 2006).

In addition to the characteristics described above, the honey bee also exhibits polyandry[3]. A queen can be inseminated by approximately 10-15 drones (Trjasko, 1951; Woyke, 1960; Kerr *et al*., 1962, Adams *et al*., 1977) which varies among different populations and species. A virgin queen begins mating about one week after exiting the colony and getting oriented outside the colony (Ruttner, 1956). A queen can have up to four mating flights (Roberts, 1944) on average, during which it flies to a drone congregation area and mates with drones. After mating flights, queens begin oviposition (laying of eggs) and never mate again during the rest of their life span (Tarpy *et al*., 2000). A drone consists of a haploid set of chromosomes. Since there is no reductional division during the production of sperms, all sperm cells produced by a drone are genetically identical and are clones of the drone. When several drones inseminate a queen, millions of copies of each drone in the form of sperms are stored in the spermathecae of the queen which can later fertilize eggs produced by the queen.

### 1.1.3. Genome

One important landmark for research in honey bees was the sequencing of the genome of *Apis mellifera* (The Honeybee Genome Sequencing Consortium, 2006). The honey bee genome consists of 16 linkage groups with an approximate length of 236 Mb (The Honeybee Genome Sequencing Consortium, 2006). The SNP dataset (approximately 1 million) is also published that can be used to map important genes (The Honeybee Genome Sequencing Consortium, 2006). As compared to other species, the honey bee genome has distinctive characteristics such as high A+T and CpG contents and the lack of major transposon families (The Honeybee Genome Sequencing Consortium, 2006), making it an interesting candidate for further research. Moreover, the honey bee exhibits an extremely high recombination rate of 19 cM/Mb (Beye *et al*., 2006; The Honeybee Genome Sequencing Consortium, 2006) which is several-folds higher than that reported for any other higher eukaryotic species. For detailed information about the

[3] Polyandry, commonly referred to as multiple mating, is a phenomenon observed in honey bee whereby a queen mates with multiple drones (average of 10 to 20 drones).

genome of the honey bee please refer to the publication by The Honeybee Genome Sequencing Consortium (2006) and the genome databases, Honey bee genome project (www.hgsc.bcm.tmc.edu/projects/honeybee/) and National center for biotechnology information for *Apis mellifera* (www.ncbi.nlm.nih.gov/genome?term=apis%20mellifera/).

### 1.1.4. *Varroa* infestation

Considering its unique biological characteristics, honey bees make an interesting candidate as a model organism for future studies. Apart from it contribution to honey and wax production, it is one of the most important pollinator of wild plants and agricultural crops. However, to be able to exploit this species, its maintenance and conservation is essential. *Apis mellifera* faces threats to survival due to several diseases and parasites. *Varroa* has emerged as one of the biggest threats to the honey bee (Figure 1.2.a). It serves as a vector for several viral infections (Allen and Ball, 1996; Nordström, 1999) and has been associated with the occurrence of acute bee paralysis (Batuev, 1979; Ball, 1985; Allen *et al*., 1986; Ball and Allen, 1988; Bakonyi *et al*., 2002), slow paralysis virus (Ball, 1989; Santillán-Galicia *et al*., 2010), deformed winged virus (Ball, 1989; Bowen-Walker *et al*., 1999; Santillán-Galicia *et al*., 2010) and Kashmir bee virus (Chen *et al*., 2004). It has been observed that heavily infested colonies have a reduced life span and a low number of workers (Figure 1.2.b). This leads to poor hygienic behaviour, giving rise to bacterial and viral infection that eventually results into the terminal collapse of colonies.

Figure 1.2. a. A brood infested with *Varroa* b. Damage caused by infection due to *Varroa*.



The use of pesticides is not the most optimal solution as they pose the hazard of contaminating the surrounding environment and the commercial products obtained from the honey bee. In

addition, the effectiveness gets diminished eventually due to the development of resistant strains (Milani, 1999). One of the best solutions for the management of honey bees is selective breeding to create resistant lines. Several researchers (Bienefeld *et al*., 1999, 2008; Büchler *et al*., 2010; Rinderer *et al*., 2010) are working to breed bees showing the 'hygienic behaviour' (Figure 1.3). Furthermore, selective breeding can be exploited to improve other economically important traits in the honey bee such as honey production, swarming, aggressiveness and calmness.

Figure 1.3. A worker bee exhibiting the hygienic behaviour.



## 1.2. Fundamental concepts related to genetic evaluation

### 1.2.1. Quantitative trait loci

Usually a phenotype is assumed to be determined by the genotype at a single gene/locus. It is important to understand that a majority of the phenotypic variation is a result of the interaction of multiple genes and the environmental conditions. This class of phenotypes is usually referred to as 'Quantitative traits' or 'Multi-factorial traits' (Hamilton, 2009). In quantitative genetics, the terms phenotype, trait and character are all considered synonymous (Hamilton, 2009). Over the years, genetic improvement of quantitative traits in important plant and animal species has been achieved through artificial selection. This has contributed greatly to the increase in quality of the phenotype which was under selection (Dekkers and Hospital, 2002).

The variation in a quantitative trait can be attributed to several loci/regions across the genome. Two theories were proposed to explain the genetic variance associated with a quantitative trait:

(1) the infinitesimal model and (2) the finite site model. According to the infinitesimal model, proposed by Fisher (1918), very many independently segregating loci additively affect the trait and it is assumed that each locus has an infinitesimal effect on the trait (Weller, 2001). This model forms the background for the estimation of breeding values. The finite site model, on the other hand, assumes that there is a finite number of loci in the genome as the genome itself consist of a finite number of genes. It has been shown that a large number of the quantitative trait loci (QTL) have an extremely small effect and very few have a large effect on the phenotype (Shrimpton and Robertson, 1988; Hayes and Goddard, 2001). In this study, the finite site model was exploited to simulate QTL effects.

**1.2.2. Breeding values**

This section explains the concept of breeding values. More details can be found in Falconer and Mackay (1996), Mrode (2005) and Hamilton (2009). In animal breeding programs, usually a planned mating of selected individual is performed with the aim to genetically improve important traits. The selection of candidate is made through 'genetic evaluation' which predicts the 'breeding values' of individuals based on information about phenotypes, pedigree or genotypes.

For a quantitative trait, a phenotypic observation of an individual is determined by genetic factors, environmental factors and interaction between genetic and environmental factors. A phenotypic observation can be expressed in the following form (Falconer and Mackay, 1996; Mrode, 2005):

$$\textit{Phenotypic observation} = \textit{Environmental effects} + \textit{Genetic effects} + \textit{Residual effects}$$

or $$y_{ij} = \mu_i + g_i + e_{ij} \qquad 1.1$$

where $y_{ij}$ is the $j^{th}$ phenotypic observation of the $i^{th}$ individual, $\mu_i$ is the identifiable fixed environmental effects on the phenotype such as apiary or bee breeder on the $i^{th}$ individual, $g_i$ is a sum of additive, epistatic and dominance genetic effects of the $i^{th}$ individual and $e_{ij}$ is the

unidentifiable random environmental effects on the $j^{th}$ phenotypic observation of the $i^{th}$ individual.

The epistatic effect is an inter-locus interaction where one gene locus affects the expression of another gene locus. The dominance effect is an intra-locus interaction where one allele affects the expression of the other allele at the same gene locus. The additive genetic effect results from the average effect (see Chapter 2) of allele substitution (Falconer and Mackay, 1996; Wu *et al*., 2007). The average effect is the only component that can be selected, as it is a function of the genes which an individual inherits from its parent (Mrode, 2005). The breeding value of an individual is equal to the sum of average effects of the alleles it carries i.e. the summation over the pair of alleles at each locus and over all loci (Falconer and Mackay, 1996). In other words, the average additive genetic effect of genes an individual inherits from its parents represents the breeding value of an individual. The epistatic and dominance effects which cannot be selected are ignored and assumed to be a part of the random environmental effects, thus the previous equation can now be represented as (Mrode, 2005):

$$y_{ij} = \mu_i + g_i^a + e_{ij}^* \qquad 1.2$$

where $e_{ij}^*$ represents the sum of random environmental effects, epistatic effects and dominance effects and $g_i^a$ represents the sum of additive genetic effects of the $i^{th}$ individual.

Since each parent contributes only one-half of its genes to the progeny, the breeding value of an individual ($a_i$) can also be defined as follows:

$$a_i = g_i^a = \frac{1}{2}a_s + \frac{1}{2}a_d + m_i \qquad 1.3$$

where $a_s$ is the additive genetic effect/breeding value of the sire, $a_d$ is the additive genetic effect/breeding value of the dam and $m_i$ is Mendelian sampling effect of the $i^{th}$ individual. For an individual, each parent explains only one-quarter of the total additive genetic variance. In other words, the variance of additive genetic effects of both sire and dam explain only one-half

of the progeny's additive genetic variance. Therefore, for individuals with the same parents, the other half of the additive genetic variance cannot be explained by the additive genetic effects of the sire and dam. This effect is termed as the Mendelian sampling effect (Quaas and Pollock, 1980; Weller, 2001). It is the specific genetic component passed to an individual that differentiates this individual from his full sibs (Weller, 2001).

### 1.2.3. Methodology of genetic evaluation

Breeding values serve as a measure on the basis of which selection is implemented. Thus, accurate estimation of breeding values is crucial for genetic improvement. Usually, the method of estimating breeding values is based on the available information about pedigree, phenotypic and/or genotyping data. Traditionally, selection of quantitative traits has been based on phenotypic and pedigree data. The advancement in high-throughput genotyping technology permits the use of molecular genetic markers, such as SNP, to be used in breeding programs for genetic improvement. Another important requirement for the estimation of breeding values is a statistical model that describes the data. Likewise, different methods, such as Selection Index, Best Linear Unbiased Prediction (BLUP) and Bayesian approaches, have been used to estimate breeding values. BLUP is the most commonly used methodology for the estimation of breeding values in almost all livestock species. BLUP, developed by Henderson (1975, 1988), allows to estimate both fixed and random effects simultaneously (see Chapter 3). It is also used for genetic evaluation in the honey bee (Bienefeld *et al*., 2007). Therefore, in order to perform genetic evaluation one requires data on phenotype, pedigree or genotype, a statistical model and a solving procedure. A brief summary is given in the following sections.

#### *1.2.3.1. Data*

**Pedigree record**

Genetic evaluation of a population is impossible without pedigree information. Pedigree records also help to standardize the breeding program. Pedigree recording requires unique identifiers for all animals, their sires and their dams. Additionally, information about the origin, breed, birth-date, number of relatives, genotyping information can also be recorded.

**Phenotypic record**

Complete and accurate recording of phenotyping data for economically important traits is necessary. Recording should be unbiased and measured objectively for all animals within a production unit (apiary, herd, ranch or flock). Phenotyping data for an animal should also include information about the date of recording and possible factors that could influence the animal's performance.

**Genotyping record**

With the rapid advancement in genotyping technology, it has become possible to easily genotype individuals and utilize this information about QTL and/or markers for genetic evaluation. Data for genotyping should include information about the allelic variations and the associated allele frequency in a population.

Apart from the pedigree, phenotyping and genotyping information described above, it is extremely important to understand the reproductive peculiarities of a species in order to design an optimum mating and selection strategy. For example, it important to have information about the following aspects (Schaeffer, 2010): the gestation length, the age at first breeding, number of offspring per female per gestation and method employed for fertilization, e.g. artificial insemination or island mating in the honey bee. All this information can be used to formulate appropriate linear models for the analysis of the data and the accurate estimation of breeding values of animals (Schaeffer, 2010).

***1.2.3.2. Linear Model***

An important requirement for the estimation of breeding values is a 'linear model'. The model assumes that different factors affect a trait in a linear fashion, i.e. each factor has an additive and independent effect on the trait (Schaeffer, 2010). These models are an approximation of how different factors affect a trait (Schaeffer, 2010). The choice of a model depends upon the animal that is to be evaluated and its suitability to describe the data. The most suitable model is the one that is able to account for the most part of the variation. These models rely strongly on statistical knowledge and require computing expertise. Mrode (2005) provided a good description of different models used for estimating breeding values depending on the type of information available on the selection candidate. A summary of most of the models is given in Table 1.1. For details please refer to Mrode (2005).

Table 1.1. Summary of different models.

| | Model | Description | Matrix notation |
|---|---|---|---|
| Uni-variate model (Single random effect) | Animal model | Breeding values are estimated for all animals in the pedigree. | $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$ |
| | Sire model | Breeding values are estimated only for sires | $\mathbf{y} = \mathbf{Xb} + \mathbf{Zs} + \mathbf{e}$ |
| | Reduced animal model | Breeding values are estimated only for parents. Breeding values for non-parents are expressed as the average of parental breeding values plus Mendelian sampling. Solutions for non-parents are obtained using the solutions for the fixed effects and parents. | Model for parents: $\mathbf{y_p} = \mathbf{X_p b} + \mathbf{Z a_p} + \mathbf{e}$ Model for non-parents: $\mathbf{y_n} = \mathbf{X_n b} + \mathbf{Z_1 a_p} + \mathbf{e}^*$ or $\begin{bmatrix} \mathbf{y_p} \\ \mathbf{y_n} \end{bmatrix} = \begin{bmatrix} \mathbf{X_p} \\ \mathbf{X_n} \end{bmatrix} \mathbf{b} + \begin{bmatrix} \mathbf{Z} \\ \mathbf{Z_1} \end{bmatrix} \mathbf{a_p} + \begin{bmatrix} \mathbf{e} \\ \mathbf{e}^* \end{bmatrix}$ |
| | Animal model with groups | This model accounts for the subpopulation structure through proper grouping of the base animals. | $\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Za} + \mathbf{e}$ |
| Models with random environmental effect (This model includes an additional random environmental effect) | Repeatability model | Applicable when multiple measurements on the same trait are recorded for an individual. Accounts for the permanent environmental effects. | $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wpe} + \mathbf{e}$ |
| | Model with common environmental effects | Common environmental effects are included in the model. It accounts for the additional covariance between members of a family and the increased variance between families. | $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wc} + \mathbf{e}$ |

| | | | |
|---|---|---|---|
| Multi-variate animal model (Multiple random effects) | Model with multiple traits | Suitable for multiple trait evaluation. | An example model with two traits:<br>$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \end{bmatrix} + \begin{bmatrix} \mathbf{Z_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} \end{bmatrix} \begin{bmatrix} \mathbf{a_1} \\ \mathbf{a_2} \end{bmatrix} + \begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \end{bmatrix}$ |
| Maternal-trait model (For maternally influenced traits such as weaning weight in cattle) | Animal model for a maternal trait | The maternal and direct breeding values are estimated for all animals in the pedigree. | $\mathbf{y = Xb + Zu + Wm + Spe + e}$ |
| | Reduced maternal model with maternal effect | The maternal and direct breeding values are estimated only for parents. Breeding values for non-parents are expressed as the average of parental breeding values plus Mendelian sampling. Solutions for non-parents are obtained using the solutions for the fixed effects and parents. | $\begin{bmatrix} \mathbf{y_p} \\ \mathbf{y_n} \end{bmatrix} = \begin{bmatrix} \mathbf{X_p} \\ \mathbf{X_n} \end{bmatrix} \mathbf{b} + \begin{bmatrix} \mathbf{Z_p} \\ \mathbf{Z_n} \end{bmatrix} \mathbf{u_p} + \mathbf{Z_2 m} + \mathbf{Z_3 pe} + \begin{bmatrix} \mathbf{e_p} \\ \mathbf{e_n} \end{bmatrix}$ |
| | Multivariate maternal animal model | Suitable for multiple trait evaluation with maternal and direct effects. | An example model with two traits:<br>$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \end{bmatrix} + \begin{bmatrix} \mathbf{Z_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} \end{bmatrix} \begin{bmatrix} \mathbf{u_1} \\ \mathbf{u_2} \end{bmatrix} + \begin{bmatrix} \mathbf{W_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{W_2} \end{bmatrix} \begin{bmatrix} \mathbf{m_1} \\ \mathbf{m_2} \end{bmatrix} + \begin{bmatrix} \mathbf{S_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S_2} \end{bmatrix} \begin{bmatrix} \mathbf{pe_1} \\ \mathbf{pe_2} \end{bmatrix} + \begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \end{bmatrix}$ |
| Non-additive animal model (Prediction of | Animal model with dominance effects | A dominance relationship matrix is required apart from the numerator relationship matrix. | $\mathbf{y = Xb + Za + Wd + e}$ |

| | | | |
|---|---|---|---|
| dominance and epistatic effects in addition to additive genetic effects) | Animal model with epistatic effects | Prediction of epistatic effects. An epistatic relationship matrix is required which can be derived from the numerator relationship matrix and the dominance relationship matrix. | $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wd} + \mathbf{Sep} + \mathbf{e}$ |

Currently, an animal model with maternal effects (Bienefeld *et al*., 2007) is used in the honey bee for genetic evaluation.

#### *1.2.3.3. Genetic evaluation approaches*

On the basis of information used for the estimation of breeding values, genetic evaluation can be classified into (1) pedigree based approaches and (2) marker based approaches. These approaches have been further categorized on the basis of the type of information available about the selection candidate.

**Pedigree based approaches**

The pedigree based approach relies on the pedigree information. The estimation of breeding values based on this approach requires knowledge about the genetic covariance between individuals in the pedigree. The genetic covariance among relatives is determined by the additive genetic variance, dominance variance and the epistatic variance (Mrode, 2005). Usually, the latter two are not included in the genetic component and instead form a part of the residual effects. The additive genetic variance is crucial to the estimation of breeding values and requires the construction of an 'additive genetic relationship matrix'. The additive genetic relationship matrix (**A**), also called as the 'numerator relationship matrix', provides an estimate of the degree of relatedness between individuals based on the probability of genes being identical by descent[4]. The **A** matrix is symmetrical. Diagonal elements are equal to $1+F_i$ ($F_i$ stands for the inbreeding coefficient for the i$^{th}$ individual) whereas off-diagonal elements denote the relationship between individuals which is equal to twice the 'Coancestry' or the 'Kinship Coefficient' (Falconer and Mackay, 1996; Mrode, 2005). The covariance between breeding values is given as $\mathbf{A}\sigma_a$, where $\sigma_a$

[4] Two alleles may be called identical by decent if they have originated from the replication of one single allele in a previous generation (Falconer and Mackay, 1996)

is the additive genetic variance (Mrode, 2005). Since honey bees have unique reproductive characteristics, Chapter 3 describes the assumptions and methodology for constructing the numerator relationship matrix for the honey bee.

**Marker based approaches**

With the advancement in genotyping technology, it has become possible to include molecular marker information for genetic evaluation. The different methodologies that employ marker data for selection are described in the following section.

Marker-assisted selection - The use of marker information together with traditional animal breeding methodology is termed marker-assisted selection (MAS). It has been utilized for the breeding programmes of crops (Davierwala *et al*., 2001, Flint-Garcia *et al*., 2003), dairy cattle (Boichard *et al*., 2002; Bennewitz *et al*., 2003), pigs (Visscher and Haley, 1995; Hayes and Goddard, 2003) and fishes (Sonesson, 2007). The methodology of MAS relies on the association of marker loci with the QTL responsible for the trait of interest. It can be based on the linkage equilibrium (LE[5]) of the QTL with markers (LE-MAS), linkage disequilibrium (LD[5]) of the QTL with markers (LD-MAS) or the causative mutation itself resulting in the QTL effect (Hayes, 2008). The concept of LE and LD are explained in more detail in Chapter 2. The methodology of incorporating marker information in the BLUP procedure for genetic evaluation was developed by Fernando and Grossman (1989). The model is given as follows:

$$y_i = x_i\beta + v_i^p + v_i^m + a_i + e_i \qquad 1.4$$

where $y_i$ is the phenotype of the i$^{th}$ animal, $\beta$ is the fixed effect, $v_i^p$ and $v_i^m$ are the effects of the paternally and maternally inherited QTL alleles of the i$^{th}$ animal and $a_i$ is the additive genetic effect of the remaining QTL not linked to the marker locus of the i$^{th}$ animal. The covariance matrix for the additive genetic effects i.e. the numerator relationship matrix **A**, requires information about the relationship among individuals. The covariance matrix for the effect of QTL alleles, $\mathbf{G_v}$, depends both on the relationship and marker information. The $\mathbf{G_v}$ matrix

[5] LE is observed if the expected value of a genotype frequency is the product of its allele frequencies. Any deviation from this equilibrium is LD which results due to the non-random association of alleles at two loci.

becomes equivalent to the **A** matrix as the distance between the markers and the QTL increases, and the marker based estimates approaches the non-marker based estimates of breeding values.

The advantage of MAS over the use of a pedigree based approach is that genetic markers can yield a more accurate estimate of the breeding values as markers capture the variance associated with genes/regions affecting the trait more precisely (Weller, 2001; Mrode, 2005; Hayes, 2008). A drawback associated with MAS is that it only takes into account those few regions that have a large effect on the trait of interest and a large number of regions having small effect on the trait are excluded (Thallman, 2009). For more information about MAS, please refer to Fernando and Grossman (1989), Weller (2001), Dekkers (2004), Mrode (2005) and Hayes (2008).

Genomic selection - An alternative to the MAS approach is the 'Genomic selection' strategy (Meuwissen *et al*., 2001) that takes into account all regions that influence a trait, thus also accounting for the relatively large number of regions with small effects. The whole genome sequencing of most agricultural animals and the availability of a large number of SNP has made it possible to use high-density marker information for genetic evaluation. Meuwissen *et al*. (2001) proposed the methodology of genomic selection that exploits this dense marker information for genetic evaluation. The high-density of marker ensures all QTL are in LD[5] with a marker or a marker-haplotype, thus accounting for the entire genetic variance across the genome. In general, the implementation of genomic selection requires the following two steps (Meuwissen *et al*., 2001; Hayes, 2008):

Step 1: Estimation of marker/haplotype[6] effects in each chromosome segment in the reference population containing both genotypic and phenotypic information.

Step 2: Prediction of the genomic breeding values in the selection candidates which have been genotyped. The genomic estimated breeding values of individuals with genotyping data can be given as $\sum_{i=1}^{n} \mathbf{X}_i \hat{\mathbf{g}}_i$, where n is the number of chromosome segments across the genome, $\mathbf{X}_i$ is a

[6] A haplotype refers to a group of alleles at adjacent marker loci on a chromosome which is assumed to be transmitted together from a parent.

design matrix allocating animals to the marker/haplotype effects in the $i^{th}$ chromosome segment and $\hat{\mathbf{g}}_i$ is a vector of marker/haplotypes effects in the $i^{th}$ chromosome segment.

It should be noted that the number of chromosome segment effects to be estimated is much larger than the number of records. As a result, there are not enough degrees of freedom to fit all effects simultaneously. Meuwissen *et al*. (2001) proposed some approaches to circumvent this shortage of the degree of freedom. These approaches are (1) Least-squares (2) BLUP and (3) Bayesian methods (BayesA and BayesB using Gibbs sampling and Metropolis-Hastings algorithm, respectively). Meuwissen *et al*. (2001) discussed the different attributes of these three approaches with respect to the estimation of allelic effects and compared them for their accuracy of predicting the breeding values. In the least squares approach, the simultaneous estimation of all chromosome segment effects is not possible, so Meuwissen *et al*. (2001) adopted a step-wise approach where the chromosome segments are fitted one by one. The chromosome segments that increase the log-likelihood by more than 14 units are assumed to be significant and included in the model. This approach makes no assumptions regarding the distribution of chromosome segment effects because it treats these effects as fixed. The BLUP approach allows simultaneous estimation of all chromosome segment effects. It is based on the assumption that the chromosome segment effects come from a distribution and the variance of effect at each chromosome segment are identical. In the Bayesian method, the chromosome segment effects are also assumed to have a prior distribution, but the variance of chromosome segment effect varies for different chromosome segments. Thus, during the estimation of the effects of haplotypes or single markers within the chromosome segments, the Bayesian approach allows to capture a more realistic situation by considering the fact that there will be chromosome segments containing QTL with large, moderate or small effects as well as chromosome segments with no QTL.

Apart from the approaches described above, the genomic selection can also be implemented by using a ‘genomic relationship matrix’ (VanRaden *et al*., 2008) instead of the numerator relationship matrix. Hayes *et al*. (2009) demonstrated that this method of predicting breeding values was equivalent to the genomic selection methodology when the effects of QTL contributing to variation in the trait were assumed to be normally distributed. The genomic

relationship matrix (**G**) is constructed from the marker information obtained from the genotyped individuals. The methodology for estimating breeding values is similar to pedigree based BLUP; however, instead of the relationship matrix (**A**) derived from pedigree, the genomic relationship matrix is used. The procedure of constructing the **G** matrix will be discussed in more detail in Chapter 3.

The unified approach - Genetic evaluation based on the genomic selection methodologies requires a multi-step procedure. This may lead to the loss of some information and to selection bias (Vitezica *et al*., 2011). Also, it is a well known fact that not all individuals in a pedigree can be genotyped due to technical and economical constraints. Thus, a single-step procedure, 'the unified approach', was proposed by Legarra *et al*. (2009) and Christensen and Lund (2010). It integrates phenotypic, pedigree and genotyping information from both genotyped and ungenotyped animals for the prediction of breeding values. The unified approach provides an optimal solution for the integration of molecular marker data for the estimation of breeding values in the honey bee. Considering the complex population structure of the honey bee, the unified approach will prove to be an advantageous method for improving the accuracy of estimation of breeding values, response to selection, genetic gain and lowering the rate of inbreeding. More details on the implementation of the unified approach in honey bees are given in Chapter 3.

# Chapter 2. A new framework for modelling and simulation of a honey bee population

Prior to performing a study with an actual genotyping dataset, it is important to validate any genetic evaluation methodology through simulations. Simulation studies require molecular genetic and pedigree datasets to ascertain selection methods. This chapter describes the generation of a dataset for a honey bee population. Sections 2.1 and 2.2 deal with the important aspects of population modelling and simulation as well as the method of simulation of key statistical quantities required for genetic evaluation. Population modelling took into account the reproductive and genetic peculiarities of the honey bee such as a high recombination rate, haplo-diploid sex determination, arrhenotoky and polyandry. At the end of Section 2.1, key population statistics such as allele frequency, minor allele frequency, Hardy-Weinberg equilibrium and linkage disequilibrium are defined. Section 2.2 describes the simulation of true breeding values, phenotypic values, negative correlation between maternal and direct effects, heritability, genetic and residual variances. Furthermore, ideas that could be exploited in future to obtain genotyping information for ungenotyped animals are presented in Section 2.3.

## 2.1. Population modelling

### 2.1.1. Base population for the honey bee

A base population is used as a starting point for simulation studies. It is composed of individuals in a pedigree for which no ancestral information is available, and is assumed to be in mutation-drift equilibrium with linkage disequilibrium (LD). Thus, a software program was developed that was capable of producing a dataset for a base population in the honey bee. It provided the possibility to investigate the effect of parameters like mutation rate, density of markers and number of individuals on the extent of LD in a population.

