Of Mites and Men:
The independent evolution of host-induced *Varroa* infertility in the drone brood of *Apis mellifera*

Dissertation
zur Erlangung des
Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der

Naturwissenschaftlichen Fakultät I – Biowissenschaften –

der Martin-Luther-Universität
Halle-Wittenberg,

vorgelegt

von Herrn Benjamin Hanson Conlon

geb. am 17/07/1991 im Liverpool, U.K.

Gutachter / in:

1. Prof. Dr. Robin F.A. Moritz (Molecular Ecology group Martin-Luther-Universität, Halle, Deutschland)

2. Prof. Dr. Robert J. Paxton (General Zoology group, Martin-Luther-Universität, Halle, Deutschland)

3. Prof. Dr. Marla Spivak (Department of Entomology, University of Minnesota, St. Paul, MN, USA)

Promotionsgesuch eingereicht am: 12.06.2018

Tag der öffentlichen Verteidigung: 03.09.2018

Contents

1. Introduction

1.1 Host-parasite coevolution

The appearance of a novel parasite can be devastating to host populations. Following an initial reduction in population size, selection is expected to favour the evolution and maintenance of parasite resistance in the host (Fries, et al., 2006). The typically shorter life cycle and faster rate of reproduction in parasites could leave the host at an evolutionary disadvantage; with the rate of evolution for parasite counter-defences exceeding that of host defences (Hamilton, et al., 1990). However sexual reproduction and epistatic interactions in the host can help balance the relationship by increasing the rate at which new genotype combinations can be created within a population (Maynard Smith, 1971; Hamilton, et al., 1990; Wilfert, et al., 2007; Kidner & Moritz, 2013).

While the selective pressures and genetic variation driving the evolution of host defences are expected to vary across a host species range (Büchler, et al., 2015; Thompson, 2005). Similar adaptations often evolve in response to the same pressure; despite different geographic and genetic origins (Locke, 2016; Oddie, et al., 2017). A promising model system for better understanding the independent evolution of a resistance trait between geographically isolated and varied host populations is the interaction between the honey bee (*A. mellifera*) and its brood parasitic mite *Varroa destructor.*

1.2 Varroa destructor

Originally a parasite of the Asian honey bee (*Apis cerana*), *Varroa* switched host to *A. mellifera* sometime during the early 20th Century (Oldroyd, 1999). *Varroa* is highly virulent on *A. mellifera* and many colonies die within three years of an initial infestation (Beaurepaire, et al., 2015; Rosenkranz, et al., 2010).

Varroa females infest honey bee larval cells shortly before capping. Five hours post-capping, the mother mite consumes her first haemolymph meal, from the pupa, and initiates oogenesis a few hours later (Garrido & Rosenkranz, 2004). If the *Varroa* mother fails to initiate oogenesis in this narrow time window, she will remain infertile for the duration of the honeybee pupation (Frey, et al., 2013). Ovary activation can be induced using cuticular hydrocarbon and haemolymph-based cues from the pre-pupae. This suggests a compound, received from the pupa during the mother mite's first blood meal, is necessary for the successful reproduction of *Varroa* (Aumeier, et al., 2002; Frey, et al., 2013). Therefore, a change in these cues could provide a pathway to the inhibition of mite reproduction.

Upon hatching, the offspring feed on the pupae and mate in the cell; this is the only time in their lives when *Varroa* will mate (Donze & Guerin, 1997; Kanbar & Engels, 2005; Rosenkranz, et al., 2010). Mature, mated female mites leave the cell with the eclosing bee while the male and any immature female mites desiccate in the cell (Donze & Guerin, 1997; Rosenkranz, et al., 2010). This creates a close relationship between pupation time and *Varroa* fitness which means mother mites will only initiate oogenesis if they infest the cell when the correct larval cues are present. If the conditions become sub-optimal, the mother mite can suspend oogenesis (Nazzi & M, 1996; Frey, et al., 2013).

1.3 The evolution of host defences

Under natural evolutionary and ecological conditions, the introduction of a novel, highly virulent parasite is expected to lead to the rapid evolution of resistance traits in the host population. This is expected to occur through a collapse and bottleneck in the host population, resulting in a reduction in the virulence of the parasite and a more stable host-parasite relationship (Locke & Fries, 2011; Fries, et al., 2006). However, in the case of *A. mellifera* and *Varroa*, the host's economic importance has changed the expected evolutionary trajectory of the parasite (Rosenkranz, et al., 2010). In Europe, the high density of managed colonies and relatively low frequency of feral colonies (Kohl & Rutschmann, 2018) means that the vast majority of honey bee colonies are treated with acaricides; removing the selective pressure for the evolution of *Varroa*-resistance (Fries & Bommarco, 2007). This means that other selective pressures, such as local environmental adaptations, are stronger drivers of evolution in European *A. mellifera* populations than *Varroa* resistance (Büchler, et al., 2015). However, a strong selective pressure for local adaptation on acaricide-treated honey bee colonies means that, when colonies are left untreated, the genetic variation from which resistance traits can evolve should differ between locations (Kefuss, et al., 2004; Fries, et al., 2006; Le Conte, et al., 2007; Wallberg, et al., 2014; Büchler, et al., 2015; Oddie, et al., 2017).

1.4 Variation and similarities in the evolution of Varroa resistance

In cases where populations of European *A. mellifera* have been managed less, and not treated with acaricides, host resistance has evolved in under a decade (Fries, et al., 2006; Rosenkranz, et al., 2010; Locke & Fries, 2011; Kefuss, et al., 2015; Locke, 2016; Oddie, et al., 2017). Despite different genetic backgrounds, the independently-evolved resistance traits in *Varroa*-resistant honey bee populations are superficially very similar (Fries, et al., 2006; Le Conte, et al., 2007; Wallberg, et al., 2014; Locke, 2016; Oddie, et al., 2017). One of the most common resistance traits in European populations of *A. mellifera* is the inhibition of *Varroa* reproduction; a trait which is also shared with *Varroa*'s original host: *A. cerana* (Oldroyd, 1999; Kefuss, et al., 2004; Fries, et al., 2006; Le Conte, et al., 2007; Oddie, et al., 2017).

This thesis will focus on the inhibition of *Varroa* reproduction in the drone brood of *A. mellifera*. The longer pupation time of drone brood means that *Varroa* is able to produce more offspring which increases its population growth rate as well as the chance that a colony does not survive the winter (Rosenkranz, et al., 2010; van Dooremalen, et al., 2012). Despite being a relatively simple trait, the inhibition of *Varroa* reproduction appears to differ between host populations in which it has evolved (Le Conte, et al., 2007; Locke & Fries, 2011; Kurze, et al., 2016; Oddie, et al., 2017). This raises the possibility that there could be multiple mechanisms by which host populations can independently evolve the resistance trait.

Using Next-Generation sequencing, I explored the genomic basis for the independent evolution of the host-induced non-reproduction of *Varroa* in two resistant populations: One from Gotland, Sweden (Chapter 2) and one from Toulouse, France (Chapter 3). However, due to the identification of several misplaced scaffolds in the *Apis mellifera* 4.5 reference genome assembly (Elsik, et al., 2014), it was first necessary to construct a *de novo* genetic map (Chapter 1) before running analyses on *Varroa*-resistance in these populations.

Increasing recombination rate estimates result from decreasing assembly accuracy in honey bee (*Apis mellifera*) reference genome updates

Running head: Recombination in the honey bee (Apis mellifera)

Benjamin H. Conlon[a], Eike Oertelt[a], Robin F. A. Moritz[a], Jarkko Routtu[a]

[a]Molecular Ecology, Institute of Biology/Zoology, Martin-Luther-University Halle-Wittenberg, Hoher Weg 4, 06099 Halle an der Saale, Germany. Email: benjamin.conlon@zoologie.uni-halle.de; eike.oertelt@student.uni-halle.de; robin.moritz@zoologie.uni-halle.de; jarkko.routtu@zoologie.uni-halle.de

Corresponding author:

Benjamin H. Conlon, Current address: Molecular Ecology, Institute of Biology/Zoology, Martin-Luther-University Halle-Wittenberg, Hoher Weg 4, 06099 Halle an der Saale, Germany. Email: benjamin.conlon@zoologie.uni-halle.de, Phone: +49 345 55 26235.

Abstract

Current Next-Generation Sequencing platforms are limited in the length of reads they are able to produce; requiring the correct order to be determined algorithmically. While assembly algorithims can present a potential error-source, genetic pedigree data can be used to identify recombination events and, as recombination events are rare locally, test the order of sequences within a genome assembly. We use high-resolution population genomic data, from 80 brother drones, to test and compare the assembly quality of the three most recent reference genome assemblies for the western honey bee (*Apis mellifera*). As a model organism, there are several reference genomes available for *A. mellifera* with estimated recombination rates ranging from 19 cM/Mb to 37 cM/Mb. We identify variation in quality between *A. mellifera* reference genome assemblies with estimated recombination rates much higher than previous estimates. After performing *de novo* genetic map constructions, the estimated recombination rates become much closer to previous *de novo* map constructions. While internal marker order within scaffolds remained stable, at least 20% of scaffolds in the current *A. mellifera* 4.5 reference genome are mis-aligned. Our results provide an explanation for the large degree of variation in estimated recombination rates between *Apis mellifera* genome assemblies. That estimated recombination rates in our *de novo* assemblies are similar to previous estimates for the *A. mellifera* 4.5 reference genome assembly, which did not calculate recombination events across scaffold boundaries, supports our conclusion that mis-aligned scaffolds are the source of very high estimates of recombination rate in *A. mellifera* 4.5 genome assembly.

Keywords

Population; Linkage; Assembly; Sequencing; Map; Scaffold

Introduction

The prevalence of reference genomes (Collins *et al*., 2003; Nygaard and Wurm, 2015; Matasci *et al*., 2014; Grigoriev *et al*., 2013) makes sequencing a sub-sample of a genome more feasible for a wider range of experiments. Large volumes of genomic data can now be generated, aligned to a reference genome, and filtered for features of interest (Conlon *et al*., 2017; Wallberg *et al*., 2014; McCormack *et al*., 2013). The process is both less labour intensive than screening and amplifying individual loci using Polymerase Chain Reactions (PCR) (McCormack *et al*., 2013; Beaurepaire *et al*., 2017; Conlon *et al*., 2016; Behrens *et al*., 2011; Solignac *et al*., 2007) and less computationally intensive than *de novo* genome assembly (Nygaard and Wurm, 2015; Conlon *et al*., 2017). However, downstream processing of the results relies on the assumption that the reference genome is correct. The short read-lengths produced by current Next Generation Sequencing (NGS) platforms and the prevalence of repeat regions means that reference genomes are made up of distinct scaffolds rather than contiguous chromosome-length sequences. These are assembled algorithmically into chromosomes (Elsik *et al*., 2014), presenting a potential for errors due to low complexity regions in genomes (Nygaard and Wurm, 2015).

The error rate, in the algorithmic assembly of a reference genome, can be reduced by pairing it with population genetic data or experimental crosses (Solignac *et al*., 2007; Artemov *et al*., 2017; Beye *et al*., 2006; Weinstock *et al*., 2006; Roesti *et al*., 2013; Utsunomiya *et al*., 2016; Zeng *et al*., 2017). Using SNP or microsatellite markers to identify recombination events, a genetic map can be created for the genome. The decay of linkage disequilibrium with increased physical distance can then be used to identify incorrectly-located regions within the reference genome and estimate their true location (Utsunomiya *et al*., 2016).

As one of the first species identified for Whole Genome Sequencing (WGS) (Weinstock *et al.*, 2006), the western honey bee (*Apis mellifera*) has four available reference genome assemblies with both *de novo* and reference-genome-based genetic maps (Table 1) (Solignac *et al.*, 2007; Elsik *et al.*, 2014; Beye *et al.*, 2006). As F2 crosses using haploid drones can be analysed, this makes it an excellent candidate model system for testing the reproducibility of genomic assemblies. While previous genetic maps, constructed *de novo* based on recombination frequency between marker positions in the scaffolds of the reference genome, have consistently reported a recombination rate of 19-22 cM/Mb (Table 1.), the most recent version of the genome (Amel_4.5) (Elsik *et al.*, 2014) has provided estimates with much higher variability (Table 1.). In producing this genome, new scaffolds were formed through merging existing scaffolds or filling intra-scaffold gaps (Elsik *et al.*, 2014). These new scaffolds were then anchored to the previous reference genome assembly (Elsik *et al.*, 2014), which was itself anchored using a microsatellite-based map (Solignac *et al.*, 2007; Beye *et al.*, 2006; Weinstock *et al.*, 2006). This raises the possibility that large increases in recombination rate seen in genetic maps constructed using Amel_4.5 genome assembly are due to assembly errors rather than genuine recombination events.

Given its importance as a model organism, we seek to test the variation in quality among the three most recent genome assemblies for *A. mellifera*. The haplodiploid sex determination, found in *A. mellifera* and throughout the hymenoptera, greatly simplifies the task of identifying recombination events as drones possess only one allele per locus: removing any ambiguities associated with heterozygosity. This, combined with a wide range of available genetic data, means *A. mellifera* is unusually well suited for a study of this kind.

Methods

Sampling and DNA analyses

We sampled 80 *A. mellifera* drone offspring from a hybrid queen (Behrens *et al*., 2011), extracting DNA with a phenol/chloroform protocol (Garnery *et al*., 1991). The resulting extracts were assessed using a Nanodrop 1000 spectrophotometer (peqlab). The 80 specimens were split into three sequencing runs: one run of 16 specimens and two runs of 32 specimens. Library preparation was conducted using the RESTseq method (Stolle and Moritz, 2013) and individually barcoded drone samples underwent single-end sequencing using an IonTorrent Personal Genome Machine (Thermo Fisher).

SNP identification

Sequencing quality was checked using FastQC (Andrews, 2010) before barcodes and poor-quality sequences, with a phred score lower than 20, trimmed with cutadapt (Martin, 2011). After trimming, reads shorter than 25bp were discarded. The resulting sequences were mapped to the *Apis mellifera* 4.5 (downloaded from NCBI), 4.0 and 3.0 (downloaded from BeeBase) genome assemblies (Elsik *et al*., 2014; Weinstock *et al*., 2006; Kitts *et al*., 2016; Elsik *et al*., 2015) using the "MEM" algorithm and default parameters of the Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2010). The mapped reads were aligned to the reference genomes, skipping indels, with SAMtools' mpileup function (Li *et al*., 2009). Variant loci were then identified using the multi-allelic and rare-variant caller in BCFtools' call function (Li, 2011).

Defining of the Phase

The phase for each locus in a linkage group was determined by using an algorithm written in R (R Core Team, 2017). Detailed methods and the phasing script are found in Additional Files 4 and 5.

Briefly, the algorithm is based on the fact that recombination events are rare locally when using high density markers such as SNPs produced by RESTseq.

By identifying a shared locus between datasets, we were able to use this as an anchor point, allowing multiple datasets to be phased and then combined. Starting at the anchor point, each marker matching the reference genome was assigned the phase “A” and each marker different to it was assigned phase “B”. To overcome the problem that the expression (matching or different to the reference genome) could switch without the phase switching as well, we used precedents for each of the phases. A precedent for phase “A” would be a match in the expression of the current marker and the previous marker in the same individual. In case of a mismatch this would be a precedent for phase “B”. This was done for each linkage group. After working through all the linkage groups, multiple datasets could be merged by the position in the genome.

Map construction

Data were filtered using the phasing script, to remove any markers with a distribution greater than 70:30 and a density under 50%, then with r/qtl (R Core Team, 2017; Broman *et al.*, 2003), to remove duplicate markers and individuals with under 1000 markers. Using the ASMap package in R (R Core Team, 2017; Taylor and Butler, 2017), recombination events and genetic distances were calculated for marker orders based on the reference genome assemblies before *de novo* map constructions were performed. Optimal marker order within a linkage group was determined by minimising the number of crossovers and genetic distances were calculated using the Kosambi map function. Due to changes in the marker order, phasing was checked and manually adjusted before re-running the *de novo* map construction. We then compared the physical location of incorrectly-placed regions to the reference genome structure.

Results

Genotyping results

Sequencing generated 16,838,258 reads across our 80 samples. Many more reads mapped to the Amel_4.5 reference genome than the earlier versions while average read length and GC% remained constant across all three (Table 2.). After filtering in R (R Core Team, 2017), our datasets contained: 49 individuals with 1456 unique markers and 92% coverage for the 3.0 genome assembly; 48 individuals with 1556 unique markers and 92% coverage for the 4.0 genome assembly and 78 individuals with 2879 unique markers and 77% coverage for the 4.5 genome assembly.

Map construction

Based on marker order in the reference genome, we calculated initial recombination rates of 42.2 cM/Mb, 38.5 cM/Mb and 98.1 cM/Mb for the 3.0, 4.0 and 4.5 refererence genome assemblies respectively. These are much higher than previous estimates (Table 1.). We identified high genetic linkage between physically distant markers with multiple recombination events between them, a sign of potential assembly errors, in all genome assemblies (Additional file 1), however, these are much more common in the 4.5 genome assembly (Additional file 1; Additional file 2); as evidenced by its recombination rate being more than double that of 3.0 or 4.0 (Table 1).

