Population Genomics of *Polistes* Wasps

Kathleen Dogantzis

A thesis submitted to the faculty of graduate studies in partial fulfillment of the requirements for the degree of Masters of Science

Graduate Program in Biology
York University
Toronto,
Ontario

December 2016

Abstract

The molecular mechanisms influencing the evolution of social behaviour in insects are of great interest and have been the focus of many recent studies. Chapter one of this thesis reviews several major hypotheses regarding the evolution of sociality. Chapter two outlines the methodological steps taken to generate a high quality population genomic data set for primitively eusocial paper wasps in the genus *Polistes*. The third chapter of the thesis uses the dataset generated in chapter two to estimate patterns of natural selection on the *Polistes* genome, and to evaluate the importance of novel and caste biased genes on the fitness of this primitively eusocial species.

Acknowledgments

I would like to thank my supervisor Dr. Amro Zayed and collaborator Dr. Amy Toth for making this project possible. They provided the foundation for this project along with valuable advice about the analyses. Additionally, thanks to Dr. Zayed for all the support and encouraging me to pursue studies beyond my Masters. Special thanks to Brock Harpur who really helped me get this project off the ground and provided me with the tools and knowledge needed to get where I am today. I would like to thank Clement Kent for his statistical input. Thanks to Dr. Laurence Packer for sitting on my advisory committee and giving me feedback on sociobiological theory. Lastly, big shout out to the Zayed lab for making my Masters experience so enjoyable.

Table of Contents

List of Tables

List of Figures

Chapter One: A Review of the Progression of Hypotheses for the Evolution of Eusociality

Introduction

Eusociality is a highly derived behaviour commonly found within the haplodiploid insect order Hymenoptera (ants, bees, and wasps). It is generally defined by three main characteristics: cooperative brood care, overlapping generations, and reproductive division of labour (Wilson 1971). Eusociality has evolved independently more than ten times in insects (Yan, et al. 2014), but the expression of sociality varies greatly among different social species (Michener 1969).

Advanced eusocial societies, as found in the honey bees (*Apis*), have distinct female division of labour. The reproductive queen caste mates and lays eggs, while the worker caste remains effectively sterile taking on the position of colony care, provisioning, and offspring rearing. Castes in advanced eusocial species have distinct morphological differences making them easily identifiable (Michener 1969; Michener 1974). On the other hand, primitively eusocial species, like paper wasps (*Polistes*), possess a less distinct female caste system. Nests are initiated by foundresses, who take on the initial role of foraging and reproduction. Once the first set of brood has developed into adult workers, colonies become cooperative and foundresses assume the role of queen. In primitively eusocial wasp societies, reproductive division of labour is maintained through physical dominance, and the worker caste, which can display no morphological distinction, remains associated with the nest contributing to brood care, provisioning, and foraging (Jandt, et al. 2014).

The reproductive altruism displayed by the worker caste was described by Darwin (1859) as being "the one special difficulty,… fatal" to his theory of natural selection. Darwin recognized that selection cannot directly act upon workers to optimize individual fitness or transmit altruistic

traits since individuals forgo reproduction. Thus, mutations that influence altruistic behaviours would go extinct as they cannot be passed on to future generations. Since then, alternative hypotheses to the traditional route of selection have been proposed in order to explain the evolution of eusociality; these include multilevel selection and kin selection.

Theories of Eusocial Evolution

Multilevel Selection

Multilevel selection theory, whose origins can be traced back to Darwin (1871), postulates that natural selection may operate at more than one level of the biological hierarchy (i.e. genes, cells, individuals, groups or populations, etc.) (Okasha 2005). A variation of multilevel selection theory called group selection has been invoked to explain the evolution of sociality. This hypothesis suggests that some individuals will have higher fitness when placed in a group rather than on their own. For example, within a group, a selfish individual may outcompete an altruist, but between groups, altruistic groups may outcompete selfish groups (Wilson and Wilson 2007). This hypothesis was rejected in the 1960's on the basis that altruistic groups were susceptible to non-altruist invaders (Smith 1964), and that group benefits were possible if they arose through individuals, but not if they arose exclusively for the benefit of the group (Williams 2008). However, there has been a recent resurgence of interest in the multilevel hypothesis in evolutionary literature, especially within the context of eusociality as a major transition in evolution (Keller 1999; Okasha 2005; Wilson and Wilson 2007; Nowak, et al. 2010; Marshall 2011). But, there is still debate over its acceptance (van Veelen, et al. 2012).

Kin Selection

Kin selection and inclusive fitness theory has been the longstanding explanation for the evolution of eusociality (Marshall 2011). The theory was popularized by W. D Hamilton and suggests that altruistic acts towards individuals with greater genetic similarity increases the chances of the actor's genes – including those responsible for altruism – being passed on to subsequent generations (Hamilton 1964). Hamilton conceptualizes this theory mathematically, explaining that altruistic behaviour evolved when the product of genetic relatedness (R) and fitness benefits of the recipient (B), outweighed the fitness cost to the altruist (C) (Hamilton 1964; Bijma and Wade 2008). Although inclusive fitness theory can be applied to any sexually reproducing organism, haplodiploidy (i.e. males are haploid while females are diploid) – a defining characteristic of Hymenoptera – is believed to have facilitated the evolution of eusociality because it results in high sister to sister relatedness (3/4). However, Trivers and Hare (1976) noted that the lower degree of sister-brother relatedness (1/4) cancels out high sister-sister relatedness, producing an average degree of relatedness as when females produce their own offspring as solitary individuals (1/2). Though, it has been argued that female biased sex ratios and worker produced males are conducive to the evolution of eusociality via kin selection (Trivers and Hare 1976; Crozier and Pamilo 1996). Despite that kin selection theory has been critiqued on its mathematical basis and the discovery of eusocial species that use diploidiploid sex determination (Nowak, et al. 2010; Marshall 2011), it still remains the most supported hypothesis for the evolution of sociality to date (Gardner, et al. 2011).

The Evolution of Eusociality – Mechanistic Hypotheses

The above-mentioned hypotheses provide evolutionary strategies for how mutations causing altruism could spread and fix; thereby leading to the evolution of eusociality from ancestrally solitary populations. Recently there has been interest in exploring social life in molecular terms to identify the underlying factors that influence worker and queen phenotypes in eusocial societies. As such, multiple hypotheses have been proposed that strive to address which genes and pathways regulate the behavioural and physiological differences among individuals found in eusocial colonies (Robinson, et al. 2005). Of course, these hypotheses are not mutually exclusive from each other and can be used in combination to better explain eusocial evolution.

Hypotheses Involving Regulation of Gene Expression

Advanced eusocial insects have notable differentiation between the behavioural, physiological, and morphological traits expressed by each caste (Michener 1969). On the other hand, primitively eusocial species often lack discrete caste differences and individuals exhibit a degree of flexibility in behaviour and physiology; future queens exhibit worker traits early in the colony cycle and workers remain totipotent (Hunt, et al. 2010; Berens, et al. 2015b). What is striking of most eusocial species is that queen and worker traits are produced from the same genome, making it one of the most noteworthy examples of phenotypic plasticity (Berens, et al. 2015b), however there are exceptions (Julian, et al. 2002; Linksvayer, et al. 2006).

Phenotypic plasticity is defined as a single genotype capable of producing multiple phenotypes. Plasticity in traits enables individuals to maximize their fitness in response to fluctuating environmental conditions by maintaining flexibility in phenotype production (Sumner, et al. 2006; Berens, et al. 2015b). The degree of differentiation in phenotypes between castes occurs as a result of differential expression of genes. Thus, knowledge of transcriptional

regulation combined with observations of caste related behaviours can be used to formulate hypotheses regarding the evolution of eusociality. Three hypotheses include: the ovarian ground plan, the maternal heterochrony hypothesis, and the genetic toolkit hypothesis.

The ovarian ground plan hypothesis, proposed by West-Eberhard (West-Eberhard and Turillazzi 1996), was inspired by field observations of wasps. The hypothesis proposes that gene networks in ancestral solitary species responsible for coordinating reproduction, brood care, and foraging, were co-opted and differentiated by natural selection to control reproductive traits of queens and sib-care and foraging of workers in eusocial societies (West-Eberhard and Turillazzi 1996). In other words, the ovarian ground plan suggests a decoupling of reproduction and foraging behaviours of solitary females - achieved through differential gene expression - to produce queen-like and worker-like individuals. Fundamentally, the hypothesis revolves around ovarian activation and suppression between castes controlled via a switch-like mechanism. Previous research has demonstrated that pheromones and hormones have a direct effect on worker suppression of ovary development through a reduction in size and oocyte development (Bloch, et al. 2002; Hoover, et al. 2003; Ronai, et al. 2015). Additionally, nutritional differences, including the quality and quantity of food, have downstream effects on ovarian development ultimately influencing queen and worker phenotypes (Asencot and Lensky 1988; Hunt and Karsai 2002; Buttstedt, et al. 2014).

The maternal heterochrony hypothesis is conceptually related to the ovarian ground plan hypothesis, however it emphasises the relationship between maternal and sibling care and its influence on caste behaviours. The hypothesis suggests that reproductive division of labour evolved through changes in the timing of expression of genes related to maternal care; predicting that maternal and sibling care should be regulated in similar patterns (Linksvayer and Wade

2005). Gene expression studies of primitively eusocial species have provided strong support for this hypothesis. Brain gene expression of *Polistes metricus* has demonstrated that workers and foundresses, who display maternal-like behaviours, possess similar patterns of expression (Toth, et al. 2007). In contrast, gynes, which emerge later in the season and will become foundresses the following year, display comparable expression patterns to queens; neither of which show maternal care (Toth, et al. 2007). Similar results have also been demonstrated in eusocial bumble bees (Woodard, et al. 2014) and incipiently social carpenter bees (Rehan, et al. 2014).

Lastly, the genetic toolkit hypothesis, which was initially developed to explain the conservation of genes involved in development, can be used to understand the mechanistic basis of sociality in insects (Toth and Robinson 2009). The hypothesis predicts that genes and genetic networks that regulate behaviour in solitary species will be co-opted by natural selection to regulate the same tasks in social insects. Support for this hypothesis was first observed through the correlation between brain expression patterns of the foraging gene in *Drosophila melanogaster* (*for*) and *Apis mellifera* (*Amfor*) (Ben-Shahar, et al. 2002). *D. melanaogaster* flies that have a high propensity to forage display elevated expression levels of the *for* gene and the product it encodes for, PKG (cGMP-dependent protein kinase). In honey bees, adult foragers display similar expression patterns of elevated *amfor* and high PKG levels when compared with adult nurses, who do not forage. This supports the idea that changes in the regulation of a conserved gene that influenced foraging in a solitary insect can be re-purposed by natural selection to regulate division of labour between different worker castes in a social insect. Previous studies comparing various advanced and primitively eusocial species have found some overlap in gene expression patterns related to aggression (Toth, et al. 2014), diapause (Amsalem, et al. 2015), and caste differentiation (Morandin, et al. 2016). However, transcriptome wide

analysis of caste determination in three hymenopteran social linages discovered little overlap in relation to specific genes, but showed similarity in the types of metabolic pathways and molecular functions; including arginine/protein metabolism and glycolysis/gluconeogenesis. Their research suggests more of a "loose toolkit", where similar changes in common metabolic pathways, but not necessary the same genes, are involved in generating caste traits in different social insects (Berens, et al. 2015a; Rehan and Toth 2015).

Hypothesis Involving Novel Protein Coding Evolution

Research on gene regulation and expression has been critical for formulating and supporting mechanistic hypotheses related to the evolution of eusocial behaviour. Most of the hypotheses discussed above involve changes in expression of conserved genes found in both solitary and social insects. However, several recent studies have highlighted the important role of novel genes in generating the novel traits found in social insects. For example, gene duplications, which lead to the expansion of gene groups, can also create genetic redundancy within the genome allowing new gene copies to accumulate molecular changes (Zhang 2003). This process can result in the neo-functionalization of loci where gene duplicates become functionally divergent (Zhang 2003; Assis and Bachtrog 2013). These molecular changes can drastically alter genome composition leading to the evolution of novel phenotypes, such as sociality.

The novel genes hypothesis proposes that novel genes are important for generating novel caste-specific traits found in eusocial insects (Johnson and Tsutsui 2011). Novel genes are taxonomically restricted genes (TRG) that are unique to a specific phylogenetic group of animals (Khalturin, et al. 2009). For example, the advanced eusocial honey bee has approximately 700 taxonomically restricted genes dispersed among the insect, hymenoptera, and species specific

lineages (Johnson and Tsutsui 2011). These TRGs are also disproportionately involved in caste differential expression, particularly those that are worker-biased; suggesting that worker traits are the product of novel genes (Johnson and Tsutsui 2011; Jasper, et al. 2014). Furthermore, TRGs have also been linked to novel physiological traits of honey bees (e.g. the hypopharyngeal glands that produce royal jelly and brood food). Jasper *et al.* (2014) established a connection between novel tissues in honey bees and the proportion of TRGs expressed in these tissues. Additionally, these TRGs expressed in novel tissues have a higher probability of being under positive selection (Jasper, et al. 2014). These genes typically have low network connectedness, and as such, their sequence is often not constrained and free to rapidly change and potentially produce novel phenotypes (Jasper, et al. 2014).

Hypothesis Involving both Novel Genes and Regulation of Gene Expression

A recently study of solitary and social insects proposed that changes in gene regulation may be important at early stages of eusociality, while gene composition changes may be more important for lineage specific adaptation (Simola, et al. 2013). Their study demonstrated that changes in both trans and cis regulatory factors were similar across some eusocial species suggesting that regulatory factors may have convergently evolved at early stages of social evolution (Simola, et al. 2013). In comparison, Hymenoptera specific changes related to gene family expansions and contractions and the discovery of novel genes suggest that gene compositional changes may be more important for lineage-specific social adaptations (Simola, et al. 2013). By expanding on this proposal and incorporating the above hypotheses, Rehan and Toth (2015) have proposed a systematic framework for studying the evolution of eusociality. Rehan and Toth (2015) invoke the 'social ladder' analogy to represent a range of social

behaviours that span from solitary to highly eusocial, where the rungs represent different stable forms of social organization observed in insects (e.g. primitive eusociality, advanced eusociality). Rehan and Toth (2015) made several predictions regarding which types of genes and molecular mechanism are involved during the early, middle, and late stages of social evolution. They predict that at early stages of evolution, changes to the timing of gene expression are initially influenced by the environment resulting in flexible phenotypes. When eusociality evolves further, expression patterns become more distinct leading to caste specific expression and specialization of genes. In this scenario, genes are increasingly free from pleiotropic constraints promoting sequence changes. Thus, transitions during later stages of eusociality are accompanied by gene sequence changes producing novel genes with novel functions (Rehan and Toth 2015).

Population Genomics – Insights into Social Evolution

Both evolutionary and mechanistic hypotheses for the rise of eusociality invoke positive selection. For example, kin-selection theory assumes that 'genes for altruism' experience positive selection if altruism increases the inclusive fitness of workers, while mechanistic hypotheses assume that some gene sets were co-opted by natural selection to cause caste specific phenotypes. While the above mentioned hypotheses have guided the framework for social evolution studies, we have lacked a means to objectively test them, mainly because it was – until recently – impossible to quantify patterns of positive selection across the genome of eusocial insects and their solitary ancestors.

Population genomics is a rapidly evolving field focused on implementing a whole genome approach to understand the biology and evolution of species. Within a sociogenomic

context, population genomics can be used to study the evolution of novel social behaviours by examining the effects of key evolutionary mechanisms, such as natural selection, across the genome (Robinson, et al. 2005). Positive selection is a non-stochastic evolutionary mechanism thought to be the primary means for adaptation to specific environments and niches. It shapes genome variation by increasing the frequency of beneficial mutations, which in consequence, enhances the fitness and adaptation of the individuals carrying them (Vitti, et al. 2013). Over time these changes produce divergent loci and alleles between populations and species causing significant phenotypic difference. As such, population genomics can provide insight into the evolutionary history of a species by detecting loci under selection to elucidate what genes and pathways potentially regulate the behavioural and physiological differences found within populations.

Various approaches have been used to identify positive selection in species on macro and micro evolutionary levels. Macro-evolutionary trends can be used to identify divergent patterns between species highlighting selective events that occurred in the past, while micro-evolutionary trends can identify selective patterns within a species highlighting more contemporary local adaptation (Nielsen 2005; Vitti, et al. 2013). Methods to detect selection at the macro-evolutionary level rely on identifying synonymous and non-synonymous substitutions between species. Synonymous substitutions are presumed to be selectively neutral (i.e. have no influence on fitness) while non-synonymous substitutions produce changes that positively or negatively affect fitness. Genomic regions between species where non-synonymous substitutions differ significantly and result in greater synonymous to non-synonymous ratios are suggested to be under positive selection (McDonald and Kreitman 1991; Vitti, et al. 2013).

Methods to detect selection at the micro-evolutionary level depend mainly on analysing changes in allele frequency within species. As beneficial mutations increase their prevalence in a population, it produces regions of the genome with reduced genetic diversity. Since populations are subject to variable environmental pressures, it will result in divergent genomic regions between populations (Vitti, et al. 2013). Depending on the question being asked, these methods provide an effective means for assessing the patterns of selection in social insect species.

So far there have been a limited number of studies assessing patterns of positive selection in social insect genomes. Population genomic analysis of advanced eusocial *Apis* has revealed that taxonomically restricted genes show higher rates of positive selection relative to hymenoptera and insect restricted genes (Harpur, et al. 2014). This study also showed that proteins demonstrating caste biased expression in workers have higher signatures of positive selection relative to queen biased proteins (Harpur, et al. 2014). This provides compelling support for the novel gene hypothesis suggesting novel genes have greater adaptive influence at higher social stages relative to conserved genes. Additionally, these results support the idea that worker traits are of major importance to the fitness of honey bees.

Population genomics provides a means to fully understand the genomic underpinnings of social evolution. While we have learned much from the population genomic study of the advanced eusocial honey bee efforts need to be made to study the early stages of social evolution. This will help put into context the patterns and mechanisms conserved across all levels of sociality, while also highlighting narrow phylogenetic genomic changes. As such, as genomic research expands, these methods can be applied to a wide variety of social and non-social species to provide a complete understanding of the evolution of social behaviours.

