THE UNIVERSITY *of* EDINBURGH

Comparative genomics and the evolution of immune genes in Drosophila

Pankaj Dhakad

Doctor of Philosophy

INSTITUTE OF ECOLOGY AND EVOLUTION
THE UNIVERSITY OF EDINBURGH

August 2025

To my family and other loved ones,

To anyone who has shown me friendship and kindness during my PhD.

Abstract

Immune genes are among the most dynamic components of animal genomes, shaped by complex trade-offs between functional constraints, pathogen selective pressure, and environment. While *Drosophila melanogaster* has provided a detailed picture of innate immunity in insects, the family Drosophilidae—spanning over 60 million years and diverse ecological niches—offers a unique opportunity to study how immune systems evolve across deep evolutionary timescales. Recent large-scale sequencing efforts have now generated hundreds of high-quality drosophilid genomes, creating unprecedented opportunities for comparative analysis. However, a major limitation has been the lack of consistent gene annotations across species. In this thesis, I address this gap by generating standardized protein-coding gene annotations for 304 drosophilid species. Using a combination of comparative (CAT) and de novo (BRAKER3) annotation, guided by RNAseq and protein homology evidence, I built a consistent, high-quality gene annotation resource. I next investigated the evolutionary rates and turnover of immune-related genes relative to non-immune genes. Using models of DNA sequence evolution and trait evolution, I show that immune genes, particularly effectors and recognition proteins, evolve more rapidly than signaling genes and non-immune genes. I identified gene-level predictors of evolutionary rate, including expression level, relative solvent accessibility, gene length, and protein/genetic interactions. Finally, I used transcriptomic data from pathogen-challenged and unchallenged individuals of three non-model drosophilids to evaluate the bioinformatic recovery of known immune genes, to investigate their conservation and evolution of immune responses and discover novel immune effectors. I identified 20 candidate antimicrobial peptides that are infection-inducible, encode short, secreted proteins, and have no clear homologs in *D. melanogaster*, highlighting extensive lineage-specific evolution of immune effectors.

Together, this work provides a comparative annotation resource for the family Drosophilidae, identifies structural, regulatory, and functional gene features driving evolution of immune genes, and integrates expression data to reveal hidden diversity in immune gene repertoires of non-model drosophilid species. The findings contribute to our understanding of how immune systems diversify and adapt across deep evolutionary timescales.

Lay Summary

Fruit flies are not only familiar insects in kitchens and orchards, but they are also powerful models for science. The fruit fly family (Drosophilidae) contains thousands of species that live in a wide range of environments, from tropical forests to deserts, and feed on very different resources such as fruits, mushrooms, and even plants. These varied lifestyles expose flies to very different microbes and pathogens, making them an excellent group for studying how immune systems adapt over time.

In this thesis, I investigated how immune genes have evolved and diversified across more than 300 species of drosophilid flies. First, I built a large, standardised catalogue of genes for these species, which is the most comprehensive resource of its kind to date. This allowed me to compare immune genes across the entire fly family in a consistent way. I then studied how immune genes evolve, and found that some change very quickly, especially genes that recognise microbes or produce antimicrobial proteins while others remain highly stable over millions of years. These differences suggest that while some components of the immune system are locked in a constant arms-race with pathogens, other components are more resistant to change. Finally, I performed experiments where I infected non-model fly species with a bacteria and measured how their genes responded. I discovered that some immune reactions are shared across species, but others vary greatly, likely reflecting the different ecological pressures these flies face.

Overall, this work provides new insight into how immune systems evolve in nature, showing that immune genes can follow very different evolutionary paths depending on their function and ecological context. It also delivers valuable genetic resources for the scientific community, which will help future studies on fly biology, immunity, and genome evolution.

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Darren Obbard. I have been incredibly lucky to have him as a mentor, always friendly, approachable, and supportive. His enthusiasm for science, thoughtful feedback, and constant availability, whether during a crisis or just to discuss ideas, have made this journey so much easier and more rewarding. I am also grateful to my thesis committee members, Dr. Konrad Lohse and Prof. Paul Sharp, for their guidance throughout my PhD. Their careful reviews of my progress and discussions have helped shape my work in important ways.

