The influence of ploidy-specific expression on selection

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Mark Christian Harrison BSc MSc

Department of Genetics

University of Leicester

2016

The influence of ploidy-specific expression on selection

Mark Christian Harrison

Abstract

More efficient selection is expected for haploid-expressed genes compared diploid-expressed genes. This is because recessive mutations can be masked from selection by a dominant allele in diploids but are always exposed to selection in haploids. The significance of this effect for haplodiploids was recognised by White in 1945, who predicted less efficient selection on genes with diploid-limited expression.

I present the first empirical support for these predictions for haplodiploids on the buff-tailed bumblebee, *Bombus terrestris*. I found evidence for weaker purifying selection on diploid-biased genes compared to haploid-expressed genes. This has led to higher protein divergence rates and polymorphism levels in diploid-biased genes compared to haploid-expressed genes. In contrast, I found no evidence of greater positive selection on haploid-expressed genes, suggesting that most new, recessive mutations may be deleterious.

In a second experiment I tested the effect of ploidy-specific selection in the plant *Arabidopsis thaliana* by comparing selection patterns between haploid pollen genes and diploid sporophytic genes. I detected evidence for a change in selection patterns possibly due to a loss of self-incompatibility. Divergence data indicate stronger positive selection within pollen genes during a period dominated by outcrossing, likely caused by pollen competition and haploid selection. Polymorphism data, on the other hand, reveal signs of relaxed selection within pollen genes, possibly due to high homozygosity levels, which reduce pollen competition and the masking of recessive mutations in diploid genes.

In a third study I used the data produced for determining ploidy-biased genes in *B. terrestris* to infer expression patterns involved in caste determination. This is the first broad scale analysis on caste determination in bumblebees. One major finding was that the expression patterns of bumblebee workers more closely resemble those of queens when reproductive compared to higher eusocial Hymenoptera, possibly due to the more plastic nature of bumblebee worker castes.

Acknowledgments

First I would like to thank both of my PhD supervisors Dr Rob Hammond and Dr Eamonn Mallon for developing such a fascinating research project. Thank you both for your brilliant support throughout the three and a half years and for always having an open door.

Further, I would like to express my gratitude to others who have supported me in different aspects of these studies. Thank you Julia for helping me collect bees. Thank you Prof Dave Twell for support on the development of the *Arabidopsis* project. Thank you Dr Harindra Amarasinghe for supporting me with advice and practical assistance in the bee-rearing experiment. Thank you Ceinwen Tilley for all lab related assistance. Thank you Dr Miranda Johnson for your support with deliveries. Also thank you Dr Toni Gossmann for help and advice with the *Arabidopsis* DoFE analysis.

Thank you to all my family for supporting me 100% when I decided to go back to university at the age of 37, especially my parents for offering financial support during some difficult patches of my studies prior to the PhD. A huge thank you, *muchas gracias* and *danke schön* go to Amparo and Georg for their unbelievable help and generosity during my write-up phase of this thesis. I am forever in your debt.

And of course, thank you again to Julia for always being there and listening to my endless ramblings on ploidy and selection. Sorry for being so moody.

This PhD was funded by the Natural Environment Research Council.

Contents

List of Tables

List of Figures

Abbreviations

α	The proportion of adaptive substitutions
$\pi(_n)$	(nonsynonymous) Nucleotide diversity
$\theta_w(_n)$	(nonsynonymous) Watterson's theta
ANCOVA	Analysis of covariance
DE	Differential expression
DoFE	Distribution of fitness effects
dN	Number of nonsynonymous substitutions per nonsynonymous site
dS	Number of synonymous substitutions per synonymous site
dN/dS	The ratio of dN and dS, or rate of protein evolution
FC	Fold change
FDR	False discovery rate
GO	Gene ontology
h	degree of dominance
indel	insertion / deletion mutation
L1-L4	Larval instar stages 1 to 4
mnc	mean normalised counts
$\mathbf{N}_e$	effective population size
$\mathbf{N}_e\mathbf{s}$	Strength of selection; s = selection coefficient
PCA	Principal component analysis
RNA	Ribonucleic acid
RNAseq	Ribonucleic acid sequencing
SEM	Standard error of the mean
SNP	Single nucleotide polymorphism
(g)VCF	(genomic) Variant calling format file
Vg	Vitellogenin
Castes/sexes, developmental stages in chapter 2	
M, Q, W, MQ	Male, Queen, Worker, Mother Queen
r/R, u, N	reproductive, undetermined reproductive status, non-reproductive
L, P, A	Larvae, Pupae, Adults

Chapter 1

Introduction

1.1 Thesis rationale

The main aim of my PhD was to investigate the influence of ploidy-specific expression on selection. In his chapter "Sex determination by male haploidy", MJD White (1945) predicted that selection would be more efficient on rare, recessive mutations in haplodiploid organisms (haploid males and diploid females) than in a diplodiploid species due to haploid selection in males. For genes, which are only expressed in the diploid sex of haplodiploids, on the other hand, rare recessive mutations can remain undetected from selection due to the masking of a dominant allele. Haldane (1924) even believed this mode of selection may have played an important role in the evolution of hymenopterans.

I decided to test these predictions on the buff-tailed bumblebee, *Bombus terrestris*, by comparing levels of selection on haploid-biased and non-biased genes to selection on diploid-biased genes. Before this analysis was possible, it was therefore necessary to determine genes with haploid-biased, diploid-biased and unbiased expression. I achieved this by conducting an RNAseq experiment on different combinations of castes, sexes and developmental stages and then measuring differential expression between these. This was the first time that a genome-wide differential expression analysis had been carried out on bumblebees. Therefore, the data which were generated by this analysis allowed me to address further interesting questions on caste determination within bumblebees. The results of this analysis are presented in chapter 2 and were published in Molecular Ecology (Harrison *et al.*, 2015).

In the following study, which is presented in chapter 3, I compared levels of selection between haploid-biased, diploid-biased and non-biased genes in *B. terrestris* to test White's predictions on haplodiploids. To further test this phenomenon I compared levels of selection between pollen-specific genes (haploid) and sporophyte-

specific genes (diploid) in *Arabidopsis thaliana*. The results of this third project are presented in chapter 4.

In this introduction I begin by discussing theoretical expectations regarding the effects of ploidy-specific expression on selection. This is followed by a summary of existing, empirical studies on the phenomenon within different taxonomical groups. In the proceeding section I offer background on the biological aspects of *B. terrestris*, which are relevant for the caste-determination study. At the end of each section within this introduction I formulate my predictions for the three studies presented in chapters 2, 3 and 4.

1.2 The effects of ploidy-specific expression on selection

1.2.1 Masking

In a diploid organism dominance affects the efficiency of selection to act on a mutation. A completely recessive, deleterious mutation will not be subject to purifying selection in heterozygous individuals and only be successfully selected against in homozygous individuals. Similarly, a recessive, advantageous mutation is less likely to be driven to fixation than a dominant mutation. This 'masking' of recessive alleles can lead to deleterious (and beneficial) mutations remaining within a diploid population at a low frequency (Otto and Goldstein, 1992; Nielsen, 2005). No such masking occurs in haploid organisms, in which mutations are always exposed to selection. Ploidy, therefore, influences the efficacy of selection to remove a deleterious mutation or to increase the frequency of an advantageous mutation (Wallace, 1963). The significantly higher adaptation rates of haploid yeast populations in a direct comparison with diploid populations appears to support this masking hypothesis (Zeyl *et al.*, 2003). The ploidy level at which a gene is expressed can vary within the same genome, for example (i) between autosomal genes and genes located on a sex-chromosome in ZZ/ZW and XX/XY organisms, (ii) between genes with male or female-specific expression in haplodiploid organisms, or (iii) between genes with expression limited to pollen or sporophyte tissues in plants. In the following sections I discuss how evolutionary rates may vary between genes with haploid and diploid-specific expression for each of these three groups of organisms.

1.2.2 Sex chromosomes & autosomes

1.2.2.1 Theoretical discussions

For organisms with the XY sex-determination system, the genes which are situated on the X-chromosome are haploid in the male, while all autosomal genes are diploid in both sexes. In the case of female heterogamety (ZZ/ZW system) Z-linked genes are haploid in the female. As early as 1924 Haldane predicted differences in selection for sex-specific genes (Haldane, 1924). He stated that selection is very ineffective on rare, recessive mutations unless they are sex-linked and expressed in the heterogametic sex. Charlesworth *et al.* (1987) began their paper on the relative rates of evolution of autosomes and sex chromosomes by citing Haldane directly:

> "*Moreover selection is ineffective on recessive characters when these are rare, except in the case of sex-linked characters, when selection is effective in the heterozygous sex.... It seems therefore very doubtful whether natural selection in random-mating organisms can cause the spread of autosomal recessive characters*"
> (Haldane, 1924, cited in Charlesworth *et al.*, 1987)

They went on to describe the likely connection of sex chromosomes with Haldane's rule, which describes how it is usually the heterogametic sex that is negatively affected in hybrids (Charlesworth *et al.*, 1987; Haldane, 1922). Empirical work was reviewed by Charlesworth *et al.* (1987) which pointed towards a special role of sex chromosomes in evolution and speciation, leading to the suggestion that genes on the X-chromosome evolve differently to autosomal genes. Models were developed which compared the evolutionary rates of sex chromosomes and autosomes with two main conclusions that are relevant for the current study:

(i) the fixation of beneficial mutations is faster on X or Z chromosomes relative to fixation on the autosomes if they are at least partially recessive ($h < 0.5$; h = degree of dominance);

(ii) the fixation of slightly deleterious, recessive mutations via random drift is slower on X and Z chromosomes due to more effective purging (Charlesworth *et al.*, 1987). This means that genes on the sex chromosomes evolve faster if the majority of non-synonymous substitutions result from the fixation of recessive, beneficial mutations or slower if most substitutions are driven by recessive, slightly deleterious mutations (Vicoso and Charlesworth, 2006). Quite often the former appears to be the case for some *Drosophila* (Baines *et al.*, 2008; Begun *et al.*, 2007) and mammalian species (Torgerson and Singh, 2003; Hvilsom *et al.*, 2012), giving rise to the term 'faster-X effect'.

1.2.2.2 Effective population size and other confounding factors

Testing the faster-X effect (or faster-Z effect in birds) empirically can be complicated by different factors. First, the effective population size of X-chromosomes (N_{eX}) is 3/4 the size of the N_e of autosomes (N_{eA}), so the efficacy of selection may be lower for X-chromosomal genes due to an increased influence of genetic drift (Wright, 1931). However, this is only true if the variation in reproductive success and the number of breeding individuals are equal between males and females (Vicoso and Charlesworth, 2009; Mank *et al.*, 2010a). A higher variance in mating success for males, as is the case in *Drosophila* due to male-male competition (Partridge *et al.*, 1987), would lower N_e for males since fewer males would contribute to the next generation than females (Vicoso and Charlesworth, 2009). This reduces the difference between N_{eX} and N_{eA} as the X-chromosome is less affected by the lower N_e of males, in which it spends only a third of its time (Vicoso and Charlesworth, 2009). In birds, the lower N_e of males has a greater effect on the Z-chromosome since it spends two thirds of its time in males and only one third in females (Mank *et al.*, 2010a). The result is an increased difference between N_{eZ} and N_{eA}, reducing the N_{eZ}:N_{eA} ratio to below 3/4 (Mank *et al.*, 2010a). Further factors which need to be considered are the overall effective population size, differences in mutation rates and dosage compensation (Mank *et al.*, 2010a; Vicoso and Charlesworth, 2009).

1.2.2.3 Empirical studies

Drosophila

One of the first empirical tests of the faster-X effect was carried out on *Drosophila melanogaster* (Betancourt *et al.*, 2002). As in many studies that followed, evolutionary rates were estimated by comparing interspecific divergence rates between orthologous sequences (in this case: *D. melanogaster* - *D. simulans*) at synonymous (dS) and nonsynonymous sites (dN). The ratio of these two rates (dN/dS), which controls for differences in the silent substitution rate, can be interpreted as the rate of protein evolution. In that first study no difference in dN/dS was found between X-chromosomal and autosomal genes, offering no evidence for the faster-X effect (Betancourt *et al.*, 2002). The authors suggested it was unlikely the lack of difference in dN/dS was due to their data set containing mainly neutrally evolving genes since in the same year Smith and Eyre-Walker (2002) had estimated that 45% of amino acid substitutions had been fixed by natural selection between *D. simulans* and *D. yakuba*. Instead Betancourt *et al.* (2002) assumed that most beneficial mutations may not be recessive so eliminating the effect of masking. Thornton *et al.* (2006) also found no evidence for the faster-X effect when comparing dN/dS between the *D. melanogaster* and *D. pseudoobscura* lineages. In *D. pseudoobscura* species the 3L

chromosome arm, which is autosomal in the *D. melanogaster* lineage, is fused with the X-chromosome. The faster-X hypothesis would predict higher dN/dS values for 3L homologues between lineages or within the *D. pseudoobscura* lineage compared to between *D. melanogaster* and *D. yakuba*, in which 3L is autosomal. This was not the case and the authors again assumed that most beneficial mutations may not be recessive or that adaptation may use standing variation rather than new mutations (Thornton *et al.*, 2006). A third study on the *Drosophila melanogaster* lineage also found no difference in dN/dS between X-chromosomal and autosomal genes (Connallon, 2007). However, they found pN/pS, which is analogous to dN/dS for intra-specific polymorphism data, to be significantly lower in X-chromosomal genes suggesting stronger purifying selection. This supports the second conclusion of Charlesworth *et al.* (1987) described in section 1.2.2.1.

Several studies have, however, found evidence for the faster-X effect in *Drosophila*. Thornton and Long (2002) calculated dN/dS between gene duplicates within the *D. melanogaster* genome and found higher evolutionary rates for paralog pairs which were both situated on the X-chromosome. Another study compared homologous chromosomal segments between *D. melanogaster* and *D. pseudoobscura*, finding higher dN rates among homologous X-chromosomal segments compared to all autosomal and autosome-X homologues (Musters *et al.*, 2006). Total divergence was higher and polymorphism lower for X-chromosomes in comparison to autosomes in *D. simulans* (Begun *et al.*, 2007). A multi-species approach found higher positive selection on X-chromosomal genes only in a few *Drosophila* species, whereas purifying selection was consistently higher for X-chromosomal genes (Singh *et al.*, 2008). Baines *et al.* (2008) found evidence for a faster-X effect in *D. melanogaster*, which was strongest in genes with male-biased expression. The faster evolution of male-biased, X-linked genes could be explained by higher rates of adaptive evolution and their greater exposure to heterogametic expression.

Mammals

The faster-X effect has also been investigated for mammals. Comparing genes from all tissues, Torgerson and Singh (2003) found no overall difference in dN/dS (human - mouse) between X-linked and autosomal genes. However, the protein evolution of X-linked sperm genes (genes only known to be expressed in spermatogonia or mature sperm) was significantly higher than autosomal sperm genes (Torgerson and Singh, 2003). Similarly, testis-specific (expressed only in testis tissue and not in brain, liver, kidney or heart), X-chromosomal genes showed significantly more differences in expression between humans and chimps than testis-specific, autosomal genes (Khaitovich *et al.*, 2005). The same pattern could not be found for genes ex-

pressed in the four other tested tissues, brain, liver, kidney and heart. Two further studies, however, managed to detect stronger positive selection for X-linked genes in general compared to autosomal genes by measuring divergence between humans and chimps (Lu and Wu, 2005; Nielsen *et al.*, 2005). Reduced diversity on the X-chromosome along with higher dN/dS rates compared to autosomal genes were also interpreted as evidence for stronger positive selection on X-linked genes in the mouse (Baines and Harr, 2007). Results of an investigation into levels of purifying and positive selection in two European rabbit subspecies (*Oryctolagus cuniculus*) were not as conclusive (Carneiro *et al.*, 2012). The proportion of substitutions between *O. c. algirus* and the Iberian hare (*Lepus granatensis*) that were fixed by positive selection (α) was higher for genes on the X-chromosome compared to autosomal genes, and the proportion of effectively neutral mutations was lower indicating stronger purifying selection on the X-chromosome. The same pattern was, however, not found for the second investigated subspecies *O. c. cuniculus* (Carneiro *et al.*, 2012).

Birds

Significantly higher dN/dS rates between chicken and zebra finch were measured on the Z-chromosome compared to autosomes (Mank *et al.*, 2007). This was interpreted as evidence for stronger adaptive evolution for Z-linked genes due to their hemizygous state in females. An alternative explanation was that the higher rate of protein divergence may be caused by relaxed purifying selection due to a lower N_e of Z-linked genes, which leads to a greater probability of the fixation of slightly deleterious mutations. That explanation was deemed improbable as pN/pS did not differ between Z and autosomal genes (Mank *et al.*, 2007).

However, a later study found evidence which supported the alternative explanation (Mank *et al.*, 2010b). Again dN/dS rates in chicken were significantly higher for Z-linked genes than autosomal genes. However, the relative difference in dN/dS between Z-linked and autosomal genes did not differ between genes with male-biased, female-biased and non-biased expression. The authors argued that in the case of stronger positive selection driving the faster protein divergence of Z-linked genes, the faster-Z effect would be greatest within female-biased genes, intermediate in non-biased genes and weakest in male-biased genes. The lack of difference in the faster-Z effect between these expression groups indicated that random drift, due to a reduced N_e for Z-linked genes, was the main cause of faster-Z in birds (Mank *et al.*, 2010b).

The reason random drift seems to cause the faster-Z effect in birds, whereas positive selection generally seems to be the accepted cause of faster-X in mammals

and *Drosophila*, is probably related to the size of N_{eZ} relative to N_{eA}. As described above (1.2.2.2), the ratio of N_{eZ}:N_{eA}, for example in birds, is expected to be smaller than N_{eX}:N_{eA} in mammals and *Drosophila*. In fact, within these groups, this ratio was found to be lowest for chicken at 0.30 (Sundström *et al.* 2004, cited in Mank *et al.* 2010a) and highest for *Drosophila* at around 1 (Connallon 2007 cited in Mank *et al.* 2010a). Thus selection efficacy is reduced for the Z-chromosomes in the chicken so that random drift becomes more dominant compared to autosomes.

1.2.3 The effects of ploidy-specific expression on selection in *Bombus terrestris*

1.2.3.1 Expectations for haplodiploids

In haplodiploids, for example Hymenoptera, males generally hatch from unfertilised eggs and are haploid (but see section 1.3.3.2), whereas females hatch from fertilised eggs and are diploid. In these taxa one would expect more efficient selection on recessive mutations compared to diploid organisms due to their haploid exposure to selection when expressed in males. However, recessive mutations should be less efficiently selected on if they are only expressed in the diploid females. This was recognised by Haldane (1922) when formulating his model of "selection of a sex-linked character in the homozygous sex only", stating that this type of selection must have played an important role in the evolution of Hymenoptera. Over 20 years later White (1945) formulated these ideas in more detail for the haplodiploids. He predicted that, in contrast to diplodiploid organisms, there would be no "reservoir of hidden variability" in haplodiploids due to haploid expression in males, except for mutations which are only expressed in females. The similarities of X-linked genes and haplodiploids regarding selection were discussed by Hedrick and Parker (1997), predicting a lower frequency of recessive, deleterious alleles and a faster increase in the frequency of favourable alleles for X-linked genes and in haplodiploids than for autosomal genes or in diploids.

1.2.3.2 Project aims

To the best of my knowledge, these predictions have not been tested empirically in haplodiploids. This is even though a test in haplodiploids would allow a direct test of the prediction without being complicated by confounding factors such as differences in N_e as is the case when testing the faster-X or faster-Z effect. Differences in N_e between sexes are irrelevant, since a gene with expression limited to the haploid sex will, even though not expressed, still be carried by diploid individuals and vice versa.

In this project I aimed to test MJD White's (1945) predictions regarding selection in haplodiploids. Specifically, I tested the hypothesis that selection is less efficient on diploid-biased genes (over-expressed in queens or workers) in *Bombus terrestris* due to the masking of rare, recessive alleles. Diploid-biased genes are genes whose expression level is significantly upregulated in female castes compared to males. I predicted stronger purifying and positive selection on haploid-biased genes (upregulated in males compared to female castes) and on non-biased genes compared to diploid-biased genes. To test this prediction I grouped genes by ploidy-bias based on expression patterns measured in an RNAseq experiment. Within each of the three gene groups - haploid-biased, diploid-biased and non-biased - I measured divergence to *B. impatiens*, common eastern bumblebee, and polymorphism levels within 27 *B. terrestris* transcriptomes. Higher protein divergence (dN/dS) would be expected for haploid and non-biased genes compared to diploid-biased genes if positive selection dominates, and lower divergence would be expected if sequence divergence is mainly driven by the fixation of slightly deleterious, recessive mutations (Vicoso and Charlesworth, 2006). An analysis of polymorphism levels would allow a distinction between these two alternatives. Lower polymorphism levels would indicate stronger directional selection so that a combination of high interspecific protein divergence and low levels of intraspecific polymorphism would suggest strong positive selection (McDonald and Kreitman, 1991). Accordingly, low intraspecific polymorphism levels and low interspecific divergence would suggest strong purifying selection.

1.2.4 The effects of ploidy-specific expression on selection in *Arabidopsis thaliana*

1.2.4.1 Expectations for plants

The expectations for selection on ploidy-biased genes in haplodiploids are applicable to selection on pollen-biased and sporophyte-biased genes in plants, which are haploid and diploid expressed, respectively. As for haplodiploids, investigations comparing selection between the two gene groups should not be confounded by differences in N_e as all genes are carried in each tissue type, even if they are not expressed. As with X-linked genes and male- or non-biased genes in haplodiploids, selection should be more efficient on pollen genes compared to sporophyte genes. These predictions have been confirmed for *Capsella grandiflora*, an outcrossing member of the mustard family Brassicaceae, (Arunkumar *et al.*, 2013) by evidence of stronger purifying and positive selection on genes with expression restricted to pollen tissues compared to sporophyte-specific genes. In that study a combination of pollen competition and

haploid selection were assumed to be responsible for the observed selection patterns. Evidence for similar patterns in *Arabidopsis thaliana* have been inconclusive. In one recent study, higher dN/dS and pN/pS values in pollen-specific compared to sporophyte-specific genes were interpreted as weaker purifying selection on pollen-specific genes (Szövényi *et al.*, 2013). Higher tissue specificity and expression noise (variance in expression level) of pollen-specific genes were offered as reasons for these unexpected findings. In a second study no difference could be found in dN/dS between pollen-biased and sporophytic genes (Gossmann *et al.*, 2013). However, evidence of stronger purifying and positive selection on pollen genes were presented. The discrepancy between these two studies could have different causes. First, as discussed by Gossmann *et al.* (2013) the effect of haploid expression should be reduced in *A. thaliana* when analysing polymorphism data. This is because *A. thaliana* became self-compatible roughly 1 million years ago (Tang *et al.*, 2007) leading to an increase in homozygosity levels due to high selfing rates (Nordborg, 2000; Wright *et al.*, 2008; Platt *et al.*, 2010). If a high proportion of loci are homozygous then the effect of masking will be reduced and the efficacy of selection will be increased for diploid genes. Divergence data, on the other hand, generally measured to *A. lyrata*, from which *A. thaliana* diverged about 13 million years ago (Beilstein *et al.*, 2010), is likely to mainly reflect selection patterns similar to outcrossing species. Second, the choice of genes may have an effect. Gossmann *et al.* (2013) chose more specific gene groups (pollen, pollen tube and sperm cell) based on a significant difference in expression level, whereas Szövényi *et al.* (2013) included all genes which were exclusively expressed in pollen tissues.

1.2.4.2 Project aims

I aimed to test the hypothesis that selection strength has changed for pollen-specific genes in *Arabidopsis thaliana* since it became self-compatible. Stronger selection should be detectable within pollen-specific genes compared to sporophyte-specific genes within divergence data due to the effect of haploid expression. The difference in selection efficacy should be reduced when incorporating polymorphism data due to a recent increase in homozygosity levels. This would help resolve the conflicting results presented in two recent studies on selection in pollen for *A. thaliana* (Szövényi *et al.*, 2013; Gossmann *et al.*, 2010), while also illustrating the effect of masking on selection.

1.3 Differential expression and caste determination in *Bombus terrestris*

1.3.1 Eusociality in Hymenoptera

Eusociality describes a system, in which reproduction is restricted to only one or a few individuals while the remaining, sterile individuals of a colony care for the brood (Andersson, 1984). Eusociality occurs in several groups of organisms, including mole rats, snapping shrimps, termites and ambrosia beetles (Nowak *et al.*, 2010). Eusociality is common in the Hymenoptera, in which it has evolved multiple times (Andersson, 1984). However, it must be noted that the majority of hymenopteran species are solitary (Goulson, 2010). Of all known horntail and sawfly species, as well as all parasitoid hymenopteran species, none are known to be eusocial (Wilson, 2008). Among the social hymenopterans, the degree of eusociality ranges from the primitively eusocial taxa, such as the paper wasp (*Polistes*), some sweat bee species (*Lasioglossum*) or the hover wasp (*Eustenogaster fraterna*), to highly eusocial species, such as the honeybees (*Apis*), the tropical stingless bees (Meliponinae) and all ant species (Formicidae; Goulson, 2010; Wilson, 2008; Sumner *et al.*, 2006; Sadd *et al.*, 2015). In highly eusocial species queens and workers differ morphologically, and their fate is determined during development. Workers are sterile and often divided into subcastes with distinct tasks such as foraging, brood care or nest-guarding (Cameron, 1989). In primitively eusocial societies, castes are determined in adulthood, are temporary and do not differ morphologically (Sumner *et al.*, 2006).

1.3.2 The intermediately eusocial insect, *Bombus terrestris*

The buff-tailed bumblebee, *Bombus terrestris* (Linnaeus), belongs to the genus *Bombus* (Latreille) within the family Apidae (bees; Cameron *et al.*, 2007; Goulson, 2010; figure 1.1). A small fraction (about 25 of the known 250 species) of the bumblebees, the cuckoo bumblebees, are not social and live within the nests of social bumblebee species (Goulson, 2010). In my further description of the bumblebees I will be referring only to the social bumblebees, to which *B. terrestris* belongs.

Figure 1.1: *Bombus terrestris* workers tending to larvae.
From Sadd *et al.* (2015)

The bumblebees are often described as primitively eusocial (Cameron, 1989; Sadd *et al.*, 2015; Kawakita *et al.*, 2004) as they do not share all the traits of highly eusocial Hymenoptera. Morphologically, queens and workers only differ in size, and compared to advanced species bumblebees form much smaller colonies that are usually annual rather than perennial (Sadd *et al.*, 2015). A *B. terrestris* colony will contain up to 350 workers (Goulson, 2010), while a honeybee (*Apis mellifera*) colony, for example, can contain 20,000 to 100,000 workers (Sadd *et al.*, 2015). A further feature, which distinguishes bumblebees from advanced eusocial species, is the greater ability of some workers to successfully lay unfertilised eggs to produce males. Bumblebees do, however, possess characteristics which place them above the primitively eusocial species. The female caste (queen or worker) is determined during development and remains permanent throughout adulthood unlike primitively eusocial species (Sumner *et al.*, 2006). Goulson (2010) does not fully accept the term 'primitively eusocial' for bumblebees as workers can communicate information on food sources and some species perform nest homeostasis. For these reasons I will use the term 'intermediately eusocial' to describe the level of sociality in bumblebees.

1.3.3 The *Bombus terrestris* castes

1.3.3.1 Workers

All three *B. terrestris* castes (males, workers and queens) are black with a golden-yellow stripe across the first thoracal segment and the second abdominal segment. The tip of the abdomen is white with more or less buff colouring towards the anterior edge (figure 1.2). *B. terrestris* workers are produced first in a colony cycle by the founding queen laying fertilised eggs. The workers are then responsible for rearing

offspring by tending to the brood, foraging for pollen and nectar, and generally maintaining the nest. Workers are on average the smallest of the three castes, although their size is more variable. Goulson *et al.* (2002) found the width of worker thoraces to vary between 2.3 and 6.9 mm and their total mass between 68 and 754 mg. In the current study adult workers also varied more than adult males and queens in terms of mass, with a coefficient of variance of 0.332 compared to 0.147 in queens and 0.217 in males (figure 2.1). The smaller workers tend to concentrate more on nest duties while larger workers forage more (Goulson *et al.*, 2002). However, there appears to be no clear division of labour among bumblebee workers, although workers differ in the relative amount of time they spend foraging and working within the nest (Foster *et al.*, 2004; Cameron, 1989). Bumblebee workers live for 2 to 6 weeks depending on species (Goulson, 2010) but little has been reported on the longevity of *B. terrestris* workers in the wild. Reported medians of 76 and 54 days for two independent replicates of groups of *B. terrestris* workers receiving sugar and pollen *ad libitum* probably exceed natural lifespans (Smeets and Duchateau, 2003). Towards the end of the colony cycle some workers will become more aggressive towards other workers and towards the queen. These workers often show greater ovarian development and begin to lay eggs (Foster *et al.*, 2004). The larvae that hatch from these eggs will develop into males as they are unfertilised (see next section for more details 1.3.3.2). Workers begin to lay eggs due to a reduction in the dominance of the queen, which otherwise suppresses worker reproduction (Röseler *et al.*, 1981).

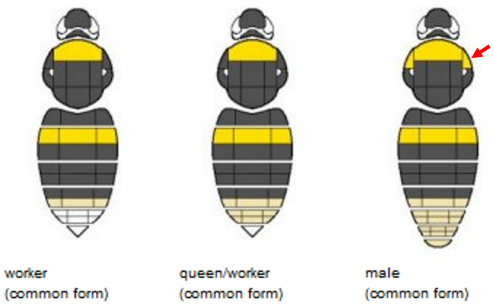


Figure 1.2: *Bombus terrestris* identification key.
Adapted from: http://www.nhm.ac.uk/research-curation/research/projects/bombus/_key_colour_british/ck_widespread.html

1.3.3.2 Males

After an initial colony growth period, the switch point is reached, after which the queen begins to lay unfertilised eggs (Duchateau and Velthuis, 1988; Goulson, 2010).

These haploid offspring develop into drones. In a large number of hymenopteran taxa, sex is determined by the genotype at a single locus called *complementary sex determination* (*csd*; Heimpel and Boer 2008). Individuals which are heterozygous at this locus are female, while homozygous or hemizygous individuals are male (Beye *et al.*, 2003). Diploid males are rare due to high allele diversity at the *csd* locus and reduced viablility of diploid males (Heimpel and Boer, 2008), so that the majority of males are haploid.

Males externally resemble the workers both in colour and size, apart from a slightly elongated abdomen and a broader thoracal stripe creating yellow 'shoulders' (figure 1.2, see arrow). Beside the lack of a sting, males can be distinguished by the number of antennal segments: males possess 13 compared to 12 on females (Prys-Jones and Corbet, 1987). Males are on average slightly larger than workers; Goulson *et al.* (2002) measured a mean thorax width of 5.63 ± 0.32 mm (standard error) compared to 4.50 ± 0.79 mm for workers. This was confirmed in the current study, in which adult males had a mean mass of 208.5 ± 45.2 mg and workers 170.3 ± 56.5 mg ($p = 4.8 \times 10^{-4}$; Mann Whitney U test; figure 2.1). Male bumblebees may live a few weeks; in one study laboratory-reared males lived an average of 30.5 days (Duchateau and Mariën, 1995). Again, the lifespan in the wild is probably lower. Males do not take on any colony tasks and leave after a few days to feed and search for a mate (Goulson, 2010). *B. terrestris* males find a mate by leaving scent-marks, which are patrolled regularly (Goulson, 2010) and may mate with several different queens (Röseler, 1973).

1.3.3.3 Queens

A diploid larva has the potential to develop into either a worker or a queen. However, new queens are only produced later in the colony cycle, around the same period, in which males are produced. It has been suggested that a dominant queen most likely suppresses the development of queens (Pereboom *et al.*, 2003) so that in the early stages of a colony all diploid larvae develop into workers. When the queen begins to lose her dominance towards the end of a colony cycle, diploid larvae have the potential of becoming queens. Queen-destined larvae are not fed differently to worker-destined larvae (Pereboom, 2000) further supporting the idea of pheromone control, or lack of control, by the queen. Caste is determined early in larval development (Pereboom *et al.*, 2005) leading to a longer development period for queen-destined larvae (Cnaani *et al.*, 1997).

Queens resemble workers but are roughly twice as large (Goulson, 2010; Duchateau and Velthuis, 1988) and the buff colouring of the tail is usually more visible (figure 1.2). Virgin queens leave the nest to forage but, although they may return,

they do not generally take on nest chores (Goulson, 2010). A *B. terrestris* queen mates only once (Schmid-Hempel and Schmid-Hempel, 2000) before entering a diapause over winter. The following spring, the queen will emerge to found a colony of her own.

1.3.3.4 Summary

The caste differences described in the previous section are summarised in table 1.1. As all three castes possess the same gene set and the sexes differ only in ploidy, it can be assumed that differences in gene expression play an important role in caste determination. This has been shown to be true for several hymenopteran species, which is discussed in the following section.

Table 1.1: Caste differences in *Bombus terrestris*

	Queen	Worker	Male
Morphology	largest in body size	identical to queen apart from size and development of ovaries	distinct: no sting, longer antennae, male sexual organs
Reproduction	responsible for producing males, workers and queens, singly mated	some produce males late in the colony cycle	reproduction is main task, multiple mates
Longevity	1 year	few weeks	few weeks
Life cycle	mating, hibernation, founding & maintaining colony	foraging & brood care	leaves nest soon after maturity to mate
Ploidy	diploid	diploid	haploid

1.3.4 Differential expression in eusocial Hymenoptera

1.3.4.1 Primitively eusocial Hymenoptera

In the primitively eusocial wasp, *Polistes*, female castes differ behaviourally and are determined during adulthood (Sumner *et al.*, 2006). A worker can, for instance, become a queen if the original queen dies (Strassmann *et al.*, 2004). Sumner *et al.* (2006) tested the plasticity of female castes in *Polistes canadensis* by comparing expression patterns between young females, workers and queens. Via suppression

subtractive hybridisation, cloning and northern blots, they found 32 differentially expressed genes. The majority of these genes (81%) were expressed in workers at intermediate levels between young females and queens. The intermediate status of the worker caste was further confirmed by expression patterns being significantly correlated between workers and queens, as well as between workers and young females. The expression patterns of young females and queens, which represent opposite ends of the caste differentiation spectrum, were not significantly correlated. Nine of the 32 caste-specific genes from the study had also been found in differential expression analyses on more advanced species *(A. mellifera*, *B. terrestris* and *Solenopsis invicta*, red imported fire ant). It was supposed that these genes may be important in the evolution of eusociality as they play a role in caste determination of different species, for which eusociality has arisen independently (Sumner *et al.*, 2006).

A further study investigated differences in gene expression between four behavioural groups of *P. metricus* (Toth *et al.*, 2010). These groups were foundresses, queens, workers and gynes. Queens were described as successful foundresses, which have raised offspring to adulthood, whereas foundresses were collected in spring after initiating a colony. Gynes are future queens produced at the end of a colony cycle. Via a microarray analysis a larger number of genes (around 3200) could be analysed than in the Sumner *et al.* (2006) study, of which 389 (12%) were differentially expressed between behavioural groups. Worker and queen expression patterns were most similar and the expression patterns of gynes were most distinct. There was a significant overlap of foraging-related genes with genes also related to foraging in *A. mellifera*, however, not for genes involved in reproduction (Toth *et al.*, 2010).

1.3.4.2 Highly eusocial Hymenoptera

The most work on caste differentiation has been performed on highly eusocial hymenopterans. Grozinger *et al.* (2007) researched the expression patterns which control morphology, physiology and behaviour in female castes of the honeybee, *A. mellifera*. Even though honeybee workers do not lay eggs due to the influence of the queen and policing of other workers, some do develop ovaries. These reproductive workers were found to become more 'queen-like' in their expression patterns, although they still most closely resembled non-reproductive workers (Grozinger *et al.*, 2007). A similar pattern was found in the ant *Temnothorax longispinosus*, for which the gene expression patterns of three worker castes and queens were investigated (Feldmeyer *et al.*, 2014).

Not all differential expression analyses have concentrated on the adult female castes. In the social wasp *Vespula squamosa* castes, sexes and developmental stages were compared (Hoffman and Goodisman, 2007). Expression patterns were most

distinct between developmental stages rather than caste or sex. This pattern was also found in the two fire ant species *Solenopsis invicta* and *S. richteri*, for which expression patterns were conserved between species (Ometto *et al.*, 2011).

1.3.4.3 Bumblebees

Via suppression subtractive hybridisation Pereboom *et al.* (2005) identified 12 genes whose expression levels differed in comparisons between workers and queens during larval stages. Nine of the genes were upregulated in queen-destined larvae during the first instar, while all 12 were upregulated in workers in the fourth instar. Eight of these genes were differentially expressed between adult female castes (Pereboom *et al.*, 2005). These results highlight the importance of the timing of gene expression in caste determination. In a further study, which presented the first transcriptome of *B. terrestris*, expression patterns were compared between adult castes, sexes and developmental stages (Colgan *et al.*, 2011). A high number of transcripts (2,185) differed between the following samples: one larva, one pupa, one adult male, one gyne and two workers. Due to a lack of replication the results of an R-STAT analysis (Stekel *et al.*, 2000) on differential expression were considered preliminary. Within workers primarily genes involved in metabolism, chitin binding and defence had expression patterns indicative of differential expression. In the gyne several transcripts associated with storage proteins such as hexamerins were over-represented. Especially genes involved in immunity and metabolism were over-expressed in the male (Colgan *et al.*, 2011).

1.3.5 Project aims

My main motivation for this investigation was to determine genes which are differentially expressed between haploids and diploids, in order to test the effect of ploidy-specific expression on selection. However, the generated data allowed me to address further questions on caste determination in bumblebees. Investigations into the expression patterns involved in caste determination exist for primitively eusocial and highly eusocial Hymenoptera. However, little is known about the specific expression patterns governing caste determination in bumblebees, which can be classed as intermediately eusocial. The two studies presented in the previous section confirm that a connection exists between gene expression and caste determination in *B. terrestris*. However, the first study identified very few genes (Pereboom *et al.*, 2005), while the second study could only offer preliminary results due to a lack of replication (Colgan *et al.*, 2011).

In this study I was interested in discovering expression patterns involved in de-

termining the distinct morphologies, physiologies and behaviours of the three *B. terrestris* castes (queens, workers and males; table 1.1), and how these patterns compare to highly and primitively eusocial species. Specifically, I aimed to detect genes involved in female reproduction by comparing expression patterns between reproductive workers, non-reproductive workers and queens. As bumblebee workers are more plastic in their behaviour than highly eusocial Hymenoptera (Cameron, 1989) it is possible that they become even more queen-like in their expression patterns when reproductive compared to honeybees and ants. Furthermore, I was interested in elucidating the genes and expression patterns involved in controlling adult behaviour and physiology compared to those associated with morphological development by comparing males and workers at different developmental stages. When is gene expression most distinct between males and workers: During larval development to control the generation of gonads and imaginal discs, during the development of adult morphology in the pupal stage or in adulthood to coordinate distinct behaviours and physiology? To address these questions I aimed to rear queens, workers and males at all developmental stages (larvae, pupae and adults) in at least three independent colonies. Adult workers were further classified by their reproductive status. I then performed an RNAseq analysis to determine differential expression between different combinations of developmental stage, caste and sex.

Chapter 2

Caste determination in *Bombus terrestris*

The results of this study have been published in Molecular Ecology:

Harrison, M.C., Hammond, R.H., and Mallon, E.B. 20015. Reproductive workers show queenlike gene expression in an intermediately eusocial insect, the buff-tailed bumble bee *Bombus terrestris*. *Molecular Ecology*, 24(12): 3043-3063.

2.1 Introduction

Eusociality, the division of labour for adult females into reproductive queens and mainly sterile workers that care for the brood, has evolved multiple times independently within the Hymenoptera (Andersson, 1984). The level of sociality varies within the Hymenoptera, ranging from non-social solitary species through primitively eusocial to highly eusocial taxa. Among highly eusocial hymenopterans, beside the clear division of reproduction between morphologically distinct workers and queens, further worker sub-castes exist (Wilson, 1978; Cameron, 1989). Members of the sub-castes may be responsible for, among others things, brood care, foraging or nest-guarding. In some ant groups worker sub-castes are morphologically distinct and display, at the least, a clear size polymorphism (Buckingham, 1911; Detrain and Pasteels, 1992). In other highly eusocial taxa worker sub-castes are monomorphic and task specialisation is determined by age (Cameron, 1989). In primitively eusocial taxa, such as the paper wasp *Polistes*, female adult castes are behaviourally distinct but monomorphic and behaviourally plastic, meaning an adult worker can potentially become the dominant, reproducing queen at any time by replacing the current queen or founding a new colony (Sumner *et al.*, 2006; Reeve *et al.*, 2000).

These distinct morphological and behavioural castes, which exist among adult females of a eusocial colony, are based on alternative expression of the same genome. The plasticity of the behavioural castes in the primitively eusocial paper wasp, *Polistes canadensis*, was demonstrated by the existence of overlapping gene expression patterns along a continuum from newly emerged females, through intermediate workers to the dominant queens (Sumner *et al.*, 2006). Most gene expression studies in this area have, however, concentrated on highly eusocial taxa. Large differences in gene expression have been recorded both between the morphologically distinct queens and workers (*Temnothorax longispinosus*: Feldmeyer *et al.*, 2014; *Vespula squamosa*: Hoffman and Goodisman, 2007; *Solenopsis invicta* & *S. richteri*: Ometto *et al.*, 2011; *Apis mellifera*: Grozinger *et al.*, 2007) and between monomorphic, behavioural worker sub-castes (*Temnothorax longispinosus*: Feldmeyer *et al.*, 2014). Under certain conditions workers of highly eusocial societies may develop their ovaries. The expression patterns of these reproductive workers become more 'queen-like' but they still remain more similar to non-reproductive workers than queens (Grozinger *et al.*, 2007; Feldmeyer *et al.*, 2014). Of the many genes found to be involved in caste differentiation *vitellogenin* has perhaps received most attention and has been shown to be differentially expressed among female castes of the honeybee and several ant species (Amdam *et al.*, 2003; Feldmeyer *et al.*, 2014; Corona *et al.*, 2013; Morandin *et al.*, 2014). Often in such studies a heavy focus has been placed on adult female castes; however, little work has been done to elucidate expression differences of males (but see Nipitwattanaphon *et al.*, 2014). The haploid males are both morphologically and behaviourally distinct from their sisters and mother, but, although they differ in their ploidy level, they otherwise share the same genes as other colony members and are therefore also alternative expressions of the same genome.

Bumblebees represent an interesting taxon in which to study the phenomenon of eusociality as they possess both highly eusocial characteristics and more primitive features. For instance, whether a female will become a queen or a worker is irreversibly determined during development, as is the case for highly eusocial taxa. However, although a clear size dimorphism exists between queens and workers, generally both female adult castes are morphologically similar as in primitively eusocial species. Workers take on distinct tasks within a colony but the division of labour is more plastic than is the case for higher eusocial bees and is generally not temporally fixed (Cameron, 1989). Furthermore, towards the end of the colony cycle the division of labour between workers and reproductive queens breaks down and queens and workers come into direct conflict over the parentage of males. At this stage some workers activate their ovaries and begin to lay eggs and in the process

become highly aggressive towards each other and also the queen (Alaux *et al.*, 2004; Bloch, 1999).

This far no broad-scale studies had been conducted which focus on the expression patterns involved in caste determination within bumblebees, although two previous studies did present some caste-specific genes (Pereboom *et al.*, 2005; Colgan *et al.*, 2011). Pereboom *et al.* (2005) investigated how and when females developed into queens or workers. They identified, using suppression subtractive hybridisation, 12 genes whose expression differed in the comparisons: (1) worker and queen 1st instar larvae; (2) worker and queen 4th instar larvae; (3) adult queens and workers; (4) reproductive and non-reproductive workers. Colgan *et al.* (2011), within their analysis of the bumblebee transcriptome, found a high number of transcripts (2,185) that differ in their expression between adult castes, sexes and developmental stages but considered their results as preliminary due to a lack of replication (1 larva, 1 pupa, 2 adult workers, 1 adult male and 1 virgin queen).

Here, using RNA-seq, I investigated genes involved in caste determination within the buff-tailed bumblebee, *Bombus terrestris*. I compared expression patterns of reproductive workers with those of non-reproductive workers and queens to isolate genes which are important for the acquisition of fertility as well as genes which may control behaviour differences compared to non-reproductive workers. Because of the flexible, plastic nature of bumblebee workers (Cameron, 1989), reproductive workers are capable of becoming more 'queen-like' not only in their fertility but also in their behaviour. I therefore tested the hypothesis that there is a greater similarity in gene expression patterns between queens and reproductive workers compared to those found in less plastic highly eusocial species.

Furthermore, I explored genes that control the specific behaviour and morphology of males. I investigated the question of when, during the ontogeny of a male bumblebee, is the difference in gene expression to workers the greatest? Is the male gene expression pattern more distinct during larval development when the gonads and imaginal discs are generated? Are more genes involved in the development of the adult morphology during the pupal phase? Or does indeed the development and control of distinct behaviours among adults require the most distinct gene expression pattern? To address these questions I compared gene expression patterns of males and workers both within larvae and pupae. In adults, I analysed differences in expression patterns between males, queens, reproductive workers and non-reproductive workers.

2.2 Materials & Methods

2.2.1 Colonies

Originally, six standard (1-6) and two small (7 & 8), colonies of the commercially available subspecies *B. terrestris audax* were obtained from Agralan Ltd. Upon arrival each standard colony contained a queen and approximately 100 adult workers, while the small colonies contained around 20 workers. All colonies were kept in wooden nest boxes with inner dimensions of 24 x 16 x 13.5 cm. The nest boxes were connected via a rubber tube to a smaller perspex box which served as a foraging and waste area. The bees were supplied with pollen (mixed polifloral pollen, www.naturallygreen.co.uk) and a sugar solution (BIOGLUC®, Biobest) *ab libitum* within the foraging box. An additional dish of pollen was placed within the wooden nesting box. All colonies were kept in identical conditions within the same room at 26°C and 60% humidity in constant darkness to imitate natural colony conditions.

The standard colonies were already too advanced for collecting samples of all three castes (worker, male and queen) at each developmental stage. In fact colony 2 had already reached the competition phase and the queen was killed by workers within the first 2-3 days. The decision was made to purchase four additional small colonies (9-12) in order to guarantee the collection of all necessary samples from at least three independent colonies. The new colonies consisted of a mother queen and 8 to 20 workers when they arrived. The new colonies were kept in identical conditions.

2.2.2 Sampling

I aimed to collect samples from 11 different combinations of caste and developmental stage, each from three independent colonies. Within larvae and pupae these were workers, males and queens, while in adults I intended to collect males, reproductive workers, non-reproductive workers, mother queens and virgin queens.

Sampling was carried out under red light conditions. I determined the sex of adults by counting antennal segments (males: 13; females: 12) and checking for the presence or absence of a sting (Prys-Jones and Corbet, 1987), while queens were identified via their superior mass (adult workers ranged from 26 to 325 mg, male adults from 127 to 347 mg and adult queens from 616 to 942 mg; figure 2.1). In order to identify reproductive adult workers, individual workers from each colony were anaesthetised by cooling to 3°C for approximately 10 minutes, and their abdomens were dissected to observe ovary development. In order to avoid loss of RNA,

dissections lasted only a few seconds and samples were immediately snap frozen in liquid nitrogen. Workers with developed ovaries were labelled 'reproductive'. I categorised the workers, in which ovaries were not visible, to be of 'undetermined reproductive status', because of the potential time-lag between the expression of reproductive genes and subsequent changes in ovary morphology.

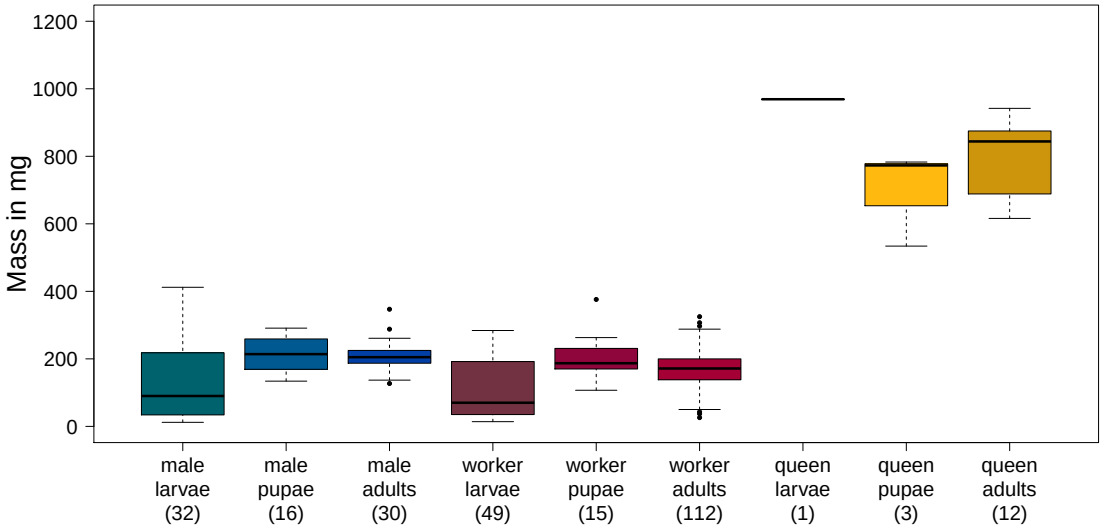


Figure 2.1: Masses of the *B. terrestris* castes (males, workers and queens) and developmental stages collected from colonies 7, 8, 9 & 11. Numbers in brackets represent sample size.

For the sampling of workers, queens and males during larval and pupal stages I followed the following protocol. I photographed the colonies at regular intervals of one to two days to monitor the emergence of new batches and their development. With the term 'batch' I refer to a single cohort of offspring laid together (figure 2.2). At intervals of at least three days larvae and pupae were sampled from each batch while ensuring at least half of each batch was allowed to develop to adulthood. I collected larvae from each of the four larva instar stages based on their weight according to Cnaani *et al.* (1997) and assuming male instar masses were similar to worker instars. Pupae were collected both shortly after pupation (pre-pupae) and later in pupal development when appendages were developed.