Having performed *de novo* genetic map constructions, the recombination rate and linkage between physically distant markers for each assembly decreases greatly (Table 1.; Additional file 1; Additional file 2). By comparing marker position in the genome assembly to the *de novo* assembly, we can see that much of the decrease has come from re-orientation of large regions within linkage groups rather than a total rearrangement of markers (Figure 1 A, B, C). The estimated recombination rates for the new assemblies are 22.8cM/Mb and 22.6cM/Mb for the 3.0 and 4.0 genome assemblies respectively and 44cM/Mb for the 4.5 genome assembly. The estimated recombination rate for the *de novo* assemblies using markers from the 4.0 and 3.0 assemblies are very close to earlier estimates using *de novo* constructions (Table 1.). While the estimate for markers from the 4.5 genome assembly is higher than has previously been estimated (Table 1.), this may be a result of the high number of mis-placed and mis-orientated scaffolds affecting the phasing method used (additional files 4 and 5).

Relation to scaffold order

We identified 72 large and clearly misplaced or inverted regions within the current *Apis mellifera* 4.5 reference genome assembly (Figure 1 C; Additional file 3). Of these, 71 contained all the markers on a scaffold suggesting an error during genome construction. For the one region, which did not contain all the markers on a scaffold, there was still a large amount of variation in placement between the scaffold's remaining markers suggesting a degree of error in the construction and that the entire scaffold is also incorrectly placed.

Discussion

We compared genome construction quality among the three most recent versions of the *Apis mellifera* genome assembly using high-density SNP-based linkage maps. After performing *de novo* map constructions, we found the linkage between adjacent markers increased greatly while the number of recombination events decreased. As linkage is expected to decrease with increasing physical distance (Utsunomiya *et al*., 2016), this supports our re-ordering of markers in the *de novo* map construction. For *Apis mellifera* 3.0 and 4.0, the two most well-constructed genome assemblies in our analysis, this results in recombination rates of 22.8cM/Mb and 22.6cM/Mb, highly similar to the 22.0cM/Mb reported in a previous *de novo* assembly (Solignac *et al*., 2007). While the estimated recombination rate for the *Apis mellifera* 4.5 genome assembly is higher (44cM/Mb), much of this difference could be explained by the misplacement or misorientation of at least 20% of scaffolds in the Amel_4.5 reference genome and the effect this would have on the phasing method used.

Despite being the current representative genome for *A. mellifera*, the Amel_4.5 (Elsik *et al*., 2014) genome assembly contained the largest regions of misplaced scaffolds in our analysis; a result, which could cause problems for the 120 studies citing it in Web of Science (Clarivate Analytics). Although the earlier genome assemblies may also contain misplaced or aligned scaffolds, the short length of these means the physical distance may not be enough for linkage to decay between the two adjacent scaffolds. The merging of shorter scaffolds and filling of gaps from previous assemblies, without the subsequent generation of a new genetic map, likely introduced errors in the assembly and contributes to our identification of such a high number of mis-placed and mis-aligned scaffolds.

While *A. mellifera* does appear to have one of the highest recombination rates of any eukaryote (Solignac *et al*., 2007; Beye *et al*., 2006; Weinstock *et al*., 2006; Liu *et al*., 2015), the very high estimates associated with the Amel_4.5 genome assembly (Liu *et al*., 2015) are likely biased by poor scaffold placement rather than representing genuine recombination events. Indeed, when studies do report recombination frequencies under 30 cM/Mb for the Amel_4.5 genome assembly (Table 1.), they have either ignored multiple recombination events between markers (Liu *et al*., 2015) or did not estimate recombination events between scaffolds (Wallberg *et al*., 2015): supporting our conclusion that misplaced scaffolds are artificially increasing the estimated recombination rate. Further evidence can be found from the production of the *Apis cerana* reference genome assembly (Park *et al.*, 2015), where the Amel_4.5 reference genome was used to test the accuracy of the construction. While synteny data were only published for Chromosome 3, comparison of a region with seeming genomic rearrangement between the two species (Park *et al*, 2015) to our own data (Figure 1C, Additional file 3) reveals that we identify a scaffold inversion in that region on Chromosome 3.

Although there appears to be a larger degree of error in the Amel_4.5 genome assembly than earlier genome assemblies, when this is known and corrected for, it should still be considered the best option for read-mapping. This is evidenced by the increased sequence lengths and number of anchored scaffolds in Amel_4.5 allowing us to map more sequences to the genome and generate an increased total number of markers as well as an increased number of markers per individual (Table 2.). In future, it is possible that new sequencing technologies, such as the long reads generated by PacBio (Rhoads and Au, 2015), or advances in assembly methods such as Hi-C (Belton *et al.*, 2012), could help to bridge the gaps between scaffolds and further improve the genome assembly.

Conclusions

The generation of reference genomes is accelerating rapidly. On the 31st of December 2017, 5983 (for 5075 unique species) eukaryotic genome assemblies were stored in the NCBI genome assembly database (Kitts *et al.*, 2016). 1345 (22.5%) of these assemblies were added in 2017, at an average rate of over 112 per month; compared to a total of 50 (<1%) genome assembiles added in the first five years and 403 (7%) in the first decade of this millennium (Kitts *et al.*, 2016). While the rapid rise and proliferation of genomic data will benefit research, our results show that even well-resolved assemblies cannot be relied on fully and that high-density linkage maps generated using NGS can be invaluable in testing assembly quality. The variation we find between reference genome assemblies also highlights the difficulties that come with comparing results to those generated using different genome assemblies.

While the availabilty of reference data does make genome-level studies cheaper and more feasible for a wide range of studies (Collins *et al.*, 2003; Nygaard and Wurm, 2015; McCormack *et al.*, 2013), caution should be excercised when using them and the placement of scaffolds should, if possible, be confirmed experimentally (Solignac *et al.*, 2007; Artemov *et al.*, 2017; Beye *et al.*, 2006; Weinstock *et al.*, 2006; Roesti *et al.*, 2013; Utsunomiya *et al.*, 2016; Zeng *et al.*, 2017). This would not only provide more reliable results for a single study but, by maximising the accuracy of each iteration of a reference genome assembly, should make comparisons between iterations more feasible than currently appears possible for *A. mellifera*.

Data avaliability

Sequence data has been deposited in the Sequence Read Archive (SRA) of the National Centre for Biotechnology Information (NCBI) under the project number: PRJNA388299

Funding

Funding was provided by Deutsche Forschungsgemeinschaft grant RO 5121/1-1 to JR

Acknowledgements

We would like to thank Barbara Locke, Eva Frey and Peter Rosenkranz for providing samples, Petra Leibe and Denise Kleber for their help with the lab work.

References

Andrews S. (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Artemov GN, Peery AN, Jiang X, Tu Z, Stegniy VN, Sharakhova MV, *et al*. (2017) The physical genome mapping of *Anopheles albimanus* corrected scaffold misassemblies and identified interarm rearrangements in genus *Anopheles*. *G3*. 7:155-164.

Beaurepaire AL, Krieger KJ, Moritz RFA. (2017) Seasonal cycle of inbreeding and recombination of the parasitic mite *Varroa destructor* in honeybee colonies and its implications for the selection of acaricide resistance. *Infect. Genet. Evol.* 50:49-54.

Behrens D, Huang Q, Geßner C, Rosenkranz P, Frey E, Locke B, *et al*. (2011) Three qtl in the honey bee *Apis mellifera* L. Suppress reproduction of the parasitic mite *Varroa destructor*. *Ecol. Evol*. 1:451-458.

Belton JM, McCord RP, Gibcus J, Naumova N, Zhan Y, Dekker J. (2012) Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods*. 58:268-276.

Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, *et al*. (2006) Exceptionally high levels of recombination across the honey bee genome. *Genome Res.* 16:1339-1344.

Broman KW, Wu H, Sen S, Churchill GA. (2003) R/qtl: QTL mapping in experimental crosses. *Bioinformatics*. 19:889-890.

Collins FS, Morgan M, Patrinos A. (2003) The human genome project: lessons from large-scale biology. *Science*. 300:286-290.

Conlon BH, Mitchell J, de Beer ZW, Carøe C, Gilbert MT, Eilenberg J, *et al*. (2017) Draft genome of the fungus-growing termite pathogenic fungus *Ophiocordyceps bispora* (Ophiocordycipitaceae, Hypocreales, Ascomycota). *Data in Brief*. 11:537-542.

Conlon BH, de Beer ZW, De Fine Licht HH, Aanen DK, Poulsen M. (2016) Phylogenetic analyses of diverse *Podaxis* specimens from Southern Africa reveal hidden diversity and new insights into associations with termites. *Fungal Biol.* 120:1065-1076.

Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, *et al*. (2014) Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics.* 15:86.

Elsik CG, Tayal A, Diesh CM, Unni DR, Emery ML, Nguyen HN *et al*. (2015) Hymenoptera Genome Database: intergrating genome annotations in HymenopteraMine. *Nucleic Acids Res.* 44:D793-D800.

Garnery L, Vautrin D, Cornuet JM, Solignac M. (1991) Phylogenetic relationships in the genus *Apis* inferred from mitochondrial DNA sequence data. *Apidologie*. 22:87-92.

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, *et al.* (2013) Mycocosm portal: gearing up for 1000 genomes. *Nucleic Acids Res.* 42:D699-D704.

Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, *et al.* (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44:D73-D80.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*. 25:2078-2079.

Li H, Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754-1760.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 27:2987-2993.

Liu H, Zhang X, Huang J, Chen JQ, Tian D, Hurst LD, *et al.* (2015) Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. *Genome Biol.* 16:15.

Martin M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 17:10.

Matasci N, Hung LH, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, *et al.* (2014) Data access for the 1,000 plants (1KP) project. *Gigascience*. 3:17.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. (2013) Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* 66:526-538.

Nygaard S and Wurm Y. (2015) Ant genomics (Hymenoptera: Formicidae): challenges to overcome and opportunities to seize. *Myrmecol. News.* 21:59-72.

Park D, Jung JW, Choi BS, Jayakodi M, Lee J, Lim J, *et al*. (2015) Uncovering the novel charachteristics of Asian honey bee, *Apis cerana*, by whole genome sequencing. *BMC Genomics*. 16:1

R Core Team. (2017) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Rhoads A, Au KF. (2015) PacBio sequencing and it's applications. *Genomics Proteomics Bioinformatics*. 13:278-289.

Roesti M, Moser D, Berner D. (2013) Recombination in the threespine stickleback genome – patterns and consequences. *Mol. Ecol.* 22:3014-3027.

Solignac M, Mougel F, Vautrin D, Monnerot M, Cornuet J-M. (2007) A third-generation microsatellite-based linkage map of the honey bee, *Apis mellifera*, and its comparison with the sequence-based physical map. *Genome Biol.* 8:R66.

Stolle E, Moritz RFA. (2013) RESTseq – efficient benchtop population genomis with RESTriction fragment SEQuencing. *PLOS one.* 8:e63960.

Taylor J, Butler D. (2017) ASMap: linkage map construction using the MSTmap Algorithm. R package version 0.4-7. Accessed Jan, 2017.

Utsunomiya AT, Santos DJ, Boison SA, Utsunomiya YT, Milanesi M, Bickhart DM, *et al.* (2016) Revealing misassembled segments in the bovine reference genome by high resolution linkage disequilibrium scan. *BMC* Gen*omics*. 17:705.

Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, *et al*. (2014) A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature Genet*. 41:1081-1088.

Wallberg A, Glémin S, Webster MT. (2015) Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genet.* 11:e1005189.

Web of Science. Clarivate Analytics, USA. (2018) https://www.webofknowledge.com/. Accessed 12th March 2018.

Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, Maleszka R, *et al*. (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*., *Nature*. 443:931-949.

Zeng Q, Fu Q, Li Y, Waldbieser G, Bosworth B, Liu S, *et al.* (2017) Development of a 690K SNP array in catfish and its application for genetic mapping and validation of the reference genome sequence. *Sci. Rep.* 7:40347.

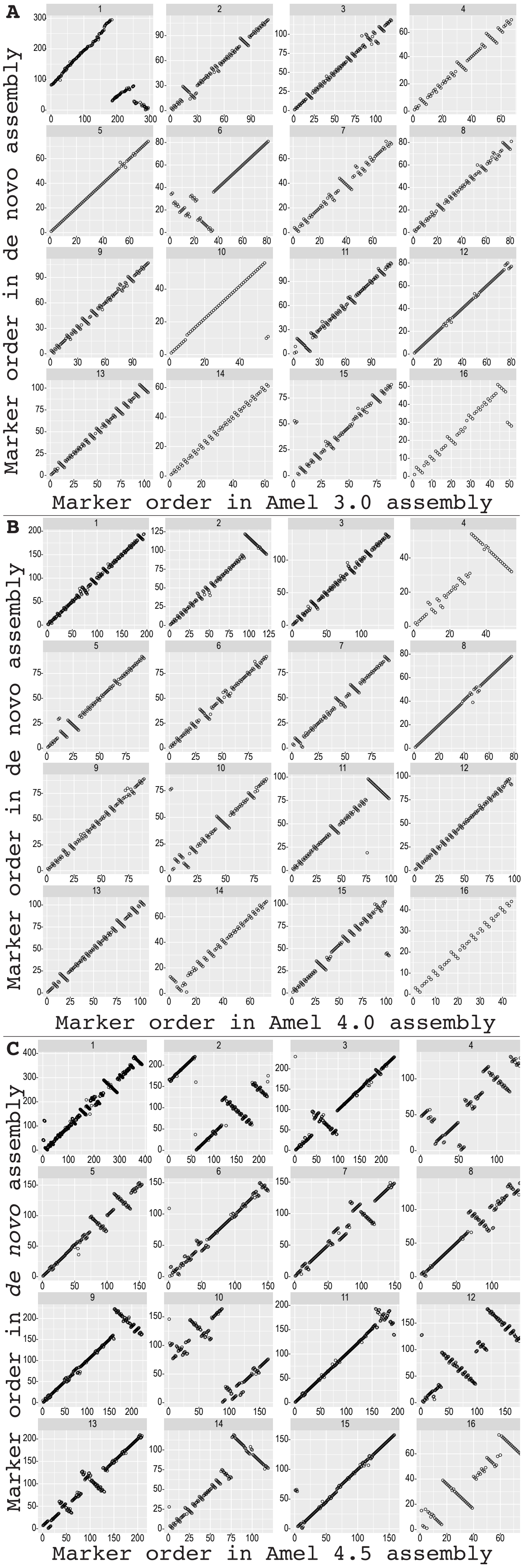
Figure legends

Figure 1.

Comparison of marker order in the genome assembly vs the *de novo* map construction for (A)

Amel_3.0 (n = 1456), (B) Amel_4.0 (n = 1556) and (C) Amel_4.5 (n = 2879).

Figure 1


A
Marker order in de novo assembly
1
0 100 200 300
2
0 30 60 90
3
0 25 50 75 100
4
0 20 40 60
5
0 20 40 60
6
0 20 40 60 80
7
0 20 40 60
8
0 20 40 60 80
9
0 30 60 90
10
0 20 40
11
0 30 60 90
12
0 20 40 60 80
13
0 25 50 75 100
14
0 20 40 60
15
0 25 50 75
16
0 10 20 30 40 50
Marker order in Amel 3.0 assembly
B
Marker order in de novo assembly
1
0 50 100 150 200
2
0 25 50 75 100 125
3
0 50 100
4
0 20 40
5
0 25 50 75
6
0 25 50 75
7
0 25 50 75
8
0 20 40 60 80
9
0 25 50 75
10
0 25 50 75
11
0 25 50 75 100
12
0 25 50 75 100
13
0 25 50 75 100
14
0 20 40 60
15
0 25 50 75 100
16
0 10 20 30 40
Marker order in Amel 4.0 assembly
C
Marker order in de novo assembly
1
0 100 200 300 400
2
0 50 100 150 200
3
0 50 100 150 200
4
0 50 100
5
0 50 100 150
6
0 50 100 150
7
0 50 100 150
8
0 50 100
9
0 50 100 150 200
10
0 50 100 150
11
0 50 100 150 200
12
0 50 100 150
13
0 50 100 150 200
14
0 25 50 75 100
15
0 50 100 150
16
0 20 40 60
Marker order in Amel 4.5 assembly

Tables with legends

Table 1.

Paper	Data-type	Genome assembly version	Recombination rate (cM/Mb)
Beye *et al.*, 2006	Microsatellite	2.0	19.0
Solignac *et al.*, 2007	Microsatellite	*de novo*	22.0
Wahlberg *et al.*, 2015	SNP	4.5	26.0
Liu *et al.*, 2015	SNP	4.5	37.0
Liu et al., 2015 – ignoring multiple recombination events	SNP	4.5	24.5
Conlon et al. – 3.0	SNP	3.0	42.2
Conlon et al. – *de novo*	SNP	3.0	22.8
Conlon et al. – 4.0	SNP	4.0	38.5
Conlon et al. – *de novo*	SNP	4.0	22.6
Conlon et al. – 4.5	SNP	4.5	98.1
Conlon et al. – *de novo*	SNP	4.5	44.0

Comparison of recombination rate estimates, for previous *de novo* and reference-genome-based genetic maps, to our estimates.

Table 2.

Assembly	Reads mapped	Average read length	GC%	Number of SNPs
3.0	11,876,551	154	44	465,991
4.0	11,976,915	154	44	471,754
4.5	15,049,559	155	44	533,863

Mapping statistics for the three reference genome assemblies.

Additional files

Additional file 1.

Conlon_*et_al*_SupMat1_Recombination_fractions_for_all_assemblies.pdf

Additional file 2.

Conlon_*et_al*_SupMat2_Log10_Recombination_Events.pdf

Additional file 3.

Conlon_*et_al*_SupMat3_Location_of_misaligned_scaffolds.pdf

Additional file 4.

Conlon_*et_al*_SupMat4_PhaseHaploid_Instructions_and_pseudocode.pdf

Additional file 5.