I would like to thank the Ashworth (Institute of Ecology and Evolution) for being such a wonderful place to work. The DDH coffee area, and all the events organized by people in the Ashworth building created an environment that kept me motivated. Along the way, I have been fortunate to find amazing friends and colleagues, Hend, Martha, Katie, Julie, Shravan, Neelakshi, Neelima, Oumie, and Dhobasheni, you all made every day more enjoyable. Manas and Adnan, I enjoyed all our debates over lunch. And thanks to all the people who truly became part of my life beyond work.

Special thanks goes to Ankita, my flatmate number one, friend, and bestie, for always being there for me (and for feeding me amazing food). Fawad, my flatmate number two, thanks for being my cooking buddy and the intensity of Ind–Pak cricket matches, which will always stay with me. I am also grateful to Viggi, Anjitha, Mehak, Ardra, Tamoghna, Mukesh, Priyanka and Prachi for their friendship and support.

Finally, none of this would have been possible without my family. Their love, encouragement, and unwavering belief in me have carried me through all the ups and downs of this PhD journey.

Contents

Figures and Tables

Figures

Tables

Chapter 1

General Introduction

This chapter was written with minor comments from Prof. Darren Obbard. It provides an overview of the evolutionary and genomic context relevant to this thesis, with a particular focus on comparative genomics, annotation methods, and *Drosophila* immunity.

The chapter is intended to orient readers to key concepts and tools that underpin the subsequent data chapters. As following data chapters were originally written as standalone manuscripts for publication, there may be minor repetition of background information across chapters, particularly in the introductory sections.

1.1 Drosophilidae: a resource for comparative genomics

The fruit fly, *Drosophila melanogaster*, has been a cornerstone of genetic and evolutionary research for more than a century. From Thomas Hunt Morgan's early work on the chromosomal theory of inheritance to modern applications in animal genetics, developmental, behavioural, evolutionary biology, and human disease, *D. melanogaster* has served as a model system that bridges fundamental biological disciplines (Morgan 1915; Hoffmann 1995; Bergman et al. 2017; Irion and Nusslein-Volhard 2022). *Drosophila*'s appeal lies in its short generation time, ease of laboratory maintenance, established genetic tools, and rich genomic and functional resources. However, the importance of Drosophila as a model extends far beyond *D. melanogaster*, encompassing a diverse family, Drosophilidae, which offers unparalleled opportunities for comparative and evolutionary genomics.

1.1.1 Historical and foundational contributions

The utility of *Drosophila* began with classical genetics. Morgan and colleagues used visible mutations in *D. melanogaster* to map genes to chromosomes and sex-linked inheritance was demonstrated, revolutionizing our understanding of inheritance (Morgan 1910; Morgan 1915). Throughout the 20th century, studies in *Drosophila* provided insights into sex determination (Cline 1983; Salz et al. 1987; Salz et al. 1989), dosage compensation (Muller 1932; Mukherjee and Beermann 1965), gene regulation (Rogina et al. 1998), and the genetic basis of body plans (Illmensee and Mahowald 1974; Nüsslein-volhard and Wieschaus 1980). Discoveries such as the homeobox gene cluster (Antennapedia and Bithorax; Gehring and Hiromi 1986), circadian rhythm gene (*period*; Bargiello and Young 1984; Bargiello et al. 1984; Reddy et al. 1984; Zehring et al. 1984), and genes controlling neurodevelopment (Campos-Ortega 1998) were all pioneered in flies. Over the last century, Drosophila research has been recognised with six Nobel prizes, reflecting its foundational role in advancing genetics, development, and molecular biology.