Figure 2.2: Inside colony 8. Centrally the queen can be seen (arrow) and several workers tending to at least four batches of larvae and pupae. The numbers show four batches of increasing development: 1-3 contain larvae (L2-L4) & 4 contains pupae.

Importantly, sex and caste of all sampled larvae and pupae were confirmed by isolating batches after pupation and sexing all emerging adults. Only if 100% of the unsampled adults emerging from a batch belonged to the same sex and caste would the samples from that batch be considered for analysis. All samples were collected between the hours of 9am and 5pm as soon as they became available. They were immediately weighed, snap-frozen in liquid nitrogen and then stored at -80°C.

Worker larvae and pupae were obtained from batches laid and reared in young colonies in the presence of the queen. After sufficient worker batches were available the mother queen was removed from each colony for sampling. All batches laid in the presence of the queen but hatched shortly before or after the removal of the queen were considered potential queen batches (Pereboom *et al.*, 2005). Any batches which were laid after queen removal were considered male batches. Additional male larvae and pupae were reared by isolating two to three groups of five workers from each colony in separate, small perspex boxes containing pollen, sugar water and cat litter due its hygienic yet absorbent nature. The majority of male larvae and pupae (mean 67.2% ± 8.4% SEM) and all male adults were sampled from the main colonies.

As samples of the first larval stage were not obtained for workers from three separate colonies, L1 samples were excluded from all libraries. Adult virgin queens were only obtained from one colony, and the batches from which they emerged also produced adult workers. Therefore, larvae and pupae were only confirmed as queens if (1) they were sampled from batches from which adult queens emerged, and (2) if they exceeded 500mg (no sampled male or worker larva, pupa or adult exceeded

420 mg; figure 2.1).

2.2.3 RNA extractions

Whole bodies were used for sampling for two reasons. First, I had no prior assumptions regarding the tissues within which genes would be differentially expressed between castes, sexes and developmental stages. Second, I wished to detect as many differentially expressed genes as possible across all comparisons. All samples were homogenised directly from -80°C. This was done within the Eppendorf tube with a plastic pestle for most larvae, and with a ceramic mortar and pestle for large larvae, all pupae, and all adults. The mortar was filled with liquid nitrogen to keep the samples frozen during homogenisation. This was not necessary for the homogenisations which took place in Eppendorf tubes as the process was completed quickly. Total RNA was extracted from all samples using a GenElute Mammalian Total RNA Miniprep kit (Sigma-Aldrich) following the manufacturer's protocol. The quality and concentration of RNA were estimated with an Agilent 2100 Bioanalyzer.

2.2.4 RNA pooling

A total of 27 RNA pooled samples were constructed that covered all 11 combinations of caste and developmental stage from one or three colonies (table 2.1). Based on the concentrations estimated with the Bioanalyzer the larval libraries were prepared so as to contain equal quantities of RNA from each of the three larval stages 2 - 4 and equal quantities per individual within each larval stage. The same was also true for pre-pupae and pupae within the pupal libraries.

Table 2.1: The 27 RNA libraries and the number of pooled individuals contained in each.

Caste	Developmental stage	Colonies			
		7	8	9	11
Worker	Larva (L2-L4)	18	13	18	
	Pupa	6	4	5	
	Reproductive adult	1	1	1	
	Undetermined reproductive adult		1	1	1
Queen	Larva (L4)	1			
	Pupa	3			
	Mother queen	1	1	1	
	Virgin queen	1			
Male	Larva (L2-L4)	16		9	7
	Pupa	4		9	3
	Adult	1		1	1

2.2.5 Sequencing & mapping

Via poly-A extraction, mRNA was isolated from the 27 RNA samples by the Edinburgh Genomics facility of the University of Edinburgh. The resulting cDNA libraries were sequenced on three lanes of an Illumina HiSeq 2500 system in rapid mode. After quality control and raw read processing, the reads were mapped to the *B. terrestris* transcriptome, BT_transcriptome_v2 (Colgan *et al.*, 2011), using bwa_0.6.1. The transcriptome was used for mapping since the genome had not been released at the time I conducted this study. Only reads which mapped uniquely were considered for further analysis. Counts per transcript were subsequently calculated for each library by Edinburgh Genomics using custom scripts.

2.2.6 Differential expression analysis

The Blast2GO java program (Conesa *et al.*, 2005) was used to annotate the transcriptome with gene descriptions and Gene Ontology (GO) terms (blastx against the nr database with $e < 0.001$). Differential expression analyses were carried out with the DESeq package (1.16.0; Anders and Huber, 2010) in R (3.1.1; Team, 2012).

A neighbour-joining tree was created based on expression differences between each of the 27 libraries. The distance matrix for the tree was calculated with the DESeq package and contained Euclidean distances between each library based on

variance stabilisation transformed counts. The tree was created with Phylip (3.695, Felsenstein, 2005). A principal components analysis was performed on all adult libraries within the DESeq package on variance stabilisation transformed data. Euler diagrams were created with the R package venneuler (Wilkinson and Urbanek, 2011).

In order to guarantee sufficient read depth and to increase confidence in base calls, transcripts with a total of 50 reads or fewer across all 27 libraries were removed before performing the differential expression (DE) analyses. All remaining transcripts were tested for differential expression in each comparison. For each DE analysis, standard comparisons were performed between two conditions on normalised count data and with dispersion accounted for. Only transcripts with a Benjamini-Hochberg corrected p value (FDR) < 0.05 were considered as significantly, differentially expressed. For comparisons between castes within developmental stages, colonies were considered as replicates. No comparisons were made against queen pupae, queen larvae or adult virgin queens, as in each case only one replicate existed. These libraries were, however, included in comparisons of expression between developmental stages.

Gene function enrichment analyses (Fisher exact test) were carried out on DE transcripts with the R package topGO (2.16.0; Alexa and Rahnenfuhrer, 2010). Enriched GO terms (FDR < 0.01) were subsequently summarised to meaningful clusters using Revigo (Supek *et al.*, 2011). This method reduces redundancy of GO terms.

2.3 Results

2.3.1 Transcriptome mapping

A total of 469.3 million 50 base pair, single-end reads were generated, ranging from 13.9 to 23.7 million reads per library. The reads mapped to the *Bombus terrestris* transcriptome at an average of 85.27% (75.25% to 92.47%) per library. The transcripts ranged in length from 101 to 26,110 bases (mean: 1,102; median: 721; fig. 2.3). Average read depth across the 27 libraries ranged from 0 to 47,420 (mean: 181; median: 16; fig. 2.3). All transcripts, to which a total of 50 or fewer reads (10,089, 27.8%) had been mapped across all libraries, were removed, leaving 26,265 (72.2%) transcripts for the differential expression analyses. The normalised counts per transcript ranged from 0 to 391,971 (median 34.41, mean 247.81) per library.

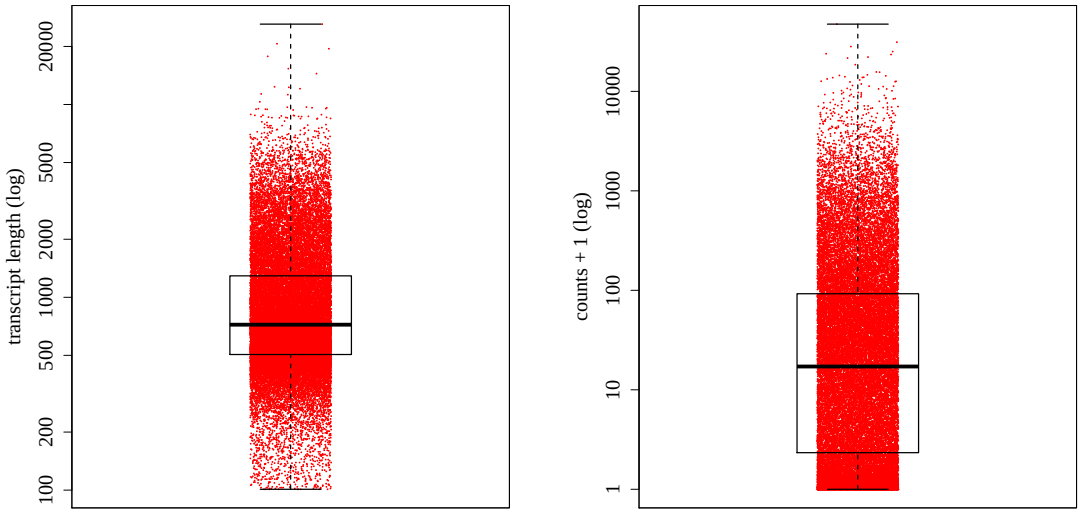


Figure 2.3: Range of transcript length in the transcriptome used for mapping (left) and the average read depth per transcript across 27 libraries as raw read counts plus 1 (right). A log scale is used in both plots.

2.3.2 Overview of gene expression patterns

All replicates, i.e. libraries from the same caste and developmental stage but from different colonies, showed low variation in their gene expression patterns and thus grouped together well in a neighbour-joining tree (fig. 2.4). The main clusters in the tree were formed by developmental stage (larvae, pupae and adults) rather than by caste. A differentiation in expression pattern between sexes becomes more apparent in adults, where males form a distinct cluster. Among female adult castes a further

clustering seems to have occurred. All reproductive workers and mother queens clustered together, and two of the workers with undetermined reproductive status ($W_{Au}8$ and $W_{Au}11$) formed a separate branch, while the adult virgin queen remained more distant to all other female adult groups. The adult worker with undetermined reproductive status from colony 9 ($W_{Au}9$), on the other hand, grouped together with reproductive workers and mother queens. A principal component analysis (PCA) performed on all adult libraries indicated that $W_{Au}9$ was indeed reproductive although ovaries had not been visible (fig. 2.5). In the analysis $W_{Au}9$ clusters strongly with all reproductive workers and mother queens. $W_{Au}8$ and $W_{Au}11$ form a distinct group, well separated from the reproductive workers and mother queens. This clustering pattern could mean that $W_{Au}9$ was reproductive and ovaries had not yet been visible in dissection. $W_{Au}8$ and $W_{Au}11$ were most likely non-reproductive.

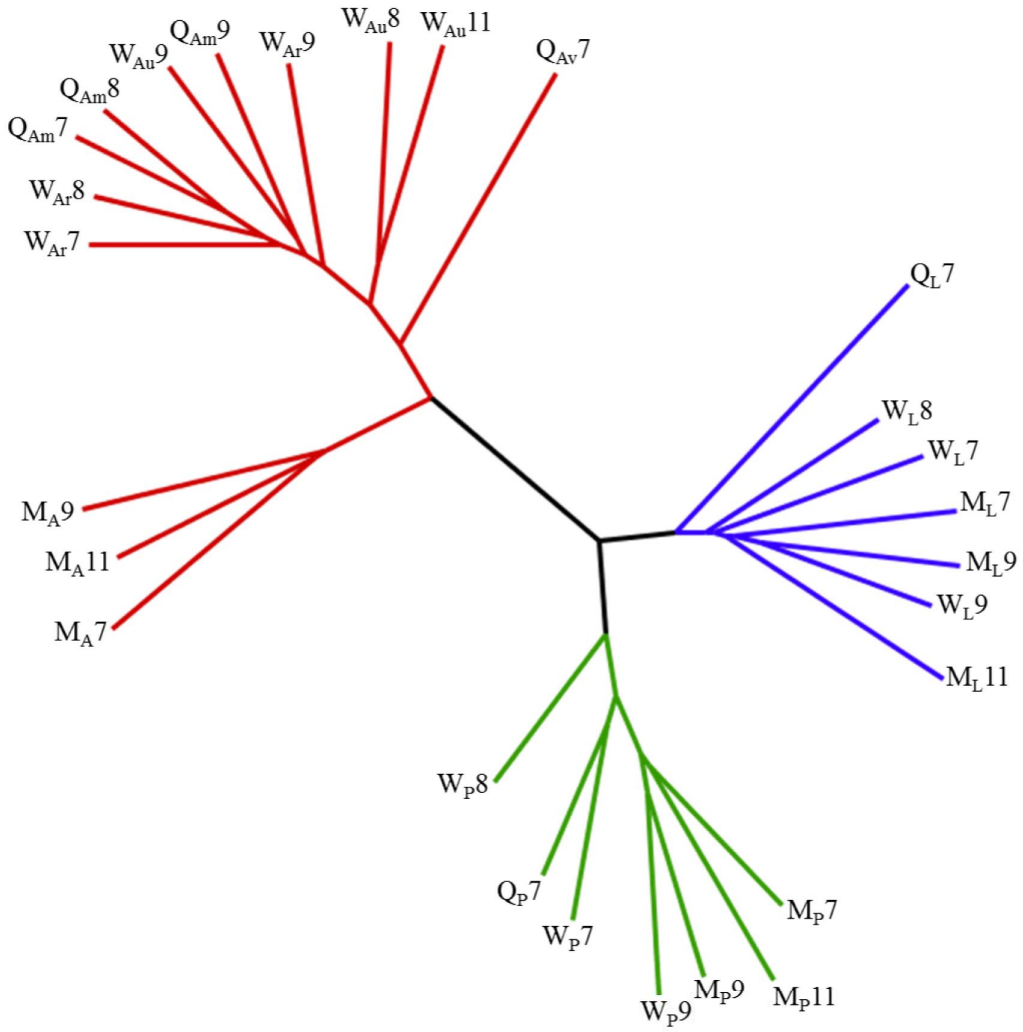


Figure 2.4: Neighbour-joining tree representing relationships between colonies, developmental stages, sexes and castes based on expression pattern. Distances are euclidean and based on variance stabilisation transformed counts. Numbers represent colonies; M = male; Q = queen; W = worker; $_L$ = larva; $_P$ = pupa; $_A$ = adult; $_r$ = reproductive; $_u$ = undetermined reproductive status; $_m$ = mother; $_v$ = virgin. Blue: larvae; green: pupae; red: adults.

These conclusions were further supported by an over-representation of *Apis mellifera* reproductive genes (Cardoen *et al.*, 2011) within $W_{Au}9$ but not in $W_{Au}8$ or $W_{Au}11$. The expression of the 299 genes, which were over-expressed in reproductive honeybee workers compared to non-reproductive workers, was lower in $W_{Au}8$ (median 122.8; mean 363.2) and $W_{Au}11$ (median 117.2; mean 389.4) than all three of the reproductive workers in this study ($W_{Ar}7$: median 194.0, mean 638.9; $W_{Ar}8$: median 212.3, mean 627.6; $W_{Ar}9$: median 138.9, mean 457.0). These differences were significant compared to $W_{Ar}7$ (compared to $W_{Au}8$: $p = 0.0045$; compared to $W_{Au}11$: $p = 0.0083$) and $W_{Ar}8$ (compared to $W_{Au}8$: $p = 0.0080$; compared to

$W_{Au}11$: p = 0.0018; Mann-Whitney U test; fig. 2.6). Expression of the Cardoen reproductive genes was significantly higher in $W_{Au}9$ (median: 196.5; mean: 542.5) than in both $W_{Au}8$ and $W_{Au}11$ (p = 0.0238 & 0.0376 respectively; Mann-Whitney U test; fig. 2.6). For these reasons $W_{Au}9$ was considered reproductive and $W_{Au}8$ and $W_{Au}11$ were classed as non-reproductive for all further analyses.

The patterns shown in the neighbour-joining tree (fig. 2.4) and PCA of adult castes (fig. 2.5) were reflected in the number of differentially expressed (DE) transcripts found between developmental stages and castes. From 6,289 to 7,483 (mean 7,019) transcripts were differentially expressed between developmental stages. Only 71 and 162 DE transcripts were found between males and workers within larvae and pupae respectively, while a mean of 4,114 DE transcripts were found within adult comparisons ranging from 111 between reproductive workers and mother queens to 8,706 between adult males and mother queens (fig. 2.7).

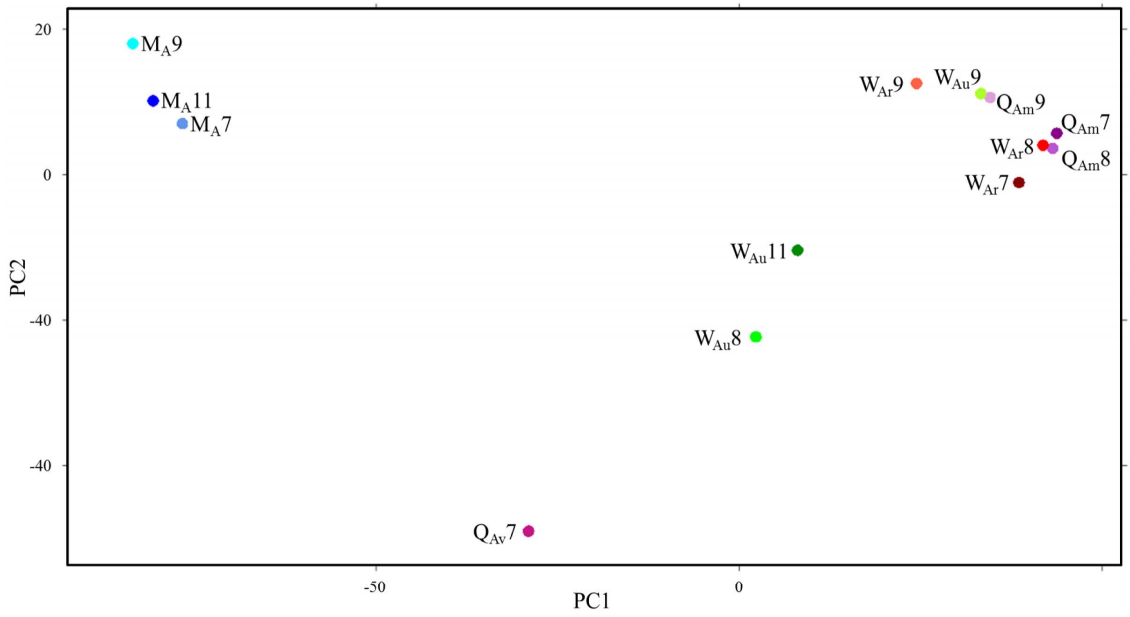


Figure 2.5: A principal components analysis of expression patterns among adult castes. The first two components explain 75.8% of variance. Distances are euclidean and based on variance stabilisation transformed counts. Numbers represent colonies; M = male; Q = queen; W = worker; $_{L}$ = larva; $_{P}$ = pupa; $_{A}$ = adult; $_{r}$ = reproductive; $_{u}$ = undetermined reproductive status; $_{m}$ = mother; $_{v}$ = virgin.

For some analyses of differential expression, colonies were not uniformly distributed, e.g. adult males (colonies 7, 9 & 11) versus non-reproductive workers (colonies 8 & 11). For these, ANCOVAs were performed to test for significant colony effects (table B.1; Appendix). In only one out of 9 cases (larvae vs. pupae) a significant

main effect of colony was found. But a separate significant effect also existed for the important group difference (developmental stage).

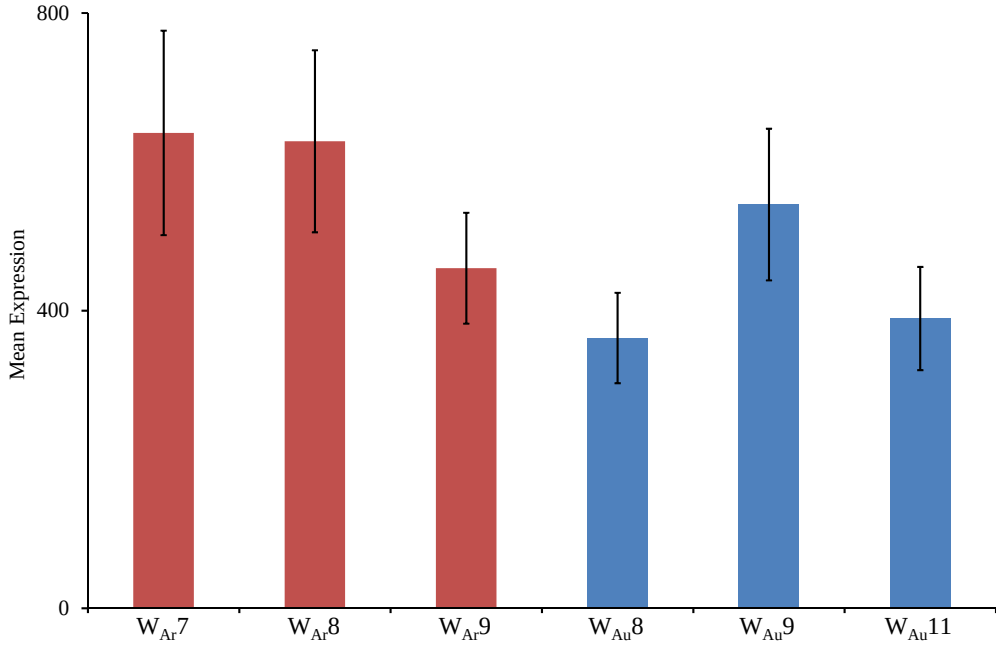


Figure 2.6: Mean expression level of honeybee reproductive genes identified in the study by Cardoen *et al.* (2011) in bumblebee workers of the present study. N = 299; error bars are standard error of the mean; W = worker; $_{A}$ = adult; $_{r}$ = reproductive; $_{u}$ = undetermined reproductive status; numbers = colony.

2.3.3 Developmental stages

A total of 12,218 DE transcripts were recorded in the three comparisons between larvae, pupae and adults (fig. 2.8). As already suggested by the neighbour-joining tree (fig. 2.4), adults differed most greatly from the other two developmental stages, confirmed by 3,237 transcripts which were differentially expressed compared to both pupae and larvae. A Gene Ontology (GO) term enrichment analysis showed that a heightened cellular metabolism distinguishes larvae from pupae and adults. The three main clusters of significantly over-represented GO terms (Fisher's exact test, FDR <0.01) in a Revigo treemap were "translation", "oxidative phosphorylation" and "ribosomal biogenesis" (fig. 2.9). Most over-represented GO terms among transcripts up-regulated in pupae either related to cell communication and movement, "signal transduction" and "cellular component organisation", or the development of morphological features, "anatomical structure morphogenesis" (fig. 2.10). Most

enriched adult GO terms belonged to the supercluster "G-protein coupled receptor signalling pathway" (fig. 2.11). This cluster included sub-clusters such as, "phototransduction", "detection of stimulus" and "cell surface receptor signalling pathway", highlighting the higher sensory capabilities and requirements of adults. 42.6% of larval, 58.7% of pupal and 48.3% of adult DE transcripts either received no significant blast hit or were linked to genes of unknown function. Detailed results of the Fisher tests can be found in the appendix (tables B.2, B.3 & B.4).

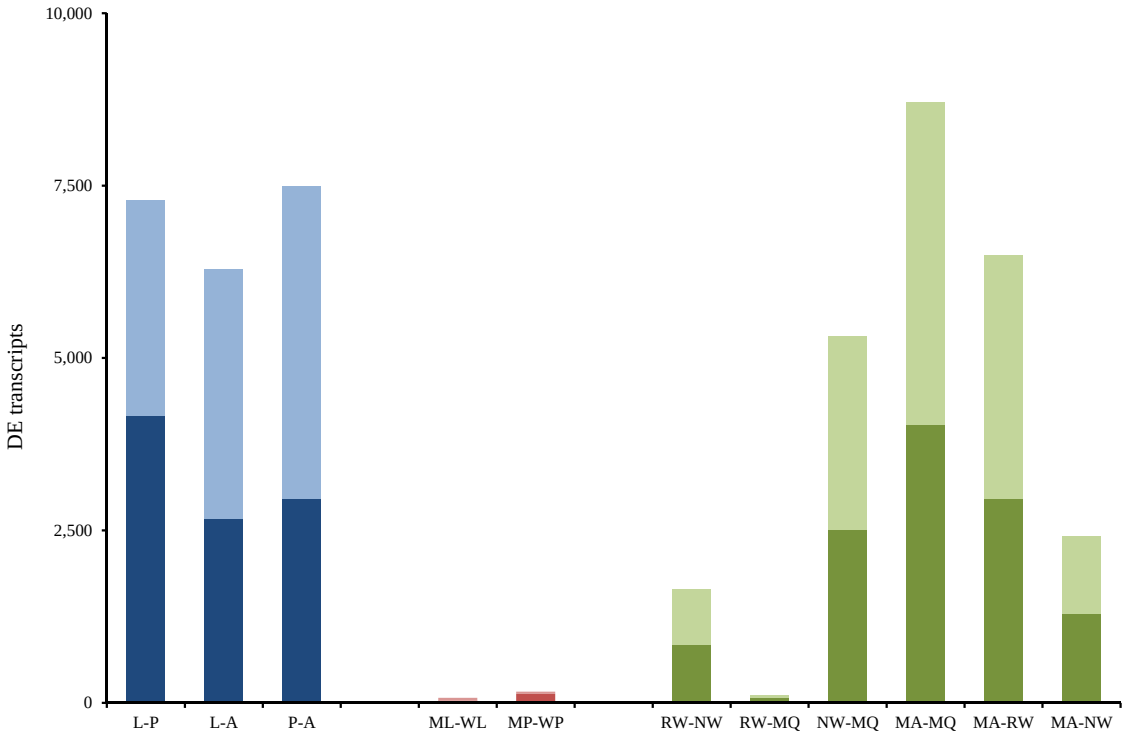


Figure 2.7: Differentially expressed transcripts within and between developmental stages. Darker colours: up-regulation in first named caste; lighter colours: up-regulated in second named caste. M = male; W = worker; MQ = mother queen; L = larva; P = pupa; A = adult; R/N = reproductive/non-reproductive.

This test was repeated using lists of unique genes rather than transcript lists. All tendencies and the largest GO clusters remained unchanged. However, the number of significantly enriched GO terms was reduced. This was most likely due to the reduced number of genes in the test as a consequence of a high number of transcripts without an annotated gene match.

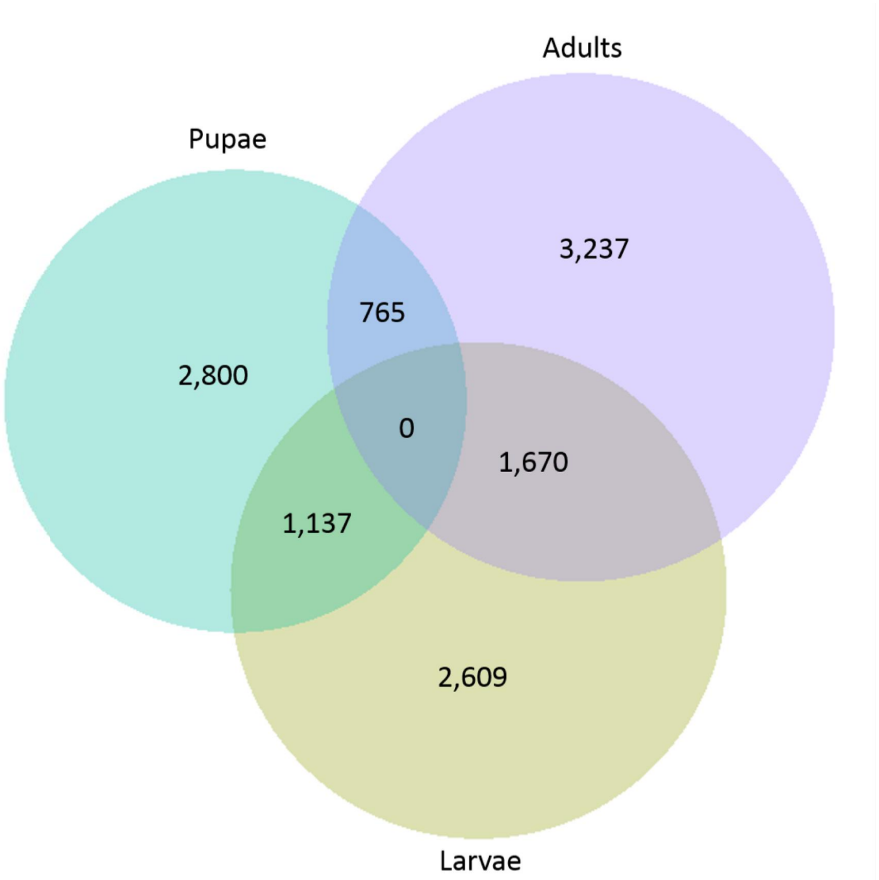


Figure 2.8: Number of transcripts which were differentially expressed between developmental stages.

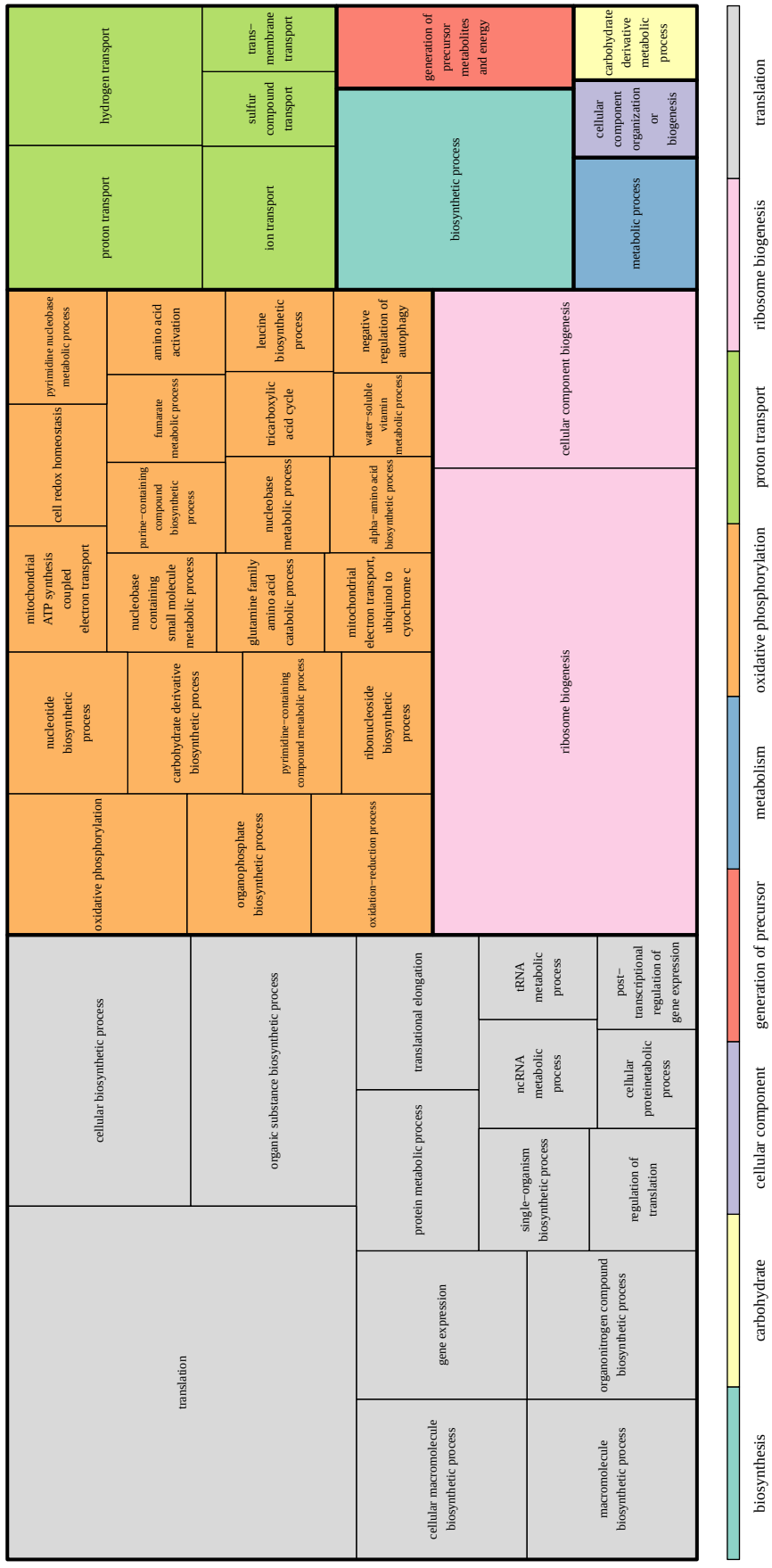


Figure 2.9: Most highly represented GO terms within larval DE genes (compared to pupae and adults).

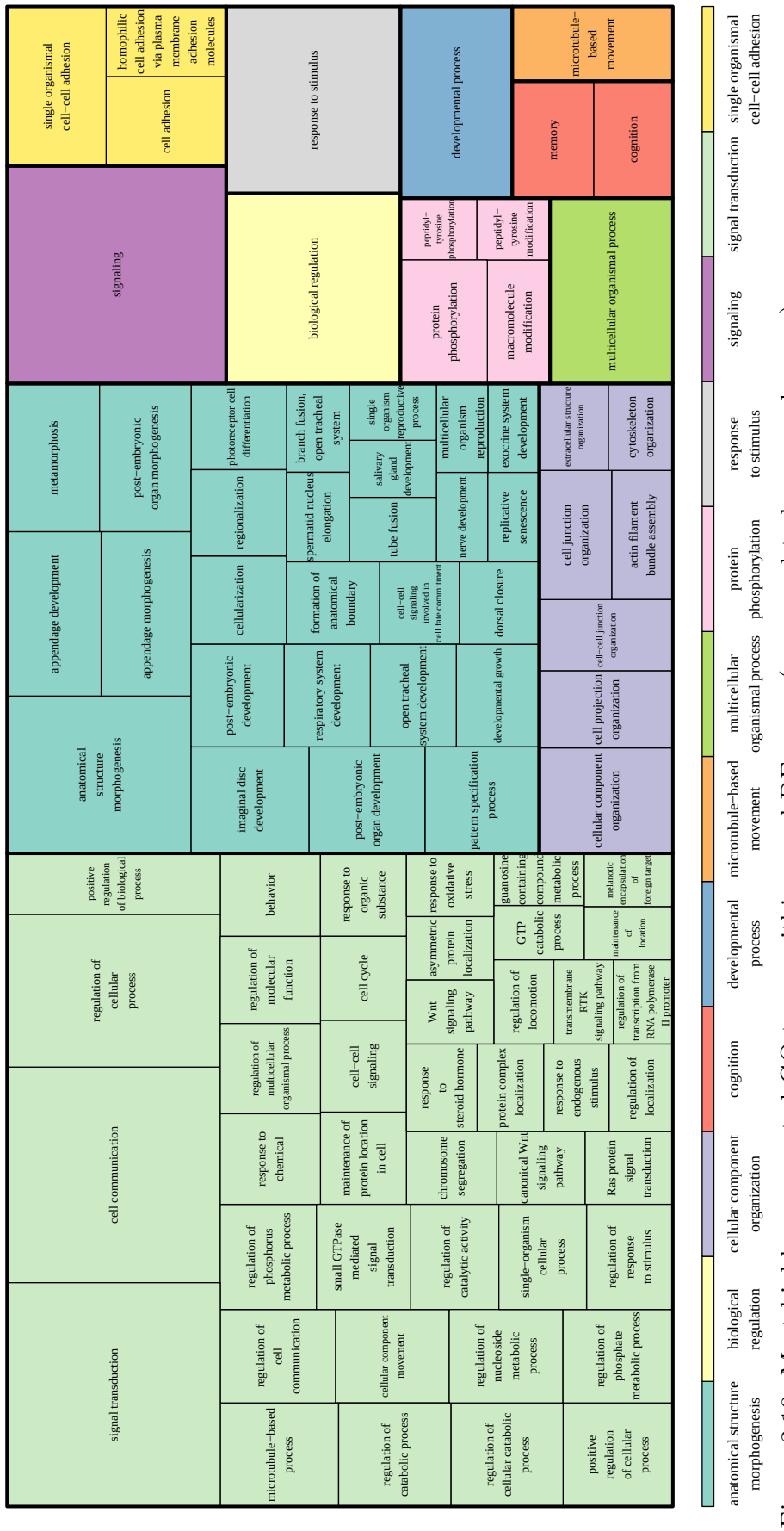


Figure 2.10: Most highly represented GO terms within pupal DE genes (compared to larvae and pupae).

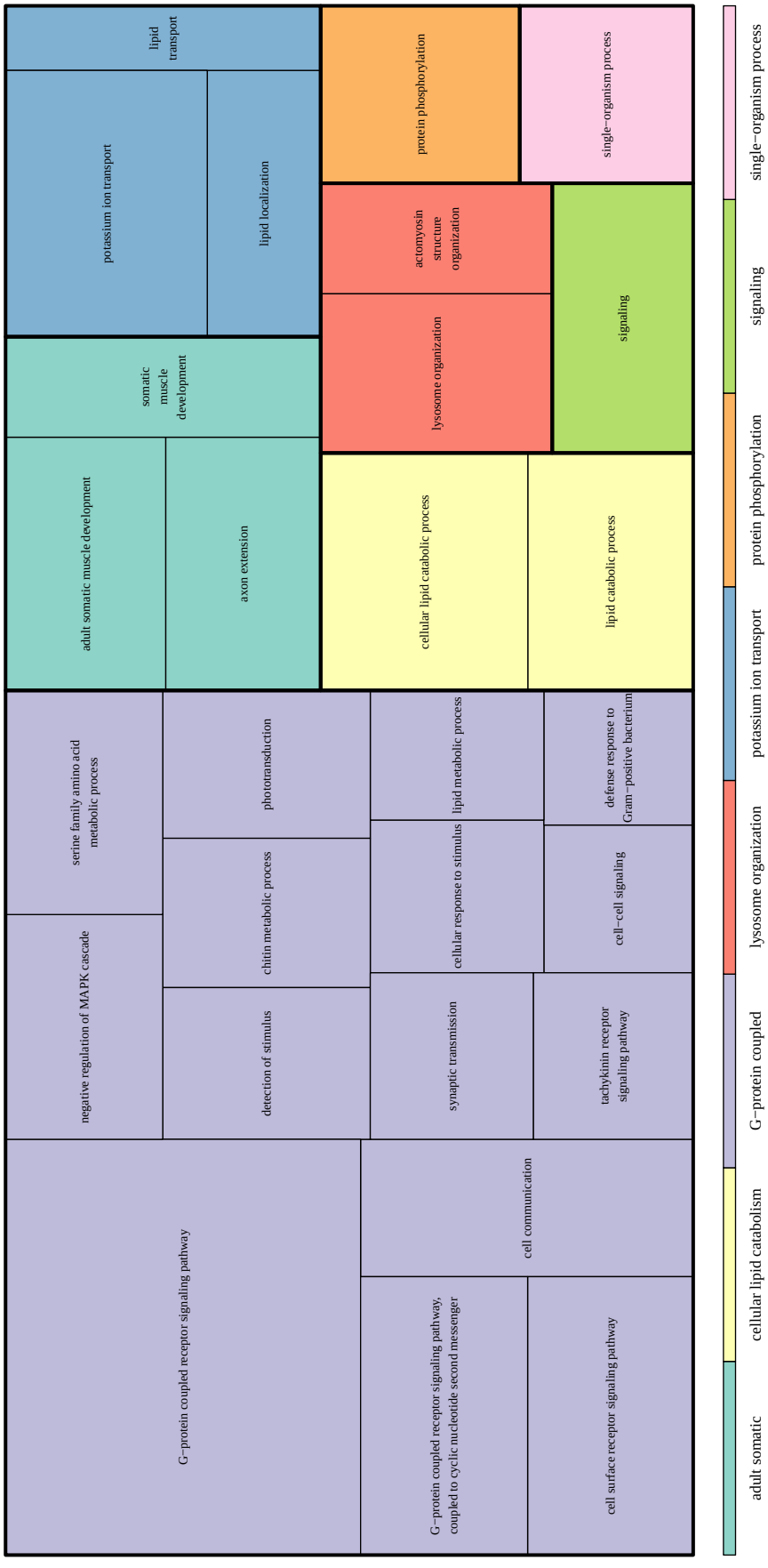


Figure 2.11: Most highly represented GO terms within adult DE genes (compared to larvae and pupae).

2.3.4 Male versus worker larvae

Within larvae only a relatively small group of transcripts proved to be differentially expressed between males and workers (32 and 39 up-regulated transcripts respectively). Within the list of male larvae DE transcripts *nose resistant to fluoxetine protein 6-like*, *nrf-6*, appeared six times with a fold change (FC) ranging from 3.86 - 25.34 and expression of 48 to 4,576 mean normalised counts (mnc; table B.5; Appendix). Nrf-6 is a transmembrane protein present in the intestine of various invertebrates (Choy and Thomas, 1999; Yao *et al.*, 2014) and has been reported as up-regulated in the gut of *Ostrinia nubilalis* larvae (Lepidoptera) in response to a bacterial toxin (Yao *et al.*, 2014). The presence of a further transcript within this list which encodes *cytochrome p450 6k1-like* (BTT39618_1; 2.07 FC; 1,682 mnc; table B.5; Appendix) provides possible further evidence for an infection within the male larvae. Riddell *et al.* (2014) found in *B. terrestris* that the expression of 16 different cytochrome p450 transcripts was altered post infection.

Takeout-like (XP_003397291.1; transcript BTT15842_1) was also strongly up-regulated in male larvae compared to worker larvae (5.75 FC; 2,693 mnc; table B.5; Appendix). A close homolog to this transcript (blastp: 68% identity, e-value $3e^{-126}$) has been characterized for *A. mellifera* (Hagai *et al.*, 2007). Takeout (to) was reported to be involved in the regulation of maturation in worker honeybees. In that study only adult workers were investigated so that any sex or developmental effects are as yet unknown for Hymenoptera. However, the *to* gene family is known to be over-expressed in adult *Drosophila* males, affecting courtship behaviour (Dauwalder *et al.*, 2002).

The majority of the worker larvae DE genes (24 out of 39; 8 of the top 10 in terms of FC) were either of unknown function or received no significant blast hits (table B.6; Appendix). One *vitellogenin* transcript (BTT24408_1; 4.35 FC; 62 mnc; table B.6; Appendix) was over-expressed in worker larvae compared to male larvae.

2.3.5 Male versus worker pupae

Differentiation was somewhat greater between males and workers during the pupal phase compared to the larval phase. 128 transcripts were significantly up-regulated in male pupae and 34 in worker pupae. The pupal list contained a high number of uncharacterised transcripts: 84 (66%) male and 24 (71%) worker pupae transcripts (tables B.7 & B.8; Appendix).

Six male DE transcripts coded for *tubulin* related genes (3 α-tubulin transcripts, 1 β-tubulin transcript and 2 tubulin-tyrosine ligases; 6.28 - 490.54 FC; 56 - 1,200

mnc; table B.7; Appendix). The tubulin-tyrosine ligase is involved in the post-transcriptional modification of α-tubulin (Ersfeld *et al.*, 1993), so it appears tubulin transcripts, especially α, may be important for male pupal development. The same *vitellogenin* transcript up-regulated in worker larvae (BTT24408_1) was also up-regulated in worker pupae compared to male pupae (5.83 FC; 29 mnc; tables B.8; Appendix).

2.3.6 Fertility genes

Within the comparisons between adult castes (males, reproductive workers, non-reproductive workers and mother queens), reproductive workers and mother queens were most similar with only 111 DE transcripts (64 up-regulated in reproductive workers and 47 in mother queens; fig. 2.7). Non-reproductive workers, on the other hand, were distinct from both mother queens (2,499 up-regulated in non-reproductive workers, 2,817 in mother queens) and, to a lesser extent, reproductive workers (844 up-regulated in reproductive, 810 in non-reproductive workers). The majority (791, 93.7%) of the transcripts up-regulated in reproductive workers compared to non-reproductive workers were also up-regulated in mother queens compared to non-reproductive workers. As the common difference between non-reproductive and reproductive workers and between non-reproductive workers and mother queens is their fertility status, I have named these 791 transcripts 'fertility genes' (fig. 2.12).

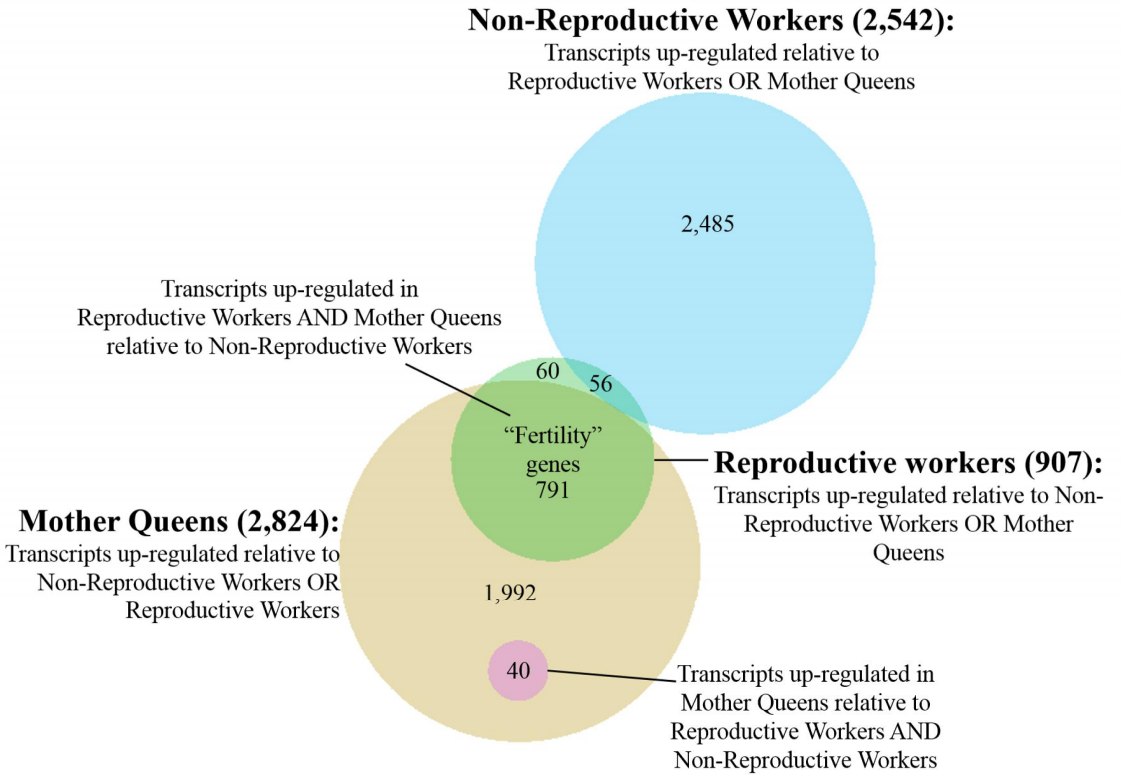


Figure 2.12: The number of transcripts which were differentially expressed between female adult castes. Overlapping areas indicate shared transcripts.

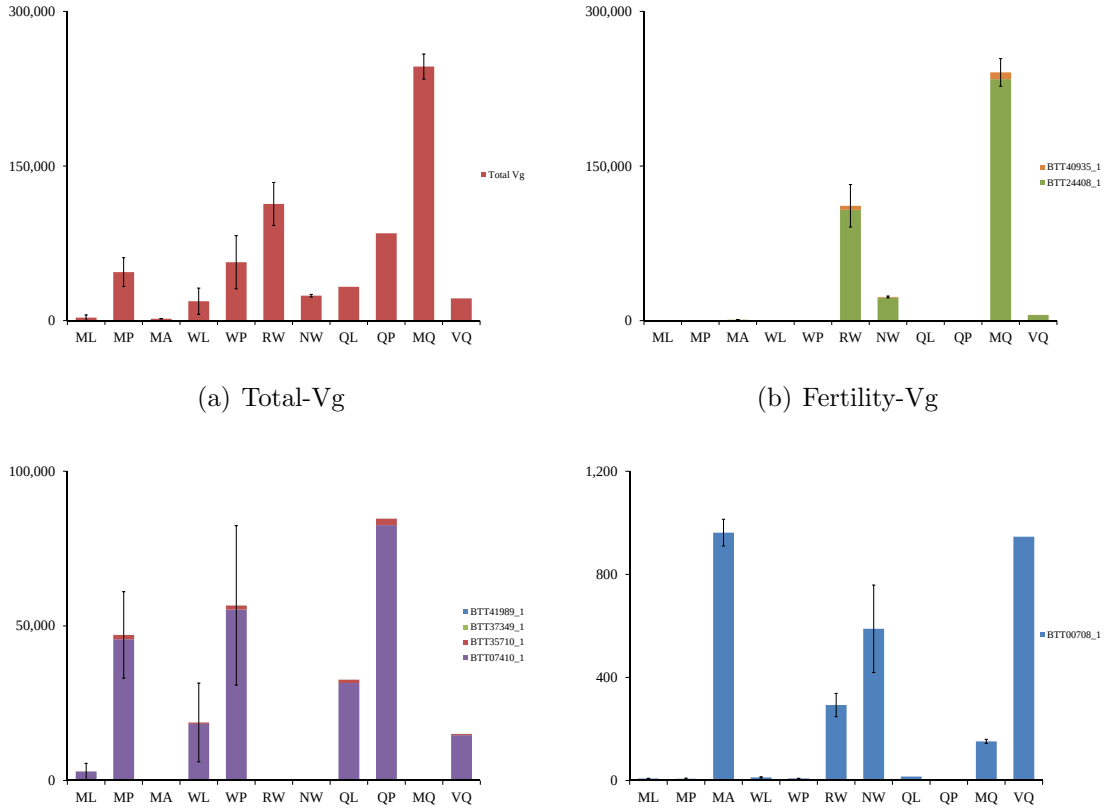
All differential expression values in this section are based on the comparison of reproductive and non-reproductive workers, although all transcripts were also up-regulated in mother queens versus non-reproductive workers. 267 (33.8%) of the fertility transcripts were of unknown function (1.78 - 336.18 FC; 18 - 39,927 mnc; table B.9; Appendix). A large number of transcripts were involved in protein synthesis activity: a total of 54 up-regulated gene transcripts contained the labels "transcription", "translation", "RNA polymerase", "ribosomal", "ubiquitin", "helicase" or "ribonucleoprotein" (1.85 - 11.68 FC; 68 - 29,132 mnc; table B.9; Appendix). Seven *tudor* transcripts, a gene known to be involved in the formation of female germ cells in *Drosophila* (Boswell and Mahowald, 1985), were significantly higher expressed in reproductive workers with a fold change ranging from 1.88 to 2.44 (144 - 491 mnc; table B.9; Appendix). Sixty-one of the fertility transcripts (1.77 - 9.53 FC; 12 - 9,973 mnc; table B.9; Appendix) were direct homologs of genes up-regulated in honeybee reproductive workers in a similar comparison (Cardoen *et al.*, 2011). These transcripts encoded genes with functions such as oocyte meiosis, oocyte axis specification, oogenesis and female gonad development (table B.9;

Appendix).