Conlon_*et_al*_SupMat5_PhaseHaploid_script.pdf

The role of epistatic interactions underpinning resistance to parasitic Varroa mites in haploid honeybee drones

Journal: *Journal of Evolutionary Biology*

DOI: 10.1111/jeb.13271

Benjamin H. Conlon[a], Eva Frey[b], Peter Rosenkranz[b], Barbara Locke[c], Robin F. A. Moritz[a], Jarkko Routtu[a]

[a]Molecular Ecology, Institute of Biology/Zoology, Martin-Luther-University Halle-Wittenberg, Hoher Weg 4, 06099 Halle an der Saale, Germany. Email: benjamin.conlon@zoologie.uni-halle.de; robin.moritz@zoologie.uni-halle.de; jarkko.routtu@zoologie.uni-halle.de

[b]University of Hohenheim, Apicultural State Institute, 70599, Stuttgart, Germany. Email: eva.frey@uni-hohenheim.de; peter.rosenkranz@uni-hohenheim.de

[c]Department of Ecology, Swedish University of Agricultural Sciences, PO Box 7044, 750 07, Uppsala, Sweden. Email: barbara.locke@slu.se

Corresponding author:

Benjamin H. Conlon, Current address: Molecular Ecology, Institute of Biology/Zoology, Martin-Luther-University Halle-Wittenberg, Hoher Weg 4, 06099 Halle an der Saale, Germany. Email: benjamin.conlon@zoologie.uni-halle.de, Phone: +49 345 55 26235.

Running title: Epistasis in honeybee resistance to *Varroa*

A modified honey bee Ecdysone pathway inhibits reproduction in *Varroa*

Benjamin H. Conlon[a], John Kefuss[b], Adriana Aurori[c], Daniel S. Dezmirean[c], Robin F.A. Moritz[a,c,d], Jarkko Routtu[a]

[a] Molecular Ecology, Institute of Biology/Zoology, Martin-Luther-University Halle-Wittenberg, Hoher Weg 4, 06099 Halle an der Saale, Germany

[b] 49 Rue Jonas, Le Rucher D'Oc, Toulouse, France

[c] University of Agriculture Sciences and Veterinary Medicine, Cluj-Napoca, Romania

[d] Dept of Zoology and Entomology, University of Pretoria, South Africa

Corresponding author:

Benjamin H. Conlon, Current address: Molecular Ecology, Institute of Biology/Zoology, Martin-Luther-University Halle-Wittenberg, Hoher Weg 4, 06099 Halle an der Saale, Germany. Email: benjamin.conlon@zoologie.uni-halle.de, Phone: +49 345 55 26235.

Summary

The brood-parasitic mite *Varroa destructor* devastates colonies of the honey bee (*Apis mellifera*). *Varroa* population size is a significant predictor of colony death. However, resistance can evolve rapidly. A common trait among resistant colonies is the inhibition of *Varroa* reproduction in pupal cells but the mechanism has not been identified. Using resistant and susceptible haploid drone offspring of a single queen in a high-density genome wide association analysis, we show that an *ecdysone*-induced gene is significantly linked to resistance in our mapping population. Different *ecdysone*-related genes are present at resistance loci, from a lower-resolution study, in a different population. *Ecdysone* both triggers pupation in the bee and initiates reproduction in *Varroa*. The *Varroa* genome lacks a complete pathway for *ecdysone* biosynthesis but active *ecdysone* analogues initiate ovary action. If *Varroa* co-opts *ecdysone* ingested from the pupae to initiate its own reproduction, modifications to this pathway in resistant pupae could physiologically inhibit the parasites reproduction.

Article

The transition to a parasitic lifestyle is often accompanied by a reduction in overall or functional genome size [1, 2]. While Acari can exhibit a loss of function in some metabolic pathways [3, 4], this is not always accompanied by a reduction in genome size [5]. The brood parasitic mite *Varroa destructor* is an excellent example of this. *Varroa* possesses a much larger genome than many insects, including its host: *Apis mellifera* [6, 5]. However, *Varroa* also exhibits reduced metabolic pathways when compared to other Acari and Arthropods [3, 4]. One of these functionally reduced pathways, the *ecdysone* biosynthesis pathway, is important for the initiation of a female mite's reproductive cycle [7, 4]

Varroa completes its entire reproductive cycle within the pupal cells of *A. mellifera* [8]. A *Varroa* mother lays one male and up to four female eggs, which develop sequentially at 30-hour intervals, inside the sealed pupal cell[8]. The *Varroa* mother, and her developing offspring, will feed on the haemolymph of the developing pupae; reducing the bee's adult lifespan and making it less able to forage and support the colony [9]. The reduction in lifespan of *Varroa*-parasitised pupae translates into negative fitness consequences for the colony as a whole with *Varroa* population size in the autumn acting as a significant predictor of colony overwinter mortality [10]. Although the widespread use of acaricides can reduce colony mortality, it rapidly selects for the evolution of acaricide resistant *Varroa* while removing the selective pressure for the evolution of *Varroa*-resistance in *A. mellifera* [11, 12, 13]. However, when populations of *A. mellifera* are left untreated with acaricides, *Varroa* resistance can evolve rapidly [14, 15, 16, 17].

Despite different geographic and genetic origins, host-induced inhibition of *Varroa* reproduction is a shared trait of many *Varroa*-resistant *A. mellifera* populations across the globe as well as the original host *A. cerana* [18, 8, 19, 20, 17, 21, 22]. The reduced reproduction of *Varroa* appears to make an important contribution to the survival of untreated colonies and contributes to colony overwinter survival by reducing *Varroa* population size in the autumn [10, 17]. In some resistant populations, it has been shown that the reduced reproduction of *Varroa* is a genetic, heritable, trait of the host pupae [20, 22]. While it is yet to be shown exactly how a honey bee pupa is capable of inducing its parasite not to reproduce [23], experimental manipulations show *Varroa* will suspend reproduction when the conditions inside the cell are not optimal [24]. This suggests the induction of non-reproduction may not be a classical immune response but an act of physiological manipulation by the pupa.

We investigated the genomic basis for the host-induced non-reproduction of *Varroa*. We screened the haploid drone offspring of queens in a resistant population in search of a ~50% rate of non-reproduction of *Varroa*. This is indicative of single-gene control inherited from a heterozygous mother queen. The unusually high recombination rate in *A. mellifera* (19 cM/MB) [6] means that, even with a relatively small sample size, we were able to identify genomic regions and physiological pathways are linked to the resistance trait.

Colony screening

We collected 69 infested pupae with 46% of mites in singly-infested cells not reproducing. This was not significantly different from a Medelian 1:1 segregation expected under single locus control ($X^2 = 0.010$, df = 1, p = 0.920).

Sequencing

After mapping and filtering to remove individuals with low coverage, markers with <90% genotype coverage, a sequencing depth outside of 15-50 reads per individual and an allelic distribution greater than 35-65%, gave us a dataset containing 45 individuals (19 Resistant, 26 Susceptible) and 112,976 SNPs for the Fst analysis. We identified 29022 unique SNPs to be used in the QTL analysis.

Candidate locus identification

The presence of a single locus for resistance was supported by our identification of one peak from 7.42-7.45 Mbp on Chromosome 15 in the Fixation Index (F_{ST}) analysis (Figure 1; SNPs = 20; Mean $F_{ST} = 0.338$). The peak is 2.27 Mbp from a QTL peak identified in the Gotland population of *Varroa*-resistant *A. mellifera* (Figure 2) [22]. QTL analysis identified one significant locus, and no significant interactions, (LOD = 4.21, df = 1, p = 0.029, phenotype explained = 40%) ranging from 7.42-7.53 Mbp on Chromosome 15 and overlapping with the F_{ST} peak (Extended data figure 1). In total, the overlapping F_{ST} and 1.5 LOD window contained 74 SNPs and 14 INDELs.

Candidate gene identification

The analytical power of using the haploid honey bee drones and very high recombination rate [6] allowed us to narrow our resistant locus down to 10 functional SNPs in 4 genes across a 43 Kb window which result in a non-synonymous change in the amino acid sequence (Table 1). With 10-20% of *Varroa* expected not to reproduce regardless of host genotype [25, 26], there is an unavoidable 10-20% error rate in the identification of resistant pupae. This is reflected in our data (Figure 3) and means that 40% is likely to be an underestimation of the percentage of the phenotype explained.

The presence of *Mblk-1* as the best-segregating gene at the significant locus suggests a potential pathway for resistance in the Toulouse population of *Varroa*-resistant *A. mellifera*. The regulation of metamorphosis by conserved regions of *Mblk-1* is conserved across both holo- and hemi-metabolous insects [27]. With *Varroa* being experimentally shown to suspend reproduction when pupal cues, related to the initiation of morphogenesis, are not optimal [24], a change in the action of *Mblk-1* could therefore induce *Varroa* to suspend reproduction. A similar, albeit less clear, mechanism has been suggested for reduced *Varroa* reproduction in a resistant honey bee population from the island of Gotland, Sweden [22].

In the early stages of pupation, the prepupae releases a pulse of ecdysteorid hormones, including *ecdysone*, which serve to initiate morphogenesis [28]. *Ecdysone* and its derivatives also act as a trigger for vitellogenesis in the Acari [7, 4, 29]. In the Gotland population of *Varroa*-resistant honey bees, *cytochrome P450 18a1* (*Cyp18a1*) and *cytochrome P450 306a1* (*phm*) are both linked to the host-induced inhibition of *Varroa* reproduction [22]. Both genes are involved in the *ecdysone* biosynthesis pathway and *Cyp18a1* is involved in lowering the titre of *ecdysone* during the transition from prepupa to pupa [30]. This suggests the regulation of *ecdysone*-linked genes could represent a common pathway for the inhibition of *Varroa* reproduction across independently-evolved populations. Fascinatingly, although *Varroa* shows increased expression of genes involved in the production of *ecdysone* when initiating reproduction [4, 29], the pathway is incomplete with only three of the seven genes from the *ecdysone* biosynthetic pathway present in the *V. destructor* genome [5, 4]. Functional forms of *ecdysone* are capable of ingestion by *Varroa* [31]; suggesting the reduced number of genes may be an adaptation of the mite to its parasitc lifestyle and missing compounds are acquired through its haemolymph diet [4]. This raises the possibility that the pulse of prepupal ecdysteroids is not a signal but a necessary physiological component for the successful initation of reproduction in *V. destructor*. A change in the regulation of genes involved in the production of or induced by *ecdysone* could reduce the amount available for ingestion by *Varroa* rendering it incapable of initiating oogenesis. In this sense, the host-induced inhibition of *Varroa* reproduction may represent a case of the host wresting back control of its extended phenotype; preventing its cooption by the parasite and increasing its own fitness.

The inhibition of *Varroa* reproduction appears to play an important role in colony survival for the Toulouse and other *Varroa*-resistant populations [19, 17, 22, 20]. While inhibition in the Toulouse and Gotland resistant populations may be linked to the manipulation of *Varroa* using the *ecdysone* signalling cascade, they achieve the same result using different methods [22]. This raises the possibility that the *ecdysone* pathway represents a common link for the inhibition of *Varroa* reproduction in independently-evolved resistant populations and the cooption of prepupal ecdysteroids may be an important physiological trigger for the initiation of *Varroa* reproduction.

References

1 Mounsey K.E. *et al.* Quantitative PCR-based genome size estimation of the astigmatid mites Sarcoptes scabiei, Psoroptes ovis and Dermatophagoides pteronyssinus. *Parasit. Vectors.* **5**, 3, (2012).

2 Poulin, R. & Randhawa, H. S. Evolution of parasitism along convergent lines: from ecology to genomics. *Parasitology.* **142**, S6-S15, (2015).

3 Grbić, M. *et al.* The genome of Tetranychus urticae reveals herbivorous pest adaptations, *Nature.* **479**, 487-492, (2011).

4 Cabrera, A. R. *et al*. Three Halloween genes from the Varroa mite, Varroa destructor (Anderson & Trueman) and their expression during reproduction. *Insect Mol. Biol.* **24**, 277-292 (2015).

5 Cornman, R. S. *et al.* Genomic survey of the ectoparasitic mite Varroa destructor, a major pest of the honey bee Apis mellifera. *BMC Genomics.* **11,** 602 (2010).

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

7 Roe, R. M. *et al.* Hormonal regulation of metamorphosis and reproduction in ticks. *Front. Biosci.* **13,** 7250-7268 (2008).

8 Rosenkranz, P., Aumeier, P. & Ziegelmann, B. Biology and control of Varroa destructor. *J. Invertebr. Pathol.* **103,** S96–S119 (2010).

9 Annoscia, D., Del Piccolo, F., Covre, F. & Nazzi, F. Mite infestation during development alters the in-hive behaviour of adult honeybees. *Apidologie.* **46**, 306-314 (2015).

10 van Dooremalen, C. *et al.* Winter Survival of Individual Honey Bees and Honey Bee Colonies Depends on Level of Varroa destructor Infestation. *PLoS One.* **7**, e36285 (2012).

11 Fries, I. & Bommarco, R. Possible host-parasite adaptations in honey bees infested by Varroa destructor mites. *Apidologie.* **38**, 525–533 (2007).

12 González-Cabrera, J. *et al.* Novel Mutations in the Voltage-Gated Sodium Channel of Pyrethroid-Resistant Varroa destructor Populations from the Southeastern USA. *PLoS One.* **11**, e0155332 (2016).

13 Beaurepaire, A. L., Krieger, K. J. & Moritz R. F. A. Seasonal cycle of inbreeding and recombination of the parasitic mite Varroa destructor in honeybee colonies and its implications for the selection of acaricide resistance. *Infect. Genet. Evol.* **50**, 49-54 (2017).

14 Kefuss, J., Vanpoucke, J., Bolt, M. & Kefuss, C. Selection for resistance to Varroa destructor under commercial beekeeping conditions. *J. Apic. Res.* **54**, 563-576 (2015).

15 Fries, I., Imdorf , A. & Rosenkranz, P. Survival of mite infested (Varroa destructor) honey bee (Apis mellifera) colonies in a Nordic climate. *Apidologie.* **37**, 564–570 (2006).

16 Le Conte, Y. *et al.* Honey bee colonies that have survived Varroa destructor. *Apidologie,* **38**, 566–572 (2007).

17 Oddie, M. A. Y., Dahle, B. & Neumann, P. Norwegian honey bees surviving Varroa destructor mite infestations by means of natural selection. *PeerJ.* **5**, e3956 (2017).

18 Oldroyd, B. P. Coevolution while you wait: Varroa jacobsoni, a new parasite of western honeybees. *Trends Ecol. Evol.* **14**, 312-315 (1999).

19 Locke, B. Natural Varroa mite-surviving Apis mellifera honeybee populations. *Apidologie.* **47**, 467-482 (2016).

20 Locke, B. Inheritance of reduced Varroa mite reproductive success in reciprocal crosses of mite-resistant and mite-susceptible honey bees (Apis mellifera). *Apidologie.* **47**, 583-588 (2016).

21 Nganso, B. T. *et al*. Low fertility, fecundity and numbers of mated female offspring explain the lower reproductive success of the parasitic mite Varroa destructor in African honeybees. *Parasitology.* 1-7 (2018=.

22 Conlon, B. H. *et al*. The role of epistatic interactions underpinning resistance to parasitic Varroa mites in haploid honey bee (Apis mellifera) drones. *J. Evol. Biol.* https://doi.org/10.1111/jeb.13271 (2018).

23 Nazzi, F. & Le Conte, Y. Ecology of Varroa destructor, the major ectoparasite of the western honey bee, Apis mellifera. *Annu. Rev. Entomol.* **61**, 417-432 (2016).

24 Frey, E., Odemer, R., Blum, T. & Rosenkranz, P. Activation and interruption of the reproduction of Varroa destructor is triggered by host signals (Apis mellifera). *J. Invertebr. Pathol.* **113**, 56-62 (2013).

25 Martin, S., Holland, K. & Murray, M. Non-reproduction in the honeybee mite a Varroa jacobsoni. *Exp. Appl. Acarol.* **21**, 539–549 (1997).

26 Bienefeld, K., Haberl, M. & Radtke, J. Does the Genotype of Honeybee Brood Influence the Attractiveness for Varroa Jacobsoni And/or the Reproduction of This Parasite?. *Hereditas.* **129**, 125-129 (1998).

27 Takayanagi-Kiya, S., Kiya, T., Kunieda, T. & Kubo, T. Mblk-1 Transcription Factor Family: Its Roles in Various Animals and Regulation by NOL4 Splice Variants in Mammals. *Int. J. Mol. Sci.* **18**, 246 (2017).

28 Lee, C-Y *et al*. E93 Directs Steroid-Triggered Programmed Cell Death in Drosophila. *Mol. Cell.* 6, 433-443 (2000).

29 Mondet, F. *et al*. Transcriptome profiling of the honeybee parasite Varroa destructor provides new biological insights into the mite adult life cycle. *BMC Genomics.* **19**, 328 (2018).

30 Rewitz, K. F., Yamanaka, N. & O'Connor, M. B. Steroid Hormone Inactivation Is Required during the Juvenile-Adult Transition in Drosophila. *Dev. Cell.* **19**, 895–902 (2010).

31 Cabrera, A. R., Shirk, P. D. & Teal, P. E. A. A feeding protocol for delivery of agents to assess development in Varroa mites. *PLoS One.* **12**, e0176097 (2017).

32 Garnery, L., Vautrin, D., Cornuet, J. M. & Solignac, M. Phylogenetic relationships in the genus Apis inferred from mitochondrial DNA sequence data. *Apidologie.* **22**, 87-92 (1991).