In order to generate a dataset according to the requirements specific to honey bee populations, the software program allows to input: (1) number of generations, (2) number of sire queens, (3) number of dam queens, (4) number of marker loci, (5) forward and backward mutation rates, (6) minor allele frequencies and (7) number of marker loci to be selected as SNP on the basis of

minor allele frequency. Further details for implementing the program are provided in the appendix. The features of the software program and modelling of the population structure, the genome and the evolutionary processes are described below.

#### *2.1.1.1. Structure of the base population*

The software program constructs a population structure according to the provided input. The input data includes number of sire queens and dam queens (with a ratio of 10:1), number of generations and total number of marker loci to be simulated (assumed to be bi-allelic). In order to model the haploid drones, a sire queen is defined that represents a drone-producing colony. Two matrices with a size equal to number of individuals by number of marker loci represent the genome of diploid sire queens and dam queens, respectively. The population size remains constant in every generation according to the Fisher-Wright population model. Furthermore, all simulated generations are non-overlapping.

#### *2.1.1.2. Genome*

A diploid genome, consisting of 16 linkage groups, is simulated for sire and dam queens. The length of each chromosome is simulated according to the actual length of all honey bee chromosomes. The number of marker loci ( $N$ ) to be simulated along the genome can be provided as input. In the software program, the number of marker loci to be distributed per chromosome, $N_i$ ($i = 1,2,.....16$), is based on the actual proportion of SNP loci present on each honey bee chromosome and is computed using the following formula:

Number of marker loci on the $i^{th}$ chromosome, $N_i = NR_i$ 2.1

where $R_i$ is equal to the actual ratio between number of SNP loci on the $i^{th}$ chromosome and total number of SNP loci in the honey bee genome. Positions of all loci on all chromosomes are sampled from a uniform distribution. The number of SNP loci per chromosome and the length of all 16 chromosomes were obtained from the honey bee SNP database (www.hgsc.bcm.tmc.edu/projects/honeybee/;http://www.ncbi.nlm.nih.gov/genome?term=apis%20mellifera) as shown in Table 2.1.

Table 2.1. Summary of the chromosome length, number of SNP and $R_i$.

Chromosome length and SNP data were obtained from the honey bee genome database; $R_i$ is the actual ratio between the number of SNP on the $i^{th}$ chromosome and the total number of SNP in the honey bee genome; this information is used to simulate the genome and to distribute markers across the chromosomes.

| **Chromosome** | **Length (in base-pairs)** | **Number of SNP** | **$R_i$** |
|---|---|---|---|
| 1 | 29,893,408 | 140,148 | 0.1414 |
| 2 | 15,549,267 | 62,801 | 0.0633 |
| 3 | 13,234,341 | 70,577 | 0.0712 |
| 4 | 12,718,334 | 55,407 | 0.0559 |
| 5 | 14,363,272 | 62,750 | 0.0633 |
| 6 | 18,472,937 | 78,086 | 0.0788 |
| 7 | 13,219,345 | 59,210 | 0.0597 |
| 8 | 13,546,544 | 61,811 | 0.0623 |
| 9 | 11,120,453 | 55,302 | 0.0558 |
| 10 | 12,965,953 | 50,243 | 0.0507 |
| 11 | 14,726,556 | 68,972 | 0.0696 |
| 12 | 11,902,654 | 57,616 | 0.0581 |
| 13 | 10,288,499 | 50,380 | 0.0508 |
| 14 | 10,253,655 | 48,322 | 0.0487 |
| 15 | 10,167,229 | 38,452 | 0.0388 |
| 16 | 7,207,165 | 31,295 | 0.0316 |
| Total | 219,629,612 | 991,372 | |

***2.1.1.3. Evolution***

To simulate an evolutionary process, recombination and mutation are implemented during the process of gamete formation in every generation. Multiple mating is modelled in the parental generations. The processes are briefly described below.

**Recombination**

Recombination is the exchange of chromosomal segments between paternal and maternal chromosomes. It is implemented in the software program as follows. The recombination probability ($\theta$) between two adjacent loci on a chromosome is calculated from the Haldane mapping function (Haldane, 1919), which is the most commonly used mapping function. It is based on the assumption that crossovers in any given chromosomal segment follow a Poisson distribution, with no interference between crossovers. The recombination probability is calculated using the following expression:

$$\theta = \frac{1}{2}\left[1 - \exp\left(-2 \mid x \mid\right)\right] \qquad 2.2$$

where exp denotes the exponential function and $\mid x \mid$ stands for the absolute value of the map distance between adjacent loci. The Haldane mapping function requires that distances are expressed in Morgan units, therefore, distances between two loci are converted from base-pairs to Morgan using the reported recombination rate of 19 cM/Mb (Beye *et al*., 2006, The Honeybee Genome Sequencing Consortium, 2006).

**Mutation**

Mutation is implemented in the software program to create polymorphisms. All loci in all individuals belonging to generation zero have a single allele coded as 1. Both forward and backward mutations are modelled, allowing each locus to mutate from allele 1 to allele 2 and from allele 2 to allele 1. The required rates of forward and backward mutations can be specified in the input as mutation rates per locus per gamete per generation. The advantage of modelling a bi-directional mutation is that different values of forward and backward mutation rates can be chosen. Setting the backward mutation rate to zero will result in an infinite site model of mutation (Kimura, 1969) where each locus can only mutate once over all generations and mutation will result in the formation of allele 2. The infinite site model of mutation can be useful

when simulating an extremely high initial marker density with a low mutation rate and a large number of generations as shown in studies by Sonesson and Meuwissen (2009) and Calus *et al*. (2008), where 1 million and 300,000 marker loci were simulated for a genome of size 10 M and 3 M, respectively.

**Multiple mating**

Figure 2.1 describes the general mating scheme followed during the simulation. Polyandry, commonly referred to as multiple mating, is a phenomenon observed in honey bee whereby a queen mates with multiple drones (average of 10 to 20 drones). To model this situation in the software program, a dam queen and a group of 10 sire queens (a sire queen represents a drone-producing colony) are assumed as mating partners. To form groups, all sire queens are randomly permuted and thereafter divided into groups consisting of 10 sire queens.

A detailed mating scheme, showing how gametes from a dam queen are combined with the drones from a group of sire queens is illustrated in Figure 2.2. A dam queen generates a total of 11 gametes, of which 10 give rise to sire queens and one to a dam queen in the next generation. Since a gamete produced by a sire queen is regarded as a drone, it is assumed to occur in multiple copies. One of the 10 sire queens of a group contributes a drone, which combines with a gamete from the dam queen to produce a new dam queen for the next generation. In addition to the drone generated for the formation of a dam queen, each of the 10 sire queens of a group produces one drone, thus a group contributes a total of 11 drones. During the formation of a sire queen, all 11 drones of a group have an equal probability to be drawn as a gamete. Since drones in a set are sampled with replacement, the resulting progenies are related as super-sibs (coefficient of relatedness = 0.75), full-sibs (coefficient of relatedness = 0.5) or half-sibs (coefficient of relatedness = 0.25).

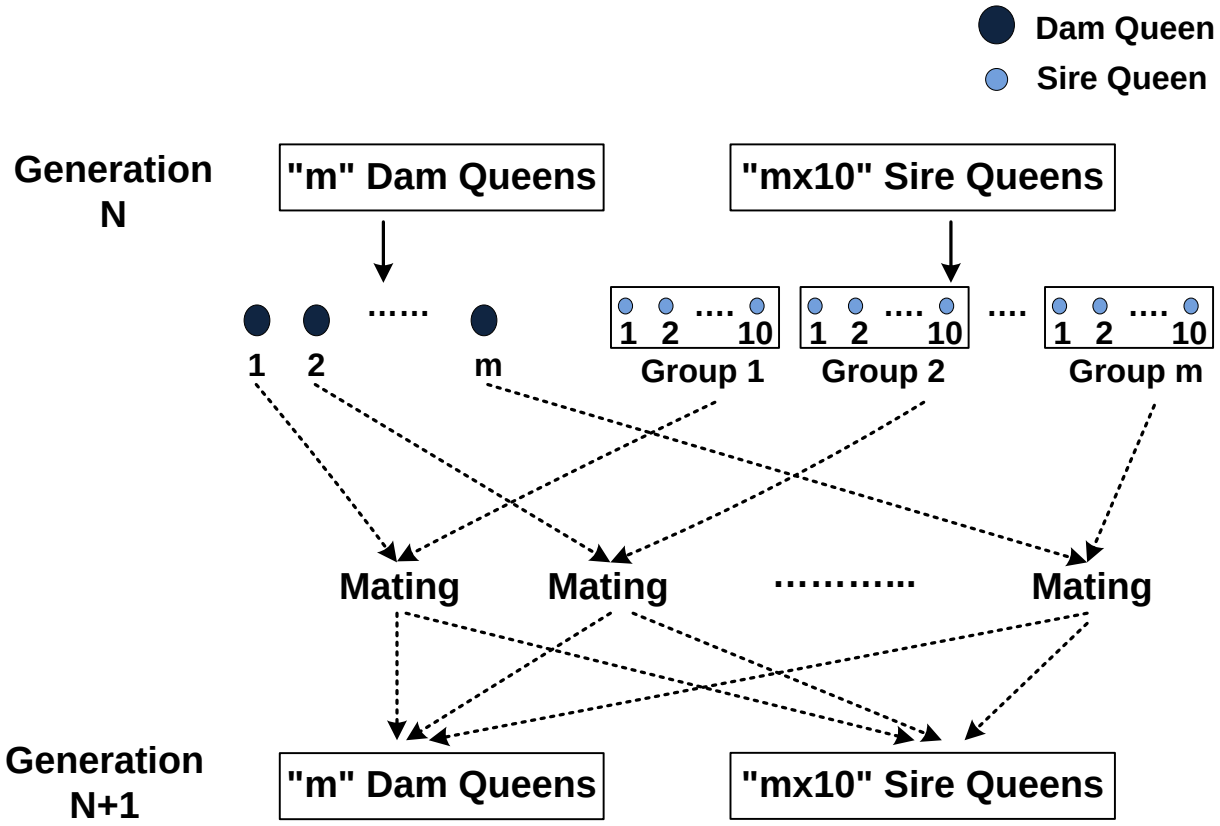


Figure 2.1. General mating scheme for the base population.

m = total number of dam queens; since there is a 1:10 ratio between number of dams and sires, the number of sire queens is “mx10”; in every generation, all sire queens are randomly permuted and grouped; each group consists of 10 sire queens; a dam queen and a group of sire queens are the mating partners; all generations are non-overlapping and the population size is kept constant across all generations.

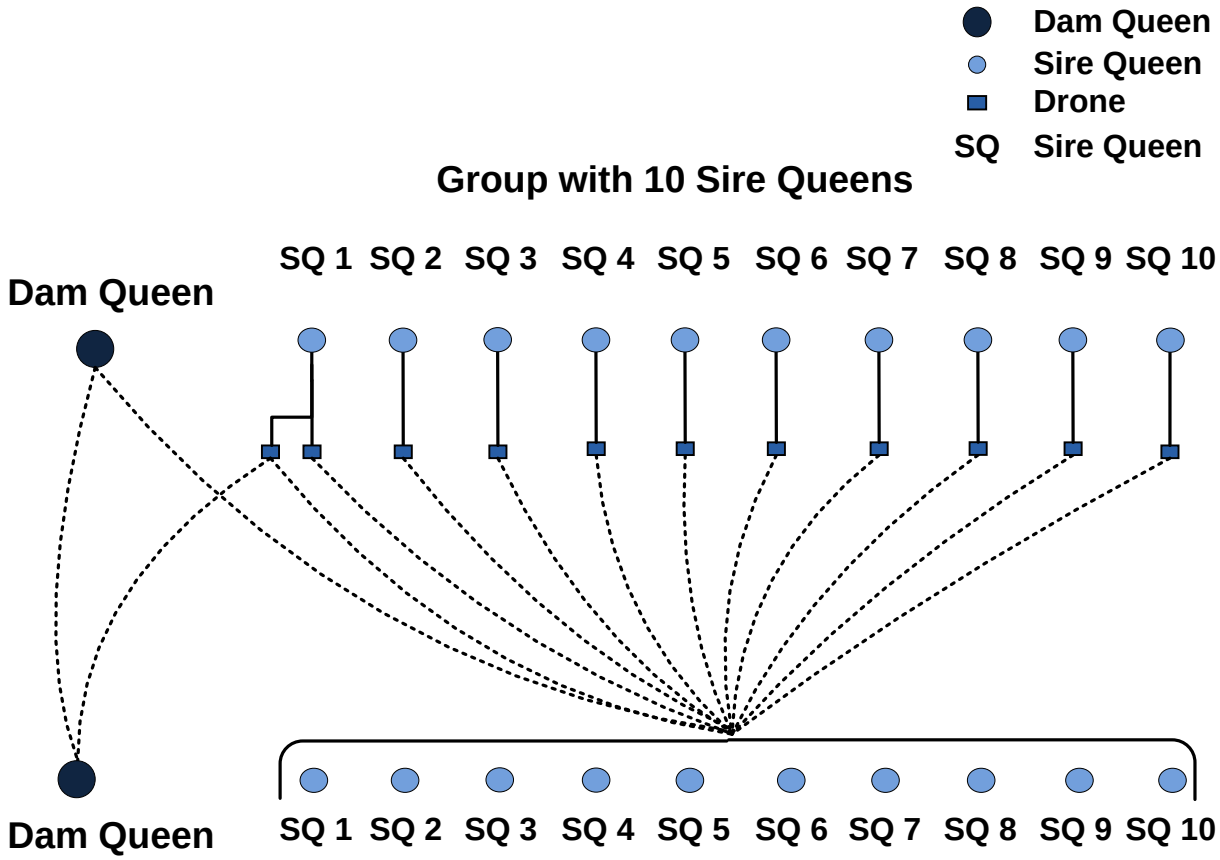


Figure 2.2. Multiple mating in the base population.

Multiple mating between a dam queen and 11 drones from a group. The resulting offspring consist of one dam queen and 10 sire queens; all drones are sampled with replacement which models the phenomenon of producing multiple copies of identical gametes by a drone.

***2.1.1.4. Base population structure for this study***

For this study, a population with 550 queens, 500 sire queens and 50 dam queens, were simulated with 100,000 marker loci (Table 2.2) for 1000 generations to obtain a base population in mutation-drift equilibrium. The forward and backward mutation rates were taken to be 0.0025. In the resulting base population, 44,000 marker loci (Spötter *et al*., 2012) with the highest minor allele frequency were chosen. Out of these 44,000 marker loci, 250 with the highest minor allele frequency were chosen as QTL and the remaining as SNP. Consequently, the average distance between adjacent SNP loci was approximately 0.001 M. QTL alleles received an effect drawn from a normal distribution $N(0,1)$.

In addition, a validation was performed with populations consisting 220 and 550 queens. The achieved LD was compared with the theoretical value of LD. Results are presented in Chapter 5.

Table 2.2. A summary of the simulated number of markers on each chromosome of the honey bee.

| **Chromosome** | **Number of marker** |
|---|---|
| 1 | 14,137 |
| 2 | 6,335 |
| 3 | 7,119 |
| 4 | 5,589 |
| 5 | 6,330 |
| 6 | 7,877 |
| 7 | 5,973 |
| 8 | 6,235 |
| 9 | 5,578 |
| 10 | 5,068 |
| 11 | 6,957 |
| 12 | 5,812 |
| 13 | 5,082 |
| 14 | 4,874 |
| 15 | 3,879 |
| 16 | 3,155 |
| Total | 100,000 |

### 2.1.2. Modelling a 'natural' population structure - Subsequent populations from the base population

Five additional overlapping generations were simulated from the base population. A realistic structure was modelled for the subsequent generations since they composed the population that was used to test the methodology of genetic evaluation. In the following section, modelling of these five additional generations is described.

#### *2.1.2.1. Mating and selection scheme*

As described earlier, the base population consisted of 550 unrelated queens in total. From these 550 queens, 50 queens were selected as dam queens and the remaining queens were used as drone-producing queens. Each of the generations 1-5 consisted of 500 potential-dam queens and 250 drone-producing queens. From these 500 potential-dam queens, 10% (i.e. 50 queens) were randomly selected (Figure 2.3) as dam queens.

In each generation, the 50 selected dam queens produced a total of 500 potential-dam queens. In addition, half of the dam queens (i.e. 25 dam queens) also produced 250 drone-producing queens. As a result, the population size in each generation remained constant with 500 potential-dam queens and 250 drone-producing queens (Figure 2.3). To summarize, the complete pedigree from the base population up to the fifth generation consisted of 4300 queens i.e. 50 dam queens and 500 drone-producing queens in the base population and 2500 potential-dam queens and 1250 drone-producing queens from generations 1-5.

In most of the simulation studies, the number of progenies produced by a dam/sire is usually fixed, but here a more realistic situation was modelled by allowing all dam queens to have an equal probability to be sampled as a dam for the next offspring. For the current study only 10% individuals were selected. The selection percentage could be varied depending on the requirements of the study.

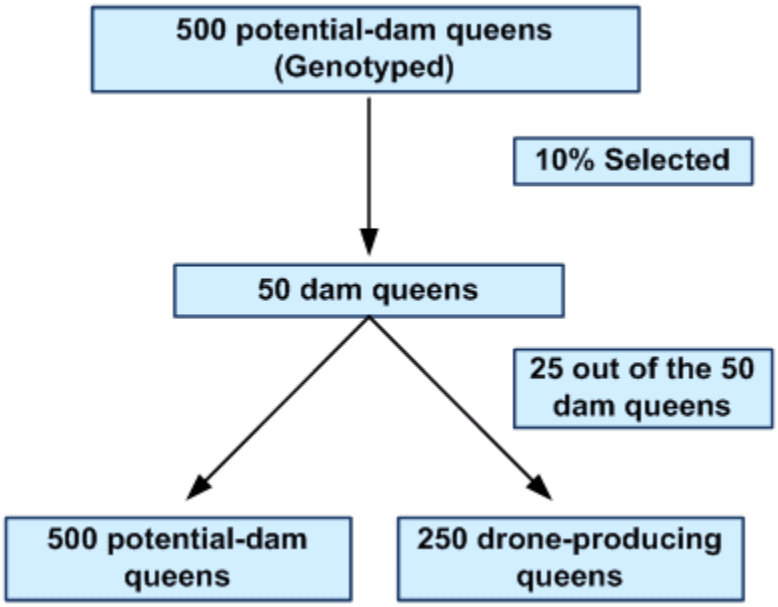


Figure 2.3. Selection scheme.

The selection scheme shows that in each generation 10% of the potential-dam queens were selected which produced offspring for the next generation.

***2.1.2.2. Population characteristics specific to the honey bee***

To construct a population similar to that used for genetic evaluation in the honey bee (Bienefeld *et al*., 2007), a dummy sire and an average worker was constructed. Generations following the base population were overlapping and mating was polyandrous as in the real breeding population. These characteristics are described in more detail in the following section.

**Construction of a dummy sire**

As a consequence of polyandry in the honey bee, offspring have an unclear paternal descent. To overcome the problem of representing the paternal descent, Bienefeld *et al*. (2007) suggested using a dummy sire in the pedigree. A dummy sire represents a group of sister colonies (approximately 8-10 sister colonies) which are maintained at the mating stations (Nolan, 1937; Rothenbuhler, 1958) with the purpose of producing only drones. This way, a controlled mating is ensured. An example pedigree depicting a dummy sire is shown in Figure 2.4 and Table 2.3. For the current study, it was assumed that a dummy sire consisted of 10 drone-producing queens; thus, each generation consisted of 25 dummy sires formed by 250 drone-producing queens. It should be noted that in generations 1-5, the 10 drone-producing queens that formed a dummy

sire were related as sisters as they had the same dam queen and dummy sire (Figure 2.4), a situation similar to mating stations used in several European countries.

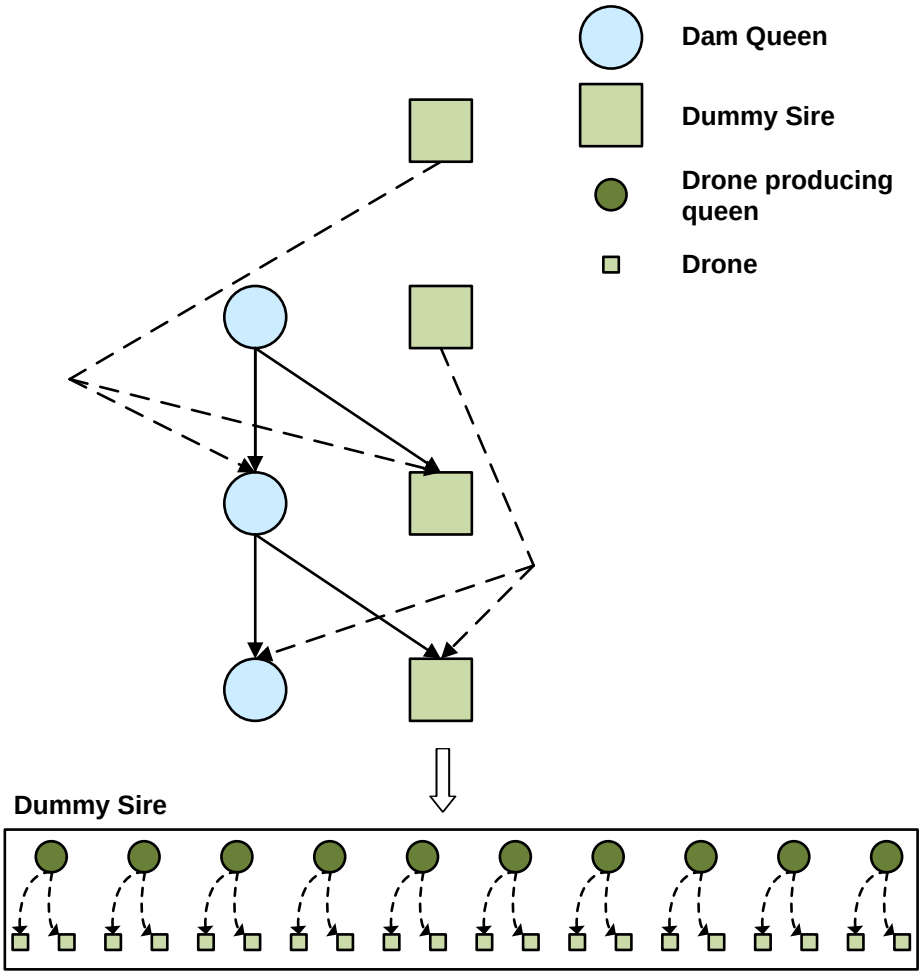


Figure 2.4. Pedigree diagram.

In this pedigree diagram, dummy sires are represented by the larger square box in light green whereas dam queens are represented by the circle in blue. The expanded rectangular box shows a dummy sire consisting of 10 drone-producing sister queens represented by smaller circle in dark green. Each drone-producing sister queen contributes two drones which are represented by the smaller square box in light green. The pedigree shows that mating takes place between overlapping generations. All drone-producing sister queens comprising a dummy sire have a common dam queen and dummy sire, thus, they are related as sisters.

Table 2.3. An example pedigree.

The example pedigree shown here consists of queens, average workers and dummy sires belonging to three generations. All animals have been assigned a fictitious animal ID. It also shows that mating takes place between overlapping generations.

| Generation | Animal ID | Dam Queen | Dummy Sire |
|---|---|---|---|
| Generation 1 | Queen1 | - | - |
| | Queen2 | - | - |
| | Queen3 | - | - |
| | Queen4 | - | - |
| | Queen5 | - | - |
| | Queen6 | - | - |
| | Dummy Sire1 | - | - |
| | Dummy Sire2 | - | - |
| | Average Worker1 | Queen1 | - |
| | Average Worker2 | Queen2 | - |
| | Average Worker3 | Queen3 | - |
| | Average Worker4 | Queen4 | - |
| | Average Worker5 | Queen5 | - |
| | Average Worker6 | Queen6 | - |
| Generation 2 | Queen7 | Queen1 | - |
| | Queen8 | Queen6 | - |
| | Queen9 | Queen2 | - |
| | Queen10 | Queen3 | - |
| | Queen11 | Queen1 | - |
| | Queen12 | Queen2 | - |
| | Dummy Sire3 | Queen3 | - |
| | Dummy Sire4 | Queen6 | - |
| | Average Worker7 | Queen7 | Dummy Sire1 |
| | Average Worker8 | Queen8 | Dummy Sire1 |
| | Average Worker9 | Queen9 | Dummy Sire2 |
| | Average Worker10 | Queen10 | Dummy Sire2 |
| | Average Worker11 | Queen11 | Dummy Sire1 |
| | Average Worker12 | Queen12 | Dummy Sire2 |
| Generation 3 | Queen13 | Queen9 | Dummy Sire2 |
| | Queen14 | Queen11 | Dummy Sire1 |
| | Queen15 | Queen7 | Dummy Sire1 |

| | | | |
|---|---|---|---|
| | Queen16 | Queen10 | Dummy Sire2 |
| | Queen17 | Queen7 | Dummy Sire1 |
| | Queen18 | Queen10 | Dummy Sire2 |
| | Dummy Sire5 | Queen10 | Dummy Sire2 |
| | Dummy Sire6 | Queen11 | Dummy Sire1 |
| | Average Worker13 | Queen13 | Dummy Sire4 |
| | Average Worker14 | Queen14 | Dummy Sire4 |
| | Average Worker15 | Queen15 | Dummy Sire3 |
| | Average Worker16 | Queen16 | Dummy Sire3 |
| | Average Worker17 | Queen17 | Dummy Sire4 |
| | Average Worker18 | Queen18 | Dummy Sire3 |

**Construction of an average worker**

A colony is formed by a queen and its progeny comprising several thousand workers. Since it is impossible to include all workers of a colony for genetic evaluation, an average worker was constructed that represented all workers of a colony. It was assumed that one progeny average worker existed for each potential-dam queen/dam queen in the pedigree (Table 2.3). As a result both maternal and direct effects were taken into account.

**Modelling polyandry and overlapping generations**

In each generation, 50 dam queens and 25 dummy sires were randomly selected as mating partners. A dam queen mated with one specific dummy sire, whereas a dummy sire mated with more than one dam queen. To model polyandry, each dummy sire provided 20 drones (two from each sister queen) to the dam queen for mating.