The sequencing and assembly of the *D. melanogaster* genome in 2000 (Adams et al. 2000) marked the beginning of the post-genomic era for the genus *Drosophila*. This milestone was accomplished through a collaborative effort between Celera Genomics and the Berkeley Drosophila Genome Project (BDGP), utilizing whole genome shotgun sequencing (WGS), a DNA sequencing technique earlier applied to *Caenorhabditis elegans*, the first sequenced metazoan genome, as well as smaller viral and bacterial genomes (Sequencing Consortium 1998). WGS provided insights into the structural and complex nature of the *D. melanogaster* genome, comprising roughly 14,000 protein-coding genes, as well as a wide diversity of non-coding elements, including microRNAs and transposable elements (Misra et al. 2002). It was soon followed by sequencing of the *D. pseudobscura* genome, enabling one of the earliest genome-level comparisons among metazoans (Richards et al. 2005). With an estimated divergence time (at that time) of 25–40 million years from *D. melanogaster* (Russo et al. 1995; Tamura et al. 2004), *D. pseudoobscura* provided a valuable phylogenetic contrast for investigating both coding sequence evolution and the conservation of cis-acting regulatory elements (CREs; Bergman et al. 2002). This comparison revealed a surprisingly high degree of micro-synteny between the two species and helped refine gene annotations while also offering new perspectives on chromosomal rearrangement and structural evolution. However, it quickly became evident that fine mapping of cis-acting regulatory elements would require a broader phylogenetic sampling (Kellis et al. 2003; Richards et al. 2005). This realization motivated the Drosophila 12 Genomes Project, which expanded comparative efforts by sequencing 10 additional species representing a range of ecological niches, life histories, and phylogenetic

distances (Clark et al. 2007; Lin et al. 2007; Stark et al. 2007). Species sampling aimed to, at least in part, capture the extraordinary biodiversity of the *Drosophila* genus and thus allow the inference of general principles underlying speciation, adaptation, and ultimately evolution. The chosen species included close and well studied relatives of *D. melanogaster* and *D. pseudoobscura* within the *Sophophora* radiation, such as *D. simulans*, *D. sechellia*, and *D. persimilis*, as well as more distantly related taxa from the *Drosophila* subgenus, such as *D. mojavensis*, *D. virilis*, and *D. grimshawi* (a strikingly divergent species, with body length around twice that of *D. melanogaster*; see Figure 1.1). Selection criteria included not only phylogenetic distance, but also ecological specialization—for example, *D. sechellia*, which is endemic to the Seychelles and specializes on host plants that are toxic to other species (Louis J. 1986). The project provided a more comprehensive understanding of gene structures, conservation of non-coding RNAs, and a better understanding of chromosomal evolution across lineages. Furthermore, comparative studies also revealed striking differences in transposable element (TE) content among species, ranging from as low as about 2.7% in *D. simulans* to about 25% in *D. ananassae*. These differences, along with lineage-specific losses of TE families like Galileo and 1360, offered compelling evidence for the dynamic turnover and possible horizontal transmission of mobile elements during drosophilid evolution. In subsequent years, large-scale sequencing efforts (Model Organism Encyclopedia of DNA Elements project, modENCODE; Roy et al. 2010) and additional genome sequencing efforts by individual labs have expanded the *Drosophila* genomic landscape even further. The modENCODE project, in particular, generated hundreds of genome-wide datasets for *D. melanogaster*, encompassing information on chromatin organization, epigenetics, transcription factor binding profiles, non-coding RNAs, and regulation of gene expression (Roy et al. 2010; Négre et al. 2011). All these resources and foundational studies have transformed *Drosophila* into one of the best-characterized eukaryotic model systems and laid the foundation for more impactful comparative studies of genome structure, gene regulation, and evolutionary dynamics across the genus.