33 Elsik, C. G. *et al*. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics,* **15**, 86 (2014).

34 Li, H. & Durbin, R. Fast and accurate long-read align- ment with Burrows-Wheeler transform. *Bioinformatics.* **25**, 1754–1760 (2010).

35 Broad Institute, Picard Tools. GitHub Repository. (2015) Available: http://broadinstitute.github.io/picard/.

36 McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).

37 Poplin, R. *et al*. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv.* 201178 (2017).

38 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491-498 (2011).

39 van der Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics.* **11**, 11.10.1–11.10.33 (2013).

40 Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.,* **32**, 244–257 (2015).

41 R Core Team, R: A language and environment for statistical computing, (2017) Available: https://www.R-project.org/.

42 Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *biorXiv.* doi: 10.1101/005165 (2014).

43 Broman, K. W., Wu, H. Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics.* **19**, 889-890 (2003).

44 Taylor, J. & Butler, D. R Package ASMap: Efficient Genetic Linkage Map Construction and Diagnosis. *J. Stat. Softw.* **79**, 1-29 (2017).

45 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly,* **6**, 80-92 (2012).

46 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457-D462 (2015).

47 Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25-29 (2000).

48 The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331-D338 (2017).

49 The Uniprot Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158-D169 (2016).

Extended data figures

Extended data figures are linked to the online version of the paper at www.nature.com/nature.

Acknowledgements

We thank M. Kardell for her help with the initial colony screening, P. Leibe and D. Kleber for their help in the lab and E. Stolle for helpful discussions on the data. This research was funded by the

Deutsche Forschungsgemeinschaft RO 5121/1-1 (to J.R.) and the PCE project GRAL (to R.F.A.M.).

Author contributions

Fieldwork was performed by B.H.C., and J.K.; B.H.C. performed the labwork; A.A. and D.S.D contributed to the sequencing and data handling; B.H.C. analysed the data; R.F.A.M. and J.R. supervised the project. All authors wrote the manuscript.

Author information

Reprints and permissions information is available online at ww.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.H.C. (benjamin.conlon@zoologie.uni-halle.de) or J.R. (jarkko.routtu@zoologie.uni-halle.de).

Table legends

Table 1.

The ten SNPs causing in changes in the amino acid sequence with their genomic position, bases, gene name and amino acid change.

Figure legends

Figure 1.

Genome-wide comparision of SNP Fst between susceptible and resistant pupae. The suggestive line was calculated as the 99.99th percentile. The suggestive region, at the 99.99th percentile, has a maximum Fst of 0.391. SNP windows which overlap the QTL in the Toulouse population are highlighted in green.

Figure 2.

Fst, calculated using 20 SNP windows, on Chromosome 15. The maximum Fst for the Toulouse population was 0.391. SNP windows which overlap the QTL in the Toulouse population are highlighted in green, the QTL in the Gotland population is highlighted with a red box. The maximum Fst in the Gotland QTL region was 0.04.

Figure 3.

The proportion of reproducing and non-reproducing *Varroa* for each allele at the QTL peak.

Materials and Methods

Colony screening

Colony screening took place from May-June 2017 near the village of Le Born, Haute Garonne, France (43°54'N 1°32'E). Drone brood cells from the white-eyed pupal stage onwards were opened and phenotyped based on the number of *Varroa* offspring. Cells in which it was not possible to unambiguously phenotype *Varroa* reproduction were excluded from further analyses. *Varroa* was considered to have successfully reproduced if it produced at least one daughter and one son while reproduction was considered unsuccessful if the mite produced no offspring or only sons[20,25]. We identified one colony in which ~50% of *Varroa*-infested cells did not reproduce and performed a Chi-Squared goodness-of-fit test to identify whether the distribution of successful vs unsuccessfully reproducing mites was significantly different to 50:50. The pupae and mature mites were stored together in 96% ethanol at -80°C for genetic analysis.

DNA extraction, sequencing and genotyping

DNA was extracted from the thorax of *Varroa*-infested brother-drone pupae using a phenol/chloroform extraction [32]. The resulting extracts were assessed using a Nanodrop 1000 spectrophotometer (peqlab) and underwent 20X, 150bp, paired-end sequencing on Illumina HiSeq with Novogene (Hong Kong).

Variant loci calling and analysis

DNA sequences were mapped to the scaffolds of *Apis mellifera* 4.5 reference genome [33] using the BWA "MEM" algorithm [34]. Variant loci were identified using Picard [35]and GATK [36, 37]. Base quality score recalibration, indel realignment, duplicate removal and SNP and INDEL discovery and genotyping was performed for all samples simultaneously using standard hard filtering parameters following GATK best practices [38, 39]. SNPs and Indels were called into separate files for further analyses.

SNPs and INDELs were filtered to remove loci with fewer than 90% genotyped individuals and where the allelic distribution was greater than 35-65%. F_{ST} was then calculated using a window of 10-20 SNPs with a maximum window size of 50,000 bp using popgenwindows [40]. F_{ST} values were analysed in R [41] using the qqman package [42]. The threshold for suggestive loci was calculated as the 99.99th percentile of F_{ST}.

The filtered SNPs were phased using the R script *PhaseHaploid* (B.H.C., Oertelt, E., R.F.A.M., J.R. Increasing recombination rate estimates result from decreasing assembly accuracy in honey bee (Apis mellifera) reference genome updates, *J. Hered.*, manuscript in review) and a genetic map constructed using the Rqtl and ASMap packages [43, 44, 41]. Unique genotypes then underwent single- and multi-locus qtl analyses to identify loci and interactions linked to resistance. QTL size was then estimated using the 1.5 LOD score.

Candidate SNP and gene analysis

The peak regions from the F_{ST} and QTL analyses were used to create a list of candidate SNPs and genes from the *A. mellifera* Official Gene Set v3.2. SNPeff [45] was used to identify which SNPs created a change in the amino acid sequence. The functions of genes in which a SNP changed the amino acid sequence were analysed further using the KEGG [46], Gene Ontology [47, 48] and UniProt [49] databases.

Data availability

Sequence data has been deposited in the Sequence Read Archive (SRA) of the National Centre for Biotechnology Information (NCBI) under the BioProject accession number: PRJNA473430.

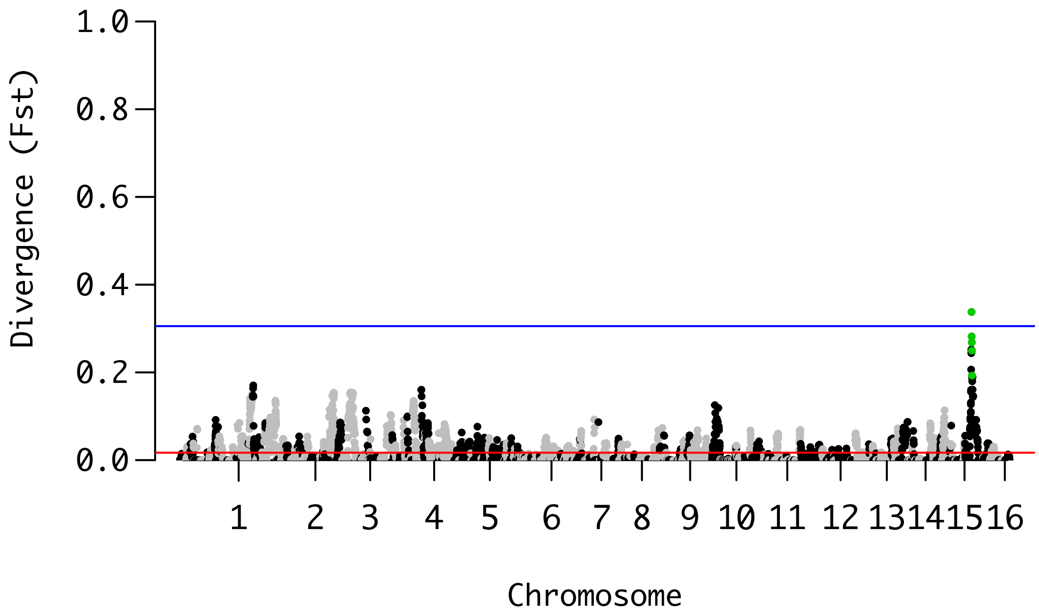


Figure 1.
Genome-wide comparision of SNP Fst between susceptible and resistant pupae. The suggestive line was calculated as the 99.99th percentile. The suggestive region, at the 99.99th percentile, has a maximum Fst of 0.391. SNP windows which overlap the QTL in the Toulouse population are highlighted in green.

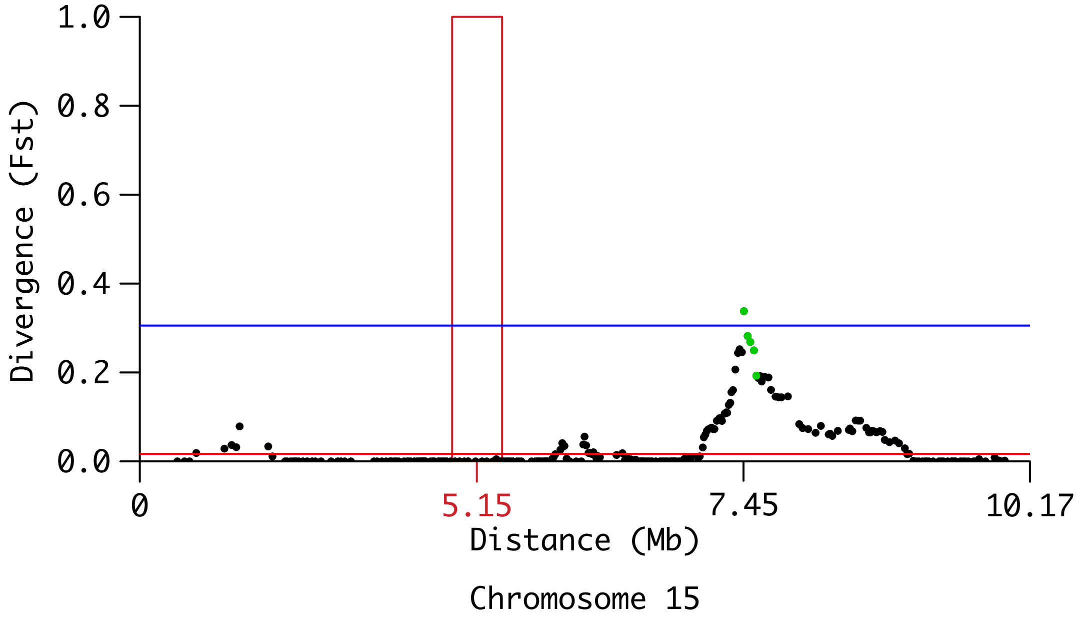


Figure 2.
Fst, calculated using 20 SNP windows, on Chromosome 15. The maximum Fst for the Toulouse population was 0.391. SNP windows which overlap the QTL in the Toulouse population are highlighted in green, the QTL in the Gotland population is highlighted with a red box. The maximum Fst in the Gotland QTL region was 0.04.

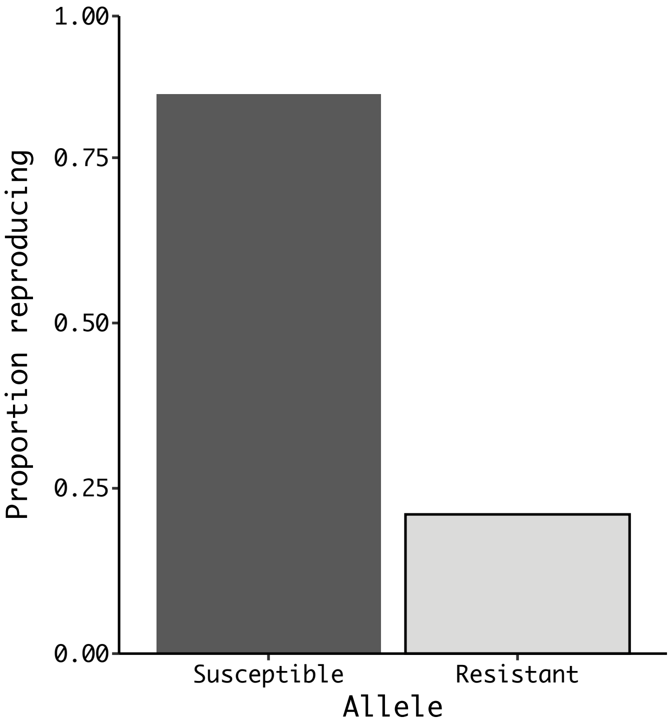
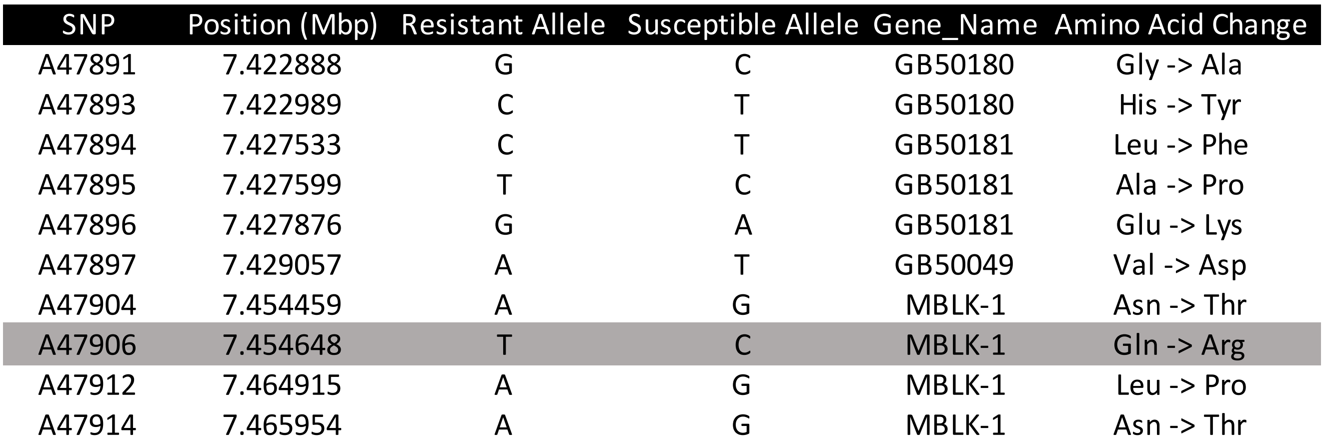


Figure 3.
The proportion of reproducing and non-reproducing Varroa for each allele at the QTL peak.



SNP	Position (Mbp)	Resistant Allele	Susceptible Allele	Gene_Name	Amino Acid Change
A47891	7.422888	G	C	GB50180	Gly -> Ala
A47893	7.422989	C	T	GB50180	His -> Tyr
A47894	7.427533	C	T	GB50181	Leu -> Phe
A47895	7.427599	T	C	GB50181	Ala -> Pro
A47896	7.427876	G	A	GB50181	Glu -> Lys
A47897	7.429057	A	T	GB50049	Val -> Asp
A47904	7.454459	A	G	MBLK-1	Asn -> Thr
A47906	7.454648	T	C	MBLK-1	Gln -> Arg
A47912	7.464915	A	G	MBLK-1	Leu -> Pro
A47914	7.465954	A	G	MBLK-1	Asn -> Thr

Table 1.

The ten SNPs causing in changes in the amino acid sequence with their genomic position, bases, gene name and amino acid change.

5. Conclusion

This thesis sought to explore the genomic basis for the host-inhibition of *Varroa* reproduction in independently-evolved resistant populations of *A. mellifera*. With *Varroa* population size in the autumn acting as a significant predictor of overwintering colony death (van Dooremalen, et al., 2012), a reduction in *Varroa* reproductive success is expected to reduce population sizes in the autumn and contribute to colony survival. Although both the original host (*A. cerana*) and many resistant populations of *A. mellifera* exhibit elevated rates of non-reproduction compared to susceptible colonies (Oldroyd, 1999; Kefuss, et al., 2004; Fries, et al., 2006; Le Conte, et al., 2007; Oddie, et al., 2017), the mechanism by which they achieve this appears to differ slightly between independently-evolved populations (Kurze, et al., 2016).

Aided by the identification, and correction for, multiple misplaced and mis-orientated scaffolds in the *A. mellifera* 4.5 reference genome assembly (Elsik, et al., 2014), the results of the studies on the Gotland and Toulouse populations support the suggestion that the mechanism of resistance differs between populations. Although a resistance locus on Chromosome 15 was identified in both populations, the distance between these two loci (2.27 Mbp) means linkage breaks down and they appear to represent two separate resistance loci. However, despite possessing different resistance loci, both populations may use the same physiological pathway to inhibit resistance.

5. 1 Varroa reproductive physiology

There is a body of experimental evidence to suggest that *Varroa* requires a kairomonal cue from the cuticle and the haemolymph of the pupa before it will initiate oogenesis (Garrido & Rosenkranz, 2004; Aumeier, et al., 2002; Rosenkranz, et al., 2010; Frey, et al., 2013) and that it will suspend oogenesis if the conditions are suboptimal (Frey, et al., 2013). The results from the studies on the Toulouse and Gotland populations provides strong evidence that this cue could be linked to the steroid hormone *ecdysone*. In both populations, significant loci contained genes linked to the *ecdysone* pathway (Figure 1).