For generations to be overlapping, queens chosen to become dam queens were sampled from the $n^{th}$ generation and queens constituting a dummy sire came from $n\text{-}1^{th}$ generation, i.e. one generation preceding to the dam queen's generation (Figure 2.5). The described mating scheme was consistent with the scheme followed by most bee breeders in Europe. It resulted in the offspring being related as 'super-sibs', 'full-sibs' or 'maternal half-sibs'. Super-sibs or full-sibs have a common mother and a dummy sire. A paternal gamete comes from a single drone in case of super-sibs and different drones derived from the same queen in case of full-sibs. Maternal half-sibs also share the same mother and dummy sire, but a paternal gamete comes from different drones derived from two sister queens.

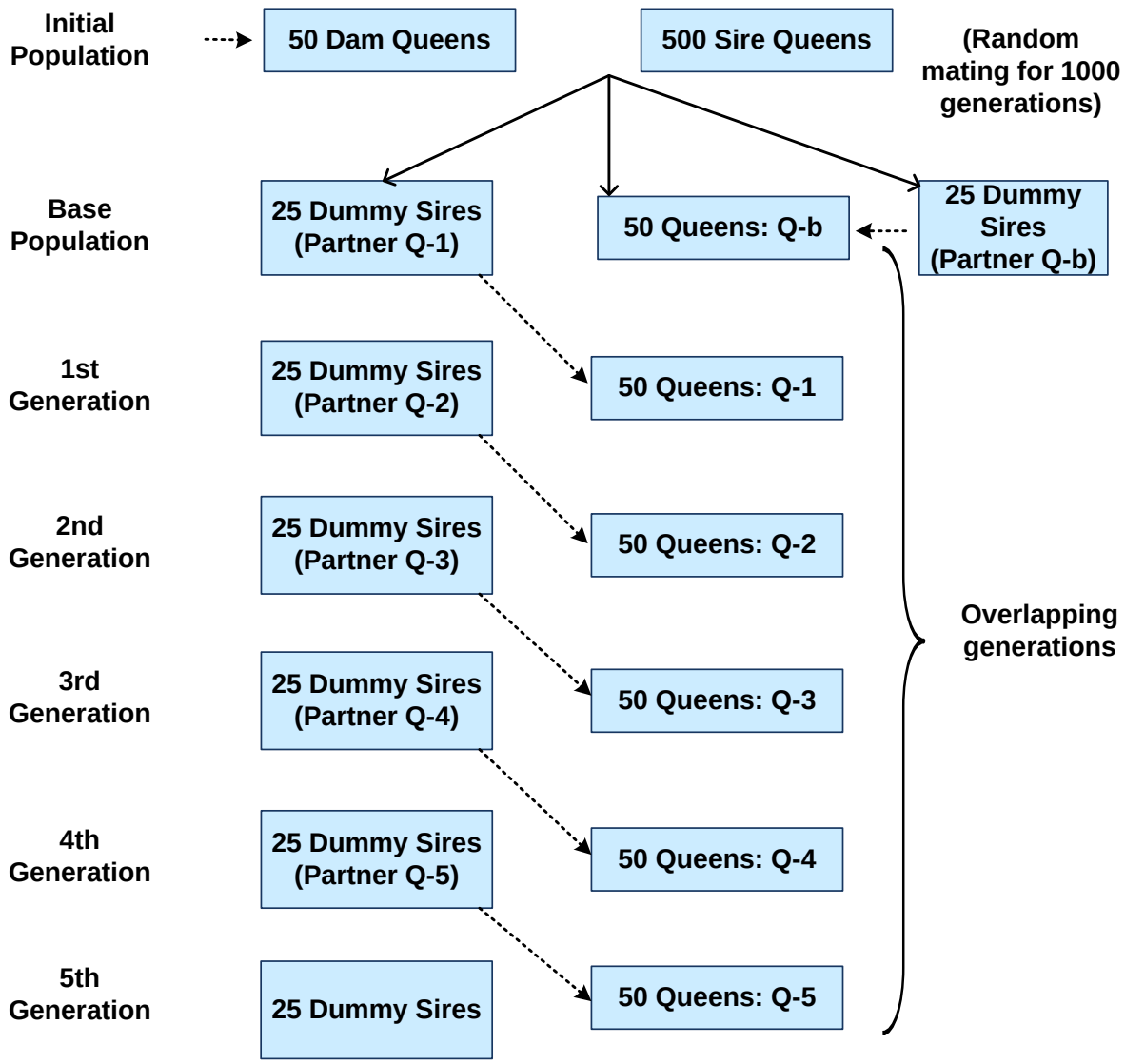


Figure 2.5. Mating scheme in the simulated population.

The mating scheme in the simulated generations 1-5 is illustrated in this figure. Dummy sires connected through dashed arrow to queens are the mating partners. Q-b are the queens belong to the base population and Q-1 to Q-5 are queens belonging to generations 1 to 5. Each dummy sire is equivalent to 10 drone-producing sister colonies. Therefore, in total 25 dummy sires represent 250 queens. This mating scheme is similar to the mating scheme followed by several bee breeders in European countries.

### 2.1.3. Population statistics

Statistics for the allele frequency, Hardy-Weinberg equilibrium, minor allele frequency, LD etc. allow inferences to be made about an evolving population. Allele frequency data is a requisite for any population; Hardy-Weinberg equilibrium and minor allele frequency are the usual criteria to

evaluate the informativeness of marker loci. Most studies based on genome-wide marker data rely on the assumption that a marker and the locus affecting the trait are in LD (Slatkin, 2008).

For this study, these statistics were also calculated and used during the simulation of the datasets. For example, the software program for the base population calculated statistics for the allele frequency, minor allele frequency, Hardy-Weinberg equilibrium and LD. In the last generation, the software program calculates the average LD value for selected SNP with the highest minor allele frequency. In addition, generation-wise LD values for all simulated marker locus pairs are calculated and plotted in a graph. As a measure of LD, $r^2$ (Hill and Robertson, 1968; Hill, 1975) is used in the software program:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b} \qquad 2.3$$

where $D = p_{AB} p_{ab} - p_{Ab} p_{aB}$ ; $p_{AB}$ , $p_{ab}$ , $p_{Ab}$ , $p_{aB}$ and $p_A$ , $p_a$ , $p_B$ , $p_b$ are the observed frequencies of haplotypes AB, ab, Ab, aB and of alleles A, a, B, b, respectively, in the population. Allele frequency, minor allele frequency and Hardy-Weinberg equilibrium as well as the concept of LD, different LD measures and factors affecting it are explained briefly in the following sections.

#### *2.1.3.1. Allele frequency*

It is one of the most important statistics in a population genetics study. Allele frequency refers to the proportion of a gene variant in a population (Falconer and Mackay, 1996; http://www.nature.com/scitable/definition/allele-frequency-298). It can be expressed in different forms, i.e. as a percentage or fraction. In this study, information about the allele frequency of marker loci in the base population was used to construct the genomic relation matrix.

#### *2.1.3.2. Minor allele frequency*

Minor allele frequency is the frequency of the less frequent allele of a SNP in the population. It is an important criterion for selecting informative SNP. Usually, SNP with a MAF of less than 0.05 are discarded in the genome-wide association studies as it requires a strong statistical power to make a useful prediction about such rare alleles. As described previously, for this study,

43,750 loci with the highest MAF were chosen as SNP (the remaining 250 out of 44,000 were chosen as QTL) and were used for the genetic evaluation.

#### *2.1.3.3. Hardy-Weinberg equilibrium*

The Hardy-Weinberg equilibrium states that in a random mating population, the allele and genotype frequencies remain constant in every generation, provided that there is no selection, mutation and migration. (Falconer and Mackay, 1996; Hamilton, 2009; http://www.nature.com/scitable/definition/hardy-weinberg-equilibrium-122). The Hardy-Weinberg equilibrium is often used as a null-hypothesis to test the effect of evolutionary forces on the allele and genotype frequencies. The software program for simulating the base population also provides the $\chi^2$ statistics for Hardy-Weinberg equilibrium that can be used as a criterion to select informative marker loci.

#### *2.1.3.4. Linkage disequilibrium*

Linkage disequilibrium (also known as gametic phase disequilibrium, gametic disequilibrium or allelic association) is defined as the non-random association of alleles at two loci. LD has been exploited in marker-assisted selection, genomic selection strategies as well as for mapping QTL associated with the trait of interest.

To explain LD, an example by Hayes (2008) is considered. Let's assume that a pair of loci A and B occurs on a chromosome with two alleles A/a and B/b. Thus, the four haplotypes for this pair of loci are AB, Ab, aB and ab. If the allele frequencies of A, a, B and b in the population are 0.5, then the expected frequencies of each of the four haplotypes in the population is 0.25. Any deviation of the haplotype frequencies from 0.25 is LD (Hayes, 2008). In other words, if the expected value of a haplotype frequency is the product of its allele frequencies, the loci are in linkage equilibrium i.e. $p_{AB} = p_A p_B$, $p_{ab} = p_a p_b$, $p_{Ab} = p_A p_b$ and $p_{aB} = p_a p_B$ (Table 2.4). Any deviation from this equilibrium is disequilibrium.

Table 2.4. Allele and genotype frequencies.

| **Alleles** | **A** | **a** | **Allele frequency** |
|---|---|---|---|
| **B** | $p_{AB}$ | $p_{aB}$ | $p_B$ |
| **b** | $p_{Ab}$ | $p_{ab}$ | $p_b$ |
| **Allele frequency** | $p_A$ | $p_a$ | 1 |

$p_{AB}$, $p_{aB}$, $p_{Ab}$ and $p_{ab}$ are the haplotype frequencies; $p_A$ and $p_a$ and $p_B$ and $p_b$ are the allele frequencies for locus A and B, respectively.

Table 2.5. Example data for allele and genotype frequencies.

| **Alleles** | **A** | **a** | **Total** |
|---|---|---|---|
| **B** | 0.25 | 0.25 | 0.5 |
| **b** | 0.25 | 0.25 | 0.5 |
| **Total** | 0.5 | 0.5 | 1 |

Linkage Equilibrium

| **Alleles** | **A** | **a** | **Total** |
|---|---|---|---|
| **B** | 0.1 | 0 | 0.1 |
| **b** | 0 | 0.9 | 0.9 |
| **Total** | 0.1 | 0.9 | 1 |

Linkage Disequilibrium

$r^2 = 1$; 'Perfect LD'

**Measures of LD**

Different measures of LD have been proposed (Devlin and Risch, 1995; Zhao *et al*., 2007). A concise description of the important and commonly used measures of LD is as follows:

$D$ – This measure of LD gives the difference between the observed and expected haplotype frequencies (Lewontin and Kojima, 1960; Hill, 1968, 1975, 1981; Hayes, 2008). For the above example (Table 2.4, 2.5), when the two loci are in LD, the deviation in the expected frequencies can be given as: $p_{AB} = p_A p_B + D$, $p_{Ab} = p_A p_b - D$, $p_{aB} = p_a p_B - D$ and $p_{ab} = p_a p_b + D$. The commonly used expression for $D$ is given as following:

$$D = p_{AB} p_{ab} - p_{Ab} p_{aB} \qquad 2.4$$

This expression can be derived in the following manner:

$$p_{AB} p_{ab} = (p_A p_B + D)(p_a p_b + D)$$

$$= p_A p_B p_a p_b + p_A p_B D + p_a p_b D + D^2 \qquad 2.5$$

$$p_{Ab}p_{aB} = (p_A p_b - D)(p_a p_B - D)$$

$$= p_A p_B p_a p_b - p_A p_b D - p_a p_B D + D^2 \qquad 2.6$$

On subtracting equations 2.5 and 2.6, one obtains:

$$p_{AB}p_{ab} - p_{Ab}p_{aB} = D(p_A p_B + p_a p_B + p_a p_b + p_A p_b)$$

Since $p_A p_B + p_a p_B + p_a p_b + p_A p_b = 1$, therefore,

$$p_{AB}p_{ab} - p_{Ab}p_{aB} = D \qquad 2.7$$

The range of $D$ is $-0.25 \le D \le +0.25$.

A major disadvantage associated with the measure $D$ is that it is highly influenced by the allele frequency.

$D'$ – This measure of LD was proposed by Lewontin (1964). The value of $D'$ is obtained by standardizing $D$ with its maximum value. The expression can be given as following:

$$D' = \frac{|D|}{D_{\max}} \text{ where } D_{\max} = \begin{cases} \min(p_A p_b, p_a p_B) \text{ if } D>0 \\ \min(p_A p_B, p_a p_b) \text{ if } D<0 \end{cases} \qquad 2.8$$

The range of $D'$ is $0 \le D' \le 1$.

$r^2$ –This measure was proposed by Hill and Robertson (1968). It is the square of correlation of allele frequencies in the population and is expressed as follows:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b} \qquad 2.9$$

It is the most commonly used and preferred measure of LD. The range of $r^2$ is $0 \le r^2 \le 1$.

$\chi^{2'}$ measure - This is a multi-locus measure of LD (Zhao *et al*., 2005; Hayes, 2008) and is given as follows:

$$\chi^{2'} = \frac{1}{(l-1)} \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\mathrm{D_{ij}}^2}{\mathrm{pA_i pB_j}} \qquad 2.10$$

where $D = pA_iB_j - pA_i pB_j$, $pA_iB_j$ is the frequency of the haplotype $A_iB_j$, $pA_i$ is the frequency of the $i^{th}$ allele at locus A, $pB_j$ is the frequency of the $j^{th}$ allele at locus B and $l$ is the minimum of the number of alleles at locus A and locus B.

**Factors affecting LD**

Several factors, i.e. recombination, mutation, migration, population admixture, effective population size, genetic drift, selection and inbreeding, affect the extent of LD in a population (Balding *et al*., 2007; http://bio.classes.ucsc.edu/bio107/Class%20pdfs/W05_lecture15.pdf).

Recombination – It is one of the most crucial factors that determine the level of LD in a population. The decay of LD is controlled by the rate of recombination in a species. The general expression for the decay of LD with time for a random mating population in the absence of other evolutionary forces can be given as follows:

$$D_t = D_0(1-\theta)^t$$

where $D_t$ and $D_0$ are the levels of LD at the $t^{th}$ and $0^{th}$ generations, respectively, and $\theta$ is the recombination fraction (see Equation 2.2, Haldane mapping function).

Thus, when two loci are located further apart, LD will tend to be smaller and will decrease over time as a result of recombination. LD gives a general indication of the frequency of recombination, hence, the physical distance between two loci.

Mutation – Mutation rates are generally very small, thus mutation itself can only introduce small changes in the allele frequency. However, on an evolutionary time scale it might contribute to giving rise to LD.

Migration and population admixture – Migration occurring between populations that greatly differ in allele frequencies will lead to the establishment of a larger LD as compared to populations with similar allele frequencies.

Finite population size and genetic drift – The smaller the effective population size, the larger is the effect of genetic drift and the greater is the value of LD.

Selection – LD can be caused due to selection of one combination of alleles over another. However, the effect is localized around specific genes, thus the average LD for the whole genome may not be affected significantly.

Inbreeding – Inbreeding decreases the rate of decay of LD in a population because it reduces the frequency of double heterozygotes.

## 2.2. Simulation

This section begins with the derivation of breeding values for a single locus and introduces the concept of population mean and average effect of an allele. This will help to understand the simulation of true breeding values and phenotypic values in the later sections. Furthermore, the methodology of the calculation of key parameters such as negative correlation between maternal and direct effects, genetic variances (including determination of weighing factors for maternal and direct effects), phenotypic variance, residual variance and heritability of maternal and direct effects is explained.

### 2.2.1. Breeding value for a single locus

In Chapter 1, the concept of breeding values has already been explained. This section further elaborates this concept and describes the method of obtaining breeding values for a single locus using information based on the allele frequency and allele substitution effects (Falconer and Mackay, 1996; Hamilton, 2009). This sets the background for understanding the simulation of true breeding values and phenotypic values.

#### *2.2.1.1. Population mean*

The value observed when a characteristic is measured on an individual is called its phenotypic value. Alternatively, '*mean phenotypic*' value refers to the average phenotype of a population of individuals. The mean phenotype is a result of the effect of genes and environment. Thus, the

components of this mean phenotypic value are *mean genotypic value* and *environmental deviation* (Falconer and Mackay, 1996; Hamilton, 2009). Usually, the mean environmental deviation is taken to be zero, thus in that case, the mean phenotypic value would be equal to the mean genotypic value. The term '*population mean*' refers to this 'mean phenotypic' or 'mean genotypic' value (Falconer and Mackay, 1996; Hamilton, 2009). Consider a gene with alleles $A_1$ and $A_2$ occurring at a frequency of $p$ and $q$. It is assumed that $A_1A_1$, $A_1A_2$ and $A_2A_2$ are associated with the genotypic values $+a$, $d$ and $-a$, respectively, where '$a$' refers to the additive genetic effect and '$d$' is the dominance effect (Table 2.6).

Table 2.6. Population mean.

(Source: Falconer and Mackay, 1996)

| Genotype | Frequency | Value | Frequency x Value |
|---|---|---|---|
| $A_1A_1$ | $p^2$ | $+a$ | $p^2a$ |
| $A_1A_2$ | $2pq$ | $d$ | $2pqd$ |
| $A_2A_2$ | $q^2$ | $-a$ | $-q^2a$ |
| | | | Population mean = $a(p-q) + 2dpq$ |

The population mean is equal to $a(p-q) + 2dpq$, and is obtained by multiplying genotypic value with its frequencies and then summing over all genotypes (Falconer and Mackay, 1996).

***2.2.1.2. Average effect of an allele and breeding values***

The term average effect is used to assign a value to an allele (Hamilton, 2009). The average effect of an allele is defined as the mean phenotypic deviation from the population mean of that group of individuals which received that allele from one parent and the other allele from a parent drawn at random from the population (Falconer and Mackay, 1996; Hamilton, 2009). In simpler terms, the average effect is a 'deviation' that measures the difference between the value of all genotypes that contain a given allele and the population mean (Hamilton, 2009). The average effect of an allele or an allele substitution ($\alpha$) depends on the genotypic value $a$ and $d$ as well as the genotypic and allele frequencies in the population (Hamilton, 2009).

For an illustration, one can consider an example from Falconer and Mackay (1996) that assumes two alleles $A_1$ and $A_2$ at a locus occurring at a frequency of $p$ and $q$ with the average effect of

the allele $A_1$ as $\alpha_1$. If gametes carrying only allele $A_1$ unite with gametes from the population then the frequency of genotype $A_1A_1$ will be $p$ and that of genotype $A_1A_2$ will be $q$. The genotypic value of $A_1A_1$ is $+a$ and that of $A_1A_2$ is $d$, thus, the proportion in which they occur in the population is $pa + qd$. The average effect of the allele $A_1$ is obtained by subtracting this mean value from the population mean, as given below (Falconer and Mackay, 1996):

$$\alpha_1 = pa + qd - [a(p - q) + 2dpq]$$

$$= q[a + d(q - p)] \qquad 2.11$$

Similarly, the average effect of the allele $A_2$ is given as:

$$\alpha_2 = -p[a + d(q - p)] \qquad 2.12$$

The average change in value due to an allelic substitution ('the average effect due to substitution' from alleles $A_2$ to $A_1$) is:

$$\alpha = \alpha_1 - \alpha_2 = a + d(q - p) \qquad 2.13$$

On comparing equations 2.11, 2.12 and 2.13, one gets:

$$\alpha_1 = q\alpha \text{ and } \alpha_2 = -p\alpha \qquad 2.14$$

The breeding value of an individual can be defined as a value associated with the genes carried by the individual and transmitted to its offspring. Thus, based on the explanation of the average allele effects, the breeding value of an individual is equal to the average effect of the genes it carries, i.e. the sum of effects over both alleles at each locus and over all loci (Falconer and Mackay, 1996). Thus, for a single locus with two alleles the breeding values of the genotypes are given in Table 2.7:

Table 2.7. Breeding values for a single locus.

| **Genotype** | **$A_1A_1$** | **$A_1A_2$** | **$A_2A_2$** |
|---|---|---|---|
| Breeding value | $2q\alpha$ | $(q - p)\alpha$ | $-2p\alpha$ |
| Adjusted breeding value | $\alpha$ | 0 | $-\alpha$ |

Dekkers (1999) suggested that when selection is within a generation, breeding values can for simplicity be deviated from the breeding value of the heterozygote without changing the ranking of individuals by subtracting $(q - p)\alpha$. Thus, the adjusted breeding values (Table 2.7) for each genotype will be equal to $+\alpha$, 0 and $-\alpha$. From equation 2.13 it is clear that in case $d = 0$, then $\alpha$ will be equal to '$a$'.

**2.2.2. Simulation of the true breeding values and phenotypic values**

The true breeding value of an individual is usually expressed as a sum of the average effect of allele substitutions over all QTL. In honey bees, a colony trait e.g. honey production, wax production, aggressiveness is comparable to maternally influenced traits in mammals such as birth and weaning weight. Thus, it can be partitioned into the maternal additive genetic effects of the queen and the direct additive genetic effects of the progeny workers. Therefore, the total true breeding of a queen is the sum of the maternal and direct true breeding values. A strategy to simulate the maternal and direct true breeding values for an individual from its QTL information is presented below.

As described earlier (Sub-section 2.1.1.4), a total of 250 QTL were simulated across the genome. To model the maternal and direct effects, it was assumed that of the total 250 QTL, 86 loci controlled direct effects, 78 pleiotropic loci controlled both direct and maternal effects and the remaining 86 loci controlled maternal effects (Figure 2.6). This scheme for distributing QTL allowed simulating the maternal and direct true breeding values in a simple way.

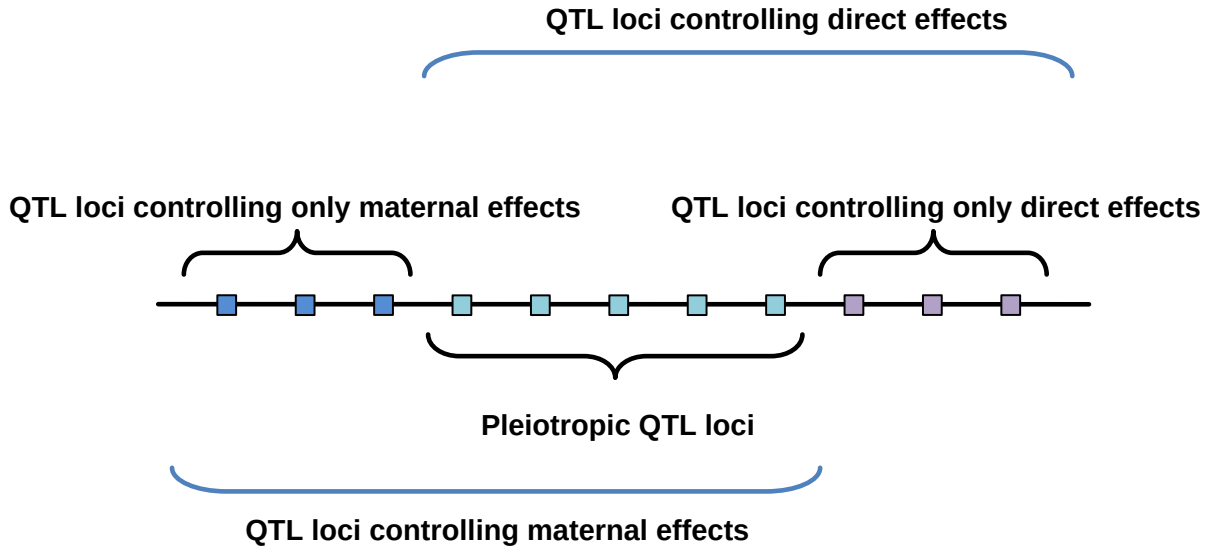


Figure 2.6. A scheme for distributing QTL to simulate maternal and direct effects.

The maternal and direct true breeding value of a queen was taken as a sum of the average effect of allele substitution over all QTL controlling maternal and direct effects, respectively. True breeding values for maternal ($TBV_q$) and direct effects ($TBV_w$) for a queen were calculated using the formula $TBV_q^i = \sum_j q_q^{ij} a^j$ and $TBV_w^i = \sum_k q_w^{ik} a^k$ where $TBV_q^i$ and $TBV_w^i$ are the maternal and direct true breeding values for the $i^{th}$ queen. $q_q^{ij}$ and $q_w^{ik}$ are QTL genotypes of the $i^{th}$ queen at the $j^{th}$ and $k^{th}$ QTL controlling maternal and direct effects, respectively. It has a value of -1 or 1 for homozygous genotypes and 0 for the heterozygous genotypes. $a^j$ and $a^k$ are average effects of allele substitution at the $j^{th}$ and $k^{th}$ QTL, respectively. Maternal and direct true breeding values were simulated for all dam queens of the base population and all potential-dam queens in generations 1 to 5.

The sum of maternal and direct true breeding values gave the overall true breeding value of a queen. The phenotype of each queen was obtained by adding its overall true breeding value to a residual value drawn from a normal distribution $N(0, \sigma_e^2)$.

### 2.2.3. Simulation of correlation between maternal and direct effects

Studies in the honey bee (Bienefeld and Pirchner, 1991; Ehrhardt and Bienefeld, unpublished data) have shown that there is a strong negative correlation between maternal and direct effects. To model this negative correlation, signs of QTL effects for maternal and direct effects were chosen opposite to each other at the pleiotropic loci. The level of negative correlation was determined by the number of pleiotropic loci. No correlation between maternal and direct effects was obtained by randomly choosing signs for QTL effects for maternal and direct effects. The estimate of the simulated value of correlation ($r_{qw}$) was obtained by calculating the correlation between the maternal and direct true breeding values which is the ratio of the covariance between maternal and direct effects to the product of the standard deviations of maternal and direct effects, $r_{qw} = \frac{\sigma^2_{qw}}{\sigma_q \sigma_w}$. To the best of the knowledge, this is the first study that analysed the impact of using marker information on the accuracy of estimating breeding values for traits with a negative correlation between maternal and direct effects.

### 2.2.4. Variance component and heritability of trait

The variance of residual effects ($\sigma^2_e$) and the total genetic effects ($\sigma^2_g$) determine the variance of phenotypes ($\sigma^2_p$). The estimation of these variance components (Searle *et al*., 1992) is essential to the estimation of breeding values. Moreover, they are also required for estimating heritability of different traits and the correlation between them. The estimation of variances is based on algorithms such as maximum-likelihood (ML) (Harville, 1977), restricted maximum likelihood (REML) (Patterson and Thompson, 1971) and average information REML (AI-REML) (Gilmour *et al*., 1995; Johnson and Thompson, 1995). These algorithms are implemented using different software. With real datasets, the amount of information is very large and it is desirable to use these efficient algorithms for estimating breeding values. For this study, the variances were calculated directly from the simulated dataset. It was assumed that the estimates of breeding values did not differ significantly from the estimates obtained using these algorithms. Nevertheless, for estimating breeding values, software optimised for the special pedigree and numerator relationship matrix of the honey bee should be developed in future.