The list also contained two *vitellogenin* (4.95 & 6.03 FC; 4,103 & 111,595 mnc) and four vitellogenin receptor transcripts (1.94 - 3.39 FC; 222 - 11,577 mnc; table B.9; Appendix). The two *vitellogenin* transcripts had, on average across all libraries, a total expression level of 45,294 mnc, making up 69.4% of all *vitellogenin* transcripts on average per individual (97.7% in mother queens and 98.4% in reproductive workers; fig. 2.13). The *vitellogenin* transcripts (BTT24408_1 and BTT40935_1) are closely related to the 1,772 amino acid vitellogenin genes ACQ91623 and ACU00433 of *B. ignitus* and *B. hypocrita* respectively (table 2.2). These genes correspond to the conventional *Vg1* gene described by Morandin *et al.* (2014; blastp: E = 0.0, Id = 33%). The four receptor transcripts corresponded to the two *B. terrestris* genes *vitellogenin receptor-like isoform 1* and *isoform 2* (XP_003402703 and XP_003402704).

Table 2.2: Seven vitellogenin transcripts and the castes in which they are up-regulated.

Transcript	Caste specificity	Top blastx hit	e value	Protein length	Mean expression across 27 libraries
BTT24408_1	Reproductive female adults	ACQ91623 (*B. ignitus*) vitellogenin	0.0	1,772	43,917
BTT40935_1		ACU00433 (*B. hypocrita*) vitellogenin	5e-62		1,377
BTT07410_1	Larvae and pupae	XP_003400264 (*B. terrestris*) vitellogenin-6-like	0.0	1,514	18,374
BTT35710_1			1e-164		477
BTT37349_1			0.0		9
BTT41989_1			3e-29		10
BTT00708_1	Adult males and non-reproductive adult females	XP_003393940 (*B. terrestris*) vitellogenin-like	0.0	319	250



(a) Total-Vg

(b) Fertility-Vg

(c) Developmental-Vg

(d) Adult-Vg

Figure 2.13: Vitellogenin expression levels within different castes and developmental stages. Expression is mean of normalised counts across replicates; error bars are standard error of the mean. (a) the summed expression level of 18 vitellogenin transcripts; (b) two vitellogenin transcripts up-regulated in reproductive workers and mother queens versus non-reproductive workers;(c) four vitellogenin transcripts up-regulated in larvae and pupae versus adults; (c) one transcript up-regulated in all adults compared to larvae and pupae but down-regulated in reproductive adults. M=male, W=worker, Q=queen, L=larvae, P=pupae, A=adult, NR/R=non-/reproductive, M=mother, V=virgin; All castes include samples from three colonies, except RW: 4; NRW: 2; QL, QP & VQ: 1.

Vitellogenin was, however, not restricted to female reproductive castes. The second highest expressed *vitellogenin* transcript across all libraries, BTT07410_1, constituted on average 28.1% of vitellogenin transcripts. This transcript together with three further transcripts (BTT35710_1, BTT41989_1 and BTT37349_1) is associated with the *B. terrestris* gene XP_003400264 (*vitellogenin-6-like*), which is 1,514 amino acids in length and corresponds with the *Vg-like-A* gene described by Morandin *et al.* (2014; blastp: E = 0.0, Id = 44%; table 2.2). These four transcripts appear to be involved in development and independent of sex as they were

up-regulated in all larvae and pupae samples compared to adults irrespective of caste and sex (fig. 2.13(c)). One *vitellogenin* transcript (BTT00708_1) was significantly up-regulated in adults compared to pupae and larvae but was down-regulated by reproductive workers (significantly compared to male adults) and mother queens (significantly compared to non-reproductive workers and male adults; fig. 2.13(d)). This transcript is coded by the *B. terrestris* vitellogenin-like gene XP_003393940, which is much shorter than the two previously discussed *vitellogenin* genes (319 amino acids) and is similar to *Vg-2* of *Apis mellifera* (blastp: 66% identity, e-value $1e^{-142}$) and the *Vg-C-like* homolog described in Morandin *et al.* (2014; blastp: E = $1e^{-134}$, Id = 57%; table 2.2).

Seven *α-glucosidase* transcripts were differentially expressed within the fertility genes (6.93 - 9.14 FC; 16 - 35,260 mnc). An analysis of all ten *α-glucosidase* transcripts within the *B. terrestris* transcriptome across all libraries showed raised expression levels for reproductive workers and mother queens compared to all other castes and developmental stages. Non-reproductive workers had the third highest levels of the 11 combinations of caste and developmental stage but *α-glucosidase* transcripts were 8 times more abundant in reproductive workers and mother queens (fig. 2.14). Four *glucose dehydrogenase* transcripts (BTT01220_1, BTT08099_1, BTT18258_1 & BTT20465_1), on the other hand, were down-regulated in mother queens and reproductive workers, although up-regulated in all adults compared to larvae and pupae (fig. 2.15). These transcripts all related to the *B. terrestris glucose dehydrogenase* gene XP_003395668.1.

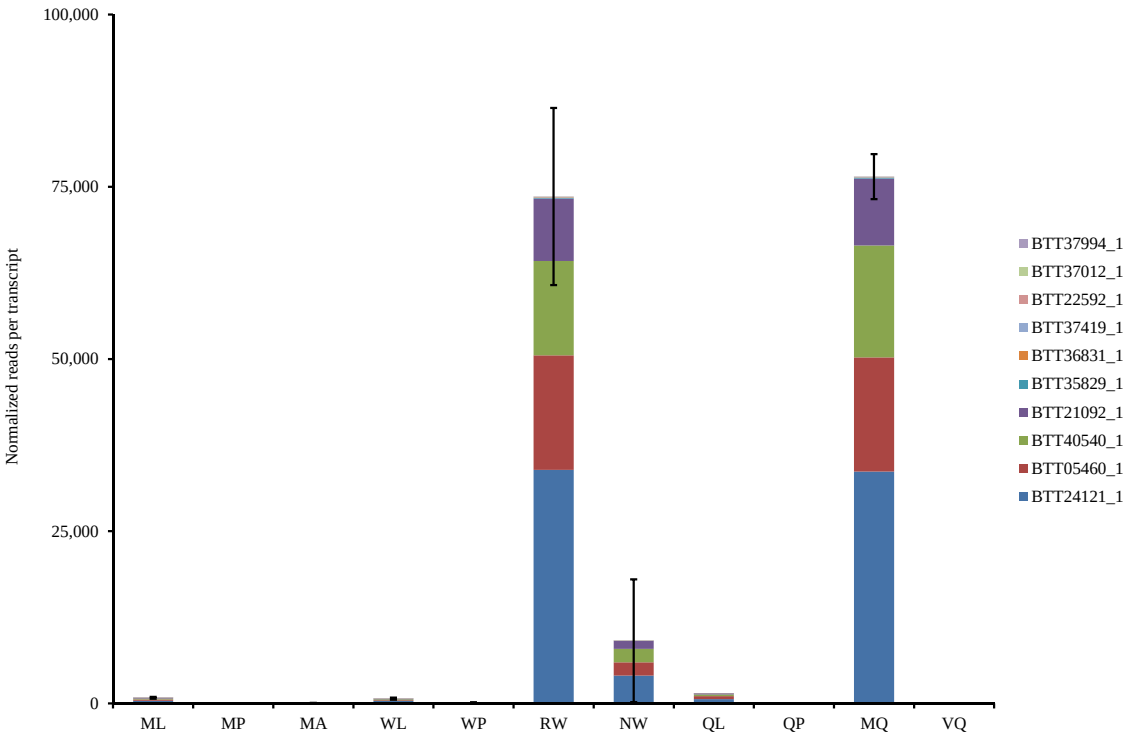


Figure 2.14: Expression levels of ten α-glucosidase transcripts within 11 different castes or developmental stages. Expression is mean of normalised counts across replicates; error bars are standard error of the mean. M=male, W=worker, Q=queen, L=larvae, P=pupae, A=adult, NR/R=non-/reproductive, M=mother, V=virgin; All castes include samples from 3 colonies, except RW: 4; NRW: 2; QL, QP & VQ: 1.

Interestingly, mean expression of the 10 *α-glucosidase* transcripts correlated significantly and positively with mean expression of the two *vitellogenin* transcripts (BTT24408_1 and BTT40935_1), which were also up-regulated in the fertility genes ($\rho = 0.7247$; p = 1.91 x 10^{-5}; Spearman's rho). Similarly, mean expression of the four *glucose-dehydrogenase* transcripts, down-regulated in fertility genes, significantly positively correlated with the down-regulated *vitellogenin* transcript (BTT00708_1; $\rho = 0.7888$; p = 1.02 x 10^{-5}; Spearman's rho).

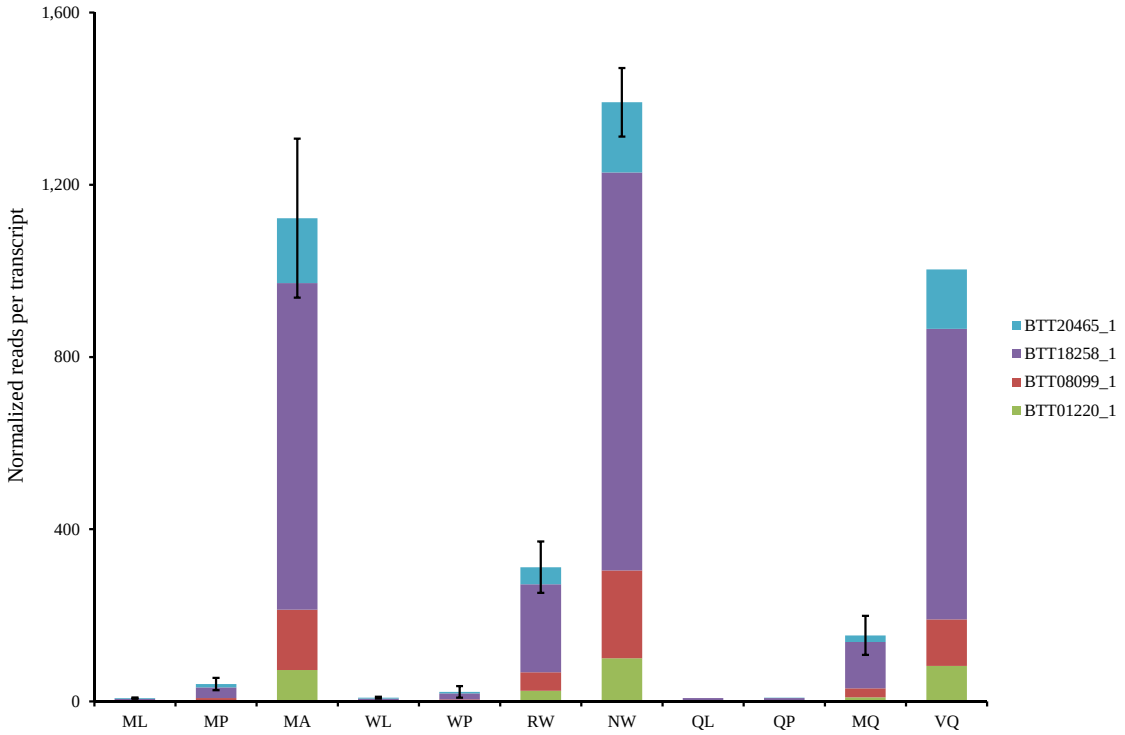


Figure 2.15: Expression levels of 4 glucose dehydrogenase transcripts within 11 different castes or developmental stages. Expression is mean of normalized counts across replicates; error bars are standard error of the mean. M=male, W=worker, Q=queen, L=larvae, P=pupae, A=adult, NR/R=non-/reproductive, M=mother, V=virgin; All castes include samples from 3 colonies, except RW: 4; NRW: 2; QL, QP & VQ: 1.

Two transcripts (BTT20241_1 & BTT33633_1; 67.16 & Inf FC; 5,528 & 37 mnc), which encode laccase-2-like, were up-regulated in reproductive versus non-reproductive workers but not in mother queens versus non-reproductive workers. Laccase 2 is a protein involved in the sclerotisation of extracellular structures in invertebrates (Arakane *et al.*, 2005).

2.3.7 Non-reproductive workers

For the majority (465 out of 810; 57.4%) of the transcripts up-regulated in non-reproductive workers compared to reproductive workers the function was unknown (table B.10; Appendix). 19 of the non-reproductive worker genes were direct homologs of genes up-regulated in non-reproductive *A. mellifera* workers (Cardoen *et al.*, 2011). Eight of those (1.77 - 3.74 FC; 72 - 653 mnc) had been attributed to the effect of the queen mandibular pheromone (QMP) in a previous study (Grozinger *et al.*, 2003).

2.3.8 Adult queens

Transcripts which were up-regulated in mother queens compared to both reproductive and non-reproductive workers were considered 'queen genes' (fig. 2.12). The 40 queen transcripts ranged in fold change compared to reproductive workers from 1.68 to 8.87 (29 - 245,472 mnc; table B.11; Appendix). Eleven of the transcripts (27.5%) were of unknown function. Most notable among the queen genes were five transcripts relating to serine protease inhibitors, SPI (2.92 - 8.87 FC; 2,145 - 10,596 mnc). These five SPIs were expressed together at a mean of 27,758 mnc ±1,247 SEM in mother queens compared to only 5,026 mnc in the virgin queen (fig. 2.16; table B.11; Appendix). The second highest levels were found in non-reproductive workers (7,555 mnc ±1,527 SEM) followed by reproductive workers (6,435 mnc ±699 SEM).

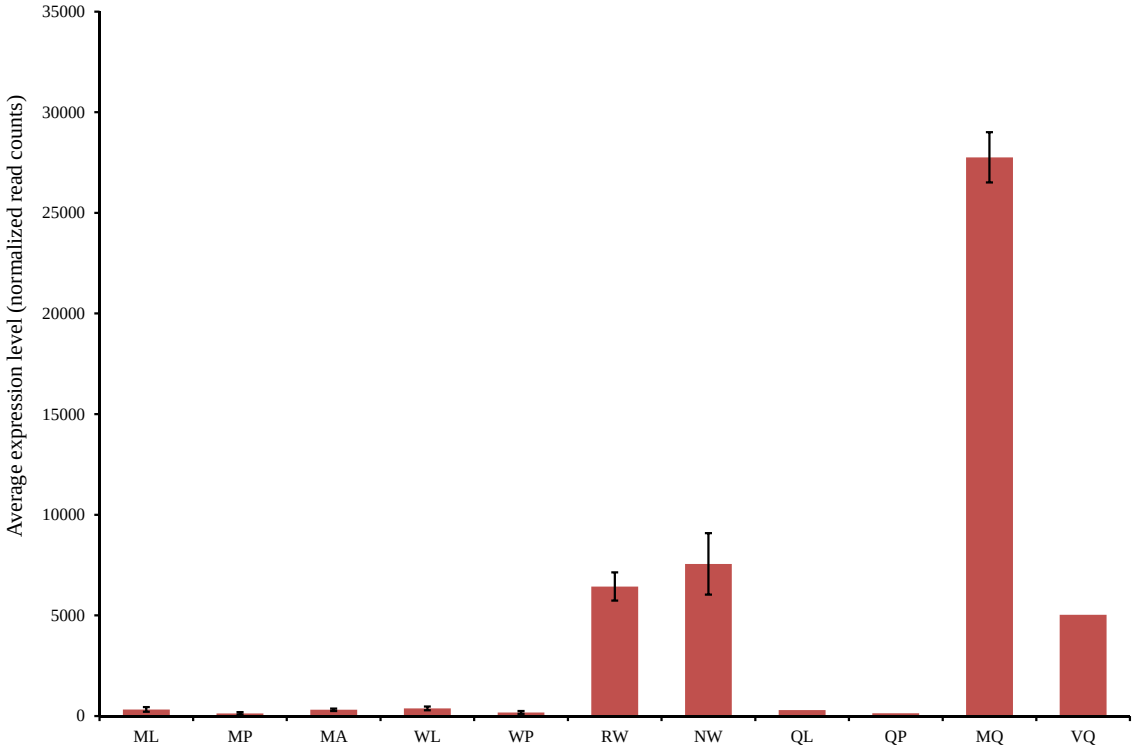


Figure 2.16: Expression levels of 6 serine protease inhibitors. Expression is mean of normalised counts across replicates; error bars are standard error of the mean. M=male, W=worker, Q=queen, L=larvae, P=pupae, A=adult, NR/R=non-/reproductive, M=mother, V=virgin; All castes include samples from three colonies, except RW: 4; NRW: 2; QL, QP & VQ: 1.

2.3.9 Adult males

In males compared to non-reproductive workers 1,280 transcripts were up-regulated, of which 526 (41.1%) were of unknown function (table B.12; Appendix). A high

number of male transcripts (190), containing the tags "mitochond", "cytochrome", "pyruvate", "NADH dehydrogenase" or "quinone", were involved in the mitochondrial metabolism (1.85 - 41.66 FC; 8 - 62,872 mnc; table B.12; Appendix). 37 transcripts were involved in muscle development (*myosin*, *troponin*, *twitchin* and *titin*; 2.42 - 28.60 FC; 10 - 5,877mnc) and a further 16 in the fatty acid metabolism (1.94 - 202.24 FC; 6 - 3,935 mnc; table B.12; Appendix).

2.3.10 Comparison with previous studies on *Bombus terrestris*

The top ten transcripts up-regulated in larvae in the study carried out by Colgan *et al.* (2011) related to cuticle proteins, the storage protein hexamerin and the metabolic proteins carbonic anhydrase and cytochrome p450. In the present study five cuticle, two hexamerin (70c and 70b), ten carbonic anhydrase and twelve cytochrome p450 related transcripts were also up-regulated in larvae compared to pupae and adults. The ten transcripts listed in Colgan *et al.* (2011) could be linked to one GO term (GO:0042302: "structural constituent of cuticle"), which was also attributed to 17 of the larvae transcripts (up-regulated relative to pupae and adults). In a further study a cuticle protein and hexamerin were also present in larvae but absent in adults; pupae were not included in the analysis (Pereboom *et al.*, 2005). The *vitellogenin* transcript BTT07410_1, which I found to be up-regulated in larvae and pupae, was also over-expressed in pupae in the Colgan *et al.* study (2011), however, was not detected in larvae. All seven of the GO terms which were associated with the top ten pupal genes in the Colgan *et al.* study (2011) were also present in my list of up-regulated pupae transcripts ($p = 1.3 \times 10^{-4}$, hypergeometric test).

In workers Colgan *et al.* (2011) found over-expressed genes associated with flight, defence and metabolism (*cytochrome p450*, *lipase* and *α-glucosidase*). In the present study flight muscles were also over-represented in non-reproductive workers and the metabolism genes *cytochrome p450*, *lipase* and *α-glucosidase* were more highly expressed in workers than in males. Twenty of the 36 GO terms associated with the worker transcripts in the Colgan *et al.* (2011) study were also found in the transcripts up-regulated in non-reproductive workers relative to adult males in the current study ($p = 5.8 \times 10^{-9}$, hypergeometric test). The genes differentially expressed between adult female castes and sub-castes in the Pereboom *et al.* study (2005), *60-S ribosomal protein*, *chymotrypsin*, *cytochrome oxidase*, *peroxiredoxin*, *fatty acyl CoA-desaturase* and *ATP synthase beta subunit*, could not be confirmed with my data.

Colgan *et al.* (2011) found transcripts of the flight muscle gene *titin* to be over-

represented in male adults, as well as several immunity genes. Many flight muscle proteins were also up-regulated in my study, however, I could not confirm the over-representation of immunity genes among the transcripts with known function. All 17 GO terms present in the top ten male transcripts of the Colgan *et al.* study (2011) were also present in my list of male transcripts (up-regulated in adult males relative to non-reproductive workers; $p = 2.3 \times 10^{-14}$, hypergeometric test).

2.4 Discussion

I compared gene expression patterns both between developmental stages and between castes within each developmental stage for the buff-tailed bumblebee *Bombus terrestris*. The number of differentially expressed transcripts ranged from 71 between male and worker larvae to 8,706 between adult males and mother queens. I found gene expression patterns to differ more between developmental stages than between caste or sex. Genes up-regulated in larvae were associated with a high cellular metabolism, whereas in pupae over-expressed genes were associated with cell communication and the development of morphological features. Most of the over-represented GO terms in adults were related to the G-protein coupled receptor signaling pathway. G-proteins are cell-surface receptors, which respond to extracellular stimulants with an intracellular signal cascade (Dohlman, 2002; Strader *et al.*, 1994).

The number of genes differentially expressed became progressively larger through the three developmental stages as each caste became more distinct. These findings suggest that a comparatively low number of genes are required to create distinct morphological castes compared to the high number involved in distinct behaviours between adult castes. Sex grouped more strongly than caste. Similar findings have been presented for the social wasp *Vespula squamosa*, for which workers, queens and males clustered clearly into developmental stages (Hoffman and Goodisman, 2007). A study on the two fire ant species *Solenopsis invicta* and *S. richteri* also found expression patterns between developmental stages to differ more than between sex followed by caste and species (Ometto *et al.*, 2011).

My data confirmed, to some extent, previous findings for *B. terrestris* (Pereboom *et al.*, 2005; Colgan *et al.*, 2011). Several associations of gene functions with specific castes or developmental stages detected by Colgan *et al.* (2011) were also found in the present study. Discrepancies can be explained by, in contrast to my study, a lack of replication in the 2011 study or a difference in analysis structure; Colgan *et al.* (2011) implemented R-STAT (Stekel *et al.*, 2000) to calculate differential expression of a contig within all libraries, whereas I performed specific pairwise comparisons. Little overlap could be found with an older study on caste determination in *B. terrestris* (Pereboom *et al.*, 2005). However, due to the method implemented in that study, suppression subtractive hybridisation, only a few differentially expressed genes could be isolated, and also, due to a different focus, fewer comparisons were performed than in this study (Pereboom *et al.*, 2005).

2.4.1 Reproductive workers closely resemble queens

Towards the end of a bumblebee colony cycle a queen-worker conflict develops, in which reproductive workers compete with the mother queen for male parentage (Alaux *et al.*, 2004; Bloch, 1999). The expression patterns observed in this study support my hypothesis that when bumblebee workers become reproductive they would, in comparison to highly eusocial species, more strongly resemble queens in their behaviour and physiology due to the more plastic nature of worker castes in bumblebees. Of all adult expression patterns, those of reproductive workers and mother queens were most similar, in fact more similar than between reproductive and non-reproductive workers. Only 111 transcripts differed significantly between reproductive workers and mother queens compared to 1,654 between reproductive and non-reproductive workers. Non-reproductive workers differed from mother queens even more strongly (5,316 DE transcripts). These findings are in strong contrast to patterns found in two highly eusocial hymenopteran species. In *A. mellifera* over 2,000 genes differed significantly in both comparisons between queens and either reproductive or non-reproductive workers; the expression of only 221 genes differed significantly between the two worker castes (Grozinger *et al.*, 2007). Similarly, 2,785 genes were significantly up- or down-regulated between queens and reproductive workers in the myrmicine ant *Temnothorax longispinosus* compared to only 571 between reproductive and non-reproductive workers (Feldmeyer *et al.*, 2014).

Feldmeyer *et al.* (2014) suggested the high similarity between reproductive and non-reproductive workers in these two hymenopteran taxa indicates that a relatively low number of genes are required for ovary activation and egg laying compared to the high number involved in further physiological or behavioural differences which exist between queens and workers. Based on this assumption, my data indicate a greater similarity in behaviour and general physiology between bumblebee queens and reproductive workers than is the case for honeybees or myrmicine ants. The division of labour among bumblebee workers is not as clearly temporally or morphologically fixed as in the highly eusocial honeybees and ants, indicating the capability of individual bumblebee workers to flexibly adapt their current role (e.g. from forager to nurse) to changing conditions within a colony at any given time (Cameron, 1989). In honeybees a shift towards a more 'queen-like' expression pattern was recorded in reproductive workers (Grozinger *et al.*, 2007); but it is possible that the more flexible nature of the bumblebee worker roles in my study allowed a much stronger shift in behaviour and physiology, allowing the reproductive workers to more strongly resemble a queen.

2.4.2 Male expression patterns are most distinct among adults

Males, in contrast to both queens and all workers, do not possess a sting and their antennae contain an additional segment. Their sexual organs naturally also differ. It was therefore surprising that the expression of comparatively few transcripts significantly differed during development. Within the larval stage no clear clusters could be formed based on expression patterns, and only 71 transcripts differed significantly in their expression levels between males and workers. During the pupal stage, when morphological features are being generated, expression patterns became more distinct with 162 transcripts differentially expressed. However, it was only in adulthood that the expression pattern of males became truly distinct from the other castes. In male adults between 2,411 and 8,706 transcripts were either up- or down-regulated compared to the three adult female castes mother queen, reproductive worker and non-reproductive worker. This indicates that a much greater number of genes may be required to control behaviour and the physiology of reproduction than to develop morphologies.

A high number (69; 59.5%) of the transcripts up-regulated in male pupae, and therefore likely to contain some genes linked to the development of the male morphology, were of unknown function. The six α- and β-*tubulin* transcripts, which were over-represented in male pupae, are possibly linked to spermatogenesis as both α*2*- and β-*tubulin* are known to be testis specific in *Drosophila* (Theurkauf *et al.*, 1986; Kemphues *et al.*, 1979). 190 transcripts were involved in mitochondrial processes and a further 37 were associated with genes linked to muscle development. These 37 transcripts are related to the proteins myosin, troponin, twitchin and titin, which are all integral parts of insect muscles (Hooper and Thuma, 2005). In their mating flights males have been recorded as covering significantly larger distances than workers from the same colony (Kraus *et al.*, 2009). The apparent greater need for muscle development and higher energy levels in males compared to workers are possibly linked to their greater flight distances.

2.4.3 Vitellogenin

Vitellogenin was originally thought to be limited to reproductive egg-laying females due to its function as a yolk precursor in all oviparous animals, though it is now known to fulfil various functions in hymenopterans (Amdam *et al.*, 2003). The reproductive ground plan model proposed by Amdam *et al.* (2004) describes how pleiotropic associations of reproductive genes, above all vitellogenin, with genes that

control sensory perception, longevity and foraging behaviour have been utilised to control behaviour patterns in honeybee worker sub-castes.

Previously only one vitellogenin gene had been described for honeybees, which is differentially expressed in female castes (Amdam *et al.*, 2012). However, in a more recent study on *Formica* ants four vitellogenin homologs were found within the genome of all ant and bee species included in the study (Morandin *et al.*, 2014). These vitellogenin homologs were classed as conventional vitellogenin (*Vg-1*), *Vg-like-A*, *Vg-like-B* and *Vg-like-C*, which were expressed at different levels and differently between queens and workers. Four copies of *Vg-1* have been found in *Solenopsis invicta* (Wurm *et al.*, 2011) and *Temnothorax longispinosus* (Feldmeyer *et al.*, 2014) and two in *Pogonomyrmex barbatus* (Corona *et al.*, 2013). In each of these cases, the gene copies showed differential expression between adult female castes.

Here I have found only one copy of *Vg-1* and two further vitellogenin genes which are closely related to *Vg-like-A* and *Vg-like-C*. *Vg-1*, as in *Formica* adults (Morandin *et al.*, 2014), was the highest expressed of the three vitellogenin genes discovered in this study. I found *Vg-1* to be highly up-regulated in mother queens and reproductive workers compared to all other castes and developmental stages, which suggests it has maintained its conventional function in reproductive egg-laying females for *B. terrestris*. This also appeared to be the case for three out of seven *Formica* species, in which *Vg-1* was up-regulated in queens compared to workers (Morandin *et al.*, 2014). Workers were not grouped according to reproductive status in the Morandin *et al.* study (2014), which could explain the lack of significant differences between castes in more than three species. The comparison of expression between queens and workers for *Vg-A-like* differed among *Formica* species (up-regulated in queens for three and in workers for one species; Morandin *et al.* 2014). In *B. terrestris* the homolog of *Vg-A-like*, XP_003400264, appears to play a lesser role in adults, as it was up-regulated in larvae and pupae of both sexes compared to adults. Expression of *Vg-C-like* was significantly higher in workers than queens in all seven *Formica* species (Morandin *et al.*, 2014). In the current study the homolog of *Vg-C-like*, XP_003393940, was also down-regulated in mother queens but also in reproductive workers compared to higher levels in non-reproductive workers and adult males.

Here I have shown that three copies of vitellogenin genes are not only differentially expressed between adult females castes as shown for other hymenopteran taxa (Amdam *et al.*, 2004; Morandin *et al.*, 2014; Wurm *et al.*, 2011; Feldmeyer *et al.*, 2014; Corona *et al.*, 2013), but that they are differentially expressed across all adult castes and between developmental stages.

2.4.4 Carbohydrate processing enzymes

I found the expression of two carbohydrate processing enzymes to be differentially expressed among adult castes. Expression of *α-glucosidase* was almost exclusively restricted to female adults but with levels eight times higher in mother queens and reproductive workers than in non-reproductive workers. This is in contrast to honeybees for which *α-glucosidase* is down-regulated in reproductive compared to non-reproductive honeybee workers (Cardoen *et al.*, 2011). In honeybees α-glucosidase catalyses the splitting of the sucrose present in nectar in the production of honey (Kubota *et al.*, 2004; Ohashi *et al.*, 1999). The apparent restriction of this protein to reproductive workers and mother queens may indicate a different role for this protein in *B. terrestris* compared to honeybees. *Glucose dehydrogenase*, on the other hand, was present in all *B. terrestris* adults but was down-regulated in reproductive workers and mother queens. The similar protein glucose oxidase is specifically found in the hypopharyngeal gland of forager honeybees and converts the glucose of nectar to gluconic acid and hydrogen peroxide in honey production (Ohashi *et al.*, 1999). Glucose dehydrogenase may perform a similar function in *B. terrestris* as it also catalyses the oxidation of glucose to gluconic acid but without the by-product hydrogen peroxide (Bak, 1967). Expression of *α-glucosidase* significantly correlated positively with *Vg-1* while expression patterns of *glucose dehydrogenase* significantly correlated positively with *Vg-C-like*. These correlations indicate interactions between *vitellogenin* and the two carbohydrate enzymes, which may be associated with distinct foraging preferences among adult castes.

2.4.5 Further caste-specific genes

One highly represented gene in the list of transcripts over-expressed in mother queens compared to reproductive workers was serine protease inhibitor. Serine proteases have been detected in the venom of a variety of Hymenoptera species (Hoffman and Jacobson, 1996; Winningham *et al.*, 2004). One possibility is that serine protease inhibitor was produced to counteract the effect of stings, either as a reaction to sting attacks or as a preventative measure. This could be linked to the high aggression shown towards a bumblebee queen by workers late in a colony cycle often resulting in her death (Bourke and Ratnieks, 2001).

Workers can become reproductive in queenright (containing dominant queen) conditions, but whether workers or queens control worker reproduction is unresolved (Alaux *et al.*, 2007). Intriguingly I found eight transcripts up-regulated in non-reproductive individuals (BTT06229 1, BTT09963 1, BTT20486 1, BTT15870 1,

BTT22989 1, BTT27276 1, BTT17949 1 and BTT09790 1) whose expression is believed to be regulated by queen mandibular pheromone in *Apis mellifera* and where expression shows similar patterns (Grozinger *et al.*, 2003; Cardoen *et al.*, 2011). It is clear that further research is needed to understand the relationship between pheromonal signalling and ovary development (Amsalem *et al.*, 2009).

In each of the caste comparisons performed in this study large numbers of differentially expressed transcripts either could not be associated with any known gene or were related to genes with so far unknown function. These range from 1,636 to 2,609 (32.0% - 54.4%) up-regulated transcripts when comparing between developmental stages. The number of differentially expressed transcripts was much lower between male and worker larvae (34 & 39) and pupae (128 & 34), but still the majority of these transcripts (58.7%) were of unknown function. 267 of the 791 fertility transcripts, i.e. up-regulated in reproductive workers and mother queens compared to non-reproductive workers, belonged to uncharacterised genes, while 465 and 526 transcripts in the comparison between non-reproductive workers and adult males were of unknown function. Clearly, further research is required in these areas.

Chapter 3

The effects of ploidy-specific expression on selection in the buff-tailed bumblebee, *Bombus terrestris*

3.1 Introduction

The ploidy level at which a gene is expressed can affect the efficiency of selection acting on that gene. For recessive alleles, selection should be more efficient on a haploid expressed gene, since a mutation is always exposed to selection. On the other hand, recessive mutations within diploid expressed genes will be masked from selection by a dominant allele when heterozygous (Haldane, 1924; Vicoso and Charlesworth, 2006). In 1924 Haldane predicted differences in selection for sex-specific genes due to differences in ploidy level. He expected selection on rare, recessive alleles only to be effective when they are expressed in the heterogametic sex and believed this mode of selection to have played an important role in the evolution of hymenopterans, where the males are haploid and the females diploid (Haldane, 1924). A little over 20 years later, these expectations for haplodiploids were discussed in further detail by White (1945). White expected there to be no "reservoir of hidden variability" in haplodiploids, since all recessive mutations would be selected upon in males. For genes with expression limited to the diploid sex, however, mutations would be expected to persist in populations even if they were deleterious (White, 1945).

To the best of my knowledge these predictions have, so far, not specifically been

tested for haplodiploids (although see Hunt *et al.* (2011)). Most empirical studies, which aimed to test Haldane's predictions (1924), have centred around comparing levels of selection between genes on sex chromosomes and autosomal genes (Vicoso and Charlesworth, 2006; Meisel and Connallon, 2013). For this comparison of selection efficacy Charlesworth *et al.* (1987) predicted (i) a faster fixation of recessive, beneficial mutations and (ii) a slower fixation of slightly deleterious mutations via random drift on the X-chromosome compared to autosomes. This is because recessive genes on the X-chromosome are more exposed to selection due to their hemizygous expression in males (Charlesworth *et al.*, 1987). If most substitutions arise from the fixation of recessive, beneficial mutations then a faster evolution of X-chromosomal genes would be expected (Vicoso and Charlesworth, 2006). Indeed, a 'faster-X' effect has been detected both in *Drosophila* (Baines *et al.*, 2008; Counterman *et al.*, 2004; Begun *et al.*, 2007; Thornton and Long, 2002) and in mammals (Torgerson and Singh, 2003; Hvilsom *et al.*, 2012; Lu and Wu, 2005; Carneiro *et al.*, 2012), which has often been explained by increased positive selection on X-linked genes compared to autosomal genes. The effect appears to be stronger within male-biased genes (Baines *et al.*, 2008).

However, broadly varying estimates of the relative divergence rates of X-linked and autosomal genes leave room for debate regarding the extent of the effect (Meisel and Connallon, 2013). Also, isolating the faster-X effect is confounded by differences in mutation rates between sexes: males generally have higher mutation rates due to a high number of cell divisions in spermatogenesis (Ellegren, 2007). These differences in mutation rate contribute to X-chromosomes and autosomes differently as an X-chromosome spends 2/3 of its time in females and only 1/3 in males (Vicoso and Charlesworth, 2009; Mank *et al.*, 2010a). Furthermore, the effective population size (N_e) differs between the X-chromosome and the autosomes, which influences the contribution of genetic drift (Vicoso and Charlesworth, 2009; Mank *et al.*, 2010a). In birds, where females are the heterogametic sex, the discrepancy in N_e between the sex chromosome (Z) and autosomes is thought to be greater than in taxa with male heterogamety (Mank *et al.*, 2010a). This difference in N_e and the resulting increased effect of genetic drift seems to be the cause of a faster-Z effect in birds (Mank *et al.*, 2010b; Wright *et al.*, 2015).

Investigations into the effect of ploidy-specific expression on selection have not been entirely limited to the sex-linked genes of heterogametic species. In an experiment directly comparing haploid and diploid populations of the budding yeast, *Saccharomyces cerevisiae*, large haploid populations had significantly increased adaptation rates compared to diploids (Zeyl *et al.*, 2003). A faster evolution of genes

exclusively expressed in pollen, and therefore haploid, has also been observed in *Capsella grandiflora*, a member of the mustard family (Arunkumar *et al.*, 2013). Both stronger purifying and positive selection were detected for pollen genes, explained by a combination of sexual selection via pollen competition and haploid expression.

In this study I aimed to test the predictions made by MJD White (1945) over 70 years ago on the buff-tailed bumblebee, *Bombus terrestris*. I expected to find evidence for different levels of selection efficacy among genes which correspond to patterns of ploidy-specific expression. Specifically, I predicted greater purifying and positive selection to exist for haploid-biased genes (upregulated expression in males), in which recessive mutations will not be masked by a dominant allele. In contrast, genes with diploid-biased expression (upregulated in workers and queens) should experience weaker purging of recessive, deleterious mutations and less efficient fixation of recessive, adaptive mutations. The effect may be intermediate for non-biased genes (equally expressed in haploids and diploids). On the one hand, selection should be more efficient for non-biased genes than in diploid-biased genes due to their expression in haploid males. However, the effect has been known to be stronger for male-biased genes (Baines *et al.*, 2008), so that purifying and positive selection may be more efficient in the haploid-biased genes than in non-biased genes.

In a first analysis, to infer levels of purifying selection in the three gene groups (genes with haploid-biased, diploid-biased and non-biased expression), I estimated the distribution of fitness effects of new nonsynonymous mutations (Eyre-Walker and Keightley, 2009). Stronger purifying selection within haploid-biased and non-biased genes should be detectable in lower levels of effectively neutral and slightly deleterious mutations compared to diploid-biased genes. Using the same software, I also estimated the proportion of substitutions which have been fixed through positive selection (α). A higher value for α should be measurable in haploid and non-biased genes. I measured the effect of these selection patterns on the evolutionary rates of proteins (dN/dS) as well as intra-specific polymorphism levels. If positive selection dominates then higher dN/dS levels should be detectable for haploid-biased and non-biased genes. The opposite would be true if purifying selection dominates. In both cases, stronger selection (positive or purifying) should be visible in lower levels of polymorphism for haploid and non-biased genes.

3.2 Methods

3.2.1 Determination of expression bias

In order to identify ploidy-biased genes, I first aligned 24 of the 27 RNAseq libraries from chapter 2 (table 2.1), disregarding three non-replicated libraries (table 3.1), to the *Bombus terrestris* genome (version Bter_1.0; RefSeq assembly accession: GCF_000214255.1) using STAR version 2.4.2a (Dobin *et al.*, 2013) with default settings. With the resulting bam files I generated a counts matrix (number of reads per gene for each library) using the protocol outlined in the Bioconductor workflow entitled 'rnaseqGene' (http://www.bioconductor.org/help/workflows/rnaseqGene/; accessed Sep. 2015).

Table 3.1: The RNA libraries with which differential expression was calculated.

Caste	Developmental stage	Colonies			
		7	8	9	11
Worker	Larva (L2-L4)	18	13	18	
	Pupa	6	4	5	
	Adult	1	2	2	1
Mother queen	Adult	1	1	1	
Male	Larva (L2-L4)	16		9	7
	Pupa	4		9	3
	Adult	1		1	1

Numbers represent number of pooled individuals in each library.

To identify ploidy-bias I first compared expression levels between adult males and adult mother queens. I conducted a differential expression analysis on the counts matrix using DESeq2 (Love *et al.*, 2014). I classed genes that were significantly upregulated in males compared to queens (adjusted $p < 0.05$) with a fold change > 5 as haploid-biased. Genes significantly upregulated in queens by a fold change > 5 were considered diploid-biased. To ensure I was not including genes whose ploidy-bias was reversed in different castes or developmental stages, I carried out three further comparisons between males and workers at the different developmental stages (larvae, pupae and adults). All genes that were upregulated in one comparison but downregulated in another of the four comparisons for the same sex were excluded from all analyses. Any additional ploidy-biased genes which were found in this process were added to the two ploidy-biased gene lists. All genes which were non-significant

in all comparisons, were regarded as non-biased. All analyses were performed on these three gene groups - haploid-biased, diploid-biased and non-biased genes.

To test the validity of these methods for defining ploidy-bias, I compared the expression level of haploid- and diploid-biased genes in all male and female libraries (table 3.1). To do this I calculated the relative difference in mean expression between male and female libraries for each gene: (mean male expression - mean female expression) / (mean male expression + mean female expression). The mean relative difference in expression was 0.51 (median: 0.46) for haploid-biased genes and -0.65 (-0.67) for diploid-biased genes, which confirmed the analyses had reliably detected ploidy-biased genes (figure 3.1).

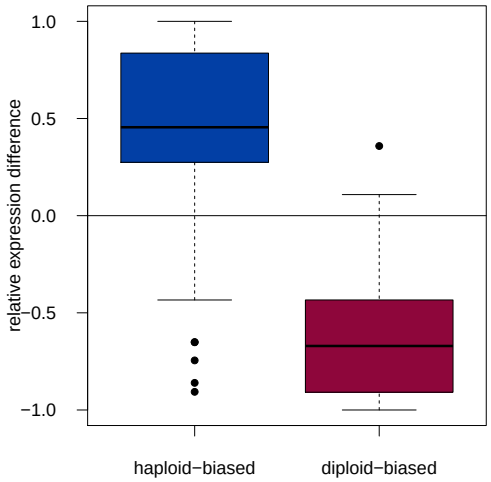


Figure 3.1: Relative expression difference between male and female libraries within haploid-biased and diploid-biased genes.
Relative expression difference: (male expression - female expression)/(male expression + female expression)

3.2.2 Generation of polymorphism data

3.2.2.1 Rationale for generating worker transcriptomes

Polymorphism data were required for coding sequences. For this I decided to produce worker transcriptomes. The main reason for choosing transcriptome data over genomic data was to reduce cost while increasing the depth of coding sequence data. Workers were collected because they are more abundant and more readily collected in the wild. One possible drawback of using worker transcriptomes was that haploid-biased genes would be, by definition, under-represented. However, the majority of the haploid-biased genes (97.2%) were also detected in the six worker libraries from

chapter 2 albeit at a lower level than in males. It therefore seemed reasonable to assume that the majority of the genes from each of the three gene groups would be reliably detectable in most worker transcriptomes.

3.2.2.2 Methodology

Wild *B. terrestris* workers were collected at 27 different locations in England and Wales (table B.13; Appendix). The bees were euthanised in liquid nitrogen then stored at -80°C. I extracted total RNA from whole bees with a GenElute Mammalian Total RNA Miniprep Kit (Sigma-Aldrich). I homogenised each bee from frozen within a small volume of liquid nitrogen using a mortar and pestle. For the total RNA extraction I followed the manufacturer's protocol although I required 2mL of lysis solution per sample, rather than the recommended 500μL. I also required three filter columns and four binding columns per sample, rather than the recommended one of each, to accommodate the larger volume of solution.

From the total RNA samples, mRNA was extracted via poly-A selection and cDNA libraries constructed at Eurofins Genomics. Single-end, 125-bp sequences were generated for the 27 polymorphism libraries on two lanes of an Illumina 2500 instrument. I performed quality control and raw read processing on the fastq files using FastQC. I generated variation data by following the GATK protocol entitled "Calling variants in RNAseq" (https://www.broadinstitute.org/gatk/). Within this method fastq files were aligned to the *B. terrestris* genome using the STAR 2-pass method (Dobin *et al.*, 2013). With this approach splice junctions are identified in a first alignment step, from which a new genome index is created. This new index then guides the final alignment. To further improve the first alignment I added the option of a gff file, from which splice junctions can also be identified. In the variant calling step (step 6) of the GATK protocol I added the -GVCF option. This way a genomic variant calling format file (gVCF) was generated for each of the samples rather than a VCF file. The gVCF file contained information on all base positions rather than just variable bases, which allowed the differentiation between a reference base (no polymorphism) and missing data when the files were merged. I merged the gVCFs with GATK's CombineGVCFs tool to produce a file which contained genotype information across the 27 individuals as well as information on variable bases such as the type of variant (i.e. SNP or indel), quality scores and read depths. I filtered this file according to step 7 of the GATK protocol, that is, variants with a Fisher strand bias > 30 and a quality by depth value $<$ 2 were removed. Quality by depth normalises the variant confidence quality score by read depth as deep coverage may inflate the quality score. I filtered the gVCF file to contain only coding sequences using the –bed and –recode options in vcftools

(version 0.1.14; Danecek *et al.* 2011). The –bed option filters the input file to only include variants within a set of coordinates contained in a bed-file, while –recode creates a new reduced VCF (or gVCF) file. I created the bed-file with a custom-made Perl script (script C.1; Appendix) which retrieved the coordinates of all coding sequences from the gff file.

3.2.2.3 Confirmation of suitability of worker transcriptomes

Within the raw bam files, which were produced in step 1 of the GATK protocol outlined above in section 3.2.2.2, the read depth varied between the three gene groups. A mean read depth of 1034.1 ± 45.1 (median: 298.7) per gene and per library was achieved in the full data set. Read depth was highest for the non-biased genes (mean: 1069.8 ± 67.6; median: 317.8) and intermediate for diploid-biased genes (mean: 924.1 ± 370.2; median: 53.4). The lower read depth for haploid-biased genes reconfirmed their lower, downregulated expression among workers. However, due to high overall coverage the mean read depth for haploid-biased genes was 639.6 ± 76.4 (median: 22.0). The overall high read depth allowed, after filtering, a high fraction of genes from each group to be recovered by the variant calling protocol (haploid-biased: 67.2%; diploid-biased: 73.0%; non-biased: 83.7%; see RESULTS 3.3.4).

3.2.3 Detecting levels of purifying selection and proportions of adaptive substitutions

In order to make inferences about levels of purifying selection I estimated the distribution of fitness effects of non-synonymous polymorphisms with the DoFE software (downloaded from http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/WebsiteSoftware.html in November 2013) for the three groups of genes (haploid-biased, diploid-biased and non-biased). The method described by Eyre-Walker and Keightley (2009) was implemented, for which the following data columns were required in the input files:

1. Gene
2. Number of chromosomes

Within intra-specific polymorphism data:

3. Number of nonsynonymous sites
4. Nonsynonymous allele frequency spectrum

5. Number of synonymous sites
6. Synonymous allele frequency spectrum

Within inter-specific divergence data:

7. Number of nonsynonymous sites
8. Number of nonsynonymous substitutions
9. Number of synonymous sites
10. Number of synonymous substitutions

As suggested by Gossmann *et al.* (2013), I randomly sampled twenty alleles at each site without replacement to speed up computing. In a first step, I identified sites with an allele sample size below 20 with the vcftools command –counts and retained genes with a minimum of 20 alleles at each site using a combination of two custom Perl scripts (C.2 & C.3; Appendix). I included all remaining ploidy-biased genes (204 haploid- and 194 diploid-biased genes) and randomly sampled 500 of the remaining 1629 non-biased genes without replacement using the sample function in R (version: 3.2.2; Team 2012), as the analysis failed to converge for a larger gene set. For each gene group, I achieved the allele sampling by shuffling (Perl module: List::Util::shuffle) all available alleles for a particular site and then keeping the first twenty elements of the Perl array (script C.4, Appendix).

I used 0-fold and 4-fold degenerate sites to represent nonsynonymous and synonymous sites. These were identified using a custom-made Perl script (C.5, Appendix). Using a further custom Perl script (C.6, Appendix) I counted the numbers of SNPs (polymorphism data) and substitutions (divergence data) at 4-fold and 0-fold sites. Any codons containing more than one polymorphism or substitution were removed and the total number of potential sites (4-fold and 0-fold) were recalculated. From the SNP data I calculated the allele frequency spectra with a custom Perl script (C.8, Appendix). The data in each column were then summed across all genes to speed up computing as suggested in previous studies (Eyre-Walker and Keightley, 2009; Gossmann *et al.*, 2013).

The software produces two main sets of results. First, the proportions of mutations within four categories of strength of selection (N_es: < 1, 1-10, 10-100 and $>$ 100; where N_e = effective population size & s = selection coefficient) are calculated based on the polymorphism data. Then this output, combined with the divergence data, is used to calculate the proportion of adaptive substitutions (α). I combined the last two N_es categories and classed the three categories as effectively neutral

(< 1), slightly deleterious (1-10) and strongly deleterious (> 10; Eyre-Walker and Keightley 2009; Gossmann *et al.* 2013; Arunkumar *et al.* 2013).

3.2.4 Evolutionary rates

To estimate evolutionary rates of proteins, I calculated dN/dS ratios (ratio of non-synonymous to synonymous substitution rates relative to the number of corresponding non-synonymous and synonymous sites) per gene for all orthologous genes between *B. terrestris* and *B. impatiens* using PAML (Yang, 2007). I detected orthologues following the methods described in Szövényi *et al.* (2013). Following this protocol, I found orthologues by performing reciprocal blastp searches (Altschul *et al.*, 1997) between *B. terrestris* and *B. lyrata* protein sequences, which were downloaded from the RefSeq database. Pairs with mutual best hits showing at least 30% identity along 150 aligned amino acids were retained (Rost, 1999). I aligned proteins with MUSCLE (Edgar, 2004) at default settings and mapped the protein alignments onto the corresponding, aligned mRNA pairs with pal2nal (Suyama *et al.*, 2006). I used the codeml program with runmode -2, model 2 and 'NSsites' set to 0 for each alignment. Genes with a dS greater than 2 were removed from analyses.