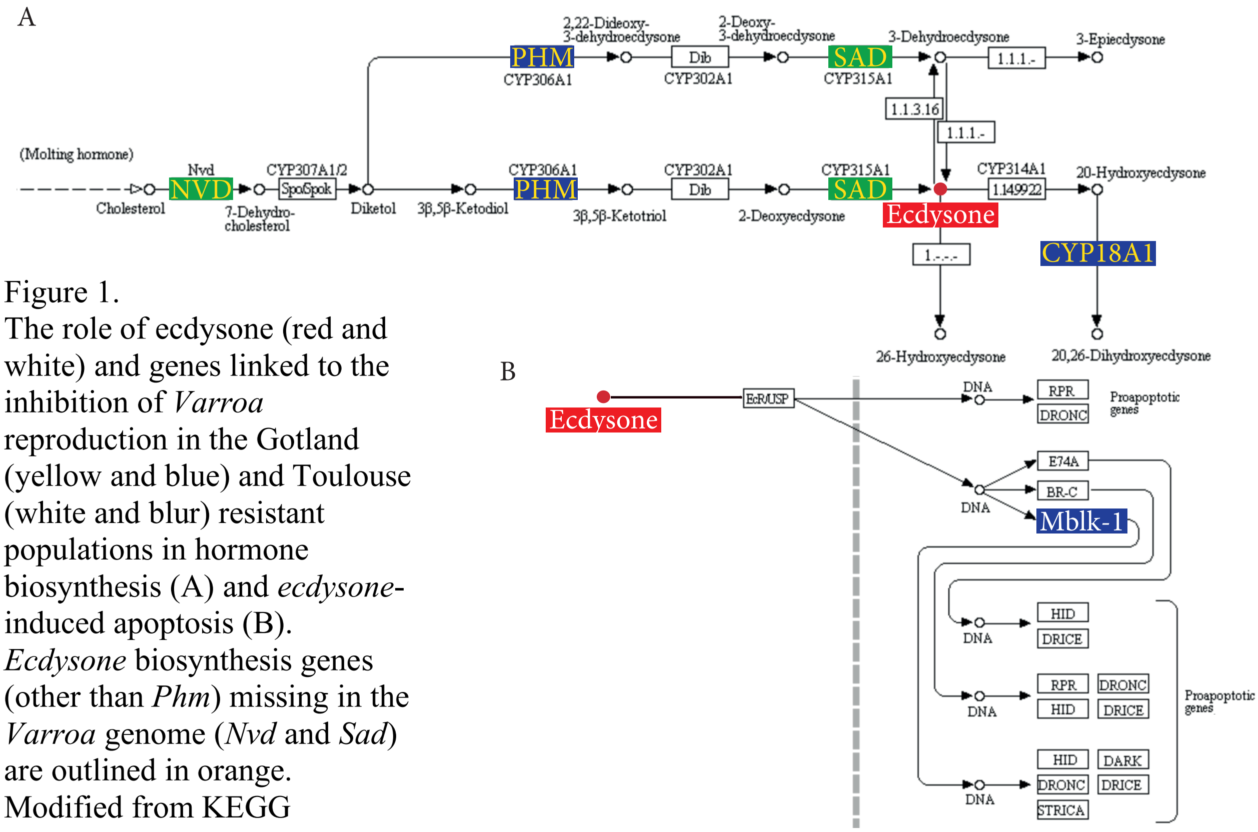


Figure 1.
The role of ecdysone (red and white) and genes linked to the inhibition of *Varroa* reproduction in the Gotland (yellow and blue) and Toulouse (white and blur) resistant populations in hormone biosynthesis (A) and *ecdysone*-induced apoptosis (B). *Ecdysone* biosynthesis genes (other than *Phm*) missing in the *Varroa* genome (*Nvd* and *Sad*) are outlined in orange. Modified from KEGG (Kanehisa, et al., 2015).

A pulse of ecdysteroids, including *ecdysone* and its derivatives, acts as a conserved trigger for metamorphosis in insect prepupae (Lee, et al., 2000; Takayanagi-Kiya, et al., 2017). With the *Varroa* mother feeding on the haemolymph of the prepupae before initiating reproduction (Rosenkranz, et al., 2010), the presence of *ecdysone*-linked genes, and the elevated titres in prepupae (Lee, et al., 2000) suggests this could act as a trigger for the initiation of *Varroa* reproduction. *Ecdysone* not only play an important role in the initiation of metamorphosis, they have also been shown to be important for the initiation of reproduction in *Varroa* and other Acari (Cabrera, et al., 2015; Roe, et al., 2008). With the *Varroa* genome possessing a reduced *ecdysone* biosynthesis pathway compared to other Acari (Grbić, et al., 2011; Cabrera, et al., 2015), and with *ecdysone* analogues capable of ingestion in a functional form by *Varroa* (Cabrera, et al., 2017), it is possible that *Varroa* co-opts the prepupal *ecdysone* pulse to initiate its own reproduction. This could explain the lack of a complete pathway for *ecdysone* biosynthesis in the *Varroa* genome (Cabrera, et al., 2015).

A *Varroa* mother's fitness is tightly linked to the pupation time of the bee (Rosenkranz, et al., 2010). The mother possesses a finite amount of sperms and eggs and her diploid daughter mites will only mate with haploid sons in the natal cell (Rosenkranz, et al., 2010). This provides a very strong selective pressure to lay eggs only when there is enough time for them to develop and mate before the bee ecloses. Should *Varroa* require *ecdysone* from the prepupal pulse to initiate oogenesis, this would biologically prevent the mother from wasting her finite sperms and eggs if the timing for reproduction is not optimal. However, this also suggests that minor alterations in the *ecdysone* pathway could prevent *Varroa* from successfully reproducing despite the timing being optimal. That the four loci identified across the genomes of the Gotland and Toulouse populations do not overlap, suggests that, although they may be linked to the same physiological pathway, there are multiple ways in which the inhibition of *Varroa* reproduction can evolve.

5.2 The evolution of resistance

That resistance loci appear to differ between the Gotland and Toulouse populations is perhaps unsurprising. The differences between the Nordic and Mediterranean climates means both populations of bees and mites will have been exposed to very different ecological conditions (Calis, et al., 1999; Büchler, et al., 2015). Historically, the previous use of acaricide treatments, as well as their still widespread use in the area surrounding the Toulouse population, means the pressure to adapt to the local conditions is likely to have outweighed that for *Varroa* resistance traits (Fries & Bommarco, 2007; Büchler, et al., 2015). With differences in the initial genetic variation for selection to act on, it is fascinating that the same method of resistance has evolved so many times in so many different locations (Fries, et al., 2006; Le Conte, et al., 2007; Wallberg, et al., 2014; Locke, 2016; Oddie, et al., 2017).

5.3 Consequences for apiculture

The capability of *A. mellifera* to evolve effective defences to *Varroa*, when the conditions are allowed, highlights the importance of evolutionary thinking in agriculture. Selection for local adaptation, and the presence of *Varroa* susceptible genetic material in surrounding treated populations, raises issues for attempts to transfer or breed from resistant populations in different environments. However, the repeated evolution of *Varroa*-resistance from differing genetic stocks suggests that, if selection is not inhibited with acaricide treatments, most European honey bee populations could develop resistance. The development of commercially viable breeding protocols (Kefuss, et al., 2015) means this could be possible without unsustainable levels of colony losses. An increase in the density of resistance populations could lead to the migration of resistance genes between populations. This increase in variance may contribute to the development of a balanced host-parasite relationship by reducing the likelihood of *Varroa* developing resistance to a single trait.

6. References for Introduction and Conclusion

Aumeier, P., Rosenkranz, P. & Francke, W., 2002. Cuticular volatiles, attractivity of worker larvae and invasion of brood cells by Varroa mites. A comparison of Africanized and European honey bees. *Chemoecology,* Volume 12, pp. 65-75.

Büchler, R. et al., 2015. The influence of genetic origin and its interaction with environmental effects on the survival of Apis mellifera L. colonies in Europe. *J. Apic. Res.,* 53(2), pp. 205-214.

Beaurepaire, A. L. et al., 2015. Host Specificity in the Honeybee Parasitic Mite, Varroa spp. in Apis mellifera and Apis cerana. *PLoS One,* 10(8), p. e0135103.

Cabrera, A. R. et al., 2015. Three Halloween genes from the Varroa mite, Varroa destructor (Anderson & Trueman) and their expression during reproduction. *Insect Mol. Biol.,* 24(3), pp. 277-292.

Cabrera, A. R., Shirk, P. D. & Teal, P. E. A., 2017. A feeding protocol for delivery of agents to assess development in Varroa mites. *PLoS One,* 12(4), p. e0176097.

Calis, J., Fries, I. & Ryrie, S., 1999. Population modelling of Varroa jacobsoni Oud.. *Apidologie,* Volume 30, pp. 111-124.

Donze, G. & Guerin, P. M., 1997. Time-Activity Budgets and Space Structuring by the Different Life Stages of Varroajacobsoni in Capped Brood of the Honey Bee, Apis mellifera. *J. Insect Behav.,* 10(3), pp. 371-393.

Elsik, C. G. et al., 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC genomics,* 15(1), p. 86.

Frey, E., Odemer, R., T, B. & P, R., 2013. Activation and interruption of the reproduction of Varroa destructor is triggered by host signals (Apis mellifera). *J. Invert. Pathol.,* Volume 113, pp. 56-62.

Fries, I. & Bommarco, R., 2007. Possible host-parasite adaptations in honey bees infested by Varroa destructor mites. *Apidologie,* Volume 38, pp. 525-533.

Fries, I., Imdorf, A. & Rosenkranz, P., 2006. Survival of mite infested (Varroa destructor) honey bee (Apis mellifera) colonies in a Nordic climate. *Apidologie,* Volume 37, pp. 564-570.

Garrido, C. & Rosenkranz, P., 2004. Volatiles of the honey bee larva initiate oogenesis in the parasitic mite Varroa destructor. *Chemoecology,* Volume 14, pp. 193-197.

Grbić, M. et al., 2011. The genome of Tetranychus urticae reveals herbivorous pest adaptations. *Nature,* Volume 479, pp. 487-492.

Hamilton, W. D., Axelrod, R. & Tanese, R., 1990. Sexual reproduction as an adaptation to resist parasites (A Review). *Proc. Nat. Acad. Sci. USA,* Volume 87, pp. 3566-3573.

Kanbar, G. & Engels, W., 2005. Communal use of integumental wounds in honey bee (Apis mellifera) pupae multiply infested by the ectoparasitic mite Varroa destructor. *Genet. Mol. Res.,* 4(3), pp. 465-472.

Kanehisa, M. et al., 2015. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.,* 44(D1), pp. D457-D462.

Kefuss, J., Vanpoucke, J., Bolt, M. & Kefuss, C., 2015. Selection for resistance to Varroa destructor under commercial beekeeping conditions. *J. Apic. Res.,* 54(5), pp. 563-576.

Kefuss, J., Vanpoucke, J., de Lahitte, J. & Ritter, W., 2004. Varroa Tolerance in France of Intermissa Bees From Tunisia And Their Naturally Mated Descendants: 1993-2004. *Am. Bee J.,* 144(7), pp. 563-568.

Kidner, J. & Moritz, R. F. A., 2013. The Red Queen Process does not Select for High Recombination Rates in Haplodiploid Hosts. *Evolutionary Biology,* 40(3), pp. 377-384.

Kohl, P. L. & Rutschmann, B., 2018. The neglected bee trees: European beech forests as a home for feral honey bee colonies. *PeerJ,* Volume 6, p. e4602.

Kurze, C., Routtu, J. & Moritz, R. F. A., 2016. Parasite resistance and tolerance in honeybees at the individual and social level. *Zoology,* Volume 119, pp. 290-297.

Le Conte, Y. et al., 2007. Honey bee colonies that have survived Varroa destructor. *Apidologie,* Volume 38, pp. 566-572.

Lee, C.-Y.et al., 2000. E93 Directs Steroid-Triggered Programmed Cell Death in Drosophila. *Mol. Cell,* Volume 6, pp. 433-443.

Locke, B., 2016. Natural Varroa mite-surviving Apis mellifera honeybee populations. *Apidologie,* 47(3), pp. 467-482.

Locke, B. & Fries, I., 2011. Characteristics of honey bee colonies (Apis mellifera) in Sweden surviving Varroa destructor infestation. *Apidologie,* Volume 42, pp. 533-542.

Maynard Smith, J., 1971. What use is sex?. *J. Theor. Biol.,* Volume 30, pp. 319-335.

Nazzi, F. & M, M., 1996. The presence of inhibitors of the reproduction of Varroa jacobsoni Oud. (Gamasida: Varroidae) in infested cells. *Exp. Appl. Acarol.,* Volume 20, pp. 617-623.

Oddie, M. A. Y., Dahle, B. & Neumann, P., 2017. Norwegian honey bees surviving Varroa destructor mite infestations by means of natural selection. *PeerJ,* Volume 5, p. e3956.

Oldroyd, B. P., 1999. Coevolution while you wait: Varroa jacobsoni, a new parasite of western honeybees. *Trends Ecol. Evol.,* 14(8), pp. 312-315.

Roe, R. M., Donohue, K. V., Khalil, S. M. S. & Sonenshine, D. E., 2008. Hormonal regulation of metamorphosis and reproduction in ticks. *Front. Biosci.,* Volume 13, pp. 7250-7268.

Rosenkranz, P., Aumeier, P. & Ziegelmann, B., 2010. Biology and control of Varroa destructor. *J. Invert. Pathol.,* Volume 103, p. S96–S119.

Takayanagi-Kiya, S., Kiya, T., Kunieda, T. & Kubo, T., 2017. Mblk-1 Transcription Factor Family: Its Roles in Various Animals and Regulation by NOL4 Splice Variants in Mammals. *Int. J. Mol. Sci.,* Volume 18, p. 246.

Thompson, J. N., 2005. Coevolution: The Geographic Mosaic of Coevolutionary Arms Races. *Curr. Biol.,* 15(24), pp. R992-R994.

van Dooremalen, C. et al., 2012. Winter Survival of Individual Honey Bees and Honey Bee Colonies Depends on Level of Varroa destructor Infestation. *PLoS One,* 7(4), p. e36285.

Wallberg, A. et al., 2014. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera. *Nature Genet.,* 46(10), pp. 1081-1087.

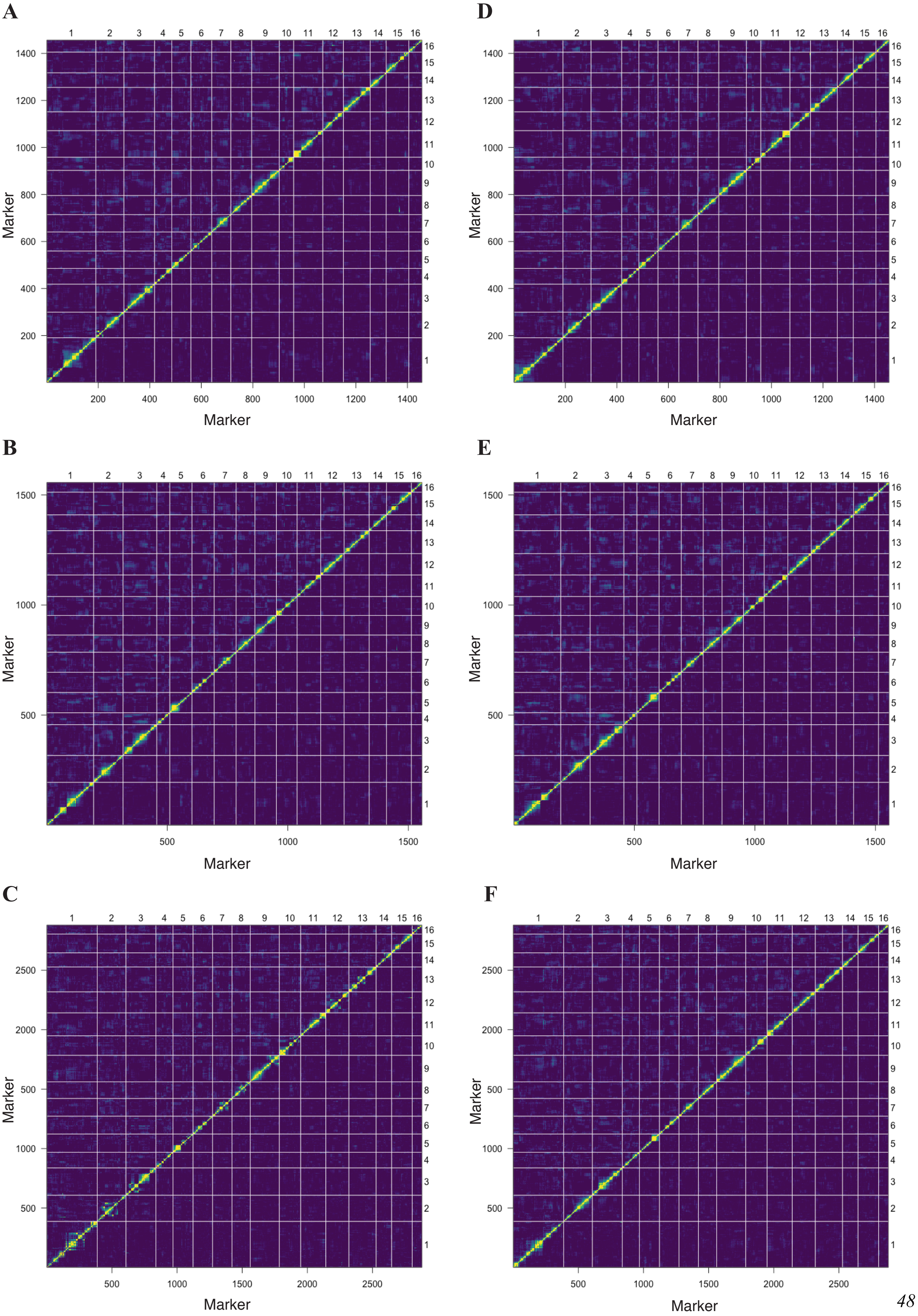
Wilfert, L., Gadau, J. & Schmid-Hempel, P., 2007. The genetic architecture of immune defense and reproduction in male Bombus terrestris bumblebees. *Evolution,* 61(4), pp. 804-815.