### *2.2.4.1. Simulation of genetic variances*

**Variance and covariance of maternal and direct genetic effects**

For the calculation of the breeding values, estimates of the variance of maternal ($\sigma_q^2$) and direct ($\sigma_w^2$) effects and the covariance ($\sigma_{qw}$) between maternal and direct effects are required. Thus, the variance of maternal and direct effects were obtained by calculating the variance of the simulated maternal breeding values and direct true breeding values over all the generations (including the base population). Similarly, the covariance between maternal and direct effects was obtained by calculating the covariance between the maternal and direct true breeding values. The general formula is summarized as follows:

$$\sigma_q^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(TBV_q^i)^2 - \frac{\left(\sum_{i=1}^{n} TBV_q^i\right)^2}{n}\right] \quad 2.15$$

$$\sigma_w^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(TBV_w^i)^2 - \frac{\left(\sum_{i=1}^{n} TBV_w^i\right)^2}{n}\right] \quad 2.16$$

$$\sigma_{qw} = \frac{1}{n-1}\left[\sum_{i=1}^{n} TBV_q^i TBV_w^i - \frac{\sum_{i=1}^{n} TBV_q^i \sum_{i=1}^{n} TBV_w^i}{n}\right] \quad 2.17$$

**Total genetic variance - Weights for maternal and direct effects**

Assuming that there are '$n$' correlated random variables $X_1, X_2, X_3 \cdots X_n$ and $a_1, a_2, a_3 \cdots a_n$ are '$n$' constants. Let $Y$ represent the weighted sum of random variables, then a general expression for the variance of the weighted sum of correlated random variables is given as (Albright *et al*., 2011):

$$\text{var}(Y) = \sum_{i=1}^{n} a_i^2 \,\text{var}(X_i) + \sum_{i<j} 2a_i a_j \,\text{cov}(X_i, X_j)$$

Similarly, the variance of total genetic effects with two dependent random variables can be written as:

$$\sigma_g^2 = wt_q^2\sigma_q^2 + wt_w^2\sigma_w^2 + 2wt_q wt_w\sigma_{qw} \qquad 2.18$$

where $wt_q$ and $wt_w$ are the weights for maternal and direct effects.

It is important to determine the weights that can be assigned to maternal and direct effects. Usually a breeding value is defined as twice the expected deviation of an individual's progeny mean value from the population mean value, or twice the 'transmitting ability' of an individual (Falconer and Mackay, 1996; Hamilton, 2009). If one considers a complete colony as 'offspring' of a queen, then this colony comprises a daughter (the queen) and a family of granddaughters (the workers). These animals express one-half of the mother's maternal breeding value and one-fourth of the grand-dam's direct breeding value. In this case, the breeding value would be defined as twice the one-half of the maternal breeding value of a queen plus twice the one-fourth of its direct breeding value (i.e. two times the expected deviation of progeny's mean value from the population mean value, provided all other relatives have average breeding values of zero). Thus, the maternal and direct breeding values get a weight of 1 and 0.5, respectively. The total genetic variance ($\sigma_g^2$) will be $\sigma_q^2 + 0.25\sigma_w^2 + \sigma_{qw}$ (the latter from $2\times1\times0.5\times\sigma_{qw}$). However, for the sake of easy comparison and interpretation, the breeding value was taken as a sum of the direct and maternal breeding values of a queen in a manner analogous to mammals. Thus, for this study, equal weights (= 1) were assigned to both maternal and direct effects, and the total genetic variance was taken as a sum of variance of maternal effects, direct effects and twice the covariance between them ($\sigma_g^2 = \sigma_q^2 + \sigma_w^2 + 2\sigma_{qw}$).

#### *2.2.4.4. Simulation of phenotypic variance, residual variance and maternal and direct heritability*

A colony trait in honey bee is determined by the heritability of maternal ($h_m^2$) and direct effects ($h_d^2$). Table 2.8 shows the values of simulated maternal heritability and achieved direct heritability at different correlations between maternal and direct effects. In this study, a fixed maternal heritability of 0.15, 0.25 and 0.35 was simulated that can be expressed as the ratio of the variance of maternal (queen) effects to the phenotypic variance and is given as follows:

$$h_m^2 = \frac{Variance\ of\ \ maternal\ effect}{Phenotypic\ variance} = \frac{\sigma_q^2}{\sigma_e^2 + \sigma_g^2} = \frac{\sigma_q^2}{\sigma_p^2} \qquad 2.19$$

After rearranging this expression, one obtains $\sigma_p^2 = \frac{\sigma_q^2}{h_m^2}$. Thus, the phenotypic variance ($\sigma_p^2$) was obtained from the ratio of the variance of maternal effects ($\sigma_q^2$) and a fixed value of the maternal heritability ($h_m^2$).

Table 2.8. Heritability of direct effects at different values of simulated heritability of maternal effects and correlation between maternal and direct effects.

| **Simulated heritability for the maternal effect** | **Correlation between maternal and direct effects** | **Achieved heritability for the direct effect (standard deviation)** |
|---|---|---|
| 0.150 | 0 | 0.162 (0.021) |
| 0.150 | -0.46 | 0.155 (0.022) |
| 0.250 | 0 | 0.270 (0.035) |
| 0.250 | -0.46 | 0.259 (0.037) |
| 0.350 | 0 | 0.377 (0.049) |
| 0.350 | -0.46 | 0.362 (0.051) |

If a linear mixed model is applied, then the variance of observations is equal to $\mathbf{ZRZ}'+\mathbf{I}\sigma_e^2$ with $\mathbf{Z} = [\mathbf{Z_1}\ \mathbf{Z_2}]$, where $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are design matrices for direct and maternal genetic effects and **I** is the identity matrix. Here, **R** is the Kronecker product of the relationship matrix (e.g. numerator relationship matrix derived from pedigree or combined relationship matrix derived from pedigree and genomic data) with a 2 by 2 matrix **S**, where $s_{11}$ is the direct genetic variance, $s_{22}$ is the maternal genetic variance and $s_{12}$ $(= s_{21})$ is the covariance. Given the above relationship, the variance of phenotypes ($\sigma_p^2$) is a sum of the variance of residual effect ($\sigma_e^2$) and the total genetic effect ($\sigma_g^2$) which can be expressed as $\sigma_e^2 + \sigma_g^2$. Based on this relationship, the residual variance

was obtained by subtracting the total genetic variance from the phenotypic variance and can be written as $\sigma_e^2 = \sigma_p^2 - \sigma_g^2$.

The ratio of the variance of direct effects to the phenotypic variance provided a measure of the heritability of direct (worker) effects and is as given below.

$$h_d^2 = \frac{Variance\ of\ \ direct\ effect}{Phenotypic\ variance} = \frac{\sigma_w^2}{\sigma_e^2 + \sigma_g^2} = \frac{\sigma_w^2}{\sigma_p^2} \qquad 2.20$$

## 2.3. Future Ideas: Methods for obtaining approximate genotypic information for an average worker

### 2.3.1. Genotyping a sample of workers from a colony

The average genotyping information for an average worker of a colony can be obtained through genotyping a sample of workers from each colony. An example for obtaining an average genotype on the basis of 10 marker loci for a sample of five workers from a colony is provided below:

Table 2.9. Fictitious genotyping information for five workers at 10 loci.

| | **L1** | **L2** | **L3** | **L4** | **L5** | **L6** | **L7** | **L8** | **L9** | **L10** |
|---|---|---|---|---|---|---|---|---|---|---|
| Genotype worker 1 | 11 | 12 | 12 | 21 | 11 | 22 | 22 | 22 | 12 | 11 |
| Genotype worker 2 | 22 | 12 | 22 | 11 | 11 | 21 | 22 | 21 | 21 | 11 |
| Genotype worker 3 | 11 | 22 | 22 | 22 | 11 | 21 | 12 | 12 | 12 | 21 |
| Genotype worker 4 | 22 | 12 | 22 | 12 | 21 | 11 | 22 | 22 | 11 | 21 |
| Genotype worker 5 | 21 | 12 | 22 | 22 | 12 | 11 | 21 | 12 | 11 | 21 |
| Average frequency of the first allele A1 | 5/10 | 4/10 | 1/10 | 4/10 | 8/10 | 6/10 | 2/10 | 3/10 | 7/10 | 7/10 |
| Average frequency of the second allele A2 | 5/10 | 6/10 | 9/10 | 6/10 | 2/10 | 4/10 | 8/10 | 7/10 | 3/10 | 3/10 |

*Loci are denoted by L1-L10

As shown in Table 2.9, the average genotype can be estimated from the frequency of alleles in a sample of workers that are genotyped.

**2.3.2. Genotyping information based on the approach of Israel and Weller**

Israel and Weller (1998) derived the probabilities of the genotype when only certain individuals in a pedigree are genotyped. Regarding the special case of honey bees where only queens can be easily genotyped, this methodology could be used to derive genotype probabilities at all marker loci for an average worker. The methodology proposed by Israel and Weller (1998) has been summarized as follows.

For individuals without genotype information available, the probability of the corresponding genotypes of a locus Q are $P_Q^2$, $P_q^2$ and $2P_QP_q$. For individuals that have not been genotyped but are progeny of genotyped individuals, the probability of the corresponding genotypes are $\left(P_Q^sP_Q^d\right),\left(P_Q^sP_q^d+P_q^sP_Q^d\right)$ and $\left(P_q^sP_q^d\right)$ where $P_Q^s$ and $P_q^s$ are the probabilities of the two alleles of the sire and $P_Q^d$ and $P_q^d$ are the probabilities of the two alleles of the dam of the individual under consideration. The probabilities of Q and q for the parents can be derived from $P_{QQ}+0.5P_{Qq}$ and $0.5P_{Qq}+P_{qq}$ where $P_{QQ}$, $P_{Qq}$ and $P_{qq}$ are the probabilities that the parent's genotype is QQ, Qq or qq. The probability of the two alleles is 1 and 0 for a homozygote or 0.5 for a heterozygote in case parent's genotype is available. If parents are not genotyped, then the probabilities of the two alleles are assumed to be equal to the allele frequencies in the population. If the parent is not genotyped, however some of its ancestors are genotyped, then the probability of its possible genotypes is computed from the genotyping information of its ancestors using the formula described earlier. Similar to the example pedigree from Israel and Weller (1998), a small example pedigree for the honey bee is shown below (Table 2.10) with the genotype probability for one locus. It is assumed that only dam queens in the 1$^{st}$ generation are genotyped and generations are overlapping.

Table 2.10. Probability of genotypes for ungenotyped animals derived from genotyped ancestors.

| | | | | | Genotype Probability | | |
|---|---|---|---|---|---|---|---|
| **Generation** | **Animal ID** | **Dam Queen** | **Dummy Sire*** | **Genotype** | **QQ** | **Qq** | **qq** |
| Gen 1 | Q1 | 0 | 0 | qq | 0 | 0 | 1 |
| | Q2 | 0 | 0 | Qq | 0 | 1 | 0 |
| | Q3 | 0 | 0 | qq | 0 | 0 | 1 |
| | Q4 | 0 | 0 | Qq | 0 | 1 | 0 |
| | Q5 | 0 | 0 | Qq | 0 | 1 | 0 |
| | Q6 | 0 | 0 | QQ | 1 | 0 | 0 |
| | DS1 | 0 | 0 | - | 0.49 | 0.42 | 0.09 |
| | DS2 | 0 | 0 | - | 0.49 | 0.42 | 0.09 |
| | AvgW1 | Q1 | 0 | - | 0 | 0.8 | 0.2 |
| | AvgW2 | Q2 | 0 | - | 0.4 | 0.5 | 0.1 |
| | AvgW3 | Q3 | 0 | - | 0 | 0.8 | 0.2 |
| | AvgW4 | Q4 | 0 | - | 0.4 | 0.5 | 0.1 |
| | AvgW5 | Q5 | 0 | - | 0.4 | 0.5 | 0.1 |
| | AvgW6 | Q6 | 0 | - | 0.8 | 0.2 | 0 |
| Gen 2 | Q7 | Q1 | 0 | - | 0 | 0.8 | 0.2 |
| | Q8 | Q6 | 0 | - | 0.8 | 0.2 | 0 |
| | Q9 | Q2 | 0 | - | 0.4 | 0.5 | 0.1 |
| | Q10 | Q3 | 0 | - | 0 | 0.8 | 0.2 |
| | Q11 | Q1 | 0 | - | 0 | 0.8 | 0.2 |
| | Q12 | Q2 | 0 | - | 0.4 | 0.5 | 0.1 |
| | DS3 | Q3 | 0 | - | 0 | 0.8 | 0.2 |
| | DS4 | Q6 | 0 | - | 0.8 | 0.2 | 0 |
| | AvgW7 | Q7 | DS1 | - | 0.28 | 0.54 | 0.18 |
| | AvgW8 | Q8 | DS1 | - | 0.63 | 0.34 | 0.03 |
| | AvgW9 | Q9 | DS2 | - | 0.455 | 0.44 | 0.105 |
| | AvgW10 | Q10 | DS2 | - | 0.28 | 0.54 | 0.18 |
| | AvgW11 | Q11 | DS1 | - | 0.28 | 0.54 | 0.18 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | AvgW12 | Q12 | DS2 | - | 0.455 | 0.44 | 0.105 |
| Gen 3 | Q13 | Q9 | DS2 | - | 0.455 | 0.44 | 0.105 |
| | Q14 | Q11 | DS1 | - | 0.28 | 0.54 | 0.18 |
| | Q15 | Q7 | DS1 | - | 0.28 | 0.54 | 0.18 |
| | Q16 | Q10 | DS2 | - | 0.28 | 0.54 | 0.18 |
| | Q17 | Q7 | DS1 | - | 0.28 | 0.54 | 0.18 |
| | Q18 | Q10 | DS2 | - | 0.28 | 0.54 | 0.18 |
| | DS5 | Q10 | DS2 | - | 0.28 | 0.54 | 0.18 |
| | DS6 | Q11 | DS1 | - | 0.28 | 0.54 | 0.18 |
| | AvgW13 | Q13 | DS4 | - | 0.6075 | 0.36 | 0.0325 |
| | AvgW14 | Q14 | DS4 | - | 0.495 | 0.46 | 0.045 |
| | AvgW15 | Q15 | DS3 | - | 0.22 | 0.51 | 0.27 |
| | AvgW16 | Q16 | DS3 | - | 0.22 | 0.51 | 0.27 |
| | AvgW17 | Q17 | DS4 | - | 0.495 | 0.46 | 0.045 |
| | AvgW18 | Q18 | DS3 | - | 0.22 | 0.51 | 0.27 |

Gen denotes the generation; Q denotes a queen; DS denotes a dummy sire; AvgW denotes an average worker.

*A queen mates with a dummy sire that belongs to the generation that precedes the queen's generation. Therefore, the allele frequency for the dummy sire mating with a queen is taken from the generation preceding the queen's generation. In this example, frequencies of alleles Q and q in the population for fictitious dummy sires that mates with queens in generation 1 are assumed to be 0.8 and 0.2, respectively. Similarly, frequencies of alleles Q and q in the population to which dummy sires 1 and 2 belongs (in generation 1) are assumed to be 0.7 and 0.3, respectively.

**2.3.3. Genotyping information based on the approach of Gengler *et al*.**

According to Gengler *et al*. (2007), the method given by Israel and Weller (1998) ignores information on descendents. Gengler *et al*. (2007) proposed a methodology to derive the conditional expectation of gene content of ungenotyped animals, given molecular data from genotyped individuals and pedigree data. The proposed methodology could also be adapted for honey bees, and can be especially helpful when the genomic matrix is to be constructed for ungenotyped animals (workers, dummy sires as well as ungenotyped queens) in the pedigree. The methodology proposed by Gengler *et al*. (2007) has been summarized as follows.

Let $q$ be the gene content at a particular locus, $\mu$ be the population mean and $d$ be the deviation of gene content from the population mean i.e. $d = q - \mu$. It is assumed that the gene content is a continuous variable, the relationship between gene contents is linear and the covariance between gene contents is proportional to the additive genetic relationship between individuals. Based on these assumptions, the conditional expectation of gene contents for all ungenotyped individuals ($\mathbf{q_x}$) given genotyping and pedigree data can be derived using the expression given below:

$$\mathbf{q_x} = \begin{pmatrix} \mathbf{1} & \mathbf{A_{xy}A_y^{-1}} \end{pmatrix} \begin{pmatrix} \mu \\ \mathbf{q_y} - \mathbf{1}\mu \end{pmatrix} \qquad 2.21$$

where $\mathbf{q_x}$ is a vector of gene contents for ungenotyped individuals, $\mathbf{q_y}$ is a vector of gene contents for genotyped individuals, $\mathbf{A_{xy}}$ is the additive genetic relationship matrix between ungenotyped and genotyped individuals and $\mathbf{A_y}$ is the additive genetic relationship matrix for genotyped individuals.

To provide a more accurate and practical method of computation, Gengler *et al*. (2007) incorporated the probability of errors in marker phenotypes under the incomplete penetrance model. This was achieved by assuming an error variance ($\sigma_e^2$) in the variability of **q**. The following mixed model[7] equations were proposed to obtain the solution for ungenotyped individuals:

$$\begin{pmatrix} \mathbf{1'1} & \mathbf{1'M} \\ \mathbf{M'1} & \mathbf{M'M} + \mathbf{A}^{-1}\varepsilon \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{d}}_\mathbf{y} \\ \hat{\mathbf{d}}_\mathbf{x} \end{pmatrix} = \begin{pmatrix} \mathbf{1'q_y} \\ \mathbf{M'q_y} \end{pmatrix} \qquad 2.22$$

Here, $\hat{\mu}$, $\hat{\mathbf{d}}_\mathbf{y}$ and $\hat{\mathbf{d}}_\mathbf{x}$ stand for the estimated values. $\mathbf{d_y}$ is a vector of deviations for genotyped individuals, $\mathbf{d_x}$ is a vector of deviations for ungenotyped individuals, **M** is the incidence matrix

[7] Mixed model is a statistical model containing both fixed and random effects.

relating $\mathbf{q}_y$ to $\begin{pmatrix}\mathbf{d}_y\\ \mathbf{d}_x\end{pmatrix}$ and can be written as $\begin{pmatrix}\mathbf{1}_y & \mathbf{0}_x\end{pmatrix}$, $\mathbf{A}$ is the numerator relationship matrix with $\mathbf{A}=\begin{pmatrix}\mathbf{A}_{yy} & \mathbf{A}_{yx}\\ \mathbf{A}_{xy} & \mathbf{A}_{xx}\end{pmatrix}$ and $\varepsilon=\frac{\sigma_e^2}{\sigma_d^2}$.

# Chapter 3. A new approach of genetic evaluation in the honey bee

Genetic evaluation is crucial for the genetic improvement of agriculturally important animals; as a result, various approaches were employed over the last several years for selecting genetically superior animals. In line with this fact, a chronological review of the progress of genetic evaluation in honey bees along with an introduction to the BLUP methodology and mixed model equations is provided at the beginning of this chapter. Section 3.3 explains the influence of maternal effects on the phenotype which forms a part of the linear model for genetic evaluation in honey bees. Finally, Sections 3.4 presents the implementation of the following genetic evaluation methods: (1) traditional approach using pedigree and phenotypic data (PED_BLUP) and (2) the unified approach using marker, pedigree and phenotypic data (UNI_BLUP). In addition, ideas on improving the numerator relationship matrix in the honey bee are discussed in Section 3.5.

## 3.1. Review: The progress of genetic evaluation in honey bees

Improvement for traits such as honey yield, swarming tendency, calmness and resistance to *Varroa* can be achieved by selecting animals with higher breeding values. However, as compared to other animals, the genetic evaluation is more complex in the honey bee. It requires taking into account the negative genetic correlation between maternal and direct effects (Bienefeld and Pirchner, 1991) and the peculiar genetic and reproductive biology of the honey bee. Nevertheless, genetic evaluation is crucial to the genetic improvement programs of honey bees.

In 1982, Chevalet and Cornet proposed a genetical model based on selection index [8] that was analogous to models used for maternally influenced traits in mammals. A collective value was given to a colony that was assumed to be equivalent to the sum of the queen contribution and of an average contribution of the workers. This model took into account the peculiarities of the

[8] It is a method of estimating the breeding values using all available information on the individual and its relatives (Mrode, 2005).

honey bee reproduction and genetics. Based on the selection index, Cornuet and Chevalet (1987) illustrated a selection scheme for the improvement of honey production in the stock.

Rinderer (1986) proposed the following expression for the selection index in the honey bee with respect to 'two traits':

$$I = Z_1 V\left(\frac{h_1^2}{h_2^2}\right) + Z_2\left(1 - r_g\right) \qquad 3.1$$

where $Z_1$ is the z-score for trait 1, $Z_2$ is the z-score for trait 2, V is the economic value of trait 1 relative to trait 2, $h_1^2$ is the heritability of trait 1, $h_2^2$ is the heritability of trait 2 and $r_g$ is the genetic correlation that exists between traits 1 and 2. The z-score is obtained by using the expression:

$$Z = \frac{X - M}{s} \qquad 3.2$$

where $X$ is the colony's score, $M$ is the apiary's average score and $s$ is the standard deviation of the apiary's scores.

Under the assumption that accurate heritability and correlation estimates are unavailable, van Engelsdorp and Otis (2000) used an approximation of the equation given by Rinderer (1986) to provide a modified selection index for several traits for honey bees. A survey of commercial beekeepers was used to estimate the economic values (V-values) of several traits of the honey bee colonies. The expression is given as follows:

$$I_{modified} = \sum_{1}^{n} Z_i V_i \text{ where } n \text{ is the number of traits.} \qquad 3.3$$

Bienefeld and Pirchner (1991) also derived a selection index for several traits which simultaneously considered queen and worker effects. Bienefeld *et al*. (2007) pointed out that the use of selection indices is becoming less common, as environmental influences cannot be sufficiently corrected for. In 2007, Bienefeld *et al*. proposed a BLUP based methodology for

genetic evaluation in the honey bee, which is widely used in other agricultural species. The methodology was adjusted to the peculiar reproductive behaviour and genetics of the honey bee. A shortcoming of the selection index was that records had to be pre-adjusted for fixed effects, however in BLUP, both fixed effects of environment and random genetic effects are estimated simultaneously, and the differences due to fixed effects such as apiary or bee breeder are accounted for. Moreover, during genetic evaluation, the genetic merits of all relatives plus the animal's own performance are used to estimate the animal's genetic merit. Genetic evaluation based on BLUP can also be conveniently applied to complex multivariate models. This is contributed to the advancement in computational methodologies and availability of advanced software such as PEST (Groeneveld *et al*., 1990 http://www.fli.bund.de/no_cache/de/startseite/institute/institut-fuer-nutztiergenetik/wissenschaftler/dr-dr-eildert-groeneveld.html), BREEDPLAN for beef cattle (http://www.breedplan.une.edu.au/) and MiXBLUP (http://www.mixblup.eu/).

Until now, the BLUP based genetic evaluation in the honey bee has relied on pedigree and phenotypic data. It has been shown in several studies (Dekkers and Hospital, 2002; de Roos *et al*., 2007; Calus *et al*., 2008; Legarra *et al*., 2008; Sonesson and Meuwissen, 2009) that the incorporation of molecular genetic data can significantly improve the accuracy of the estimates of breeding values, increase the genetic response and lower the rate of inbreeding. In this thesis, the impact of incorporating of SNP marker into the genetic evaluation in the honey bee was analysed. The integration of marker data was achieved through the 'unified approach' proposed by Legarra *et al*. (2009) and Christensen and Lund (2010) which combines full pedigree and genomic information from both genotyped and ungenotyped individuals. The unified approach provides a straightforward extension to the BLUP methodology for the estimation of breeding values. It is also advantageous for honey bees as genomic information for genotyped queens can be integrated with pedigree information from all animals.

### 3.2. Development of BLUP and the mixed model equations by Henderson

The BLUP methodology developed by Henderson (1950) allowed estimating fixed and random effects simultaneously. It is widely used for the estimation of breeding values in important animal and plant species. The acronym BLUP summarizes the properties of this methodology which are as follows (Henderson, 1975; Robinson, 1991; Mrode, 2005): (1) Best - correlation

between the true ($u$) and predicted breeding value ($\hat{u}$) is maximized or in other word, the prediction error variance (PEV) is minimized, (2) Linear - estimates of the random variables $u$ are linear functions of the data, $y$, (3) Unbiased - the average value of the estimate is equal to the average value of the quantity being estimated, i.e. E($\hat{u}$ = u) and (4) Predictor – it is predicting the true breeding values. Consider the following linear model (Henderson, 1975, 1984):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad 3.4$$

where **y** is a vector of records, **b** is a vector of fixed effects, **u** is a random vector for additive genetic effects, i.e. breeding values, **e** is a random vector of residual effects and **X** and **Z** are known design matrices. The expected value of the variables are $\mathrm{E}(\mathbf{y}) = \mathbf{Xb}$ and $\mathrm{E}(\mathbf{u}) = \mathrm{E}(\mathbf{e}) = 0.$ The variances are $\mathrm{var}(\mathbf{u}) = \mathbf{A}\sigma_u^2 = \mathbf{G}$ where **A** is the numerator relationship matrix and $\mathrm{var}(\mathbf{e}) = \mathbf{I}\sigma_e^2 = \mathbf{R}.$ It is assumed that $\mathrm{cov}(\mathbf{u},\mathbf{e}) = 0.$ The variance for **y** is given as $\mathrm{var}(\mathbf{y}) = \mathbf{V} = \mathrm{var}(\mathbf{Zu} + \mathbf{e}) = \mathbf{ZGZ}' + \mathbf{R}$. The covariance of **y** and **u** is $\mathrm{cov}(\mathbf{y},\mathbf{u}) = \mathrm{cov}(\mathbf{Zu} + \mathbf{e},\mathbf{u}) = \mathbf{ZG}$ and the covariance of and **y** and **e** is $\mathrm{cov}(\mathbf{y},\mathbf{e}) = \mathrm{cov}(\mathbf{Zu} + \mathbf{e},\mathbf{e}) = \mathbf{R}$.