Chapter Two: Producing High Quality Data Sets for Population Genomic Analyses

Summary

Genome analyses have become increasingly prevalent due to rapid advances in sequencing technology. The extensive amount of data produced by these procedures allows for a broader range of applications, but calls for stricter quality control. In this study we sequenced the genomes of ten *Polistes dominula* and two *P. gallicus* samples using Illumina technology. To establish a high quality dataset for population genomic analysis, single nucleotide polymorphisms were called and filtered using a series of rigorous techniques. Quality control of the data set was estimated using independently derived SNPs from a transcriptomic dataset to confirm common variants and by calculating the transition to transversion ratio that can be used as an approximate measure of quality.

Introduction

Population genomics is a rapidly evolving field focused on implementing a whole genome approach to understand the biology and evolution of species. Facilitated by advancements in sequencing technologies and software availability (Catchen, et al. 2013), this approach to studying organisms has drastically increased the amount of information attainable of target species. The accessibility to thousands of loci has removed the requirement of needing prior knowledge of the neutral or adaptive significance of candidate loci. As such, population genomics is enabling research that was not practical or attainable in the past.

One of the primary goals of population genomics is to identify and characterize signatures of selection acting across a genome using data from single nucleotide polymorphisms

(SNPs). SNPs occur extensively throughout the genome and are estimated to occur every 200-500 base pairs in non-coding DNA and every 500-1000 base pairs in coding DNA (Brumfield, et al. 2003). Genome wide coverage of SNP markers attained through genomic sequencing improves the resolution of population studies and increases the likelihood of discovering genomic regions influenced by natural selection (Freeland, et al. 2011). However, identifying genetic variants at the genome scale depends on the automation of genotyping, which can produce false or poor quality SNPs due to poor alignments and DNA quality (Laurie, et al. 2010). As such, the construction of genome SNP datasets requires post genotype filtering in order to implement quality control measures.

Quality control is a critical step in any analysis to ensure data integrity, avoid spurious results, and to reduced false positives (Laurie, et al. 2010; Gondro, et al. 2014). In instances of investigating the prevalence and patterns of selection, false positives could cause misinterpretations of deviations from neutrality. To ensure the integrity and quality of the variant calls, stringent filtering thresholds are applied to SNP datasets. SNP filtering criteria are often based on variant annotations, which include parameters such as strand biases, mapping quality and base quality, and read depth (Auwera, et al. 2013). Additional filters can be employed to remove variants that fall within extremes of quality control variables including deviations from Hardy-Weinberg equilibrium, minor allele frequencies, and percentage of missing data (Burton, et al. 2007; Manolio, et al. 2007; Sladek, et al. 2007; Unoki, et al. 2008; Pongpanich, et al. 2010). After filtering SNPs based on quality control measures and predetermined thresholds, SNPs can be categorized based on functional annotations that can then be used to reveal the evolutionary and biological processes that have shaped social insect genomes.

Methods

Sample Preparation and Whole Genome Sequencing

Samples of *Polistes dominula* and its sister species *Polistes gallicus* were collected within their native ranges in Tuscany, Italy in August 2014. Samples were subsequently sequenced for DNA using extractions from the whole thorax. Genome sequencing was performed on ten *P. dominula* and two *P. gallicus* samples using the Illumina HiSeq 2500 system using two lanes per sample to yield approximately 30X coverage per sample based on a 200Mb genome.

Whole genome alignment

Illumina paired-end reads were trimmed of adapters and poor quality bases using the default settings of Trimmomatic (Bolger, et al. 2014). Reads (Raw FASTQ) were then aligned to the unmasked *P. dominula* reference sequence (PdomGDB r1.2) (Standage, et al. 2016) using the default parameters of the Burrows-Wheeler aligner MEM algorithm (Li 2013). Alignments were sorted and output as BAM files using Samtools (Li 2011). The sorted BAM files were marked for duplicates with validation stringency set to silent, and read groups were replaced using Picard (http://broadinstitute.github.io/picard/). Subsequent files were indexed to BAM files using Samtools. The resulting BAM files were realigned around indels using default parameters in GATK 3.5-0-g36282e (McKenna, et al. 2010) (Appendix A – Figure 1). Coverage for each of the BAM files was calculated by averaging the coverage per base pair of each BAM file using Depth of Coverage in GATK 3.5-0-g36282e (Appendix A – Table 1).

Whole Genome Variant Calling and Filtration

Variants were detected using GATK's HaplotypeCaller using all species-specific BAM files in unison. Variants identified with GATK were initially filtered for a Minimum Base Quality (MBQ) (MBQ > 20), followed by filtering based on the variant call annotations produced in the subsequent VCF file. SNPS were discarded (Fail SNPS) if they had a poor Mapping Quality (MQ < 40) or inconsistent base qualities between the reference and alternative alleles (MQRankSum < -12.5), and if there was strand (FS > 60) or position bias (ReadPosRankSum < -8.0) for alternative allele calls. SNPs were also discarded if there was poor variant confidence (QD < 5) and if they had an unusually high or low depth of coverage (DP < 100 & > 350 for *P. dominula*, and DP < 20 & > 83 for *P. gallicus*). SNPs positioned within and around five base pairs of unfiltered indels were also discarded. From the variants remaining, SNPs were filtered based on minimum frequency threshold for missing data. Since there was a small sample size, any variant that had missing data was discarded from the analysis. Additionally, regions with highly repetitive sequences or recently duplicated genes were filtered from the analysis. This is because loci with high sequence homology to more than one segment of the genome can cause complications during the alignment process. Highly homologous regions were identified by dividing the reference genome of *P. dominula* into 150-bp fragments and using Blastn to match those segments back onto the reference genome. SNPs that were present within segments which resulted in a Blastn match to two or more regions with a corresponding E-value of $10e^{-40}$ or higher were removed from the analysis (Appendix A – Figure 2). Lastly the discarded SNPs (Fail SNPs) from *P. gallicus* were removed from *P. dominula* and vice versa to avoid skewing fixed and polymorphic variant ratios by retaining potentially poor

quality SNPs. All remaining SNPs (Pass SNPs) were used in subsequent analyses (Appendix A – Table 2-3).

Determination of Filters

The thresholds for the MBQ, MQ, FS, MQRankSum, and ReadPosRankSum were adopted from GATK best practices (DePristo, et al. 2011; Auwera, et al. 2013), which have also been employed in recent genome analysis (Gudbjartsson, et al. 2015). Through visual inspection of the raw VCF file, the threshold for QD was adjusted from GATK best practice recommendations in light of the typical QD scores found in the dataset.

Upper limit thresholds for depth of coverage were determined by calculating the 1.5*IQR (interquartile range) for the total depth of coverage for each variant. SNPs lying outside of the upper limit were considered outliers and discarded from the analysis. The lower threshold was determined by allowing an average depth of coverage of ten reads per base pair per individual (Appendix A – Figure 3-4).

The E-value for Blastn results was determined by plotting the percentage of SNPs that would be removed from the Pass SNPs and Fail SNPs to determine when the threshold plateaus. The highest E-value that can be called is $10e^{-72}$, at which point no segments would have a corresponding match other than to its original segment, thus no SNPs would be removed from the analysis. As the E-value becomes less conservative more SNPs are removed from the analysis. The percentage of Fail SNPs overlapping with repetitive segments was used to determine the adequacy of the SNP filters. In theory, the filters should be removing poor quality SNPs that would be arising in mismatching repetitive regions (Appendix A – Figure 5).

Deviation from Hardy-Weinberg equilibrium (HWE) is often employed to detect gross genotyping errors in large SNP datasets. Studies which have used this as a filtering method have

noted a wide range of P-value thresholds ranging from 5.7×10^{-7} to 0.001 (Burton, et al. 2007; Sladek, et al. 2007; Evans, et al. 2014; Gudbjartsson, et al. 2015). In this study, HWE was not used as a filtering parameter as none of the variants found within genes showed major deviations from HWE to be discarded from the analysis. To determine this, a measure of confidence thresholds was determined using the Bonferroni adjustment (α/n) where α is the significance threshold and n is the number of tests (Noble 2010). According to the Bonferroni correction criteria, deviations from Hardy-Weinberg based on a statistical significance of α=0.05 would require a P-value less than $0.05/86{,}000 = 6 \times 10^{-7}$. The smallest P-values determined from the Chi^2 tests was $> 1 \times 10^{-3}$, thus none of the SNPs would have shown significant departures from HWE after or before Bonferroni corrections. We also employed a Benjamini-Hochberg FDR (false discovery rate) adjustment for multiple testing and no SNP showed significant deviations from HWE at $p < 0.05$.

The minimum allele frequency threshold was also not used because the proportion of minor alleles is within the normal range found in other studies (Crawford and Lazzaro 2012; Manske 2012) (Appendix A – Figure 6).

Variant Annotation

Variants that passed all of the filtering criteria where annotated for predicted effects on genes using SnpEff (Cingolani, et al. 2012). SnpEff utilizes the reference sequence and genome annotation file (GFF3) to make predictions regarding the effects of the variant on the gene. Variant annotations are classified as: intergenic, intronic, synonymous/stop, nonsynonymous/start, 3' UTR, 5' UTR, upstream and downstream SnpEff also indicates

potential annotation errors with a warning sign. Annotated variants were filtered for synonymous and non-synonymous predictions and filtered of SNPs with warnings (Appendix A – Figure 8).

Technical validation

Validation and Confidence of Variant Calls

In order to verify the accuracy of the SNP calls and filtering thresholds, the genomic dataset was compared to SNP calls generated from transcriptomic data. Paired-end transcriptomic reads of ten *P. dominula* samples generated from a two lane Illumina HiSeq 2000 run (SRX1122234, SRX1124050, SRX1124051, SRX1124052, SRX1124053, SRX1124054, SRX1124056, SRX1124059, SRX1124060, SRX1124061) were downloaded from GenBank and aligned to the unmasked *P. dominula* reference sequence (PdomGDB r1.2) (Standage, et al. 2016) using the default parameters of STAR's 2-pass method (Dobin, et al. 2013). The STAR 2-pass method performs two alignments in which the splice junctions detected in the first alignment are used to guide the second alignment. The resulting BAM files were then marked for duplicates with validation stringency set to silent and read groups were replaced using Picard. Subsequent files were imported into Samtools and indexed to BAM files. Following GATK's best practices for RNAseq data (DePristo, et al. 2011), reads within the BAM files were split into exons and trimmed of overhang into intronic regions and reassigned mapping quality scores as per the Split'N'Trim step. Resulting BAM files were realigned around indels using default parameters in GATK 3.5-0-g36282e4 (McKenna, et al. 2010) (Appendix A – Figure 1).

Variant calling was performed using GATKs UnifiedGenotyper while filtering for base pairs with a MBQ > 20. Variant calls were filtered using the variant call annotation produced in the subsequent VCF file and we discard SNPs based on strand bias (FS > 30), poor variant confidence (QD < 2) (DePristo, et al. 2011; Auwera, et al. 2013), and low depth of coverage (DP

< 40). Variants that passed all filtering criteria where annotated for predicted effects on genes using SnpEff (Cingolani, et al. 2012) and filtered for synonymous and non-synonymous predictions discarding SNPs with warnings. The remaining call set was then compared to the synonymous and non-synonymous genomic dataset in order to determine the degree of overlap within annotated coding exonic regions.

In total there were 86,108 synonymous and non-synonymous SNPs for the *P. dominula* genomic dataset and, 58,137 synonymous and non-synonymous SNPs for the *P. dominula* transcriptomic datasets. Datasets overlapped by 38,058 common SNP calls comprising 44.2% of genomic calls and 65.5% of transcriptomic calls (Appendix A – Figure 7). Discrepancies in the total number of variants and the percentage of overlap between each dataset can be explained by the differences in sample type. The genomic dataset was derived from populations of *P. dominula* obtained within the species' native range in Europe (Tuscany, Italy), while variants from the transcriptomic dataset were derived from an invasive population of *P. dominula* in Pennsylvania. Studies have shown that native Tuscany populations of *P. dominula* show higher levels of genetic diversity compared to introduced populations based on measures of expected and observed heterozygosity and allelic richness (Liebert, et al. 2006). Native populations also demonstrate a significantly higher number of private alleles compared with introduced North American populations (Liebert, et al. 2006). The differences in genetic variability and private alleles may therefore account for the greater percentage and diversity of SNPs in the genomic dataset relative to the transcriptomic dataset. Additionally, the transcriptomic dataset lacks coverage across every gene as not all genes are expressed at each life history stage, consequently reducing the observable number of variants.

Estimation of the Transition/Transversion Ratio

Transition/transversion (Ti/Tv) ratios were calculated for each annotation group determined from SnpEff. Transitions are defined as base pair changes between purine or pyrimidine pairs, while transversions are base pair changes between a purine and a pyrimidine. Ratios are determined by dividing the number of transitions by the number of transversions. Ti/Tv ratios can be used as an approximate measure of quality; variant calls with higher Ti/Tv ratios are associated with fewer false positives (Liu, et al. 2012; Carson, et al. 2014; Wang, et al. 2014). Generally, the ratio of transitions to transversions will be 1:2 because there are twice as many possible transversions. Typically, we can expect a genome Ti/Tv ratio of 2.0-2.1 for humans (DePristo, et al. 2011; Gudbjartsson, et al. 2015).

Whole genome Ti/Tv ratios improved following filtering, increasing from 1.69 to 1.91 for *P. dominula* (Appendix A – Table 4, Appendix A – Figure 8). We can expect whole genome ratios to be lower for *P. dominula*, than humans, due to low GC (30.8%) and high AT (69.2%) content, lower gene to genome coverage, and reduced CpG regions (Standage, et al. 2016). High GC content regions, especially those with higher CpG regions that experience methylation, demonstrate higher transition frequencies (Keller, et al. 2007). In an AT rich genome with lower genic regions and reduced CpG, we can expect fewer transitions and more transversions lowering the overall transition and transversion ratios.

Intergenic regions of *P. dominula,* which accounts for approximately 75% of the overall dataset, reveal a Ti/Tv ratio of 1.86 (Appendix A – Table 4, Appendix A – Figure 8). Synonymous SNPs had the highest Ti/Tv ratio of 5.25 which is within the range found in humans (Gudbjartsson, et al. 2015), while the synonymous ratio for *P. gallicus* (3.62) was lower (Appendix A – Table 4, Appendix A – Figure 8). High quality exome datasets typically yield

Ti/Tv ratios between 2.8-3.5 for humans (Liu, et al. 2012; Carson, et al. 2014). Exome regions for *P. dominula* and *P. gallicus* revealed Ti/Tv ratios of 3.8 and 2.42 respectively, falling just outside the predicted human data range (Appendix A – Table 4, Appendix A – Figure 8). (Gudbjartsson, et al. 2015). Lastly, Ti/Tv ratios for failed or discarded SNP calls revealed ratios of 1.39 and 1.29 suggesting poor quality SNPs were removed from the dataset (Appendix A – Table 4, Appendix A – Figure 8).

Chapter Three: Using Population Genomics to Explore the Evolution of Eusociality in Primitively Eusocial Papers Wasps

Summary

Eusociality is a major evolutionary transition that independently evolved several times in insects and is defined by three main characteristics – overlapping generations, cooperative brood care, and reproductive division of labour. The mechanisms underlying the evolution of eusociality are not well understood. Here, we carried out a population genomic approach study of the primitively eusocial *Polistes* to identify the genes and traits with signs of adaptive evolution. We found no significant difference in the strength of positive selection on novel and conserved genes, which is consistent with the hypothesis that novel genes are mostly important during the latter stages of eusocial evolution. Genes associated with queen traits showed marginal enrichment for signs of positive selection relative to workers, emphasizing a greater adaptive role of queen traits in *Polistes* evolution. Finally, we found that genes under positive selection in the eusocial honey bees, bumble bees, and paper wasps included functions related to immunity and communication, indicating that these genes play important roles in social evolution during both early and late stages of eusociality.

Introduction

Understanding the origin and elaboration of eusociality is a major goal of evolutionary biology. Eusociality has independently evolved several times in the Hymenoptera (Brady, et al. 2006; Hines, et al. 2007; Schwarz, et al. 2007) and is characterized by overlapping generations, cooperative brood care, and reproductive division of labour (Batra 1966; Michener 1969; Wilson 1971). The expression of eusociality is variable and ranges between solitary species, that show

no social tendencies, to highly advanced social species (Michener 1969; West-Eberhard 1969). It is often assumed that eusociality evolved along a 'ladder' whereby social lineages become more complex (i.e. greater colony sizes), and caste divergences become more pronounced over evolutionary timescales (Szathmáry and Smith 1995; Rehan and Toth 2015).

Recently, interest has been focused on exploring social life in molecular terms to identify the mechanisms influencing the evolution and adaptation of social insect societies. Particularly, are there different patterns of molecular evolution associated with the rise and elaboration of eusociality? As such, multiple mechanisms have been proposed to address this question citing instances of gene regulation changes (West-Eberhard 1989; West-Eberhard and Turillazzi 1996; Linksvayer and Wade 2005; Sumner, et al. 2006; Berens, et al. 2015a) and protein coding modification (Johnson and Tsutsui 2011; Woodard, et al. 2011; Harpur, et al. 2014; Jasper, et al. 2014). This has raised further questions regarding the degree to which novel and conserved genes play a role in social evolution, in addition to the effects of queen and worker phenotypes on colony fitness.

The above questions were until recently difficult to address as it was impossible to estimate patterns of positive selection across the genome of non-model organisms. With diminishing costs and advancements in bioinformatics tools, whole genome sequencing and analyses are attainable. As such, population genomics now provides a viable opportunity to study the evolutionary forces shaping the genomes of social insects. The first population genomic study carried out on a social insect involved the advanced eusocial honey bee *Apis mellifera*, and provided much needed insight into the evolution of this socially-complex genus (Harpur, et al. 2014). In particular, this study highlighted that, relative to queen traits, worker traits were highly enriched for signs of positive selection, suggesting that their contributions to colony fitness is

immense (Harpur, et al. 2014). Additionally, this study showed that novel genes were highly enriched for signs of positive selection compared to conserved genes elucidating on the impact of taxonomically restricted genes in honey bee evolution (Harpur, et al. 2014). Novel genes tend to be greatly expressed in workers (Johnson and Tsutsui 2011) and this study adds to the notion that worker traits controlled by novel genes are critical for colony fitness and adaptation of *Apis*.