7. Acknowledgments

I'd like to thank: Jarkko and Robin for their supervision and guidance during my thesis; Peter Rosenkranz and Claudia Haußermann for introducing me to *Varroa* sampling; Axel Brockmann and the members of the Brockmann lab at the NCBS in Bengaluru for being so welcoming and helpful; Shishira Devraj and Marina Kardell for assisting in fieldwork; John Kefuss for allowing me to stay with him and sample his colonies; Christina Jenkins and Eckart Stolle for their advice on data handling and analysis; Silvio Erler and Rositta Mothes for their help with the University and local administrations; Petra Leibe and Denise Kleber for their help in the lab; everybody in the Moritz and Paxton labs for helpful discussions and a great workplace; finally, my family – and, in particular, Asta – who have supported me and put up with all the bee talk.

Additional File 1

Heatmap plotting the pair-wise recombination fractions against the LOD linkage score for the (A) Amel_3.0 (n = 1556), (B) Amel_4.0 (n = 1556) and (C) Amel_4.5 (n = 2879) genome assemblies and the (D) de novo 3.0 (n = 1456), (E) de novo 4.0 (n = 1556) and (F) de novo 4.5 (n = 2879) maps. Yellow represents high genetic linkage and a low recombination rate between two markers while blue represents low genetic linkage and a high recombination rate between two markers.



Additional File 3.

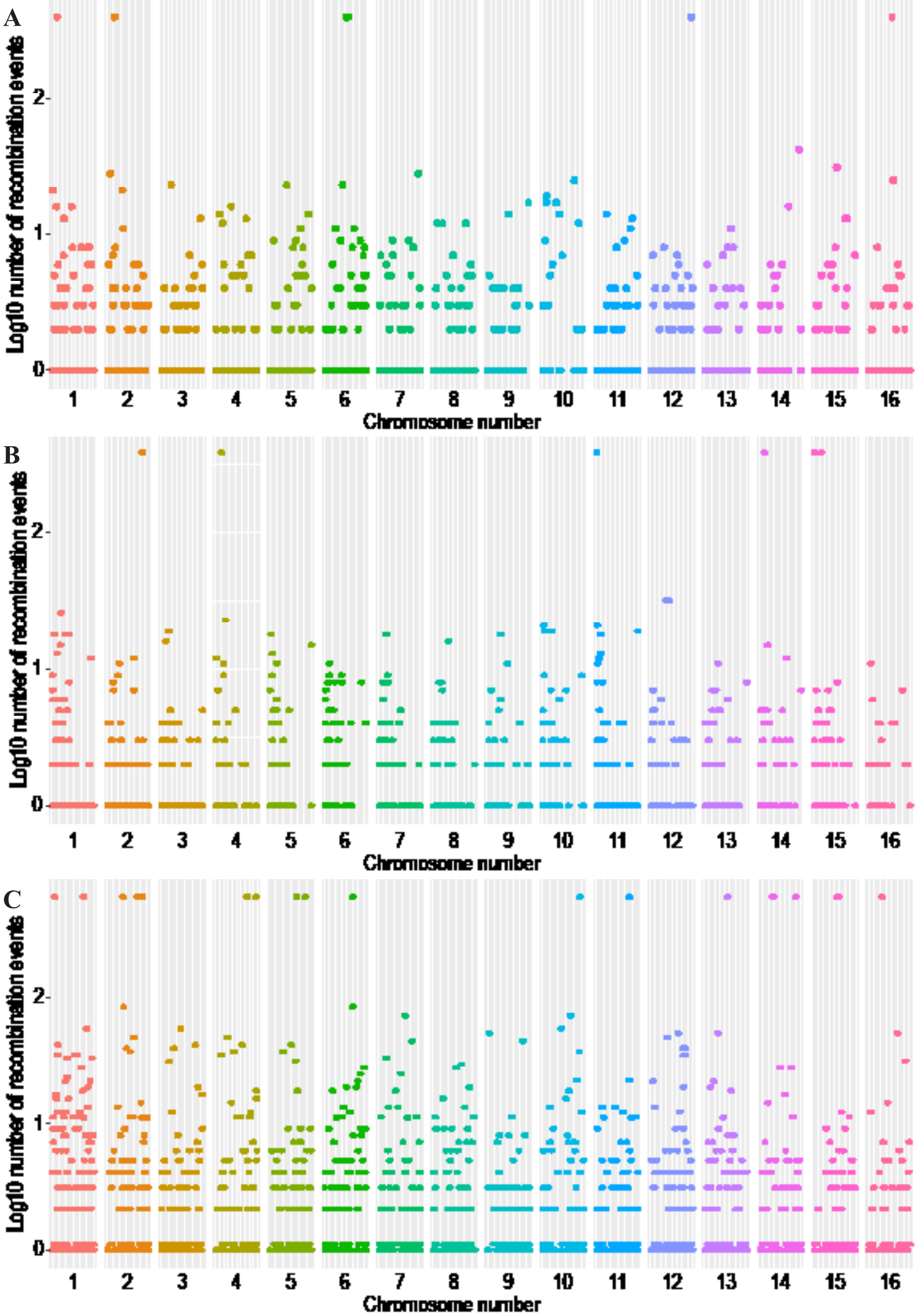
Location of identified mis-aligned scaffolds in the Apis mellifera 4.5 reference genome assembly.

Assembly	Chromosome	Upper region (Mbp)	Lower region (Mbp)	Upper Marker	Lower Marker	Scaffold	Entire Scaffold (Y/N)	Inverted or misplaced
4.5	1	17.79	16.12	A212	A184	29	Y	Inverted
4.5	1	18.14	18.12	A216	A213	30	Y	Inverted
4.5	1	18.73	18.37	A224	A217	31	Y	Inverted
4.5	1	24.67	22.62	A362	A293	37	Y	Inverted
4.5	1	29.79	28.56	A482	A444	43	Y	Inverted
4.5	2	15.49	12.54	A783	A711	20	Y	Misplaced
4.5	2	8.42	7.47	A647	A574	15	Y	Inverted
4.5	2	10.21	8.72	A702	A657	17	Y	Inverted
4.5	3	6.12	4.12	A910	A834	8	Y	Inverted
4.5	3	11.4	10.73	A1017	A1034	15	Y	Inverted
4.5	4	5.42	5.02	A1140	A1130	8	Y	Misplaced
4.5	4	4.96	4.76	A129	A1127	7	Y	Misplaced
4.5	4	4.19	3.79	A1126	A1123	6	Y	Misplaced
4.5	4	1.73	2.98	A1118	A1082	5	Y	Misplaced
4.5	4	0.78	0.78	A1079	A1079	3	Y	Misplaced
4.5	4	1.2	0.93	A1081	A1080	4	Y	Inverted and misplaced
4.5	4	0.47	0.13	A1078	A1075	1	Y	Misplaced
4.5	4	5.65	7.4	A1155	A1142	9	Y	Inverted
4.5	4	8.19	7.94	A1159	A1156	10	Y	Inverted
4.5	4	11.58	9.38	A1207	A1167	13	Y	Inverted
4.5	4	12.2	11.9	A2108	A1223	16	Y	Inverted
4.5	5	8.56	7.65	A1359	A1327	12	Y	Inverted
4.5	5	11.95	9.36	A1412	A1374	14	Y	Inverted
4.5	7	6.26	5.42	A1698	A1692	17	Y	Inverted
4.5	7	10.45	8.36	A1791	A1737	21	Y	Inverted
4.5	8	8.58	7.17	A1957	A1917	9	Y	Inverted
4.5	8	12.38	11.62	A2015	A1995	17	Y	Inverted
4.5	9	2.96	2.82	A2027	A2026	7	Y	Inverted and misplaced
4.5	9	1.73	1.73	A2023	A2024	5	Y	Inverted and misplaced
4.5	9	11.07	9.63	A2321	A2223	12		Inverted

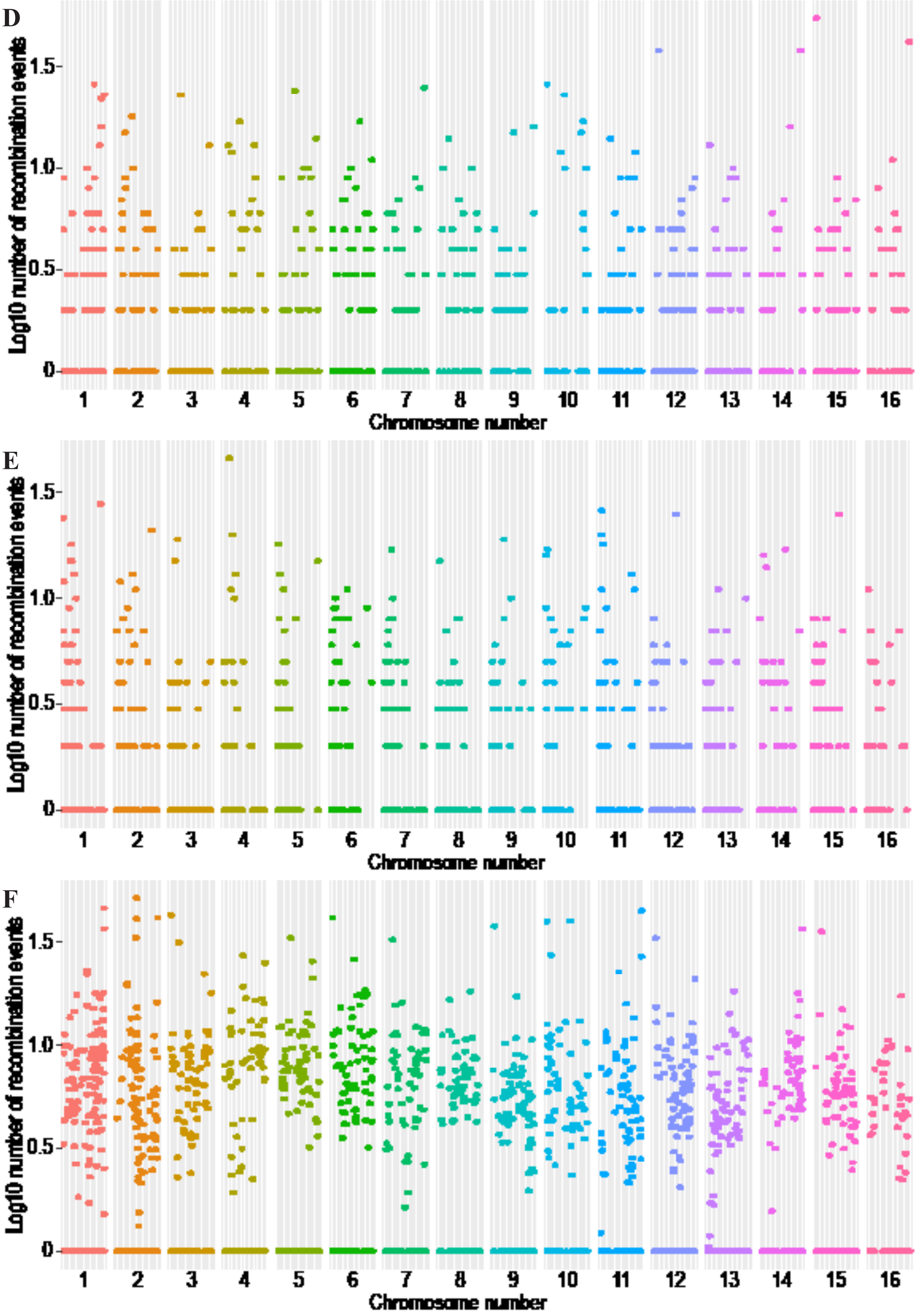
Assembly	Chromosome	Upper region (Mbp)	Lower region (Mbp)	Upper Marker	Lower Marker	Scaffold	Entire Scaffold (Y/N)	Inverted or misplaced
4.5	10	12.84	6.03	A2521	A2414	23 – 29	Y	Misplaced
4.5	10	4.8	0.04	A2412	A2322	1 – 18	Y	Misplaced
4.5	11	14.38	13.87	A2775	A2742	20	Y	Inverted
4.5	12	7.59	4.84	A2901	A2822	13	Y	Inverted
4.5	12	8.91	8.23	A2935	A20910	16	Y	Inverted
4.5	12	11.74	9.14	A3016	A2936	17	Y	Inverted
4.5	13	2.13	1.49	A3086	A3052	4	Y	Inverted
4.5	13	5.48	3.26	A3169	A3113	7	Y	Inverted
4.5	14	0.56	0.36	A3284	A3280	1	Y	Inverted
4.5	14	3.08	2.68	A3314	A3306	8	Y	Inverted
4.5	14	8.39	7.71	A3373	A3361	14	Y	Inverted
4.5	14	10.06	8.82	A3421	A3375	15	Y	Inverted
4.5	16	0.24	0.01	A3640	A3642	1	Y	Inverted
4.5	16	0.99	0.46	A3643	A3664	2	Y	Inverted
4.5	16	4.09	2.32	A3700	A3667	4	Y	Inverted
4.5	16	4.89	4.91	A3704	A3702	5	Y	Inverted
4.5	16	5.74	5.49	A3724	A3716	6	N	Inverted
4.5	16	7.06	6.24	A3748	A3727	9	Y	Inverted

Additional File 2

Log10 of recombination events between adjacent markers along the (A) Amel_3.0 (n = 1456), (B) Amel_4.0 (n = 1556) and (C) Amel_4.5 (n = 2879) genome assemblies and the (D) de novo 3.0 (n = 1456), (E) de novo 4.0 (n = 1556) and (F) de novo 4.5 (n = 2879) maps.



Additional File 2

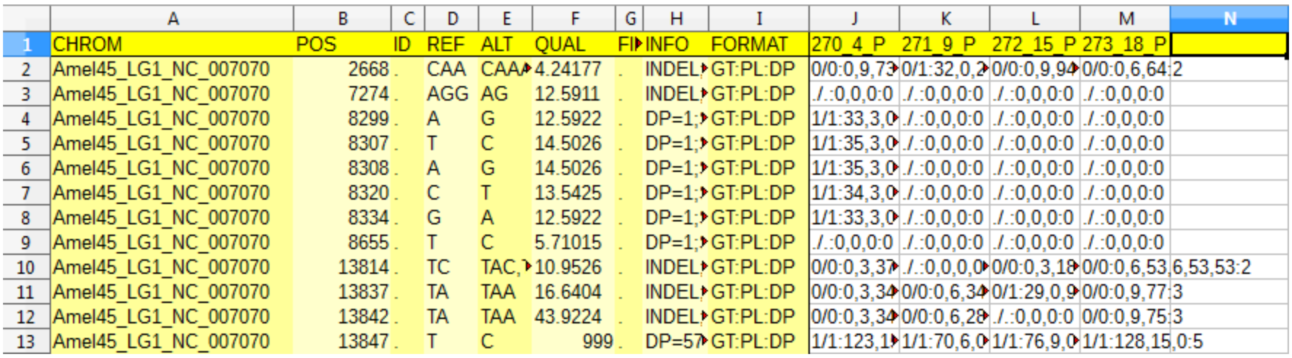


Additional File 4.

Methods and example for the use of R script *PhaseHaploid*, in phasing vcf formatted genotype data.

1. Application example

1.1 Data format



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	CHROM	POS	ID	REF	ALT	QUAL	FI	INFO	FORMAT	270_4_P	271_9_P	272_15_P	273_18_P	
2	Amel45_LG1_NC_007070	2668	.	CAA	CAA	4.24177	.	INDEL	GT:PL:DP	0/0:0,9,73	0/1:32,0,2	0/0:0,9,94	0/0:0,6,64:2	
3	Amel45_LG1_NC_007070	7274	.	AGG	AG	12.5911	.	INDEL	GT:PL:DP	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	
4	Amel45_LG1_NC_007070	8299	.	A	G	12.5922	.	DP=1;	GT:PL:DP	1/1:33,3,0	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	
5	Amel45_LG1_NC_007070	8307	.	T	C	14.5026	.	DP=1;	GT:PL:DP	1/1:35,3,0	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	
6	Amel45_LG1_NC_007070	8308	.	A	G	14.5026	.	DP=1;	GT:PL:DP	1/1:35,3,0	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	
7	Amel45_LG1_NC_007070	8320	.	C	T	13.5425	.	DP=1;	GT:PL:DP	1/1:34,3,0	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	
8	Amel45_LG1_NC_007070	8334	.	G	A	12.5922	.	DP=1;	GT:PL:DP	1/1:33,3,0	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	
9	Amel45_LG1_NC_007070	8655	.	T	C	5.71015	.	DP=1;	GT:PL:DP	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	./.:0,0,0:0	
10	Amel45_LG1_NC_007070	13814	.	TC	TAC,T	10.9526	.	INDEL	GT:PL:DP	0/0:0,3,37	./.:0,0,0,0	0/0:0,3,18	0/0:0,6,53,6,53,53:2	
11	Amel45_LG1_NC_007070	13837	.	TA	TAA	16.6404	.	INDEL	GT:PL:DP	0/0:0,3,34	0/0:0,6,34	0/1:29,0,9	0/0:0,9,77:3	
12	Amel45_LG1_NC_007070	13842	.	TA	TAA	43.9224	.	INDEL	GT:PL:DP	0/0:0,3,34	0/0:0,6,28	./.:0,0,0:0	0/0:0,9,75:3	
13	Amel45_LG1_NC_007070	13847	.	T	C	999	.	DP=57	GT:PL:DP	1/1:123,1	1/1:70,6,0	1/1:76,9,0	1/1:128,15,0:5	

The dataset should resemble the one shown in the picture in the following aspects:
- the first line (and only the first line) should be a heading, the text in the heading is irrelevant and may also contain empty strings
- the first column should contain the identifier for the chromosomes, it is mandatory that the strings are the same within each chromosome
- the second column should contain the physical distance on the chromosome, if this information is not given, the column should contain zeros
- Genotypes should be formatted as diploidß
- the other columns are irrelevant

1.2 Input

The file should be saved as either a comma or tab delimited file. Other file formats may also work but were never tested.