For the selection index, it is assumed that **b** is known whereas in reality, certain elements of **b** may be completely unknown. Therefore, **Xb** must be estimated by some method, such as least squares, to obtain a selection index (Henderson, 1963). Thus, the accuracy of the selection index depends upon the choice of the estimate of **b**. Another big drawback of the selection index evaluation is that not all records are used for estimating breeding values. To overcome these problems, Henderson (1963) derived the BLUP methodology in which the predictions are unaffected by **b**. This requires predicting a linear function of **b** and **u**, i.e. **k'b + u**, using a linear function of **y**, i.e. **L'y**, given that **k'b** is estimable (Mrode, 2005). The predictor **L'y** is chosen so that it is unbiased (implying that the average value of the estimate is equal to the average value of the quantity being estimated) and 'best' in the sense that the PEV is minimal (Mrode, 2005). The BLUP of **u** and the best linear unbiased estimate (BLUE) of **b** is given as:

$$\mathrm{BLUP}(\mathbf{u}) = \hat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) \qquad 3.5$$

$$\mathrm{BLUE}(\mathbf{b}) = \hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \text{ where } \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-} \text{ stands for the generalized inverse} \qquad 3.6$$

$$\mathbf{L}'\mathbf{y} = \mathbf{k}'\hat{\mathbf{b}} + \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}\right) \qquad 3.7$$

As per convention, BLUE refers to the estimates of fixed effects whereas BLUP to the estimates of random effects (Henderson, 1975) although BLUE and BLUP are similar in properties.

The solutions for **u** and **b** in 3.5 and 3.6 require the inverse $\mathbf{V}^{-1}$, which cannot be computed in some cases. Henderson (1950) presented a method to estimate solutions for the fixed effects (**b**) and to predict solutions for random effects (**u**) simultaneously without having to compute $\mathbf{V}^{-1}$. This was realized by maximizing the joint density of **y** and **u** for variations in **b** and **u** under the assumption of normality (Henderson, 1984). The resulting equations are known as the mixed model equations (MME) and are given as follows:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \qquad 3.8$$

$\mathbf{R}^{-1}$ can be factorized from both sides of the equation to give:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \qquad 3.9$$

Henderson *et al*. (1959) proved that $\hat{\mathbf{b}}$ from the MME (3.9) is the same as $\hat{\mathbf{b}}$ from Equation 3.6 i.e., a GLS solution. Similarly, Henderson (1963) proved that $\hat{\mathbf{u}}$ from MME (3.9) is the same as $\hat{\mathbf{u}}$ from Equation 3.5 i.e., BLUP.

It should be noted that BLUP is equivalent to the selection index method when $\hat{\mathbf{b}}$ is substituted with **b**. Thus, the mixed model equations with **b** assumed known provides a convenient method for obtaining the selection index on the basis of all records for estimating breeding values.

Henderson (1984) explained that the MME is easier to compute because of the following properties. Firstly, $\mathbf{R}^{-1}$ is easier to obtain than $\mathbf{V}^{-1}$, even though both have the same dimension. This is because **R** has a simpler form such as $\mathbf{I}\sigma_e^2$ or is block diagonal. Secondly, $\mathbf{G}^{-1}$ is simpler to compute ($\mathbf{G} = \mathbf{A}\sigma_u^2$) as $\mathbf{A}^{-1}$ can be computed directly without constructing the **A** matrix

(Henderson, 1976). Lastly, the coefficient matrix usually exhibits diagonal or block diagonal dominance thereby resulting in equations that are well suited to iterative solution.

## 3.3. Maternally influenced traits

A mother's ability to provide a suitable environment, e.g. better nutrition, contributes to the phenotypic expression of some traits in the offspring. This ability is partly genetic and partly environmental (Willham, 1963, 1972). Consequently, the resulting maternally influenced trait in the offspring is a sum of the direct genetic effects due to the individual, maternal genetic effects due to its dam and environmental effects. In livestock species, several traits are maternally influenced, e.g. birth and weaning weight in beef cattle. A similar situation exists in honey bees where 'colony traits', e.g. honey yield, wax production and calmness, result from the activity of the queen (contributing to maternal effects) and several thousand workers inhabiting the colony (contributing to direct effects).

A complexity associated with the estimation of breeding values for maternally influenced traits is that the maternal and direct effects are usually negatively correlated (Bienefeld and Pirchner, 1991; Larsgard and Olesen, 1998; Splan *et al*., 2002; Safari *et al*., 2005) which severely impedes the response to selection (Willham, 1972; Roehe and Kennedy, 1993; Mousseau and Fox, 1998; Räsänen and Kruuk, 2007). Since the use of marker information improves the accuracy of the estimation of breeding values, it can be particularly advantageous in case of maternally influenced traits with negative correlation between maternal and direct effects. Therefore, this study also assesses the impact of marker information on traits with negative correlation and no correlation between maternal and direct effects.

## 3.4. Method of genetic evaluation

### 3.4.1 Traditional pedigree based approach (PED_BLUP)

Currently, genetic evaluation in the honey bee is performed using a BLUP approach which requires pedigree and phenotypic data. The following sections describe the method of genetic evaluation through the construction of a numerator relationship matrix and the associated linear mixed model equations.

***3.4.1.1. Construction of the numerator relationship matrix***

The genetic evaluation based on BLUP makes use of the numerator relationship matrix (**A**) which indicates the 'additive genetic relationship' between individuals (Falconer and Mackay, 1996). For the prediction of breeding values, the inverse $\mathbf{A}^{-1}$ is required which can be obtained through different algorithms proposed by Henderson (1976), Meuwissen and Luo (1992) and Quaas (1976, 1995).

The **A** matrix was constructed for all individuals in the pedigree. The elements of the **A** matrix were calculated according to the method developed by Bienefeld *et al*. (2007) for honey bees which includes a paternal path coefficient ( $P_p$ ) of 0.367 to account for polyandry. The details for constructing an **A** matrix recursively are given below.

Assuming that $s$ and $d$ denote the indices of the sire and dam of the i$^{th}$ individual, then:

i. If both sire and dam are known

$$a_{ji} = a_{ij} = 0.5a_{jd} + P_p\left(a_{js}\right) \qquad for\ \ j = 1\ to\ (i-1) \tag{3.10}$$

$$a_{ii} = 1 + 0.5\left(a_{sd}\right) \tag{3.11}$$

ii. If sire is known and assumed to be unrelated to the dam

$$a_{ji} = a_{ij} = P_p\left(a_{js}\right) \qquad for\ \ j = 1\ to\ (i-1) \tag{3.12}$$

$$a_{ii} = 1 \tag{3.13}$$

iii. If dam is known and assumed to be unrelated to the sire

$$a_{ji} = a_{ij} = 0.5a_{jd} \qquad for\ \ j = 1\ to\ (i-1) \tag{3.14}$$

$$a_{ii} = 1 \tag{3.15}$$

iv. If both parents are unknown and assumed to be unrelated

$$a_{ji} = a_{ij} = 0 \qquad for\ j = 1\,to\,(i-1) \tag{3.16}$$

$$a_{ii} = 1 \tag{3.17}$$

The **A** matrix can be partitioned into $\mathbf{A_{11}}$, $\mathbf{A_{22}}$, $\mathbf{A_{12}}$ and $\mathbf{A_{21}}$, where subscripts 1 and 2 denote genotyped and ungenotyped individuals. As described by Christensen and Lund (2010), $\mathbf{A}^{-1}$ for the partitioned matrix is given by the following expression:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1} \\ -\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1} \end{bmatrix} \tag{3.18}$$

***3.4.1.2. Mixed model equations***

An approach to include maternal effects into the MME was initially presented by Quaas and Pollak (1980). In 2007, Bienefeld *et al*. adapted the BLUP-animal model with maternal and direct genetic effects for genetic evaluation in the honey bee. The model takes into account the influence of maternal effects on traits and is adapted to the peculiarity of honey bees. The linear model is given as follows:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z_1u_1} + \mathbf{Z_2u_2} + \mathbf{e} \tag{3.19}$$

where **y** is a vector of records/traits of the colonies, **b** is a vector of fixed effects, $\mathbf{u_1}$ is a vector of random direct genetic effects (i.e. breeding values for direct effects), $\mathbf{u_2}$ is a vector of random maternal genetic effects (i.e. breeding values for maternal effects), **e** is a vector of random residual effects, **X** is a known incidence matrix relating observations to the corresponding environmental effects, $\mathbf{Z_1}$ is a known incidence matrix relating observations to the corresponding direct effects and $\mathbf{Z_2}$ is a known incidence matrix relating observations to the corresponding maternal effects.

The variance of **y** is given as follows:

$$\text{var}(\mathbf{y}) = \text{var}(\mathbf{Z_1u_1} + \mathbf{Z_2u_2} + \mathbf{e})$$

$$= \mathbf{Z_1}\,\mathrm{var}(\mathbf{u_1})\mathbf{Z_1}' + \mathbf{Z_2}\,\mathrm{var}(\mathbf{u_2})\mathbf{Z_2}' + \mathrm{var}(\mathbf{e}) + \mathrm{cov}(\mathbf{Z_1u_1}, \mathbf{Z_2u_2}) + \mathrm{cov}(\mathbf{Z_2u_2}, \mathbf{Z_1u_1}) + \mathrm{cov}(\mathbf{Z_1u_1}, \mathbf{e}) + \mathrm{cov}(\mathbf{Z_2u_2}, \mathbf{e}) + \mathrm{cov}(\mathbf{e}, \mathbf{Z_1u_1}) + \mathrm{cov}(\mathbf{e}, \mathbf{Z_2u_2})$$

$$= \mathbf{Z_1}\,\mathrm{var}(\mathbf{u_1})\mathbf{Z_1}' + \mathbf{Z_2}\,\mathrm{var}(\mathbf{u_2})\mathbf{Z_2}' + \mathrm{cov}(\mathbf{Z_1u_1}, \mathbf{Z_2u_2}) + \mathrm{cov}(\mathbf{Z_2u_2}, \mathbf{Z_1u_1}) + \mathrm{var}(\mathbf{e})$$

$$= \mathbf{Z_1}g_{11}\mathbf{A}\mathbf{Z_1}' + \mathbf{Z_2}g_{22}\mathbf{A}\mathbf{Z_2}' + \mathbf{Z_1}g_{12}\mathbf{A}\mathbf{Z_2}' + \mathbf{Z_2}g_{21}\mathbf{A}\mathbf{Z_1}' + \mathbf{I}\sigma_e^2$$

$$= \begin{bmatrix}\mathbf{Z_1} & \mathbf{Z_2}\end{bmatrix}\begin{bmatrix} g_{11}\mathbf{A} & g_{12}\mathbf{A} \\ g_{21}\mathbf{A} & g_{22}\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{Z_1}' \\ \mathbf{Z_2}'\end{bmatrix} + \mathbf{I}\sigma_e^2 \qquad 3.20$$

where $g_{11}$ is the additive genetic variance of direct effects, $g_{22}$ is the additive genetic variance of maternal effects, $g_{12}$ and $g_{21}$ is the additive genetic covariance between direct and maternal effects, $\sigma_e^2$ is the residual error variance and $\mathbf{A}$ is the numerator relationship matrix.

The mixed model equations to obtain the best linear unbiased prediction for $\mathbf{u_1}$ and $\mathbf{u_2}$, and the best linear unbiased estimate for **b** are as follows:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z_1} & \mathbf{X'Z_2} \\ \mathbf{Z_1'X} & \mathbf{Z_1'Z_1} + \mathbf{A}^{-1}\alpha_1 & \mathbf{Z_1'Z_2} + \mathbf{A}^{-1}\alpha_2 \\ \mathbf{Z_2'X} & \mathbf{Z_2'Z_1} + \mathbf{A}^{-1}\alpha_2 & \mathbf{Z_2'Z_2} + \mathbf{A}^{-1}\alpha_3 \end{bmatrix}\begin{bmatrix}\mathbf{b} \\ \mathbf{u_1} \\ \mathbf{u_2}\end{bmatrix} = \begin{bmatrix}\mathbf{X'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y}\end{bmatrix} \qquad 3.21$$

where $\begin{bmatrix}\alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_3\end{bmatrix} = \sigma_e^2\mathbf{G}_{\mathrm{var\_cov}}^{-1} = \sigma_e^2\begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22}\end{bmatrix}$

Here, $\mathbf{G}_{\mathrm{var\_cov}}^{-1} = \begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22}\end{bmatrix}$ and its non-inverse form is $\mathbf{G}_{\mathrm{var\_cov}} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22}\end{bmatrix}$.

Thus, based on the mixed model given above, the estimates of direct ($\mathbf{u_1}$) and maternal ($\mathbf{u_2}$) breeding values are obtained.

### 3.4.2. Unified approach - the integration of genomic, pedigree and phenotypic data (UNI_BLUP)

In the honey bee pedigree, a dummy sire and an average worker represent group of individuals, thus, it is not possible to get genotyping data for all individuals. Using the unified approach can

be advantageous for honey bees because genomic information for genotyped queens can be integrated together with pedigree information resulting in a combined relationship matrix **H**.

***3.4.2.1. Construction of the combined relationship matrix***

The method of construction of the **A** matrix has been described previously in the Sub-section 3.4.1.1. A genomic matrix (**G**) was constructed for genotyped queens consisting of dam queens in the base population and potential-dam queens in the following generations. Based on the approach, the source of information about the allele frequency and/or the use of various scaling parameter, different methods have been proposed for constructing the **G** matrix (VanRaden, 2008; Aguilar *et al*., 2010; Meuwissen *et al*., 2011). A methodology proposed by VanRaden (2008) was employed that has been used in several other studies (Legarra *et al*., 2009; Misztal *et al*., 2009; Aguilar *et al*., 2010; Christensen and Lund, 2010). The **G** matrix was obtained from the formula $\mathbf{ZZ}'/\left[2\sum p_i(1-p_i)\right]$, where **Z** is equal to $\mathbf{M}-\mathbf{P}$. **M** is the matrix specifying marker alleles inherited by each individual and is equal -1, or 1 for the homozygous genotypes and 0 for the heterozygous genotype. **P** is equal to $2(p_i-0.5)$ with $p_i$ being the frequency of second allele at locus $i$ in the 'base population'.

The following example illustrates the construction of the **G** matrix:

| | **Marker 1** | **Marker 2** | **Marker 3** | **Marker 4** |
|---|---|---|---|---|
| Animal 1 | 11 | 12 | 22 | 11 |
| Animal 2 | 11 | 22 | 11 | 12 |
| Animal 3 | 11 | 12 | 22 | 22 |
| $p_i$ | 1/10 | 1/4 | 3/4 | 1/2 |

For this simple example, the **M** and **P** matrices can be constructed as follows:

$$\mathbf{M}=\begin{pmatrix}-1 & 0 & 1 & -1\\ -1 & 1 & -1 & 0\\ -1 & 0 & 1 & 1\end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} 2p_1 - 1 & \cdots & 2p_n - 1 \\ \vdots & \ddots & \vdots \\ 2p_1 - 1 & \cdots & 2p_n - 1 \end{pmatrix} = \begin{pmatrix} -4/5 & -1/2 & 1/2 & 0 \\ -4/5 & -1/2 & 1/2 & 0 \\ -4/5 & -1/2 & 1/2 & 0 \end{pmatrix}$$

Since $\mathbf{Z} = \mathbf{M} - \mathbf{P}$ and $\mathbf{G} = \dfrac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$, the **Z** and **G** matrices for this example are given as follows:

$$\mathbf{Z} = \begin{pmatrix} -0.2 & 0.5 & 0.5 & -1 \\ -0.2 & 1.5 & -1.5 & 0 \\ -0.2 & 0.5 & 0.5 & 1 \end{pmatrix} \text{ and } \mathbf{G} = \begin{pmatrix} 1.08 & 0.03 & -0.32 \\ 0.03 & 3.17 & 0.03 \\ -0.32 & 0.03 & 1.08 \end{pmatrix}$$

The **G** matrix could be singular (VanRaden, 2008; Legarra *et al*., 2009; Aguilar *et al*., 2010); therefore, it was modified with a weighing factor *w* to $\mathbf{G_w}$, given as $\mathbf{G_w} = w\mathbf{G} + (1-w)\mathbf{A_{11}}$. Christensen and Lund (2010) suggested that $(1-w)$ could be interpreted as the relative weight on polygenic effect. For this study, the value of $w$ was taken as 0.99 (Christensen and Lund, 2010).

For computing the **H** matrix, the methodology described by Christensen and Lund (2010) and Legarra *et al*. (2009) was followed and is given as $\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{G_w} - \mathbf{A_{11}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$.

The inverse of the combined relationship matrix ($\mathbf{H}^{-1}$) with integrated pedigree and genomic information was obtained using the following formula:

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{G_w^{-1}} + \mathbf{A_{11}^{-1}A_{12}}\left(\mathbf{A_{22}} - \mathbf{A_{21}A_{11}^{-1}A_{12}}\right)^{-1}\mathbf{A_{21}A_{11}^{-1}} & -\mathbf{A_{11}^{-1}A_{12}}\left(\mathbf{A_{22}} - \mathbf{A_{21}A_{11}^{-1}A_{12}}\right)^{-1} \\ -\left(\mathbf{A_{22}} - \mathbf{A_{21}A_{11}^{-1}A_{12}}\right)^{-1}\mathbf{A_{21}A_{11}^{-1}} & \left(\mathbf{A_{22}} - \mathbf{A_{21}A_{11}^{-1}A_{12}}\right)^{-1} \end{bmatrix} \quad 3.22$$

The above equation can also be written as:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{G_w^{-1}} - \mathbf{A_{11}^{-1}} & 0 \\ 0 & 0 \end{bmatrix} \quad 3.23$$

***3.4.2.2. Mixed model equations***

In order to estimate the breeding values for all individuals in the pedigree, the unified approach was implemented using the following modified mixed model equations that include the $\mathbf{H^{-1}}$ matrix.

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z_1} & \mathbf{X'Z_2} \\ \mathbf{Z_1'X} & \mathbf{Z_1'Z_1}+\mathbf{H}^{-1}\alpha_1 & \mathbf{Z_1'Z_2}+\mathbf{H}^{-1}\alpha_2 \\ \mathbf{Z_2'X} & \mathbf{Z_2'Z_1}+\mathbf{H}^{-1}\alpha_2 & \mathbf{Z_2'Z_2}+\mathbf{H}^{-1}\alpha_3 \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u_1} \\ \mathbf{u_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \end{bmatrix} \qquad 3.24$$

The symbols $\mathbf{y}$, $\mathbf{X}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$, $\mathbf{u_1}$, $\mathbf{u_2}$, $\mathbf{b}$, $\alpha_1$, $\alpha_2$ and $\alpha_3$ have already been defined in Sub-section 3.4.1.2.

**3.4.3. Estimation of breeding values and solving the mixed model equations**

For the estimation of breeding values it was assumed that pedigree records were available for all generations. A phenotypic value in the honey bee represents an observation for the whole colony and thus, cannot be decomposed into individual phenotypic values of a queen and an average worker; therefore, both the queen and an average worker of a colony were assigned the same colony phenotypic value. Phenotypes were available for all dam queens (and the corresponding average worker) in the base generation and for all potential-dam queens (and the corresponding average worker) in every but the last generation. Genotyping information was available for all dam queens in the base generation and all potential-dam queens.

The simulated values of genetic and residual variance were used for estimating breeding values. The direct and maternal breeding values were estimated for all individuals in the pedigree using the UNI_BLUP and PED_BLUP approaches. The overall breeding value for each individual was obtained from the sum of its direct and maternal breeding values. For both the PED_BLUP and the UNI_BLUP approach, accuracies were calculated for the following: (1) the overall estimated breeding values, (2) the maternal estimated breeding values and (3) the direct estimated breeding values. The accuracy of estimated breeding values were calculated for ‘juvenile queens’ constituted by potential-dam queens in the last generation and for ‘all queens’ constituted by dam queens in the base population and potential-dam queens in all generations. The accuracy was reported as the correlation between the estimated and the true breeding values (Mrode, 2005). All

calculations were performed in MATLAB. The summary statistics is based on 20 replicated simulations.

## 3.5. Future Ideas: Modification of the numerator relationship matrix to account for the composite structure of the dummy sire and average worker

Both a dummy sire and an average worker represent groups of individuals. Therefore, the diagonal elements of the numerator relationship matrix for dummy sires and average workers can be modified to account for this structure. This modification will help to improve the relationship matrix in the future studies. A basic concept is suggested below that shows how modifications can be included.

### 3.5.1. Relationship of a dummy sire with itself

The relationship of a dummy sire with itself ( $a_{DD}$ ) can be calculated from the methodology used for calculating coancestry ( $S_{DD}$ ). Assuming that a dummy sire is composed of three sister queens (I, II and III) as shown in Figure 3.1, $S_{DD}$ can be written as follows:

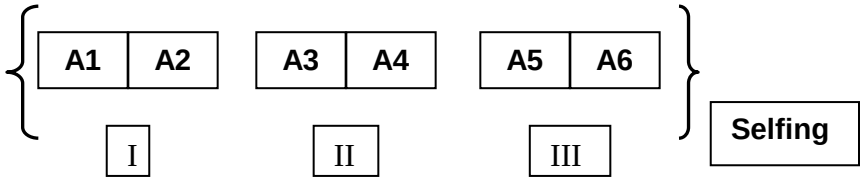


Figure 3.1. Sister queens constituting a dummy sire.

Sister queens are denoted by I, II, III. A1/A2, A3/A4 and A5/A6 correspond to the alleles at a locus in the queens I, II and III, respectively.

$$
\begin{aligned}
S_{DD} = 1/36\,[&p(A1=A1) + p(A1=A2) + p(A1=A3) + p(A1=A4) + p(A1=A5) + p(A1=A6) + \\
&p(A2=A1) + p(A2=A2) + p(A2=A3) + p(A2=A4) + p(A2=A5) + p(A2=A6) + \\
&p(A3=A1) + p(A3=A2) + p(A3=A3) + p(A3=A4) + p(A3=A5) + p(A3=A6) + \\
&p(A4=A1) + p(A4=A2) + p(A4=A3) + p(A4=A4) + p(A4=A5) + p(A4=A6) + \\
&p(A5=A1) + p(A5=A2) + p(A5=A3) + p(A5=A4) + p(A5=A5) + p(A5=A6) + \\
&p(A6=A1) + p(A6=A2) + p(A6=A3) + p(A6=A4) + p(A6=A5) + p(A6=A6)] \\
&= 1/36\left[6 + 2FI + 2FII + 2FIII + 24S_{ij}\right] \\
&= 1/36\left[6 + 6F + 24S_{ij}\right] \\
&= 1/6\left(1+F\right) + 24/36\,S_{ij} \\
&= 1/6\left(1+F\right) + 2/3\,S_{ij} \\
&= 1/6\left(1+0.5a_{sd}\right) + 2/3\,S_{ij} \quad (\text{since } F = 0.5a_{sd})
\end{aligned}
$$

where A1/A2, A3/A4 and A5/A6 are alleles at a locus in individuals I, II and III, respectively. $p(A_x = A_y)$ is the probability that the two alleles are identical by descent. FI, FII and FIII (assuming that FI, FII and FIII = F) are the inbreeding coefficients of individuals I, II and III, respectively. $S_{ij}$ is the coancestry of the $i^{th}$ queen with the $j^{th}$ queen and $a_{sd}$ is the relationship between sire and dam of the $i^{th}$ queen.

Let a dummy sire be composed of '$n$' queens, the 'general formula' for $S_{DD}$ can be given as follows:

$$S_{DD} = \frac{1}{2n}\left(1 + 0.5a_{sd}\right) + \left(\frac{n-1}{n}\right)S_{ij} \qquad 3.25$$

Since relationship is twice of coancestry, therefore,

$$a_{DD} = 2 \times S_{DD} = \frac{1}{n}\left(1 + 0.5a_{sd}\right) + \left(\frac{n-1}{n}\right)S_{ij} \times 2 \qquad 3.26$$

From this general equation, $a_{DD}$ can be obtained for the base population and normal population generated from the base population as shown below:

- For the base population, it can be assumed that all queens were unrelated and non-inbred; therefore, $S_{ij} = 0$, hence $S_{DD} = \frac{1}{2n}(1 + 0.5a_{sd})$. As $a_{sd}$ is also zero in the base generation, $S_{DD} = \frac{1}{2n}$ and $a_{DD} = 2S_{DD} = \frac{1}{n}$.

- In the normal population generated from the base population $a_{DD}$ is given as:

  $$a_{DD} = 2S_{DD} = \frac{1}{n}\left(1 + 0.5a_{sd}\right) + \left(\frac{n\text{-}1}{n}\right)\left(S_{ij} \times 2\right)$$

  The above equation includes the relationship between sister queens/colonies ($S_{ij}$) that constitute a dummy sire; thus, it is important to estimate the average

relationship/coancestry between two sister queens. Since sister queens are usually related as super-sibs, full-sibs or half-sibs, the average relationship between sister queens can be assumed to be equal to $\frac{0.75+0.5+0.25}{3}=0.5$; therefore, $S_{ij}=\frac{0.5}{2}=0.25$. A weighted probability can also be used instead of the simple usage of 1/3. Thus, after including the average relationship between sister queens, the value of relationship in diagonal elements for the dummy sire can be written as:

$$a_{DD}=2S_{DD}=\frac{1}{n}\left(1+0.5a_{sd}\right)+\left(\frac{n\text{-}1}{n}\right)\left(0.25\times 2\right)=\frac{1}{n}\left(1+0.5a_{sd}\right)+\left(\frac{n\text{-}1}{n}\right)0.5$$

### 3.5.2. Relationship of between workers of a colony

Table 3.1. An example relationship matrix between workers of a single colony assuming that all drones are from a single queen.

| | | | | |
|---|---|---|---|---|
| 1 0.75 - 1 | $\frac{1}{2}$ | $\frac{1}{2}$ | ..... | $\frac{1}{2}$ |
| $\frac{1}{2}$ | 1 0.75 - 1 | | ..... | $\frac{1}{2}$ |
| $\frac{1}{2}$ | $\frac{1}{2}$ | 1 0.75 - 1 | ..... | $\frac{1}{2}$ |
| : : | : : | : : | . . | : : |
| $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ | ..... | 1 0.75 - 1 |

Table 3.1 shows an example for a relationship matrix ($\mathbf{A}_{\mathbf{worker}}$) between workers of a single colony. The colony size ($CS$) is defined as the number of workers in a colony and $d$ is the

number of drones mating with the dam of the workers. For now, all drones are assumed to come from a single queen. Furthermore, if it is assumed that the number of worker offspring from each drone is $\frac{CS}{d}$, then the size of each diagonal block (shown in gray) is equal to $\left(\frac{CS}{d}\right)^2$. It is supposed that all workers in a colony come from a single non-inbred queen, unrelated to the dummy sire. The values in Table 3.1 denote the following:

1 – relationship of a worker with itself

0.75 – relationship between super sisters

0.5 – relationship between full sisters

The relationship between workers of a colony, $\bar{a}_{worker,worker}$, is given as:

$$\bar{a}_{worker,worker} = \frac{1}{CS^2}\left\{\left[\left(\frac{CS}{d}\right)+\frac{3}{4}\left(\left(\frac{CS}{d}\right)^2-\frac{CS}{d}\right)\right]d+\frac{1}{2}\left(CS^2-d\left(\frac{CS}{d}\right)^2\right)\right\}$$

On solving further, it reduces to $\bar{a}_{worker,worker} = \frac{1}{4CS}+\frac{1}{4d}+\frac{1}{2}$

As $CS$ ~ 50,000, $\frac{1}{4CS}$ can be neglected. Thus, the relationship can be given as:

$$\bar{a}_{worker,worker} \approx \frac{1}{2}+\frac{1}{4d} \qquad 3.27$$

Considering that drones may come from more than one queen, the above relationship can be extended by including '$q$' queens in the dummy sire (i.e. all drones come from '$q$' queens). Table 3.2 shows the relationship between workers after including this modification.