Although the honey bee population genomic study resulted in several interesting conclusions, it is difficult to extrapolate this knowledge to other social species that have lower complexity (e.g. smaller colony size, less distinct queen-worker divergence). Wasps within the family Vespidae are ideal for studying the evolution of eusociality (Jandt, et al. 2014). The Vespidae contain a range of complexity spanning between solitary and advanced eusocial species making comparative studies possible (Hunt 2007; Hunt 2012; Jandt, et al. 2014). Thus far, most studies of the Vespidae have concerned the genus *Polistes*, which displays an intermediate level of social behaviour referred to as primitive eusociality.

In *Polistes*, the colony life cycle is initiated by one or multiple foundresses who begins colony construction, egg laying, and provisioning for developing brood (Jandt, et al. 2014) . Once the first set of brood has developed into workers, the foundress takes on the queen role, and the workers begin to provide brood care and provisioning needs (Jandt, et al. 2014). In late season, the last set of female brood develops into gynes who do not engage in colony tasks and instead mate and overwinter in aggregations to emerge as foundresses the following year (Dapporto and Palagi 2006). The division of labour in primitively eusocial paper wasps is maintained by a dominance hierarchy and workers remain totipotent.

The European paper wasp (*Polistes dominula*) is a primitively eusocial species whose unique life history coupled with emerging genomic and transcriptomic datasets (Standage, et al.

2016) make it a strong candidate for population genomic studies. Recently, methlyomic and transcriptomic research (Patalano, et al. 2015; Standage, et al. 2016) has suggested that *Polistes* show unexpected patterns of gene regulation when compared with social Apoidea species (Grozinger, et al. 2007; Lyko, et al. 2010; Sadd, et al. 2015; Rehan, et al. 2016) suggesting that genomic patterns could also diverge from what is expected. Thus, a population genomic assessment could reveal interesting patterns pertaining to the evolution of eusociality.

Here, we used a population genomic approach to test several mechanistic hypotheses related to the evolution of social insects. Our goal was to assess the patterns and prevalence of selection across the genome to identify the types of loci influencing evolution and adaptation in a primitively eusocial *Polistes* wasp. We focused our efforts on addressing four research objectives. The first goal of this study was to identify genome regions with signs of positive selection. Genes with signs of positive selection were subjected to gene ontology (GO) enrichment analysis to better understand the biological, cellular, and molecular functions of adaptively evolving genes.

The second goal of the study was to evaluate patterns of selection on taxonomically restricted genes (TRGs) relative to genes found throughout all insects. In advanced eusocial honey bees TRGs at the *Apis* and Apoidea levels were highly enriched for positive selection relative to less taxonomically exclusive genes (Harpur, et al. 2014). As such, we can hypothesize that TRGs in *Polistes* will also have higher signatures of positive selection if novel genes are universally important for the evolution of social behaviour. Alternatively, if TRGs only play a role in advanced eusocial species, they should not be enriched for positive selection as per the social ladder hypothesis (Rehan and Toth 2015).

The third goal of this study was to evaluate if queen and worker phenotypes contribute equally to the evolution of social insects. We used transcriptomic studies (Standage, et al. 2016) to classify genes as either queen-biased or worker-biased and examined if these two genes sets differed with respect to patterns of positive selection. In the advanced eusocial honey bees, worker biased proteins were found to be enriched for signs of positive selection, suggesting that worker traits are a primary means of adaptive evolution in advanced eusocial species (Harpur, et al. 2014). As such, we can hypothesize that worker biased genes will have higher signatures of positive selection if the worker phenotype is disproportionately significant to the evolution of social insects.

The final goal of this study was to compare loci under positive selection in *Polistes* to *Apis* and *Bombus* to identify shared gene sets with signs of adaptive evolution in these two independently derived eusocial lineages. Positive selection on common genes may indicate that sociality was caused by positive selection on a core set of genes.

Methods

Genome Alignment, Variant Calling and Filtration

Methods for genome alignment, SNP detection, and filtering are described in detail in Chapter 2. In summary, paired-end Illumina genome sequencing was performed on ten *P. dominula* and two *P. gallicus* female worker samples. Reads were aligned to the unmasked *P. dominula* reference genome (PdomGDB r1.2) (Standage, et al. 2016) using BWA-MEM (Li and Durbin 2009). Duplicates were marked using Picard (http://broadinstitute.github.io/picard/) and indel realignment was performed with GATK (McKenna, et al. 2010; DePristo, et al. 2011). Variants were detected with GATK's HaplotypeCaller using all species-specific alignments in

unison. Variants were subsequently filtered of poor quality SNPs based on GATK's hard filter recommendations, upper and lower depth limits, missing data, and regions of high sequence homology. In addition to the previously mentioned filtering criteria, a percent threshold for low to no coverage base pairs across coding regions was established. Genes that had poor coverage for > 0.1 (10%) of the coding sequence were removed from further analysis (Appendix B – Figure 1).

Variant Annotation

Variants that passed all of the filtering criteria were annotated for predicted effects on genes using SnpEff (Cingolani, et al. 2012). SnpEff utilizes the reference sequence and genome annotation file (GFF3) to make predictions regarding the effects of the variant on the gene. Variant annotations are classified as: intergenic, intronic, synonymous/stop, non-synonymous/start, 3' UTR, 5' UTR, upstream and downstream. Genes were removed from the analysis if they contained warnings for an incomplete transcript, multiple stop codons, and no start codon. Additionally, genes were removed if they possessed annotations for lost stop codons, gain of stop codon, loss of start codon, or non-synonymous start variants. Tri-allelic variants were also discarded to limit the chances of retaining SNPs with potential sequencing error and to simplify downstream analysis for the large dataset (Wang, et al. 2013). Additionally, variants labeled as non-synonymous stops were discarded from the analysis. The remaining variants were used in subsequent analyses.

Quantifying Selection

A Bayesian implementation of the McDonald-Kreitman test (Eilertson, et al. 2012) was use to estimate the prevalence of selection acting on genes in *Polistes*. This test relies on

interspecies divergence and intraspecies polymorphisms based on synonymous and non-synonymous substitutions to estimate selection over intermediate timescales (McDonald and Kreitman 1991). The test works by comparing the synonymous sites, which are inferred to be neutral, with non-synonymous sites to estimate the degree and direction of selection (Vitti, et al. 2013). The null hypothesis for the McDonald-Kreitman test is neutral evolution, whereby mutations are predicted to have no effect on the fitness of a species. As such, the null hypothesis predicts that the ratio of non-synonymous (Pn) to synonymous (Ps) variation within a species should be equal to the ratio of non-synonymous (Dn) to synonymous (Ds) divergence between species. Accordingly, positive selection is inferred when the Dn/Ds ratio is greater than the Pn/Ps ratio.

Synonymous and non-synonymous substitutions were determined using the predicted gene annotations from SnpEff. Divergence data is based on fixed mutations between *P. dominula* and *P. gallicus* gene sequences, while polymorphisms were based on variable mutation sites in both species. Sites were removed if the comparison resulted in a triallelic variant and in cases where both species were fixed for the same allele. The Bayesian implementation of the McDonald-Kreitman test, SnPIRE, makes use of genome wide information and doesn't require a priori knowledge of species divergence parameters. The selection coefficient, gamma ($\boldsymbol{\gamma}$), is calculated for each gene, where gamma represents the average selection coefficient on non-synonymous mutations in a gene scaled by the effective population size ($\boldsymbol{\gamma} = 2N_e s$) (Eliertson *et al.* 2012). Here, we can quantify the degree of selection acting upon each gene by classifying the $\boldsymbol{\gamma}$ values into ranges: $\boldsymbol{\gamma} > 1$ strong positive selection, $0 < \boldsymbol{\gamma} < 1$ nearly neutral to weak positive selection, $-1 < \boldsymbol{\gamma} < 0$, nearly neutral to weak negative selection, and $\boldsymbol{\gamma} < -1$ strong negative selection (Torgerson, et al. 2009).

Gene ontology

Gene ontology (GO) was performed with target loci to identify enriched terms and annotation clusters related to the biological, cellular, or molecular function of genes. The gene ontology assessment was achieved using the DAVID 6.8 (2013-2016) (Huang, et al. 2009) web program using *Drosophila melanogaster* fly base gene IDs. *Polistes* sequences orthologous to *Drosophila* were identified using reciprocal Blastp matches with an E-value threshold of $1e^{-10}$. *Drosophila* orthologs were found for 6741 genes (57%) of the total 11815 annotated genes in PdomGDB r1.2.

Ortholog Hierarchical Analysis

Polistes genes were classified into taxonomic groups using OrthoDB V9, a web based catalogue that strives to classify protein coding genes into groups of orthologs from genes descended from the last common ancestor (Kriventseva, et al. 2015). The OrthoDB V9 flat files were downloaded and merged into one data file based on common field entries. The Metazoa (OrthoID: 33208), Arthropoda (OrthoID: 61921), Insecta (OrthoID: 50557), Hymenoptera (OrthoID: 7399), Aculeata (OrthoID: 22080), and Vespoidea (OrthoID: 34725) classification levels were extracted from the files and then orthologs containing a *P. dominula* association were assessed for lowest possible taxonomic level. Orthologs restricted to Aculeata species was classified as an Aculeata gene. Orthologs associated with Aculeata and at least one Hymenoptera were classified at the Hymenoptera level. Genes associated with Aculeata, Hymenoptera, and at least one other Insect, Arthropod, or Metazoan, were classified as Insecta or older. Genes that could not be categorized were assessed for orthology between *P. dominula* and *P. canadensis* to find putative *Polistes* restricted genes. *Polistes* orthologs were identified using reciprocal Blastp matches with an E-value threshold of $1e^{-10}$, also considering multiple perfect match hits.

Differential Gene Expression

Differential expression between queens and workers was examined from twelve Illumina paired-end RNA-Seq libraries of six *P. dominula* queen and worker whole heads (Standage, et al. 2016). Sequences were downloaded from GenBank (SRX1124061, SRX1124060, SRX1124059, SRX1124057, SRX1124056, SRX1124054, SRX1124053, SRX1124052, SRX1124051, SRX1124050, SRX1124049, SRX1122234) and aligned to the unmasked *P. dominula* reference genome (PdomGDB r1.2) (Standage, et al. 2016) using the default parameters of STAR's 2-pass method (Engström, et al. 2013). The STAR 2-pass method performs two alignments in which the splice junctions detected in the first alignment are used to guide the second alignment. Using the resulting BAM files, transcriptomes were assembled using Cufflinks (Trapnell, et al. 2012). Transcriptomes were produced with bias detection and correction algorithms using the *P. dominula* reference sequence (PdomGDB r1.2), and were quantitated using the annotated reference genome GFF3. Assembled transcriptomes were merged using Cuffmerge to concatenate overlapping regions that agree in splice and orientation position. Cuffdiff was then used to find significant changes in gene level expression between worker and queen castes. Sample SRX1124050 was excluded from the analysis after examination with cummerbund (Goff, et al. 2012) revealed the distribution of the FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values to be highly skewed compared to the remaining replicates. The sample was removed to avoid skewing the results and negatively affecting the differential expression analysis. Loci were removed from the analysis if they could not be identified after the merge step due to: the direct overlap of genes located on opposite strands, genes on the same strand that were merged due to shared transcripts, or clustering of three of more genes.

Remaining genes were determined to be significantly differentially expressed if they passed the following thresholds: FDR < 0.05, and FPKM > 1.

Comparative Genomics

To evaluate the correlation of selection patterns across other eusocial lineages we completed a cross species comparison with *Polistes* (Standage, et al. 2016), *Bombus* (Sadd, et al. 2015), and *Apis* (Elsik, et al. 2014). Orthologs between species pairs were identified using reciprocal Blastp matches with an E-value threshold of $1e^{-10}$. Reciprocal matches with corresponding gamma values for *Polistes* and *Apis* (Harpur, et al. 2014), *Polistes* and *Bombus* (Harpur, et al. in prep), and *Bombus* and *Apis*, revealed 8573, 8178, and 9434 ortholog matches respectively. A three-way reciprocal blast between all three species produced 4548 gene orthologs.

Results

Overview

We sequenced the genomes of ten *P. dominula* and two *P. gallicus* female workers collected in Tuscany Italy. The genome alignments of *P. dominula* individuals yielded an average sequencing depth of 23X, while *P. gallicus* had a depth of 20X. SNPs called independently for each species were filtered using strict criteria outlined in chapter two. Post filtering, there were 1,941,335 SNPs in the *P. dominula* dataset and 2,903,697 SNPs in the *P. gallicus* dataset. We removed 506 genes due to incomplete transcripts, multiple stop codons, or no start codon, and 1641 genes due to a loss of stop codon, gain of stop codon, loss of start codon, non-synonymous start variants, or poor gene coverage. The remaining genes and variants were used to conduct the population genomic analyses.

Quantifying Selection

We used a Bayesian implementation of the McDonald-Kreitman test to estimate selection on *Polistes* genes. We calculated the selection coefficient for 9668 (81.8%) genes from the 11815 annotated genes in the *Polistes dominula* genome (PdomGDB r1.2) (Standage, et al. 2016). We found that 81% of genes (7815) possessed a $\boldsymbol{\gamma}$ value between 0 and 1, and 988 genes (10.25%) possessed a value consistent with strong positive selection $\boldsymbol{\gamma} > 1$ (Appendix B – Figure 2). The genome wide selection pattern on coding genes revealed an average $\boldsymbol{\gamma}$ value of 0.48, indicating a slightly positive average selection coefficient. A comparable average has also been observed in *Bombus* ($\boldsymbol{\gamma} = 0.56$) (Harpur, et al. in prep) which contrasts with the nearly neutral average of *Apis* ($\boldsymbol{\gamma} = 0.14$) (Harpur, et al. 2014).

Genes that showed signs of being under strong positive selection ($\boldsymbol{\gamma} > 1$) were assessed for gene ontology terms to identify the function of adaptively evolving loci. We found that genes under positive selection were mostly associated with transcription and gene expression, particularly zinc finger domains and RNA polymerases. Additionally, there were clusters of terms associated with fatty acid synthesis and metabolism, post-transcriptional regulation, and metal binding (S1–Table 1.1-1.2).

Ortholog Hierarchical Analysis

Using OrthoDB v9 (Kriventseva, et al. 2015) we classified *P. dominula* genes into hierarchical orthologous groups. In total 9369 were classified into the Metazoa taxonomic level, 358 into Athropoda, 450 into Insecta, 200 into Hymenoptera, 157 into Aculeata, and 70 into *Polistes*. There were 1211 genes or 10% of the genome that lacked sequence homology to other species lineages and were placed into an unclassified group. In order to determine whether

conserved or taxonomically restricted genes demonstrate a greater effect on the adaptive evolution of *Polistes*, we assessed selection coefficient for each taxonomic level. The Metazoa, Athropoda, and Insecta categorized genes were concatenated into a group of 'old' or 'conserved' genes, while the narrow taxonomic groups are considered to be independent taxonomically restricted gene (TRGs) groups.

Considering a selection coefficient could only be estimated for 9668 genes, there were 8673, 155, 119, and 45 genes assigned to Insecta and Older, Hymenoptera, Aculeata, and *Polistes* gene groups respectively. We found that *P. dominula* genes with signs of positive selection ($\boldsymbol{\gamma} > 0$) had the highest average in the *Polistes* group (0.598 ± SEM 0.060), while the Aculeata ortholog group contained the lowest (0.510 ± SEM 0.034). There was, however, no significant difference between the four classification groups ($F_{3,8147} = 0.644$, $p = 0.587$) (Appendix B – Figure 3). We also found that the proportion of genes indicating strong positive selection ($\boldsymbol{\gamma} > 1$) was highest in the *Polistes* (17.8%, n=8) and lowest in the Hymenoptera (6.5%, n=10); Insecta (10.6%, n=919), Aculeata (7.6%, n=9). However, the proportions of genes with $\boldsymbol{\gamma} > 1$ did not significantly diverge from expected outcomes ($\chi^2 = 5.761$, df=3, p=0.124).

Worker and Queen Phenotypes

We evaluated the expression patterns of six queens and five workers to establish queen and worker biased gene lists. Overall there were 8946 (75.7%) genes that were successfully tested and could be individualized, while 7612 (85.1%) of those genes had a corresponding gamma value. There were 7211 genes that were non-differentially expressed between the two castes, along with 114 and 287 genes determined to be up-regulated in queens and workers respectively (Appendix B – Figure 4). We found that genes that were up-regulated in queens had a higher average $\boldsymbol{\gamma}$ value (0.593 ± SEM 0.0493) when compared with those that were up-

regulated in workers (0.496 ± SEM 0.0268) and non-differentially expressed genes (0.481 ± SEM 0.00517) (Appendix B – Figure 5). Overall there was a significant difference among the three groups ($F_{2,7609}$ = 3.735, p=0.024), particularly between the queen and non-differentially expressed genes (TukeyHSD p=0.02), but not between queen and workers (TukeyHSD p=0.11), or worker and non-differentially expressed genes (TukeyHSD p=0.85). We also found that the proportion of genes indicating strong positive selection ($\boldsymbol{\gamma} > 1$) was highest in queen biased genes (19.3%, n=22) compared to worker biased (11.2%, n=32), and non-differentially expressed genes (10.1%, n=728) (χ^2=9.47, df=2, p=0.008). In particular, there was a significantly higher proportion of genes with strong positive selection in queen biased genes relative to worker biased genes (χ^2=4.02, df=1, p=0.045), and worker and queen biased genes relative to non-differentially expressed genes (χ^2=4.20, df=1, p=0.040).

We also examined whether queen and worker biased genes where enriched for TRGs. We found a significantly higher proportion of genes classified at the *Polistes* taxonomic level associated with worker biased genes relative to queen and non-differentially expressed genes (χ^2=59.41, df=1, p<0.0001), but there was no significant association with either of the other three taxonomic categories (χ^2, p>0.60).