3.2.5 Polymorphism analyses

I counted and compared the numbers of genes containing indels (all insertions or deletions) between the gene groups using vcftools. Subsequently, I reduced the polymorphism data set to only include single nucleotide polymorphisms (SNPs) with the vcftools filter option –remove-indels. I performed the diversity analyses on three sets of data: all SNPs, SNPs at synonymous sites and SNPs at non-synonymous sites. Again, I used 4-fold and 0-fold sites to represent synonymous and non-synonymous nucleotide positions. For the analyses on SNPs at 4-fold and 0-fold degenerate sites, I filtered the gVCF file to only contain the coordinates of these sites. I removed any codons containing more than one SNP from the analyses on 4-fold and 0-fold sites. For each gene within each category of site, I counted the total number of SNPs using custom Perl scripts (C.6 & C.7; Appendix). Watterson's theta (θ_w; Watterson 1975) was then calculated from the numbers of SNPs per gene based on the average number of available alleles per gene, which was obtained with the vcftools function –counts. To obtain relative values per gene, I then either divided the sums of SNPs and θ_w by gene length (coding regions only) or by the total number of considered 4-fold or 0-fold degenerate sites for each gene (the latter is outputted from Perl script C.6; Appendix). To analyse intra-specific selective constraints, I calculated

pN/pS (ratio of nonsynonymous to synonymous polymorphism rates; Schloissnig *et al.* 2013) per gene by dividing the relative numbers of 0-fold SNPs by the relative numbers of 4-fold SNPs. I excluded genes with no 4-fold SNPs from this calculation.

3.2.6 Controlling for confounding factors

It is known that at least expression level and codon bias (Drummond *et al.*, 2006) are associated with protein evolution. Codon bias may also be correlated with gene length (Akashi, 2001). Here I estimated the possible influence of five genomic parameters on dN/dS and pN/pS. These were expression level, codon bias variance, gene density, GC-content and gene length. I extracted expression level from the output of the DESeq2 analysis (column 'baseMean'). I used RSCU (relative synonymous codon usage) to measure codon bias, which I calculated for each codon of each locus with the R package 'seqinr' (uco() function; version 3.1-3; Charif and Lobry 2007). I calculated gene density with custom Perl and R scripts by counting the number of genes within each block of 100kb and then attributed gene densities to each gene according to its corresponding 100kb window. I downloaded a Perl script for determining GC content, which was originally written by Dr. Xiaodong Bai (http://www.oardc.ohio-state.edu/tomato/HCS806/GC-script.txt). I calculated gene length directly from information contained in the gff file.

To condense the factors (expression level, codon bias variance, gene density, GC-content and gene length) into one variable I performed a principal components regression analysis (Mandel, 1982) using the pcr() function in the pls package (version 2.4-3; Mevik and Wehrens 2007) on the five parameters with dN/dS or pN/pS as the dependent variable. As described in Drummond *et al.* (2006) all input variables were log transformed and scaled to zero mean and unit variance using the scale() function in R. The significance of each factor was tested with a jack knife test and all non-significant factors ($p < 0.05$) were removed. I then used the principal component which explained the largest amount of variance for the dependent variable, as the continuous variable in an ANCOVA analysis with dN/dS or pN/pS as the dependent variable and ploidy-bias (haploid, diploid or non-biased) as the co-variable. With this analysis the significance of observed differences in dN/dS and pN/pS between gene groups (haploid-biased, diploid-biased and non-biased) can be tested in the context of one rather than five inter-correlated predictors.

3.2.7 Statistical analyses

I carried out all statistical analyses in R (version: 3.2.2; Team 2012). I tested the significance of differences between groups with a two-sided Mann-Whitney U test. In case of multiple testing, I corrected all p-values with the Bonferroni method. I carried out Spearman rank partial correlations with the function pcor.test() from the ppcor package (version 1.0; Kim 2012).

3.3 Results

3.3.1 Over 13% of genes ploidy-biased

I investigated differential expression for a total of 10,794 genes. In the four comparisons between males and females (male & worker larvae, pupae and adults, adult males and queens) 220 genes were diploid-biased and haploid-biased in different comparisons and were excluded. Of the remaining 10574 genes, 818 (7.7%) were haploid-biased (significantly upregulated with a fold change > 5), 622 (5.9%) were diploid-biased and 2925 (27.7%) genes were non-biased (not differentially expressed: corrected $p \geq 0.05$).

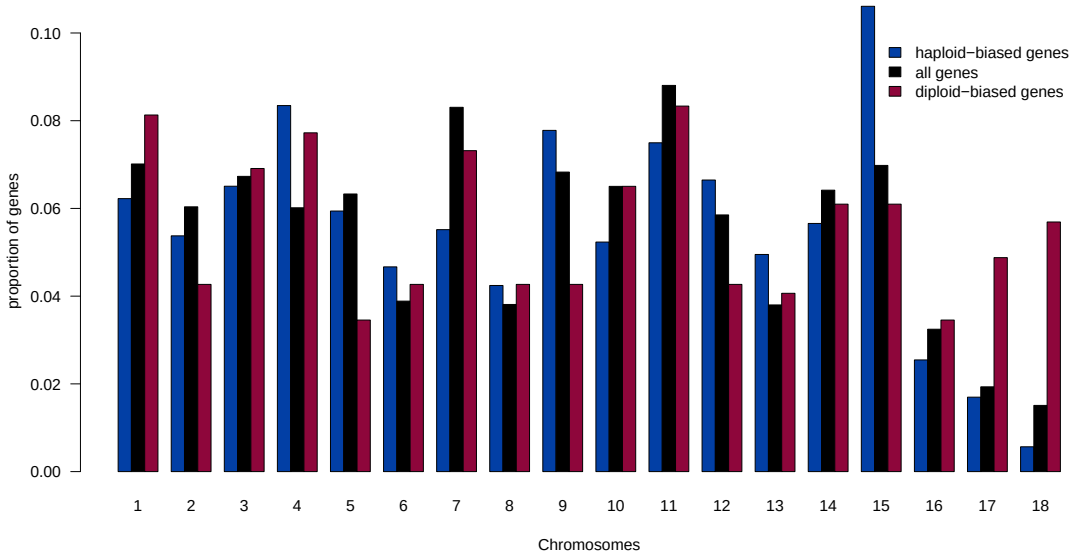


Figure 3.2: Proportions of haploid-biased and diploid-biased genes compared to all genes within each of the 18 chromosomes.

Not all scaffolds of the draft genome have been assigned to the 18 chromosomes. A higher number of non-biased (87.0%) and haploid-biased genes (86.4%) have been assigned to the 18 chromosomes compared to 79.1% of the diploid-biased genes. The remaining genes are mapped to unplaced scaffolds. I analysed the distribution of the haploid, diploid and non-biased genes among the 18 chromosomes in order to control for possible differences in mutation rates among the chromosomes. Within the genes which have been assigned to a chromosome, a χ^2 test showed a significant difference in the distribution of the ploidy-biased genes among the 18 chromosomes compared to all other genes (haploid-biased: $\chi^2 = 50$, $p < 0.001$, df: 17; diploid-biased: $\chi^2 = 126$, $p < 0.001$, df: 17). Non-biased genes did not differ in their distribution among

the chromosomes from all other genes ($\chi^2 = 26$, $p > 0.05$, df: 17). The distribution of haploid-biased genes differed from that of diploid-biased genes ($\chi^2 = 296$, $p < 0.001$; df: 17).

Haploid-biased genes were over-represented on chromosomes 4, 6, 13 and 15 (38.7%, 20.1% 30.3% and 52.0% higher than the total proportion of genes), while diploid-biased genes were also over-represented on chromosome 4 and especially on chromosomes 17 and 18 (28.4%, 152.4% and 277.1% higher than the total proportion of genes; figure 3.2). The median chromosomal positions did not differ between haploid and diploid-biased genes on any of the 18 chromosomes. However, non-biased genes did differ significantly from haploid-biased genes in their distribution along chromosome 9 ($p = 0.022$) and from diploid-biased genes along chromosomes 13 and 17 ($p = 0.014$ & 0.007, respectively; figure 3.3).

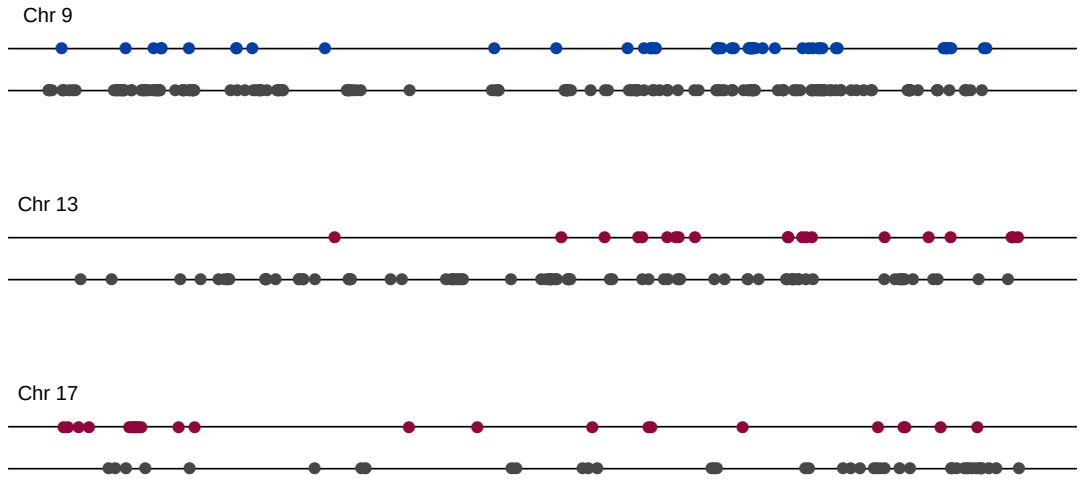


Figure 3.3: Comparison of chromosomal positions of haploid-biased (blue) and non-biased genes (grey) along chromosome 9; and of diploid-biased (red) and non-biased genes (grey) along chromosomes 13 and 17.

3.3.2 Selection strength and type differ between gene groups

The distribution of fitness effects (Eyre-Walker and Keightley, 2009) showed the highest proportion of mutations with $N_e s > 10$ (selection strength; N_e = effective population size; s = selection coefficient), and therefore strongly deleterious, for haploid-biased genes (71.0%) followed by non-biased genes (64.3%; figure 3.4). The proportion of strongly deleterious alleles was lowest for diploid-biased genes (58.6%). Diploid-biased genes showed a higher proportion of effectively neutral mutations

($N_e s < 1$; 14.3%) compared to haploid-biased genes (9.8%) and non-biased genes (12.3%), indicating weaker purifying selection within the diploid-biased genes (figure 3.4). The proportion of slightly deleterious mutations ($1 < N_e s < 10$) was also highest within diploid-biased genes (27.1%) offering further evidence of weaker purifying selection. The proportion of slightly deleterious mutations was lowest within haploid-biased genes (19.3%) and intermediate within non-biased genes (23.4%; figure 3.4).

In order to corroborate these interpretations regarding purifying selection I compared the frequencies of rare alleles within each group of genes at 4-fold and 0-fold sites. For this I analysed all genes for which 100% coverage was available, i.e. 54 alleles (170 haploid-biased, 129 diploid-biased and 1767 non-biased genes). As to be expected in the case of stronger purifying selection, I found a significantly higher proportion of singletons within the 0-fold sites for haploid-biased (22.3%) and non-biased genes (21.0%) compared to diploid-biased genes (14.2%; χ^2 tests, p= 0.008 & 0.004 respectively). This difference was non-significant for 4-fold sites (haploid: 18.2%; diploid: 16.3% & non-biased: 17.6%).

Using the same DoFE software I estimated the proportion of adaptive substitutions (α) within each gene group. α was highest within diploid-biased (65.4%), slightly lower within haploid-biased genes (62.1%) and lowest within non-biased genes (56.1%; figure 3.5).

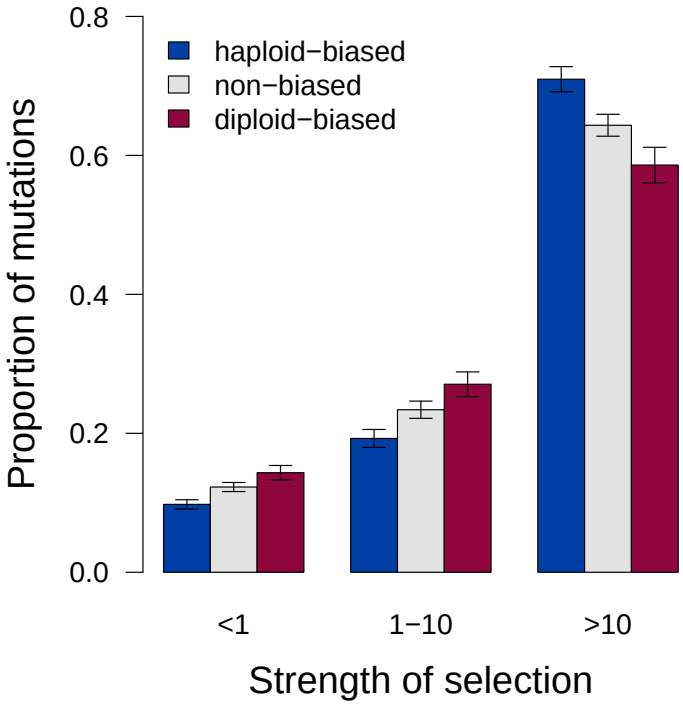


Figure 3.4: Distribution of fitness effects for haploid-biased, non-biased and diploid-biased genes. Shown are the mean proportions of mutations in three $N_e s$ (selection strength) ranges with SDs (see methods 3.2.3).

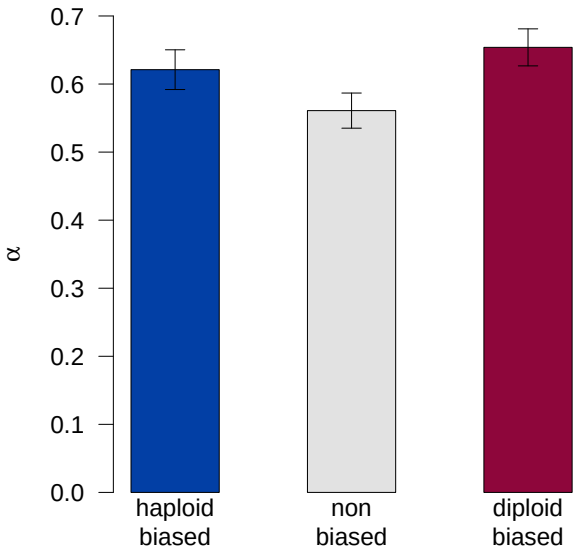


Figure 3.5: The proportion of adaptive subsitutions (α) for haploid-biased, non-biased and diploid-biased genes. Shown are means and SDs.

3.3.3 Increased evolutionary rates in diploid-biased genes

I estimated inter-specific sequence divergence for 8627 orthologues between *Bombus terrestris* and *Bombus impatiens*, of which 634 were haploid-biased, 360 diploid-biased and 2275 were non-biased. In agreement with my estimation of stronger purifying selection, in the previous section (3.3.2), the evolutionary rate of proteins (dN/dS) was significantly slower within haploid-biased (median: 0.176) and non-biased genes (median: 0.128; p = 1.3 x 10^{-4} & 1.6 x 10^{-21} respectively) compared to diploid-biased genes (median: 0.253; figure 3.6). dN/dS was significantly higher within haploid-biased genes compared to non-biased genes (p = 1.3 x 10^{-11}). These dN/dS patterns were mainly driven by differences in nonsynonymous (dN) rather than synonymous divergence (dS), for which the relative differences between gene groups were much greater (table 3.2).

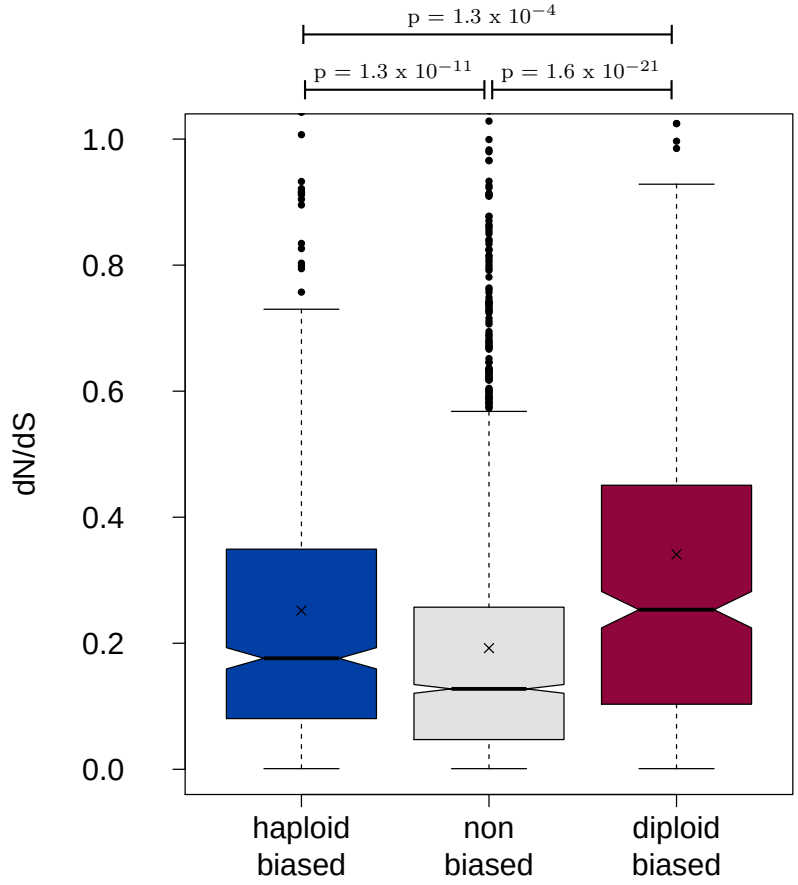


Figure 3.6: dN/dS for ploidy-biased and non-biased genes.
Whiskers range to 1.5 x interquartile range; "x" denotes mean.

Table 3.2: Divergence rates at nonsynonymous (dN) and synonymous (dS) sites. The top half of the table shows median values for haploid-biased (H), diploid-biased (D) and non-biased (N) genes. The bottom half shows the relative difference between the gene groups.
* $p < 0.05$; ** $p < 10^{-3}$; *** $p < 10^{-6}$

	dN	dS	dN/dS
haploid-biased	0.011	0.060	0.176
diploid-biased	0.019	0.073	0.253
non-biased	0.008	0.065	0.128
H:D	39%***	18%***	30%**
H:N	38%***	7%*	38%***
D:N	127%***	13%***	98%***

An ANOVA comparing divergence measures between the 18 chromosomes showed a significant variation in dN and dS among chromosomes ($p = 0.004$ & $p = 0.028$) but no significant variation in dN/dS. The variation in divergence between gene groups could therefore be caused by their non-random distribution among the chromosomes or within chromosomes rather than differences in ploidy-bias due to variation in mutation rates. To test for this I removed all chromosomes on which haploid-biased or diploid-biased genes were over-represented or positions varied between groups (4, 6, 9, 13, 15, 17 and 18; see section 3.3.1), and repeated the analyses. In the reduced data set (358 haploid-biased, 212 diploid-biased and 1422 non-biased genes) none of the tendencies changed and all differences remained significant (table 3.3).

Table 3.3: Divergence rates at nonsynonymous (dN) and synonymous (dS) sites within a reduced data set not including chromosomes 4, 6, 9, 13, 15, 17 and 18. The top half of the table shows median values for haploid-biased (H), diploid-biased (D) and non-biased (N) genes. The bottom half shows the relative difference between the gene groups.
* $p < 0.05$; ** $p < 10^{-3}$; *** $p < 10^{-6}$

	dN	dS	dN/dS
haploid-biased	0.011	0.060	0.165
diploid-biased	0.019	0.077	0.236
non-biased	0.008	0.065	0.117
H:D	37%***	22%***	30%*
H:N	38%**	7%*	41%***
D:N	120%***	19%***	102%***

The greater proportion of adaptive substitutions within haploid and diploid-biased genes, as indicated by an $\alpha > 60\%$ (figure 3.5), can explain the significantly higher dN/dS rates within both ploidy-biased groups ($p = 1.3 \times 10^{-11}$ & 1.6×10^{-21} respectively; figure 3.6) compared to non-biased genes. This tendency increased with increasing ploidy bias (figure 3.7). Genes with low ploidy bias (fold change of expression < 5) did not show significantly higher dN/dS rates than non-biased genes, in fact genes with low diploid bias had significantly lower evolutionary rates ($p = 0.013$). However, for both medium (fold change 5-10) and highly diploid-biased (fold change >10) genes, dN/dS was significantly higher than in non-biased genes ($p = 1.6 \times 10^{-5}$ and 5.5×10^{-20} respectively; figure 3.7). For haploid-biased genes dN/dS was also significantly higher than non-biased genes in the medium and high bias categories ($p = 3.4 \times 10^{-4}$ and 2.2×10^{-10} respectively; figure 3.7).

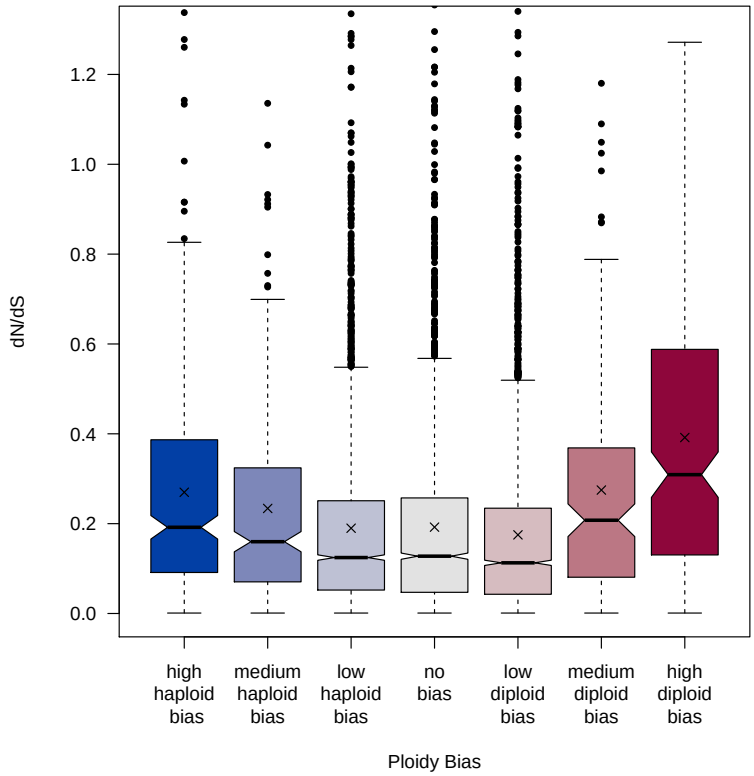


Figure 3.7: dN/dS for genes grouped by level of ploidy bias.
High bias: > 10 FC; medium: 5-10 FC; low: 0-5 FC; no bias: no significant difference in expression. FC = fold change in expression level.
Number of genes per group (from left to right): 238, 264, 2383, 2993, 2420, 143, 201. Whiskers range to max 1.5 x interquartile range; "x" denotes mean.

I tested the influence of five genomic parameters (expression level, codon bias variance, GC content, gene length and gene density) on the dN/dS patterns I ob-

served. Expression level was significantly higher within diploid than haploid-biased genes ($p = 6.2 \times 10^{-10}$; table 3.4) and correlated significantly and negatively with dN/dS (-0.109, $p = 2.3 \times 10^{-23}$; table 3.5) when controlling for the remaining four parameters. Gene density was also significantly negatively correlated with dN/dS (table 3.5) and was higher within haploid-biased genes (table 3.4). GC content was significantly higher within haploid-biased genes but did not correlate with dN/dS. Gene length and codon bias variance, on the other hand, did not differ between the two ploidy-biased groups but correlated significantly with dN/dS.

Table 3.4: Differences in 5 genomic variables between haploid and diploid-biased genes. Values are medians (means); significance tested with Mann Whitney U test; p values Bonferroni corrected.

Parameter	haploid-biased		diploid-biased	significance
Expression level	132.2 (2128.4)	<	385.4 (2879.6)	6.2×10^{-10}
Codon bias variance[1]	0.2482 (0.2898)	=	0.2659 (0.3049)	not signif.
GC-content	42.7 (42.6)	>	37.9 (38.4)	1.4×10^{-29}
gene length (bp)	3646.5 (12429.3)	=	3373.0 (13240.5)	not signif.
gene density[2]	7 (8.0)	>	6 (7.1)	0.0047

[1]Variance of RSCU; [2]Number of genes per 100kb window.

Table 3.5: Partial correlations of 5 genomic paramters with dN/dS. Spearman rank partial correlation; p values Bonferroni corrected.

Parameter	correlation with dN/dS	significance
Expression level	-0.109	2.3×10^{-23}
Codon bias variance	0.052	5.6×10^{-6}
GC-content	no correlation	
gene length	-0.061	9.1×10^{-8}
gene density	-0.057	6.0×10^{-7}

In order to control for the possible influence of these five parameters (expression level, codon bias variance, GC content, gene length and gene density) on the measured differences in dN/dS between haploid-biased, diploid-biased and non-biased genes, I carried out a principal component regression. This analysis allowed me to condense any significant predictors of dN/dS into one principal component (PC). PC2 which explained the largest portion of variation in dN/dS (2%) was then further investigated for the significance of all predictors. Gene density and GC content had no significant effect on the variation in PC2 so were removed. The remaining three significant parameters explained only 3.2% of variation in dN/dS, the majority of

which was contributed via PC1 (2.1%). I used PC1 as the continuous variable in an ANCOVA analysis with dN/dS as the dependent variable and ploidy bias (haploid, diploid or non-biased) as the co-variable. When controlling for the three genomic factors via PC1, dN/dS remained significantly lower for haploid-biased ($p = 0.0023$) and non-biased genes ($p = 1.7 \times 10^{-19}$) compared to diploid-biased genes (figure 3.8).

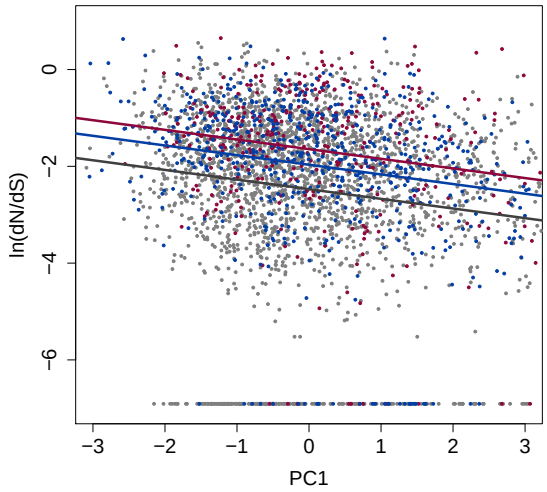


Figure 3.8: ANCOVA analysis with PC1 as continuous variable and ln(dN/dS) as the dependent variable plotted on three levels of bias. Red: haploid-biased genes and regression line; blue: diploid-biased genes and regression line; grey: non-biased genes and regression line.
Haploid-biased regression line significantly lower than diploid-biased; $p = 0.0023$. Non-biased regression line significantly lower than diploid-biased; $p = 1.7 \times 10^{-19}$.

3.3.4 Diversity levels lower in haploid-biased genes

I analysed the 27 transcriptomes of wild workers for differences in polymorphism levels between the three groups. After filtering (see METHODS 3.2.2), sequence data was available for 9203 genes (87.0%) in total and for 2447 of the non-biased genes (83.7%). A slightly higher fraction of the diploid-biased genes could be recovered (73.0%; 454) compared to that of the haploid-biased genes (67.2%; 550).

Total polymorphism levels, estimated with Watterson's theta (θ_w), were significantly higher within diploid-biased genes compared to both haploid-biased ($p = 0.003$) and non-biased genes ($p = 1.7 \times 10^{-9}$; figure 3.9). Polymorphism levels were also slightly higher among haploid-biased compared to non-biased genes ($p = 0.049$; figure 3.9). These high polymorphism levels in diploid-biased genes could

be attributed to a significantly higher θ_w at 0-fold degenerate sites (compared to haploid-biased genes: p = 3.8 x 10^{-9}; compared to non-biased genes: p = 1.6 x 10^{-8}; figure 3.9). The lower θ_w within haploid-biased genes compared to non-biased genes was only marginally significant (p = 0.0497). When I restricted the analysis to 4-fold sites, only the significant difference between haploid-biased and non-biased genes remained (p = 0.026; figure 3.9).

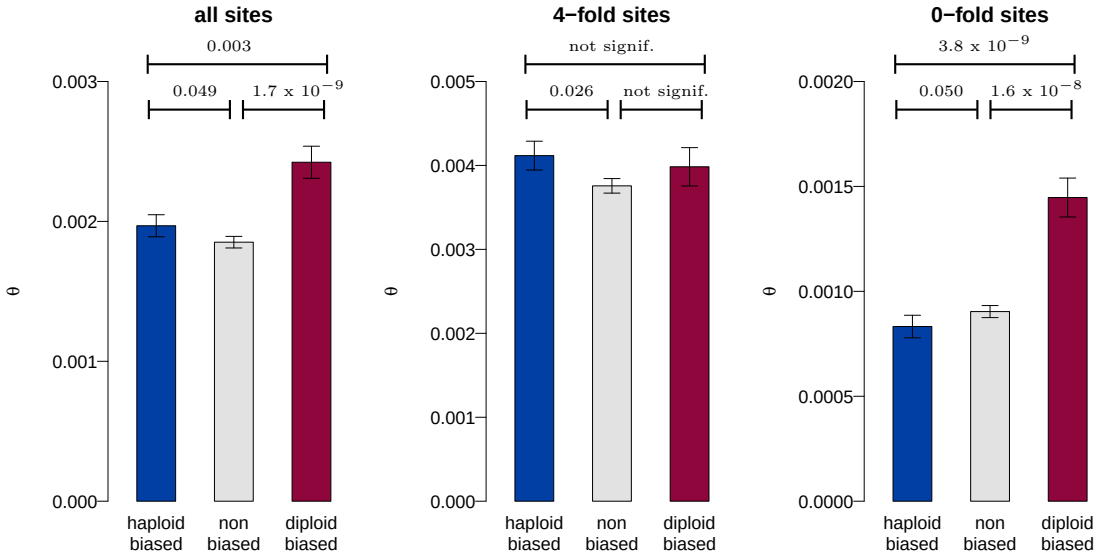


Figure 3.9: Watterson's theta at all sites, 4-fold and 0-fold sites for haploid-biased, non-biased and diploid-biased genes. Shown are means and standard errors.

As an indicator of the strength of purifying selection within the three gene groups I compared the number of genes which contained indel mutations. 25.3% of non-biased genes and 20.5% of diploid-biased genes contained indels compared to only 12.0% of haploid-biased genes. I measured protein diversity levels by comparing proportions of SNPs at 0-fold relative to proportions of SNPs at 4-fold sites (pN/pS) between gene groups. pN/pS was lowest within haploid-biased genes (mean: 0.200 ± 0.016) and highest within diploid-biased genes (mean: 0.375 ± 0.026). Non-biased genes had an intermediate level of diversity (mean: 0.278 ± 0.010). All differences were significant (haploid-diploid: p = 1.1 x 10^{-9}; haploid-non: p = 1.1 x 10^{-5}; diploid-non: p = 1.6 x 10^{-4}; figure 3.10).

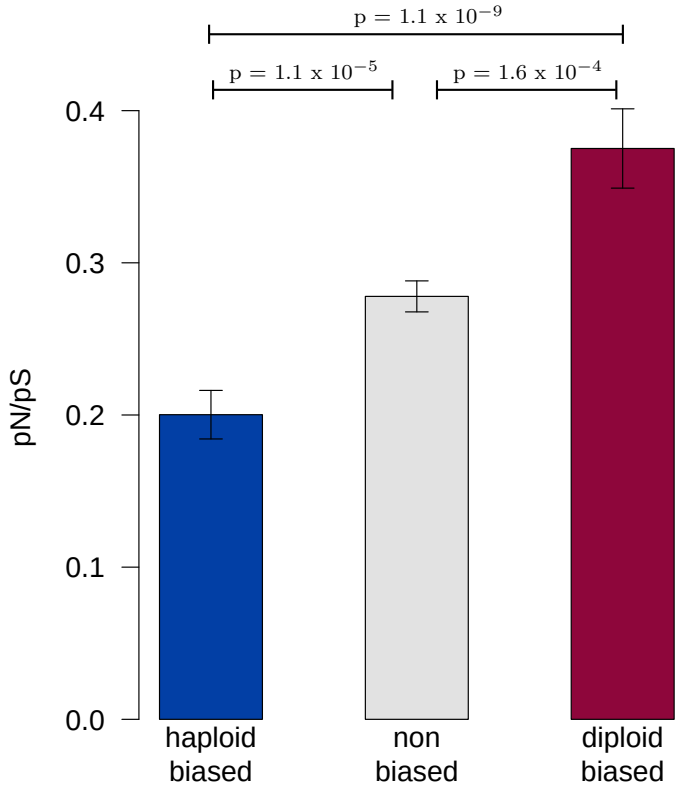


Figure 3.10: pN/pS within the three gene groups.
pN/pS: proportion of SNPs at 0-fold sites relative to proportion of SNPs at 4-fold sites.

In a principal components regression analysis to estimate the potential influence of factors other than ploidy bias on the observed variation in pN/pS, codon bias variance, GC-content and gene length contributed significantly to 2.5% of variation. The largest principal component (PC 2; 1.4% of variation in pN/pS) was plotted against pN/pS in an ANCOVA analysis on the three levels of ploidy bias: haploid, diploid and non-bias. Within diploid-biased genes pN/pS remained significantly higher than haploid-biased genes ($p = 7.0 \times 10^{-6}$) and, to a lesser extent, non-biased genes ($p = 0.014$; figure 3.11).

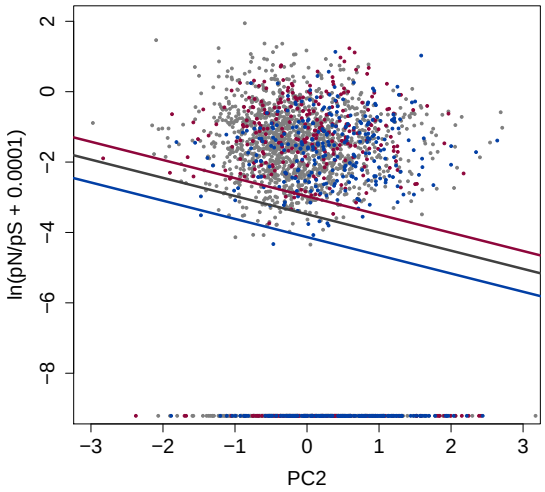


Figure 3.11: ANCOVA analysis with PC2 as continuous variable and ln(pN/pS + 10^{-4}) as the dependent variable plotted on three levels of bias. Red: haploid-biased genes and regression line; blue: diploid-biased genes and regression line; grey: non-biased genes and regression line.
Haploid-biased regression line significantly lower than diploid-biased; p = 7.0 x 10^{-6}.
Non-biased regression line significantly lower than diploid-biased; p = 0.014.

The allele sample size, that is the number of reliably sequenced alleles (maximum 54), did not differ significantly between haploid-biased (median: 44.0; mean: 38.0) and diploid-biased genes (median: 42.0; mean: 38.4). However, these levels were significantly lower than for non-biased genes (median: 52.0; mean: 46.8). In order to account for this variation, which could influence the sensitivity for detecting polymorphisms within the two ploidy-biased gene groups relative to the non-biased genes, I randomly sampled 20 alleles at each site for each gene and repeated the analyses. Of all the polymorphism values (θ_w at each site type and pN/pS) only θ_w at 0-fold sites was slightly lower than in the full data-set for diploid-biased genes (mean of 0.0010 compared to 0.0014; p = 0.040) and non-biased genes (mean of 8.7 x 10^{-4} compared to 9.0 x 10^{-4}; p = 0.010).

Within this reduced data-set, most tendencies remained between ploidy-biased and non-biased genes for all analyses (figures A.1, Appendix & 3.12). The only exception was that θ_w at 0-fold sites was now lowest rather than intermediate for non-biased genes (figure A.1, Appendix). All differences between non-biased and either of the two sets of biased genes were now non-significant after reducing the data-set, except for θ_w at all sites, which was significantly lower compared to diploid-biased genes (p = 0.025). It is possible that the loss of significance in these comparisons was the result of a reduction in power due to reduced gene sample sizes (haploid-biased

genes reduced from 529 to 204; diploid-biased genes reduced from 431 to 194).

Because allele sample size did not differ between haploid- and diploid-biased genes their comparison in the full data set is valid. Nevertheless, when reducing the data set, the difference in pN/pS remained significantly higher within diploid-biased compared to haploid-biased genes showing the robsutness of this result (p = 0.019; figure 3.12). For θ_w at each of the site types the difference between haploid- and diploid-biased genes was reduced (figure A.1). The large, observable difference in θ_w at 0-fold sites, which remained between haploid-biased (mean: 8.9 x 10^{-4}; median: 3.5 x 10^{-4}) and diploid-biased genes (mean: 1.0 x 10^{-3}; median: 6.2 x 10^{-4}; figure A.1), was no longer significant after correcting (uncorrected p = 0.034; corrected p = 0.102). Again this was possibly due to a reduction in power.

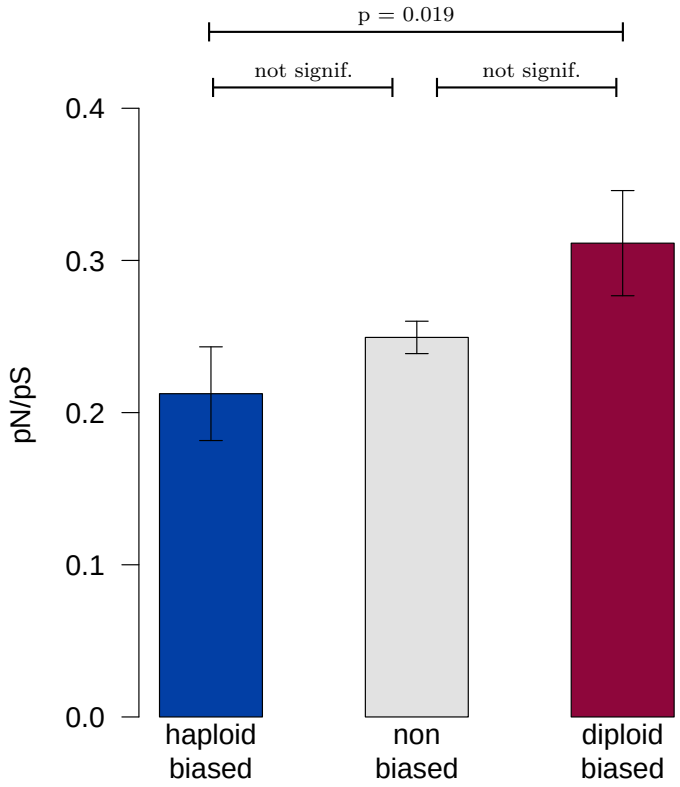


Figure 3.12: pN/pS within the three gene groups for 20 resampled alleles.
pN/pS: proportion of SNPs at 0-fold sites relative to proportion of SNPs at 4-fold sites.

3.4 Discussion

In this study I tested White's (1945) hypothesis regarding the effect of ploidy on selection in haplodiploids on the buff-tailed bumblebee, *Bombus terrestris*. To the best of my knowledge, this is the first empirical study of its kind conducted on Hymenoptera, or even on any haplodiploid organism. In agreement with my expectations I found evidence for higher levels of purging within genes which are expressed in haploids compared to diploid-biased genes. Contrary to my expectations, I found no evidence of weaker positive selection within diploid-biased genes. Positive selection was generally higher within both haploid and diploid-biased genes compared to non-biased genes.

3.4.1 Stronger purging and a slow haploid effect

Within the haploid-biased and non-biased genes a larger proportion of mutations were strongly deleterious ($N_e s > 10$) and a lower proportion were effectively neutral ($N_e s < 1$) or slightly deleterious ($1 < N_e s < 10$) compared to diploid-biased genes. This is evidence of stronger purifying selection acting upon genes expressed in haploid males (Eyre-Walker and Keightley, 2009; Gossmann *et al.*, 2010, 2013). This observation of a higher frequency of effectively neutral and slightly deleterious mutations within diploid-biased genes, confirms my expectations and the predictions of Haldane (1924), White (1945) and Charlesworth *et al.* (1987) that recessive alleles will be less efficiently purged when diploid than haploid. The less efficient purging of recessive, deleterious mutations in diploid-biased genes due to masking has possibly led to a greater accumulation of slightly deleterious substitutions due to random drift compared to haploid-biased genes. This higher proportion of deleterious, nonsynonymous substitutions in diploid-biased genes is detectable in a significantly higher evolutionary protein rate (dN/dS). Also, less efficient purging is detectable in a raised level of intra-specific, non-synonymous diversity (θ_w), a higher pN/pS rate and a higher proportion of indel mutations within diploid-biased genes.

Most faster-X studies have attempted to detect positive selection by testing for differences in dN/dS, with the results of these tests showing variable support (Betancourt *et al.*, 2002; Thornton *et al.*, 2006; Thornton and Long, 2002; Musters *et al.*, 2006). However, in at least two studies, stronger purifying selection has been detected for X-linked genes (Connallon, 2007; Singh *et al.*, 2008). In the first of these two studies, as in the current study, stronger purifying selection was identified via lower pN/pS levels within X-linked genes (Connallon, 2007), and in the second via a

lower codon bias (Singh *et al.*, 2008). X-linked genes are analogous to haploid-biased and non-biased genes in the present study as they are subject to haploid expression in the male. In the plant *Capsella grandiflora* stronger purifying selection was found in genes expressed in pollen tissues, which are haploid, compared to genes limited to the diploid, sporophytic tissues (Arunkumar *et al.*, 2013). This, as in the current study, was evident in a lower proportion of neutral ($N_e s < 1$) and slightly deleterious ($N_e s$ 1-10) mutations in the haploid-biased, pollen genes.

3.4.2 No fast haploid effect

One unexpected finding was the apparent lack of a 'faster-haploid' effect, the rapid accumulation of adaptive substitutions within haploid- compared to diploid-expressed genes. The proportion of substitutions which were adaptive (α) was not higher within haploid-biased genes as would be expected in the case of more efficient positive selection. In fact α was slightly higher within diploid-biased genes (65.4% versus 62.1%). The frequency of non-neutral mutations which are adaptive is probably rare (Eyre-Walker and Keightley, 2007). This may explain why the effect of haploid expression appears to be stronger for purifying selection acting on slightly deleterious, recessive mutations than for positive selection in *B. terrestris*. The scarcity of recessive, beneficial mutations has already been suggested as the possible cause for an overall weak faster-X effect in *Drosophila* (Mank *et al.*, 2010b).

A higher relative frequency of slightly deleterious, recessive mutations compared to beneficial, recessive mutations may explain the stronger effect of haploid expression on purifying selection. However, it does not account for the slightly higher proportion of adaptive substitutions within diploid-biased genes. In fact, stronger positive selection is often detected in male-biased genes rather than in female-biased genes. For instance, a study on the faster-X effect in *Drosophila* found that not only X-linked but also male-biased, autosomal genes evolved faster than female-biased, autosomal genes due to positive selection (Baines *et al.*, 2008). Similarly, a comparative analysis of genes over-expressed in male and female reproductive tissues in *Arabidopsis* found the highest levels of purifying and positive selection in the male tissues (Gossmann *et al.*, 2013).

The fact that positive selection is not superior in males to females in this study may suggest that male sexual selection does not play such a dominant role in *B. terrestris*. It has often been suggested that sexual selection may be weaker in monogamous, eusocial taxa (Boomsma, 2007), such as *B. terrestris* (Schmid-Hempel and Schmid-Hempel, 2000), and the results I present here may support these suggestions. This statement is speculative, however, since male bumblebees do show specific ad-

aptations, indicative of effective sexual selection (Baer, 2003), and the results I present here need to be corroborated in further hymenopteran species.

To speculate further, based on the results at hand, it is possible that selection is generally stronger on female castes in *B. terrestris*. This could explain the apparent lack of a haploid effect on adaptive, recessive mutations, and that in fact a slightly higher proportion of adaptive substitutions existed for diploid-biased genes. This may be caused by the broader range of selective pressures which a queen or worker must endure in its lifetime compared to a male. Bumblebee queens, for instance, must after mating first survive winter by hibernating before finding a suitable nesting place and then founding and maintaining a new colony the following spring. Workers fulfill several tasks including foraging, nest-guarding and nursing. In contrast, males are not required to fulfill any tasks within the colony. They are short-lived, being produced towards the end of a colony cycle, and their primary task is to reproduce. Stronger adaptive evolution would also explain the higher diversification of workers and queens within the eusocial Hymenoptera, whereas the males are generally quite similar in their life cycles. The effect of stronger selection on haploids on account of their unmasked exposure to selection may, therefore, be counterbalanced by the greater selection on diploid-expressed genes because of their exposure to a wider range of selective pressures.

3.4.3 Ploidy-biased genes under stronger positive selection

The proportion of mutations within non-biased genes, which were neutral or slightly deleterious, was intermediate between the levels measured in haploid- and diploid-biased genes. This indicates purifying selection is more efficient within non-biased compared to diploid-biased genes. This is to be expected for this group of genes since they too are expressed within haploid males. However, the question of why purifying selection on non-biased genes is weaker than in haploid-biased genes needs to be addressed. A consideration of positive selection, on which the haploid effect is apparently weak, may offer a solution. It seems positive selection is stronger in both haploid-biased and diploid-biased genes compared to non-biased genes, revealed by a higher α (62%, 65% & 56%, respectively) and higher dN/dS rates (medians: 0.176, 0.253 & 0.128, respectively) in ploidy-biased genes. The higher level of positive selection is most likely linked to sex-bias rather than ploidy-bias. Higher rates of adaptive evolution have often been found for genes involved in reproduction (Turner and Hoekstra, 2008; Swanson and Vacquier, 2002). Also, both within autosomal and X-linked genes, lower levels of positive selection have been found for non-biased genes in *Drosophila melanogaster* (Baines *et al.*, 2008; Pröschel *et al.*, 2006). If

beneficial mutations have a strong selective effect on sex-biased genes then deleterious mutations are also likely to have a stronger selective effect than in non-biased genes. Therefore the intermediate position of non-biased genes between haploid and diploid-biased genes in terms of purifying selection can be explained by a combination of haploid expression (stronger in non-biased genes than diploid-biased genes) and sexual selection (stronger in ploidy-biased genes). A combination of stronger purifying selection and positive selection may have led to an intermediate dN/dS level for haploid-biased genes, while stronger purifying selection but low positive selection have caused the lowest dN/dS rate in non-biased genes. A possible alternative explanation is that caste-biased genes are generally under relaxed purifying selection, which would lead to raised dN/dS level compared to the non-biased genes (Hunt *et al.*, 2011).

3.4.4 Compatibility with faster-X effect

In contrast to faster-X and faster-Z studies, analyses are not complicated by differences in N_e in the current study. Males and females share the same chromosomes; only the ploidy level differs between sexes. Mutation rates possibly vary between the groups of ploidy-biased genes due to their uneven distribution among chromosomes. However, I have shown that this does not affect divergence rates. The difference in evolutionary rates between ploidy-biased genes are therefore most likely caused by sex or ploidy. One can assume that selection differs between sexes due to their very distinct biologies. If, however, sex alone were responsible, if for example stronger selective pressures existed for males, then one would expect stronger selection on both deleterious and beneficial mutations. Sex alone does not explain the more efficient purging of slightly deleterious alleles in males but the marginally more efficient positive selection of adaptive mutations in females. If, on the other hand, I assume recessive, slightly deleterious mutations are more frequent than recessive, beneficial mutations, then ploidy can sufficiently explain the selection patterns presented in this study. Due to more efficient purifying selection, as a result of haploid expression and the relative frequency of recessive, slightly deleterious alleles, male-biased genes evolve more slowly than female-biased genes in the haplodiploid species *B. terrestris*. This supports the prediction formulated by MJD White over 70 years ago (1945).

Chapter 4

The effects of ploidy-specific expression on selection in *Arabidopsis thaliana*

4.1 Introduction

In plants, more efficient selection can be expected for genes which are expressed in pollen tissues, which are haploid, compared to genes with expression limited to sporophytic tissues, which are diploid. This is because, in contrast to pollen genes, recessive mutations may be masked by a dominant allele in sporophyte-limited genes when heterozygous. Additionally, pollen competition is likely to strengthen selection on pollen-limited genes compared to sporophytic genes (Charlesworth and Charlesworth, 1992). These effects of haploid expression and pollen competition were shown recently for *Capsella grandiflora*, a member of the mustard family, for which the rate of protein evolution was significantly higher for pollen genes compared to sporophytic genes (Arunkumar *et al.*, 2013). Evidence was presented for stronger purifying and positive selection within genes exclusively expressed in pollen compared to sporophyte-specific genes.

For *Arabidopsis thaliana*, the self-compatible close relative of *C. grandiflora*, the findings of two recent studies have not been as clear-cut (Szövényi *et al.*, 2013; Gossmann *et al.*, 2013). Szövényi *et al.* (2013) also found a significantly higher rate of protein evolution within pollen-specific genes compared to sporophyte-specific genes for *A. thaliana* (Szövényi *et al.*, 2013). However, the detection of higher polymorphism levels within pollen genes led to the conclusion that relaxed purifying selection on pollen genes had caused the higher divergence rates, rather than stronger positive selection, which would have reduced intra-specific polymorphism levels. Higher

tissue specificity and expression noise (variance in expression level) compared to sporophytic genes were considered the likely causes of relaxed selection on pollen genes.

In contrast, Gossmann *et al.* (2013) found no difference in protein divergence between pollen genes and sporophytic genes. Also, despite using a larger number of accessions to measure polymorphism than in the Szövényi *et al.* study (80 compared to 19), Gossmann *et al.* did not detect any difference in nucleotide diversity between non-reproductive genes and pollen-specific genes in general, although nucleotide diversity was significantly lower for sperm cell-specific genes (2013). However, by comparing polymorphism to divergence data with a modified version of the McDonald-Kreitman test (McDonald and Kreitman 1991, Distribution of Fitness Effects Software, DoFE; Eyre-Walker and Keightley 2009) a higher proportion of non-synonymous sites were found to be under purifying and adaptive selection for pollen genes compared to both female-biased and non-reproductive genes (Gossmann *et al.*, 2013).