1.1.2 Diversity and radiation of family Drosophilidae

The family Drosophilidae comprises over 4,000 described species, distributed across every continent except Antarctica (Kim et al. 2024). They occupy a diverse range of ecological niches, ranging from opportunistic feeders on decaying fruits to specialists adapted to sap fluxes, mushrooms, flowers, and cacti—and even more bizarre lifestyles such as inquilines of mining bees, ectoparasites of cecopid nymphs, whitefly predators, or commensalistic behavior with crabs (Ashburner 1981; Markow and O'Grady 2008; Detcharoen et al. 2025). In addition to the temperate environments commonly associated with human-commensal species such as *D.*

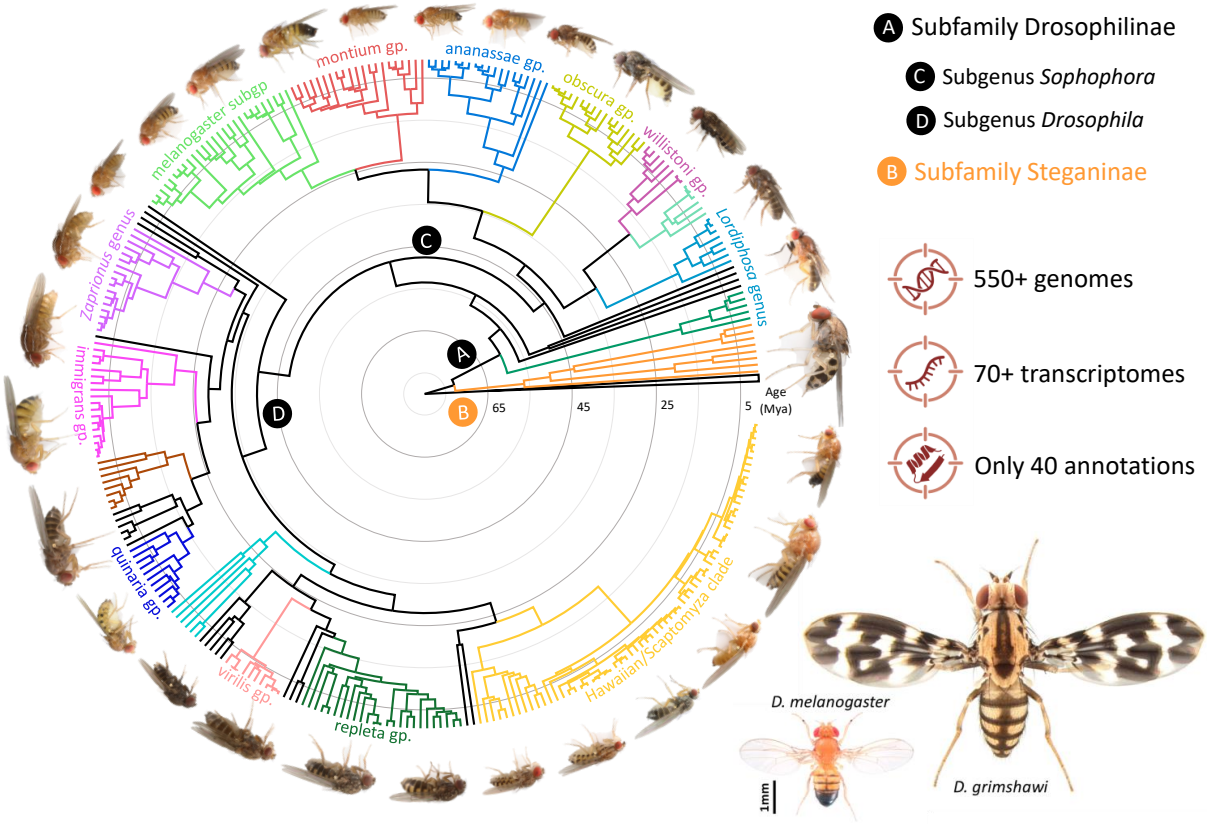

Figure 1.1: Phylogeny and diversity of the family Drosophilidae.
Time calibrated phylogeny of 305 Drosophilidae species spanning ∼72 million years of evolution, reproduced from Dhakad et al. (2025b). The tree highlights major radiations across the family, including representatives from both subfamilies, Drosophilinae and Steganinae. Fly images around the tree illustrate species from key lineages, showcasing dramatic variation in body size and morphology. The comparative image of *D. grimshawi* and *D. melanogaster* (mrca ∼40 million years ago) illustrates size variation within the family. Fly images around the tree were provided by Prof. Darren Obbard. The comparative images are reproduced from FlyBase and are © Nicolas Gompel under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