Comparative Genomics

We compared the proportion of overlapping genes under strong positive selection in three eusocial species across two independent lineages: *P. dominula, B. impatiens*, and *A. mellifera*. Shared orthologs with corresponding gamma values between *A. mellifera* and *P. dominula* included 7285 genes, while *P. dominula* and *B. impatiens* shared 7147 genes. There were 4472 gene orthologs shared among all three species. We tested for over-representation of overlapping genes with strong positive selection in each species pair and found a slightly larger and more

strongly associated overlap for *A. mellifera* and *B. impatiens* (FET $p< 2.2e^{-16}$, OR= 2.700135) compared with *B. impatiens* and *P. dominula* (FET $p< 2.2e^{-16}$, OR= 2.158572), and *A. mellifera* and *P. dominula* (FET $p=2.25e^{-07}$, OR= 1.855172) (Appendix B – Figure 6-7). A similar pattern was also observed in the three-way comparison that saw the most significant overlap and association between *A. mellifera* and *B. impatiens* (FET p< 2.2e-16, OR=2810058) compared with *P. dominula* and *B. impatiens* (FET $p= 1.823e^{-11}$, OR=2.121377), or *A. mellifera* and *P. dominula* (FET $p= 3.481e^{-05}$, OR=1.834916) (Appendix B – Figure 6-7).

Using the overlapping gene sets for each pairwise and three-way comparison, gene ontology analysis was performed to identify the function of adaptively evolving loci. While there was not strong enrichment for GO terms and functional clusters, we found that genes under positive selection overlapping between *P. dominula* and *A. mellifera* were associated with functions related to transmembrane proteins and receptor and signalling activity. There was also evidence of functions related to learning and memory, olfactory learning, as well as immune response and defence. Likewise, genes with positive selection were associated with transmembrane proteins, immunity, and defence response were shared between *P. dominula* and *B. impatiens* as were terms for RNA polymerase, fatty acid metabolism, and zinc fingers (S1–Table 2.1-2.4).

Discussion

Patterns of selection

The patterns and prevalence of selection across the *Polistes* genome revealed the average selection coefficient of adaptively evolving loci to be weakly positively selected. Likewise, a similar genomic pattern has also been observed in *Bombus* which is also classified as a primitively eusocial species (Harpur, et al. in prep). These observations contrast with results previously described for advanced eusocial *Apis*, whose average genome selection coefficient is nearly neutral (Harpur, et al. 2014). The correlation between the level of sociality and the average genome selection coefficient could be a result of effective population size (N_e) or an increase in selection. The effective population size of social insects has been shown to decrease as sociality increases (Romiguier, et al. 2014); resulting in a lower appearance and fixation of beneficial mutations (Galtier 2015). Thus, lower Ne would result in lower overall selection coefficients (Eyre-Walker 2002). Alternatively, relaxed constraint enables the acquisition of mutations, which may facilitate rapid directional (positive) selection when the variant is beneficial. In a eusocial context, increased selection would allow genes to be adopted for specialized tasks contributing to queen and worker phenotypes (Gadagkar 1997). Once specialization has been established we could expect genomic constraint to increase, and selection to decrease in order to establish stabilising selection in the species (Ghalambor, et al. 2007). While these hypotheses could explain the variation in gamma values, it is difficult to compare species directly.

It has been proposed that changes in gene regulation are important during the earlier stages of social evolution (Rehan and Toth 2015). Our study lends some support to this hypothesis because many of the genes found to be under positive selection in primitively

eusocial *Polistes* have functions associated with regulation of transcription, including zinc fingers and RNA polymerase. Zinc fingers are a class of small regulatory proteins that can bind DNA, RNA, and proteins; the largest group of which are C_2H_2 characterised proteins (Iuchi 2001). This group is known to act as regulatory proteins by binding to DNA and controlling transcription of target genes. RNA polymerase is involved in producing primary transcript RNA, and thus, in the control of gene transcription. Additionally, we found evidence for mRNA processing including polyadenylation and 3' end processing which are involved in post-transcriptional regulation (Leff and Rosenfeld 1986). As such, positive selection on genes related to transcription in *Polistes* may highlight the importance of gene regulation in generating queen and worker phenotypes during the early stages of social evolution.

Gene Hierarchy

Taxonomically restricted genes have been implicated in the development of novel traits in social insect lineages, including the formation of castes and eusocial behaviour (Johnson and Tsutsui 2011; Ferreira, et al. 2013; Sumner 2014). Previous studies on advanced eusocial *Apis* have shown that novel genes are significantly enriched for positive selection, relative to conserved genes (Harpur, et al. 2014). Additionally, novel genes in *Apis* have been shown to be disproportionately associated with the worker phenotype, which exhibits highly derived traits thought to have evolved since eusociality (Johnson and Tsutsui 2011; Jasper, et al. 2014). Indeed, enrichment of positive selection on novel genes correlates with increased reliance on worker contributions to colony development. Unlike the honey bee study (Harpur, et al. 2014), our results suggest that novel genes may not be very important during the early stages of social evolution. In *Polistes*, we found no significant differences in average gamma for conserved

versus taxonomically restricted genes – in stark contrast to the honey bee dataset. *Polistes* queens and workers lack the same level of highly derived attributes that are proposed to be extensively influenced by novel genes, which could thus explain our findings. This result is in broad agreement with a recent hypothesis suggesting that novel genes are more important during the last stages of social evolution (Rehan and Toth 2015).

Queen and Worker Phenotypes

The queen and worker phenotypes that have evolved in eusocial species display prominent behavioural, physiological, and often morphological specialization. Differential expression is thought to account for much of these differences, resulting in caste biasedly expressed genes. Previous research on advanced eusocial *Apis* has shown that worker biased proteins are enriched for positive selection suggesting the worker phenotype is the primary means of adaptation in honey bees (Harpur, et al. 2014). We tested the concordance of this observation in *Polistes* and found a noticeable contrast. In *Polistes* we found that worker biased genes did not indicate signs for enriched positive selection. In comparison, queen biased genes were marginally enriched for signs of adaptive evolution. These differences perhaps reflect that the success of primitively eusocial colonies is more dependent on queens, as reflected in *Polistes*. On the other hand, workers in advanced eusocial species, as in the honey bee, are present during the entire colony cycle and have a significantly larger role in colony success.

Previous studies of *Polistes* species have indicated that caste differentially expressed genes are enriched for functions related to lipid metabolism (Sumner, et al. 2006; Hunt, et al. 2010; Toth, et al. 2010; Standage, et al. 2016). We found genes that showed signs of positive selection also showed evidence for gene functions related to fatty acid biosynthesis and

metabolism. Functions related to the production and breakdown of lipid stores are essential for species that rely on the success of overwintering gynes and foundress colony initiation (Dapporto and Palagi 2006; Dapporto, et al. 2006; Kovacs and Goodisman 2012). When late summer *Polistes* gynes emerge, they possess elevated fat stores the metabolism of which is expected to provide energy during quiescence (Toth, et al. 2009). Related findings have also been seen in a primitively eusocial population of the socially polymorphic halictid bee *Lasioglossum albipes,* which exhibits an expansion of the inositol monophosphatase gene family associated with lipid metabolism (Kocher, et al. 2013). These results suggest that traits associated with overwintering play an important role for the fitness of future queens in *Polistes*.

Comparative Genomics

We compared loci found to be under strong positive selection in *Polistes* to orthologous genes under strong positive selection in advanced eusocial *Apis* and primitively eusocial *Bombus* to explore the degree of evolutionary convergence across independently evolved lineages. We found that there was a greater overlap in genes under strong positive selection between *Polistes* and *Bombus* compared to *Polistes* and *Apis*. This is an intriguing result because even though Vespidae diverged from Apidae 170-140 MYA, the evolutionary distance between *Polistes* and *Bombus* is equal to that between *Polistes* and *Apis* (Cardinal and Danforth 2013; Biewer, et al. 2015). As such, this suggests that species of equal social status have greater similarity in selection patterns. This is a compelling result supporting the idea that sociality in *Polistes* and *Bombus* involves positive selection on a few key genes, or that sociality can drive patterns of adaptive evolution for a few key genes.

Genes under strong positive selection that overlapped with *Polistes*, *Apis*, and *Bombus* indicated functions related to transmembrane proteins, receptors and signalling activity, as well as immune response and defence. There was also presence of terms related to learning and memory, and olfactory learning between *Apis* and *Polistes*, and RNA polymerase, fatty acid metabolism, and zinc fingers between *Bombus* and *Polistes*. Receptor and signalling activity and transmembrane proteins included genes associated with G-protein couple receptors which are believed to be involved in the sensitivity of odor detection, making them important in chemical communication and behavioural response (Hildebrand and Shepherd 1997; Bonasio, et al. 2010). Additionally, olfactory learning is involved in long lasting adaptive behavioural change in response to olfactory cues (Davis 2004). Chemical cues and pheromones are especially important in social evolution as they influence nest (Breed 1998) and brood recognition (Slessor, et al. 2005) defence (Breed, et al. 2004; Slessor, et al. 2005), and ovary development of workers (Hoover, et al. 2003).

Additionally, immune related functions including immune response and defence are shared among lineages. Immune response is a biological process that functions in reaction to a potential invasive threat, while defence is a trigger to the presence of a foreign body or the occurrence of an injury to mitigate damage and infection. Immune response mechanisms are essential for social insects as increased colony size and interactions affect the necessity to mitigate pathogen transmission (Chen, et al. 2006; Cremer, et al. 2007). While there was not a strong enrichment of GO terms, these results suggest that gene functions related to communication and immune defense are shared among lineages and could be a key component in social insect evolution.

Conclusion

Our population genomic study of the primitively eusocial *Polistes* generated findings that were different, and often opposite, to those found in a recently published *Apis* study. The findings are compelling as they suggest that patterns of molecular evolution may reflect differences in social organization and complexity. Although it is difficult to reach any strong conclusions based on only two population genomic studies, the diminishing costs of sequencing will enable more population genomic studies of insects with different types of social organization. This will enable us to fully understand the molecular evolution process involved in the evolution and elaboration of eusociality, and how sociality influences patterns of molecular evolution.

Citations

Amsalem E, Galbraith DA, Cnaani J, Teal PEA, Grozinger CM. 2015. Conservation and modification of genetic and physiological toolkits underpinning diapause in bumble bee queens. Molecular Ecology 24:5596-5615.

Asencot M, Lensky Y. 1988. The effect of soluble sugars in stored royal jelly on the differentiation of female honeybee (Apis mellifera L.) larvae to queens. Insect biochemistry 18:127-133.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in Drosophila. Proceedings of the National Academy of Sciences 110:17409-17414.

Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Current protocols in bioinformatics:11.10. 11-11.10. 33.

Batra S. 1966. Nests and social behavior of halictine bees of India (Hymenoptera: Halictidae). Indian Journal of Entomology 28:375-393.

Ben-Shahar Y, Robichon A, Sokolowski MB, Robinson GE. 2002. Influence of Gene Action Across Different Time Scales on Behavior.pdf. Science 296:741-744.

Berens AJ, Hunt JH, Toth AL. 2015a. Comparative Transcriptomics of Convergent Evolution: Different Genes but Conserved Pathways Underlie Caste Phenotypes across Lineages of Eusocial Insects. Molecular biology and evolution 32:690-703.

Berens AJ, Hunt JH, Toth AL. 2015b. Nourishment level affects caste-related gene expression in Polistes wasps. BMC Genomics 16:235.

Biewer M, Schlesinger F, Hasselmann M. 2015. The evolutionary dynamics of major regulators for sexual development among Hymenoptera species. Frontiers in genetics 6:1-11.

Bijma P, Wade MJ. 2008. The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. Journal of evolutionary biology 21:1175-1188.

Bloch G, Wheeler DE, Robinson GE. 2002. Endocrine influences on the organization of insect societies. In: Pfaff D, editor. Hormones, brain, and behavior. San Diego: Academic Press.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics:btu170.

Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, et al. 2010. Genomic Comparison of the Ants Camponotus floridanus and Harpegnathos saltator. Science 329:1068-1071.

Brady SG, Schultz TR, Fisher BL, Ward PS. 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. Proceedings of the National Academy of Sciences 103:18172-18177.

Breed M. 1998. Recognition Pheromones of the Honey Bee. BioScience 48:463-470.

Breed MD, Guzmán-Novoa E, Hunt GJ. 2004. Defensive behavior of honey bees: organization, genetics, and comparisons with other bees. Annual Reviews in Entomology 49:271-298.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV. 2003. The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology and Evolution 18:249-256.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.

Buttstedt A, Moritz RFA, Erler S. 2014. Origin and function of the major royal jelly proteins of the honeybee (Apis mellifera) as members of the
yellow gene family. Biological Reviews 89:255-269.

Cardinal S, Danforth B. 2013. Bees diversified in the age of eudicots. Proceedings of the Royal Society B Biological Sciences 280:20122686.

Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen J-B, Frazer KA. 2014. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC bioinformatics 15:125.
Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124-3140.
Charles R. 1859. Darwin, On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. In: London: John Murray.
Chen Y, Evans J, Feldlaufer M. 2006. Horizontal and vertical transmission of viruses in the honey bee, Apis mellifera. Journal of invertebrate pathology 92:152-159.
Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6:80-92.
Crawford JE, Lazzaro BP. 2012. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. Frontiers in genetics 3:66.
Cremer S, Armitage SA, Schmid-Hempel P. 2007. Social immunity. Current biology 17:R693-R702.
Crozier RH, Pamilo P. 1996. Evolution of social insect colonies : sex allocation and kin selection. Oxford ; New York: Oxford University Press.
Dapporto L, Palagi E. 2006. Wasps in the shadow : Looking at the pre-hibernating clusters of Polistes dominulus. Annales Zoologici Fennici 43:583-594.
Dapporto L, Palagi E, Cini A, Turillazzi S. 2006. Prehibernating aggregations of Polistes dominulus: An occasion to study early dominance assessment in social insects. Naturwissenschaften 93:321-324.
Darwin C. 1888. The descent of man, and selection in relation to sex: Murray.
Davis RL. 2004. Olfactory learning. Neuron 44:31-48.
DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43:491-498.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21.
Eilertson KE, Booth JG, Bustamante CD. 2012. SnIPRE: selection inference using a Poisson random effects model. PLoS computational biology 8:e1002806.
Elsik CG, Worley KC, Bennett AK, Beye M, Camara F, Childers CP, de Graaf DC, Debyser G, Deng J, Devreese B. 2014. Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics 15:1.
Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, et al. 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. Nature methods 10:1185-1191.
Evans LM, Slavov GT, Rodgers-Melnick E, Martin J, Ranjan P, Muchero W, Brunner AM, Schackwitz W, Gunter L, Chen J-G, et al. 2014. Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. Nature genetics advance on:1089-1096.
Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics 162:2017-2024.
Ferreira PG, Patalano S, Chauhan R, Ffrench-Constant R, Gabaldón T, Guigó R, Sumner S. 2013. Transcriptome analyses of primitively eusocial wasps reveal novel insights into the evolution of sociality and the origin of alternative phenotypes. Genome biology 14:R20.
Freeland JR, Petersen SD, Kirk H. 2011. Molecular Ecology: Wiley.
Gadagkar R. 1997. The evolution of caste polymorphism in social insects: Genetic release followed by diversifying evolution. Journal of Genetics 76:167-179.

Galtier N. 2015. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. Journal of Chemical Information and Modeling:1-23.
Gardner A, West SA, Wild G. 2011. The genetical theory of kin selection. Journal of evolutionary biology 24:1020-1043.
Ghalambor CK, McKay JK, Carroll SP, Reznick DN. 2007. Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. Functional Ecology 21:394-407.
Goff L, Trapnell C, Kelley D. 2012. cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R package version 2.
Gondro C, Porto-Neto LR, Lee SH. 2014. snpqc - an R pipeline for quality control of Illumina SNP genotyping array data. Animal Genetics:758-761.
Grozinger CM, Fan Y, Hoover SER, Winston ML. 2007. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (Apis mellifera). Molecular ecology 16:4837-4848.
Gudbjartsson DF, Sulem P, Helgason H, Gylfason A, Gudjonsson SA, Zink F, Oddson A, Magnusson G, Halldorsson BV, Hjartarson E, et al. 2015. Sequence variants from whole genome sequencing a large group of Icelanders. Scientific data 2:150011.
Hamilton WD. 1964. The genetical evolution of social behaviour. II. Journal of theoretical biology 7:17-52.
Harpur B, Dey A, Albert J, Patel S, Hines H, Hasselmann M, Packer L, Zayed A. in prep. Contribution of queen and worker traits to adaptive evolution differs between bumble bees and honey bees.
Harpur Ba, Kent CF, Molodtsova D, Lebon JMD, Alqarni AS, Owayss Aa, Zayed A. 2014. Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. Proceedings of the National Academy of Sciences of the United States of America 111:2614-2619.
Hildebrand JG, Shepherd GM. 1997. MECHANISMS OF OLFACTORY DISCRIMINATION : Converging Evidence for Common Principles Across Phyla.595-631.
Hines HM, Hunt JH, O'Connor TK, Gillespie JJ, Cameron Sa. 2007. Multigene phylogeny reveals eusociality evolved twice in vespid wasps. Proceedings of the National Academy of Sciences of the United States of America 104:3295-3299.
Hoover SER, Keeling CI, Winston ML, Slessor KN. 2003. The effect of queen pheromones on worker honey bee ovary development. Naturwissenschaften 90:477-480.
Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4:44-57.
Hunt JH. 2012. A conceptual model for the origin of worker behaviour and adaptation of eusociality. Journal of evolutionary biology 25:1-19.
Hunt JH. 2007. The Evolutionof Social Wasp.259.
Hunt JH, Karsai I. 2002. Food Quantity Affects Traits of Offspring in the Paper Wasp Polistes metricus (Hymenoptera : Vespidae). Population Ecology 31:99-106.
Hunt JH, Wolschin F, Henshaw MT, Newman TC, Toth AL, Amdam GV. 2010. Differential gene expression and protein abundance evince ontogenetic bias toward castes in a primitively eusocial wasp. PloS one 5:e10674.
Iuchi S. 2001. Three classes of C2H2 zinc finger proteins. Cellular and molecular life sciences : CMLS 58:625-635.
Jandt JM, Tibbetts EA, Toth AL. 2014. Polistes paper wasps: a model genus for the study of social dominance hierarchies. Insectes Sociaux 61:11-27.
Jasper WC, Linksvayer TA, Atallah J, Friedman D, Chiu JC, Johnson BR. 2014. Large-Scale Coding Sequence Change Underlies the Evolution of Postdevelopmental Novelty in Honey Bees. Molecular biology and evolution 32:334-346.