As pointed out by Gossmann *et al.* (2013), inter-specific divergence and intra-specific polymorphisms likely arose under different selection regimes for *A. thaliana*. The divergence of *A. thaliana* from its closest relative *A. lyrata* happened largely during a period of outcrossing, since speciation occurred approximately 13 million years ago (Beilstein *et al.*, 2010), whereas *A. thaliana* became self-compatible roughly one million years ago (Tang *et al.*, 2007). Divergence patterns between *A. thaliana* and *A. lyrata* should therefore be similar to outcrossing species, because for 25/26 of the evolutionary time, since *A. thaliana* and *A. lyrata* shared a common ancestor, evolution has been in an outcrossing environment. Divergence patterns should therefore reveal stronger selection on pollen genes due to selection in the haploid stage and pollen competition. Existing, intra-specific polymorphisms, on the other hand, are expected to be influenced by high selfing rates in *A. thaliana* populations that have led to high levels of homozygosity across the whole genome (Nordborg, 2000; Wright *et al.*, 2008; Platt *et al.*, 2010). One can expect the outcome to be a reduction in the masking of deleterious alleles in diploid sporophyte stages (because of high homozygosity) compared to the haploid gametophyte stage. Furthermore, selfing will result in fewer genotypes competing for fertilisation so lowering the magnitude of pollen competition and reducing the strength of selection acting on pollen (Charlesworth and Charlesworth, 1992). One can expect this reduction in pollen competition to lead to a relaxation of purifying selection and a build-up of deleterious variation.

The aim of this study was to attempt to resolve apparently conflicting results

concerning selection on pollen and sporophyte genes in *A. thaliana* (Szövényi *et al.*, 2013; Gossmann *et al.*, 2013) and to test the hypothesis that patterns of selection have changed for *A. thaliana* since it became self-compatible. I expected divergence data to show evidence for stronger selection on pollen-specific genes (expressed only in pollen tissues) compared to sporophyte-specific genes (expression limited to sporophytic tissues) due to a combination of pollen competition and the masking of recessive mutations in the sporophyte. The effect of masking and pollen competition were expected to have become weaker since *A. thaliana* became self-compatible and homozygosity levels have increased. This should lead to similar levels of selection in the pollen-specific and sporophyte-specific genes which should manifest in similar overall patterns of polymorphism.

In a first step I estimated the protein divergence of 1,552 pollen and 5,494 sporophytic genes between *A. thaliana*, *A. lyrata* and *Capsella rubella*, a close relative of *Arabidopsis*, in terms of interspecific dN/dS. Then, by combining polymorphism data with divergence data, a DoFE analysis was conducted to discover the proportion of sites under positive and negative selection within the two groups. As the polymorphism and divergence data likely reflect periods of differing selection regimes (divergence under self-incompatibility, polymorphism under self-compatibility), I additionally detected sites under positive selection using a site model of the Phylogenetic Analysis by Maximum Likelihood software (PAML 4.6; Yang 2007). This method does not require polymorphism data and detects sites under positive selection by allowing dN/dS to vary within genes. In a further step, to investigate more recent selection patterns, I analysed intra-specific polymorphism levels (nucleotide diversity and Watterson's theta) within pollen-specific and sporophyte-specific genes, and then compared existing levels of putative deleterious alleles (premature stop codons and frameshift mutations) between pollen genes and sporophytic genes.

4.2 Methods

4.2.1 Genomic data

Publicly available variation data were obtained for 269 inbred strains of *A. thaliana*. Beside the reference genome of the Columbia strain (Col-0), which was released in 2000 (*Arabidopsis*, Genome Initiative), I downloaded 250 from the 1001 genomes data center (http://1001genomes.org/datacenter/; accessed September 2013), 170 of which were sequenced by the Salk Institute (Schmitz *et al.*, 2013) and 80 at the Max Planck Institute, Tübingen (Cao *et al.*, 2011). I obtained a further 18 from the 19 genomes project (http://mus.well.ox.ac.uk/; accessed September 2013; Gan *et al.* 2011). These 269 files contained information on SNPs and indels recorded for separate inbred strains compared to the reference genome. I applied a quality filter to all files in order to retain only SNPs and indels with a phred score of at least 25. For further analyses, I created gene sequences for each of these strains based on coding sequence information contained in the TAIR10 gff3 file.

4.2.2 Expression data

I obtained normalised microarray data, covering 20,839 genes specific to different developmental stages and tissues of *A. thaliana* (table 4.1), from Borg *et al.* (2011). The expression data consisted of seven pollen and ten sporophyte data sets (table 4.1). Four of the pollen data sets represented expression patterns of the pollen developmental stages, uninucleate, bicellular, tricellular and mature pollen grain, one contained expression data of sperm cells and the remaining two were pollen tube data sets. There was a strong, significant correlation between the two pollen tube data sets ($\rho = 0.982$; $p < 2.2\text{x}10^{-16}$; Spearman's rank correlation), so I combined both and used the highest expression value of the two sets for each gene. Each of the ten sporophyte data sets contained expression data for specific sporophytic tissues (table 4.1).

Table 4.1: Expression data sets.

	Dataset	Description	Microarray chips	Original source
Haploid	UNM	Uninucleate microspore	2	Honys and Twell, 2004
	BCP	Bicellular Pollen	2	Honys and Twell, 2004
	TCP	Tricellular Pollen	2	Honys and Twell, 2004
	MPG	Mature Pollen	2	Honys and Twell, 2004
	GP*	Pollen Tube Grouped	6	Qin *et al.*, 2009; Wang *et al.*, 2008
	PT4*	Pollen Tube Grouped	6	Qin *et al.*, 2009; Wang *et al.*, 2008
	SPC	Sperm Cell	3	Borges *et al.*, 2008
Diploid	SL	Silique	30	NASC
	LF	Leaves	36	NASC
	GC	Guard Cell	3	NASC
	PT	Petiole	3	NASC
	ST	Stems	2	NASC
	HP	Hypocotyl	8	NASC
	XL	Xylem	3	NASC
	CR	Cork	3	NASC
	RT	Roots	11	NASC
	RH	Root hair elongation zone	3	NASC

Taken from Borg *et al.* (2011).
NASC: Nottingham Arabidopsis Stock Centre.
* GP and PT4 were combined to one data set, selecting the highest expression level of the two for each gene.

Each expression data point consisted of a normalised expression level (ranging from 0 to around 20,000, scalable and linear across all data points and data sets) and a presence score ranging from 0 to 1 based on its reliability of detection across replicates, as calculated by the MAS5.0 algorithm (Borg *et al.*, 2011). In my analyses I conservatively considered expression levels as present if they had a presence score of at least 0.9, while values < 0.9 were regarded as zero expression.

I classed genes as either pollen- or sporophyte-specific genes if expression was reliably detectable in only pollen or only sporophyte tissues or developmental stages. The highest expression value across all tissues or developmental stages was used to define the expression level of a particular gene.

4.2.3 Evolutionary Rates

To estimate evolutionary rates of proteins, I calculated dN/dS ratios (ratio of non-synonymous to synonymous substitution rates relative to the number of corresponding non-synonymous and synonymous sites) for all orthologous genes between *A. thaliana*, *A. lyrata* and *Capsella rubella* using the codeml program within the PAML package (Yang, 2007). Orthologues were found between each species pair using the protocol described in Szövényi *et al.* (2013) by performing reciprocal blastp searches (Altschul *et al.*, 1997) between protein sequences and retaining pairs with mutual best hits showing at least 30% identity along 150 aligned amino acids (Rost, 1999). For each pair of species I aligned orthologous proteins with MUSCLE (Edgar, 2004) at default settings and mapped the protein alignments onto the corresponding mRNA alignments with pal2nal (Suyama *et al.*, 2006). The codeml program was run with runmode -2, model 2 and 'NSsites' set to 0 for each alignment within each species pair. Genes with a dS > 2 were removed from the analysis.

In order to detect genes that contain codon sites under positive selection, I performed a likelihood-ratio test (LRT) between models 7 (null hypothesis; dN/dS limited between 0 and 1) and 8 (alternative hypothesis; additional parameter allows dN/dS > 1) by using runmode 0, model 0 and setting 'NSsites' to 7 & 8. An LRT statistic (twice the difference in log-likelihood between the two models) greater than 5.991 indicated a significant difference ($p < 0.05$) between the two models suggesting the existence of sites under positive selection within the tested gene (Anisimova *et al.*, 2003; Yang, 2007). This was performed on a three species alignment between *A. thaliana*, *A. lyrata* and *C. rubella*, which was produced in the same manner as the two-species alignments above.

4.2.4 Detecting levels of purifying selection and proportions of adaptive substitutions

Levels of purifying and positive selection were estimated with the Distribution of Fitness Effects Software (DoFE 3.0) using the Eyre-Walker and Keightley (2009) method. For the input files, the following data columns were required:

1. Gene
2. Number of chromosomes

   Within intra-specific polymorphism data:

3. Number of nonsynonymous sites

4. Nonsynonymous allele frequency spectrum
5. Number of synonymous sites
6. Synonymous allele frequency spectrum

Within inter-specific divergence data:

7. Number of nonsynonymous sites
8. Number of nonsynonymous substitutions
9. Number of synonymous sites
10. Number of synonymous substitutions

I used four-fold sites to represent synonymous positions and zero-fold degenerate sites to represent nonsynonymous positions. I detected and isolated four-fold and zero-fold sites with adapted Perl scripts used in chapter 3 (scripts C.5 and C.6; Appendix); any codons containing more than one SNP were removed from the analysis. I randomly sampled 20 alleles at each site without replacement using the Perl module 'shuffle'. I calculated synonymous and non-synonymous site spectra using the Pegas package (Paradis, 2010) in R (version 3.2.0; Team 2012). All input values (numbers of sites and allele frequencies) were summed across all genes to speed up computing. The software produces two main sets of results. First, the proportions of mutations within four categories of strength of selection ($N_e s$; < 1, 1-10, 10-100 and > 100) are calculated based on the polymorphism data. Then this output, combined with the divergence data, is used to calculate the proportion of adaptive substitutions (α).

4.2.5 Intra-specific polymorphism

I calculated nucleotide diversity (π) and Watterson's θ for non-synonymous sites using the R package PopGenome (version 2.1.6; Pfeifer *et al.* 2014). The diversity.stats() command was implemented and the subsites option was set to "nonsyn". Subsequently, I divided both values by gene length.

4.2.6 Putatively deleterious alleles

To quantify the frequency of deleterious mutations for each gene, I calculated the occurrence of premature stop codons and frameshifts for each gene locus across all 268 strains compared to the reference genome using custom Perl scripts. Stop

codons were recorded as the number of unique alternative alleles occurring within the 269 strains as a result of a premature stop codon. Frameshifts were calculated as a proportion of the strains containing a frameshift mutation for a particular gene. Both values were normalised by dividing by coding gene length. All analyses of coding regions were based on the representative splice models of the *A. thaliana* genes (TAIR10 genome release, www.arabidopsis.org).

4.2.7 Controlling for confounding factors

I investigated six genomic parameters as possible predictors of dN/dS, polymorphism levels and frequency of deleterious mutations. These were expression level, GC-content, codon bias variance, gene density, average gene length and average intron length. Expression level is described above in the section "Expression data". Average gene length and average intron length were calculated using custom-made Perl scripts which extracted information from the genomic gff file. GC content was calculated with a downloaded Perl script, which was originally written by Dr. Xiaodong Bai (http://www.oardc.ohio-state.edu/tomato/HCS806/GC-script.txt). RSCU (relative synonymous codon usage) was used to measure codon bias (Kotlar and Lavner, 2006). It was calculated for each codon of each locus with the R package 'seqinr' (uco() function; version 3.1-3; Perriere, 2014). As the mean value per gene varied very little between loci but varied by site within genes, I used RSCU variance as a measure for codon bias. I calculated gene density with custom Perl and R scripts by counting the number of genes within each block of 100kb along each chromosome. Gene densities were then attributed to each gene depending on the 100kb window in which they were situated.

As most of the genomic parameters investigated here (gene expression, GC-content, codon bias variance, gene density, average gene length and average intron length) generally differed between groups of genes (see 4.3), it was important to control for their possible influence on divergence, polymorphism and frequencies of deleterious mutations. The six parameters were also inter-correlated, so I decided to implement principal component regression analyses (pcr() command, pls package, version 2.4-3; Mevik and Wehrens, 2007) in order to combine these parameters into independent predictors of the variation in the investigated dependent variable (e.g. dN/dS) as described by Drummond *et al.* (2006). All variables, including the dependent variable, were log transformed (0.0001 was added to gene length and average intron length due to zero values) and scaled to zero mean and unit variance using the scale() function in R. I subsequently performed a jack knife test (jack.test()) on each set of principal component regression results to test if the

contribution of each predictor was significant. Non-significant predictors were then removed and the analyses were repeated. The principal component (PC), which explained the highest amount of variation in the dependent variable, was then used to represent the genomic predictors in an ANCOVA (e.g. lm(log(dN/dS) ~ PC1 * ploidy)) with life-stage as the binary co-variate.

4.2.8 Statistical analyses

All analyses were performed in R (version 3.2.0; Team 2012). To measure statistical difference between groups I used the non-parametric Mann Whitney U test (wilcox.test() function). In case of multiple testing, all p-values were corrected with the Bonferroni method using the function p.adjust(). For correlations either the Spearman rank test (rcorr() function of Hmisc package; version 3.16-0; Jr and others 2015) or Spearman rank partial correlation (pcor.test() function; ppcor package; version 1.0; Kim 2012) was carried out.

4.3 Results

4.3.1 Life-stage limited genes

Within the total data-set, containing 20,839 genes, 4,304 (20.7%) had no reliably detectable expression (score < 0.9; see methods) in any of the analysed tissues and were removed from the analysis. Of the remaining 16,535 genes, 1,552 genes (9.4%) were expressed only in pollen and a further 5,494 (33.2%) were limited to sporophytic tissues (referred to as pollen-specific genes and sporophyte-specific genes in this study). The pollen-specific genes were randomly distributed among the five chromosomes but the sporophyte-specific genes were non-randomly distributed ($\chi^2 = 15.4$, $p = 0.004$, df: 4; table 4.2). Their median positions within the chromosomes did not differ significantly from each other (table 4.3).

Table 4.2: Chi squared test of the distribution of pollen and sporophyte-specific genes among the five nuclear *A. thaliana* chromosomes. Degrees of freedom: 4.

Chromosome	Non pollen genes	Pollen genes	Non sporophyte genes	Sporophyte genes
1	3,928	379	2,840	1,467
2	2,237	236	1,628	845
3	2,958	329	2,254	1,033
4	2,218	207	1,605	820
5	3,516	352	2,635	1,233
Σ	14,857	1,503	10,962	5,398
χ^2		5.609		15.353
p		0.230		0.004

Table 4.3: A test of the median position of pollen and sporophyte-specific genes within each of the 5 nuclear chromosomes. Mann Whitney U test.

Chromosome	W	p
1	2.79×10^5	0.137
2	1.00×10^5	0.071
3	1.72×10^5	0.315
4	8.54×10^4	0.267
5	2.31×10^5	0.241

Expression level was roughly twice as high within pollen-specific genes (mean: 2562.3 ± 86.5) compared to sporophyte-specific genes (mean: 1256.2 ± 23.8; $p = 1.2$

x 10^{-63}; table 4.4). GC-content was significantly higher within sporophyte-specific genes (median: 44.6%) than in pollen-specific genes (median: 43.8%; p = 1.0 x 10^{-19}; table 4.4). Sporophyte-specific genes were significantly longer and contained significantly longer introns than pollen-specific genes (table 4.4). Gene density was slightly but significantly higher in pollen-specific genes. Codon bias variance did not differ significantly (table 4.4).

4.3.2 Pollen-specific proteins evolve at a faster rate than sporophyte-specific proteins

The rate of evolution of *Arabidopsis thaliana* proteins relative to *Arabidopsis lyrata* orthologues was estimated using interspecific dN/dS. Protein divergence was significantly higher for pollen-specific genes than sporophyte-specific genes (p = 4.3 x 10^{-24}; figure 4.1(c). This was mainly due to a significant difference in the non-synonymous divergence rate for which the median was 30.8% higher in pollen-specific genes (dN; p = 2.4 x 10^{-27}; figure 4.1(a)). Synonymous divergence (dS) was only 3.7% higher in pollen-specific genes and the difference was less significant (p = 1.6 x 10^{-4}; figure 4.1(b)). The non-random distribution of sporophyte-specific genes among the five chromosomes was not responsible for the difference in dN/dS, since dN/dS was significantly higher for pollen-specific genes within each chromosome (table 4.6).

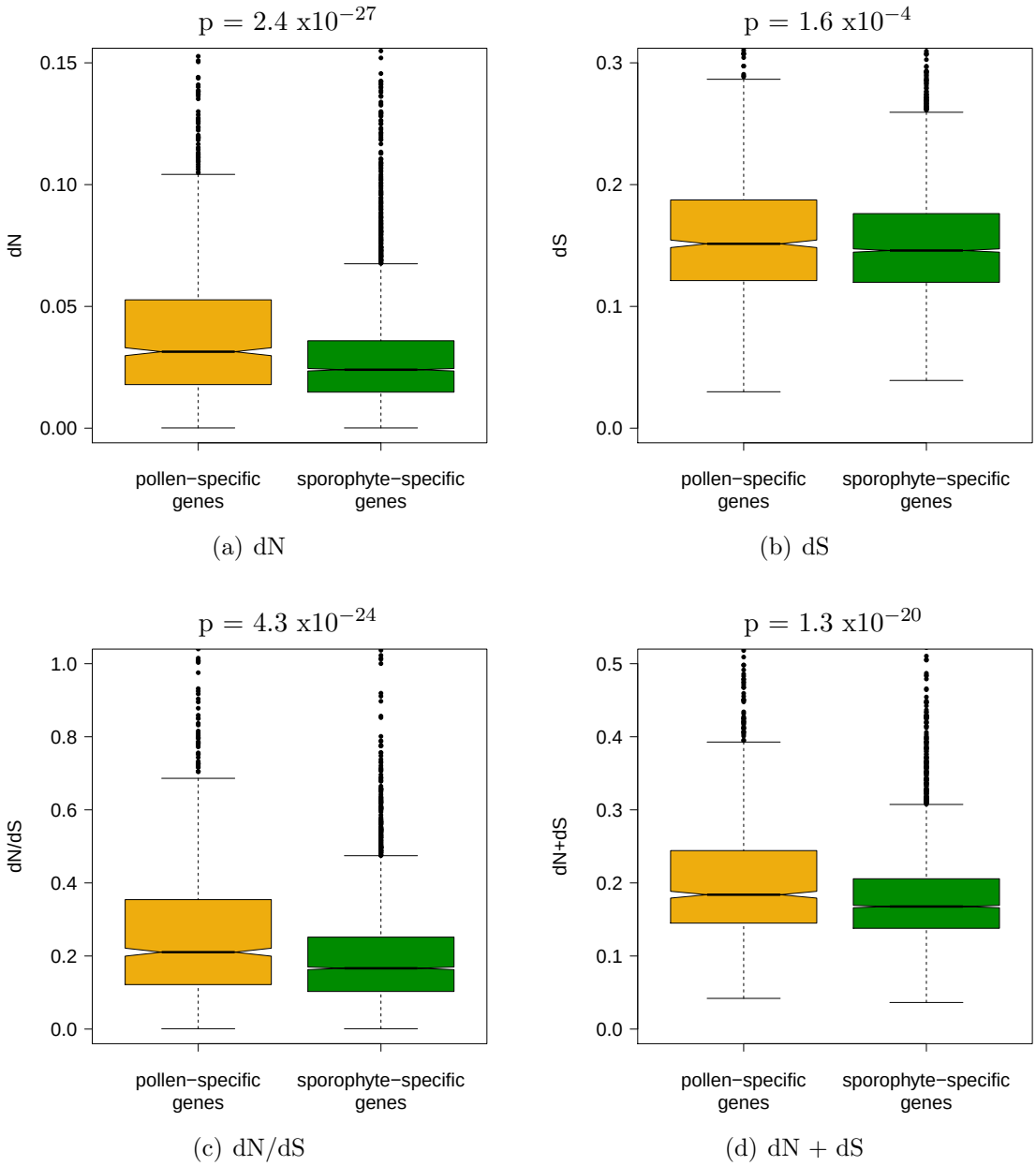
Table 4.4: Differences in six genomic variables between pollen-specific and sporophyte-specific genes. Values are means ± standard error of the mean.

Genomic variable	Pollen-specific genes			Sporophyte-specific genes		p
Expression level	2,562.30	± 86.49	>	1,256.21	± 23.80	1.2×10^{-63}
GC content (%)	44.20	± 0.08	<	45.08	± 0.04	1.0×10^{-19}
Codon bias variance	0.46	± 0.01	=	0.43	± 0.00	not significant
gene length	1,570.30	± 24.41	<	1,634.39	± 11.62	2.3×10^{-4}
average intron length	124.44	± 3.23	<	160.08	± 2.49	8.6×10^{-10}
gene density (per 100kb)	29.99	± 0.12	>	29.57	± 0.07	1.5×10^{-3}

Table 4.5: Partial correlations of 6 genomic variables with dN/dS, θ_n, π_n, frequency of premature stop codons and frameshift mutations. Spearman rank correlations controlling for remaining 5 variables.

Genomic variable	dN/dS		θ_n		π_n		stop codons		frameshifts	
Expression level	-0.232	***	-0.131	***	-0.086	***	not significant		-0.090	***
GC content (%)	-0.145	***	-0.192	***	-0.166	***	-0.180	***	-0.143	***
Codon bias variance	-0.104	***	-0.210	***	-0.161	***	-0.124	***	-0.088	***
gene length	-0.108	***	0.325	***	0.181	***	0.136	***	-0.037	*
average intron length	-0.061	***	-0.191	***	-0.123	***	0.084	***	-0.109	***
gene density (per 100kb)	0.039	*	-0.137	***	-0.116	***	-0.054	***	-0.029	*

*p<0.01; **p<10^{-6}; ***p<10^{-9}



(a) dN (b) dS

(c) dN/dS (d) dN + dS

Figure 4.1: Non-synonymous divergence (a), synonymous divergence (b), dN/dS (c) and total nucleotide divergence (d) within pollen-specific and sporophyte-specific genes between *A. thaliana* and *A. lyrata*.

Table 4.6: Median dN/dS for pollen and sporophyte specific genes by chromosome. See table 4.2 for sample sizes.

Chromosome	Pollen	Sporophyte	p
1	0.208	0.166	8.3×10^{-6}
2	0.220	0.171	1.4×10^{-3}
3	0.227	0.163	2.5×10^{-8}
4	0.189	0.160	0.025
5	0.211	0.168	1.5×10^{-6}

Both expression level (ρ = -0.232; p = 5.6×10^{-169}) and GC-content (ρ = -0.145; p = 4.3×10^{-64}) were significantly negatively correlated with dN/dS while controlling for other factors (codon bias variance, gene length, average intron length and gene density; table 4.5). Codon bias variance, gene length and average intron length also each correlated significantly and negatively with dN/dS, while gene density was weakly, positively correlated (table 4.5).

In order to determine how the life-stage to which the expression of a gene is limited may be contributing to the measured difference in dN/dS, it was important to control for the six previously mentioned genomic variables (expression level, GC-content, codon bias variance, gene length, average intron length and gene density). This was important since five of the six genomic variables differed significantly between pollen- and sporophyte-specific genes (table 4.4) and all six were significantly correlated to dN/dS (table 4.5). I conducted a principal component regression to allow these predictors of dN/dS to be condensed into independent variables. I first included all six predictors in the principal component regression model, and they explained 9.10% of dN/dS variation. Principal component (PC) 2 explained the largest amount of variation in dN/dS at 6.15%. A jack knife test on this PC revealed significant p-values (< 0.05) only for expression, GC content and codon bias variance. After removal of the non-significant predictors (gene length, average intron length and gene density) codon bias variance was also no longer significant. The first PC of a model containing expression and GC content as the predictors of dN/dS had an explanation value of 7.15% (total 7.24%). This first PC was used as the continuous variable in an ANCOVA with dN/dS as the dependent variable and life-stage as the binary co-variable. The pollen regression line was higher than for the sporophyte-specific genes for the majority of the PC1 range (fig. 4.2). As the slopes differed significantly (p = 4.4×10^{-4}), I measured the difference in dN/dS between pollen-specific and sporophyte-specific genes within five equal bins along the PC1 axis. In the first four quantiles dN/dS was significantly higher within pollen-specific

genes, and within the fifth quantile dN/dS was also higher for pollen-specific genes but not significantly (table 4.7).

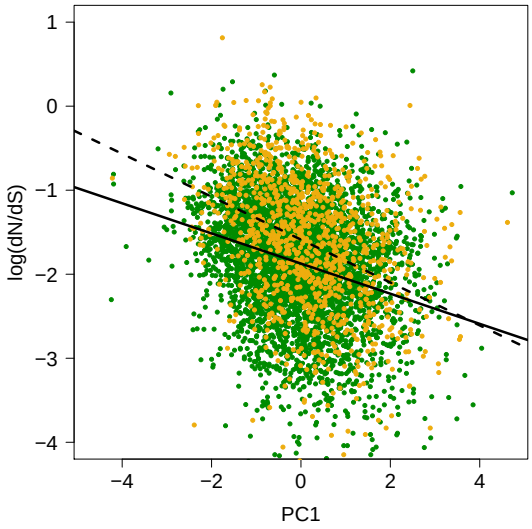


Figure 4.2: ANCOVA analysis of dN/dS within pollen-specific (yellow points and dashed line) and sporophyte-specific genes (green points and solid line) with PC1 as the continuous variable .
Slopes differ significantly; p = 4.4 x10^{-4}.

Table 4.7: dN/dS within five equal bins along the PC1 axis. Shown are means (medians)

	< 20%	20% - 40%	40% - 60%	60% - 80%	> 80%
Pollen	0.269 (0.347)	0.249 (0.346)	0.210 (0.276)	0.173 (0.214)	0.144 (0.198)
Sporophyte	0.220 (0.247)	0.173 (0.362)	0.160 (0.350)	0.146 (0.192)	0.132 (0.160)
p	2.1 x 10^{-23}	7.0 x 10^{-15}	2.5 x 10^{-10}	1.4 x 10^{-3}	non-sig.

4.3.3 Divergence data indicate higher levels of positive selection within pollen-specific genes

I investigated whether the higher divergence of pollen-specific proteins compared to sporophyte-specific proteins was restricted to *Arabidopis*, and possibly fuelled by selection in either *A. thaliana* or *A. lyrata*, by investigating the protein divergence of both from *Capsella rubella*. Divergence was significantly higher for pollen-specific

proteins in all three comparisons (table 4.8). Between branches only one comparison of divergence values differed significantly for sporophyte-specific proteins: *A. thaliana*-*A. lyrata* dN/dS > *A. lyrata*-*C. rubella* dN/dS ($p = 0.046$); all other differences between branches were non-significant.

Table 4.8: dN/dS between *A. thaliana*, *A. lyrata* and *C. rubella*. Values are means (and medians).

	Pollen	Sporophyte	p value
A. thaliana vs. *A. lyrata*	0.2689 (0.2106)	0.1963 (0.1664)	4.3×10^{-24}
A. thaliana vs. *C. rubella*	0.2409 (0.2036)	0.1801 (0.1567)	8.8×10^{-22}
A. lyrata vs. *C. rubella*	0.2370 (0.1945)	0.1818 (0.1568)	1.3×10^{-15}

A higher dN/dS value, which is still lower than 1, generally indicates weaker purifying selection (Yang and Bielawski, 2000). Only 41 out of the total gene set (13,518 genes) had a dN/dS value greater than 1 and 65.1% of genes had a dN/dS less than 0.2. However, gene-wide estimates of dN/dS can be inflated by a few codon sites under positive selection (> 1) even if purifying selection is otherwise prevalent.

In order to test whether the higher dN/dS within pollen-specific genes was driven by relaxed purifying selection or increased positive selection I first conducted an analysis to investigate levels of positive selection. For the calculation of divergence on a multi-sequence alignment (*A. thaliana*, *A. lyrata* and *C. rubella*), I allowed dN/dS to vary among sites in order to detect sites under positive selection using codeml in PAML (Yang, 2007). This analysis suggested a much higher proportion of pollen-specific genes contained sites under positive selection (15.6%) compared to sporophyte-specific genes (9.5%). This difference is significant ($p = 3.0 \times 10^{-7}$, Fisher's exact test). As expected, dN/dS was significantly higher within the genes containing sites under positive selection compared to genes with no evidence for positive selection (median of 0.338 compared to 0.179 for pollen-specific genes, $p = 3.8 \times 10^{-21}$; 0.228 compared to 0.154 in sporophyte-specific genes, 3.9×10^{-24}). It appears, therefore, that at least in part, the difference in dN/dS may be caused by a higher rate of adaptive fixations in pollen-specific genes.

4.3.4 Polymorphism data indicate weaker purifying selection within pollen-specific genes

An analysis of the distribution of fitness effects of new mutations using the DoFE software (DoFE 3.0; Eyre-Walker and Keightley 2009) produced findings which contradict those based on divergence data (section 4.3.3). The distribution of fitness

effects showed that a smaller fraction of non-synonymous mutations were strongly deleterious (selection strength $N_e s > 10$; where N_e = effective population size & s = selection coefficient) within pollen-specific genes (36.7%) compared to sporophyte-specific genes (51.0%; figure 4.3). Also, a higher proportion of mutations within pollen-specific genes were effectively neutral ($N_e s < 1$: 40.2%) or slightly deleterious ($1 < N_e s < 10$: 23.1%) compared to sporophyte-specific genes (35.8% & 13.2%; figure 4.3). This indicates weaker purifying selection within the pollen-specific genes (Eyre-Walker and Keightley, 2009).

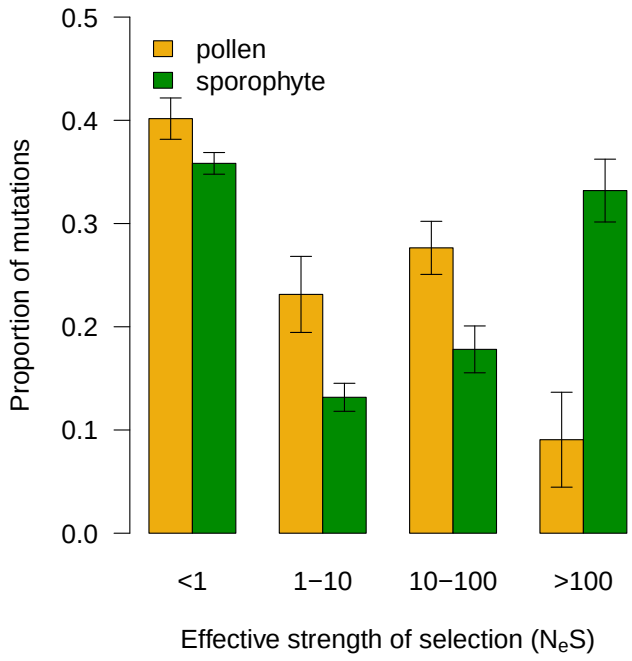
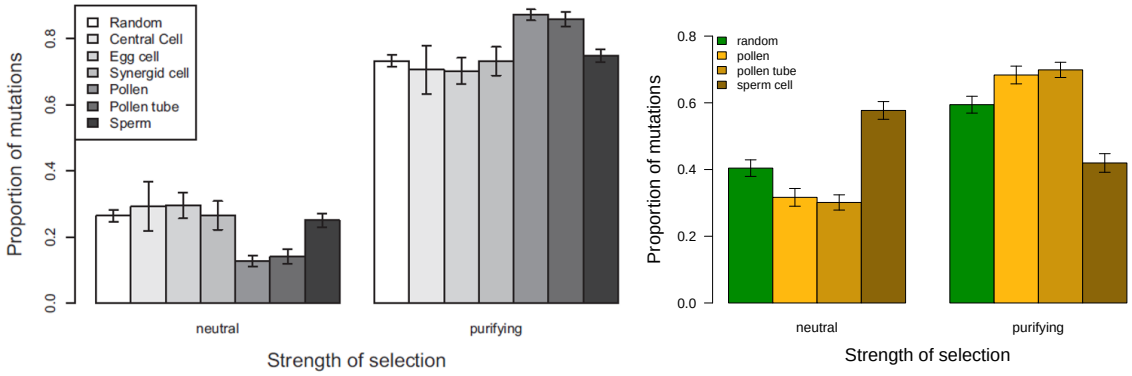


Figure 4.3: Distribution of fitness effects for pollen and sporophyte-specific genes. Shown are the mean proportions of mutations in four $N_e s$ ranges with SDs

As these values differed quite strongly from those presented by (Gossmann *et al.* 2013; figure 4.4(a)), I repeated the analysis with the random gene set and three male gene sets within the 80 accessions from that study. The proportion of mutations, which were effectively neutral, was higher than in the original study: For instance, the proportion of effectively neutral mutations was 30.1% in pollen tube genes and 31.7% in pollen genes (figure 4.4(b)), compared to 14 and 13% in the original study (figure 4.4(a)). However, the relative relationships among the gene groups were quite similar to the original study: the highest proportion of neutral sites was found within sperm cell genes, while pollen and pollen tube genes were very similar. Also here as in the original study, the value for random genes was higher than pollen and

pollen tube genes (figure 4.4).



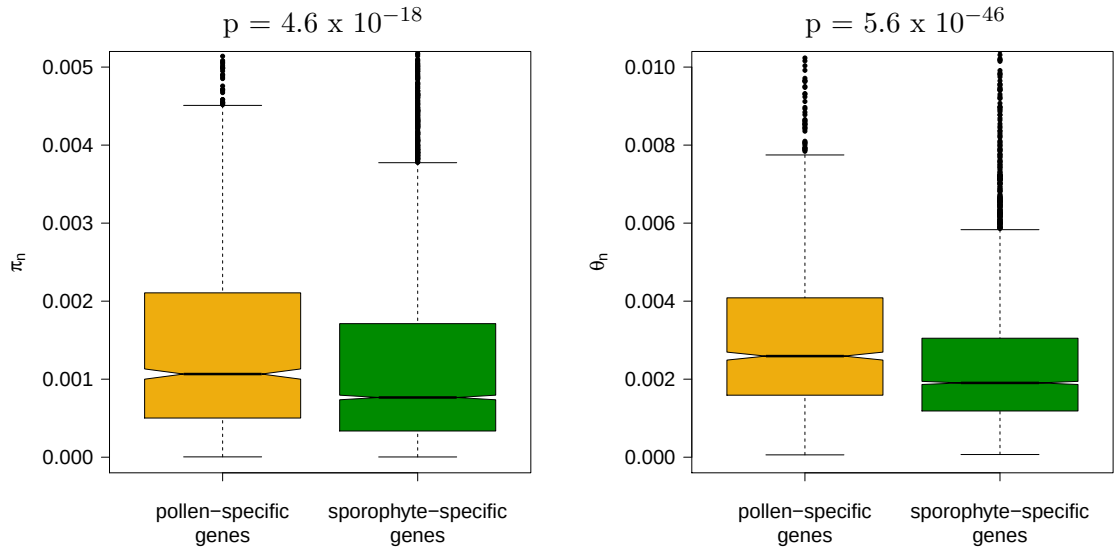
(a) (b)

Figure 4.4: Distribution of fitness effects for gene lists analysed in Gossmann *et al.* (2013). The original figure (a) contains seven gene groups; (b) shows results for the random gene set and the three male gene sets when compiled with the methods in the current study.
Shown are the mean proportions of mutations with SDs

By combining polymorphism frequencies and divergence rates the same DoFE software produces an estimation of α (proportion of sites under positive selection). With this method I was unable to find evidence for positive selection for either group of genes since α was not significantly greater than zero (mean: -0.8 in pollen-specific; -1.2 in sporophyte-specific genes).

4.3.5 Pollen-specific genes are more polymorphic than sporophyte-specific genes

Pollen-specific genes were more polymorphic than sporophyte-specific genes with both non-synonymous nucleotide diversity (π_n) and non-synonymous Watterson's theta (θ_n) significantly higher in pollen-specific genes (fig. 4.5). The two polymorphism measures (π_n and θ_n) were significantly higher for pollen-specific genes within each of the five chromosomes (tables 4.9 and 4.10).



(a) (b)

Figure 4.5: Non-synonymous nucleotide diversity (a) and non-synonymous Watterson's theta (b) among pollen-specific and sporophyte-specific genes.

Table 4.9: Median π_n for pollen and sporophyte-specific genes by chromosome. See table 4.2 for sample sizes.

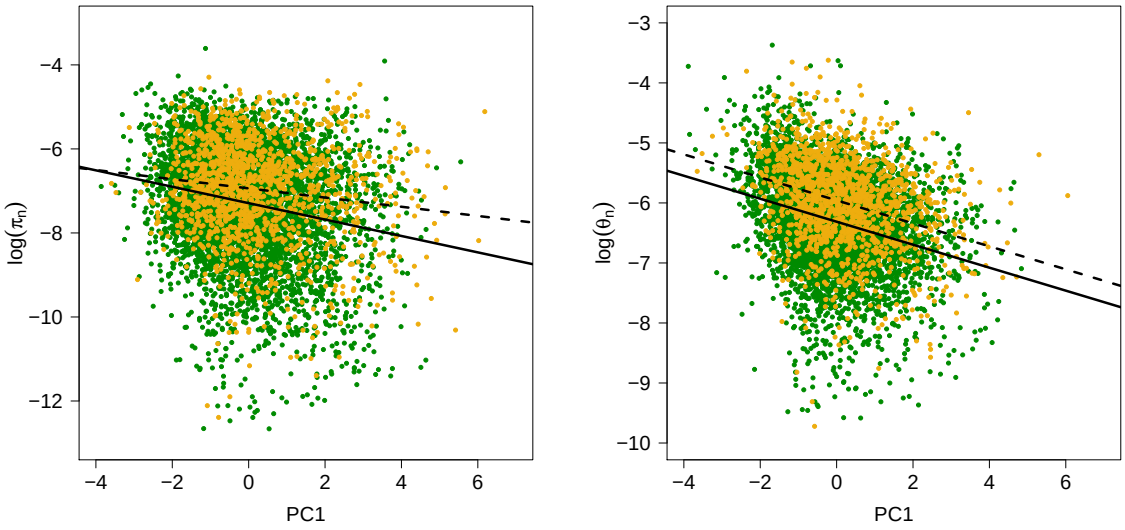
Chromosome	Pollen	Sporophyte	p
1	9.5×10^{-4}	8.2×10^{-4}	6.4×10^{-3}
2	9.5×10^{-4}	6.9×10^{-4}	1.3×10^{-3}
3	1.1×10^{-3}	6.8×10^{-4}	1.4×10^{-6}
4	1.2×10^{-3}	8.6×10^{-4}	8.1×10^{-4}
5	1.2×10^{-3}	8.4×10^{-4}	4.2×10^{-4}

Table 4.10: Median θ_n for pollen and sporophyte-specific genes by chromosome. See table 4.2 for sample sizes.

Chromosome	Pollen	Sporophyte	p
1	2.6×10^{-3}	2.0×10^{-3}	5.7×10^{-8}
2	2.3×10^{-3}	1.8×10^{-3}	1.0×10^{-6}
3	2.8×10^{-3}	1.8×10^{-3}	1.3×10^{-15}
4	2.8×10^{-3}	2.1×10^{-3}	5.1×10^{-8}
5	2.5×10^{-3}	1.8×10^{-3}	4.1×10^{-12}

Each of the six correlates of dN/dS listed above also correlated significantly with π_n and θ_n (all negatively except gene length; table 4.5). Five of the six variables (average intron length was not significant) explained 8.57% of variation in π_n in a

principal component regression. The first PC contributed most (3.11%). Four of the six factors (expression level, GC content, codon bias variance, and gene density) explained a total of 7.76% of the variation in θ_n (first PC: 7.38%). For each model the first PC was implemented in an ANCOVA testing the influence of life-stage as a co-variate. θ_n remained significantly higher for pollen-specific genes (p = 6.4 x 10^{-61}; fig. 4.6(b)). PC1 had a significantly greater influence on π_n for sporophyte-specific genes (slope: -0.195) than for pollen-specific genes (slope: -0.109; p = 7.2 x 10^{-4}; figure 4.6(a)). I therefore tested the significance of difference in π_n within five equal bins along the PC1 axis. In the 2nd to the 5th 20% quantiles π_n was significantly higher within pollen-specific genes, there was no difference in the first quantile (table 4.11).



(a) (b)

Figure 4.6: ANCOVA analysis of π_n (a) and θ_n (b) within pollen-specific (yellow points and dashed line) and sporophyte-specific genes (green points and solid line) with PC1 as the continuous variable.
Slopes significantly differ in (a); p = 7.2 x 10^{-4}.
Regression line is significantly higher for pollen in (b); p = 6.4 x 10^{-61}

Table 4.11: Nonsynonymous π within five equal bins along the PC1 axis. Shown are medians (means).

	< 20%	20% - 40%	40% - 60%	60% - 80%	> 80%
Pollen	1.0×10^{-3} (1.7×10^{-3})	1.1×10^{-3} (1.7×10^{-3})	1.1×10^{-3} (1.7×10^{-3})	1.1×10^{-3} (1.6×10^{-3})	8.4×10^{-4} (1.5×10^{-3})
Sporophyte	1.0×10^{-3} (1.7×10^{-3})	8.6×10^{-4} (1.4×10^{-3})	7.2×10^{-4} (1.2×10^{-3})	6.7×10^{-4} (1.1×10^{-3})	6.0×10^{-4} (1.0×10^{-3})
p	non-sig.	1.1×10^{-3}	1.1×10^{-8}	5.8×10^{-6}	1.0×10^{-5}

4.3.6 Higher frequency of deleterious mutations in pollen-specific genes

Higher polymorphism levels may indicate relaxed purifying selection on pollen-specific genes. To test this hypothesis further I investigated the frequency of putatively deleterious mutations - premature stop codons and frameshift mutations - within the 269 *A. thaliana* strains. Stop codon frequency, defined here as the relative number of unique alternative alleles due to premature stop codons occurring within the 269 strains, was significantly higher within pollen-specific genes (mean: 0.063 $\pm$0.004; sporophyte mean: 0.049 $\pm$0.002; $p = 4.1 \times 10^{-15}$; fig. 4.7). The frequency of strains containing at least one frameshift mutation was also significantly higher for pollen-specific genes (mean: 0.021 $\pm$0.002) compared to sporophyte-specific genes (mean: 0.014 $\pm$0.001; $p = 6.6 \times 10^{-22}$; fig. 4.7). The frequency of frameshift mutations was significantly higher for pollen-specific genes than for sporophyte-specific genes within each of the five chromosomes (table 4.13). Stop codon mutations were significantly higher for pollen-specific genes within chromosomes 2 to 5 but did not significantly differ on chromosome 1 (table 4.12).

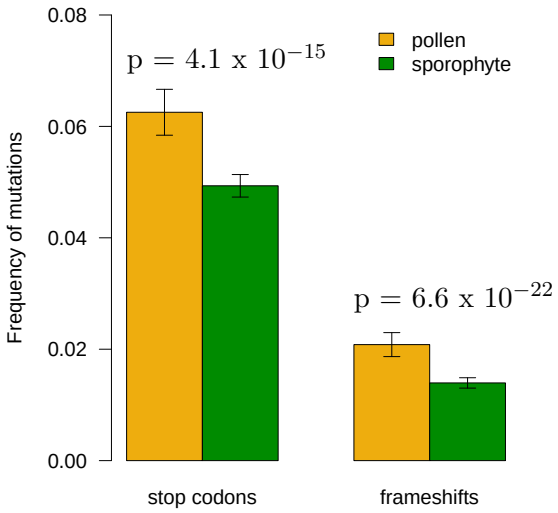


Figure 4.7: Frequency of alleles containing premature stop codon mutations and frameshift mutations in pollen-specific and sporophyte-specific genes.

Table 4.12: Mean frequency of premature stop codons for pollen and sporophyte-specific genes by chromosome.
See table 4.2 for sample sizes.

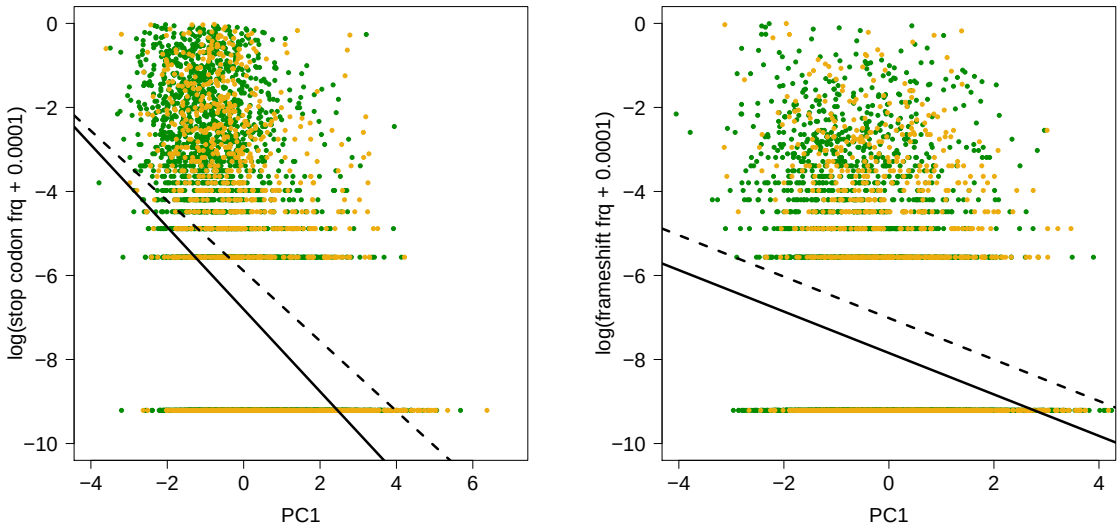
Chromosome	Pollen	Sporophyte	p
1	0.047	0.052	0.081
2	0.062	0.043	0.028
3	0.062	0.049	2.5×10^{-5}
4	0.087	0.052	1.7×10^{-3}
5	0.065	0.049	5.5×10^{-5}

Table 4.13: Mean frequency of frameshift mutations for pollen and sporophyte-specific genes by chromosome.
See table 4.2 for sample sizes.

Chromosome	Pollen	Sporophyte	p
1	0.023	0.014	5.7×10^{-7}
2	0.021	0.013	4.4×10^{-5}
3	0.021	0.013	3.7×10^{-7}
4	0.023	0.016	0.045
5	0.017	0.014	1.3×10^{-3}

Significant correlations existed between these measures of deleterious mutations and the six correlates of dN/dS (table 4.5). In a principal component regression

analysis all six predictors (expression level, codon bias variance, GC content, gene length, average intron length and gene density) were significantly correlated with stop codon frequency. The six predictors explained a total of 20.04% of the variation in stop codon frequency, 17.42% explained by the first PC. Within an ANCOVA with life-stage as the binary co-variant the frequency of premature stop codons remained higher within pollen-specific genes for the majority of PC1 (fig. 4.8(a)). The slopes differed significantly but the frequency of stop codons was significantly higher for pollen genes within the second to fifth 20% quantiles (table 4.14).



(a) (b)

Figure 4.8: ANCOVA analysis of stop codon frequency (a) and the frequency of frameshift mutations (b) within pollen-specific (yellow points and dashed line) and sporophyte-specific genes (green points and solid line) with PC1 as the continuous variable.
Slopes significantly differ in (a); p = 0.014.
Regression line is significantly higher for pollen in (b); p = 2.5 x 10^{-29}.

Table 4.14: Frequency of stop codons within five equal bins along the PC1 axis. Shown are means (medians).

	< 20%	20% - 40%	40% - 60%	60% - 80%	> 80%
Pollen	0.113	0.111	0.056	0.033	0.015
	(0.028)	(0.011)	(0.004)	(0)	(0)
Sporophyte	0.106	0.063	0.045	0.022	0.006
	(0.015)	(0.004)	(0)	(0)	(0)
p	non-sig.	5.7 x 10^{-6}	1.1 x 10^{-5}	6.3 x 10^{-8}	8.3 x 10^{-11}

Four of the predictors (expression level, GC content, gene length and gene density) were also significantly correlated with the frequency of frameshift mutations in a principal components regression. The four variables explained a total of 5.49% of variation (first PC 5.08%). In an ANCOVA analysis frameshift mutations remained significantly more frequent within pollen-specific genes when controlling for the predictors via the first PC (fig. 4.8(b)).

4.3.7 Tissue-specific genes

Tissue specificity has been shown to be strongly positively correlated with the evolutionary rates of proteins (Duret and Mouchiroud, 2000; Liao *et al.*, 2006). The average greater tissue specificity in pollen-specific genes compared to sporophyte-specific genes could therefore explain the higher dN/dS values found in pollen-specific genes. In order to control for this potential bias I compared dN/dS in pollen-specific genes with a group of 340 genes with expression limited to a single sporophyte cell type (guard cell, xylem or root hair). To further test for the effect of tissue specificity, these groups were also compared to 2,543 genes which were expressed in at least five sporophytic tissues.

In this tissue-specificity controlled comparison, dN/dS did not differ between pollen-specific genes and the tissue-specific sporophyte gene set. However, dN/dS was significantly higher in pollen-specific genes ($p = 1.7 \times 10^{-27}$) and tissue-specific sporophyte genes ($p = 1.0 \times 10^{-9}$; fig. 4.9) compared to broadly expressed sporophyte-specific genes. In a principal components regression only expression level and GC content had a significant effect on dN/dS, explaining 8.63% of variation. The PC1 (8.60%) was then mapped against dN/dS in an ANCOVA on the two levels of pollen-specific genes and tissue-specific, sporophytic genes. dN/dS was significantly higher for pollen-specific genes than tissue-specific, sporophytic genes when controlling for PC1 ($p = 1.4 \times 10^{-3}$; figure 4.10).