Table 3.2. An example relationship matrix between workers of a single colony assuming that drones come from different queens.

| | | | | |
|---|---|---|---|---|
| $\frac{1}{2}+\frac{1}{4d_1}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | ..... | $\frac{1}{4}$ |
| $\frac{1}{4}$ | $\frac{1}{2}+\frac{1}{4d_2}$ | $\frac{1}{4}$ | ..... | $\frac{1}{4}$ |
| $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{2}+\frac{1}{4d_3}$ | ..... | $\frac{1}{4}$ |
| : | : | : | . | : |
| : | : | : | . | : |
| $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | ..... | $\frac{1}{2}+\frac{1}{4d_n}$ |

The values in Table 3.2 denote the following:

$\frac{1}{2}+\frac{1}{4d_i}$ – the relationship between workers having the same drone-producing queen as sire.

$\frac{1}{4}$ – the relationship between workers having different drone-producing queens as sire (maternal half-sibs).

Here, each diagonal block refers to a group of workers which are offspring of different unrelated queens constituting the dummy sire. Assuming that $d_1, d_2, d_3, \cdots d_n$ are the number of drones contributed by each queen ($d_i = d$ is assumed). If $q$ is the number of queens representing the dummy sire, then

$$\bar{a}_{worker,worker} = \frac{1}{q^2}\left(q\left[\frac{1}{2}+\frac{1}{4d}\right]\right)+\frac{1}{q^2}\left(\left[q^2-q\right]\frac{1}{4}\right)$$

On solving further, the above equation reduces to:

$$\bar{a}_{worker,worker} = \frac{1}{4}\left(\frac{1}{q}+\frac{1}{qd}+1\right) \qquad 3.28$$

Further calculations can be performed in a similar manner for other cases of relationship in the honey bee population.

# Chapter 4. Towards analysis of real data: Development of a 44k SNP assay

The definitive version is available at www.blackwell-synergy.com

This chapter outlines the procedure for the development of a 44k SNP assay for the honey bee. One of the essential applications of a SNP assay is to genotype several thousand SNP loci across the genome simultaneously and to identify any DNA variant associated with the disease in genome-wide association studies. Genotyping information is also crucial to the application of genomic selection and other molecular marker based methods for the estimation of breeding values. Since a high density of markers spanning the entire genome is required in future association studies and genetic evaluation, the developed 44k SNP assay will serve as an indispensible tool. The SNP assay was designed with a focus on the 'hygienic behaviour' (see Chapter 1), however, the density of SNP markers in the assay is high enough, making it suitable for association studies for other traits as well. As shown in the pipeline (Figure 4.1), different experimental and data analyses steps were performed in this study. A summary of these steps is provided in this chapter. For more information please refer to Spötter *et al*. (2012).

## 4.1. Experimental work

### 4.1.1. SNP selection and capture microarray design

A total of 70,000 SNP were analysed for their suitability as genetic marker in future association studies for hygienic behaviour in the honey bee. For the construction of microarray and the selection of SNP positions, SNP regions identified and published by the Honey Bee Genome Project were chosen (www.hgsc.bcm.tmc.edu/ftp-archive/Amellifera/snp/). For further consideration, only those SNP that were assigned to a linkage group in the honey bee were selected. A repeat masking filter was applied to prevent designing oligos with uncertain genomic localization further downstream. A total of 77,565 regions were selected with an even distribution across the genome. The average distance between SNP loci was about 4 kb, covering approximately 75% of the genome. Based on a previous study of 245 microsatellite loci (M. Brink, M. Solignac, K. Bienefeld, unpublished data), which identified QTL for the trait 'removal

of Varroa-infested brood', the remaining 25% of the genome had a denser spacing with an average distance of 2 kb between adjacent SNP positions in the region of identified QTL.

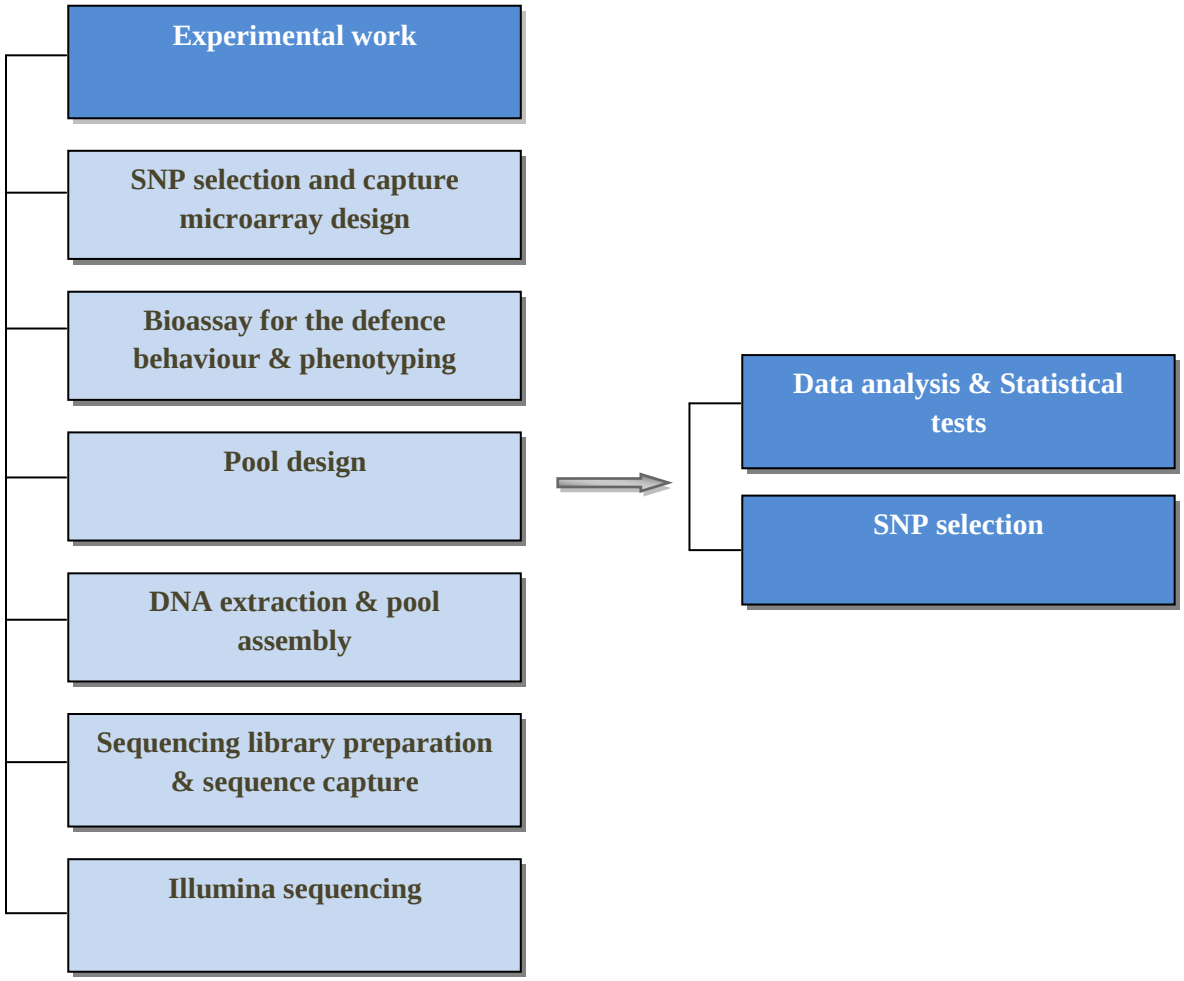


Figure 4.1. A pipeline showing the procedure of development of the 44k SNP assay.

**4.1.2. Bioassay for the defence behaviour and phenotyping**

In honey bees, hygienic behaviour involves inspection, uncapping (i.e. opening of the brood cell) and removal of diseased and dead brood from the colony. As described earlier, the SNP assay was developed with a special focus on the hygienic behaviour. Worker bees were screened for this phenotype during a bioassay and later selected to form a part of a DNA pool. Only freshly hatched worker bees (0–12 hours old) were marked and used for the defence behaviour bioassays at the age of 4 days. The defence behaviour bioassay consisted of the following steps:

1. 2000 freshly hatched worker bees were marked and transferred into one caged experimental comb (Thakur *et al.*, 1997). These combs were derived from unrelated and *Varroa*-free colonies. The observation area for each comb consisted of 169 brood cells.

2. Out of 169 brood cells, 45 were artificially infested with a *Varroa* mite. Seven cells were kept empty, 75 brood cells were untreated (as normal control) and 43 cells were opened and resealed without inserting a mite (as a control for the cap-manipulation).

3. The caged experimental combs were integrated into the hive of a helper colony that provided an environment that was close to natural conditions.

4. An infra-red camera was installed in front of the comb which allowed the recording of hygienic behaviour in the absence of light without disturbing the bees. Recording was carried out for seven days.

5. Recordings from the infra-red camera were analysed to observe for the hygienic behaviour. Bees that started to open a cell were called 'beginner' whereas those that expanded the existing holes were called 'helper'.

The defence behaviour bioassay was run 14 times, five runs in the year 2005, seven runs in the year 2007 and two runs in the year 2008. About 80% of the colonies used as sources for experimental animals belonged to a line selected for *Varroa* tolerance since 1997 in the Institute for Bee Research, Hohen Neuendorf, Germany. The remaining 20% of the colonies were obtained from breeders distributed all over Germany. These colonies were also bred for *Varroa* tolerance based on the estimated breeding values (http://www2.hu-berlin.de/bienenkunde/ZWS/).

**4.1.3. Pool design**

Three DNA pools, the trait exhibiting pool (A) and two control pools (B and C) were constructed. The pool A consisted of DNA from 50 top performing individuals for the hygienic behaviour, selected from about 28,000 tested bees. The selection criterion for the top performing bee was based on the number of uncapping action directed against *Varroa*-infested cells. These bees were involved directly in at least one uncapping action and acted as a helper in at least one uncapping event. The number of their uncapping and helping action against *Varroa*-infested cells

was at least two times higher than their action against the control cells. The pool B consisted of DNA from 50 workers belonging to the same colony as the bees chosen for pool A, but did not show any hygienic behaviour. The pool C consisted of DNA from 50 workers belonging to colonies where not a single bee showed the hygienic behaviour.

A comparison between pool A and B helped to identify the difference existing due to the hygienic behaviour. It also helped to account for any difference due to population stratification. On the other hand, comparison between pool A and C helped to identify any causative genes that have a low degree of penetrance, because in such a case no significant difference between pool A and pool B could be identified. The individuals used to construct pool C were derived from all over Germany. This allowed obtaining information about the allele distribution and the degree of polymorphism in *Apis mellifera carnica* population.

#### 4.1.4. DNA extraction and pool assembly

In order to re-sequence target regions for validating SNP for the hygienic behaviour, DNA was extracted from selected worker bees for each pool. Overall, the DNA extraction and pool assembly process consisted of the following steps:

1. Extraction of DNA: DNA was extracted using the NucleoSpin Tissue kit (Macherey-Nagel, Düren, Germany) from the heads and thoraces of the worker bee stored in 96% ethanol.

2. DNA quality and concentration check: The quality and concentration of the DNA samples were examined on 0.8% agarose gels and measured photometrically at an optical density (OD) of 260/280 using a NanoDrop 2000 (NanoDrop products, Wilmington, DE, USA). DNA samples were used only if the OD260/280 was between 1.7 and 2.1 and the degree of DNA degradation was small.

3. Preparation of the working solution: For all DNA samples, working solutions of a concentration of 20 ng/µL were prepared. Sample concentrations were checked using a NanoDrop 2000 and adjusted if necessary.

4. DNA pool assembly: From the working solutions of DNA samples, 400 ng of DNA was used to assemble all pools. This was done by taking 20 µL of the working solution by volume.

Since each of the three pools consisted of DNA from 50 workers, the resulting pool contained 20 μg DNA in a volume of 1000 μL.

**4.1.5. Sequencing library preparation and sequence capture**

The standard illumina sequencing library preparation methodology was employed for the library preparation (Meyer and Kircher, 2010) for each of the DNA pool. Sequencing library preparation and sequence capture consisted of the following steps:

1. Adaptor ligation and barcoding: A different self-assembled single read adaptor was used for each library. The adaptor contained a 4 base-pairs barcode tag at the end which was ligated to the genomic DNA fragment. This enabled the multiplexing of a sample in the same flow cell channels during sequencing and backtracking of the reads to the respective DNA pools with the help of barcodes.

2. DNA fractionation: Fragments of DNA were fractionated according to size ranging between 150-250 bp.

3. Amplification: After adaptor ligation and fractionation, the fragments of DNA were amplified to yield about 5 μg of DNA. In order to avoid the risk of bias in fragment representation, amplification was carried out with limited cycle numbers and in multiple reactions.

4. Sequence capture: Sequence capture allows parallel enrichment of target regions in a single experiment and helps to eliminate setting up thousands of PCR reactions (Meyer and Kircher, 2010). For this step, the material was hybridised individually for each sample to the customized 1 million feature Agilent array representing honey bee SNP positions (Hodges *et al*., 2009). The captured DNA was washed, eluted and amplified again in order to obtain adequate material for subsequent illumina sequencing.

5. Template for illumina sequencing: In order to establish templates for illumina sequencing, after quantitation the DNA sample was mixed in equimolar concentrations from all three capture libraries.

#### 4.1.6. Illumina sequencing

The pooled, barcoded sequencing libraries were loaded onto a full illumina 'Flow cell channel', and then single-end-sequenced with 72-bp reads on a Genome Analyzer GA-IIx, using Chrysalis 36 cycles v4.0 chemistry and the RTA SCS.2.6/CASAVA 1.6 data analysis pipeline.

## 4.2. Data analysis

Pearson's chi-square and pFDR tests were employed to identify SNP associated with the trait 'uncapping of *Varroa*-infested brood'. These statistical tests were performed for the SNP which met the following two criteria: The first criterion was that only bi-allelic SNP were chosen. For SNP having three or four alleles, the two alleles with the highest minor allele frequencies were taken as real alleles as long as they were consistent with the reference alleles from the SNP list (www.hgsc.bcm.tmc.edu/ftp-archive/Amellifera/snp/). Consequently, the other third and/or forth allele with low minor allele frequencies were discarded. The second criterion for selecting SNP was that the coverage depth, which refers to the number of sequences analysed per SNP, was 15-fold or greater. A coverage depth of 15-fold allowed to select the highest possible number of validated SNP for the assay to be developed without compromising quality.

The allele frequencies of SNP between pools A and B and pools A and C were tested SNP by SNP for significant differences of allele frequencies between pools by employing Pearson's chi-square test. For each SNP, the null hypothesis of equal allele frequencies in two compared pools was $H_0$: $p_1 - p_2 = 0$, where $p_1$ and $p_2$ are the allele frequencies in both pools. It was tested against the two-sided alternative of a non-zero difference. A chi-squared test statistics with a single degree of freedom was calculated together with the associated p-values. Multiple testing was accounted for by calculating the expected proportion of falsely rejecting the null hypothesis pFDR (Storey, 2002) for all comparisons between pools A and B as well as pools A and C. Up to a pFDR of 0.05 SNP were reported as differing significantly in their frequency between pools. The calculations were performed using the MULTTEST procedure of the SAS statistical software package (SAS 2003, 9.1; Inst., Inc., Cary, NC, USA).

## 4.3. Selection of SNP for the 44k SNP assay

Based on these experiments and data analysis, 36,000 SNP were chosen for constructing the SNP assay. Furthermore, to obtain an even distribution of SNP across the genome, additional SNP

which were not validated for their suitability as markers for hygienic behaviour were included for the design of the 44k SNP assay. For position selection and array design, information about these additional SNP was obtained from the list published by the Honey Bee Genome Project (www.hgsc.bcm.tmc.edu/ftp-archive/Amellifera/snp/). The SNP assay has been made publicly available through AROS Applied Biotechnology AS, Aarhus, Denmark.

# Chapter 5. Results and Discussion

The main aim of this study was to incorporate marker data into the genetic evaluation of the honey bee. Thus, to provide an insight into the benefits of including marker data in the genetic evaluation, comparative evaluation was performed using (1) the traditional approach based on pedigree data and (2) the unified approach based on both pedigree and marker data through simulation. In order to perform this comparative study, a population with genotyping, pedigree and phenotypic data had to be simulated. Thus, a completely new framework was developed for simulating a population based on the special genetic and reproductive characteristics of the honey bee. Most importantly, the influence of maternal effects on the trait, negative correlation between maternal and direct effects and uncertain paternity were also addressed in this study, thus making it relevant for other species as well. Results from this comparative study between the unified and pedigree based approach reports the accuracy of the estimates of overall, direct and maternal breeding values. Furthermore, the influence of the heritability of the trait as well as the genetic correlation between maternal and direct effects on the accuracy of the estimated breeding values were investigated. The results show that the unified approach performed better than the pedigree based approach. Additionally, a 44k SNP assay was developed that can be used to incorporate high-density marker information into the genetic evaluation of the honey bee in future studies.

## 5.1. Results

This section describes the validation results for the simulation of a base population and the results for the accuracy of the unified and pedigree based genetic evaluation approaches.

### 5.1.1. Validation of the software program for simulating a base population

To perform a validation of the developed software program, the achieved LD was compared with theoretical LD. For this, two simulations were performed, the first one consisted of 500 sire and 50 dam queens and the second one of 200 sire and 20 dam queens. A total of 100,000 marker loci were simulated for 2000 generations. The forward and backward mutation rates were set to 0.0025, a value similar to that used by Meuwissen *et al*. (2001), allowing a high probability of polymorphic marker loci. Information on the level of recombination and on the effective

population size ($N_e$) in the honey bee was obtained from Beye *et al*. (2006) and Estoup *et al*. (1995), respectively. The expected average LD was compared to the achieved average LD for 44,000 loci with the highest minor allele frequencies. The expected average LD in a population was calculated as follows (Hill, 1975):

$$r^2 = \frac{5 + 2N_e c}{11 + 26N_e c + 8N_e^2 c^2} \qquad 5.1$$

where c is the recombination fraction between adjacent loci and $N_e$ is the effective population size. Since the total size of the simulated genome was 219,629,612 base-pairs (Table 2.1) and the approximate recombination rate was taken as 19 cM/Mb (Beye *et al*., 2006; The Honeybee Genome Sequencing Consortium, 2006), the size of the simulated genome was 41.73 M. Thus, for a genome of 41.43 M, c was approximately 0.001 for 44,000 SNP. The honey bee population has a wide range of effective population sizes (Estoup *et al*., 1995); therefore, two scenarios were simulated, one with 220 queens and the other with 550 queens. The effective population size in the honey bee was calculated using the following expression for a haplo-diploid population (Wright, 1933; Kerr, 1967):

$$N_e = \frac{9N_f N_m}{2\left(2N_m + N_f\right)} \qquad 5.2$$

where $N_f$ is the number of queens (which is equal to the number of colonies since each colony is headed by a single queen) and $N_m$ is the number of males. In the simulation, it is assumed that each queen is inseminated by 11 drones, therefore $N_m = 11\,N_f$.

For 220 and 550 colonies, $N_e$ was approximately 473 and 1184, respectively. With $N_e$ = 473, the expected LD was 0.24 and the achieved LD was 0.23. Similarly, with $N_e$ = 1184, the expected and achieved LD were equal to 0.14 and 0.11, respectively. These values show that the software program is able to model the honey bee population with good accuracy.

Figures 5.1 and 5.2 show the establishment of LD for the simulated datasets consisting of 220 and 550 queens. The average LD (averaged over all simulated pairs of marker loci across the genome, not preselected on the basis of minor allele frequency, in each generation) was plotted against the number of generations. These graphs show that the software program simulates a random mating honey bee population till the mutation-drift equilibrium is reached and a stable value of LD is established.

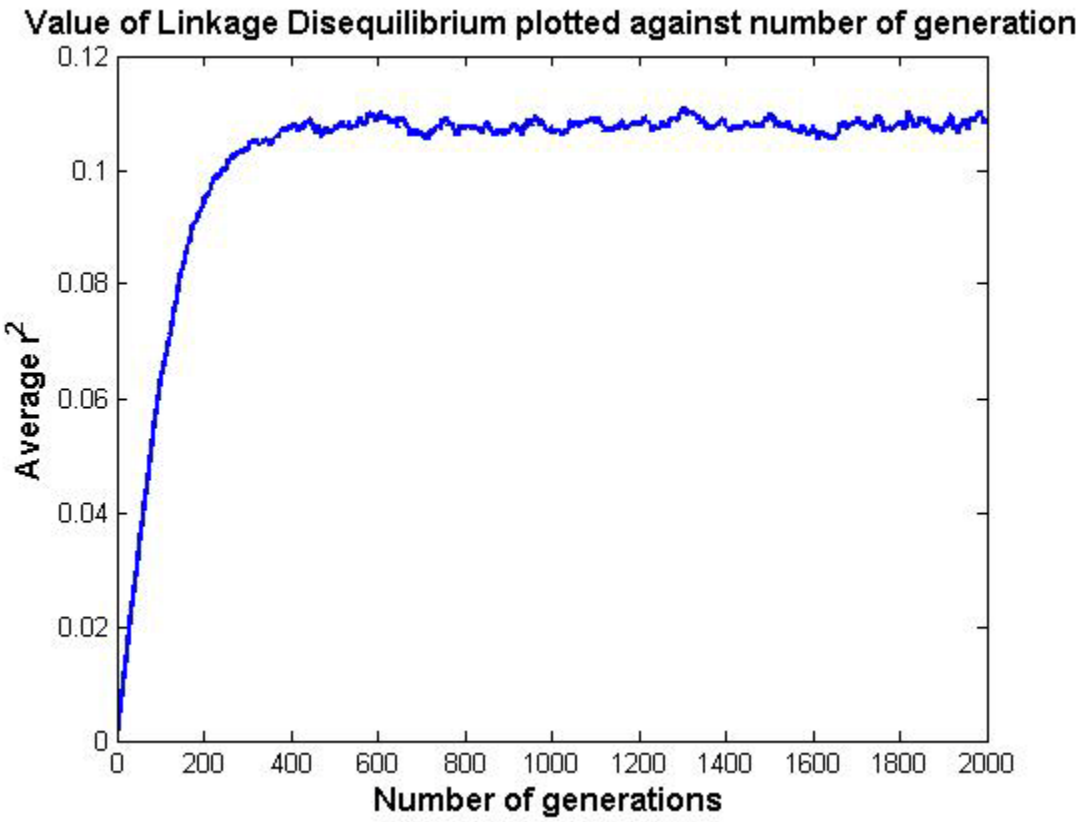


Figure 5.1. The average value of $r^2$ plotted against the number of generations for a population consisting of 220 queens.

Simulation was performed for 2000 generations with a forward and backward mutation rate of 0.0025 for 100,000 marker loci and 220 colonies (20 dam queens and 200 sire queens); with the parameter values chosen here, a stable LD was reached after random mating.

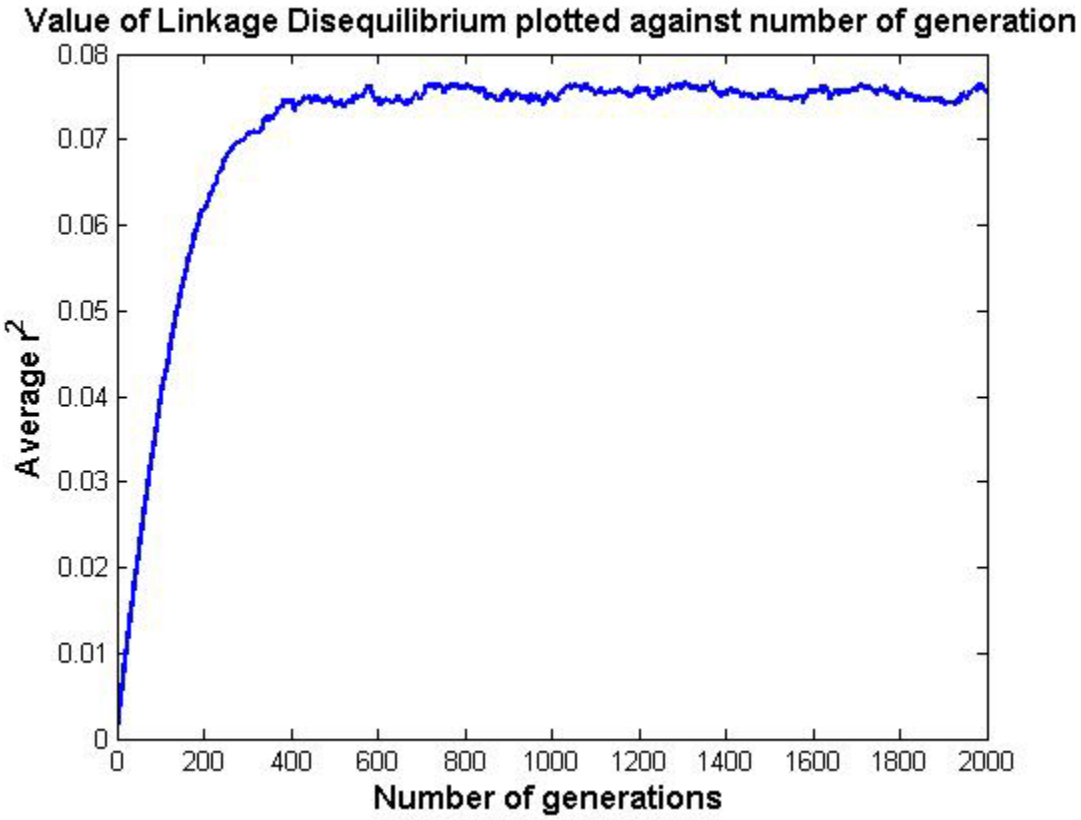


Figure 5.2. The average value of $r^2$ plotted against the number of generations for a population consisting of 550 queens.

Simulation was performed for 2000 generations with a forward and backward mutation rate of 0.0025 for 100,000 marker loci and 550 colonies (50 dam queens and 500 sire queens); with the parameter values chosen here, a stable LD was reached after random mating.

### 5.1.2. Results from the implementation of the unified approach

#### *5.1.2.1. Accuracy of the overall estimated breeding values*

In the honey bee breeding programs, the overall breeding values, a sum of the maternal and direct estimated breeding values, is used for selecting queens. Table 5.1 shows the accuracy achieved for the overall estimated breeding values with the UNI_BLUP and PED_BLUP approaches.