Johnson BR, Tsutsui ND. 2011. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC Genomics 12:164.
Julian GE, Fewell JH, Gadau J, Johnson RA, Larrabee D. 2002. Genetic determination of the queen caste in an ant hybrid zone. Proceedings of the National Academy of Sciences 99:8157-8160.
Keller I, Bensasson D, Nichols RA. 2007. Transition-transversion bias is not universal: A counter example from grasshopper pseudogenes. PLoS Genetics 3:0185-0191.
Keller L. 1999. Levels of selection in evolution: Princeton University Press.
Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics 25:404-413.
Kocher SD, Li C, Yang W, Tan H, Yi SV, Yang X, Hoekstra HE, Zhang G, Pierce NE, Yu DW. 2013. The draft genome of a socially polymorphic halictid bee, Lasioglossum albipes. Genome biology 14:R142.
Kovacs JL, Goodisman MA. 2012. Effects of size, shape, genotype, and mating status on queen overwintering survival in the social wasp Vespula maculifrons. Environmental entomology 41:1612-1620.
Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Sim??o FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM. 2015. OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. Nucleic Acids Research 43:D250-D256.
Laurie C, Doheny K, Mirel D. 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic … 34:591-602.
Leff SE, Rosenfeld MG. 1986. Complex transcriptional units: diversity in gene expression by alternative RNA processing. Annual review of biochemistry 55:1091-1117.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-2993.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 25:1754-1760.
Liebert aE, Gamboa GJ, Stamp NE, Curtis TR, Monnet KM, Turillazzi S, Starks PT. 2006. Genetics, behavior and ecology of a paper wasp invasion: Polistes dominulus in North America. Annales Zoologici Fennici 43:595-624.
Linksvayer Ta, Wade MJ. 2005. The evolutionary origin and elaboration of sociality in the aculeate Hymenoptera: maternal effects, sib-social effects, and heterochrony. The Quarterly review of biology 80:317-336.
Linksvayer TA, Wade MJ, Gordon DM. 2006. GENETIC CASTE DETERMINATION IN HARVESTER ANTS: POSSIBLE ORIGIN AND MAINTENANCE BY CYTO-NUCLEAR EPISTASIS. Ecology 87:2185-2193.
Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. 2012. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics 13:1.
Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R. 2010. The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. PLoS Biology 8.
Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, Donnelly P, Faraone SV, Frazer K, Gabriel S. 2007. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nature genetics 39:1045-1051.
Manske M. 2012. Analysis of Plasmodium falciparum diversity in natural infections by deep sequencing. Nature 487:375-379.
Marshall JAR. 2011. Group selection and kin selection: formally equivalent approaches. Trends in ecology & evolution 26:325-332.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20:1297-1303.
Michener C. 1969. Comparative social behavior of bees. Annual review of entomology 14:299-342.
Michener CD. 1974. The social behavior of the bees. A comparative study. Massachusetts: Harvard University Press.
Morandin C, Tin MMY, Abril S, Gómez C, Pontieri L, Schiøtt M, Sundström L, Tsuji K, Pedersen JS, Helanterä H, et al. 2016. Comparative transcriptomics reveals the conserved building blocks involved in parallel evolution of diverse phenotypic traits in ants. Genome Biology 17:43.
Nielsen R. 2005. Molecular signatures of natural selection. Annual review of genetics 39:197-218.
Noble WS. 2010. How does multiple testing correction work? 27:1135-1137.
Nowak MA, Tarnita CE, Wilson EO. 2010. The evolution of eusociality. Nature 466:1057-1062.
Okasha S. 2005. Multilevel Selection and the Major Transitions in Evolution. Philosophy of Science 72:1013-1025.
Patalano S, Vlasova A, Wyatt C, Ewels P, Camara F, Ferreira PG, Asher CL, Jurkowski TP, Segonds-Pichon A, Bachman M, et al. 2015. Molecular signatures of plastic phenotypes in two eusocial insect species with simple societies. Proceedings of the National Academy of Sciences of the United States of America 112:13970-13975.
Pongpanich M, Sullivan PF, Tzeng J-Y. 2010. A quality control algorithm for filtering SNPs in genome-wide association studies. Bioinformatics (Oxford, England) 26:1731-1737.
Rehan SM, Berens AJ, Toth AL. 2014. At the brink of eusociality: transcriptomic correlates of worker behaviour in a small carpenter bee. BMC Evolutionary Biology 14:260.
Rehan SM, Glastad KM, Lawson SP, Hunt BG. 2016. The genome and methylome of a subsocial small carpenter bee, Ceratina calcarata. Genome Biology and Evolution:evw079.
Rehan SM, Toth AL. 2015. Climbing the social ladder: the molecular evolution of sociality. Trends in ecology & evolution 30:426-433.
Robinson GE, Grozinger CM, Whitfield CW. 2005. Sociogenomics: social life in molecular terms. Nature Reviews Genetics 6:257-270.
Romiguier J, Lourenco J, Gayral P, Faivre N, Weinert La, Ravel S, Ballenghien M, Cahais V, Bernard a, Loire E, et al. 2014. Population genomics of eusocial insects: The costs of a vertebrate-like effective population size. Journal of evolutionary biology 27:593-603.
Ronai I, Barton Da, Oldroyd BP, Vergoz V. 2015. Regulation of oogenesis in honey bee workers via programed cell death. Journal of Insect Physiology 81:36-41.
Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, Gadau J, Grimmelikhuijzen CJP, Hasselmann M, Lozier JD, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. Genome biology 16:76.
Schwarz MP, Richards MH, Danforth BN. 2007. Changing paradigms in insect social evolution: insights from halictine and allodapine bees. Annu. Rev. Entomol. 52:127-150.
Simola DF, Wissler L, Donahue G, Waterhouse RM, Helmkampf M, Roux J, Nygaard S, Glastad KM, Hagen DE, Viljakainen L, et al. 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. Genome research 23:1235-1247.
Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881-885.
Slessor KN, Winston ML, Le Conte Y. 2005. Pheromone communication in the honeybee (Apis mellifera L.). Journal of chemical ecology 31:2731-2745.
Smith JM. 1964. Group selection and kin selection. Nature 201:1145-1147.

Standage DS, Berens AJ, Glastad KM, Severin AJ, Brendel VP, Toth AL. 2016. Genome, transcriptome and methylome sequencing of a primitively eusocial wasp reveal a greatly reduced DNA methylation system in a social insect. Molecular Ecology 25:1769-1784.
Sumner S. 2014. The importance of genomic novelty in social evolution. Molecular Ecology 23:26-28.
Sumner S, Pereboom JJM, Jordan WC. 2006. Differential gene expression and phenotypic plasticity in behavioural castes of the primitively eusocial wasp, Polistes canadensis. Proceedings. Biological sciences / The Royal Society 273:19-26.
Szathmáry E, Smith JM. 1995. The major evolutionary transitions. In. Nature. p. 227-232.
Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, Sninsky JJ, Cargill M, Adams MD, Bustamante CD, et al. 2009. Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. PLoS Genetics 5.
Toth aL, Bilof KBJ, Henshaw MT, Hunt JH, Robinson GE. 2009. Lipid stores, ovary development, and brain gene expression in Polistes metricus females. Insectes Sociaux 56:77-84.
Toth AL, Robinson GE. 2009. Evo-devo and the evolution of social behavior: brain gene expression analyses in social insects. Cold Spring Harbor symposia on quantitative biology 74:419-426.
Toth AL, Tooker JF, Radhakrishnan S, Minard R, Henshaw MT, Grozinger CM. 2014. Shared genes related to aggression, rather than chemical communication, are associated with reproductive dominance in paper wasps (Polistes metricus). BMC Genomics 15:75.
Toth AL, Varala K, Henshaw MT, Rodriguez-Zas SL, Hudson ME, Robinson GE. 2010. Brain transcriptomic analysis in paper wasps identifies genes associated with behaviour across social insect lineages. Proceedings of the Royal Society B: Biological Sciences 277:2139-2148.
Toth AL, Varala K, Newman TC, Miguez FE, Hutchison SK, Willoughby DA, Simons JF, Egholm M, Hunt JH, Hudson ME, et al. 2007. Wasp Gene Expression Supports an Evolutionary Link Between Maternal Behavior and Eusociality. Science 318:441-444.
Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols 7:562-578.
Trivers RL, Hare H. 1976. Haplodiploidy and the evolution of the social insects. Science 191:249-263.
Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jørgensen T. 2008. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nature genetics 40:1098-1102.
van Veelen M, García J, Sabelis MW, Egas M. 2012. Group selection and inclusive fitness are not equivalent; the Price equation vs. models and statistics. Journal of Theoretical Biology 299:64-80.
Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. Annual review of genetics 47:97-120.
Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. 2014. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics 31:318-323.
Wang Y, Lu J, Yu J, Gibbs Ra, Yu F. 2013. An integrative variant analysis pipeline for accurate genotype / haplotype inference in population NGS data An integrative variant analysis pipeline for accurate genotype / haplotype inference in population NGS data.833-842.
West-Eberhard M. 1969. The Social Biology of Polistine Wasps. University of Michigan Ann Harbor:1-101.
West-Eberhard MJ. 1989. Phenotypic plasticity and the origins of diversity. Annual Review of Ecology and Systematics 20:249-278.
West-Eberhard MJ, Turillazzi S. 1996. Natural history and evolution of paper wasps: New.
Williams GC. 2008. Adaptation and natural selection: a critique of some current evolutionary thought: Princeton University Press.

Wilson DS, Wilson EO. 2007. Rethinking the theoretical foundation of sociobiology. The Quarterly review of biology 82:327-348.
Wilson EO. 1971. The insect societies. Cambridge, Mass.,: Belknap Press of Harvard University Press.
Woodard SH, Bloch GM, Band MR, Robinson GE. 2014. Molecular heterochrony and the evolution of sociality in bumblebees (Bombus terrestris). Proceedings. Biological sciences / The Royal Society 281:20132419.
Woodard SH, Fischman BJ, Venkat A, Hudson ME, Varala K, Cameron Sa, Clark AG, Robinson GE. 2011. Genes involved in convergent evolution of eusociality in bees. Proceedings of the National Academy of Sciences of the United States of America 108:7472-7477.
Yan H, Simola DF, Bonasio R, Liebig J, Berger SL, Reinberg D. 2014. Eusocial insects as emerging models for behavioural epigenetics. Nature reviews. Genetics 15:677-688.
Zhang J. 2003. Evolution by gene duplication: an update. Trends in ecology & evolution 18:292-298.

Appendix A: Chapter 2 Tables and Figures

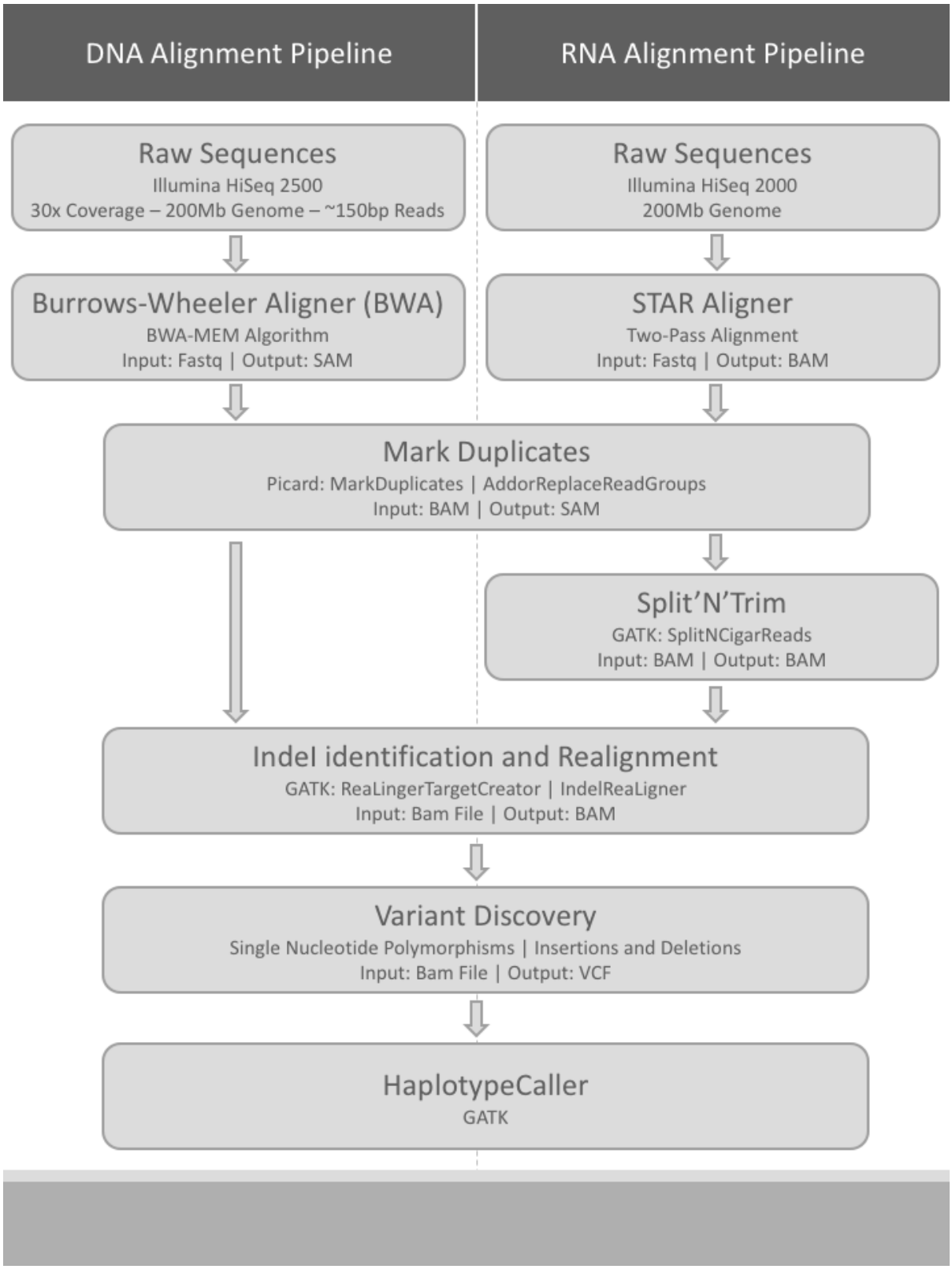

Figure 1: Pipeline for calling SNPs in genomic and transcriptomic datasets

Table 1 – Average depth of coverage of each aligned BAM file.

SAMPLE ID	AVERAGE DEPTH OF COVERAGE
***POLISTES DOMINULA* 1**	19.9179
***POLISTES DOMINULA* 2**	27.4886
***POLISTES DOMINULA* 3**	24.669
***POLISTES DOMINULA* 4**	22.2869
***POLISTES DOMINULA* 5**	22.2226
***POLISTES DOMINULA* 6**	18.4522
***POLISTES DOMINULA* 7**	24.9255
***POLISTES DOMINULA* 8**	25.08
***POLISTES DOMINULA* 9**	22.7826
***POLISTES DOMINULA* 10**	23.7727
***POLISTES GALLICUS* 4**	22.1792
***POLISTES GALLICUS* 8**	18.5674

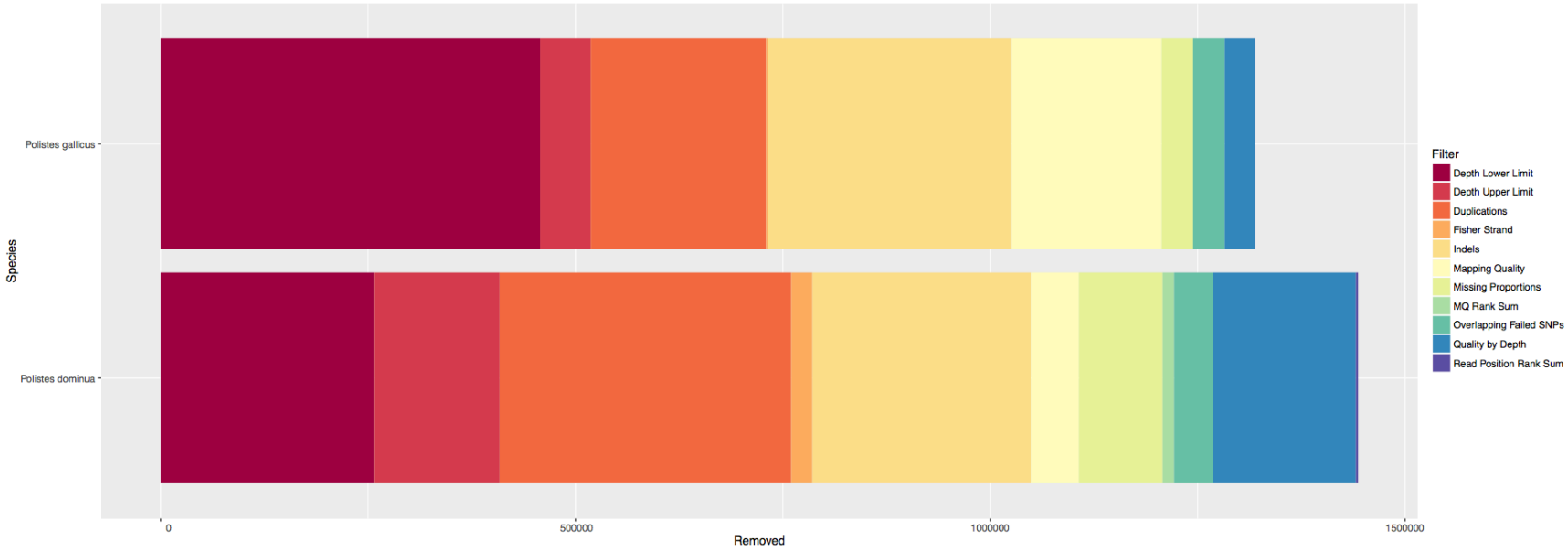

Figure 2: Proportion of removed SNPs for *Polistes dominula* and *gallicus* by filter type.

Table 2 – Number of SNPs removed after applying each filter for *Polistes dominula* (some SNPs have multiple labels).