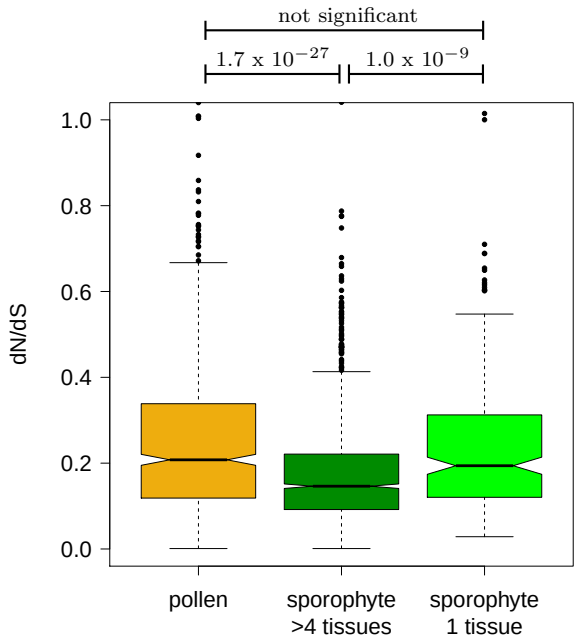


Figure 4.9: dN/dS within pollen-specific genes, broadly expressed sporophytic genes (at least five tissues) and tissue specific genes (expression restricted to guard cell, xylem or root hair tissues).

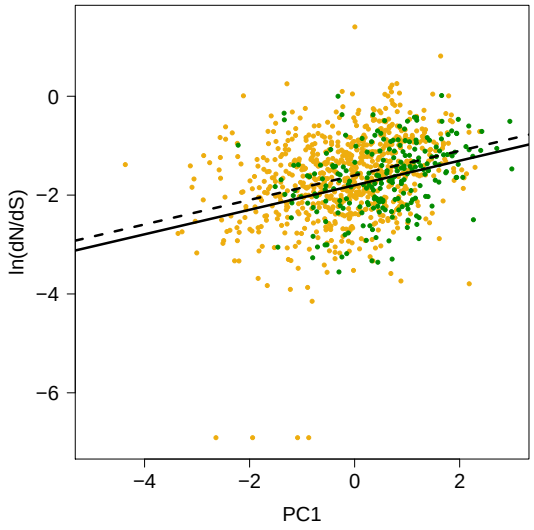


Figure 4.10: ANCOVA analysis of dN/dS within pollen-specific (yellow points and dashed line) and tissue-specific, sporophyte genes (green points and solid line) with PC1 as the continuous variable.
The pollen regression line was significantly higher; $p = 1.4 \times 10^{-3}$.

It has been suggested that selection pressure is generally weaker on tissue-specific genes compared to broadly expressed genes (Duret and Mouchiroud, 2000). In order to test whether the high tissue-specificity of pollen genes is causing the higher

polymorphism levels and higher frequency of deleterious mutations, I repeated the tissue-specific analyses described above for π_n, θ_n, the frequency of stop codons and the frequency of frameshift mutations. Similarly, π_n and θ_n did not differ between pollen-specific and the tissue-specific sporophyte gene set. However, they were both significantly higher in pollen-specific genes ($p = 1.6 \times 10^{-30}$ & 8.4×10^{-75}, respectively) and tissue-specific sporophyte genes ($p = 7.1 \times 10^{-13}$ & 2.7×10^{-26}; fig. 4.11) compared to broadly expressed sporophyte-specific genes. In a principal components regression, expression level and GC content had a significant effect on π_n, explaining 5.30% of variation. The first PC (5.06%) was plotted against π_n in an ANCOVA within pollen-specific and tissue-specific sporophytic genes. π_n was significantly higher for pollen-specific genes compared to tissue-specific, sporophytic genes when controlling for PC1 ($p = 6.5 \times 10^{-3}$; figure 4.12). In a similar analysis for θ_n, all six parameters (expression level, GC content, codon bias variance, gene length, average intron length and gene density) significantly contributed to 18.82% of variation. The second PC was largest (9.55%) and was plotted against θ_n in an ANCOVA (figure 4.12). When controlling for PC2 θ_n was significantly higher within pollen-specific genes ($p = 2.9 \times 10^{-7}$) compared to tissue-specific, sporophytic genes.

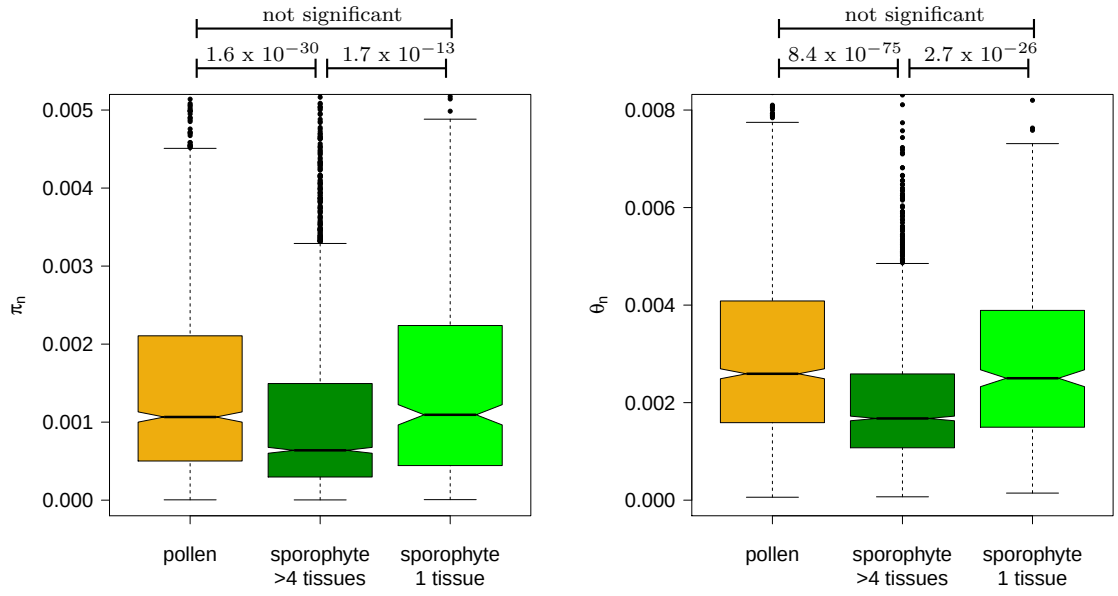
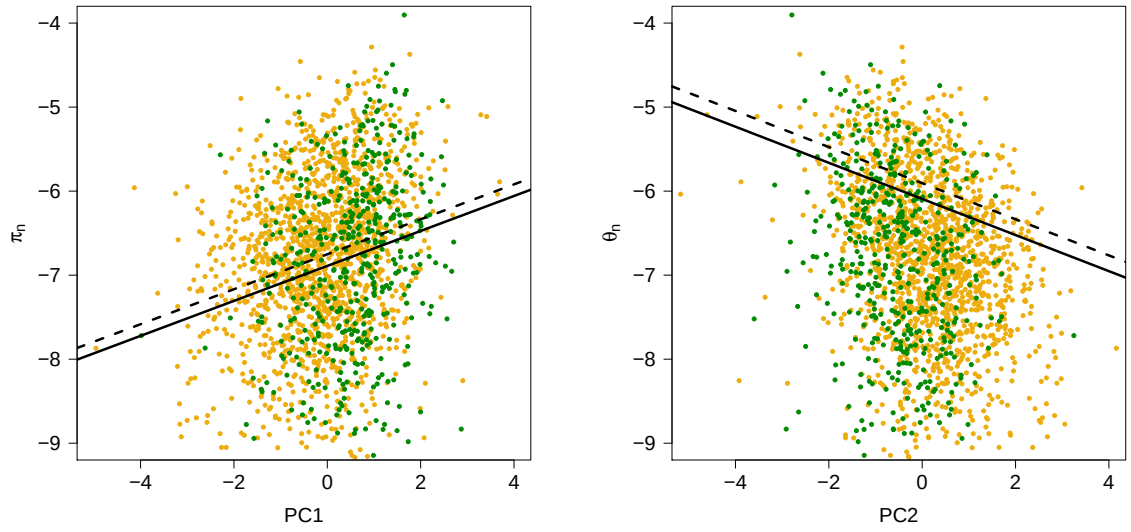


Figure 4.11: Non-synonymous nucleotide diversity (a) and non-synonymous Watterson's theta (b) within pollen-specific genes, broadly expressed sporophyte-specific genes and genes specific to guard cells, xylem or root hair.



(a) (b)

Figure 4.12: ANCOVAs comparing π_n (a) and θ_n (b) within pollen-limited genes (yellow points and dashed line) to tissue-specific, sporophytic genes (green points and solid line) while controlling for the largest PC of a principal components regression. π_n and θ_n were significantly higher within pollen-specific genes; p = 6.5 x 10^{-3} & 2.9 x 10^{-7} respectively.

Premature stop codons remained significantly more frequent in pollen-specific genes than in sporophytic, tissue-specific genes (p = 0.033), and broadly expressed, sporophytic genes (p = 3.0 x 10^{-14}; fig. 4.13). Premature stop codons were not significantly more frequent in tissue-specific, sporophytic tissues compared to broadly expressed sporophytic genes. There was no significant difference in the frequency of frameshift mutations between pollen-specific genes and tissue-specific, sporophytic genes but the frequency was significantly higher in both groups compared to broadly expressed, sporophytic genes (p = 2.0 x 10^{-34} & 1.7 x 10^{-14}; figure 4.13). In a principal components regression GC content, codon bias variance, gene length, average intron length and gene density had a significant effect on 19.03% variation in the frequency of stop codons. The first PC was largest (16.44%) and was implemented in an ANCOVA as the continuous variable (figure 4.14). Due to a significant interaction between the two independent variables, I measured differences in stop codon frequencies within five equal bins along the PC1 axis. The frequency of stop codon mutations did not differ significantly within the first four quantiles but was significantly higher within pollen-specific genes in the fifth quantile (PC1 > 1.14; p = 5.7 x 10^{-5}; table 4.15). The analysis was repeated for the frequency of frameshift mutations. Expression level, GC content, codon bias variance and gene length explained a total of 6.35% variation. PC2 was largest with 3.24% so was implemented in

an ANCOVA. The frequency of frameshift mutations was significantly higher within pollen-specific genes compared to tissue-specific, sporophytic genes when controlling for PC2 ($p = 0.017$; figure 4.14).

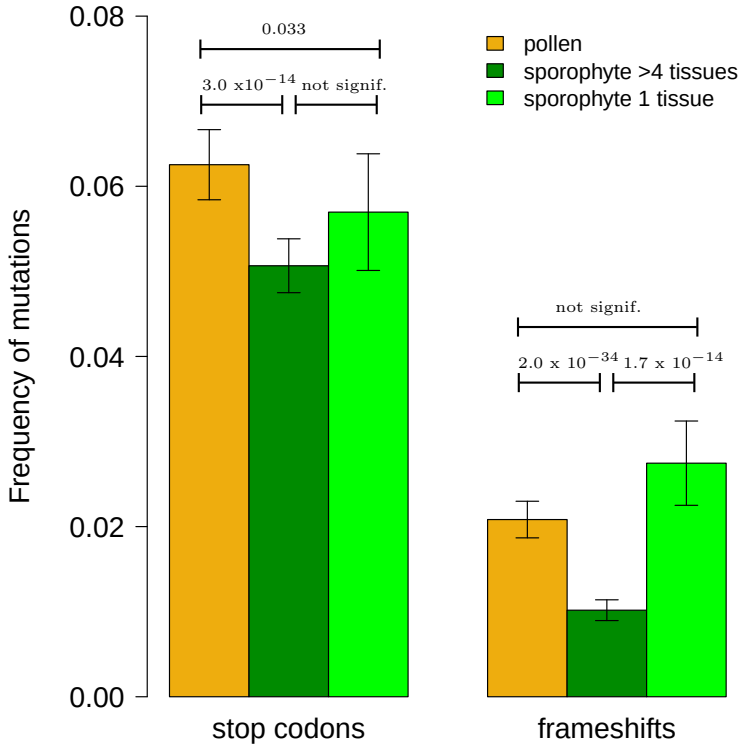
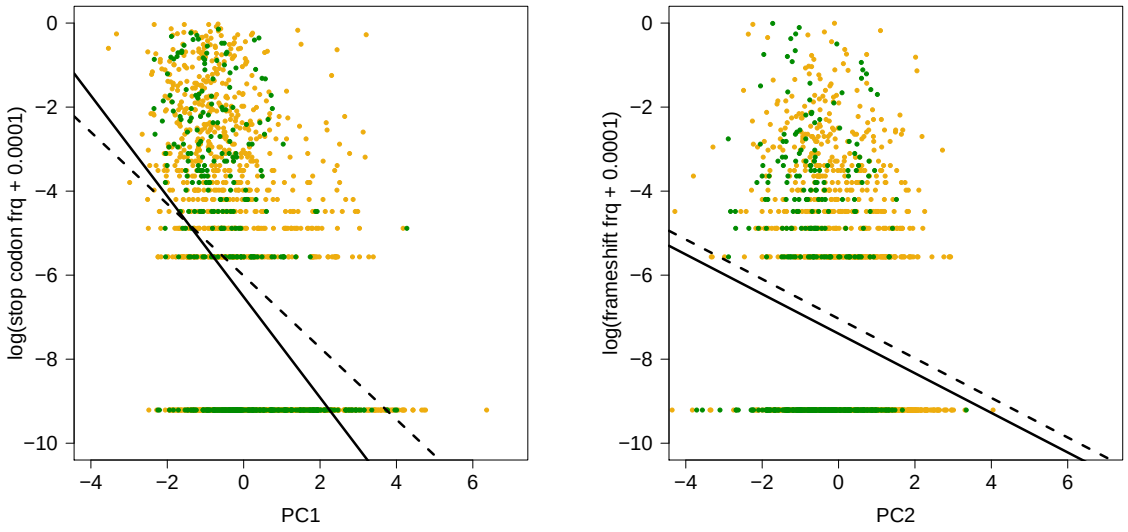


Figure 4.13: Frequency of stop codon and frameshift mutations within pollen-specific genes, broadly expressed sporophytic genes (at least five tissues) and tissue-specific genes (expression restricted to guard cell, xylem or root hair tissues;).



(a) (b)

Figure 4.14: ANCOVAs comparing the frequency of stop codon mutations (a) and frameshift mutations (b) within pollen-limited genes (yellow point and dashed line) to tissue-specific, sporophytic genes (green points and solid line) while controlling for the largest PC of a principal components regression.
Slopes significantly differ in (a); p = 0.004.
Regression line is significantly higher for pollen in (b); p = 0.017.

Table 4.15: Stop codon frequency within five equal bins along the PC1 axis for pollen-specific and sporophytic, tissues-specific genes. Shown are means (medians)

	< 20%	20% - 40%	40% - 60%	60% - 80%	> 80%
Pollen	0.110 (0.026)	0.103 (0.007)	0.050 (0.004)	0.035 (0)	0.016 (0)
Sporophyte, tissue-specific	0.136 (0.045)	0.076 (0.009)	0.038 (0)	0.036 (0)	4.8×10^{-4} (0)
p	non-sig.	non-sig.	non-sig.	non-sig.	5.7×10^{-5}

4.4 Discussion

In this study I have shown that protein divergence, polymorphism levels and the frequency of deleterious mutations are significantly higher within pollen-specific genes compared to sporophyte-specific genes. These differences remained when controlling for expression level, GC content, codon bias variance, gene length, average intron length and gene density.

4.4.1 Evolutionary rates higher within pollen-specific genes

Protein divergence rates (dN/dS) were on average 37% higher in pollen-specific genes compared to sporophyte-specific genes. This is comparable to the findings presented by Szövényi *et al.* (2013), who found dN/dS to be 39% or 81% higher in pollen genes for *A. thaliana* depending on the data set. In a further paper, no difference in dN/dS could be found between pollen-specific and non-reproductive genes for *A. thaliana* (Gossmann *et al.*, 2013). This discrepancy was most likely caused by the method of gene selection. In the Szövényi *et al.* (2013) study, as in the current study, genes with exclusive expression within sporophytic or pollen tissues were analysed. In the Gossmann *et al.* (2013) paper, on the other hand, genes were selected more inclusively, labelling a gene as pollen-enriched if expression was significantly higher at a fold change greater than 4 within different comparisons. It appears then that the exclusivity of expression to either pollen or sporophytic tissues, rather than just expression bias, may in part be responsible for the observed difference in dN/dS rates.

In fact, the difference in rates of protein evolution between pollen-specific and sporophyte-specific genes could be explained to a large extent by tissue specificity. dN/dS did not differ between pollen-specific and tissue-specific, sporophytic genes (expression limited to xylem, guard cell or root hair). However, dN/dS was significantly higher in both gene groups compared to broadly expressed ($>$ 4 tissues), sporophytic genes. This suggests that the specificity of pollen genes to a small set of tissues may be responsible for their elevated rates of protein evolution rather than their specific association with pollen tissues. However, this only partially explains the raised dN/dS levels, as, when controlling for differences in expression level and GC content, dN/dS remained significantly higher within pollen-specific genes.

A higher dN/dS rate can either be caused by greater positive selection or relaxed purifying selection. The higher dN/dS in pollen-specific genes in the current study can, in part, be explained by stronger positive selection acting on pollen-specific genes compared to sporophyte-specific genes. This is revealed by a greater

proportion of pollen-specific genes containing sites under positive selection (15.2% compared to 9.3%). An analysis of the distribution of fitness effects of new nonsynonymous mutations using the DoFE software (Eyre-Walker and Keightley, 2009), on the other hand, suggests purifying selection may be more relaxed within pollen-specific genes compared to sporophyte-specific genes. A higher proportion of nonsynonymous mutations are effectively neutral or slightly deleterious within pollen-specific genes, suggesting the higher dN/dS rate within pollen-specific genes may have been caused by a greater proportion of slightly deleterious substitutions due to random drift. Within the same software via a method, which combines the distribution of fitness effects with divergence data, a significant α (proportion of adaptive substitutions) could not be found for pollen- or sporophyte-specific genes.

A similar, recent analysis of the distribution of fitness effects for different groups of reproductive and non-reproductive genes in *A. thaliana*, however, found higher levels of purifying and positive selection in pollen genes compared to female genes and non-reproductive genes (Gossmann *et al.*, 2013). This again was likely due to their more inclusive choice of genes, as I achieved a similar pattern when repeating the analyses on the specific gene sets and within the same accessions used within their study (Gossmann *et al.*, 2013). The generally higher levels of effectively neutral sites I estimated in this study compared to the Gossmann *et al.* study possibly indicate the sensitivity of the software to how input data are prepared. In the current study, synonymous and nonsynonymous sites were represented by 4 and 0-fold sites and only polymorphisms with a phred score of at least 25 were included; the method implemented by Gossmann *et al.* is unclear. This may be leading to a systematic over-representation of nonsynonymous polymorphisms. There is, however, no indication that any systematic bias should affect pollen- and sporophyte-specific genes differently. This is shown by the relationships between groups from the Gossmann *et al.* (2013) study being well maintained in the repeat analysis presented here. Therefore, the absolute level of effectively neutral sites may not be as high as my results suggest but the level seems to be higher within pollen-specific compared to sporophyte-specific genes.

4.4.2 Polymorphism levels suggest relaxed selection on pollen-specific genes

Polymorphism levels were significantly higher within pollen-specific genes. Both Watterson's θ and π of non-synonymous sites remained significantly higher within pollen-specific genes when controlling for expression and five further genomic differences (GC content, codon bias variance, gene length, average intron length and

gene density). In one of two recent studies higher polymorphism rates were also found in pollen-specific genes for *A. thaliana* (Szövényi *et al.*, 2013). In the second study, however, no difference was found between pollen-specific genes in general and random, non-reproductive genes in terms of nucleotide diversity (Gossmann *et al.*, 2013), which, as discussed in the previous section (4.4.1), is possibly due to their more inclusive choice of genes.

I also found significantly higher levels of putatively deleterious alleles (premature stop codons and frameshift mutations) within pollen-specific genes. This supports the conclusions of Szövényi *et al.* (2013) that the raised polymorphism levels indicate relaxed purifying selection on pollen-specific genes. In other words, comparatively weaker selective constraints are allowing deleterious alleles to accumulate at a greater rate within pollen-specific genes compared to those whose expression is restricted to the sporophyte.

4.4.3 Evidence for a recent shift in selection strength

The patterns in my data are compatible with a change in selection efficacy that is likely to have taken place since *A. thaliana* became self-compatible. The relatively recent switch from self-incompatibility to self-compatability in *A. thaliana* (ca. 1MYA; Tang *et al.* 2007) explains why I have observed evidence for relaxed selection in polymorphism levels but stronger positive selection in divergence data for pollen-specific genes. The divergence data, used to calculate dN/dS and the number of genes containing sites under positive selection, mainly represent a prolonged period of outcrossing ($\sim$ 12 million years), since the speciation of *A. thaliana* from *A. lyrata* occurred roughly 13 million years ago (Beilstein *et al.*, 2010). This would explain the similarity of the dN/dS values recorded for *A. thaliana* in both pollen and sporophyte-specific genes to those estimated between obligatory outcrossers *A. lyrata* and *Capsella rubella*, reported in this study. In all comparisons dN/dS was significantly higher in pollen-specific genes than in sporophyte-specific genes. Divergence data suggest the higher dN/dS in pollen-specific genes is caused by stronger positive selection. These findings are similar to the protein divergence rates reported for the outcrossing *Capsella grandiflora* (Arunkumar *et al.*, 2013), for which higher dN/dS levels as well as more efficient purifying and positive selection were reported for pollen genes. In contrast, the polymorphism data and frequencies of putative deleterious alleles reflect a recent relaxation of selection due to high selfing rates. This may also explain why more relaxed purifying selection was discovered for pollen-specific genes by the DoFE analysis, since it also relies on polymorphism data.

4.4.4 Why is current selection weaker on pollen genes?

The more efficient purifying and adaptive selection on pollen genes found for *C. grandiflora* was linked to two possible factors: haploid expression and pollen competition (Arunkumar *et al.*, 2013). *A. thaliana* is a highly self-fertilising species with selfing rates generally in the range of 95 - 99% (Platt *et al.*, 2010), so haploid expression is unlikely to improve the efficacy of selection on pollen-specific genes relative to sporophyte genes. This is because most individuals found in natural populations are homozygous for the majority of loci, reducing the likelihood that deleterious alleles are masked in the heterozygous state when expressed in a diploid tissue (Platt *et al.*, 2010). A reduction in pollen competition can also be expected due to the probably limited number of pollen genotypes that compete for pollination on single stigmata in highly selfing populations (Charlesworth and Charlesworth, 1992; Mazer *et al.*, 2010).

So if I assume both masking and pollen competition are negligible forces when comparing selection on pollen-specific genes to sporophyte-specific genes, why is selection more relaxed on pollen-specific genes than sporophyte-specific genes rather than similar?

I have shown here that tissue specificity partly explains why selection is more relaxed on pollen genes. The full set of sporophyte-specific genes contains genes expressed across several tissues, and broadly expressed genes have been known to be under more efficient selection than tissue-specific genes due to their exposure to a higher number of selective constraints (Duret and Mouchiroud, 2000). Both pollen-specific genes and genes limited to one of three sporophytic tissues (xylem, guard cell or root hair) showed raised levels of dN/dS, polymorphism and frequency of deleterious mutations compared to broadly expressed sporophyte-specific genes (expressed in at least five tissues). Tissue specificity appeared to explain, to a certain extent, the reduced selection efficacy in pollen-specific genes as there was no longer a significant difference in polymorphism levels (θ_n and π_n) or the frequency of frameshift mutations in pollen-specific genes compared to the tissue-specific, sporophytic genes (the frequency of stop codon mutations remained significantly higher). However, tissue specificity alone only partly explains the apparent relaxation of selection within pollen-specific genes. Once further genomic features (expression level, GC content, codon bias variance, gene length, average intron length, gene density) were controlled for, all measures remained higher in pollen-specific genes even when compared to genes restricted to only one sporophytic tissue, except for stop codon frequency.

Recent similar findings indicating relaxed purifying selection in pollen-specific

genes in *A. thaliana* (Szövényi *et al.*, 2013) were explained as possibly resulting from a combination of high tissue specificity and higher expression noise (variance in expression level) in pollen compared to sporophytic genes. However, the authors did not compare selection on pollen genes to tissue-specific sporophyte genes to isolate the effect of tissue specificity. I have shown here that tissue specificity does appear to play an important role but does not alone explain the difference in selection strength between both groups of genes. Higher expression noise could then be an important factor influencing the level of deleterious alleles which exist for pollen genes in *A. thaliana*. Expression noise has been found to reduce the efficacy of selection substantially and is expected to be considerably higher for haploid expressed genes (Wang and Zhang, 2011). Therefore, given that pollen competition and masking are likely to be negligible forces, it is possible that expression noise and high tissue specificity become dominant factors for pollen-specific genes of selfing plants.

4.4.5 Conclusions

The findings presented in this study offer an explanation for seemingly contradictory results presented in two recent publications on selection in pollen and sporophytic tissues (Szövényi *et al.*, 2013; Gossmann *et al.*, 2013). The more inclusive selection of genes to represent pollen in the Gossmann *et al.* paper may explain the lack of difference in dN/dS between pollen genes and sporophytic genes. Genes exclusively expressed in each tissue, on the other hand, do show a significant difference, as shown in this study and in Szövényi *et al.* (2013). A further, possibly more important factor is that polymorphism and divergence data appear to represent different selection regimes. Szövényi *et al.* correctly interpreted the higher polymorphism levels (pN/pS) in pollen as evidence for weaker selection - this was further confirmed in this study not only by higher levels of nucleotide diversity and Watterson's theta but also greater levels of putatively deleterious mutations (premature stop codon and frameshift mutations) in pollen-specific genes. Szövényi *et al.* interpreted these polymorphism levels as evidence for relaxed purifying selection having caused the higher dN/dS levels in pollen-specific genes. However, I present evidence in this study which suggests stronger positive selection is responsible for the greater evolutionary rate of pollen proteins. I therefore offer evidence for a shift in selection patterns, possibly caused by the recent loss of self-incompatibility of *A. thaliana* and the resulting high homozygosity levels. This study presents further evidence for the effect of pollen competition and haploid expression on selection in plants. Furthermore the importance of masking is illustrated and how the relative effect of haploid selection is reduced when masking becomes negligible.

Chapter 5

Thesis conclusions

5.1 Thesis aims

I set out to investigate the influence of ploidy-specific expression on selection in the buff-tailed bumblebee, *Bombus terrestris*. This was to empirically test the predictions made by MJD White in 1945. I predicted lower levels of positive and purifying selection on genes which are mainly expressed in the diploid female castes, as recessive mutations will be masked from selection when heterozygous. All genes which are expressed in the haploid male, on the other hand, should show higher levels of selection, as recessive mutations will always be exposed to selection.

To test the generality of the effect of ploidy-specific expression on selection, I also investigated the effect in the plant *Arabidopsis thaliana*. I compared selection levels between genes with expression limited to haploid pollen tissues or the diploid sporophyte. Due to its recent loss of self-incompatibility, *A. thaliana* allowed me to test the importance of masking. I expected to find evidence for stronger selection on pollen-specific genes within divergence data, since they represent a period dominated by outcrossing. In polymorphism data, on the other hand, the current effects of selfing should be prevalent. I predicted a weaker effect of haploid selection to be visible in the polymorphism data due to high homozygosity levels which should reduce the effect of masking in diploid tissues.

A third study arose from the availability of a large set of RNAseq data for *B. terrestris*, which I had generated to determine ploidy-specific genes. I used these data to investigate expression patterns involved in caste determination in *B. terrestris*. This was the first broad scale study of its kind for bumblebees, which can be classed as intermediately eusocial, and allowed me to address the following hypotheses. Due to the plastic nature of bumblebee worker behaviour (Cameron, 1989), I hypothesised that when reproductive they become much more queen-like than

workers of highly eusocial species. This would be visible in a greater overlap in over-expressed genes between reproductive workers and queens than has been observed for the honeybee, *Apis mellifera*, (Grozinger *et al.*, 2007) and the myrmicine ant, *Temnothorax longispinosus* (Feldmeyer *et al.*, 2014). This would confirm my designation of bumblebees as intermediately eusocial, since the plasticity of worker castes is a primitive feature (Sumner *et al.*, 2006), and otherwise bumblebees possess many highly eusocial features, such as the irreversible determination of castes within development. Furthermore, I was interested in understanding expression patterns which control the different behaviour, physiology and morphology of queens, workers and males at each stage of development.

In the next sections I summarise the previous status of knowledge in these areas of study. This is followed by a summary of my findings detailed within this thesis and how they complement current knowledge. I also discuss possible future research.

5.2 Effects of ploidy and expression on selection

5.2.1 Previous evidence

The effect of ploidy-specific expression has been investigated for several groups of organisms, most notably in the context of the faster-X effect in *Drosophila* (Betancourt *et al.*, 2002; Thornton *et al.*, 2006; Musters *et al.*, 2006; Begun *et al.*, 2007; Baines *et al.*, 2008) and mammals (Torgerson and Singh, 2003; Khaitovich *et al.*, 2005; Baines and Harr, 2007; Carneiro *et al.*, 2012). The faster-X effect describes the faster evolution of genes, which are situated on the X-chromosome, due to positive selection. This phenomenon is explained by the exposure of X-linked genes to selection in males due to hemizygous expression compared to autosomal genes, in which recessive mutations are masked when heterozygous (Vicoso and Charlesworth, 2006; Meisel and Connallon, 2013). Initially, no effect could be found for *Drosophila* (Betancourt *et al.*, 2002; Thornton *et al.*, 2006), although later studies were able to confirm higher levels of positive selection and faster evolution for X-linked genes (Musters *et al.*, 2006; Begun *et al.*, 2007; Baines *et al.*, 2008). In mammals a significant effect has been found for primates (Lu and Wu, 2005; Nielsen *et al.*, 2005), mice (Baines and Harr, 2007) and rabbits (Carneiro *et al.*, 2012). Although, in the latter a faster-X effect was only found for one of two sub-species.

The apparent difficulty to detect a faster-X effect and the large variation in results

may be related to differences in the effective population size (N_e) between autosomal and X-linked genes. X-chromosomes are 3/4 as frequent as autosomes, which is likely to increase the influence of genetic drift on X-chromosomes relative to autosomes (Wright, 1931). Stronger genetic drift along with a relaxation of purifying selection within X-linked genes caused by a lower N_e would also lead to a faster evolution of X-linked genes due to the greater accumulation of slightly deleterious substitutions (Vicoso and Charlesworth, 2006). The ratio of N_e between sex-chromosomes and autosomes is even smaller in birds due to the stronger influence of mating variance in males on the sex chromosome, as the Z-chromosome spends 2/3 of its time in the male (Mank *et al.*, 2010a). Weaker purifying selection due to this difference in N_e seems to be causing a greater divergence of Z-linked genes compared to autosomal genes in birds (Mank *et al.*, 2010b).

Ploidy-specific expression has also been investigated in yeast and plants. In the budding yeast, *Saccharomyces cerevisiae*, adaptation rates were significantly higher in haploid populations compared to diploid populations (Zeyl *et al.*, 2003). In the plant *Capsella grandiflora* positive and purifying selection were stronger on genes with expression limited to pollen tissues, and therefore haploid, than on sporophytic, diploid genes (Arunkumar *et al.*, 2013). The authors named a combination of two causes for the stronger selection on pollen genes: sexual selection via pollen competition and haploid expression. For the self-compatible plant *Arabidopsis thaliana* conflicting results have been published (Szövényi *et al.*, 2013; Gossmann *et al.*, 2013). Szövényi *et al.* (2013) found purifying selection to be weaker on pollen-specific genes, leading to a greater rate of protein divergence. In contrast, Gossmann *et al.* (2013) found no difference in protein divergence rates between pollen and sporophytic genes but found evidence for stronger purifying and positive selection within pollen genes. The difficulty in finding the effect of haploid expression in *A. thaliana* is likely caused by high homozygosity levels due to selfing (Nordborg, 2000). This will reduce the masking effect of recessive mutations in sporophytic genes. As the ability to self-fertilise arose quite recently for *A. thaliana*, roughly 1MYA (Tang *et al.*, 2007), the reduction of the masking effect will only be detectable within polymorphism data, whereas divergence rates will reflect a regime of outcrossing.

5.2.2 Summary of findings

5.2.2.1 *Bombus terrestris*

I have researched the effects of ploidy within *B. terrestris* by comparing diploid-biased genes (upregulated in workers and queens) to haploid-biased genes (upregulated in males) and non-biased genes (equal expression in haploids and diploids). I

have found that purifying selection is weaker within genes, which are preferentially expressed in workers and queens, as these genes are not exposed to haploid selection in males. New, recessive mutations will generally be heterozygous in the diploid castes and so are masked by a dominant allele against selection. The stronger purifying selection within haploid-biased and non-biased genes was detectable in a lower proportion of slightly deleterious alleles and lower divergence rates relative to *B. impatiens* compared to diploid-biased genes. This is the first confirmation of the predictions made by MJD White in 1945.

Interestingly, I found no evidence for weaker positive selection within diploid-biased genes. In fact, the proportion of adaptive substitutions was slightly higher within diploid-biased genes compared to haploid-biased genes indicating positive selection is stronger on female castes. It seems therefore that haploid selection does not affect positive selection to the same degree as purifying selection, which suggests that new, recessive mutations are rarely adaptive. The scarcity of new, recessive, adaptive mutations has been suggested by Mank *et al.* (2010b) as the cause of the overall low faster-X effect in *Drosophila*. Also, when Betancourt *et al.* (2002) and Thornton *et al.* (2006) found no evidence of a faster-X effect in *Drosophila*, they suggested most beneficial mutations may not be recessive or that adaptation may use standing variation rather than new mutations. In a further study, similar to the results presented here, no evidence for a difference in positive selection could be found between X-chromosomal and autosomal genes in *Drosophila* but purifying selection was stronger on the X-chromosome (Connallon, 2007).

Positive selection was stronger in both diploid-biased and haploid-biased genes compared to non-biased genes. This was likely caused by sex-bias and sexual selection rather than ploidy bias. Stronger positive selection has often been reported for genes involved in reproduction (Turner and Hoekstra, 2008; Swanson and Vacquier, 2002). The pattern observed here is comparable to the lower levels of positive selection in non-biased genes for both autosomal and X-linked genes in *D. melanogaster* (Baines *et al.*, 2008; Pröschel *et al.*, 2006). Positive selection appears to be slightly higher within diploid-biased genes than in haploid-biased genes. This may mean that female castes (workers and queens) show a greater rate of adaptation compared to males. This is understandable when considering the greater selection for a greater variety of roles which queens and workers face compared to males. Whereas queens and workers are responsible for reproduction (mainly queens), colony development and maintenance, raising offspring, foraging and nest guarding, males are mainly produced to reproduce and perform no tasks within the colony.

The combination of stronger positive selection on haploid-biased and diploid-

biased genes but weaker purifying selection on diploid-biased genes has led to higher divergence rates in diploid-biased genes, intermediate divergence rates in haploid-biased genes and the lowest rates of divergence in non-biased genes.

5.2.2.2 *Arabidopsis thaliana*

I have found evidence for a change in selection patterns since *A. thaliana* became self-compatible. Divergence data, which represent a period of outcrossing, show evidence of stronger positive selection within pollen-specific genes compared to sporophyte-specific genes. This has led to a greater protein divergence rate of pollen genes than sporophytic genes as also recorded for the obligate outcrosser *Capsella grandiflora* (Arunkumar *et al.*, 2013). Polymorphism data, on the other hand, which represent a period of inbreeding and high levels of homozygosity (Nordborg, 2000; Wright *et al.*, 2008), show lower levels of selection in pollen-specific genes than in sporophyte-specific genes. This is observable in higher levels of polymorphism (π and θ) and putatively deleterious mutations (frameshift mutations and premature stop codons).

These findings confirm the effect of haploid expression on selection although the contribution of sexual selection from pollen competition should not be neglected. As assumed by Arunkumar *et al.* (2013) for *Capsella grandiflora* haploid expression likely enhances the effect of pollen competition. These results also highlight the importance of masking. Probably due to high levels of high homozygosity, and the corresponding low levels of heterozygosity, masking possibly no longer plays an important role in selection for *A. thaliana*, reducing the difference in selection efficacy between haploid-expressed genes and diploid-expressed genes. Additionally, high homozygosity would reduce pollen competition due to the lower number of competing pollen genotypes (Charlesworth and Charlesworth, 1992; Mazer *et al.*, 2010). A combined reduction in pollen competition and masking have therefore likely led to a reduction in selection on pollen-specific genes.

I have also highlighted the importance of tissue specificity and further genomic factors for selection. When analysing genes with expression limited to only one sporophytic tissue, differences in divergence and polymorphism levels to pollen-specific genes were reduced. However, when controlling for additional factors, such as expression level, GC-content and codon bias, the differences remained significant.

5.2.3 Overall conclusions

New mutations are generally rare and therefore heterozygous in diploid individuals, so are masked from selection if recessive. In haploids, on the other hand, new, recessive mutations will always be exposed to selection when expressed. This effect of

ploidy-specific expression on selection is widespread among a broad range of taxa. Before this study, a significant effect had been found for *Drosophila*, mammals (primates, mouse, rabbit), *S. cerevisiae* and the plant *Capsella grandiflora*. However, the isolation of the effect of ploidy-specific expression is often complicated by confounding factors. The difference in N_e between sex-linked genes and autosomal genes, for instance, explains why an effect has not been found in birds and has been difficult to isolate in *Drosophila*. The additional influence of sexual competition on the faster-X effect can be controlled for by comparing selection between X-linked and autosomal genes within male-biased, female-biased and non-biased genes (Baines *et al.*, 2008). Such a control is not possible for the comparison of pollen genes to sporophytic genes in plants, since pollen genes are inextricably linked to reproduction, so that a combination of pollen competition and haploid expression must be assumed as the cause of elevated selection (Arunkumar *et al.*, 2013).

It appears the effect may be even more widespread after now finding further evidence for stronger selection within haploid-expressed genes compared to genes with diploid-specific expression in another insect group, the bumblebees, and in a second plant, *Arabidopsis thaliana*. However, the results are more clear-cut for *B. terrestris*, since heightened selection within haploid-expressed genes could be detected both in divergence and polymorphism data. For *A. thaliana* only divergence data revealed stronger selection on pollen-specific genes. Furthermore, the comparisons between haploid-biased or non-biased genes with diploid-biased genes in *B. terrestris* were not strongly hindered by confounding factors. N_e does not differ between groups of genes since all individuals carry all genes. The role of sexual selection could also be controlled for by comparing both male-biased (haploid) and non-biased (expressed equally in males and females) genes to female-biased (diploid) genes. In both haploid-biased and non-biased genes, purifying selection was stronger than in diploid-biased genes, indicating an individual effect of haploid-expression on selection. The individual effect of haploid-expression on selection can not be isolated from the effect of pollen competition in *A. thaliana*, however. Although, the *A. thaliana* study may show the importance of the effect of masking. Since *A. thaliana* became self-compatible and heterozygosity levels decreased, this reduced the effect of masking, which, in part, may explain the strong reduction in selection efficacy on pollen-specific genes in polymorphism data compared to divergence data. The suitability of the polymorphism data used in the *A. thaliana* analysis for detecting purifying and positive selection is questionable since they are based on inbred strains rather than natural variation. Using one sequence to represent the variation in a population and then comparing between these strains is likely to exaggerate levels of

polymorphism since variation within populations is likely to be lower than between populations. However, these data have often been used in this manner (Gossmann *et al.*, 2013; Szövényi *et al.*, 2013). Distorted polymorphism levels may hinder the ability of the DoFE analysis to detect true levels of positive and purifying selection, although the relationship between pollen and sporophytic genes should not be affected.

5.2.4 Future research

In the future it would be interesting to test how prevalent this effect is within Hymenoptera and haplodiploids in general. A test within a non-social hymenopteran could eliminate the possibility that a peculiarity of castes within bumblebees may be causing the patterns observed within this study rather than ploidy-specific expression. It may be advisable to search for stronger purifying selection in X and Z-linked genes rather than positive selection, if, as this study suggests, most new recessive mutations are deleterious. The availability of genomic data, transcriptomic data and newly available software and techniques such as the DoFE analysis implemented in this thesis allow the detection of purifying and positive selection. For *A. thaliana*, it would be interesting to estimate levels of purifying and positive selection within natural variation data collected from wild populations. It is possible that this would offer a more precise depiction of current levels of selection.

5.3 Caste determination in *Bombus terrestris*

5.3.1 Previous research

Several studies exist which investigate the determination of castes in Hymenoptera via differential expression. For example, in the primitively eusocial paper wasp, *Polistes canadensis*, expression patterns overlap to a large extent between female adult castes (Sumner *et al.*, 2006). This can be explained by the high plasticity of the behavioural castes of primitively eusocial Hymenoptera; for example a colony worker has the ability to become a queen depending on the needs of a colony. In highly eusocial species, in which castes are very distinct and permanent, such as the honeybee, *Apis mellifera* (Grozinger *et al.*, 2007), or the ant, *Temnothorax longispinosus* (Feldmeyer *et al.*, 2014), expression patterns are very distinct between queens and workers. Even when a worker becomes reproductive its expression pattern still very

closely resembles that of a non-reproductive worker rather than a queen. In terms of expression, a worker remains a 'worker' (Grozinger *et al.*, 2007; Feldmeyer *et al.*, 2014). Most studies on caste determination to date have concentrated on highly eusocial Hymenoptera. Although large differences in expression patterns have been found between adult workers and queens in these studies (Grozinger *et al.*, 2007; Feldmeyer *et al.*, 2014; Hoffman and Goodisman, 2007; Ometto *et al.*, 2011), two studies found the largest difference to exist between developmental stage rather than between castes (Hoffman and Goodisman, 2007; Ometto *et al.*, 2011).

The categorisation of bumblebees as primitively or highly eusocial is disputed as they exhibit characteristics of both groups (Sadd *et al.*, 2015; Goulson, 2010). For example, workers can successfully produce offspring, they are plastic in their colony tasks and closely resemble the queen, which are typical primitive characteristics. The irreversible determination of female castes within development, on the other hand, is a highly eusocial characteristic. An in depth study of expression patterns may help clarify this. However, so far, very few studies had been conducted to explain caste determination in bumblebees. One study using suppression subtractive hybridisation techniques found 12 genes which were differentially expressed between workers and queens in *B. terrestris* (Pereboom *et al.*, 2005). A second study which concentrated on the construction of the *B. terrestris* transcriptome identified a higher number of differentially expressed transcripts (2,185) but described these as preliminary due to a lack of repetition (Colgan *et al.*, 2011).

5.3.2 Summary of findings

In this study I conducted the first large scale RNA-seq analysis into caste differentiation within the genus *Bombus*. I have described expression patterns within *B. terrestris*, which I believe confirm the placement of the caste determination system of this bumblebee species as intermediate along the evolutionary trajectory between primitively eusocial and highly eusocial Hymenoptera. Here I have shown that the expression patterns of reproductive workers much more closely resemble the expression patterns of queens than of non-reproductive workers for *B. terrestris*. This suggests the bumblebee worker caste may be more flexible than within highly eusocial species and that it becomes more queen-like when reproductive. This is similar to the flexible nature of primitively eusocial castes (Sumner *et al.*, 2006). This primitive feature suggests that this bumblebee is intermediately eusocial since bumblebees also possess advanced eusocial characteristics such as the non-reversible determination of castes during development.

Additionally, I found three vitellogenin genes to be differentially expressed in *B.*

terrestris. These appear not only to be differentially expressed within adult female castes as has previously been found for other hymenopteran taxa (Amdam *et al.*, 2004; Morandin *et al.*, 2014; Wurm *et al.*, 2011; Feldmeyer *et al.*, 2014; Corona *et al.*, 2013) but also within developmental stages. Furthermore, I found that expression patterns between males and workers become progressively more distinct throughout development, with the largest difference in adulthood. This suggests that a greater number of genes are necessary for controlling adult behaviours and physiology than for developing morphologies.

5.3.3 Future research

For future research, I suggest investigating the role of differential splicing. It is possible that not only the expression of different genes but also the expression of different splice variants of the same gene are important in the development of distinct castes. This analysis could be directly conducted on the data generated within this study. Furthermore, the annotation of many unknown genes, which were differentially expressed in my analysis, and further research on *B. terrestris* following the recent release of the genome (Sadd *et al.*, 2015) will help us to better understand how distinct castes are created, maintained or altered within this important species. Further research may be able to determine whether the findings presented here concerning the plasticity of worker expression patterns are restricted to *B. terrestris* or if they are linked to the more plastic nature of workers in bumblebee taxa in general.

Appendix A

Figures

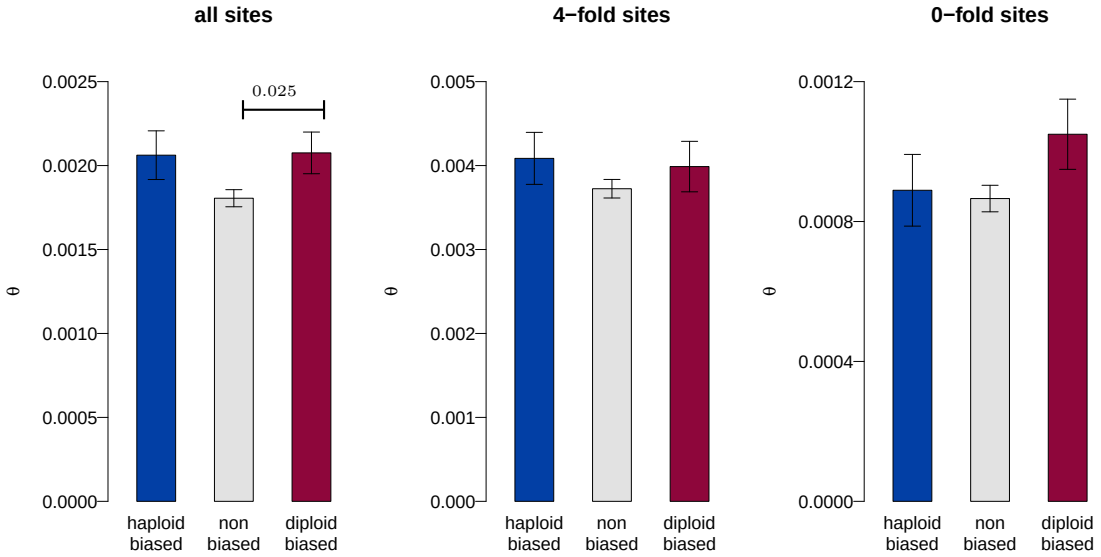


Figure A.1: Watterson's theta at all sites, 4-fold and 0-fold sites for male-biased, non-biased and queen-biased genes for 20 resampled alleles. Shown are means and standard errors.

Appendix B

Tables

Table B.1: ANCOVAs controlling for the effect of colony (section 2.3.2)

1) Larvae vs Adults	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sqrt(Expression)	1	3,661	3,661	487	< 2e-16
Caste	1	1	1	0	0.770
Colony	3	4	1	0	0.910
sqrt(Expression):Caste	1	71,195	71,195	9,477	< 2e-16
sqrt(Expression):Colony	3	1,297	432	58	< 2e-16
Caste:Colony	3	77	26	3.40	0.017
sqrt(Expression):Caste:Colony	3	1,643	548	72.89	< 2e-16
Residuals	1.2e05	9.0e05	8		

2) Larvae vs Pupae	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sqrt(Expression)	1	63,466	63,466	14,456	< 2e-16
Caste	1	18,836	18,836	4,290	< 2e-16
Colony	3	89	30	7	1.5e-04
sqrt(Expression):Caste	1	9,183	9,183	2,091	< 2e-16
sqrt(Expression):Colony	3	415	138	32	< 2e-16
Caste:Colony	3	28	9	2	0.093
sqrt(Expression):Caste:Colony	3	137	46	10	7.9e-07
Residuals	1.0e05	4.5e05	4		

3) Pupae vs Adults	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sqrt(Expression)	1	50	50	8	0.005
Caste	1	-	-	-	0.996
Colony	3	-	-	0	1.000
sqrt(Expression):Caste	1	97,540	97,540	15,586	< 2e-16
sqrt(Expression):Colony	3	390	130	21	2.0e-13
Caste:Colony	3	102	34	5	0.001
sqrt(Expression):Caste:Colony	3	269	90	14	2.5e-09
Residuals	1.5e05	9.2e05	6		

4) Male Larvae vs Worker Larvae	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Expression	1	68	68	11	0.001
Sex	1	0	0	0	0.775
Colony	3	0	0	0	1.000
Expression:Sex	1	1	1	0	0.723
Expression:Colony	3	5	2	0	0.834
Sex:Colony	1	0	0	0	0.941
Expression:Sex:Colony	1	0	0	0	0.920
Residuals	414	2.5e03	6		

5) Male Pupae vs Worker Pupae	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Expression	1	98	98	12	4.4e-04
Sex	1	-	0	0	0.970
Colony	3	0	0	0	1.000
Expression:Sex	1	1	1	0	0.686
Expression:Colony	3	2	1	0	0.976
Sex:Colony	1	-	-	0	0.994
Expression:Sex:Colony	1	1	1	0	0.704
Residuals	882	6.9e03	8		

6) Reproductive Workers vs Queens	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Expression	1	17	17	7	0.008
Caste	1	0	0	0	0.899
Colony	2	-	0	0	1.000
Expression:Caste	1	1	1	0	0.468
Expression:Colony	2	0	0	0	0.990
Caste:Colony	2	0	0	0	0.998
Expression:Caste:Colony	2	1	0	0.15	0.857
Residuals	751	1,827	2		

7) Non-reproductive Workers vs Queens	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sqrt(Expression)	1	2,598	2,598	1,028	< 2e-16
Caste	1	7	7	3	0.091
Colony	3	2	1	0	0.832
sqrt(Expression):Caste	1	2,826	2,826	1,118	< 2e-16
sqrt(Expression):Colony	3	69	23	9	4.8e-06
Residuals	2.6e04	6.6e04	3		

8) Reproductive workers vs Non-reproductive Workers	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sqrt(Expression)	1	349	348.94	125	<2e-16
Reproductive	1	0	0	0	0.816
Colony	3	3	1	0	0.810
sqrt(Expression):Reproductive	1	1,360	1,360	487	<2e-16
sqrt(Expression):Colony	3	4	1	0	0.729
Residuals	9.4e03	2.6e04	3		

9) Males vs. Non-reprodroductive Workers	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sqrt(Expression)	1	887	887	676	< 2e-16
Sex	1	1	1	1	0.412
Colony	3	2	1	1	0.640
sqrt(Expression):Sex	1	4,960	4,960	3,782	< 2e-16
sqrt(Expression):Colony	3	26	9	7	1.8e-04
Residuals	1.3e05	1.7e05	1		

Table B.2: Top 50 GO terms (out of 142) for transcripts upregulated in larvae compared to pupae and adults (section 2.3.3).
Significance estimated with Fisher's exact test.