1.3 Using the script

I. Run the *PhaseHaploid* script (Additional file 4) in R.

II. Run one of the following functions and set the parameters:

a) phase(data, ExprBeginAtCol, DistanceCol, MinDensity, DistributionVaryBy, IncludeC, RemoveArtifacts)

b) phase_two(data1, data2, ExprBeginAtCol, DistanceCol, MinDensity, DistributionVaryBy, IncludeC, RemoveArtifacts)

1. 4 Arguments

data	datasets saved in variant call format.
ExprBeginAtCol	column of the first sample.
DistanceCol	column that carries the physical distance information.
MinDensity	a float value between 0 and 1. Used to filter out rows with low data density, where 0 means 'use every row' and 1 means 'use row with full data density only'.
DistributionVaryBy	a float value to drop rows which do not follow a 50:50 segregation. 0 means 'only strickt 50:50 segregation' and 1 means 'no filtering by segregation at all'.
IncludeC	a boolean value to determine whether phase 'C' should be tried to include to either 'A' or 'B'. Inclusion is based on the distance to the nearest phase.
RemoveArtifacts	a boolean value to determine whether expressions with higher-than-one ploidy should be removed from the dataset.

1.4 Output

The script outputs a .csv file in the current working directory.

2. Methods

2.1 Selecting the expression table

Since the VCF also stores non-genotypic annotations, it is necessary to define at which column the expression table begins. The associated command is called *ExprBeginAtCol* and demands a numeric input for the first column containing genotypic data. Additionally, it is mandatory that the dataset has a single-line headline. As part of the VCF, this should be given automatically. If the dataset does not have a headline, it should be inserted. The content of the cells is irrelevant.

2.2 Filtering the dataset

To accelerate the calculation and prevent lines with no information value from bloating the result, *PhaseHaploid* can filter rows with low data density. *MinDensity* sets the minimum proportion of missing genotypes before a row is excluded from the dataset. The default value is 0.3. *PhaseHaploid* can also filter data by dropping lines which do not follow a 50:50 segregation between the two phases. *DistributionVaryBy*, like *MinDensity*, requires a value between 0 and 1, with 0 returning a ‘strict 50:50 distribution' and 1 returning unfiltered data.

2.3 Phasing

The phasing is initiated by assigning a phase to both modes ('mode' means equal/not equal to the reference genome) in the first row. Phasing proceeds by comparing each row to the one before. To do so *PhaseHaploid* iterates the cells and sums up all positions where the mode has not changed. This is necessary because the mode can switch without the phase switching as well. If there are more indicators that the mode has switched it will adapt for the phasing of the next row. The algorithm always opts for the most parsimonious solution. Each chromosome is phased individually.

2.4 Inclusion of ambiguous data

Since NGS can produce ambiguous data (“phase C”), *PhaseHaploid* can attempt to reconcile this by checking upstream and downstream of the ambiguity to identify the nearest determined phase. The ambiguity is then assigned to this phase, as recombination events are expected to be rare.

2.5 Removal of artifacts

Due to errors or contaminations in sampling it is possible that, for a given locus, the ploidy of the sample is greater than one. The rows containing these data points are considered artifacts and removed from the dataset. This can be changed using *RemoveArtifacts*, for which the default is: *TRUE*.

2.6 Estimating recombination frequency

Useful for determining genetic distances, the script will count the crossing over events in the dataset. This is added in a column at the end of the table with the heading “RECOMB”, containing a vector of numbers indicating the amount of recombination events from the previous row to the actual one.

3. Pseudocode for phasing script

3.1 Phasing one dataset

```
for( 'all Linkage Groups' ){
        'subset all rows belonging to this Linkage Group'
        'in the first line assign A or B depending on relation to reference genome'

        for( 'for all rows in this Linkage Group'){
            # AB defines which cells get assigned which class
            'compare this line to the previous line’
            AB = 'depending on the last rows distribution'

            'assign A or B to this row depending on AB'
        }
}
```

3.2 Phasing two datasets

```
for( 'all Linkage Groups' ){
        'subset all rows belonging to this Linkage Group"

        'find a shared locus'
        'in the first line assign A or B depending on relation to reference genome'

        for( 'for all rows in this Linkage Group" ){
            # AB defines which cells get assigned which class
            'compare this line to the previous line’
            AB = 'depending on the last rows distribution'

            'assign A or B to this row depending on AB'
        }
    'merge both datasets depending on their physical distance'
}
```

Additional File 5

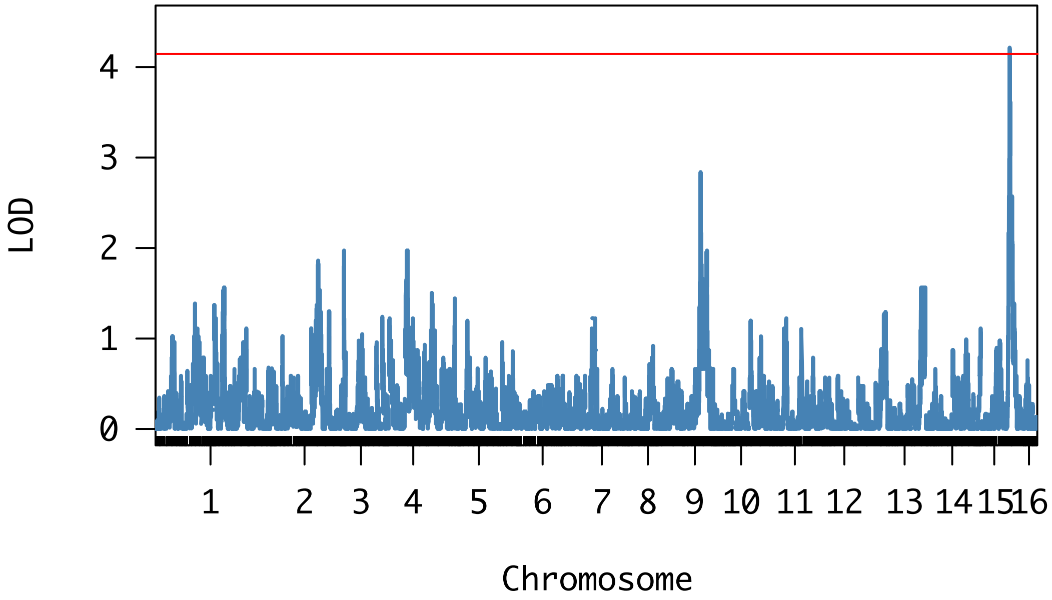
```
#####################################################################################
################################

# Functions

#####################################################################################
################################

# main
phase = function(data, ExprBeginAtCol, DistanceCol, MinDensity=0.2,
DistributionVaryBy=0.3, IncludeC=T, RemoveArtifacts=T){

  distribution = rcountAB(data)
  density = get_density(distribution, length(data[1,ExprBeginAtCol : length(data[1,])]))
  selection = data[c(1,which(((density > MinDensity) & (distribution[,1]
                                        < (1 + DistributionVaryBy)) & (distribution[,1]  > (1 -
DistributionVaryBy))))),]
  result = phasing(selection, ExprBeginAtCol)

  if(RemoveArtifacts == TRUE) {result = remove_artifacts(result, ExprBeginAtCol)}
  if(IncludeC == TRUE) {result = includeC(result, ExprBeginAtCol, DistanceCol)}

  result = cbind(result, cbind(get_distance(result),countrecomb(result, ExprBeginAtCol)))

  utils::write.table(result, file  = "result.csv", row.names=FALSE, col.names=FALSE, sep=";")

  return("Job complete, file saved")
}

phase_two = function(data1, data2, ExprBeginAtCol, DistanceCol, MinDensity=0.2,
DistributionVaryBy=0.3, IncludeC=T, RemoveArtifacts=T){

  distribution1 = rcountAB(data1)
  density1 = get_density(distribution1, length(data1[1,ExprBeginAtCol : length(data1[1,])]))
  selection1 = data1[c(1,which(((density1 > MinDensity) & (distribution1[,1]
                                          < (1 + DistributionVaryBy)) & (distribution1[,1]  > (1 -
DistributionVaryBy))))),]

  distribution2 = rcountAB(data2)
  density2 = get_density(distribution2, length(data2[1,ExprBeginAtCol : length(data2[1,])]))
  selection2 = data2[c(1,which(((density2 > MinDensity) & (distribution2[,1]
                                          < (1 + DistributionVaryBy)) & (distribution2[,1]  > (1 -
DistributionVaryBy))))),]

  result = phasing_and_merge(selection1, selection2, ExprBeginAtCol)
  if(RemoveArtifacts == TRUE) {result = remove_artifacts(result,ExprBeginAtCol)}
  if(IncludeC == TRUE) {result = includeC(result, ExprBeginAtCol,DistanceCol)}

  result = cbind(result, cbind(get_distance(result),countrecomb(result, ExprBeginAtCol)))

  utils::write.table(result, file  = "result.csv", row.names=FALSE, col.names=FALSE, sep=";")

  return("Job completet, file saved")
}


# preprocessing
rcountAB = function(dataset) {
  # this function counts the occurrences of the 2 phases for each row and returns a 3 column
matrix
  # first column: the A to B ratio
  # second column: number of As
  #thrid column: number of Bs


  # a matrix that will later carry the result
  # three columns and a rows depending on the length of the dataset
  erg = matrix(0,length(as.matrix(dataset[,1])),3)

  # we need to transform the data frame into a matrix, because data frames tend to behave
strange
  dataset = as.matrix(dataset)

  for(i in 1:length(dataset[,1])) {
    # count the As and Bs
    nrA  = length(which(grepl("1/1", dataset[i,])))
    nrB  = length(which(grepl("0/0", dataset[i,])))

    # if there are no Bs we need to manually set the value of this row to catch an error
    if(nrB == 0) {
      erg[i,1] = 0
      erg[i,2] = nrA
      erg[i,3] = nrB
      next
    }

    erg[i,1] = nrA/nrB
    erg[i,2] = nrA
    erg[i,3] = nrB
  }
  return(erg)
}
```

```
get_density = function (dataset, nrExpr) {
  # calculates the density of markers in a row
  # therefor it needs the result of the rcountAB function

  erg = matrix(0,length(dataset[,1]), 1)
  end = length(dataset[,1])
  lastc = length(dataset[1,])
  for (i in  1:end) {
    erg[i]= (as.numeric(dataset[i,lastc]) + as.numeric(dataset[i, lastc-1])) / nrExpr
  }
  return (erg)
}

get_distance = function(dataset) {
  # calculates the distances of each position to the previous position

  dataset = as.matrix(dataset)
  erg = matrix(0,length(dataset[,1]),1)
  erg[1] = "DISTANCE"

  NrLGs = length(as.character(unique(dataset[,1]))) #number of LGs

  for(k in 1:NrLGs) {
    vec = which(dataset[,1] == as.character(unique(dataset[,1])[k]))

    if(length(vec)>1){
      for(i in vec[-1]) {
        erg[i,1] = as.numeric(dataset[i,2]) - as.numeric(dataset[i-1,2])
      }
    }
  }
  return(erg)
}


# phasing
initialise = function(datarow, colstart, A, B) {
  # changes all cells in a given row to the phase it belongs

  end = length(datarow)
  datarow = as.matrix(datarow)
  for ( i in colstart:end) {
    if(grepl(A, as.character(datarow[i]))) {
      datarow[i] = "A"
      next
    }
  if(grepl(B, as.character(datarow[i]))) {
   datarow[i] = "B"
   next
  }
  if(grepl("[.]", as.character(datarow[i]))) {
   datarow[i] = "-"
   next
  }
  if(grepl("0/1", as.character(datarow[i]))) {
   datarow[i] = "C"
   next
  }
 }
 return (datarow)
}

get_phase = function(dataset, row, AB) {
 # tries to infer the phase for a row
 # to do this it compares a line with the already phased previous line

 eq = 0 # the number of precedents that the second line has the same AB as the first
 neq = 0 # the number of precedents that the second line has NOT the same AB as the first

 # compares the cells of the previous row with the ones from the actual row
 # depending on which expression we find either eq or neq is increased
 for(i in which(grepl(AB[1], as.matrix(dataset[row-1, ])))) {
  if(grepl(AB[1], dataset[row, i])) {
   eq = eq + 1
  }
  if(grepl(AB[2], dataset[row, i])) {
   neq = neq + 1
  }
 }

 if(eq > neq) {
  return(c(AB[1],AB[2]))
 }
 return(c(AB[2],AB[1]))
}

get_phase_reversed = function(dataset, row, AB) {
 # does the same as the above function, just in the opposite direction

 eq = 0
 neq = 0
 for(i in which(grepl(AB[1], as.matrix(dataset[row+1, ])))) {
  if(grepl(AB[1], dataset[row, i])) {
      eq = eq + 1
    }
    if(grepl(AB[2], dataset[row, i])) {
      neq = neq + 1
    }
  }

  if(eq > neq) {
    return(c(AB[1],AB[2]))
  }
  return(c(AB[2],AB[1]))
}

find_first_shared_loci = function(dataset1, dataset2, vec1, vec2) {
  # searches for the first shared loci in the two datasets
  # gets two vectors containing all rows for a LG and the datasets
  # it just checks whether two positions are the same in this LG
  erg = intersect(dataset1[vec1,2], dataset2[vec2,2])

  #if there is no shared loci, return 0
  if(length(erg)==0) {return(0)}

  # return 0 if there s no shared locus
  return(max(erg[1],0))
}

phasing_and_merge = function(dataset1, dataset2, colstart) {
  # main function that applies the other function on the datasets
  # merges both datasets at the end

  dataset1 = as.matrix(dataset1)
  dataset2 = as.matrix(dataset2)

  NrLGs = getLGcount(dataset1) # number of LGs (equal for both datasets)
  # it is expected, that the first row is a heading

  # initialisation of the two intermediate results
  zwerg1 = dataset1
  zwerg2 = dataset2


  # this loop iterates all the LG (excluding the first, which is supposed to be a heading)
  for(k in 2:NrLGs){
    # first we extract the rows belonging to a LG
    vec1 = getVecOfLG(zwerg1, k) # vector of the given LG in dataset1
    vec2 = getVecOfLG(zwerg2, k) # vector of the given LG in dataset2
    # now we look for the first shared locus
    # this will later be our anchor point in the phasing
    start1 = which(zwerg1[,2] == find_first_shared_loci(zwerg1, zwerg2, vec1, vec2)
                & (zwerg1[vec1[1],1] == unique(zwerg2[,1])[k]))
    if(length(start1)==0) {start1 = vec1[1]}

    start2 = which(zwerg2[,2] == find_first_shared_loci(zwerg1, zwerg2, vec1, vec2)
                & (zwerg2[vec2[1],1] == unique(zwerg2[,1])[k]))
    if(length(start2)==0) {start2 = vec2[1]}

    # we started with A as 1/1 and B as 0/0
    # this variable belongs to the initialisation, it tells which value will be translated to A or rather B
    AB = c("1/1", "0/0")

    #initialising each LG at the shared loci
    zwerg1[start1, ] = initialise(dataset1[start1, ], colstart, AB[1], AB[2])
    zwerg2[start2, ] = initialise(dataset2[start2, ], colstart, AB[1], AB[2])

    #if the first loci is the shared one, we cant look up
    if(vec1[1] != start1) {
      AB = c("1/1", "0/0")
      for(i in (start1-1) : vec1[1]){
       AB = get_phase_reversed(dataset1, i, AB)
       zwerg1[i, ] = initialise(zwerg1[i, ], colstart, AB[1], AB[2])
      }
    }

    #if vec1 has just 1 entry, we already dealt with it
    if(length(vec1) > 1) {
      end = vec1[length(vec1)]
      AB = c("1/1", "0/0")
      for(i in (start1+1) : end){
       AB = get_phase(dataset1, i, AB)
       zwerg1[i, ] = initialise(zwerg1[i, ], colstart, AB[1], AB[2])
      }
    }


    #and all the same for the second dataset
    if(vec2[1] != start2) {
      AB = c("1/1", "0/0")
      for(i in (start2-1) : vec2[1]){
       AB = get_phase_reversed(dataset2, i, AB)
       zwerg2[i, ] = initialise(zwerg2[i, ], colstart, AB[1], AB[2])
      }
    }
    if(length(vec2) > 1) {
      end = vec2[length(vec2)]
      AB = c("1/1", "0/0")
      for(i in (start2+1) : end){
       AB = get_phase(dataset2, i, AB)
       zwerg2[i, ] = initialise(zwerg2[i, ], colstart, AB[1], AB[2])
      }
    }
   }

   # now the two datasets get finally merged
   erg = merge_two_datasets(zwerg1, zwerg2, colstart)
   return(erg)
  }

  phasing = function(dataset, colstart) {
   # main function that applies the other function on the datasets

   dataset = as.matrix(dataset)

   NrLGs = getLGcount(dataset) # number of LGs
   # it is expected, that the first row is a heading

   # initialisation of the intermediate result
   zwerg = dataset


   # this loop iterates all the LG (excluding the first, which is supposed to be a heading)
   for(k in 2:NrLGs){
    # first we extract the rows belonging to a LG
    vec = getVecOfLG(zwerg, k) # vector of the given LG in dataset

    # this will later be our anchor point in the phasing
    start = vec[1]

    # we started with A as 1/1 and B as 0/0
    # this variable belongs to the initialisation, it tells which value will be translated to A or
  rather B
    AB = c("1/1", "0/0")

    #initialising each LG at the shared loci
    zwerg[start, ] = initialise(dataset[start, ], colstart, AB[1], AB[2])

    # the calculation of 'end' boosts the performance of the loop
    # otherwise it would have to calculate the end at each step of the iteration again
    end = vec[length(vec)]

    AB = c("1/1", "0/0")
    if(start < end) {
      for(i in (start+1) : end){
       AB = get_phase(dataset, i, AB)
       zwerg[i, ] = initialise(zwerg[i, ], colstart, AB[1], AB[2])
      }
    }
   }

   return(zwerg)
  }

  merge_two_datasets = function(dataset1, dataset2, ExprBeginAtCol){
   # merges two datasets into one
   # they need to have the same format (start of expressiontable, etc. )

   NrLGs = length(as.character(unique(dataset1[,1]))) # number of LGs (equal for both
  datasets)

   # the first row of the result is the merging of heading from the first and all the individuals
   # from the second dataset
   erg = matrix("",length(dataset1[,1]) + length(dataset2[,1]), length(dataset1[1, ]) +
  length(dataset2[1, ExprBeginAtCol:length(dataset2[1,])])))
   erg[1, ] = c(dataset1[1, ], dataset2[1,ExprBeginAtCol:length(dataset2[1,])])

   #global row-counter, since we are using a while-loop
   i = 1

   # this loop iterates all the LG (excluding the first, which is supposed to be a heading)
   for(k in 2:NrLGs){

    # first we extract the rows belonging to a LG
    vec1 = which(dataset1[,1] == as.character(unique(dataset1[,1])[k])) # vector of the given
  LG in dataset1
    vec2 = which(dataset2[,1] == as.character(unique(dataset2[,1])[k])) # vector of the given
  LG in dataset2

    current1 = 1 # the current row for the first dataset
    current2 = 1 # the current row for the second dataset


    #calculating this befor the loop boosts the performance
    lengthvec1 = length(vec1)
    lengthvec2 = length(vec2)
    # loop that works till one of the datasets is at the end of its LG
    while((current1 <= lengthvec1) & (current2 <= lengthvec2)){
      i = i+1

      # first case: position in dataset1 < pos in dataset2
      if(as.numeric(dataset1[vec1[current1],2]) < as.numeric(dataset2[vec2[current2],2])) {

       # merge the two data frames and bind them to the result
       erg[i, ] = c(dataset1[vec1[current1], ], matrix("-", 1,
  length(dataset2[1,ExprBeginAtCol:length(dataset2[1,])])))
       current1 = current1 + 1
       next
      }


      # second case: position in dataset1 > pos in dataset2
      if(as.numeric(dataset1[vec1[current1],2]) > as.numeric(dataset2[vec2[current2],2])) {

       # merge the two data frames and bind them to the result
       erg[i, ] = c(dataset2[vec2[current2],1:(ExprBeginAtCol-1)], matrix("-", 1,
  length(dataset1[1,ExprBeginAtCol:length(dataset1[1,])])), dataset2[vec2[current2],
  ExprBeginAtCol :length(dataset2[1,])])
       current2 = current2 + 1
       next
      }


      # third case: position in dataset1 = pos in dataset2
      if(as.numeric(dataset1[vec1[current1],2]) == as.numeric(dataset2[vec2[current2],2])) {

       # merge the two data frames and bind them to the result
       erg[i, ] = c(dataset1[vec1[current1], ], dataset2[vec2[current2], ExprBeginAtCol
  :length(dataset2[1,])])
       current1 = current1 + 1
       current2 = current2 + 1
       next
      }
    }

    # bind the rest of dataset1 to the result
    if(current1 < lengthvec1) {
      for(j in current1 : lengthvec1){
       i = i + 1
       erg[i, ] = c(dataset1[vec1[j], ], matrix("-", 1,
  length(dataset2[1,ExprBeginAtCol:length(dataset2[1,])])))
      }
    }

    #bind the rest of dataset2 to the result
    if(current2 < lengthvec2) {
      for(j in current2 : lengthvec2){
       i = i + 1
       erg[i, ] = c(dataset2[vec2[j],1:(ExprBeginAtCol-1)], matrix("-", 1,
  length(dataset1[1,ExprBeginAtCol:length(dataset1[1,])])), dataset2[vec2[j],
  ExprBeginAtCol:length(dataset2[1,])])
      }
    }
   }
   return (erg)
  }

  remove_artifacts = function(dataset, ExprBeginAtCol){
   # this function removes the artifacts in a dataset, as artifact counts:
   # each row which has cells like 0/2, 1/2, 2/2, 0/3, 1/3, 2/3 or 3/3 in it

   # change the dataframe to a matrix
   dataset = as.matrix(dataset)
   len = length(dataset[1,])

   # result vector, saves all lines to delete
   erg = vector("integer", length(dataset[,1]))
   rowcounter = 1

   # iterate the dataset and save all lines with artifacts
   end = length(dataset[,1])
   for(i in  2:end){

    # now following: "best of copy paste" by eike oertelt
    # if we have at least one artifact in a row, the row will be saved and later deleted
    if(TRUE %in% grepl("0/2", dataset[i, ExprBeginAtCol:len])){
      erg[rowcounter] = i
      rowcounter = rowcounter + 1
      next
    }

    if(TRUE %in% grepl("1/2", dataset[i, ExprBeginAtCol:len])){
      erg[rowcounter] = i
      rowcounter = rowcounter + 1
      next
    }

    if(TRUE %in% grepl("2/2", dataset[i, ExprBeginAtCol:len])){
      erg[rowcounter] = i
      rowcounter = rowcounter + 1
      next
    }

    if(TRUE %in% grepl("0/3", dataset[i, ExprBeginAtCol:len])){
      erg[rowcounter] = i
      rowcounter = rowcounter + 1
      next
    }

    if(TRUE %in% grepl("1/3", dataset[i, ExprBeginAtCol:len])){
      erg[rowcounter] = i
      rowcounter = rowcounter + 1
      next
    }

    if(TRUE %in% grepl("2/3", dataset[i, ExprBeginAtCol:len])){
      erg[rowcounter] = i
      next
    }

    if(TRUE %in% grepl("3/3", dataset[i, ExprBeginAtCol:len])){
      erg[rowcounter] = i
      rowcounter = rowcounter + 1
      next
    }
   }
   #break if no artifacts are there
   if(sum(erg)==0){return(dataset)}

   dataset = dataset[-erg, ]
  }


  # misc
  getLGcount = function(data) {
   # counts all the LG in a dataset
   erg = length(as.character(unique(data[,1])))
   return (erg)
  }

  getVecOfLG = function (data, k){
   #return a vector of all the loci belonging to a LG (k) in this Dataset
   erg = which(data[,1] == as.character(unique(data[,1])[k]))
   return(erg)
  }
```

```
countrecomb = function(data, ExprBeginAtCol) {
  # counts the visual recombination events in a row

  # result: a vec of the numbers
  erg = matrix(0,length(data[,1]),1)
  erg[1] = "RECOMB.EVENTS"

  # for each LG
  end1 = getLGcount(data)
  for(k in 2 : end1){

    vec = getVecOfLG(data, k)

    if(length(vec) > 1) {
      # for all rows in this LG
      end2 = length(vec)
      for(i in 2:end2){

        #intermediate result counts the number of crossing overs
        zwerg = 0

        # for each cells in this row
        end3 = length(data[1,])
        for(j in ExprBeginAtCol : end3){

          # count recomb.events

          #if we have an A or a B in the current cell
          if((as.character(data[vec[i],j]) == "A") | (as.character(data[vec[i],j]) == "B")){

            #if the previous cell is "-" we have to check the first upstream cell that isnt a "-"

            bla = lookup(data, vec, i-1, j)
            if(bla != 0) {
              if(as.character(data[vec[i],j]) != (as.character(data[bla,j]))) {
                zwerg = zwerg + 1
              }
            }
          }
        }
        erg[vec[i]] = zwerg
      }
    }
  }
  return(erg)
}
```

```
lookup = function(data, vec, current, col) {
 # checks starting from the current position the upstream cells for their phase and returns
the row
 # returns 0 if there is no A or B inside the index

 #current means the current position in the 'vec'-vector
 ABpos = 0

 for(i in (current):1){
  if((as.character(data[vec[i], col])=="A") | (as.character(data[vec[i], col])=="B" )){
   ABpos = vec[i]
   break
  }
 }

 return(ABpos)
}

lookdown = function(data, vec, current, col) {
 # checks starting from the current position the downstream cells for their phase and
returns the row
 # returns 0 if there is no A or B inside the index

 ABpos = 0

 for(i in current:length(vec)){
  if((as.character(data[vec[i], col])=="B") | (as.character(data[vec[i], col])=="A" )){
   ABpos = vec[i]
   break
  }
 }

 return(ABpos)
}

includeC = function(data, ExprBeginAtCol, DistanceCol) {
 # we include the C by checking up- and downstrem (inside the column)

 # for each LG
 for(k in 2: (getLGcount(data))){

  vec = getVecOfLG(data, k)

  if(length(vec) > 1) {
   #for each row
   for(i in 1:length(vec)) {
        #for each cell
        end = length(data[1,])
        for(j in ExprBeginAtCol : end) {

          # check wheter the cell is a C
          if(grepl("C", as.character(data[vec[i],j]))) {

            # up and down give the row the first occurance of an A or B in the dataset
            up = lookup(data, vec, i, j)
            down = lookdown(data, vec, i, j)

            sumup = 0
            sumdown = 0

            # if up == 0, then there is no A or B and we dont need to deal with it
            if(up == 0) {
              sumup = Inf
            }
            else{
              sumup = as.numeric(data[vec[i], DistanceCol]) - as.numeric(data[up, DistanceCol])
            }

            # if down == 0, then there is no A or B and we dont need to deal with it
            if(down == 0){
              sumdown = Inf
            }
            else{
              sumdown = as.numeric(data[down, DistanceCol]) - as.numeric(data[vec[i],
DistanceCol])
            }


            if(sumup > sumdown) {
              data[vec[i],j] = data[down,j]
            }
            if(sumup < sumdown) {
              data[vec[i],j] = data[up,j]
            }
          }
        }
      }
    }
  }

  return(data)
}
```