HAPLOTYPECALLER VARIANT CALLER AND FILTRATION | *POLISTES DOMINULA*		
	Raw Variants:	SNPS: **3,105,907**
FILTERS:		
INDELS & MASK EXTENSION 5	-mask –maskExtension 5	- 263,291
MAPPING QUALITY	MQ < 40	- 57,946
FISHER STRAND	FS > 60	- 25,529
QUALITY BY DEPTH	QD < 5	- 171,927
MAPPING QUALITY RANK SUM	MQRankSum < -12.5	- 13,860
READ POSITION RANK SUM	ReadPosRankSum < -8.0	- 2,636
TOTAL DEPTH OF COVERAGE	DP < 100	- 257,005
TOTAL DEPTH OF COVERAGE	DP > 350	- 151,647
MISSING PROPORTIONS		- 100,884
REPETITIVE REGIONS & DUPLICATED GENES		- 351,535
DISCARDED SNPS OVERLAP		- 47,172
	Total Removed:	**1,164,572 or 37%**
	Total Remaining:	**1,941,335**

Table 3 – Number of SNPs removed after applying each filter for *Polistes gallicus* (some SNPs have multiple labels).

HAPLOTYPECALLER VARIANT CALLER AND FILTRATION | *POLISTES GALLICUS*		
	Raw Variants:	SNPS: **3,998,783**
FILTERS:		
INDELS & MASK EXTENSION 5	-mask –maskExtension 5	- 293,116
MAPPING QUALITY	MQ < 40	- 181,782
FISHER STRAND	FS > 60	- 2,082
QUALITY BY DEPTH	QD < 5	- 36,766
MAPPING QUALITY RANK SUM	MQRankSum < -12.5	- 67
READ POSITION RANK SUM	ReadPosRankSum < -8.0	- 44
TOTAL DEPTH OF COVERAGE	DP < 20	- 457,529
TOTAL DEPTH OF COVERAGE	DP > 83	- 61,027
MISSING PROPORTIONS		- 37,768
REPETITIVE REGIONS & DUPLICATED GENES		- 211,196
DISCARDED SNPS OVERLAP		- 38,033
	Total Removed:	**1,095,086 or 27%**
	Total Remaining:	**2,903,697**

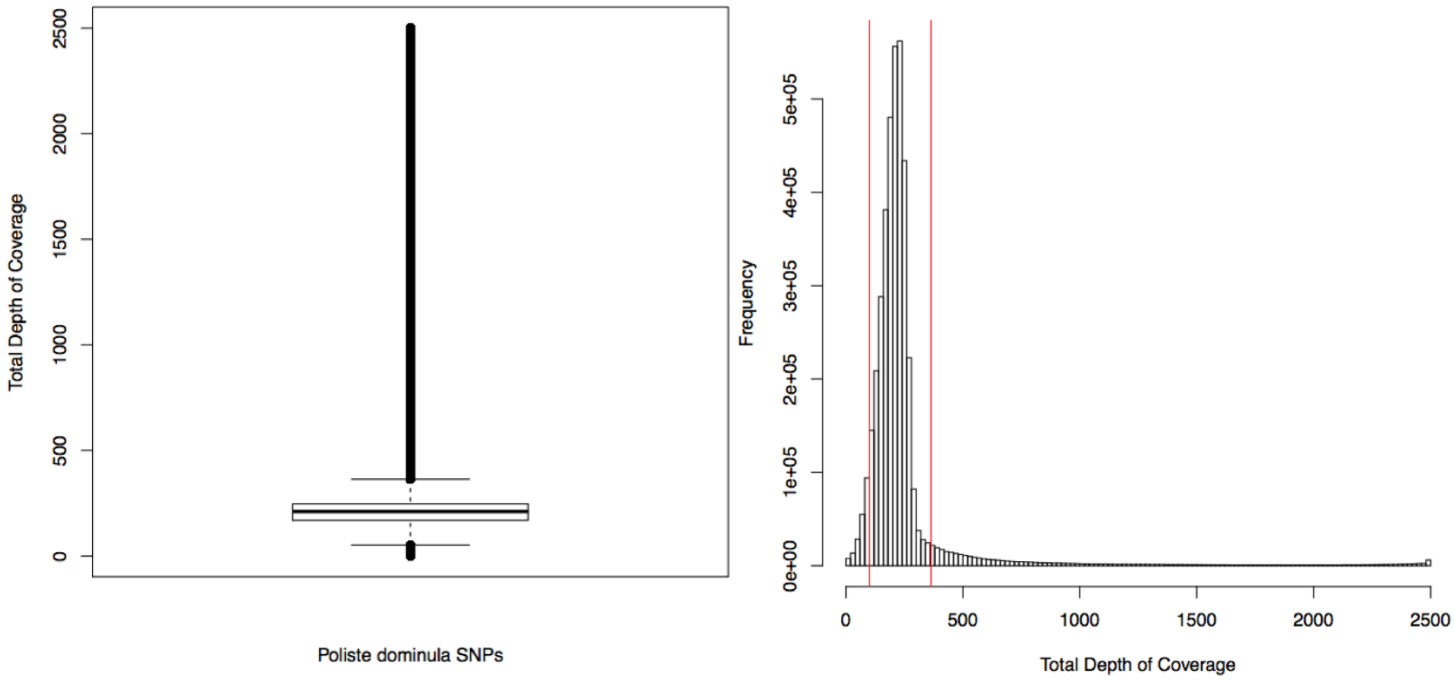

Figure 3: A) Box plot depicting the distribution of total coverage for *Polistes dominula* SNPs. B) Histogram depicting the the distribution of total coverage for *Polistes dominula*. The vertical axis shows the number of alleles corresponding to the total depth of coverage. The red lines indicated a total depth of coverage at 100 and 350 reads, which represents the threshold values chosen for DP.

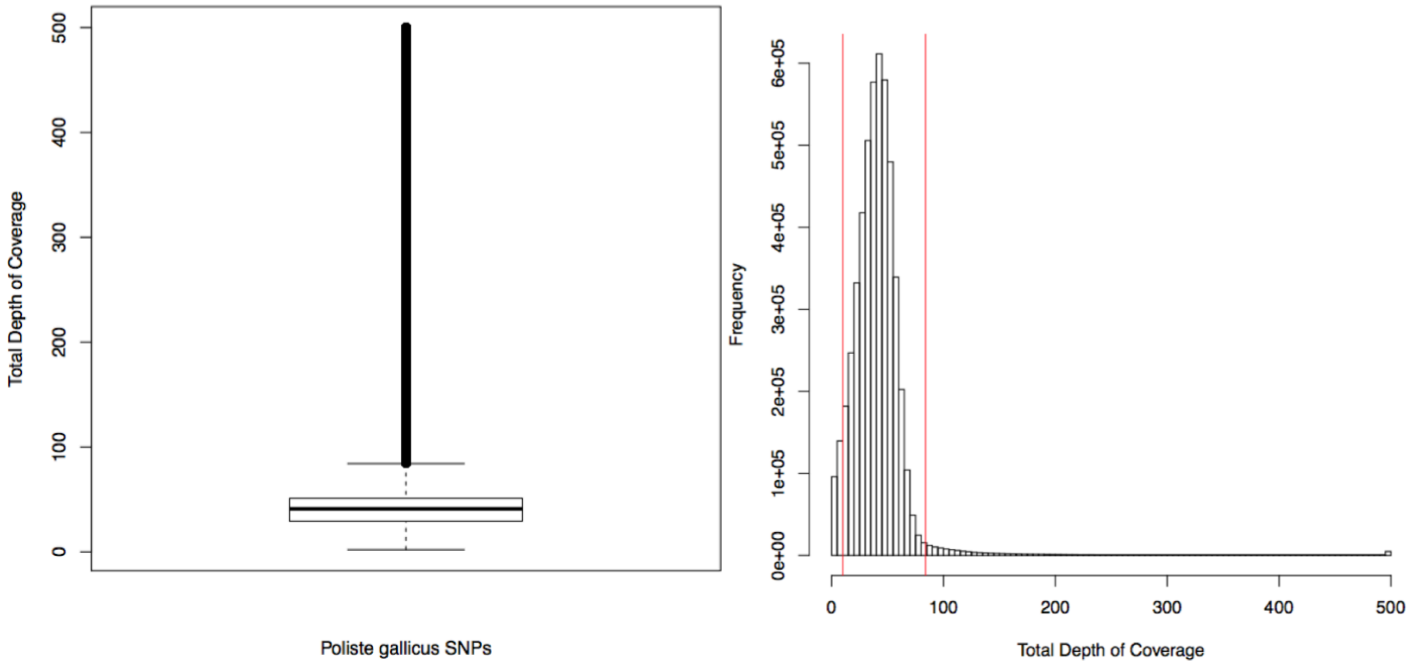

Figure 4: A) Box plot depicting the distribution of total coverage for *Polistes gallicus* SNPs. B) Histogram depicting the the distribution of total coverage for *Polistes gallicus*. The vertical axis shows the number of alleles corresponding to the total depth of coverage. The red lines indicated a total depth of coverage at 20 and 83 reads, which represents the threshold values chosen for DP.

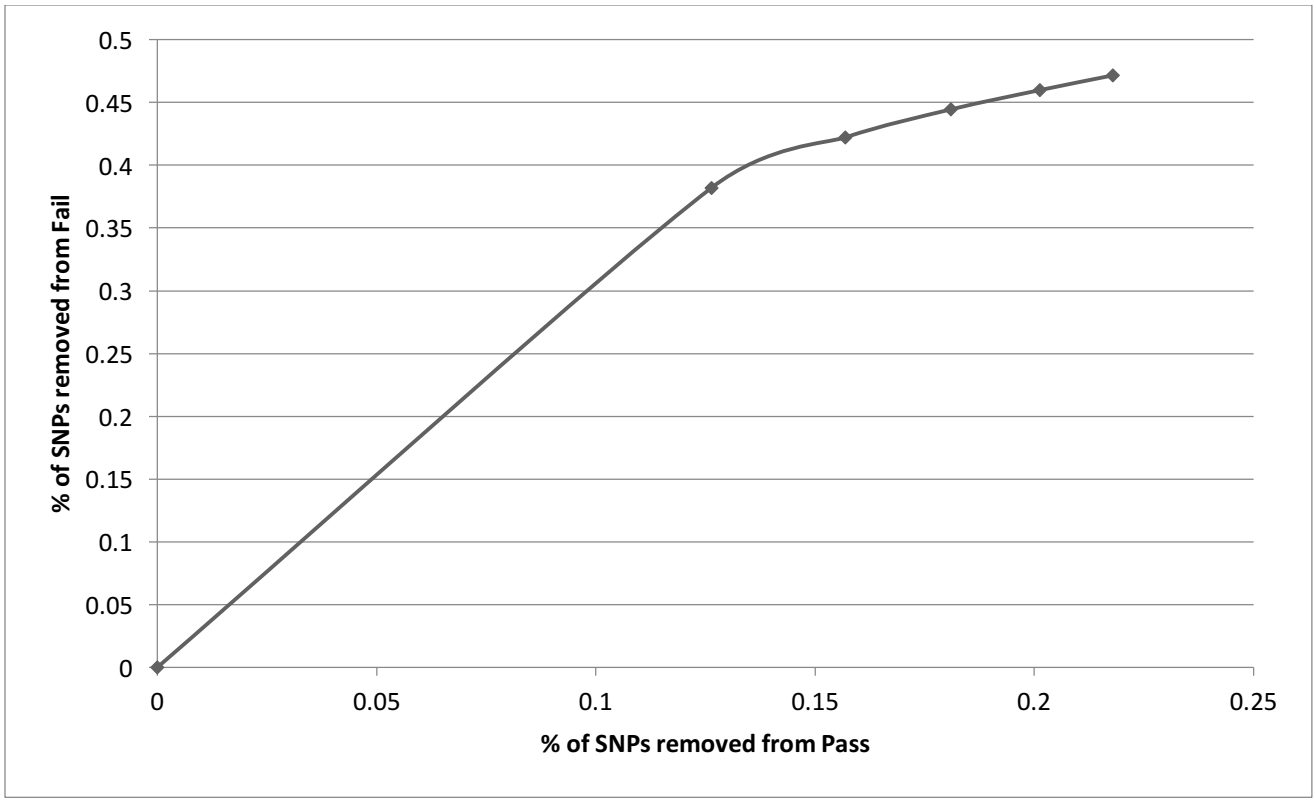

Figure 5: Percentage of SNPs that would be discarded from SNPs that passed filters and SNPs that failed filters with corresponding E-values of $10e^{-72}$, $10e^{-50}$, $10e^{-40}$, $10e^{-30}$, $10e^{-20}$, and $10e^{-10}$.

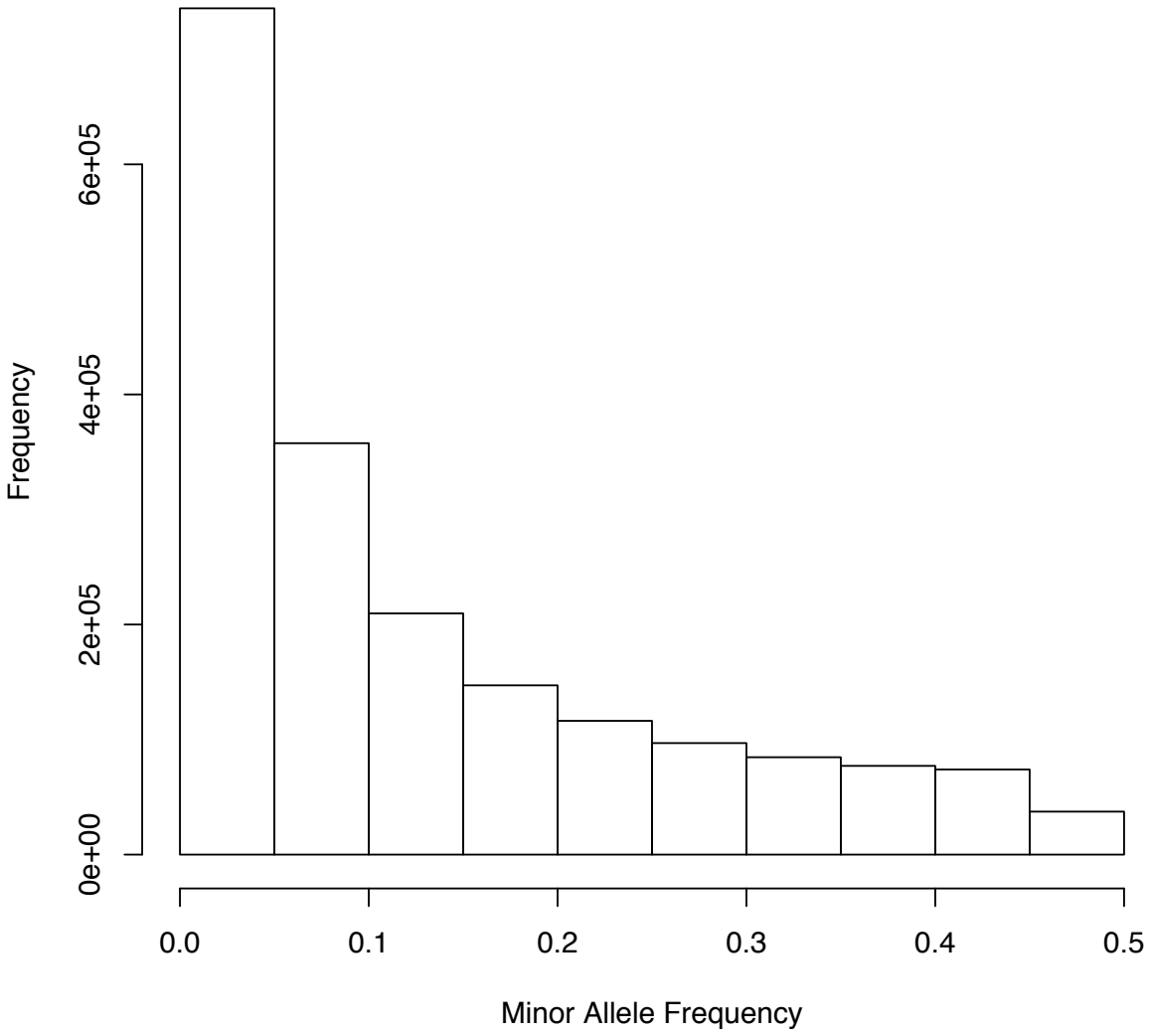

Figure 6: Minor allele frequency distribution of all 1,941,335 *Polistes dominula* pass SNPs. Vertical axis shows the number of alleles in each frequency category.

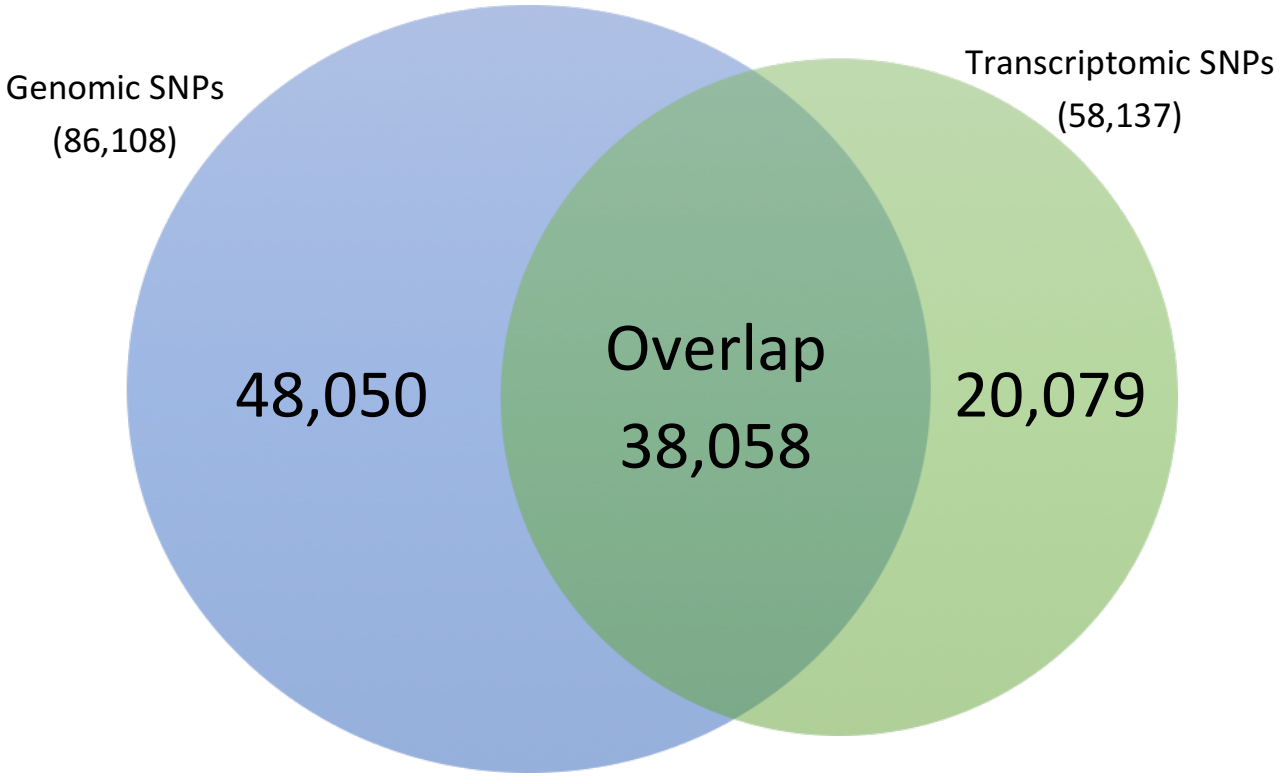

Figure 7: Venn diagram demonstrating the degree of overlap between genomic and transcriptomic SNP datasets of *Polistes dominula*.