GO.ID	Term	FDR
GO:0042254	ribosome biogenesis	< 1e-30
GO:0006412	translation	< 1e-30
GO:0022613	ribonucleoprotein complex biogenesis	< 1e-30
GO:0003735	structural constituent of ribosome	< 1e-30
GO:0005198	structural molecule activity	< 1e-30
GO:0005840	ribosome	< 1e-30
GO:0044444	cytoplasmic part	< 1e-30
GO:0005737	cytoplasm	< 1e-30
GO:0030529	ribonucleoprotein complex	2.9E-25
GO:0005739	mitochondrion	2.3E-13
GO:0044391	ribosomal subunit	7.2E-13
GO:0044249	cellular biosynthetic process	3.1E-12
GO:0009058	biosynthetic process	6.3E-12
GO:0044085	cellular component biogenesis	6.7E-12
GO:1901576	organic substance biosynthetic process	2.1E-11
GO:0043228	non-membrane-bounded organelle	8.9E-11
GO:0043232	intracellular non-membrane-bounded organ...	8.9E-11
GO:0044429	mitochondrial part	3.4E-10
GO:0044424	intracellular part	2.0E-08
GO:0043229	intracellular organelle	4.2E-08
GO:0043226	organelle	7.5E-08
GO:0022891	substrate-specific transmembrane transpo...	8.9E-08
GO:0015935	small ribosomal subunit	2.6E-07
GO:0015075	ion transmembrane transporter activity	3.4E-07
GO:0015078	hydrogen ion transmembrane transporter a...	3.4E-07
GO:0015992	proton transport	5.0E-07
GO:0034645	cellular macromolecule biosynthetic proc...	6.2E-07
GO:0006818	hydrogen transport	6.2E-07
GO:0009059	macromolecule biosynthetic process	6.2E-07
GO:0022857	transmembrane transporter activity	1.5E-06
GO:0006119	oxidative phosphorylation	1.6E-06
GO:0010467	gene expression	1.6E-06
GO:1901566	organonitrogen compound biosynthetic pro...	1.7E-06
GO:0045333	cellular respiration	4.1E-06
GO:0015980	energy derivation by oxidation of organi...	5.6E-06
GO:0015672	monovalent inorganic cation transport	5.8E-06
GO:0005759	mitochondrial matrix	1.1E-05
GO:0008324	cation transmembrane transporter activit...	1.1E-05
GO:0032991	macromolecular complex	1.5E-05
GO:0022892	substrate-specific transporter activity	1.6E-05
GO:0005811	lipid particle	1.6E-05
GO:0005743	mitochondrial inner membrane	2.0E-05
GO:0016491	oxidoreductase activity	2.3E-05
GO:0019866	organelle inner membrane	2.8E-05
GO:0019538	protein metabolic process	4.2E-05
GO:0006091	generation of precursor metabolites and ...	4.4E-05
GO:0005740	mitochondrial envelope	5.1E-05
GO:0022900	electron transport chain	5.2E-05
GO:0006811	ion transport	5.2E-05
GO:0042625	ATPase activity, coupled to transmembran...	5.5E-05

Table B.3: Top 50 GO terms (out of 257) for transcripts upregulated in pupae compared to larvae and adults (section 2.3.3).
Significance estimated with Fisher's exact test.

GO.ID	Term	FDR
GO:0023052	signaling	3.7E-22
GO:0044700	single organism signaling	3.7E-22
GO:0007165	signal transduction	3.7E-22
GO:0007154	cell communication	2.2E-21
GO:0051716	cellular response to stimulus	8.4E-17
GO:0065007	biological regulation	1.7E-15
GO:0050896	response to stimulus	2.9E-15
GO:0050789	regulation of biological process	3.0E-15
GO:0050794	regulation of cellular process	3.0E-15
GO:0009653	anatomical structure morphogenesis	1.3E-13
GO:0007399	nervous system development	2.5E-12
GO:0007275	multicellular organismal development	8.2E-12
GO:0048731	system development	1.7E-11
GO:0044707	single-multicellular organism process	5.8E-11
GO:0048856	anatomical structure development	6.0E-11
GO:0032501	multicellular organismal process	7.4E-11
GO:0048513	organ development	1.3E-10
GO:0044767	single-organism developmental process	1.7E-10
GO:0032502	developmental process	2.8E-10
GO:0060429	epithelium development	2.9E-09
GO:0048869	cellular developmental process	4.7E-09
GO:0009887	organ morphogenesis	5.0E-09
GO:0005515	protein binding	7.8E-09
GO:0016337	cell-cell adhesion	9.7E-08
GO:0048736	appendage development	1.3E-07
GO:0030154	cell differentiation	1.3E-07
GO:0007423	sensory organ development	1.9E-07
GO:0035107	appendage morphogenesis	2.7E-07
GO:0002009	morphogenesis of an epithelium	4.9E-07
GO:0016043	cellular component organization	5.9E-07
GO:0048729	tissue morphogenesis	6.5E-07
GO:0007560	imaginal disc morphogenesis	8.5E-07
GO:0007552	metamorphosis	8.5E-07
GO:0035120	post-embryonic appendage morphogenesis	8.5E-07
GO:0048563	post-embryonic organ morphogenesis	9.9E-07
GO:0035114	imaginal disc-derived appendage morphoge...	1.0E-06
GO:0048646	anatomical structure formation involved ...	1.2E-06
GO:0048737	imaginal disc-derived appendage developm...	1.2E-06
GO:0007166	cell surface receptor signaling pathway	1.3E-06
GO:0007476	imaginal disc-derived wing morphogenesis	1.3E-06
GO:0007472	wing disc morphogenesis	1.4E-06
GO:0035220	wing disc development	1.4E-06
GO:0048468	cell development	1.5E-06
GO:0048518	positive regulation of biological proces...	1.6E-06
GO:0007444	imaginal disc development	2.7E-06
GO:0009888	tissue development	2.8E-06
GO:0048569	post-embryonic organ development	3.3E-06
GO:0007017	microtubule-based process	3.6E-06
GO:0002165	instar larval or pupal development	3.7E-06
GO:0007389	pattern specification process	3.9E-06

Table B.4: All 46 GO terms for transcripts upregulated in adults compared to larvae and pupae (section 2.3.3).
Significance estimated with Fisher's exact test.

GO.ID	Term	FDR
GO:0007186	G-protein coupled receptor signaling pat...	1.0E-09
GO:0031226	intrinsic component of plasma membrane	2.6E-05
GO:0005887	integral component of plasma membrane	2.6E-05
GO:0004930	G-protein coupled receptor activity	4.9E-05
GO:0004872	receptor activity	5.3E-05
GO:0008188	neuropeptide receptor activity	5.3E-05
GO:0038023	signaling receptor activity	5.3E-05
GO:0004888	transmembrane signaling receptor activit...	5.6E-05
GO:0001653	peptide receptor activity	9.7E-05
GO:0008528	G-protein coupled peptide receptor activ...	9.7E-05
GO:0030594	neurotransmitter receptor activity	9.7E-05
GO:0015079	potassium ion transmembrane transporter ...	1.4E-04
GO:0022836	gated channel activity	1.6E-04
GO:0005249	voltage-gated potassium channel activity	1.7E-04
GO:0005215	transporter activity	2.4E-04
GO:0004672	protein kinase activity	4.2E-04
GO:0008076	voltage-gated potassium channel complex	4.7E-04
GO:0034705	potassium channel complex	4.7E-04
GO:0015276	ligand-gated ion channel activity	5.0E-04
GO:0022834	ligand-gated channel activity	5.0E-04
GO:0004968	gonadotropin-releasing hormone receptor ...	5.0E-04
GO:0016500	protein-hormone receptor activity	5.0E-04
GO:0035237	corazonin receptor activity	5.0E-04
GO:0097003	adipokinetic hormone receptor activity	5.0E-04
GO:0097004	adipokinetic hormone binding	5.0E-04
GO:0005267	potassium channel activity	6.5E-04
GO:0004348	glucosylceramidase activity	8.8E-04
GO:0016772	transferase activity, transferring phosp...	1.5E-03
GO:0005216	ion channel activity	1.8E-03
GO:0015267	channel activity	1.8E-03
GO:0022803	passive transmembrane transporter activi...	1.8E-03
GO:0022838	substrate-specific channel activity	1.8E-03
GO:0016301	kinase activity	2.7E-03
GO:0006813	potassium ion transport	3.4E-03
GO:0007218	neuropeptide signaling pathway	3.7E-03
GO:0004871	signal transducer activity	3.9E-03
GO:0060089	molecular transducer activity	3.9E-03
GO:0005261	cation channel activity	3.9E-03
GO:0005516	calmodulin binding	4.0E-03
GO:0044242	cellular lipid catabolic process	5.1E-03
GO:0007187	G-protein coupled receptor signaling pat...	5.5E-03
GO:0007166	cell surface receptor signaling pathway	5.5E-03
GO:0007154	cell communication	5.5E-03
GO:0016773	phosphotransferase activity, alcohol gro...	7.4E-03
GO:0022843	voltage-gated cation channel activity	7.8E-03
GO:0005230	extracellular ligand-gated ion channel a...	9.2E-03

Table B.5: Transcripts which were upregulated in male larvae compared to worker larvae (section 2.3.4).

BTT contig	Functional description	Expression level	log2FC	FDR
BTT05816_2	nose resistant to fluoxetine protein 6-like	52.42	4.66	3.4E-05
BTT31660_1	nose resistant to fluoxetine protein 6-like	110.46	3.66	1.5E-02
BTT32372_1	2-acylglycerol o-acyltransferase 1-like	146.80	3.44	2.4E-11
BTT05815_1	nose resistant to fluoxetine protein 6-like	135.04	3.19	7.4E-09
BTT20058_1	pancreatic lipase-related protein 2-like	5,330.14	2.96	2.4E-11
BTT35097_1	nose resistant to fluoxetine protein 6-like	156.89	2.89	5.2E-04
BTT07560_1	protein spinster homolog 1-like	32.01	2.67	1.5E-02
BTT25609_1	short-chain dehydrogenase reductase family 16c member 6-like	75.56	2.55	1.1E-02
BTT15842_1	protein takeout-like	2,693.38	2.52	6.7E-05
BTT21012_1	PREDICTED: hypothetical protein LOC100643810	170.63	2.23	4.8E-10
BTT41033_1	glucosylceramidase-like isoform 1	27.11	2.11	2.0E-02
BTT27015_1	NA	191.89	2.05	4.3E-09
BTT39743_1	nose resistant to fluoxetine protein 6-like	47.90	2.01	1.6E-03
BTT08924_1	nose resistant to fluoxetine protein 6-like	4,575.94	1.95	5.2E-10
BTT09362_1	peroxidase-like isoform 1	1,135.56	1.90	3.1E-11
BTT23692_1	peroxidase-like isoform 1	243.69	1.88	3.7E-08
BTT32452_1	sec14 domain and spectrin repeat-containing protein 1	88.54	1.78	1.4E-04
BTT07746_1	elongation of very long chain fatty acids protein 6-like	3,076.93	1.76	7.7E-05
BTT10407_1	fatty acid synthase-like	806.12	1.70	1.5E-08
BTT32213_1	elongation of very long chain fatty acids protein 6-like	106.34	1.70	1.1E-04
BTT19690_1	NA	108.88	1.59	7.0E-03
BTT14649_1	neuropeptides capa receptor-like	87.60	1.34	3.2E-02
BTT10453_1	NA	107.07	1.33	1.5E-02
BTT18106_1	PREDICTED: hypothetical protein LOC100646752	79.44	1.26	3.9E-02
BTT06052_1	dc-stamp domain-containing protein 1-like	159.97	1.26	6.5E-03
BTT08367_1	serpin b13-like	2,917.39	1.12	9.4E-03
BTT16692_1	ras-related protein rab-7a-like	22,564.75	1.11	5.7E-04
BTT06055_1	ejaculatory bulb-specific protein 3-like	450.20	1.06	9.1E-03
BTT39618_1	cytochrome p450 6b1-like	1,681.69	1.05	1.5E-02
BTT19293_1	heparan sulfate 2-o-sulfotransferase pipe-like	793.37	1.05	2.9E-02
BTT10342_1	heparan sulfate 2-o-sulfotransferase pipe-like	481.95	0.96	4.8E-02
BTT26045_1	2-acylglycerol o-acyltransferase 1-like	934.88	0.90	4.4E-02

Table B.6: Transcripts which were upregulated in worker larvae compared to male larvae (section 2.3.4).

BTT contig	Functional description	Expression level	log2FC	p value
BTT25679_1	NA	20.31	5.88	3.4E-05
BTT25839_1	NA	16.75	5.84	7.8E-05
BTT16589_1	NA	35.67	5.77	5.2E-10
BTT16238_1	PREDICTED: hypothetical protein LOC100645612	35.52	5.76	8.9E-11
BTT12404_1	PREDICTED: hypothetical protein LOC100652144	46.26	5.56	1.8E-13
BTT19436_1	NA	51.98	5.08	1.8E-13
BTT08637_2	PREDICTED: copia protein-like	86.85	5.01	1.4E-20
BTT24011_1	NA	57.30	4.99	2.4E-14
BTT27508_1	PREDICTED: hypothetical protein LOC100647826	27.35	4.67	4.6E-07
BTT18331_1	tick partial	117.67	4.39	9.4E-22
BTT34053_1	PREDICTED: uncharacterized protein LOC100906527	22.33	4.03	4.3E-05
BTT11489_1	piggybac transposable element derived 4	130.18	3.55	5.7E-04
BTT32169_1	PREDICTED: hypothetical protein LOC100645612	27.96	3.24	9.0E-05
BTT18547_1	PREDICTED: hypothetical protein LOC100645612	75.01	3.01	4.8E-10
BTT21925_1	NA	28.86	2.99	1.9E-04
BTT31177_1	PREDICTED: uncharacterized protein LOC100901753, partial	78.70	2.96	5.2E-10
BTT03588_1	NA	69.48	2.96	8.3E-10
BTT24408_1	low quality protein: vitellogenin-like	61.66	2.12	1.6E-04
BTT14131_1	lachesin- partial	238.37	2.12	1.4E-05
BTT23205_1	receptor-type tyrosine-protein phosphatase-like n-like	65.46	2.09	5.9E-05
BTT06619_1	PREDICTED: hypothetical protein LOC100645612	28.60	2.04	2.9E-02
BTT33552_1	NA	30.69	1.95	3.7E-02
BTT15020_1	PREDICTED: hypothetical protein LOC100644060 isoform 1	37.27	1.86	2.6E-02
BTT18413_1	monocarboxylate transporter 13-like	52.85	1.85	2.6E-03
BTT29171_1	NA	39.62	1.82	2.6E-02
BTT11661_1	PREDICTED: hypothetical protein LOC100742053 isoform 1	60.47	1.74	3.1E-03
BTT21572_1	PREDICTED: hypothetical protein LOC100643785	459.84	1.74	1.1E-04
BTT19170_1	protein brown-like	157.21	1.66	1.2E-05
BTT21147_1	receptor-type tyrosine-protein phosphatase-like n-like	311.01	1.51	1.5E-05
BTT10858_1	NA	202.47	1.45	8.6E-03
BTT00982_1	PREDICTED: hypothetical protein LOC100645612	71.16	1.44	1.7E-02
BTT09531_1	NA	163.26	1.42	4.8E-02
BTT09386_1	leucine-rich repeat-containing protein 15-like	562.80	1.33	1.4E-04
BTT17887_1	NA	316.78	1.28	1.1E-02
BTT27775_1	transcription factor gata-4	148.70	1.27	7.4E-03
BTT22568_1	cytochrome p450 6k1-like	164.91	1.26	4.8E-02
BTT07564_1	facilitated trehalose transporter tret1-like	1,609.60	1.00	1.6E-02
BTT08704_1	organic cation transporter	1,280.22	0.91	4.0E-02
BTT18296_1	lipase 3-like	7,888.91	0.89	5.0E-02

Table B.7: Top 50 transcripts which were upregulated in male pupae compared to worker pupae (section 2.3.5).

BTT contig	Functional description	Expression level	log2FC	FDR
BTT29597_1	hypothetical protein EAI_00375	63.69	Inf	1.4E-10
BTT01164_1	PREDICTED: hypothetical protein LOC100649137	145.72	Inf	2.0E-09
BTT15423_1	NA	113.98	Inf	9.0E-24
BTT12359_1	PREDICTED: hypothetical protein LOC100647675	639.90	10.07	4.0E-18
BTT22346_1	tubulin alpha-2 chain-like	1,199.60	8.94	2.8E-08
BTT00496_1	hypothetical protein EAG_01059	163.35	8.90	6.5E-05
BTT20446_1	PREDICTED: hypothetical protein LOC100647818	202.68	8.32	2.3E-12
BTT02878_1	PREDICTED: hypothetical protein LOC100648702	71.69	7.91	5.6E-18
BTT15953_1	PREDICTED: hypothetical protein LOC100642776	113.88	7.47	7.6E-04
BTT04893_1	PREDICTED: hypothetical protein LOC100745222	198.17	7.40	1.1E-04
BTT14059_1	PREDICTED: hypothetical protein LOC100646017	171.77	7.20	8.1E-14
BTT01187_1	PREDICTED: hypothetical protein LOC100648071 isoform 1	113.36	7.02	1.2E-05
BTT00925_1	PREDICTED: hypothetical protein LOC100649613 isoform 1	30.06	6.69	2.3E-07
BTT24086_1	tubulin–tyrosine ligase-like protein 3	91.85	6.63	3.4E-09
BTT11530_1	NA	102.60	6.44	3.9E-06
BTT18195_1	NA	96.38	6.23	6.8E-06
BTT15285_1	PREDICTED: hypothetical protein LOC100647943	39.77	5.86	6.8E-03
BTT04680_1	heavy subunit-like	201.60	5.78	4.2E-06
BTT11982_1	PREDICTED: hypothetical protein LOC100649354	31.43	5.74	9.0E-04
BTT13944_1	2-oxoglutarate mitochondrial-like	82.36	5.66	3.2E-09
BTT04534_1	carbonic anhydrase 2-like	110.60	5.55	5.0E-09
BTT13868_1	tubulin alpha chain-like	397.34	5.51	4.0E-09
BTT03936_1	PREDICTED: enolase-like	1,184.02	5.47	3.1E-30
BTT16145_1	NA	13.16	5.46	1.7E-02
BTT13922_1	tubulin–tyrosine ligase-like protein 8	55.68	5.46	8.1E-12
BTT14637_1	NA	71.91	5.34	3.8E-14
BTT15911_1	PREDICTED: hypothetical protein LOC100648296	122.08	5.19	7.7E-03
BTT14148_1	tubulin alpha chain-like	599.31	5.06	2.4E-09
BTT15891_1	NA	38.30	4.96	2.8E-03
BTT20781_1	PREDICTED: uncharacterized protein LOC100863631	63.52	4.91	3.6E-02
BTT22723_1	PREDICTED: hypothetical protein LOC100648520	926.56	4.84	6.2E-04
BTT16993_1	NA	51.03	4.81	3.6E-02
BTT21818_1	serine threonine-protein kinase ick-like	57.22	4.76	1.6E-02
BTT23207_1	mitogen-activated protein kinase 15-like	368.85	4.74	6.7E-03
BTT11481_1	mitogen-activated protein kinase 15-like	176.91	4.64	4.4E-03
BTT13249_1	NA	347.25	4.62	3.0E-11
BTT16416_1	PREDICTED: hypothetical protein LOC100645588	20.93	4.56	3.8E-02
BTT28132_1	PREDICTED: hypothetical protein LOC100744372	34.95	4.47	1.2E-04
BTT28269_1	iq and aaa domain-containing protein 1-like	26.83	4.40	8.6E-05
BTT22840_1	PREDICTED: hypothetical protein LOC100651658	66.74	4.34	2.0E-09
BTT15728_1	PREDICTED: hypothetical protein LOC100646640	230.38	4.31	2.4E-02
BTT28567_1	NA	45.88	4.23	1.1E-07
BTT14432_1	PREDICTED: hypothetical protein LOC100644232	51.35	4.22	1.2E-02
BTT10065_1	inorganic phosphate cotransporter-like	119.90	4.20	3.7E-03
BTT00100_1	PREDICTED: hypothetical protein LOC100646506	435.00	4.19	9.1E-09
BTT20349_1	NA	33.97	4.11	9.0E-06
BTT16540_1	PREDICTED: hypothetical protein LOC100642978	53.20	4.06	3.0E-04
BTT00396_1	PREDICTED: hypothetical protein LOC100649613 isoform 2	19.98	3.97	4.6E-03
BTT40187_1	proteasome inhibitor pi31 subunit-like	42.93	3.79	2.4E-06
BTT15407_1	serine threonine-protein kinase ick-like	41.20	3.72	2.7E-02

Table B.8: Transcripts which were upregulated in worker pupae compared to male pupae (section 2.3.5).

BTT contig	Functional description	Expression level	log2FC	FDR
BTT11724_1	NA	12.07	Inf	1.5E-02
BTT24011_1	NA	130.74	4.70	5.2E-19
BTT16589_1	NA	88.25	4.49	1.1E-13
BTT12404_1	PREDICTED: hypothetical protein LOC100652144	104.75	4.27	8.9E-12
BTT18331_1	tick partial	238.90	3.89	2.7E-21
BTT16238_1	PREDICTED: hypothetical protein LOC100645612	110.10	3.73	1.4E-06
BTT41909_1	NA	66.77	3.70	2.7E-04
BTT25679_1	NA	39.91	3.33	1.5E-04
BTT16940_1	NA	27.02	3.21	4.6E-03
BTT18012_1	PREDICTED: uncharacterized protein LOC100899117	46.98	3.05	5.7E-05
BTT34053_1	PREDICTED: uncharacterized protein LOC100906527	49.36	3.03	3.5E-05
BTT11489_1	piggybac transposable element derived 4	315.89	2.91	1.1E-09
BTT18547_1	PREDICTED: hypothetical protein LOC100645612	157.52	2.73	4.2E-10
BTT21925_1	NA	32.56	2.71	6.7E-03
BTT31177_1	PREDICTED: uncharacterized protein LOC100901753, partial	170.72	2.69	2.4E-06
BTT24408_1	low quality protein: vitellogenin-like	29.32	2.54	2.3E-02
BTT07228_1	NA	100.77	2.39	2.6E-06
BTT25839_1	NA	31.35	2.29	4.8E-02
BTT33552_1	NA	85.46	2.16	1.1E-04
BTT28372_1	PREDICTED: hypothetical protein LOC100644060 isoform 1	67.43	2.13	9.5E-04
BTT21809_1	PREDICTED: hypothetical protein LOC100644060 isoform 1	131.58	2.08	1.3E-05
BTT11661_1	PREDICTED: hypothetical protein LOC100742053 isoform 1	170.82	1.99	5.2E-06
BTT20123_1	hymenoptaecin- partial	57.84	1.97	4.8E-02
BTT15020_1	PREDICTED: hypothetical protein LOC100644060 isoform 1	93.18	1.73	3.5E-03
BTT18790_1	PREDICTED: hypothetical protein LOC100742070	22,761.33	1.54	1.5E-02
BTT32289_1	antithrombin-iii-like isoform 2	308.82	1.44	1.2E-02
BTT17732_1	fatty acyl- reductase cg5065-like	568.09	1.44	5.0E-02
BTT01754_1	NA	190.80	1.34	1.5E-02
BTT15079_1	NA	1,590.23	1.32	2.4E-02
BTT12082_1	PREDICTED: hypothetical protein LOC100742053 isoform 1	156.79	1.28	2.4E-02
BTT35210_1	vesicular glutamate transporter 3-like	332.07	1.23	7.3E-03
BTT04907_1	15-hydroxyprostaglandin dehydrogenase	1,887.23	1.07	5.1E-03
BTT12160_1	inorganic phosphate cotransporter-like	455.32	1.03	3.5E-02
BTT10449_1	beta-hexosaminidase subunit beta-like	4,783.84	1.02	2.4E-02

Table B.9: Top 50 fertility genes: Transcripts upregulated in queens and reproductive workers compared to non-reproductive workers (section 2.3.6).

BTT contig	Functional description	Expression level	log2FC	FDR
BTT25083_1	ccmar2 transposase	42.85	Inf	3.7E-02
BTT41628_1	bcl2 adenovirus e1b 19 kda protein-interacting protein 3-like	23.66	Inf	3.1E-06
BTT00609_1	PREDICTED: hypothetical protein LOC100743783	5,734.62	8.39	1.6E-76
BTT41624_1	PREDICTED: hypothetical protein LOC100746188	480.76	4.00	9.5E-04
BTT02648_1	NA	65.13	3.85	6.5E-08
BTT13744_1	broad-complex core protein isoforms 1 2 3 4 5-like	16.25	3.64	2.5E-02
BTT32170_1	PREDICTED: hypothetical protein LOC100646322	27.57	3.55	1.6E-04
BTT39845_1	60s ribosomal protein l10	75.19	3.55	4.2E-05
BTT41529_1	4-aminobutyrate mitochondrial-like	17.07	3.45	4.0E-03
BTT04883_1	NA	84.96	3.43	8.3E-06
BTT02512_1	PREDICTED: hypothetical protein LOC100642825 isoform 2	19.12	3.39	3.2E-03
BTT02697_1	NA	5,254.04	3.35	1.8E-02
BTT02663_1	NA	17.61	3.34	6.9E-03
BTT01389_1	NA	9,274.99	3.32	2.0E-02
BTT03164_1	PREDICTED: hypothetical protein LOC100645132	660.82	3.26	5.6E-04
BTT16271_1	NA	88.01	3.26	3.0E-04
BTT04926_1	jerky protein homolog-like	163.03	3.25	9.8E-04
BTT13690_1	tripartite motif-containing protein 71-like	134.72	3.25	3.9E-05
BTT05097_1	NA	84.37	3.25	4.2E-02
BTT00774_1	NA	49.47	3.25	1.4E-02
BTT34373_1	NA	98.07	3.20	7.3E-03
BTT02218_1	retrovirus-related pol polyprotein from transposon tnt 1-94	67.52	3.19	2.9E-03
BTT05460_1	alpha-glucosidase-like isoform 1	17,292.30	3.19	3.1E-04
BTT25957_1	PREDICTED: hypothetical protein LOC100742042	22.94	3.19	1.8E-03
BTT35759_1	pancreatic lipase-related protein 2-like	3,652.78	3.19	6.1E-05
BTT00092_1	PREDICTED: hypothetical protein LOC100646104	339.43	3.17	2.6E-02
BTT40536_1	nadh dehydrogenase	423.03	3.16	6.4E-03
BTT05212_1	PREDICTED: hypothetical protein LOC100740006	208.21	3.16	6.1E-05
BTT00344_1	mitochondrial import receptor subunit tom40-like protein	241.81	3.13	3.0E-02
BTT24121_1	alpha-glucosidase iii	35,259.82	3.13	5.4E-04
BTT01106_1	protein cip2a-like	10,029.91	3.12	4.6E-03
BTT19982_1	NA	225.20	3.12	1.1E-03
BTT02307_1	NA	3,149.49	3.11	3.4E-03
BTT19355_1	NA	3,301.90	3.06	4.7E-03
BTT21092_1	alpha-glucosidase-like isoform 1	9,406.72	3.05	3.1E-04
BTT29559_1	NA	34.38	3.04	7.5E-03
BTT01807_1	NA	1,761.25	3.02	1.2E-03
BTT02333_1	NA	459.35	3.00	1.1E-03
BTT22653_1	PREDICTED: hypothetical protein LOC100650639	46.63	2.98	1.3E-05
BTT06357_1	macro domain-containing protein 2-like isoform 2	15.50	2.96	1.7E-02
BTT02007_1	PREDICTED: translin-like	15.86	2.93	1.3E-02
BTT04356_1	sentrin-specific protease 8-like isoform 1	44.52	2.92	2.9E-05
BTT02297_1	g2 mitotic-specific cyclin-b	938.90	2.91	5.5E-03
BTT00956_1	NA	18.28	2.91	7.9E-03
BTT20455_1	jerky protein homolog-like	302.04	2.90	9.2E-03
BTT22592_1	alpha-glucosidase-like isoform 1	15.50	2.90	1.6E-02
BTT10379_1	cytochrome p450 18a1-like	178.58	2.90	4.7E-03
BTT00584_1	zinc finger protein 2-like	2,641.09	2.90	5.7E-05
BTT02380_1	PREDICTED: hypothetical protein LOC100648437	351.48	2.89	3.1E-02
BTT10310_1	NA	66.37	2.88	3.9E-04

Table B.10: Top 50 transcripts upregulated in non-reproductive workers compared to reproductive workers (section 2.3.7).

BTT contig	Functional description	Expression level	log2FC	FDR
BTT31161_1	PREDICTED: hypothetical protein LOC100642686	20.68	Inf	4.8E-12
BTT40283_1	PREDICTED: hypothetical protein LOC100649867	6.12	Inf	3.4E-02
BTT35230_1	phospholipid hydroperoxide glutathione mitochondrial-like isoform 1	14.05	7.38	8.1E-09
BTT20899_1	NA	17.30	6.55	7.7E-03
BTT32000_1	PREDICTED: hypothetical protein LOC100647540	23.77	6.13	6.2E-03
BTT17543_1	hexokinase-1-like isoform 1	44.88	5.63	4.2E-05
BTT07422_1	PREDICTED: hypothetical protein LOC100643115	1,950.48	5.45	1.3E-05
BTT07349_1	PREDICTED: hypothetical protein LOC100643262	21.57	5.10	1.9E-02
BTT05060_1	folylpolyglutamate mitochondrial-like	4.43	4.68	3.4E-03
BTT28399_1	NA	2.51	4.61	4.2E-02
BTT22515_1	PREDICTED: hypothetical protein LOC100648625	64.94	4.60	1.5E-03
BTT29026_1	NA	9.14	4.40	1.2E-02
BTT17965_1	PREDICTED: hypothetical protein LOC100644429	12.22	4.24	6.6E-07
BTT20828_1	apidaecins type 73-like	694.75	4.22	3.8E-02
BTT04091_1	NA	8.51	3.79	1.2E-04
BTT27036_1	NA	24.35	3.77	4.3E-09
BTT20802_1	abaecin isoform 1	14.55	3.74	1.2E-06
BTT10678_1	NA	3.47	3.71	4.4E-02
BTT19615_1	PREDICTED: hypothetical protein LOC100642686	34.33	3.64	1.2E-13
BTT01310_1	domon domain-containing protein cg14681-like	4.69	3.54	1.2E-02
BTT24067_1	calcium-transporting atpase sarcoplasmic endoplasmic reticulum type-like	146.96	3.48	3.4E-10
BTT15927_1	Ankyrin-2	5.17	3.41	1.2E-02
BTT24876_1	PREDICTED: uncharacterized protein LOC100878529	4.97	3.38	1.5E-02
BTT38683_1	odorant binding protein 3 precursor	26.84	3.23	1.8E-08
BTT27692_1	NA	5.71	3.21	1.1E-02
BTT31497_1	NA	4.13	3.20	3.2E-02
BTT30373_1	esterase fe4-like	20.29	3.18	1.3E-02
BTT40753_1	maltase 1-like	127.61	3.12	3.2E-05
BTT19116_1	NA	65.75	3.06	9.0E-06
BTT29003_1	esterase fe4-like	426.18	2.96	1.2E-22
BTT36530_1	leucine-rich repeat-containing protein 20-like isoform 1	11.32	2.92	4.1E-02
BTT31058_1	PREDICTED: hypothetical protein LOC100642686	30.73	2.90	1.0E-05
BTT19563_1	NA	16.86	2.86	3.9E-02
BTT12794_1	NA	5.20	2.83	2.6E-02
BTT08866_1	NA	51.53	2.83	2.2E-03
BTT21302_1	NA	42.81	2.81	2.7E-03
BTT27376_1	PREDICTED: hypothetical protein LOC100651531	21.53	2.80	2.9E-06
BTT23599_1	NA	8.05	2.80	5.1E-03
BTT22258_1	NA	22.24	2.76	5.9E-03
BTT27389_1	NA	7.97	2.75	8.6E-03
BTT02752_1	PREDICTED: hypothetical protein LOC100741449	14.99	2.73	2.7E-04
BTT01268_1	NA	15.20	2.73	1.9E-04
BTT16366_1	odorant receptor 47a-like	10.40	2.71	8.3E-03
BTT22351_1	NA	24.96	2.71	1.4E-04
BTT09885_1	NA	100.47	2.71	1.8E-08
BTT12218_1	NA	7.21	2.69	1.4E-02
BTT32044_1	PREDICTED: hypothetical protein LOC100642686	48.83	2.68	2.1E-04
BTT36345_1	low quality protein: cytochrome b5-related	20.64	2.67	1.0E-05
BTT02069_1	NA	35.13	2.66	8.6E-08
BTT00109_1	PREDICTED: hypothetical protein LOC100741425	8.17	2.65	1.8E-02

Table B.11: Top 50 transcripts upregulated in queens compared to reproductive and non-reproductive workers (section 2.3.8).

BTT contig	Functional description	Expression level	log2FC	FDR
BTT40693_1	serine protease inhibitor 3 4-like	4,166.55	3.15	1.4E-15
BTT38705_1	NA	526.30	2.85	3.1E-14
BTT00340_1	PREDICTED: hypothetical protein LOC100651685	131.03	2.71	1.4E-15
BTT14993_1	venom serine protease inhibitor	10,596.15	2.34	1.7E-08
BTT40644_1	pol protein	29.20	2.19	2.3E-03
BTT04508_1	serine protease inhibitor 3 4-like	5,034.27	2.14	6.4E-22
BTT02527_1	NA	56.15	2.02	6.5E-05
BTT04907_1	15-hydroxyprostaglandin dehydrogenase	1,025.73	1.80	2.1E-04
BTT20587_1	glutathione s-transferase	73.87	1.77	2.4E-02
BTT19248_1	NA	193.11	1.68	3.6E-08
BTT37749	serine protease inhibitor 3 4-like	7,074.30	1.63	1.7E-08
BTT41025_1	serine protease inhibitor 3 4-like	2,144.76	1.55	1.0E-10
BTT17812_1	leucine-rich repeats and immunoglobulin-like domains protein 3-like	8,246.87	1.54	8.5E-04
BTT27640_1	PREDICTED: hypothetical protein LOC100645125	326.10	1.50	4.5E-08
BTT32024_1	NA	60.43	1.49	7.0E-03
BTT09752_1	NA	935.74	1.47	3.2E-09
BTT09325_1	udp-glucuronosyltransferase 1-8-like	539.44	1.46	5.3E-08
BTT04871_2	cytochrome p450 4c1-like	104.83	1.44	1.6E-04
BTT36768_1	low quality protein: vitellogenin-like	47.32	1.44	2.4E-02
BTT10432_1	niemann-pick c1	150.60	1.31	6.7E-04
BTT38040_1	calmodulin-lysine n-methyltransferase-like	56.50	1.30	2.2E-02
BTT20281_1	NA	2,786.40	1.28	1.2E-07
BTT18070_1	leucine-rich repeats and immunoglobulin-like domains protein 3-like	1,451.46	1.26	1.3E-06
BTT17660_1	leucine-rich repeats and immunoglobulin-like domains protein 3-like	1,450.59	1.22	4.7E-06
BTT24408_1	low quality protein: vitellogenin-like	245,471.84	1.13	2.6E-03
BTT08885_1	atp-binding cassette sub-family g member 1-like	779.98	1.07	1.1E-02
BTT10800_1	collagen alpha-1 chain-like	408.62	1.05	8.0E-04
BTT12842_1	niemann-pick c1	175.54	1.04	1.6E-02
BTT20208_1	maternal protein exuperantia-like	120.49	1.01	3.5E-02
BTT15035_1	collagen alpha-1 chain-like	206.16	0.97	1.2E-02
BTT02127_1	transposase homolog	145.43	0.96	3.4E-02
BTT16132_1	NA	254.89	0.95	1.1E-02
BTT13907_1	growth differentiation factor 8	273.52	0.94	4.5E-02
BTT09930_1	PREDICTED: hypothetical protein LOC100646805	559.97	0.94	1.8E-02
BTT24237_1	wd repeat and hmg-box dna-binding protein 1-like	1,270.58	0.93	4.8E-03
BTT08050_1	glucose dehydrogenase	3,441.18	0.82	2.1E-02
BTT20759_1	translation initiation factor if-2	534.66	0.81	4.0E-02
BTT15842_1	protein takeout-like	1,509.47	0.80	2.0E-02
BTT41865_1	NA	5,293.46	0.79	2.2E-02
BTT08820_1	innexin inx1-like	1,564.05	0.75	4.5E-02

Table B.12: Top 50 transcripts upregulated in males compared to non-reproductive workers (section 2.3.9).

BTT contig	Functional description	Expression level	log2FC	FDR
BTT24171_1	NA	4.69	Inf	3.7E-02
BTT29399_1	NA	5.87	Inf	1.2E-02
BTT15379_1	NA	11.02	Inf	4.4E-04
BTT42225_1	NA	12.03	Inf	3.1E-05
BTT40182_1	NA	13.83	Inf	5.7E-06
BTT36149_1	NA	25.76	Inf	7.6E-11
BTT22362_1	NA	30.16	Inf	2.0E-12
BTT17081_1	NA	64.11	Inf	2.7E-08
BTT23808_1	NA	74.27	Inf	1.5E-05
BTT28866_1	NA	85.26	Inf	1.0E-12
BTT37171_1	NA	148.76	Inf	1.8E-30
BTT31292_1	NA	209.47	Inf	2.0E-42
BTT17425_1	NA	576.62	Inf	8.4E-26
BTT12846_1	NA	4,894.97	Inf	1.6E-49
BTT18620_1	NA	15,301.16	Inf	6.0E-104
BTT21572_1	PREDICTED: hypothetical protein LOC100643785	49.74	Inf	1.0E-05
BTT12359_1	PREDICTED: hypothetical protein LOC100647675	168.40	Inf	4.0E-05
BTT24613_1	PREDICTED: hypothetical protein LOC100649357	137.50	Inf	2.4E-34
BTT28182_1	PREDICTED: hypothetical protein LOC100649357	76.43	Inf	2.6E-18
BTT09232_1	PREDICTED: hypothetical protein LOC100742192	11,659.04	Inf	2.4E-119
BTT10278_1	PREDICTED: hypothetical protein LOC100747860	1,268.59	Inf	1.2E-09
BTT39595_1	fatty-acid amide hydrolase 2	6.24	Inf	9.7E-03
BTT28624_1	63 kda sperm flagellar membrane protein	12.21	Inf	3.9E-03
BTT41628_1	bcl2 adenovirus e1b 19 kda protein-interacting protein 3-like	16.43	Inf	3.4E-07
BTT27452_1	endocuticle structural glycoprotein bd-1-like	13.21	Inf	9.1E-06
BTT32917_1	protein spaetzle	34.87	Inf	8.3E-14
BTT10518_1	PREDICTED: hypothetical protein LOC100648786	1,031.25	12.09	4.3E-72
BTT34444_1	PREDICTED: hypothetical protein LOC100644579	535.58	11.15	8.9E-30
BTT18790_1	PREDICTED: hypothetical protein LOC100742070	2,611.30	11.11	1.0E-38
BTT14324_1	cuticular protein 19 precursor	2,568.27	10.96	2.4E-54
BTT27891_1	NA	387.60	10.38	8.9E-53
BTT20903_1	NA	13,706.62	10.32	2.7E-99
BTT09964_1	PREDICTED: hypothetical protein LOC100643884	979.44	9.72	2.3E-67
BTT18540_1	PREDICTED: hypothetical protein LOC100643884	573.88	9.36	7.2E-58
BTT29904_1	PREDICTED: hypothetical protein LOC100644579	254.47	9.07	2.8E-43
BTT12572_1	PREDICTED: hypothetical protein LOC100648148	4,101.46	9.06	1.7E-05
BTT08599_1	dual specificity testis-specific protein kinase 2-like	11,560.54	8.79	1.1E-83
BTT17626_1	tripartite motif-containing protein 2-like	694.79	8.48	9.5E-35
BTT06788_1	PREDICTED: hypothetical protein LOC100643957	426.40	8.37	7.9E-06
BTT42146_1	alpha-l1 nicotinic acetyl choline receptor	227.40	8.12	5.7E-14
BTT08735_1	PREDICTED: hypothetical protein LOC100650303	207.78	8.09	7.3E-11
BTT20742_1	NA	758.55	8.07	7.2E-57
BTT35328_1	hypothetical protein EAG_01178	75.86	8.02	1.1E-09
BTT00092_1	PREDICTED: hypothetical protein LOC100646104	5,173.23	7.96	1.3E-26
BTT20725_1	PREDICTED: hypothetical protein LOC100652047	2,105.88	7.83	6.0E-03
BTT17394_1	zinc finger protein 512b-like	231.92	7.78	8.3E-23
BTT20509_1	PREDICTED: hypothetical protein LOC100649478	381.25	7.77	2.1E-03
BTT29283_1	NA	6,791.37	7.67	6.7E-05
BTT00661_1	fatty acyl- reductase cg5065-like	331.69	7.66	2.0E-02
BTT12750_1	NA	144.24	7.56	4.8E-06

Table B.13: 27 locations where worker bumblebees were caught (section 3.2.2).

Location	Coordinates	
Barry	51°23'30.49" N	3°16'20.22" W
Blackpool	53°49'29.07" N	3°00'33.60" W
Bolton Abbey	53°59'37.03" N	1°52'57.29" W
Bristol	51°27'19.00" N	2°35'13.11" W
Batley	53°44'03.88" N	1°39'30.00" W
Bawtry	53°25'27.29" N	1°15'25.37" W
Cambridge	52°11'59.70" N	0°06'54.62" E
Cardiff	51°28'40.71" N	3°10'07.78" W
Corby	52°29'37.60" N	0°42'44.27" W
Derby	52°55'51.28" N	1°29'46.42" W
Dyffryn Ardudwy	52°47'28.52" N	4°05'50.29" W
Ely	52°23'21.08" N	0°16'13.98" E
Farcet	52°31'34.70" N	0°11'02.20" W
Harlech	52°51'34.24" N	4°06'30.76" W
Heysham	54°02'52.73" N	2°53'55.08" W
Kettering	52°23'24.76" N	0°42'27.25" W
Lancaster	54°02'59.09" N	2°48'16.11" W
Leicester	52°37'23.75" N	1°07'37.36" W
Leeds	53°49'25.73" N	1°34'49.50" W
Loughborough	52°47'02.50" N	1°13'05.04" W
Market Harborough	52°28'34.01" N	0°55'42.53" W
Minffordd	52°55'14.00" N	4°05'02.20" W
Melton Mowbray	52°45'45.61" N	0°53'08.23" W
Newport	51°35'18.74" N	2°59'38.58" W
Rutland Water	52°38'25.42" N	0°40'47.79" W
Wellingborough	52°17'43.48" N	0°40'26.52" W
York	53°57'20.74" N	1°04'30.79" W

Appendix C

Perl scripts

C.1 BED_file_creator.pl

BED file as input for vcftool is created from a gene list and gff file.

```
#!/usr/bin/perl
use strict; use warnings;
#Produce BED file containing coordinates for genes in a gene list
my ($infile, $outfile) = @ARGV; #$infile is gene list and $outfile is the
    name of the output BED file
open GENES, "<$infile";
chomp(my @genes = <GENES>);

open INFO, "<Terrestris_genomic.gff";#gff file containing coordinates
chomp(my @info = <INFO>);

open OUT, ">$outfile";
print OUT "Chrom\tstart\tend\n";

shift @genes;

foreach my $gene(@genes){
   foreach my $info(@info){
      if($info =~ /^(\S+)\s+.+gene\s+(\d+)\s+(\d+)\s+.+$gene/){
         print OUT "$1\t$2\t$3\n";
      }
   }
}
```

C.2 min_alleles.pl

With this script the minimum allele sample size is calculated per gene. The output of this script is then used to get gene coordinates (script C.3) in order to reduce vcf files to contain genes with a certain minimum allele sample size.

```
#!/usr/bin/perl
use strict; use warnings;

#get minimum allele sample size per gene. This is important for
    resampling, in which a whole gene must be removed if one of the SNPs
    has a low sample size
open GFF, "<Terrestris_genomic.gff";
chomp(my @gff = <GFF>);

my ($infile, $outfile) = @ARGV;#input is frequency count file from
    vcftools
open SNP, "<$infile";
chomp(my @snps = <SNP>);

open OUT, ">$outfile";

#store gene ranges in hash of arrays with gene name as hash key and
    ranges as elements in an array.
my %gene_chromosome;
my %start;
my %end;

foreach my $gff(@gff){
   if($gff =~ /^(\S+)\s+.+gene\s+(\d+)\s+(\d+)\s+.+Name=(\S+?);.+$/){
#$1 = chromsosome; $2 = start; $3 = end; $4 = mrna name
      push @{$gene_chromosome{$1}}, $4;#group gene names by chromosome
      $start{$4} = $2;
      $end{$4} = $3;#in order to find correct gene for a SNP we can test
          if > $start and < $end
   }
}

my $chromosome;
my $snp_pos;
```

```
my $allele;
my $gene;
my %min;

shift @snps;

foreach my $snp(@snps){
   if($snp =~ /^(\S+)\s+(\d+)\s+\S+\s+(\S+)\s+.+$/){
#$1: chromosome; $2: position; $3: number of alleles
      $chromosome = $1;
      $snp_pos = $2;
      $allele = $3;
      foreach $gene(@{$gene_chromosome{$chromosome}}){#loop through all
          genes on this chromosome
         if($snp_pos >= $start{$gene} && $snp_pos <= $end{$gene}){#to find
             correct gene
            if(exists $min{$gene}){
               if($min{$gene} > $allele){
                  $min{$gene} = $allele;
               }
            }
            else{
               $min{$gene} = $allele;
            }
         }
      }
   }
}

foreach my $key(sort keys %min){
   print OUT "$key\t$min{$key}\n";
}
```

C.3 get_coordinates_min_alleles.pl

Get coordinates for genes with a certain minimum allele sample size. Requires input created from C.2

```
#!/usr/bin/perl
use strict; use warnings;

#get coordinates genes with a minimum allele sample size
my ($infile, $number, $outfile) = @ARGV; #gene list with min number of
    alleles (output from min_alleles.pl), minimum allele sample size and
    output file name
open GENES, "<$infile";
chomp(my @genes = <GENES>);
my @list;#list of genes with minimum number of alleles
foreach(@genes){
   if($_ =~ /^(\S+)\s+(\S+)$/){
      if($2 >= $number){#
         push @list, $1;
      }
   }
}

open INFO, "<Terrestris_genomic.gff";
chomp(my @info = <INFO>);

open OUT, ">$outfile";
print OUT "Chrom\tstart\tend\n";

shift @genes;

foreach my $list(@list){
   foreach my $info(@info){
      if($info =~ /^(\S+)\s+.+.+gene\s+(\d+)\s+(\d+)\s+.+$list/){
         print OUT "$1\t$2\t$3\n";
      }

   }
}
```

C.4 vcf_sampler.pl

With this script an optional number of alleles are randomly sampled at each site

```
#!/usr/bin/perl
use strict; use warnings;
use List::Util 'shuffle';

###sample alleles at each position in a vcf and output frequency file
my ($infile, $number, $outfile) = @ARGV; # vcf file, number of alleles to
    be sampled and output file
open VCF, "<$infile";
chomp(my @vcf = <VCF>);

open OUT, ">$outfile";
print OUT "Chrom\tpos\tnumber_alleles\n";

#my %gt;#for storing genotypes at each SNP
my @gt;
my $chrom;
my $pos;
my $vcf;
my @shuffled_indexes;
my @pick_indexes;
my @sample;
my $sample_seq;
my $count;
my $sample;
my $index = $number - 1;

foreach $vcf(@vcf){
   if($vcf =~ /^(\S+)\s+(\d+)\s+.*?(\d)\/(\d):.*$/){#find genotypes (0/0,
       1/1, 0/1, 1/0, or additional alternate alleles) and one at a time
       push the individual alleles into a hash array with chromosome and
       position as hash key
      $chrom = $1;
      $pos = $2;
#      push @{$gt{$chrom}{$pos}}, ($3,$4);
      push @gt, ($3,$4);
      $vcf =~ s/\d\/\d//;#remove genotype from original string
```

```
      while($vcf =~ /.*?(\d)\/(\d):.*$/){#keep adding alleles to array
         until all removed
#      push @{$gt{$chrom}{$pos}}, ($1,$2);
      push @gt, ($1,$2);
      $vcf =~ s/(\d)\/(\d)//;
      }
####sample requested number of those alleles
      @shuffled_indexes = shuffle(0..$#gt);
      @pick_indexes = @shuffled_indexes[ 0 .. $index ];
      @sample = @gt[ @pick_indexes ];
#now get the minor allele count
      $count = 0;
      foreach $sample(@sample){
         if($sample > 0){
            $count += 1;
         }
      }
#      if($count > $number/2){
#         $count = $number - $count;
#      }
      print OUT "$chrom\t$pos\t$count\n";
   }
}
```