Table 5.1. Accuracy of the overall estimated breeding values.

| Maternal heritability | Method | Cor(M, D) | Accuracy of overall EBV for JQ | Accuracy of overall EBV for AQ |
|---|---|---|---|---|
| 0.15 | UNI_BLUP | 0 | 0.468 [a,b,c,d] | 0.661 [a,b,c,d] |
| | PED_BLUP | 0 | 0.363 | 0.603 |
| | UNI_BLUP | -0.46 | 0.381 [a,b,c,d] | 0.555 [a,b,c,d] |
| | PED_BLUP | -0.46 | 0.295 | 0.489 |
| 0.25 | UNI_BLUP | 0 | 0.542 [a,b,c,e] | 0.756 [a,b,c,e] |
| | PED_BLUP | 0 | 0.420 | 0.710 |
| | UNI_BLUP | -0.46 | 0.449 [a,b,c] | 0.640 [a,b,c,e] |
| | PED_BLUP | -0.46 | 0.348 | 0.577 |
| 0.35 | UNI_BLUP | 0 | 0.604 [a,b,d,e] | 0.832 [a,b,d,e] |
| | PED_BLUP | 0 | 0.467 | 0.800 |
| | UNI_BLUP | -0.46 | 0.498 [a,b,d] | 0.700 [a,b,d,e] |
| | PED_BLUP | -0.46 | 0.388 | 0.642 |

Cor(M, D) stands for the correlation between maternal and direct effects; JQ denotes juvenile queens; AQ denotes all queens; EBV denotes the estimated breeding values.

Significant difference in accuracy with p-values < 0.05 between: [a]UNI_BLUP and PED_BLUP; [b]no correlation and negative correlation for UNI_BLUP; [c]heritabilities 0.15 and 0.25 for UNI_BLUP; [d]heritabilities 0.15 and 0.35 for UNI_BLUP; [e]heritabilities 0.25 and 0.35 for UNI_BLUP.

For the juvenile queens (constituted by potential-dam queens of the last generation), the accuracy of the overall estimated breeding values was significantly higher with the UNI_BLUP approach (p-value < 0.05) as compared to the PED_BLUP approach for all values of heritability and correlation between maternal and direct effects. For almost all cases, the increase in accuracy was approximately 0.1 (i.e. 29%).

Similar to juvenile queens, the accuracy of the overall estimated breeding values for all queens (constituted by dam queens in the base population and potential-dam queens in all generations) was higher with the UNI_BLUP approach (p-value < 0.05) at all heritabilities and correlation between maternal and direct effects. For the case of no correlation between maternal and direct effects, the percentage increase in accuracy was approximately 9.62%, 6.48% and 4.00% at maternal heritabilities of 0.15, 0.25 and 0.35, respectively. In case of negative correlation, the percentage increase in accuracy was approximately 13.50%, 10.92% and 9.03% at maternal heritabilities of 0.15, 0.25 and 0.35, respectively.

#### *5.1.2.2. Accuracy of the maternal and direct estimated breeding values*

The accuracy of the maternal and direct estimated breeding values for juvenile queens and all queens is presented in Table 5.2. The average values of accuracy of the maternal as well as direct estimated breeding values were higher for the UNI_BLUP approach as compared to the PED_BLUP approach. However, the difference between UNI_BLUP and PED_BLUP approaches were not significant for some cases as compared to the accuracy of the overall estimated breeding values. In general, the accuracy of the maternal and direct estimated breeding values showed a trend in favour of the UNI_BLUP approach.

Table 5.2. Accuracy of the direct and maternal estimated breeding values.

| Maternal heritability | Method | Cor(M, D) | Accuracy of direct EBV for JQ | Accuracy of maternal EBV for JQ | Accuracy of direct EBV for AQ | Accuracy of maternal EBV for AQ |
|---|---|---|---|---|---|---|
| 0.15 | UNI_BLUP | 0 | 0.323 [a,b,c,d] | 0.279 [a,b,c,d] | 0.446 [a,b,c,d] | 0.420 [a,b,c,d] |
| | PED_BLUP | 0 | 0.227 | 0.225 | 0.406 | 0.381 |
| | UNI_BLUP | -0.46 | 0.115 [b,d] | 0.127 [b] | 0.225 [b,d] | 0.223 [b,d] |
| | PED_BLUP | -0.46 | 0.059 | 0.103 | 0.186 | 0.208 |
| 0.25 | UNI_BLUP | 0 | 0.373 [a,b,c] | 0.330 [a,b,c,e] | 0.510 [b,c,e] | 0.482 [a,b,c,e] |
| | PED_BLUP | 0 | 0.268 | 0.260 | 0.474 | 0.447 |
| | UNI_BLUP | -0.46 | 0.154 [b] | 0.154 [b] | 0.272 [b] | 0.257 [b] |
| | PED_BLUP | -0.46 | 0.085 | 0.125 | 0.231 | 0.240 |
| 0.35 | UNI_BLUP | 0 | 0.418 [a,b,d] | 0.371 [a,b,d,e] | 0.566 [b,d,e] | 0.527 [b,d,e] |
| | PED_BLUP | 0 | 0.307 | 0.287 | 0.538 | 0.496 |
| | UNI_BLUP | -0.46 | 0.186 [a,b,d] | 0.173 [b] | 0.308 [b,d] | 0.280 [b,d] |
| | PED_BLUP | -0.46 | 0.110 | 0.138 | 0.268 | 0.258 |

Cor(M, D) stands for the correlation between maternal and direct effects; JQ denotes juvenile queens; AQ denotes all queens; EBV denotes the estimated breeding values.

Significant difference in accuracy with p-values < 0.05 between: [a]UNI_BLUP and PED_BLUP; [b]no correlation and negative correlation for UNI_BLUP; [c]heritabilities 0.15 and 0.25 for UNI_BLUP; [d]heritabilities 0.15 and 0.35 for UNI_BLUP; [e]heritabilities 0.25 and 0.35 for UNI_BLUP.

#### *5.1.2.3. Effect of correlation and heritability*

Both low heritability and a high negative correlation lead to a lower genetic variance, thus it is expected that the accuracy will decrease with a decrease in heritability and with an increase in negative correlation. The accuracy of the overall estimated breeding values was lower for the case where maternal and genetic effects were negatively correlated in comparison to the case with no correlation (Table 5.1 and Figure 5.3, 5.4; p-value < 0.05). Similarly, the accuracy of the overall estimated breeding values increased as the heritability increased for both no correlation

and negative correlation between maternal and direct effects (Table 5.1 and Figure 5.3, 5.4; p-value < 0.05). The only exception was a single case where no significant difference was observed. Difference in the accuracy of the overall estimated breeding values between maternal heritability of 0.25 and 0.35 at negative correlation of -0.46 for juvenile queens was not significant, although the accuracy was higher for high heritability. This can possibly be explained by the fact that there will be less gain in accuracy at higher heritability. For the accuracy of the maternal and direct estimated breeding values (Table 5.2), the difference was significant for most cases between all heritability at no correlation and between heritability of 0.15 and 0.35 at negative correlation. In general, the accuracies of the maternal and direct estimated breeding values were higher for high values of heritability and no correlation between maternal and direct effects showing a trend similar to the overall estimated breeding values.

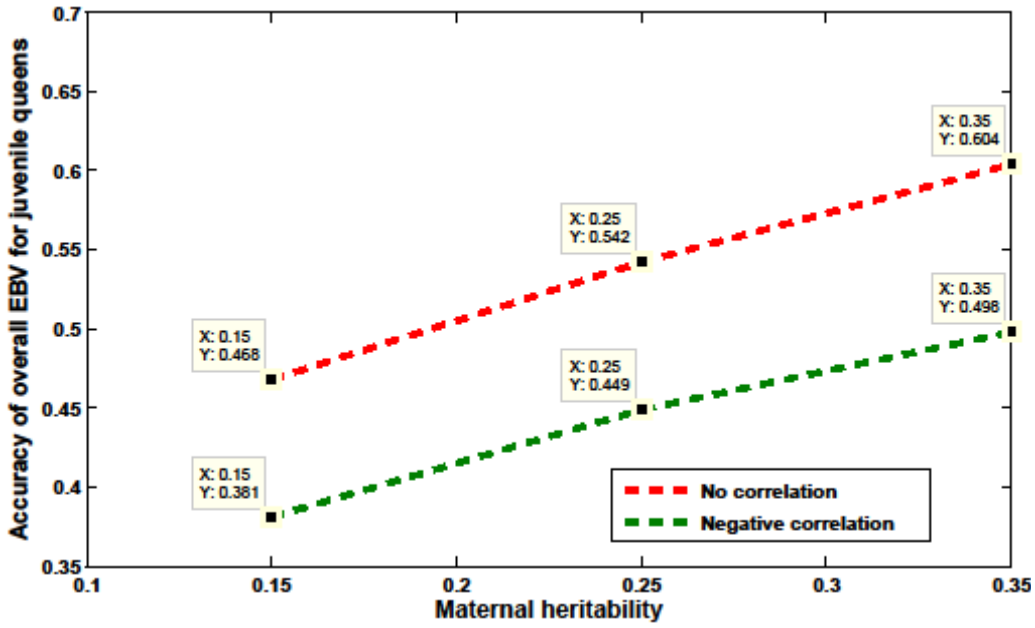


Figure 5.3. The effect of correlation between maternal and direct effects and heritability on the accuracy of the overall estimated breeding values (EBV) for juvenile queens.

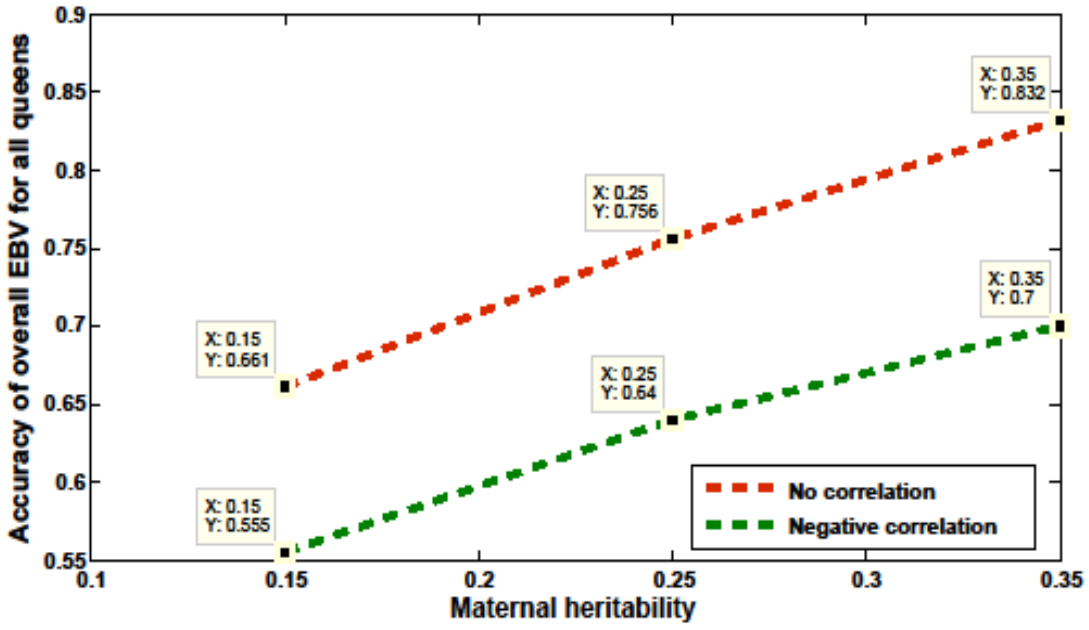


Figure 5.4. The effect of correlation between maternal and direct effects and heritability on the accuracy of the overall estimated breeding values (EBV) for all queens.

## 5.2. Discussion

Integration of marker data for genetic evaluation was achieved through the unified approach. A complex scenario was modelled that took into account several unique characteristics of the honey bee population. The study also takes into account the effect of negative correlation between maternal and direct effects and uncertain paternity, which are important factors in determining the degree of selection response in a population. The results showed that marker information improves the accuracy of the estimation of breeding values. The following section addresses different aspects related to the simulation strategy and the genetic evaluation based on the unified approach in more detail.

### 5.2.1. Base population simulating software program

The comparison of the estimated LD to the achieved LD shows that the software program is able to model the honey bee population with good accuracy. Creating a dataset for a base population is a prerequisite for any simulation study, but it can be time consuming and requires testing of optimal parameters. For this study, a software program was developed that simulates a base population for the honey bee. In most of the available software used to simulate populations (Peng *et al*., 2005), base population simulation is the preliminary stage, and is realized by

allowing the population to evolve through a 'burn-in' period until the population reaches an equilibrium from a random or uniform initial state. In the software program used for this study, all individuals in the starting generation are assumed to be unrelated. To establish LD, random mating is performed for the required number of generations and the last generation, which is in mutation-drift equilibrium, is taken as the base population. In genomic selection studies, a base population is the common starting point from which a population evolves further according to specific study requirements. To the best of the knowledge, this is the first software program that deals with evolutionary aspects in honey bees. It aims at providing an impetus to simulation studies in honey bees. It is an important initiative that could be used to implement and validate the genomic selection strategy through simulation in the honey bee. The code is written in MATLAB, but can easily be adapted to the open source alternative Octave.

**5.2.2. Implementation of the unified approach**

In a study for the pig by Forni *et al*. (2011), the accuracy of the estimated breeding values for genotyped females was reported to be 0.22 with the pedigree based approach whereas it ranged from 0.28 to 0.49 with the unified approach depending on the **G** matrix. Similarly, Christensen and Lund (2010) reported an accuracy of 0.66 with the one-step unified approach and 0.35 with the pedigree based approach on a simulated dataset for the pig. Genetic evaluation based on the unified approach of US Holstein was performed by Aguilar *et al*. (2010). The study reported the coefficient of determination (square of the correlation coefficient) with the pedigree based approach (parent average) and the unified approach (single-step GB) to be 0.24 (accuracy ~ 0.49) and 0.38 (accuracy ~ 0.62), respectively. In this study for the honey bee which additionally takes into account the effect of maternal effects unlike any previous study, comparable results were observed (Table 5.1) for the accuracy of the overall estimated breeding values. The accuracy of the overall estimated breeding values increased considerably with the unified approach for all scenarios of heritability and correlations (p-values $< 0.05$). For juvenile animals, a higher gain in the accuracy of the overall estimated breeding values was observed. It is favourable if the gain in accuracy is higher for juvenile animals as they are the subsequent candidates for selection. This may consequently help to speed up the selection procedure as a result of the reduction of the generation interval.

It has been reported in honey bees that most economically important traits have medium heritability (Bienefeld and Pirchner 1990; Boecking *et al*., 2000; Costa-Maia *et al*., 2011). For example, Bienefeld and Pirchner (1990) reported heritabilities for worker and queen effects to be 0.26 and 0.15 for honey production and 0.39 and 0.45 for wax production, respectively. Therefore, for this study, heritabilities were also simulated within the same range. The extremely negative estimates of genetic correlation between maternal and direct effects have often been a matter of discussion (Bijma, 2006; Ehrhardt and Bienefeld, unpublished results), therefore, a general value of correlation of -0.46 was simulated, which exists in other species as well (Larsgard and Olesen, 1998; Splan *et al*., 2002; Safari *et al*., 2005), and compared it to a case with no correlation between maternal and direct effects.

Roehe and Kennedy (1993) reported the accuracy of the maternal and direct estimated breeding values to be 0.21 (0.21) and 0.38 (0.28) for the case of no correlation and 0.19 (0.18) and 0.31 (0.23) for a negative correlation of -0.5 in female (male) pigs for maternal and direct heritability of 0.05 and 0.1, respectively. For the estimation of breeding values, a pedigree based complete animal model with maternal effects was used. In this study for the honey bee, the accuracies of the maternal and direct estimated breeding values for the pedigree based approach (PED_BLUP) at maternal and direct heritability of 0.15 were 0.38 and 0.41 for no correlation and 0.21 and 0.19 for a correlation of -0.46, respectively. The difference in the accuracies to that reported by Roehe and Kennedy (1993) can be a result of dissimilarities between the two studies, such as the random selection of individuals, construction of the numerator relationship matrix, value of simulated maternal and direct heritability, number of simulated generations, population structure and size. Nevertheless, the comparison of results of the PED_BLUP approach with the study from Roehe and Kennedy (1993) helps to validate the values of accuracy of the maternal and direct estimated breeding values obtained in this study for the honey bee. The accuracies of the maternal and direct estimated breeding values were higher for the UNI_BLUP approach as compared to the PED_BLUP approach, but the difference was insignificant for some cases (Table 5.2). Thus, in order to achieve maximum gain from implementing the unified approach, a proper investigation of the cost benefits and the relative improvement in genetic gain is required for traits selected solely on the basis of the maternal or direct breeding values. Nevertheless, the

sum of the maternal and direct breeding values is still the most important criterion for selection and the use of only direct or maternal breeding values is not helpful for the honey bee.

The complexity associated with the estimation of breeding values for maternally influenced traits is that the maternal and direct effects are negatively correlated in most cases. This severely impedes the response to selection (Willham, 1972; Roehe and Kennedy, 1993; Mousseau and Fox, 1998; Räsänen and Kruuk, 2007). Additionally, a negative correlation between maternal and direct effects leads to a decrease in the total genetic variance resulting in lowered accuracies. As reported earlier in the results section, the accuracy of the estimated breeding values improved significantly in the case of negative correlation with the unified approach as compared to the pedigree based approach. Thus, the use of unified approach will especially help to improve the genetic response in case of maternally influenced traits with negative correlation between maternal and direct effects. The increase in accuracies can be attributed to the genomic matrix which provides a more precise measure of genetic relatedness. The numerator relationship matrix uses pedigree information to derive the probability of genes to be identical by descent that gives an estimate of the relatedness of individuals. The genomic matrix, in contrast, uses high-density marker information, thus, it can identify genes that are identical by state[9] or genes that may be shared through common ancestor not recorded in the pedigree (Forni *et al*., 2011). Hence, it provides a more accurate measure for the relationship between individuals. It also enables better differentiation among closely related individuals since it captures Mendelian sampling with greater precision. Thus, the use of the marker based relationship matrix in the unified approach will greatly improve the accuracy of the estimated breeding values for low heritability traits and/or negatively correlated traits, e.g. traits with negatively correlated maternal and direct effects.

**5.3. The 44k SNP assay** (The definitive version is available at www.blackwell-synergy.com)
A total of 44,000 SNP were taken to construct the SNP assay out of which 36,000 are validated for the association to the trait of interest. Out of these 36,000 SNP, 813 SNP were significant between pools A and B as well as between pools A and C, 1116 SNP were significant only

[9] Two alleles may be called identical by state if the alleles are identical but they do not originate from same ancestral allele.

between pools A and B and 6,965 SNP were significant only between pools A and C (pFDR < 0.05). The number of significant SNP between pool A and C was higher as compared to pool A and B. This is probably due to the fact that the causative genes are not segregating in pool C which never showed the hygienic behaviour. The criteria for SNP selection for the assay, particularly the pFDR of <0.05, were chosen in a way that favours a certain amount of false-positive SNP. To provide an even distribution of SNP across the genome, additional 8,000 SNP were selected from the SNP list (www.hgsc.bcm.tmc.edu/ftp-archive/Amellifera/snp/) published by the Honey Bee Genome Project without prior validation. The statistical test provided a preliminary estimate of the number of genomic regions involved in the regulation of the investigated trait (Spötter *et al.*, 2012). Seven SNP are significant at a pFDR < $10^{-9}$ between pools A and B. 120 SNP are significant at a pFDR < $10^{-9}$ between pools A and C. One SNP is significant at a pFDR < $10^{-9}$ for pool between pools A and B as well as pools A and C. Next-generation sequencing was performed using the Illumina Genome Analyzer GA-IIx to validate the large subset of SNP that was used to design the 44k custom genotyping SNP assay. To the best of the knowledge, it is the first time that next generation sequencing has been used in the honey bee. In subsequent studies, these SNP will be evaluated for association with the defence behaviour.

# Chapter 6. Conclusion and Future work

## 6.1. Conclusion

This study describes the use of a high-density molecular marker data for genetic evaluation in honey bees using the unified approach. To provide a comparative evaluation between the genetic evaluation methods based on the unified approach and the pedigree based approach, a complex scenario was modelled by taking into account characteristics such as varying heritability, correlation between maternal and direct genetic effects, uncertain paternity and the honey bee specific genetic and reproductive biology. To the best of the knowledge, this is the first study which gives background knowledge about the simulation of genomic and pedigree datasets in the honey bee for genetic evaluation, thus, providing an important framework for future studies. The results showed improvement in the accuracy of the overall estimated breeding values with the unified approach for both no correlation and negative correlation between maternal and direct effects at all simulated heritabilities. The unified approach can be further used to improve the response to selection, increase genetic gain, lower the rate of inbreeding and speed up the selection procedure as a result of reduction in the generation interval. Thus, the unified approach will be a progressive step for the genetic evaluation in the honey bee.

Furthermore, the software program developed in this study is relevant for research requiring a simulated molecular genetic dataset in the honey bee such as studies aiming at optimizing the honey bee breeding programs. It can construct a base population in LD by simulating a random mating honey bee population given some input population parameters. It provides the statistics relevant to a population such as allele frequency, $r^2$ value for LD and data for marker sorting according to minor allele frequency and Hardy-Weinberg equilibrium.

The developed 44k SNP assay will be employed to perform genome-wide association studies to identify QTL associated with the hygienic behaviour in the honey bee. Previous studies (Rothenbuhler, 1964; Moritz, 1988; Lapidge *et al*., 2002; Oxley *et al*., 2010) have suggested the existence of regions in the genome controlling hygienic behaviour. However, to date, no study employing a large-scale SNP assay has been accomplished. Although, the SNP assay has been developed to identify the SNP associated with the uncapping of *Varroa*-infested brood, its usage

is not restricted to the analysis of the defence behaviour against *Varroa*. It can also be used for other traits such as swarming tendency, calmness or honey and wax production. This is due to the fact that the SNP selected for the assay are evenly distributed across the whole genome which will help in the identification and detection of QTL affecting other traits as well. Additionally, the 44k SNP assay will facilitate the implementation of marker-assisted and genomic selection methodologies that rely on high-density marker information to obtain accurate breeding value estimates.

To summarize, this study provides a comprehensive overview of a marker based genetic evaluation methodology in the honey bee through simulation and modelling. The use of SNP data in the honey bee for association studies and genetic evaluation will provide a great impetus to the genetic improvement of economically important traits in the honey bee. This study provides important background knowledge about genetic evaluation based on molecular genetic data that will significantly help future studies in the honey bee.

## 6.2. Future work

In this study only potential-dam queens were genotyped. With the fast advancement in genotyping technology, it is likely that the cost of genotyping is reduced further in future, thus making it possible for the sister queens in a multiple mating station or queens contributing drones in artificial insemination to be genotyped. This can help to improve the accuracy of the breeding values in two ways. Firstly, genotyping can help to predict the actual sires for queens instead of using a dummy sire, thus the numerator relationship matrix can be constructed precisely, and secondly, the genotyping information for the sires can be directly included in the genomic relationship matrix.

Another aspect that needs to be studied in the honey bee population is the genome-wide LD, an important parameter in all genome-wide association studies. As compared to other species, the extent of LD in the breeding population of honey bee is uncertain. The level of LD in species such as cattle, pig and sheep is reported in literature (Du *et al*., 2001; McRae *et al*., 2002; McKay *et al*., 2007). However, to the best of the knowledge, no study has been published which reports the extent of LD in the honey bee population. Factors which contribute to uncertainty in the level of LD in the honey bee are as follows:

(1) The extent of LD will vary according the effective population size of the honey bee population in a region (Estoup *et al*., 1995).

(2) The level of LD is inversely related to the genetic distance. Although the physical size of the honey bee genome is approximately 236 Mb (The Honeybee Genome Sequencing Consortium, 2006), the genetic length is much higher than in other species due to a very high recombination rate. Thus, this high genetic length can have a great influence on the extent of LD in the honey bee population.

(3) Population admixture plays an important role in creating LD. Given that honey bee population is not closed, population admixture may result from uncontrolled mating and introduction of new bees from different regions.

(4) The extent of LD may vary in different sub-species of the honey bee such as *Apis mellifera carnica* in Middle Europe, *Apis mellifera linguistica* in Italy or *Apis mellifera mellifera* in Northern Europe.

Studies in other species (Aguilar *et al*., 2010; Forni *et al*., 2011) have already optimised the approach with respect to the construction of the genomic relationship matrix and computational solving procedures. Thus, for genetic evaluation based on the unified approach and/or different genomic selection strategies, there is a scope to develop software that precisely takes into account the population structure of the honey bee.

Furthermore, the developed 44k SNP assay will be used in the future genome-wide association studies to detect QTL/genes associated with various economically important traits in the honey bee.

# Bibliography

Adams J, Rothman ED, Kerr WE, Paulino ZL (1977). Estimation of the number of sex alleles and queen matings from diploid male frequencies in a population of *Apis mellifera*. *Genetics* **86**: 583-596.

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. *Journal of Dairy Science* **93**: 743-752.

Albright SC, Winston WL, Zappe CJ (2011). *Data Analysis and Decision Making*, 4th edn. Cengage Learning, USA.

Allen MF, Ball BV (1996). The incidence and world distribution of honey bee viruses. *Bee World* **77**: 141-162.

Allen MF, Ball BV, White RF, Antoniw JF (1986). The detection of acute paralysis virus in *Varroa jacobsoni* by the use of a simple indirect ELISA. *Journal of Apicultural Research* **25**: 100-105.

Bakonyi T, Farkas R, Szendroi A, Dobos-Kovács M, Rusvai M (2002). Detection of acute bee paralysis virus by RT-PCR in honey bee and *Varroa destructor* field samples: rapid screening of representative Hungarian apiaries. *Apidologie* **33**: 63-74.

Balding DJ, Bishop M, Cannings C (2007). *Handbook of statistical genetics*, 3rd edn. John Wiley and Sons, UK.

Ball BV (1985). Acute paralysis virus isolates from honeybee colonies infested with *Varroa jacobsoni*. *Journal of Apicultural Research* **24**: 115-119.

Ball BV (1989). *Varroa jacobsoni* as a virus vector. ln: Present status of Varroatosis in Europe and progress in the *Varroa* mite control, Cavalloro R (ed.), Commission of the European Communities, Luxembourg, pp. 241-244.

Ball BV, Allen MF (1988). The prevalence of pathogens in honey bee colonies infested with the parasitic mite *Varroa jacobsoni*. *Annals of Applied Biology* **113**: 237-244.