Table 4 – Transition/Transverion (Ti/Tv) ratios of categorized SNPs from the *Polistes dominula* and *gallicus* dataset.

CONSEQUENCE	N	*POLISTES DOMINULA*	N	*POLISTES GALLICUS*
WHOLE GENOME RAW SNPS	3,105,907	1.69	3,998,783	1.52
WHOLE GENOME PASS SNPS	1,941,445	1.91	2,903,697	1.62
WHOLE GENOME FAIL SNPS	1,164,572	1.39	1,095,086	1.29
EXOME	86,005	3.80	167,229	2.42
INTERGENIC	1,452,090	1.86	2,098,976	1.60
UPSTEAM	387,933	2.01	644,651	1.64
DOWNSTREAM	366,473	2.01	616,980	1.63
5’ UTR	18,030	2.17	34,605	1.69
3’ UTR	17,130	2.07	34,389	1.64
INTRONIC	344,959	1.82	534,904	1.51
NONSYNONYMOUS	26,324	2.15	63,753	1.41
SYNONYMOUS	59,771	5.25	103,586	3.62

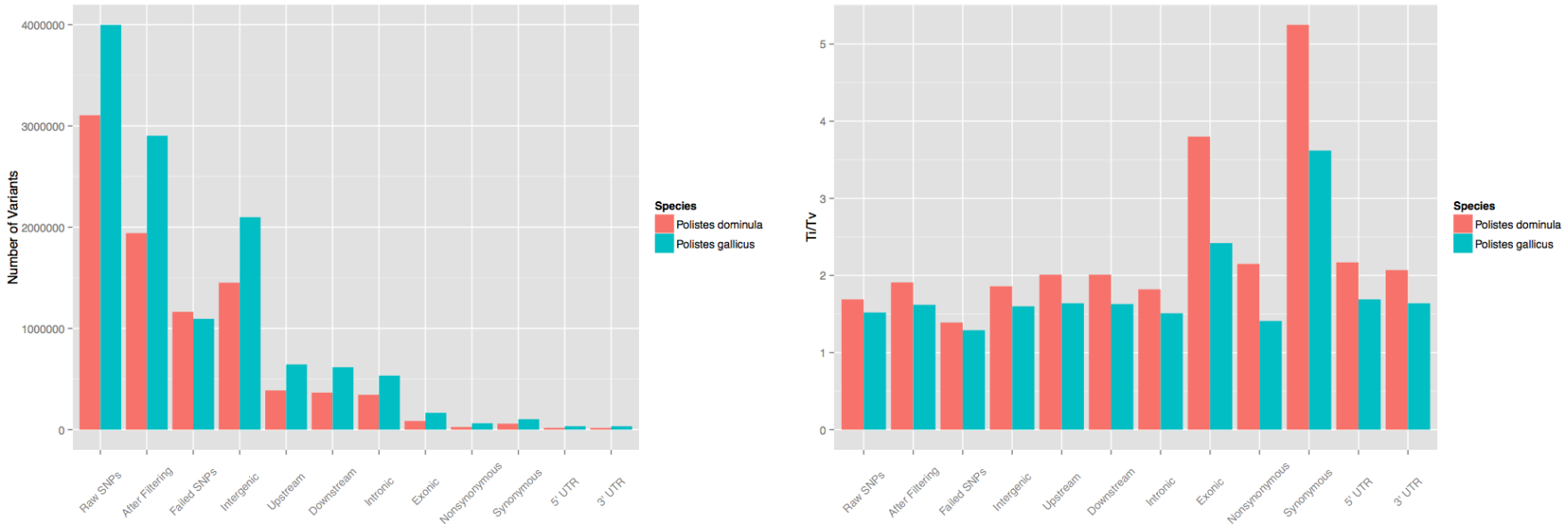

Figure 8: a) Number of variants classified by annotation type for *Polistes dominula* and *gallicus.* B) Transition/Transversion ratio for *Polistes dominula* and *gallicus* by annotation type.

Appendix B: Chapter 3 Tables and Figures

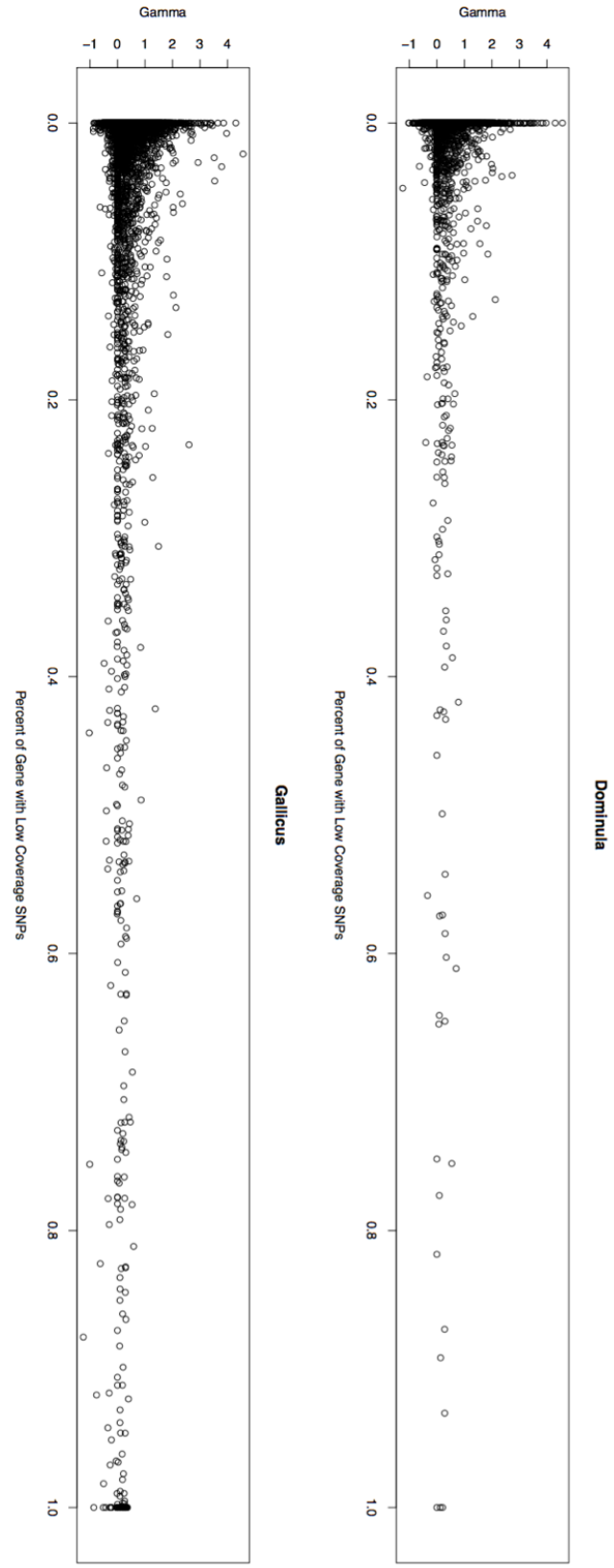

Figure 1: A) Plot of selection coefficient (**γ**) values against the percentage of low coverage for each gene. Graphs show a correlation of low gamma values with poorly mapped genes.

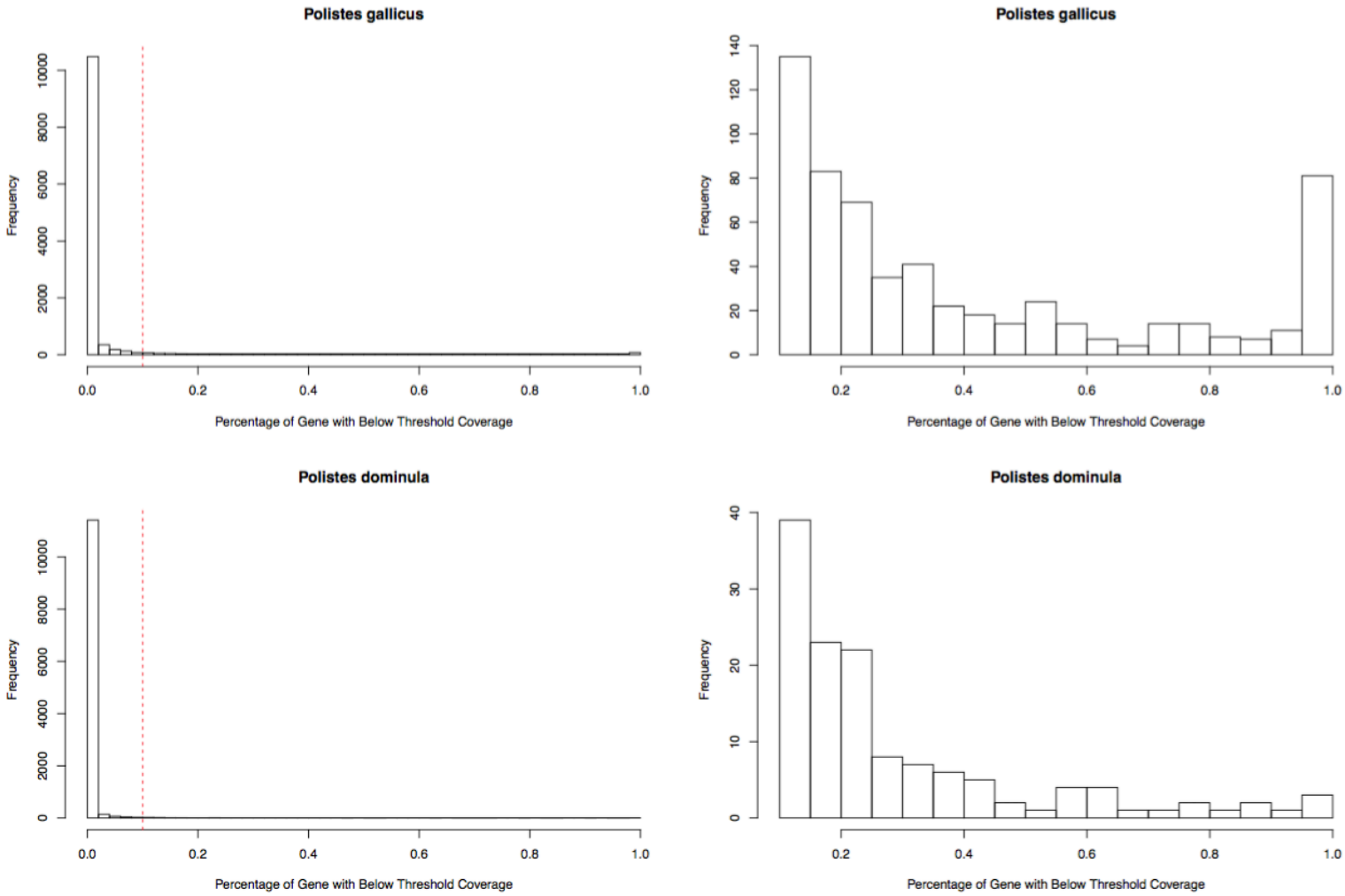

Figure 1: B) Distribution of the percent of low to no coverage of genes for *Polistes dominula* and *gallicus*. Dotted red line depicts the < 0.1 missing and low data threshold.

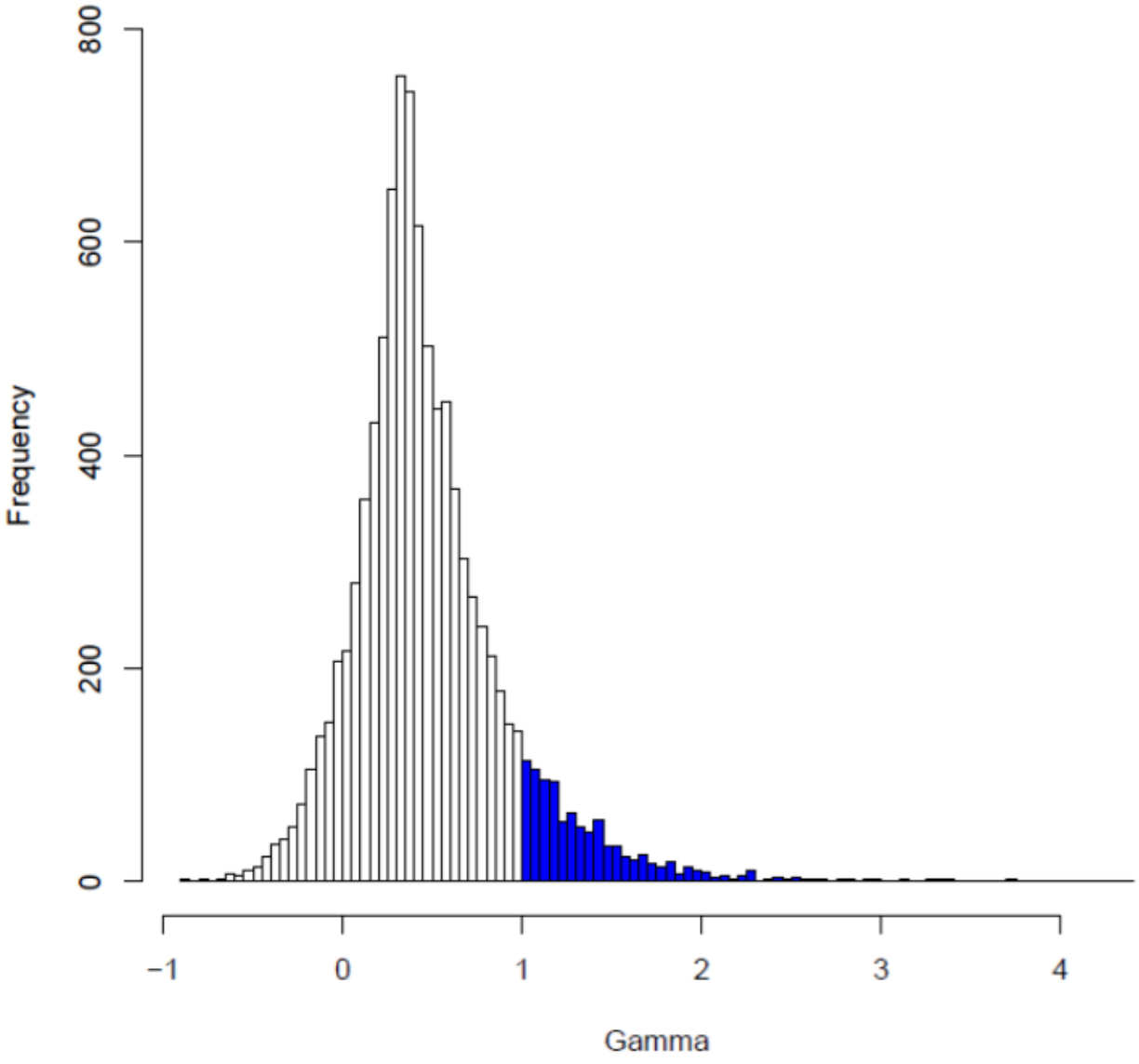

Figure 2: Distribution of the selection coefficient (gamma) across the *Polistes* genome. Segments highlighted in blue indicate the loci under strong positive selection $\boldsymbol{\gamma} > 1$.