Extended data figure 1



Genome-wide single-locus QTL analysis comparing pupae where the infesting Varroa was alive but did not produce daughters or where the mite reproduced successfully. The p = 0.05 significance threshold (LOD = 4.15) is shown by a solid red line. The highest LOD score was 4.21 (p = 0.029).

9. Curriculum Vitae

Personal information

Name: Conlon, Benjamin Hanson
Date of birth: 17th July 1991
Birthplace: Liverpool, United Kingdom
Nationality: British, Irish
Languages: English (Native), Danish (Conversational), German (Basic)

University education

2016-2018 Doctor rerum naturalium (Dr. rer. nat.), Biology – MolecularEcology,

Department of Zoology, Institute for Biology, Martin-Luther-University, Germany.
Thesis title: "Of Mites and Men: The independent evolution of host-induced *Varroa* infertility in the drone brood of *Apis mellifera*"
Supervised by: Prof. Robin F.A. Moritz and Dr. Jarkko Routtu.

2013-2015 Master of Science (M.Sc.), Biology – Ecology and Evolution, Department of Biology, University of Copenhagen, Denmark.
Thesis title: "Phyolgenetic analyses of termite-associated fungi."
Supervised by: Associate Prof. Michael Thomas-Poulsen and Assistant Prof. Henrik H. de Fine Licht.

2009-2013 Bachelor of Science (B.Sc.) with honours, Biological Sciences – Ecology, University of Stirling, Scotland.
Thesis title: "Spatial distribution of the bumble bee wax moth (*Aphomia sociella*) in Central Scotland."
Supervised by: Prof. Dave Goulson.

Publications

- 4 papers in peer-reviewed journals
- 8 papers close to submission, submitted, or in revision
- 2 invited talks at international meetings and departmental seminars
- 9 contributed oral presentations at national and international meetings/conferences?
- 6 contributed poster presentations at national and international meetings

Short-term stays abroad

2017	Fieldwork, Toulouse, France (total 2 weeks)
2017	Fieldwork, Prof. A. Brockmann's Lab, NCBS, Bangalore, India (total 6.5 weeks)
2016	Fieldwork, Prof. P. Rosenkranz's lab at the University of Hohenheim, Germany (total 1 week)
2015	Fieldwork, FABI, University of Pretoria, South Africa (total 4.5 weeks)
2011-2012	Undergraduate exchange programme to University of Victoria, British Columbia, Canada (total 9 months)

Published papers

1. **Conlon, B.H.**, Frey, E., Rosenkranz., P, Locke, B., Moritz, R.F.A., Routtu, J. (2018) The role of epistatic interactions underpinning resistance to parasitic *Varroa* mites in haploid honeybee drones. *Journal of Evolutionary Biology*. In Press. DOI: https://doi.org/10.1111/jeb.13271
2. Buys, M., **Conlon, B.H.**, De Fine Licht, H.H., Aanen, D.K., Poulsen, M. and De Beer, Z.W. (2018) Searching for *Podaxis* on the trails on early explorers in southern Africa. *South African Journal of Botany*. **115**, 317. DOI: https://doi.org/10.1016/j.sajb.2018.02.150
3. **Conlon, B.H.**, de Fine Licht, H.H., Mitchell, J., de Beer, W., Christian Carøe, M. Thomas Gilbert, Eilenberg, J. and Poulsen, M. (2017) Draft genome announcement of the fungus-growing termite pathogenic fungi *Ophiocordyceps bispora*, *Data in Brief*. **11**, 537-542. DOI: https://doi.org/10.1016/j.dib.2017.02.051
4. **Conlon, B.H.**, de Beer, W., de Fine Licht, H.H., Aanen, D.K., and Poulsen, M. (2016) Phylogenetic analyses of diverse *Podaxis* specimens from Southern Africa reveals hidden diversity and new insights into its relationship with termites, *Fungal Biology*. **120(9)**, 1065-1076. DOI: https://doi.org/10.1016/j.funbio.2016.05.011

Papers close to submission, submitted, or in revision

Conlon, B.H., Oertelt, E., Moritz, R.F.A. and Routtu, J. Increasing recombination rate estimates result from decreasing assembly accuracy in honey bee (*Apis mellifera*) reference genome updates. *Journal of Heredity*. Manuscript in review.

Conlon, B.H., Kefuss, J., Aurori, A., Dezmirean, D.S., Moritz, R.F.A. and Routtu, J. A modified honey bee *ecdysone* pathway inhibits reproduction in *Varroa*. *Nature*. Manuscript submitted.

Conlon, B.H., Devraj, S., Brockmann, A., Moritz, R.F.A. and Routtu, J. Transcriptomic analysis of the inhibition of *Varroa* reproduction in its native honey bee host *Apis cerana.* Manuscript in preparation.

Oddie, M.A.Y., Beaurepaire, A., Blacquiere, T., **Conlon, B.H.**, Dahl, B., de Miranda, J., Frey, E., Laget, D., Le Conte, Y., Locke, B., Mondet, F., Moritz, R.F.A., Moro, A., Rosenkranz, P., Routtu, J. and Neumann, P. The influence of local environmental conditions on survival rate for *Varroa*-resistant honey bee (*Apis mellifera*) colonies. Manuscript in preparation.

Moro, A., Blacquiere, T., **Conlon, B.H.**, Dahl, B., de Miranda, J., Frey, E., Laget, D., Le Conte, Y., Locke, B., Mondet, F., Moritz, R.F.A., Neumann, P., Oddie, M.A.Y., Rosenkranz, P., Routtu, J. and Beaurepaire, A. Genetic diversity of *Varroa* in resistant honey bee (*Apis mellifera*) populations. Manuscript in preparation.

Oddie, M.A.Y., Beaurepaire, A., Blacquiere, T., **Conlon, B.H.**, Dahl, B., de Miranda, J., Frey, E., Laget, D., Le Conte, Y., Locke, B., Mondet, F., Moritz, R.F.A., Moro, A., Rosenkranz, P., Routtu, J. and Neumann, P. Modern beekeeping practices as a driver of honey bee (*Apis mellifera*) colony losses: a review. Manuscript in preparation.

Conlon, B.H., Aanen, D.K., Buys, M., de Beer, W., de Fine Licht, H.H. and Poulsen, M. *Podaxis*: state of the art. (invited review). *Fungal Biology*. Manuscript in preparation.

10. Eidesstattliche Erklärung

Halle (Saale), den 12.Juni 2018

Hiermit erkläre ich an Eides statt, dass diese Arbeit, in der gegenwärtigen bzw. in einer anderen Fassung, von mir bisher weder an der Naturwissenschaftlichen Fakultät I - Biowissenschaften der Martin-Luther-Universität Halle-Wittenberg noch an einer anderen wissenschaftlichen Einrichtung zum Zweck der Promotion eingereicht wurde.

Ich erkläre weiterhin, dass ich mich bisher noch nicht um den Doktorgrad beworben habe.

Ferner erkläre ich, dass ich diese Arbeit selbstständig und nur unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel angefertigt habe. Die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen sind als solche kenntlich gemacht worden.

Benjamin H. Conlon