C.5 degenerate_site_finder.pl

This script detects all 4- and 0-fold coordinates from a fasta file containing mRNA sequences

```
#!/usr/bin/perl
use strict;
use warnings;
use List::MoreUtils qw/ uniq /;

###split sequences into 4fold and 0fold sites
# Four-fold degenerate only at 3rd postion if CT-, GT-, -C-, CG-, GG-
# Zero-fold at 3rd position for TGG and ATG only
# Zero-fold degenerate at 2nd position always
# Watch out for STOP codons at TAA, TAG and TGA

# Create hash each for 4-fold, 0-fold 1st position, and 0-fold 3rd
    position.

my %four = (
#CT-
   CTA => '',
   CTC => '',
   CTG => '',
   CTT => '',
#GT-
   GTA => '',
   GTC => '',
   GTG => '',
   GTT => '',
#CG-
   CGA => '',
   CGC => '',
   CGG => '',
   CGT => '',
#GG-
   GGA => '',
   GGC => '',
   GGG => '',
   GGT => '',
```

```
#-C-
   CCA => '',
   CCC => '',
   CCG => '',
   CCT => '',
   GCA => '',
   GCC => '',
   GCG => '',
   GCT => '',
   ACA => '',
   ACC => '',
   ACG => '',
   ACT => '',
   TCA => '',
   TCC => '',
   TCG => '',
   TCT => ''
);
# Zero-fold degenerate at 1st position if: G--, AT-, AC-, AA-, AG(T/C),
    CC-, CA-, CT(T/C), CG(T/C), TC-, TT(T/C), TA(T/C), TG(T/C/G)
my %zero = (
#1st codon position zero-degenerate
#G--
   GAA => '',
   GAC => '',
   GAG => '',
   GAT => '',
   GCA => '',
   GCC => '',
   GCG => '',
   GCT => '',
   GGA => '',
   GGC => '',
   GGG => '',
   GGT => '',
   GTA => '',
   GTC => '',
   GTG => '',
   GTT => '',
#AT-
```

```
   ATA => '',
   ATC => '',
   ATG => '',
   ATT => '',
#AC-
   ACA => '',
   ACC => '',
   ACG => '',
   ACT => '',
#AA-
   AAA => '',
   AAC => '',
   AAG => '',
   AAT => '',
#CC-
   CCA => '',
   CCC => '',
   CCG => '',
   CCT => '',
#CA-
   CAA => '',
   CAC => '',
   CAG => '',
   CAT => '',
#TC-
   TCA => '',
   TCC => '',
   TCG => '',
   TCT => '',
#AG-
   AGT => '',
   AGC => '',
#CT-
   CTT => '',
   CTC => '',
#CG-
   CGT => '',
   CGC => '',
#TT-
   TTT => '',
```

```
   TTC => '',
#TA-
   TAT => '',
   TAC => '',
#TG-
   TGC => '',
   TGG => '',
   TGT => ''

);

# Push codons into an array. then go through the array in a foreach loop,
    counting as you go.
# if the codon matches one of the four-fold codons record number of codon
    x 3 for position.
# same for 1st position for zero-fold, all 2nd positions will be kept. if
    codon matches TGG and ATG record 3rd position.
# If it matches a stop codon, stop and write warning of where it is.


open GENES, '<Terrestris_cds.fa';# fasta file containing coding sequences
chomp(my @lines = <GENES>);

open FOUR, '>four_fold';
open ZERO, '>zero_fold';
open STOP, '>prem_stops';

my $gene_string = join('',@lines);
my @genes = split (/>/, $gene_string);

my $gene;
my $name;
my @codons;
my $codon;
my $n;
my @four_positions;
my @zero_positions;
my $position;
my @second_positions;
my $length;
```

```
my %stops;
my %n_codon;

MAIN: foreach $gene(@genes){
#match sequence except first (START) and last (STOP) codon
   if($gene =~ /^(\S+\.\d)ATG(\S+)[ACTG]{3}$/){
      $name = $1;
      if(length($2) % 3 != 0){#only if full codons
print "$name not full codons\n";
         next MAIN;
      }
      $n = 1; #to account for start codon
      $length = length($2) + 3;
      @codons = ($2 =~ m/.../g);
      @four_positions = ();
      @zero_positions = ();
      %n_codon = ();
#now loop through the codons looking for 4-fold sites
      foreach $codon(@codons){
         $n += 1; #starting with second codon
#if premature stop codon, stop loop
         if($codon eq "TAA" || $codon eq "TAG" || $codon eq "TGA"){
            $stops{$name} = ();#record gene with premature stop codon to
                exclude it
         }
         else{
            if(exists $four{$codon}){
               $position = $n * 3;#3rd position in codon
               push @four_positions, $position;
            }
            if(exists $zero{$codon}){
               $position = ($n * 3) - 2;#first position in codon
               push @zero_positions, $position;
            }
            if($codon eq "TGG" || $codon eq "ATG"){
               $position = $n * 3;
               push @zero_positions, $position;
            }
            if($codon =~ /N/){#if it contains an N the codon must be
                removed, i.e. 2nd positions not included below
```

```
                $position = ($n * 3) - 1;
                $n_codon{$position} = ();
#print "$name\tN \= $position\n";
            }
         }
      }
#now to get all 2nd positions
      unless(exists $stops{$name}){
         @second_positions = ();
         for(my $i = 5; $i <= $length; $i += 3 ){
            unless(exists $n_codon{$i}){
               push @second_positions, $i;
            }
         }
         push @zero_positions, @second_positions;
         @zero_positions = sort { $a <=> $b } @zero_positions;
         @zero_positions = uniq @zero_positions;
         print FOUR "$name\t@four_positions\n";
         print ZERO "$name\t@zero_positions\n";
      }
   }

}

foreach(keys %stops){
print STOP "$_\n";
}
```

C.6 fold_splitter.pl

This scripts takes the 4- and 0-fold coordinates outputted from fold_splitter.pl and counts the 4-fold and 0-fold SNPs in a vcf file. Output are two lists (4-fold and 0-fold) containing numbers of SNPs per genes. The number of potential positions per gene are outputted in "*fourfold_lengths" and "*zerofold_lengths". The SNP positions need to be calculated beforehand using chrom_cds_matcher.pl (C.7).

```
#!/usr/bin/perl
use strict;
use warnings;
use IO::Handle;

#from vcf file extract only four-fold and zero-fold SNPs and save in
    separate files. Codons with more than one SNP will be ignored.
    Chromosomal positions of SNPs in vcf are translated to gene positions
    via snp_pos file

my ($infile1, $infile2) = @ARGV;#file 1 is vcf file and file 2 contains
    snp positions

open OUTPUT, '>', "output.txt" or die $!;
open ERROR, '>', "error.txt" or die $!;

STDOUT->fdopen( \*OUTPUT, 'w' ) or die $!;
STDERR->fdopen( \*ERROR, 'w' ) or die $!;

#open 4- and 0-fold positions and put into hash of hashes so we can test
    the existence of a SNP in each of the lists
open FOUR, "<four_fold";
chomp(my @four = <FOUR>);
my %four;
my @temp;
my $gene;
my $temp;
my $codon;

foreach my $four(@four){
   if($four =~ /^(\S+)\s+(\S+.+)$/){
      $gene = $1;
```

```
      @temp = split ' ' , $2;
      foreach $temp(@temp){
#get codon number which will be needed later on
         if($temp % 3 == 0){
            $codon = $temp/3;
         }
         else{
            $codon = int($temp/3) + 1;
         }
         $four{$gene}{$temp} = ($codon);
      }
   }
}

open ZERO, "<zero_fold";
chomp(my @zero = <ZERO>);
my %zero;

foreach my $zero(@zero){
   if($zero =~ /^(\S+)\s+(\S+.+)$/){
#   if($zero =~ /^(XM_012316519.1)\s+(\S+.+)$/){
      $gene = $1;
      @temp = split ' ' , $2;
      foreach $temp(@temp){
         if($temp % 3 == 0){
            $codon = $temp/3;
         }
         else{
            $codon = int($temp/3) + 1;
         }
         $zero{$gene}{$temp} = ($codon);
      }
   }
}



#open specific snp position file
open SNP, "<$infile2";
chomp(my @snp = <SNP>);
```

```
#filter out all double SNPs, i.e. all codons with more than one SNP.
    First store all SNPs by gene name
my %positions;
foreach my $snp(@snp){
   if($snp =~ /^\S+\s+\d+\s+(\S+)\s+(\d+)\s+.+$/){
#   if($snp =~ /^\S+\s+\d+\s+(XM_012316519.1)\s+(\d+)\s+\S\s+\S$/){
      push @{$positions{$1}}, $2; #push all SNPs of the same gene into an
          array stored in a hash with gene as hash key
   }
}
#loop through each SNP per gene and check how close they are. If
    difference is < 3, check if they are in the same codon. Only keep all
    SNPs from separate codons.
my $snp_pos;
my $last_snp;
my %double_snps;
my %double_codon;

foreach my $key(keys %positions){#for each gene
   $last_snp = 0; #reset
   foreach $snp_pos (sort { $a <=> $b } @{$positions{$key}}){
      if($last_snp == 0){
         $last_snp = $snp_pos;#record postion of this SNP for the next
             comparison
      }
      elsif($last_snp != 0 && $snp_pos - $last_snp > 2){#no problem, not
          the same codon
         $last_snp = $snp_pos;
      }
      elsif($last_snp != 0 && $snp_pos - $last_snp == 2 && $snp_pos % 3
          != 0){#if larger number is not divisible by 3, then also no
          problem
         $last_snp = $snp_pos;
      }
      elsif($last_snp != 0 && $snp_pos - $last_snp == 1 && $last_snp % 3
          == 0){
         $last_snp = $snp_pos;
      }
      else{#otherwise the current and the previous SNP are in the same
```

```
         codon
        if($snp_pos % 3 == 0){
           $codon = $snp_pos/3;
        }
        else{
           $codon = int($snp_pos/3) + 1;
        }
        $double_snps{$key}{$snp_pos} = ();#for excluding SNPs
        $double_codon{$key}{$codon} = ();#codon position <-- need this to
            calculate the total potential number of four-fold and
            zero-fold positions later on
        if($last_snp % 3 == 0){
           $codon = $last_snp/3;
        }
        else{
           $codon = int($last_snp/3) + 1;
        }
        $double_snps{$key}{$last_snp} = ();
        $double_codon{$key}{$codon} = ();
     }
  }
}

my %four_snp;
my %zero_snp;

#if snp gene positions are in 4- or 0-fold lists but not in double-snps
    list store them
foreach my $snp(@snp){
   if($snp =~ /^(\S+)\s+(\d+)\s+(\S+)\s+(\d+)\s+.+$/){
#$1 = chrom; $2 = chrom pos; $3 = gene; $4 = gene pos
      unless(exists $double_snps{$3} && exists $double_snps{$3}{$4}){
        if(exists $four{$3} && exists $four{$3}{$4}){
           $four_snp{$1}{$2} = ();
        }
        elsif(exists $zero{$3} && exists $zero{$3}{$4}){
           $zero_snp{$1}{$2} = ();
        }
     }
  }
```

```
}


#now open vcf file and print relevant lines
open VCF, "<$infile1";
$infile1 =~ /^(\S+)\.vcf$/;
my $name = $1;
chomp(my @vcf = <VCF>);

my $four_file = "$name"."_four.vcf";
my $zero_file = "$name"."_zero.vcf";

open FOUR_OUT, ">$four_file";
open ZERO_OUT, ">$zero_file";

foreach my $vcf(@vcf){
   if($vcf =~ /^#/){#print out first comment lines of file
      print FOUR_OUT "$vcf\n";
      print ZERO_OUT "$vcf\n";
   }
   if($vcf =~ /^(\S+)\s+(\d+)\s+.+$/){#print out first comment lines of
       file
      if(exists $four_snp{$1} && exists $four_snp{$1}{$2}){
         print FOUR_OUT "$vcf\n";
      }
      elsif(exists $zero_snp{$1} && exists $zero_snp{$1}{$2}){
         print ZERO_OUT "$vcf\n";
      }
      else{
         print "$vcf\n"
      }
   }
}


####
#calculate total (reduced) number of potential 4- and 0-fold sites per
    gene after removal of double SNPs
#$four{gene}{base position}=(codon position) & $zero{gene}{position}
    contain total 4- and 0-fold sites of each gene together with their
```

```
    codon positions.
#$double_codon{codon position} contains codon positions that need to be
    removed.
#This is why I linked each base position in both hashes to a codon
    position earlier on

my $four_n = 0;
my $zero_n = 0;
my $outer;
my $inner;

my $four_length_file = "$name"."fourfold_lengths";
my $zero_length_file = "$name"."zerofold_lengths";

open FOUR_N, ">$four_length_file";
open ZERO_N, ">$zero_length_file";

foreach $outer(keys %four){#foreach gene
   $four_n = 0;
   foreach $inner(keys %{$four{$outer}}){#count positions if not in
       double_snp codon
      unless(exists $double_codon{$outer} && exists
          $double_codon{$outer}{$four{$outer}{$inner}}){
         $four_n ++;
      }
   }
   print FOUR_N "$outer\t$four_n\n";
}

foreach $outer(keys %zero){#foreach gene
   $zero_n = 0;
   foreach $inner(keys %{$zero{$outer}}){
      unless(exists $double_codon{$outer} && exists
          $double_codon{$outer}{$zero{$outer}{$inner}}){
         $zero_n ++;
      }
   }
   print ZERO_N "$outer\t$zero_n\n";
}
```

C.7 chrom_cds_matcher.pl

Using a vcf file and a gff file genomic SNP positions are given relative gene positions. Produces input for script C.6.

```
#!/usr/bin/perl
use strict; use warnings;

#match chromosome positions of SNPs to base position within CDS sequence
open GFF, "<Terrestris_genomic.gff";
chomp(my @gff = <GFF>);

my ($infile, $outfile) = @ARGV;#VCF file as input
open SNP, "<$infile";
chomp(my @snps = <SNP>);

open OUT, ">$outfile";

#store CDS ranges in hash of arrays with mRNA id as hash key and ranges
    as elements in an array. Also link rna id to rna name
my %ranges;
my $range;
my %rna_name;
my %rna_chromosome;
my %start;
my %end;
my %strand;
my $start;
my $end;
my $chromosome;
my $sum;

foreach my $gff(@gff){
   if($gff =~
       /^(\S+)\s+.+mRNA\s+(\d+)\s+(\d+)\s+\S+\s+(\S+)\s+.+ID=rna(\d+);
       .+Name=(\S+?);.+$/){
#$1 = chromsosome; $2 = start; $3 = end; $4 = strand; $5 = rna id; $6 =
    mrna name
      push @{$rna_chromosome{$1}}, $6;#group mRNA names by chromosome
      $rna_name{$5} = $6;#link mRNA id to mRNA name
```

```
      $start{$6} = $2;
      $end{$6} = $3;#in order to find correct gene for a SNP we can test
          if > $start and < $end
      $strand{$6} = $4;
   }
   elsif($gff =~ /^.+CDS\s+(\d+\s+\d+)\s+.+Parent=rna(\d+);.+$/){
      push @{$ranges{$rna_name{$2}}},$1;#push range into an array with
          hash key of mRNA name
#this should produce an array of ranges for each mRNA name.
   }
}
my $snp_pos;
my $rna;
my $rna_pos;
my $base;

foreach my $snp(@snps){
   if($snp =~ /^(\S+)\s+(\d+)\s+\S+\s+(\S\s+\S,?\S?,?\S?)\s+.+$/){#only
       include single SNPs, not indels
      $chromosome = $1;
      $snp_pos = $2;
      $base = $3;
      foreach $rna(@{$rna_chromosome{$chromosome}}){#loop through all
          genes on this chromosome
         if($snp_pos >= $start{$rna} && $snp_pos <= $end{$rna}){#to find
             correct gene
#now calculate relative position depending on whether gene is on negative
    or positive strand
            $sum = 0;#reset running sum for new gene
###for gene on positive strand
           if($strand{$rna} eq "+"){#if it's on the positive strand
              foreach $range(@{$ranges{$rna}}){#loop through the cds
                  ranges
                 if($range =~ /^(\d+)\s+(\d+)$/){#start and end of cds part
                    $start = $1;
                    $end = $2;
#first test if a cds exon has already been tested for this mRNA

                    if($sum == 0){
                       if($snp_pos >= $start && $snp_pos <= $end){#if the
```

```
                        snp is in the first cds exon then calculate
                        postion and print
                        $rna_pos = $snp_pos - $start + 1;
                        print OUT
                            "$chromosome\t$snp_pos\t$rna\t$rna_pos\t$base\n";
                     }
                     else{
                        $sum = $end - $start + 1;
                     }
                  }
                  else{
                     if($snp_pos >= $start && $snp_pos <= $end){
                        $rna_pos = $snp_pos - $start + 1 + $sum;#add
                            running sum from previous cds exons
                        print OUT
                            "$chromosome\t$snp_pos\t$rna\t$rna_pos\t$base\n";
                     }
                     else{
                        $sum = $sum + ($end - $start + 1);
                     }
                  }

               }
            }
         }

###for gene on negative strand
         if($strand{$rna} eq "-"){#if it's on the negative strand
            foreach $range(@{$ranges{$rna}}){#loop through the cds
                ranges
               if($range =~ /^(\d+)\s+(\d+)$/){#start and end of cds part
                  $start = $1;
                  $end = $2;
#first test if a cds exon has already been tested for this mRNA

                  if($sum == 0){
                     if($snp_pos >= $start && $snp_pos <= $end){#if the
                        snp is in the first cds exon then calculate
                        postion and print
                        $rna_pos = $end - $snp_pos + 1;
```

```
                        print OUT
                            "$chromosome\t$snp_pos\t$rna\t$rna_pos\t$base\n";
                    }
                    else{
                        $sum = $end - $start + 1;
                    }
                }
                else{
                    if($snp_pos >= $start && $snp_pos <= $end){
                        $rna_pos = $end - $snp_pos + 1 + $sum;#add
                            running sum from previous cds exons
                        print OUT
                            "$chromosome\t$snp_pos\t$rna\t$rna_pos\t$base\n";
                    }
                    else{
                        $sum = $sum + ($end - $start + 1);
                    }
                }

            }
        }
    }

  }
 }
}
```

C.8 afs_creator.pl

Allele frequency spectra are created for DoFE analysis.

```
#!/usr/bin/perl
use strict; use warnings;

my ($infile, $outfile) = @ARGV; # freq file and output file

open FREQ, "<$infile";
chomp(my @freq = <FREQ>);

open OUT, ">$outfile";

shift @freq;
my $freq;
my %freq;

foreach $freq(@freq){
   if($freq =~ /^(\S+)\s+(\S+)\s+(\S+)$/){
      if(exists $freq{$3}){
         $freq{$3} += 1;
      }
      else{
         $freq{$3} = 1;
      }
   }
}


foreach (sort { $a <=> $b } keys(%freq)){
    print OUT "$_\t$freq{$_}\n"
}
```

Bibliography

Akashi, H. 2001. Gene expression and molecular evolution. *Current Opinion in Genetics & Development*, 11(6): 660–666.

Alaux, C., Savarit, F., Jaisson, P. and Hefetz, A. 2004. Does the queen win it all? Queen–worker conflict over male production in the bumblebee, Bombus terrestris. *Naturwissenschaften*, 91(8): 400–403.

Alaux, C., Boutot, M., Jaisson, P. and Hefetz, A. 2007. Reproductive plasticity in bumblebee workers (Bombus terrestris)—reversion from fertility to sterility under queen influence. *Behavioral Ecology and Sociobiology*, 62(2): 213–222.

Alexa, A. and Rahnenfuhrer, J. 2010. topGO: Enrichment analysis for Gene Ontology.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17): 3389–3402.

Amdam, G.V., Norberg, K., Hagen, A. and Omholt, S.W. 2003. Social exploitation of vitellogenin. *Proceedings of the National Academy of Sciences, USA*, 100(4): 1799–1802.

Amdam, G.V., Norberg, K., Fondrk, M.K. and Page, R.E. 2004. Reproductive ground plan may mediate colony-level selection effects on individual foraging behavior in honey bees. *Proceedings of the National Academy of Sciences, USA*, 101(31): 11350–11355.

Amdam, G.V., Fennern, E. and Havukainen, H. 2012. Vitellogenin in Honey Bee Behavior and Lifespan. In C. G. Galizia, D. Eisenhardt, and M. Giurfa, editors, *Honeybee Neurobiology and Behavior*, pages 17–29. Springer Netherlands.

Amsalem, E., Twele, R., Francke, W. and Hefetz, A. 2009. Reproductive competition in the bumble-bee Bombus terrestris: do workers advertise sterility? *Proceedings of the Royal Society B: Biological Sciences*, 276(1660): 1295–1304.

Anders, S. and Huber, W. 2010. Differential expression analysis for sequence count data. *Genome Biology*, 11(10): R106.

Andersson, M. 1984. The Evolution of Eusociality. *Annual Review of Ecology and Systematics*, 15: 165–189.

Anisimova, M., Nielsen, R. and Yang, Z. 2003. Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. *Genetics*, 164(3): 1229–1236.

Arakane, Y., Muthukrishnan, S., Beeman, R.W., Kanost, M.R. and Kramer, K.J. 2005. Laccase 2 is the phenoloxidase gene required for beetle cuticle tanning. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32): 11337–11342.

Arunkumar, R., Josephs, E.B., Williamson, R.J. and Wright, S.I. 2013. Pollen-Specific, but Not Sperm-Specific, Genes Show Stronger Purifying Selection and Higher Rates of Positive Selection Than Sporophytic Genes in Capsella grandiflora. *Molecular Biology and Evolution*, 30(11): 2475–2486.

Baer, B. 2003. Bumblebees as model organisms to study male sexual selection in social insects. *Behavioral Ecology and Sociobiology*, 54(6): 521–533.

Baines, J.F. and Harr, B. 2007. Reduced X-Linked Diversity in Derived Populations of House Mice. *Genetics*, 175(4): 1911–1921.

Baines, J.F., Sawyer, S.A., Hartl, D.L. and Parsch, J. 2008. Effects of X-Linkage and Sex-Biased Gene Expression on the Rate of Adaptive Protein Evolution in Drosophila. *Molecular Biology and Evolution*, 25(8): 1639–1650.

Bak, T.G. 1967. Studies on glucose dehydrogenase of Aspergillus oryzae: II. Purification and physical and chemical properties. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 139(2): 277–293.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N. et al 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biology*, 5(11): e310.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. and Mathews, S. 2010. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences, USA*, 107(43): 18724–18728.

Betancourt, A.J., Presgraves, D.C. and Swanson, W.J. 2002. A Test for Faster X Evolution in Drosophila. *Molecular Biology and Evolution*, 19(10): 1816–1819.

Beye, M., Hasselmann, M., Fondrk, M.K., Page Jr., R.E. and Omholt, S.W. 2003. The Gene csd Is the Primary Signal for Sexual Development in the Honeybee and Encodes an SR-Type Protein. *Cell*, 114(4): 419–429.

Bloch, G. 1999. Regulation of queen–worker conflict in bumble bee (Bombus terrestris) colonies. *Proceedings of the Royal Society of London B: Biological Sciences*, 266(1437): 2465–2469.

Boomsma, J.J. 2007. Kin Selection versus Sexual Selection: Why the Ends Do Not Meet. *Current Biology*, 17(16): R673–R683.

Borg, M., Brownfield, L., Khatab, H., Sidorova, A., Lingaya, M. and Twell, D. 2011. The R2r3 MYB Transcription Factor DUO1 Activates a Male Germline-Specific Regulon Essential for Sperm Cell Differentiation in Arabidopsis. *The Plant Cell Online*, 23(2): 534–549.

Borges, F., Gomes, G., Gardner, R., Moreno, N., McCormick, S., Feijó, J.A. and Becker, J.D. 2008. Comparative Transcriptomics of Arabidopsis Sperm Cells. *Plant Physiology*, 148(2): 1168–1181.

Boswell, R.E. and Mahowald, A.P. 1985. tudor, a gene required for assembly of the germ plasm in Drosophila melanogaster. *Cell*, 43(1): 97–104.

Bourke, A.F.G. and Ratnieks, F.L.W. 2001. Kin-selected conflict in the bumble-bee Bombus terrestris (Hymenoptera: Apidae). *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1465): 347–355.

Buckingham, E.N. 1911. Division of Labor among Ants. *Proceedings of the American Academy of Arts and Sciences*, 46(18): 425–508.

Cameron, S.A. 1989. Temporal Patterns of Division of Labor among Workers in the Primitively Eusocial Bumble Bee, Bombus griseocollis (Hymenoptera: Apidae)1). *Ethology*, 80(1-4): 137–151.

Cameron, S.A., Hines, H.M. and Williams, P.H. 2007. A comprehensive phylogeny of the bumble bees (Bombus). *Biological Journal of the Linnean Society*, 91(1): 161–188.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C. et al 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics*, 43(10): 956–963.

Cardoen, D., Wenseleers, T., Ernst, U.R., Danneels, E.L., Laget, D., De Graaf, D.C., Schoofs, L. and Verleyen, P. 2011. Genome-wide analysis of alternative reproductive phenotypes in honeybee workers. *Molecular Ecology*, 20(19): 4070–4084.

Carneiro, M., Albert, F.W., Melo-Ferreira, J., Galtier, N., Gayral, P., Blanco-Aguiar, J.A., Villafuerte, R., Nachman, M.W. and Ferrand, N. 2012. Evidence for Widespread Positive and Purifying Selection Across the European Rabbit (Oryctolagus cuniculus) Genome. *Molecular Biology and Evolution*, 29(7): 1837–1849.

Charif, D. and Lobry, J.R. 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Berlin Heidelberg.

Charlesworth, B., Coyne, J.A. and Barton, N.H. 1987. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *The American Naturalist*, 130(1): 113–146.

Charlesworth, D. and Charlesworth, B. 1992. The Effects of Selection in the Gametophyte Stage on Mutational Load. *Evolution*, 46(3): 703–720.

Choy, R.K.M. and Thomas, J.H. 1999. Fluoxetine-Resistant Mutants in C. elegans Define a Novel Family of Transmembrane Proteins. *Molecular Cell*, 4(2): 143–152.

Cnaani, J., Borst, D.W., Huang, Z.Y., Robinson, G.E. and Hefetz, A. 1997. Caste Determination in Bombus terrestris: Differences in Development and Rates of JH Biosynthesis between Queen and Worker Larvae. *Journal of Insect Physiology*, 43(4): 373–381.

Colgan, T.J., Carolan, J.C., Bridgett, S.J., Sumner, S., Blaxter, M.L. and Brown, M.J. 2011. Polyphenism in social insects: insights from a transcriptome-wide analysis of gene expression in the life stages of the key pollinator, Bombus terrestris. *BMC Genomics*, 12(1): 623.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M. and Robles, M. 2005. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18): 3674–3676.

Connallon, T. 2007. Adaptive Protein Evolution of X-linked and Autosomal Genes in Drosophila: Implications for Faster-X Hypotheses. *Molecular Biology and Evolution*, 24(11): 2566–2572.

Corona, M., Libbrecht, R., Wurm, Y., Riba-Grognuz, O., Studer, R.A. and Keller, L. 2013. Vitellogenin Underwent Subfunctionalization to Acquire Caste and Behavioral Specific Expression in the Harvester Ant Pogonomyrmex barbatus. *PLoS Genetics*, 9(8): e1003730.

Counterman, B.A., Ortíz-Barrientos, D. and Noor, M.A.F. 2004. Using Comparative Genomic Data to Test for Fast-X Evolution. *Evolution*, 58(3): 656–660.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. et al 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–2158.

Dauwalder, B., Tsujimoto, S., Moss, J. and Mattox, W. 2002. The Drosophila takeout gene is regulated by the somatic sex-determination pathway and affects male courtship behavior. *Genes & Development*, 16(22): 2879–2892.

Detrain, C. and Pasteels, J.M. 1992. Caste polyethism and collective defense in the ant, Pbeidole pallidula: the outcome of quantitative differences in recruitment. *Behavioral Ecology and Sociobiology*, 29(6): 405–412.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1): 15–21.

Dohlman, H.G. 2002. G Proteins and Pheromone Signaling. *Annual Review of Physiology*, 64(1): 129–152.

Drummond, D.A., Raval, A. and Wilke, C.O. 2006. A Single Determinant Dominates the Rate of Yeast Protein Evolution. *Molecular Biology and Evolution*, 23(2): 327–337.

Duchateau, M.J. and Mariën, J. 1995. Sexual biology of haploid and diploid males in the bumble beeBombus terrestris. *Insectes Sociaux*, 42(3): 255–266.

Duchateau, M.J. and Velthuis, H.H.W. 1988. Development and Reproductive Strategies in Bombus terrestris Colonies. *Behaviour*, 107(3/4): 186–207.

Duret, L. and Mouchiroud, D. 2000. Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Molecular Biology and Evolution*, 17(1): 68–70.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5): 1792–1797.

Ellegren, H. 2007. Characteristics, causes and evolutionary consequences of male-biased mutation. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1606): 1–10.

Ersfeld, K., Wehland, J., Plessmann, U., Dodemont, H., Gerke, V. and Weber, K. 1993. Characterization of the tubulin-tyrosine ligase. *The Journal of Cell Biology*, 120(3): 725–732.

Eyre-Walker, A. and Keightley, P.D. 2007. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8): 610–618.

Eyre-Walker, A. and Keightley, P.D. 2009. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, 26(9): 2097–2108.

Feldmeyer, B., Elsner, D. and Foitzik, S. 2014. Gene expression patterns associated with caste and reproductive status in ants: worker-specific genes are more derived than queen-specific ones. *Molecular Ecology*, 23(1): 151–161.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Foster, R.L., Brunskill, A., Verdirame, D. and O'Donnell, S. 2004. Reproductive physiology, dominance interactions, and division of labour among bumble bee workers. *Physiological Entomology*, 29(4): 327–334.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T. et al 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, 477(7365): 419–423.

Gossmann, T.I., Song, B.H., Windsor, A.J., Mitchell-Olds, T., Dixon, C.J., Kapralov, M.V., Filatov, D.A. and Eyre-Walker, A. 2010. Genome Wide Analyses Reveal Little Evidence for Adaptive Evolution in Many Plant Species. *Molecular Biology and Evolution*, 27(8): 1822–1832.

Gossmann, T.I., Schmid, M.W., Grossniklaus, U. and Schmid, K.J. 2013. Selection-Driven Evolution of Sex-Biased Genes Is Consistent with Sexual Selection in Arabidopsis thaliana. *Molecular Biology and Evolution*, pages 574–583.

Goulson, D. 2010. *Bumblebees: behaviour, ecology, and conservation*. Oxford University Press, Oxford, 2 edition.

Goulson, D., Peat, J., Stout, J.C., Tucker, J., Darvill, B., Derwent, L.C. and Hughes, W.O.H. 2002. Can alloethism in workers of the bumblebee, Bombus terrestris, be explained in terms of foraging efficiency? *Animal Behaviour*, 64(1): 123–130.

Grozinger, C.M., Sharabash, N.M., Whitfield, C.W. and Robinson, G.E. 2003. Pheromone-mediated gene expression in the honey bee brain. *Proceedings of the National Academy of Sciences, USA*, 100(suppl 2): 14519–14525.

Grozinger, C.M., Fan, Y., Hoover, S.E.R. and Winston, M.L. 2007. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (Apis mellifera). *Molecular Ecology*, 16(22): 4837–4848.

Hagai, T., Cohen, M. and Bloch, G. 2007. Genes encoding putative Takeout/juvenile hormone binding proteins in the honeybee (Apis mellifera) and modulation by age and juvenile hormone of the takeout-like gene GB19811. *Insect Biochemistry and Molecular Biology*, 37(7): 689–701.

Haldane, J. 1924. A mathematical theory of natural and artificial selection. Part I. *Transactions of the Cambridge Philosophical Society*, (23): 19–41.

Haldane, J.B.S. 1922. Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*, 12(2): 101–109.

Harrison, M.C., Hammond, R.L. and Mallon, E.B. 2015. Reproductive workers show queenlike gene expression in an intermediately eusocial insect, the buff-tailed bumble bee Bombus terrestris. *Molecular Ecology*, 24(12): 3043–3063.

Hedrick, P.W. and Parker, J.D. 1997. Evolutionary Genetics and Genetic Variation of Haplodiploids and X-Linked Genes. *Annual Review of Ecology and Systematics*, 28: 55–83.

Heimpel, G.E. and Boer, J.G.d. 2008. Sex Determination in the Hymenoptera. *Annual Review of Entomology*, 53(1): 209–230.

Hoffman, D.R. and Jacobson, R.S. 1996. Allergens in Hymenoptera venom XXVII: Bumblebee venom allergy and allergens. *Journal of Allergy and Clinical Immunology*, 97(3): 812–821.

Hoffman, E.A. and Goodisman, M.A. 2007. Gene expression and the evolution of phenotypic diversity in social wasps. *BMC Biology*, 5(1): 23.

Honys, D. and Twell, D. 2004. Transcriptome analysis of haploid male gametophyte development in Arabidopsis. *Genome Biology*, 5: R85.

Hooper, S.L. and Thuma, J.B. 2005. Invertebrate Muscles: Muscle Specific Genes and Proteins. *Physiological Reviews*, 85(3): 1001–1060.

Hunt, B.G., Ometto, L., Wurm, Y., Shoemaker, D., Yi, S.V., Keller, L. and Goodisman, M.A.D. 2011. Relaxed selection is a precursor to the evolution of phenotypic plasticity. *Proceedings of the National Academy of Sciences*, 108(38): 15936–15941.

Hvilsom, C., Qian, Y., Bataillon, T., Li, Y., Mailund, T., Sallé, B., Carlsen, F., Li, R., Zheng, H., Jiang, T. et al 2012. Extensive X-linked adaptive evolution in central chimpanzees. *Proceedings of the National Academy of Sciences, USA*, 109(6): 2054–2059.

Jr, F.E.H. et al 2015. Hmisc: Harrell Miscellaneous.

Kawakita, A., Sota, T., Ito, M., Ascher, J.S., Tanaka, H., Kato, M. and Roubik, D.W. 2004. Phylogeny, historical biogeography, and character evolution in bumble bees (Bombus: Apidae) based on simultaneous analysis of three nuclear gene sequences. *Molecular Phylogenetics and Evolution*, 31(2): 799–804.

Kemphues, K.J., Raff, R.A., Kaufman, T.C. and Raff, E.C. 1979. Mutation in a structural gene for a beta-tubulin specific to testis in Drosophila melanogaster. *Proceedings of the National Academy of Sciences, USA*, 76(8): 3991–3995.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M. and Pääbo, S. 2005. Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees. *Science*, 309(5742): 1850–1854.

Kim, S. 2012. ppcor: Partial and Semi-partial (Part) correlation.

Kotlar, D. and Lavner, Y. 2006. The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids. *BMC Genomics*, 7: 67.

Kraus, F.B., Wolf, S. and Moritz, R.F.A. 2009. Male flight distance and population substructure in the bumblebee Bombus terrestris. *Journal of Animal Ecology*, 78(1): 247–252.

Kubota, M., Tsuji, M., Nishimoto, M., Wongchawalit, J., Okuyama, M., Mori, H., Matsui, H., Surarit, R., Svasti, J., Kimura, A. et al 2004. Localization of alpha-Glucosidases I, II, and III in Organs of European Honeybees, Apis mellifera L., and the Origin of alpha-Glucosidase in Honey. *Bioscience, Biotechnology, and Biochemistry*, 68(11): 2346–2352.

Liao, B.Y., Scott, N.M. and Zhang, J. 2006. Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins. *Molecular Biology and Evolution*, 23(11): 2072–2080.

Love, M.I., Huber, W. and Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12): 550.

Lu, J. and Wu, C.I. 2005. Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. *Proceedings of the National Academy of Sciences of the United States of America*, 102(11): 4063–4067.

Mandel, J. 1982. Use of the Singular Value Decomposition in Regression Analysis. *The American Statistician*, 36(1): 15–24.

Mank, J.E., Axelsson, E. and Ellegren, H. 2007. Fast-X on the Z: Rapid evolution of sex-linked genes in birds. *Genome Research*, 17(5): 618–624.

Mank, J.E., Vicoso, B., Berlin, S. and Charlesworth, B. 2010a. Effective Population Size and the Faster-X Effect: Empirical Results and Their Interpretation. *Evolution*, 64(3): 663–674.

Mank, J.E., Nam, K. and Ellegren, H. 2010b. Faster-Z Evolution Is Predominantly Due to Genetic Drift. *Molecular Biology and Evolution*, 27(3): 661–670.

Mazer, S.J., Hove, A.A., Miller, B.S. and Barbet-Massin, M. 2010. The joint evolution of mating system and pollen performance: Predictions regarding male gametophytic evolution in selfers vs. outcrossers. *Perspectives in Plant Ecology, Evolution and Systematics*, 12(1): 31–41.

McDonald, J.H. and Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328): 652–654.

Meisel, R.P. and Connallon, T. 2013. The faster-X effect: integrating theory and data. *Trends in Genetics*, 29(9): 537–544.

Mevik, B.h. and Wehrens, R. 2007. The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, pages 1–24.

Morandin, C., Havukainen, H., Kulmuni, J., Dhaygude, K., Trontti, K. and Helanterä, H. 2014. Not Only for Egg Yolk—Functional and Evolutionary Insights from Expression, Selection, and Structural Analyses of Formica Ant Vitellogenins. *Molecular Biology and Evolution*, 31(8): 2181–2193.

Musters, H., Huntley, M.A. and Singh, R.S. 2006. A Genomic Comparison of Faster-Sex, Faster-X, and Faster-Male Evolution Between Drosophila melanogaster and Drosophila pseudoobscura. *Journal of Molecular Evolution*, 62(6): 693–700.

Nielsen, R. 2005. Molecular Signatures of Natural Selection. *Annual Review of Genetics*, 39(1): 197–218.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J. et al 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. *PLoS Biology*, 3(6): e170.

Nipitwattanaphon, M., Wang, J., Ross, K.G., Riba-Grognuz, O., Wurm, Y., Khurewathanakul, C. and Keller, L. 2014. Effects of ploidy and sex-locus genotype on gene expression patterns in the fire ant Solenopsis invicta. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1797): 20141776.

Nordborg, M. 2000. Linkage Disequilibrium, Gene Trees and Selfing: An Ancestral Recombination Graph With Partial Self-Fertilization. *Genetics*, 154(2): 923–929.

Nowak, M.A., Tarnita, C.E. and Wilson, E.O. 2010. The evolution of eusociality. *Nature*, 466(7310): 1057–1062.

Ohashi, K., Natori, S. and Kubo, T. 1999. Expression of amylase and glucose oxidase in the hypopharyngeal gland with an age-dependent role change of the worker honeybee (Apis mellifera L.). *European Journal of Biochemistry*, 265(1): 127–133.

Ometto, L., Shoemaker, D., Ross, K.G. and Keller, L. 2011. Evolution of Gene Expression in Fire Ants: The Effects of Developmental Stage, Caste, and Species. *Molecular Biology and Evolution*, 28(4): 1381–1392.

Otto, S.P. and Goldstein, D.B. 1992. Recombination and the evolution of diploidy. *Genetics*, 131(3): 745–751.

Paradis, E. 2010. pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26(3): 419–420.

Partridge, L., Ewing, A. and Chandler, A. 1987. Male size and mating success in Drosophila melanogaster: the roles of male and female behaviour. *Animal Behaviour*, 35(2): 555–562.

Pereboom, J.J.M. 2000. The composition of larval food and the significance of exocrine secretions in the bumblebee Bombus terrestris. *Insectes Sociaux*, 47(1): 11–20.

Pereboom, J.J.M., Velthuis, H.H.W. and Duchateau, M.J. 2003. The organisation of larval feeding in bumblebees (Hymenoptera, Apidae) and its significance to caste differentiation. *Insectes Sociaux*, 50(2): 127–133.

Pereboom, J.J.M., Jordan, W.C., Sumner, S., Hammond, R.L. and Bourke, A.F.G. 2005. Differential gene expression in queen–worker caste determination in bumblebees. *Proceedings of the Royal Society B: Biological Sciences*, 272(1568): 1145–1152.

Perriere, D.C.a.J.R.L.a.A.N.a.L.P.a.S.P.a.G. 2014. seqinr: Biological Sequences Retrieval and Analysis.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E. and Lercher, M.J. 2014. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution*, 31(7): 1929–1936.

Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Ågren, J., Bossdorf, O., Byers, D., Donohue, K. et al 2010. The Scale of Population Structure in Arabidopsis thaliana. *PLoS Genet*, 6(2): e1000843.

Prys-Jones, O.E. and Corbet, S.A. 1987. *Bumblebees*. Cambridge University Press.

Pröschel, M., Zhang, Z. and Parsch, J. 2006. Widespread Adaptive Evolution of Drosophila Genes With Sex-Biased Expression. *Genetics*, 174(2): 893–900.

Qin, Y., Leydon, A.R., Manziello, A., Pandey, R., Mount, D., Denic, S., Vasic, B., Johnson, M.A. and Palanivelu, R. 2009. Penetration of the Stigma and Style Elicits a Novel Transcriptome in Pollen Tubes, Pointing to Genes Critical for Growth in a Pistil. *PLOS Genet*, 5(8): e1000621.

Reeve, H.K., Starks, P.T., Peters, J.M. and Nonacs, P. 2000. Genetic support for the evolutionary theory of reproductive transactions in social wasps. *Proceedings of the Royal Society B: Biological Sciences*, 267(1438): 75–79.

Riddell, C.E., Lobaton Garces, J.D., Adams, S., Barribeau, S.M., Twell, D. and Mallon, E.B. 2014. Differential gene expression and alternative splicing in insect immune specificity. *BMC Genomics*, 15: 1031.

Rost, B. 1999. Twilight zone of protein sequence alignments. *Protein Engineering*, 12(2): 85–94.

Röseler, P.F. 1973. Die anzahl der spermien im receptaculum seminis von hummelköniginnen (Hym., Apoidea, Bombinae). *Apidologie*, (4.3): 267–274.

Röseler, P.F., Röseler, I. and Honk, C.G.J.v. 1981. Evidence for inhibition of corpora allata activity in workers of Bombus terrestris by a pheromone from the queen's mandibular glands. *Experientia*, 37(4): 348–351.

Sadd, B.M., Barribeau, S.M., Bloch, G., Graaf, D.C.d., Dearden, P., Elsik, C.G., Gadau, J., Grimmelikhuijzen, C.J., Hasselmann, M., Lozier, J.D. et al 2015. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biology*, 16(1): 1–32.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J. et al 2013. Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430): 45–50.

Schmid-Hempel, R. and Schmid-Hempel, P. 2000. Female mating frequencies in Bombus spp. from Central Europe. *Insectes Sociaux*, 47(1): 36–41.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J. et al 2013. Patterns of population epigenomic diversity. *Nature*, 495(7440): 193–198.

Singh, N.D., Larracuente, A.M. and Clark, A.G. 2008. Contrasting the Efficacy of Selection on the X and Autosomes in Drosophila. *Molecular Biology and Evolution*, 25(2): 454–467.

Smeets, P. and Duchateau, M. 2003. Longevity of Bombus terrestris workers (Hymenoptera: Apidae) in relation to pollen availability, in the absence of foraging. *Apidologie*, 34(4): 333–337.

Smith, N.G.C. and Eyre-Walker, A. 2002. Adaptive protein evolution in Drosophila. *Nature*, 415(6875): 1022–1024.

Stekel, D.J., Git, Y. and Falciani, F. 2000. The Comparison of Gene Expression from Multiple cDNA Libraries. *Genome Research*, 10(12): 2055–2061.

Strader, C.D., Fong, T.M., Tota, M.R., Underwood, D. and Dixon, R.A.F. 1994. Structure and Function of G Protein-Coupled Receptors. *Annual Review of Biochemistry*, 63(1): 101–132.

Strassmann, J.E., Fortunato, A., Cervo, R., Turillazzi, S., Damon, J.M. and Queller, D.C. 2004. The Cost of Queen Loss in the Social Wasp Polistes dominulus (Hymenoptera: Vespidae). *Journal of the Kansas Entomological Society*, 77(4): 343–355.

Sumner, S., Pereboom, J.J.M. and Jordan, W.C. 2006. Differential gene expression and phenotypic plasticity in behavioural castes of the primitively eusocial wasp, Polistes canadensis. *Proceedings of the Royal Society B: Biological Sciences*, 273(1582): 19–26.

Sundström, H., Webster, M.T. and Ellegren, H. 2004. Reduced Variation on the Chicken Z Chromosome. *Genetics*, 167(1): 377–385.

Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE*, 6(7): e21800.

Suyama, M., Torrents, D. and Bork, P. 2006. PAL2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(suppl 2): W609–W612.

Swanson, W.J. and Vacquier, V.D. 2002. The rapid evolution of reproductive proteins. *Nature Reviews Genetics*, 3(2): 137–144.

Szövényi, P., Ricca, M., Hock, Z., Shaw, J.A., Shimizu, K.K. and Wagner, A. 2013. Selection is no more efficient in haploid than in diploid life stages of an angiosperm and a moss. *Molecular Biology and Evolution*, page mst095.

Tang, C., Toomajian, C., Sherman-Broyles, S., Plagnol, V., Guo, Y.L., Hu, T.T., Clark, R.M., Nasrallah, J.B., Weigel, D. and Nordborg, M. 2007. The Evolution of Selfing in Arabidopsis thaliana. *Science*, 317(5841): 1070–1072.

Team, R.C. 2012. R: A language and environment for statistical computing.

Theurkauf, W.E., Baum, H., Bo, J. and Wensink, P.C. 1986. Tissue-specific and constitutive alpha-tubulin genes of Drosophila melanogaster code for structurally distinct proteins. *Proceedings of the National Academy of Sciences, USA*, 83(22): 8477–8481.

Thornton, K. and Long, M. 2002. Rapid Divergence of Gene Duplicates on the Drosophila melanogaster X Chromosome. *Molecular Biology and Evolution*, 19(6): 918–925.

Thornton, K., Bachtrog, D. and Andolfatto, P. 2006. X chromosomes and autosomes evolve at similar rates in Drosophila: No evidence for faster-X protein evolution. *Genome Research*, 16(4): 498–504.

Torgerson, D.G. and Singh, R.S. 2003. Sex-Linked Mammalian Sperm Proteins Evolve Faster Than Autosomal Ones. *Molecular Biology and Evolution*, 20(10): 1705–1709.

Toth, A.L., Varala, K., Henshaw, M.T., Rodriguez-Zas, S.L., Hudson, M.E. and Robinson, G.E. 2010. Brain transcriptomic analysis in paper wasps identifies genes associated with behaviour across social insect lineages. *Proceedings of the Royal Society B: Biological Sciences*, 277(1691): 2139–2148.

Turner, L.M. and Hoekstra, H.E. 2008. Causes and consequences of the evolution of reproductive proteins. *The International Journal of Developmental Biology*, 52(5-6): 769–780.

Vicoso, B. and Charlesworth, B. 2006. Evolution on the X chromosome: unusual patterns and processes. *Nature Reviews Genetics*, 7(8): 645–653.

Vicoso, B. and Charlesworth, B. 2009. Effective Population Size and the Faster-X Effect: An Extended Model. *Evolution*, 63(9): 2413–2426.

Wallace, B. 1963. Modes of reproduction and their genetic consequences. *Stat Genetics and Plant Breeding*, (982).

Wang, Y., Zhang, W.Z., Song, L.F., Zou, J.J., Su, Z. and Wu, W.H. 2008. Transcriptome Analyses Show Changes in Gene Expression to Accompany Pollen Germination and Tube Growth in Arabidopsis. *Plant Physiology*, 148(3): 1201–1211.

Wang, Z. and Zhang, J. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proceedings of the National Academy of Sciences of the United States of America*, 108(16): E67–E76.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2): 256–276.

White, M.J.D. 1945. *Animal Cytology and Evolution*. Cambridge University Press, Cambridge, first edition.

Wilkinson, L. and Urbanek, S. 2011. Venneuler: Venn and Euler Diagrams. *R package version*, 1-1.

Wilson, E.O. 1978. Division of Labor in Fire Ants Based on Physical Castes (Hymenoptera: Formicidae: Solenopsis). *Journal of the Kansas Entomological Society*, 51(4): 615–636.

Wilson, E.O. 2008. One Giant Leap: How Insects Achieved Altruism and Colonial Life. *BioScience*, 58(1): 17–25.

Winningham, K.M., Fitch, C.D., Schmidt, M. and Hoffman, D.R. 2004. Hymenoptera venom protease allergens. *Journal of Allergy and Clinical Immunology*, 114(4): 928–933.

Wright, A.E., Harrison, P.W., Zimmer, F., Montgomery, S.H., Pointer, M.A. and Mank, J.E. 2015. Variation in promiscuity and sexual selection drives avian rate of Faster-Z evolution. *Molecular Ecology*, 24(6): 1218–1235.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics*, 16(2): 97–159.

Wright, S., Ness, R., Foxe, J. and Barrett, S. 2008. Genomic Consequences of Outcrossing and Selfing in Plants. *International Journal of Plant Sciences*, 169(1): 105–118.

Wurm, Y., Wang, J., Riba-Grognuz, O., Corona, M., Nygaard, S., Hunt, B.G., Ingram, K.K., Falquet, L., Nipitwattanaphon, M., Gotzek, D. et al 2011. The genome of the fire ant Solenopsis invicta. *Proceedings of the National Academy of Sciences, USA*, 108(14): 5679–5684.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24(8): 1586–1591.

Yang, Z. and Bielawski, J.P. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12): 496–503.

Yao, J., Buschman, L.L., Lu, N., Khajuria, C. and Zhu, K.Y. 2014. Changes in Gene Expression in the Larval Gut of Ostrinia nubilalis in Response to Bacillus thuringiensis Cry1ab Protoxin Ingestion. *Toxins*, 6(4): 1274–1294.

Zeyl, C., Vanderford, T. and Carter, M. 2003. An Evolutionary Advantage of Haploidy in Large Yeast Populations. *Science*, 299(5606): 555–558.