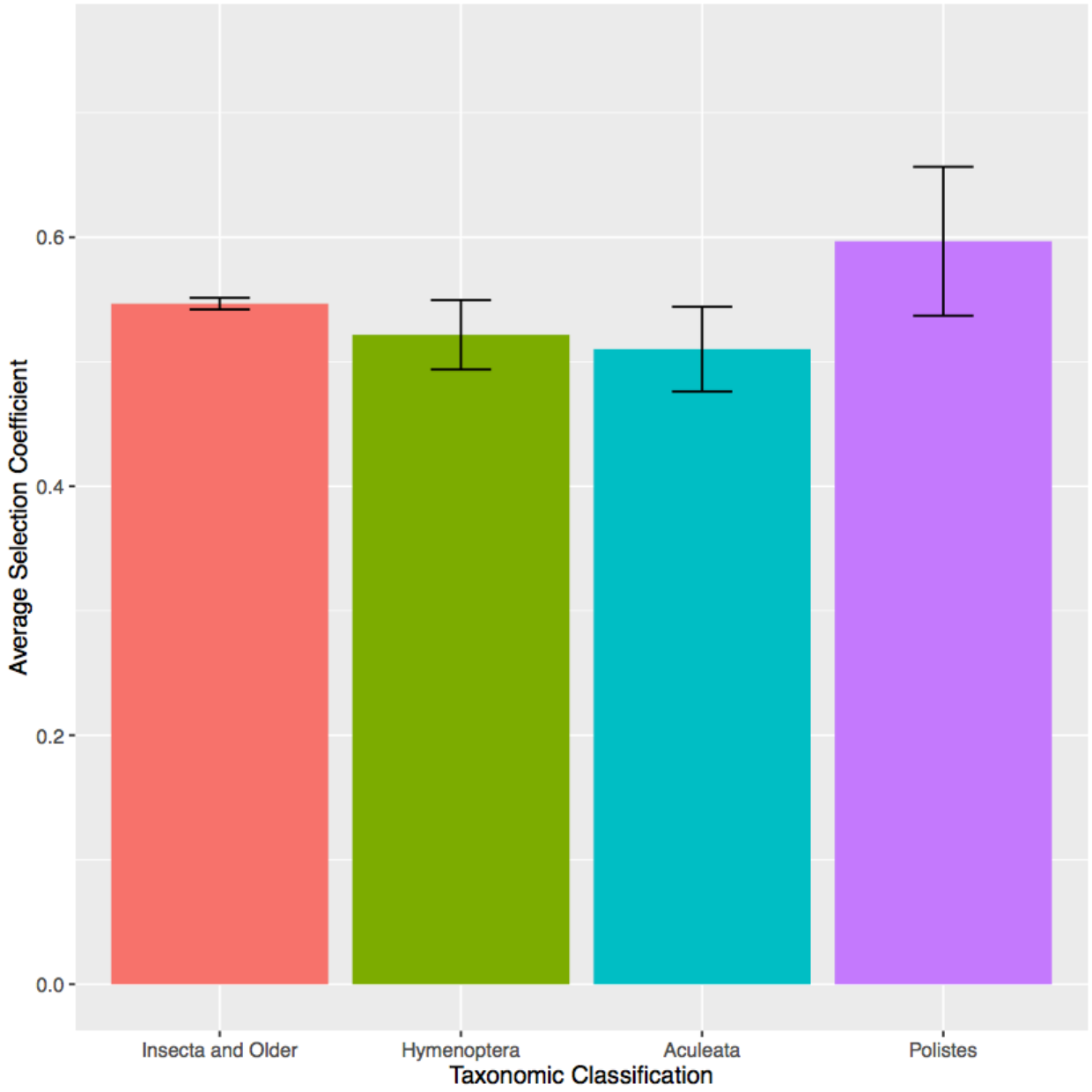

Figure 3: Average selection coefficient (γ) of genes under positive selection (γ > 0) found within the Insecta and older, Hymenoptera, Aculeata, and *Polistes* taxonomic orthologous group. Error bars represent the standard error of the mean.

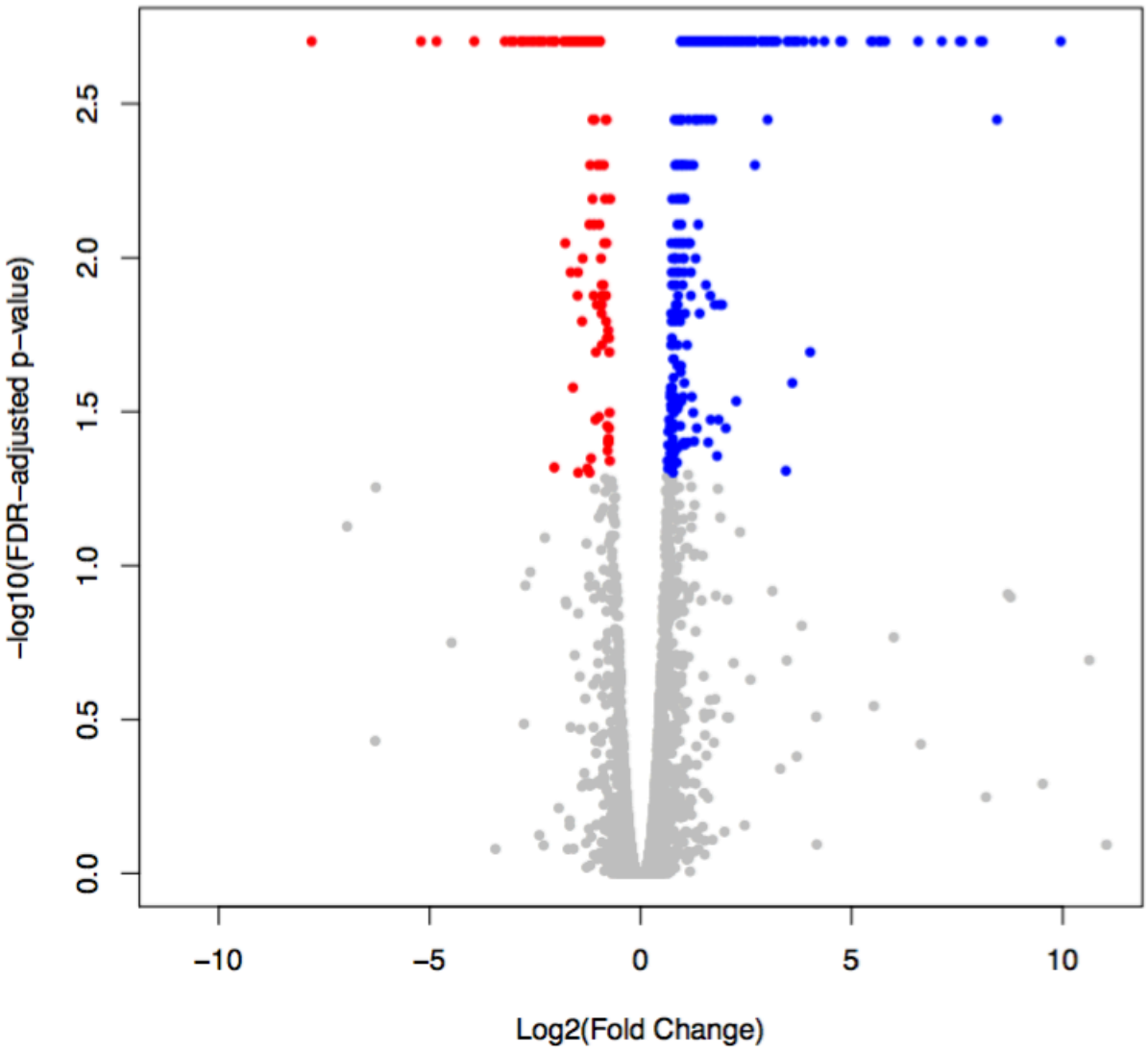

Figure 4: Volcano plot of the differentially expressed genes between *Polistes dominula* queen and worker castes. The data is plotted as the Log2 of the fold change and the –Log10 of the adjusted FDR p-value. Coloured points indicate genes that have a FDR < 0.05. Points coloured in blue are up-regulated genes in workers, while points in red are up-regulated genes in queens.

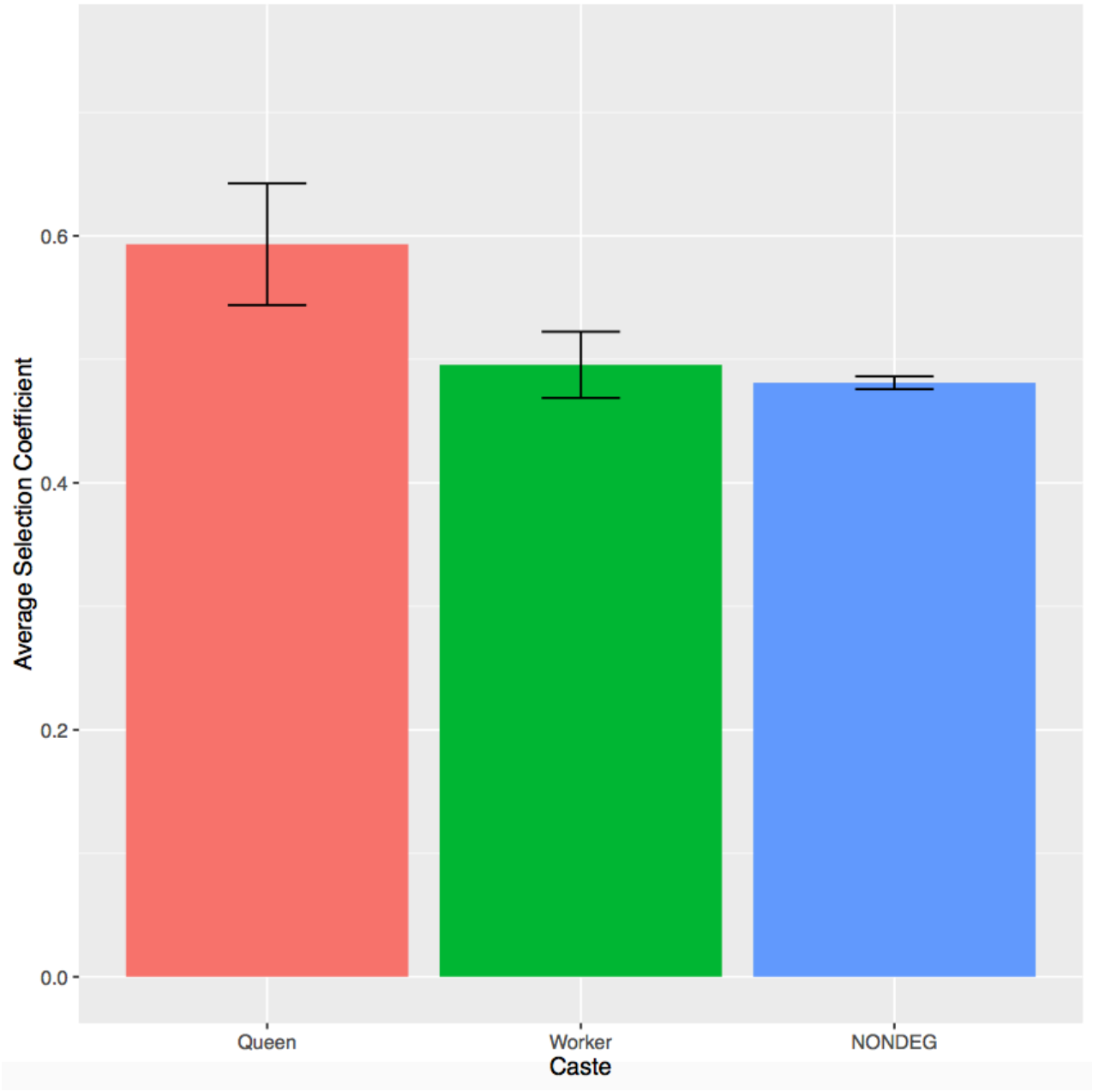

Figure 5: Average selection coefficient (γ) for queen, worker, and non-differentially expressed genes. The error bars represent the standard error of the mean.

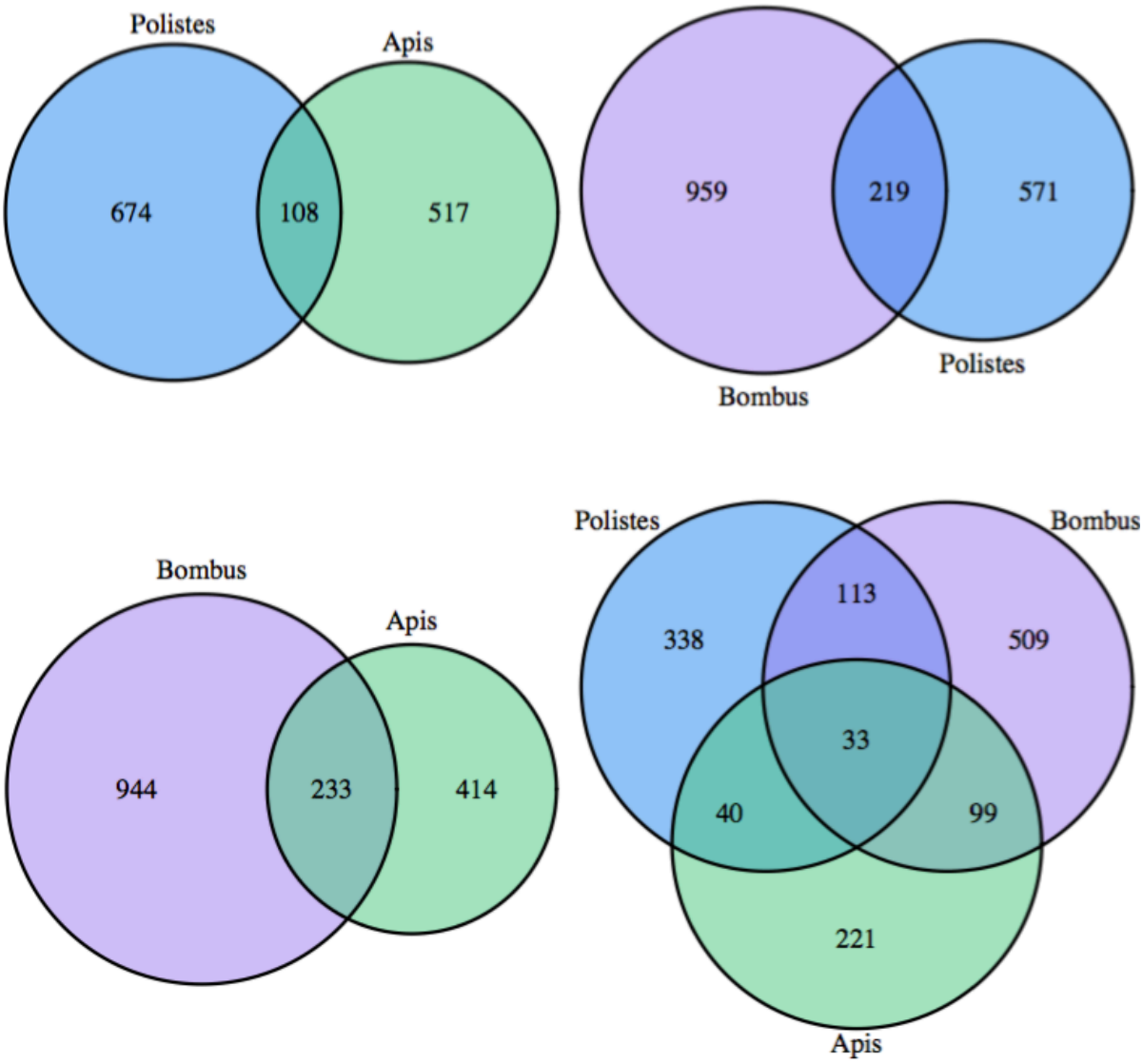

Figure 6: Venn diagram depicting the proportion of overlapping genes in *Polistes*, *Apis*, and *Bombus* that are under strong positive selection ($\gamma > 1$)

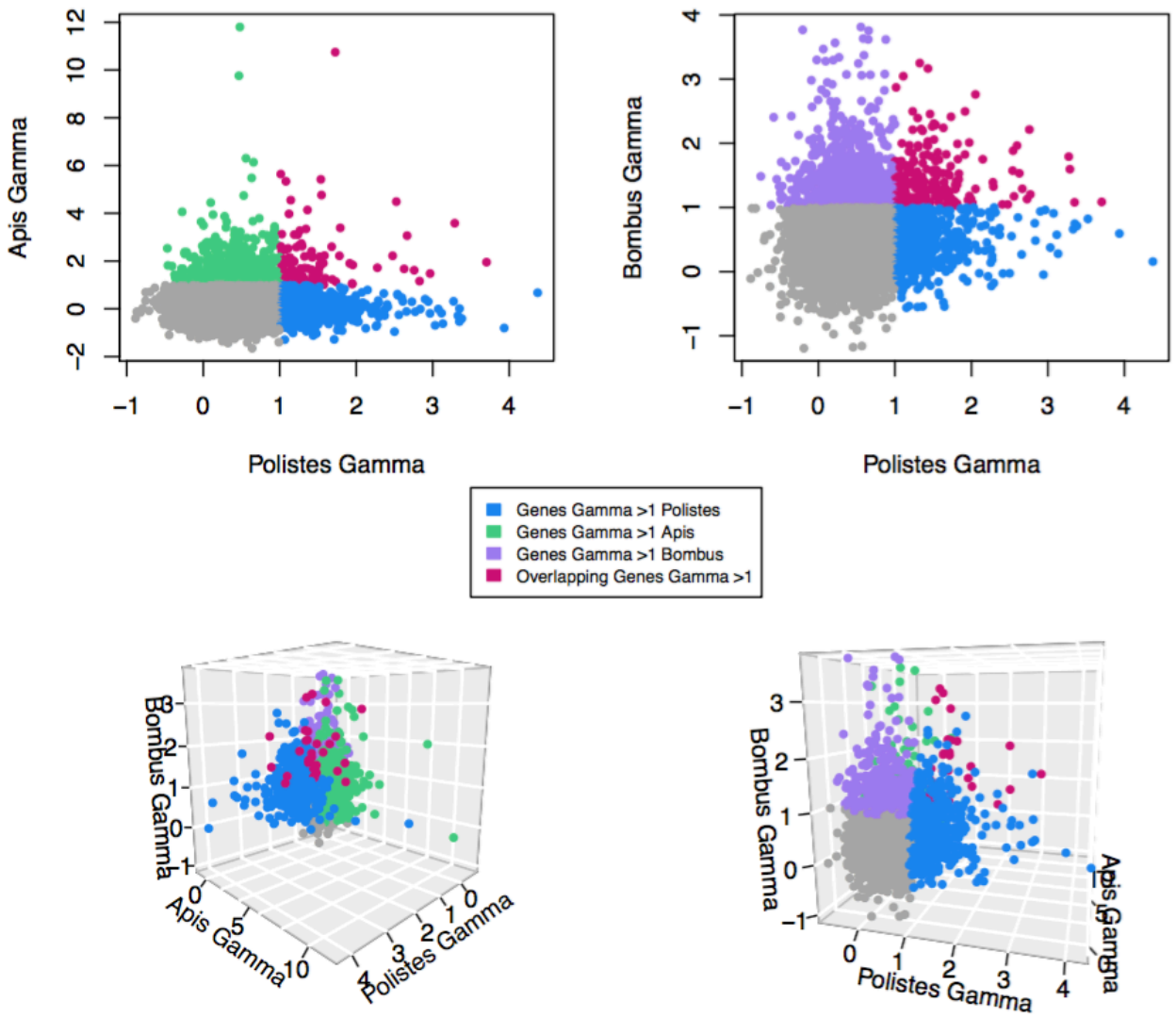

Figure 7: Venn diagrams depicting the proportion of overlapping genes in *Polistes*, *Apis*, and *Bombus* that are under strong positive selection ($\gamma > 1$). Blue data points are *Polistes* genes $\gamma > 1$, green data points are *Apis* genes $\gamma > 1$, and purple data points are *Bombus* genes $\gamma > 1$. The red data points represent overlapping genes between species pairs with $\gamma > 1$.