Batuev YM (1979). New information about virus paralysis. *Pchelovodstvo* **7**: 10-11.

Bennewitz J, Reinsch N, Thomsen H, Szyda J, Reinhart F, Kuhn C, Schwerin M, Erhardt G, Weimann C, Kalm E (2003). Marker assisted selection in German Holstein dairy cattle breeding: outline of the program and marker assisted breeding value estimation. In: Annual Meeting of European Association for Animal Production, 54th Session G1.9, Rome, Italy.

Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O, Sirviö A, Guzmán-Novoa E, Hunt G, Solignac M, Page RE Jr (2006). Exceptionally high levels of recombination across the honey bee genome. *Genome Research* **16**: 1339-1344.

Beye M, Hasselmann M, Fondrk MK, Page RE, Omholt SW (2003). The gene *csd* is the primary signal for sexual development in the honeybee and encodes an SR-type protein. *Cell* **114**: 419-429.

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). Noticeable success in honey bee selection after the introduction of genetic evaluation by 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. *Annals of the Entomological Society of America* **84**: 324-331.

Bienefeld K, Zautke F, Pronin D, Mazeed A (1999). Recording the proportion of damaged *Varroa jacobsoni* Oud. in the debris of honey bee colonies (*Apis mellifera*). *Apidologie* **30**: 249-256.

Bijma P (2006). Estimating maternal genetic effects in livestock. *Journal of Animal Science* **84**: 800-806.

Boecking O, Bienefeld K, Drescher W (2000). Heritability of the Varroa-specific hygienic behaviour in honey bees (Hymenoptera: Apidae). *Journal of Animal Breeding and Genetics* **117**: 417-424.

Boecking O, Drescher W (1992). The removal responses of *Apis mellifera* L. colonies to brood in wax and plastic cells after artificial and natural infestation with *Varroa jacobsoni* Oud. and to freeze-killed brood. *Experimental and Applied Acarology* **16**: 321-332.

Boichard D, Fritz S, Rossignol MN, Boscher MY, Malafosse A, Colleau JJ (2002). Implementation of marker-assisted selection in French dairy cattle. In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, Communication no. 22-03.

Bowen-Walker PL, Martin SJ, Gunn A (1999). The transmission of deformed wing virus between honey bees (*Apis mellifera* L.) by the ecto-parasitic mite *Varroa jacobsoni* Oud. *Journal of Invertebrate Pathology* **73**, 101-106.

Brown MJF, Paxton RJ (2009). The conservation of bees: a global perspective. *Apidologie* **40**: 410-416.

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

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**: 553-561.

Chen Y, Pettis JS, Evans JD, Kramer M, Feldlaufer MF (2004). Transmission of Kashmir bee virus by the ectoparasitic mite *Varroa destructor*. *Apidologie* **35**: 441-448.

Chevalet C, Cornuet JM (1982). Étude théorique sur la sélection du caractère “production de miel” chez l’abeille I. Modèle génétique et statistique. *Apidologie* **13**: 39-65.

Christensen OF, Lund MS (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution* **42**: 2.

Cornuet JM, Chevalet C (1987). Étude théorique sur la sélection du caractère “production de miel” chez l’abeille II. Plan de sélection combinée de reines en fécondation naturelle. *Apidologie* **18**: 253-266.

Costa-Maia FM, de Toledo VAA, Martins EN, Lino-Lourenço DA, Sereia MJ, de Oliveira CAL, Faquinello P, Halak AL (2011). Estimates of covariance components for hygienic behavior in Africanized honeybees (*Apis mellifera*). *Revista Brasileira de Zootecnia* **40**: 1909-1916.

Davierwala AP, Reddy AP, Lagu MD, Ranjekar PK, Gupta VS (2001). Marker assisted selection of bacterial blight resistance genes in rice. *Biochemical Genetics* **39**: 261-278.

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, Mullaart E, Calus MPL, Veerkamp RF (2007). Breeding value estimation for fat percentage using dense markers on *Bos taurus* autosome 14. *Journal of Dairy Science* **90**: 4821-4829.

Dekkers JCM (1999). Breeding values for identified quantitative trait loci under selection. *Genetics Selection Evolution* **31**: 421-436.

Dekkers JCM (2004). Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *Journal of Animal Science* **82**: E313-E328.

Dekkers JCM, Hospital F (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics* **3**: 22-32.

Devlin B, Risch N (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311-322.

Du FX, Clutter AC, Lohuis MM (2001). Characterizing linkage disequilibrium in pig populations. *International Journal of Biological Sciences* **3**: 166-178.

Estoup A, Garnery L, Solignac M, Cornuet JM (1995). Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* **140**: 679-695.

Falconer DS, Mackay TFC (1996). Introduction to Quantitative Genetics, 4th edn. Longmans Green, Essex, UK.

Fernando RL and Grossman M (1989). Marker-assisted selection using best linear unbiased prediction. *Genetics Selection Evolution* **21**: 467–477.

Fisher RA (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**: 399-433.

Flint-Garcia SA, Darrah LL, McMullen MD, Hibbard BE (2003). Phenotypic versus marker-assisted selection for stalk strength and second-generation European corn borer resistance in maize. *Theoretical and Applied Genetics* **107**: 1331–1336.

Forni S, Aguilar I, Misztal I (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics Selection Evolution* **43**: 1.

Franck P, Garnery L, Solignac M, Cornuet JM (1998). The origin of west European subspecies of honeybees (*Apis mellifera*): new insights from microsatellite and mitochondrial data. *Evolution* **52**: 1119-1134.

Gengler N, Mayeres P, Szydlowski M (2007). A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal* **1**: 21-28.

Gilmour AR, Thompson R, Cullis BR (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**: 1440-1450.

Groeneveld E, Kovac M, Wang T (1990). PEST, a general purpose BLUP package for multivariate prediction and estimation. In: Proceedings of the 4th World Congress on Genetics applied to Livestock Production, Edinburgh, UK, pp. 488-491.

Gupta P, Conrad T, Spötter A, Reinsch N, Bienefeld K (2012). Simulating a base population in honey bee for molecular genetic studies. *Genetics Selection Evolution* **44**: 14.

Haldane JBS (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**: 299-309.

Hamilton M (2009). Population Genetics. Wiley–Blackwell, Chichester, UK.

Harbo JR, Harris JW (1999). Heritability in honey bees (Hymenoptera: Apidae) of characteristics associated with resistance to *Varroa jacobsoni* (Mesostigmata: Varroidae). *Journal of Economic Entomology* **92**: 261-265.

Harville DA (1977). Maximum Likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**: 320-338.

Hasselmann M, Beye M (2004). Signatures of selection among sex-determining alleles of the honey bee. *Proceedings of the National Academy of Sciences* **101**: 4888-4893.

Hayes B (2008). QTL Mapping, MAS, and Genomic Selection. Course Notes. [http://www.sabre-eu.eu/Portals/0/SABRE%20Publications/Course_Notes_QTL_Mapping_MAS_and_Genomic_Selection.pdf]

Hayes B, Goddard ME (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* **33**: 209-229.

Hayes B, Goddard ME (2003). Evaluation of marker assisted selection in pig enterprises. *Livestock Production Science* **81**: 197-211.

Hayes BJ, Visscher PM, Goddard ME (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research* **91**: 47-60.

Henderson CR (1950) Estimation of genetic parameters. *Annals of Mathematical Statistics* **21**: 309-310.

Henderson CR (1963). Selection index and expected genetic advance. In: Statistical Genetics and Plant Breeding, Hanson WD, Robinson HF (eds.), Publication 982, National Academy of Sciences, National Research Council, Washington, DC, USA, pp. 141-163.

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 predicting of breeding values. *Biometrics* **32**: 69-83.

Henderson CR (1984). Best linear unbiased prediction of performance and breeding value. Notes. [http://www.poultryscience.org/docs/pba/1952-2003/1984/1984%20Henderson.pdf]

Henderson CR (1988). Theoretical basis and computational methods for a number of different animal models. *Journal of Dairy Science* **71**: 1-16.

Henderson CR, Kempthorne O, Searle SR, von Krosigk CM (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**: 192-218.

Hill WG (1975). Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. *Theoretical Population Biology* **8**: 117-126.

Hill WG (1981). Estimation of effective population size from data on linkage disequilibrium. *Genetical Research* **38**: 209-216.

Hill WG, Robertson A (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* **38**: 226-231.

Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin GD, Brizuela L, McCombie WR, Hannon GJ (2009). Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nature Protocols* **4**: 960-974.

Israel C, Weller JI (1998). Estimation of candidate gene effects in dairy cattle populations. *Journal of Dairy Science* **81**: 1653-1662.

Johnson DL, Thompson R (1995). Restricted Maximum Likelihood Estimation of variance components for univariate animal models using sparse matrix techniques and average information. *Journal of Dairy Science* **78**: 449-456.

Karaboga D (2005). An idea based on honey bee swarm for numerical optimization. Technical Report TR06, Erciyes University, Engineering Faculty, Computer Engineering Department.

Karaboga D, Akay B (2009). A survey: algorithms simulating bee swarm intelligence. *Artificial Intelligence Review* **31**: 68-85.

Kerr WE (1967). Multiples alleles and genetic load in bees. *Journal of Apicultural Research* **6**: 61-64.

Kerr WE, Zucchi R, Nakaida JT, Butolo JE (1962). Reproduction in the social bees (Hymenoptera: Apidae). *Journal of the New York Entomological Society* **70**: 265-276.

Kimura M (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61**: 893-903.

Kucharski R, Maleszka J, Foret S and Maleszka R (2008). Nutritional control of reproductive status in honeybees via DNA methylation. *Science* **319**: 1827-1830.

Lapidge KL, Oldroyd BP, Spivak M (2002). Seven suggestive quantitative trait loci influence hygienic behavior of honey bees. *Naturwissenschaften* **89**: 565–568.

Larsgard AG, Olesen I (1998). Genetic parameters for direct and maternal effects on weights and ultrasonic muscle and fat depth of lambs. *Livestock Production Science* **55**: 273-278.

Legarra A, Aguilar I, Misztal I (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* **92**: 4656-4663.

Legarra A, Robert-Granié C, Manfredi E, Elsen JM (2008). Performance of genomic selection in mice. *Genetics* **180**: 611-618.

Lewontin RC (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**: 49-67.

Lewontin RC, Kojima K (1960). The evolutionary dynamics of complex polymorphisms. *Evolution* **14**: 458-472.

McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, Crews D, Neto ED, Gill CA, Gao C, Mannen H, Stothard P, Wang Z, Van Tassell CP, Williams JL, Taylor JF2, Moore SS (2007). Whole genome linkage disequilibrium maps in cattle. *BMC Genetics* **8**: 74.

McRae AF, McEwan JC, Dodds KG, Wilson T, Crawford AM, Slate J (2002). Linkage disequilibrium in domestic sheep. *Genetics* **160**: 1113-1122.

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

Meuwissen THE, Luan T, Woolliams JA (2011). The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. *Journal of Animal Breeding and Genetics* **128**: 429-439.

Meuwissen THE, Luo Z (1992). Computing inbreeding coefficients in large populations. *Genetics Selection Evolution* **24**: 305-313.

Meyer M, Kircher M (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* **2010**: pdb.prot5448.

Milani N (1999). The resistance of *Varroa jacobsoni* Oud. to acaricides. *Apidologie* **30**: 229-234.

Misztal I, Legarra A, Aguilar I (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science* **92**: 4648-4655.

Moritz RFA (1988). A reevaluation of the two-locus model hygienic behaviour in honey bees, *Apis mellifera* L. *Journal of Heredity* **79**: 257-262.

Mousseau TA, Fox CW (1998). Maternal effects as adaptations. Oxford University Press, New York, USA.

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

Neumann P, Carreck NL (2010). Honey bee colony losses. *Journal of Apicultural Research* **49**: 1-6.

Nolan WJ (1937). Bee Breeding. United States Department of Agriculture Yearbook, pp 1396-1418. United States Department of Agriculture, Washington, DC, USA.

Nordström S, Fries I, Aarhus A, Hansen H, Korpela S (1999). Virus infections in Nordic honey bee colonies with no, low or severe *Varroa jacobsoni* infections. *Apidologie* **30**: 475-484.

Oldroyd BP, Thompson GJ (2007). Behavioural genetics of the honey bee *Apis mellifera*. *Advances in insect physiology* **33**: 1-49.

Ostersen T, Christensen OF, Henryon M, Nielsen B, Su G, Madsen P (2011). Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. *Genetics Selection Evolution* **43**: 38.

Oxley PR, Spivak M, Oldroyd BP (2010). Six quantitative trait loci influence task thresholds for hygienic behavior in honeybees (*Apis mellifera*). *Molecular Ecology* **19**: 1452-1461.

Patterson HD, Thompson R (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545-554.

Peng B, Kimmel M (2005). SimuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**: 3686-3687.

Quaas RL (1976). Computing the diagonal elements of a large numerator relationshipmatrix. *Biometrics* **32**: 949-953.

Quaas RL (1995). Fx algorithms. An unpublished note.

Quaas RL, Pollak EJ (1980). Mixed model methodology for farm and ranch beef cattle testing programs. *Journal of Animal Science* **51**: 1277-1287.

Räsänen K, Kruuk LEB (2007). Maternal effects and evolution at ecological time-scales. *Functional Ecology* **21**: 408-421.

Rinderer TE (1986). Selection. In: Bee genetics and breeding, T. E. Rinderer (ed.). Academic Press, Orlando, Florida, USA, pp. 305-321.

Rinderer TE, Harris JW, Hunt GJ, de Guzman LI (2010). Breeding for resistance to *Varroa destructor* in North America. *Apidologie* **41**: 409-424.

Roberts WC (1944). Multiple mating of queen bees proved by progeny and flight tests. *Gleanings in Bee Culture* **72**: 255-260.

Robinson GK (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**: 15-51.

Roehe R, Kennedy BW (1993). The influence of maternal effects on accuracy of evaluation of litter size in swine. *Journal of Animal Science* **71**: 2353-2364.

Rothenbuhler WC (1958). Genetics and breeding of the honey bee. *Annual Review of Entomology* **3**: 161-180.

Rothenbuhler WC (1964). Behavior genetics of nest cleaning in honey bees. IV. Responses of F1 and backcross generations to disease-killed brood. *American Zoologist* **12**: 578-583.

Roubik DW (1995). Pollination of cultivated plants in the Tropics. Food and Agriculture Organization of the United Nations, Agricultural Bulletin No. 118, Rome, Italy, pp. 196.

Ruttner F (1956). The mating of the honeybee. *Bee World* **37**: 3-15.

Ruttner F (1988). Biogeography and taxonomy of honeybees. Springer Verlag, Berlin, Germany.

Safari E, Fogarty NM, Gilmour AR (2005). A review of genetic parameter estimates for wool, growth, meat and reproduction traits in sheep. *Livestock Production Science* **92**: 271-289.

Santillán-Galicia MT, Ball BV, Clark SJ, Alderson PG (2010). Transmission of deformed wing virus and slow paralysis virus to adult bees (*Apis mellifera* L.) by *Varroa destructor*. *Journal of Apicultural Research* **49**: 141-148.

Schaeffer LR (2010). Linear models and animal breeding. Course notes. [http://www.aps.uoguelph.ca/~lrs/ABModels/DATA/EuropeNotes.pdf]

Searle SR, Casella G, McCulloch CE (1992). Variance components. John Wiley and Sons,New York, USA.

Shrimpton AE, Robertson A (1988). The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. II. Distribution of third chromosome bristle effects within chromosome sections. *Genetics* **118**: 445-459.

Slatkin M (2008). Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**: 477-485.

Sonesson AK (2007). Within-family marker-assisted selection for aquaculture species. *Genetics Selection Evolution* **39**: 301-317.

Sonesson AK, Meuwissen THE (2009). Testing strategies for genomic selection in aquaculture breeding programs. *Genetics Selection Evolution* **41**: 37.

Spivak M (1996). Hygienic behavior and defense against *Varroa jacobsoni*. *Apidologie* **27**: 245-260.

Spivak M, Gilliam M (1998). Hygienic behaviour of honey bees and its application for control of brood diseases and varroa mites. Part I: Hygienic behaviour and resistance to American foulbrood. *Bee World* **79**: 124-134.

Splan RK, Cundiff LV, Dikeman ME, Van Vleck LD (2002). Estimates of parameters between direct and maternal genetic effects for weaning weight and direct genetic effects for carcass traits in crossbred cattle. *Journal of Animal Science* **80**: 3107-3111.

Spötter A, Gupta P, Nürnberg G, Reinsch N, Bienefeld K (2012) Development of a 44K SNP assay focussing on the analysis of a varroa-specific defense behavior in honey bees (*Apis mellifera carnica*). *Molecular Ecology Resources* **12**: 323-332.

Storey JD (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society* **64**: 479-498.

Tarpy DR, Page RE (2000). No behavioral control over mating frequency in queen honey bees (*Apis mellifera* L.): Implications for the evolution of extreme polyandry. *The American Naturalist* **155**: 820-827.

Thakur RK, Bienefeld K, Keller R (1997). *Varroa* defense behavior in *A. mellifera carnica*. *American Bee Journal* **137**: 143-148.

Thallman RM (2009). Whole genome selection. University of California at Davis. [http://animalscience.ucdavis.edu/animalbiotech/Outreach/Whole_Genome_Selection.pdf]

The Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **443**: 931-949.

Trjasko WW (1951). Repeated and multiple matings of queens (in Russian). *Pchelovodstvo* **16**: 43-50.

van Engelsdorp D, Otis GW (2000). Application of a modified selection index for honey bees (Hymenoptera: Apidae). *Journal of Economic Entomology* **93**: 1606-1612.

VanRaden PM (2008). Efficient methods to compute genomic prediction. *Journal of Dairy Science* **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. *Journal of Dairy Science* **92**: 16-24.

Visscher PM, Haley CS (1995). Utilizing genetic markers in pig breeding schemes. *Animal Breeding Abstracts* **63**: 1-8.

Vitezica ZG, Aguilar I, Misztal I, Legarra A (2011). Bias in genomic predictions for populations under selection. *Genetics Research* **93**: 357-366.

Weller JI (2001). Quantitative Trait Loci Analysis in Animals. CABI Publishing,Wallingford, UK .

Willham RL (1963). The covariance between relatives for characters composed of components contributed by related individuals. *Biometrics* **19**: 18-27.

Willham RL (1972). The role of maternal effects in animal breeding: III. Biometrical aspects of maternal effects in animals. *Journal of Animal Science* **35**: 1288-1293.

Woyke J (1960). Naturalne i stzuczne unasienianie matek pszczelieh. *Pszczelnicze Zeszyty Naukowe* **4**: 183-273.

Woyke J (1963). What happens to diploid drone larvae in a honeybee colony. *Journal of. Apicultural Research* **2**: 73-75.

Wright S (1933). Inbreeding and homozygosis. *Proceedings of the National Academy of Sciences* **19**: 411-420.

Wu R, Ma CX, Casella G (2007). Statistical genetics of Quantitative traits. Springer, New York, USA.

Zhao H, Nettleton D, Soller M, Dekkers JCM (2005). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genetical Research* **86**: 77-87.

Zhao HH, Fernando RL, Dekkers JCM (2007). Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. *Genetics* **175**: 1975-1986.

# Appendix

## Instructions to use the software program for simulating a base population

The code has been written keeping in mind the biology of the honey bee. As per our knowledge, it is the first study which specifically deals with the genetic and reproductive aspects of the honey bee. The software uses a straight-forward approach for simulating a base population required for population genetics related studies in the honey bee. The code is in MATLAB but can easily be adapted to the open source version Octave. We hope that the users are familiar with MATLAB.

### Usage and Input:

The following syntax can be entered at the command window of MATLAB to execute the function:

population_simulation

(This would produce results according the default value for input arguments.)

The user can change the value for desired input arguments using the one of the following commands:

population_simulation(NoGen)

(This would allow the user to provide an input for the number of generations. For example, if population_simulation(50) is executed then the simulation will run for 50 generations and all other input arguments will use the default value.)

Similarly, values for other input arguments can be changed as follows:

population_simulation(NoGen,NoSire)

population_simulation(NoGen,NoSire,NoDam)

population_simulation(NoGen,NoSire,NoDam,NoMarker)

population_simulation(NoGen,NoSire,NoDam,NoMarker,FMrate)

population_simulation(NoGen,NoSire,NoDam,NoMarker,FMrate,BMrate)

population_simulation(NoGen,NoSire,NoDam,NoMarker,FMrate,BMrate,MAF)

population_simulation(NoGen,NoSire,NoDam,NoMarker,FMrate,BMrate,MAF,NoSelSNP)

Input includes:

1. NoGen: Number of generations to be simulated (default: 1000)
2. NoSire: Number of sire queens* (default: 200)
3. NoDam: Number of dam queens* (default: 20)
4. NoMarker: Total number of marker loci to be distributed on the genome (default: 50000)
5. FMrate: Forward mutation rate (default: 0.0025)
6. BMrate: Backward mutation rate (default: 0.0025)
7. MAF: Minor allele frequency (default: 0.05)
8. NoSelSNP: Number of marker loci with the highest minor allele frequency to be chosen as SNPs (default: 10000)

**Output**

1. File 'Graph_LD_r_square.fig' gives the plot of average LD for all simulated marker loci pair against the number of generations.
2. File ' LD_SNP_with_highest_MAF.dat' gives average LD for selected SNPs with the highest minor allele frequency in the last generation. It is a scalar value.
3. File 'Genome_m_LD.dat' contains genome of all the sire queens in the last generation. The size is: (Number of sire queens times 2) x (Number of SNP loci).
4. File 'Genome_f_LD.dat' contains genome of all the dam queens in the last generation. The size is: (Number of dam queens times 2) x (Number of SNP loci).
5. File 'Allele_frequency_LD.dat' contains allele frequency in the last generation. The size is: (Number of SNP loci) x (2).
6. File 'Avg_ld_all_gen.dat' contains the value of average LD for all simulated marker loci pair in all generations. The size is: (1) x (Number of generations).
7. File 'Chi_squ.dat' contains the value of Chi-Square statistics for all the SNP loci in the last generation. The size is: (1) x (Number of SNP loci).
8. File 'MAF_SNP.dat' contains the ID of SNP loci with minor allele frequency less than the input minor allele frequency. The size is: (1) x (Number of SNP loci with minor allele frequency less than the input value). These markers can be eliminated from the original set of markers simulated.

9. File 'Complete_snp_list.dat' contains the SNP position in base-pairs on each chromosome. SNP positions for each chromosome are concatenated one after the other. The size is: (1) x (Number of SNPs to be distributed).
10. File 'SNP_Position.dat' contains the information about a) Total number of SNPs on each chromosome and b) Position of SNP on each chromosome in base-pairs. The output file format is: Chromosome number (Number of SNPs on that chromosome) Positions of SNPs.

**Remark**

- The ratio between dam queen and sire queen has been assumed to be 1:10 therefore the values for number of sire queen and dam queen need to be assigned in the same ratio.
- The reported recombination rate of 19 cM/Mb is used in the code.
- Since the source code is provided, the values of input can be changed in the code itself using an editor.
- In the source code, random seed can be changed according to the requirement.
- Please provide a valid number of marker loci to be simulated. The genome consists of 16 chromosomes; therefore simulate a legitimate number of marker loci from which required number of SNPs can be chosen. The default value of ‘NoMarker’ is 10,000.

# Zusammenfassung

Über die vergangenen Jahre hinweg ist der Bestand der am häufigsten domestizierte Honigbiene (*Apis mellifera*) drastisch zurückgegangen. Dies ist hauptsächlich auf den Befall mit der ectoparasitären Milbe *Varroa destructor* zurückzuführen. Selektives Züchten von genetisch überlegenen Bienen kann dabei helfen, resistente Abstammungslinien zu erzeugen und somit den Verlusten aufgrund des Parasiten vorzubeugen. Desweiteren können auch andere ökonomisch wichtige Merkmale verbessert werden, z.B. Honigleistung, Schwarmtrieb und Angriffsverhalten. Ein solcher Ansatz setzt jedoch ein robustes Zuchtprogramm und die präzise Schätzung genetischer Zuchtwerte voraus, mit denen genetisch überlegene Individuen identifiziert und selektiert werden können. Diese Bewertung kann anhand verschiedener Informationsquellen erfolgen, z.B. Phänotypen, Genotypen und Abstammung der Individuen.

Diese Arbeit befasste sich mit der Eingliederung von Informationen über dichtverteilte Einzelnukleotid-Polymorphismen ('single nucleotide polymorphism', SNP) in den 'unified approach' zur Zuchtwertschätzung bei Honigbienen. Mit Hilfe von Simulationen wurden das Potential und die Anwendbarkeit dieses Ansatzes auf Honigbienen untersucht. Es musste eine Grundstruktur zur Simulation einer Honigbienenpopulation entwickelt werden, welche die honigbienenspezifischen Reproduktions- und Genomeigenschaften mit einbezog. Das beinhaltete eine hohe genetische Rekombinationsrate, haplo-diploide Geschlechtsbestimmung, Polyandrie, ungewisse Paternität und negative genetische Korrelation zwischen maternalen und direkten Effekten. Dadurch konnten Datensätze für die Abstammung, die Genotypen und die Phänotypen aller Individuen einer Honigbienenpopulation generiert werden, welche für die Implementierung des 'unified approach' erforderlich waren. Die linearen 'mixed model' Gleichungen wurden mit einem weit verbreiteten Zuchtwertschätzverfahren ('best linear unbiased prediction', BLUP) auf Grundlage des 'unified approach' gelöst. Ein besonderes Augenmerk lag auf den Auswirkungen der maternalen Effekte, negativer Korrelation zwischen maternalen und direkten Effekten, ungewisser Paternität und unterschiedlicher Heritabilität von maternalen und direkten Effekten. Dadurch sind die Ergebnisse dieser Studie auch wertvoll für die Untersuchung anderer Zuchttiere mit ähnlichen Eigenschaften. Zusätzlich wurde ein Testverfahren auf der Basis eines 44.000 SNP

Arrays entworfen, welches für genomweite Assoziationsstudien und markergestützte Selektionsstrategien verwendet werden kann.

Diese ist die erste Studie, die die Details zur Modellierung und Simulation von Genom- und Abstammungsdatensätzen für die Zuchtwertschätzung von Honigbienen untersucht. Die dadurch erworbenen Kenntnisse bieten eine solide und wertvolle Grundlage für zukünftige Untersuchungen auf diesem Gebiet. Die Umsetzung des 'unified approach' bietet eine fortschrittliche Verbesserung der genetischen Bewertung von Honigbienen. Daher ist diese Studie wegweisend für die aktuelle Forschung auf dem Gebiet der markergestützten Zuchtwertschätzung von Honigbienen und anderen Zuchttieren.