melanogaster, drosophilids are found in high-altitude montane regions, arid deserts, and humid tropical rainforests (Izumitani et al. 2016). In addition, high reproductive rate, short generation time, and rapid adaptability have facilitated drosophilids to expand beyond their ancestral habitats (Markow and O'Grady 2008). These life-history traits, coupled with ecological opportunity, have promoted not only continental diversification but also rapid adaptive radiations in island systems. The Hawaiian picture-wing clade is a striking example, having undergone spectacular adaptive radiations involving host shifts, mating behavior diversification, and morphological innovation (Bryan et al. 1988; Edwards et al. 2007; Magnacca and Price 2015).

According to a fossil calibrated phylogeny, Drosophilidae began diversifying approximately 60-70 million years ago, shortly after the Cretaceous–Paleogene boundary (Suvorov et al. 2022). The family includes two subfamilies: Drosophilinae, which contains the majority of species including *D. melanogaster*, and Steganinae, a smaller group with distinct morphological and ecological characteristics. Within Drosophilinae, further subdivision includes large genera such as *Drosophila*, *Scaptodrosophila*, *Hirtodrosophila*, and *Zaprionus*, among others (Figure 1.1). The genus *Drosophila* is typically divided into *Sophophora* and *Drosophila* subgenera (Desalle 1992; Van Der Linde et al. 2010; Yassin 2013). Within the subgenus *Drosophila*, two major lineages have been recognized: the virilis-repleta radiation (virilis section) and the immigrans–tripunctata radiation (quinaria section). To address the paraphyly of the group, recent taxonomic revisions have suggested the elevation of the virilis–repleta clade to a separate subgenus, *Siphlodora*, while retaining the immigrans–tripunctata group in the revised subgenus *Drosophila* (Yassin 2013). The latter now includes 21 species groups (Yassin 2013), 15 of which belong to the immigrans–tripunctata clade (Van Der Linde et al. 2008), with broad biogeographic distributions across the Old and New Worlds (Izumitani et al. 2016). Parallel patterns of geographic disjunction between Old and New World lineages are observed across multiple drosophilid radiations, including the *Scaptodrosophila*, *Hirtodrosophila*, *Sophophora*, *Siphlodora*, and immigrans–tripunctata radiation (Izumitani et al. 2016).

The breadth of ecological diversity within Drosophilidae has fostered repeated and independent evolution of many traits, creating natural replicates for testing evolutionary hypotheses. Traits such as pigmentation (Massey and Wittkopp 2016), courtship behaviour (Dai et al. 2008; Ahmed et al. 2019), host specialization (Dworkin and Jones 2009; Rondón et al. 2022), and immune responses (Sackton et al. 2007; Hanson et al. 2023), offers an exceptional framework for studying the genetic basis of convergent evolution. This allows researchers to test whether the same genes or pathways are reused during adaptation, or whether different molecular mechanisms underlie similar phenotypic outcomes. Furthermore, closely related species often differ markedly in phenotype and ecology despite minimal genome-wide divergence, allowing

the use of comparative genomics to identify the specific genes or regulatory elements involved in adaptive divergence. These features make drosophilids uniquely suited to dissect the interplay between genotype, phenotype, and environment across evolutionary timescales. In this way, the family provides a uniquely powerful framework for integrating comparative genomics, ecology, and evolutionary theory across timescales—from recent ecological divergence to deep phylogenetic splits.

1.1.3 Genomic resources in Drosophilidae

The remarkable ecological radiation of Drosophilidae, combined with deep phylogenetic sampling and growing genomic resources, positions the family as a premier model for comparative evolutionary genomics. As of July 2025, there are over 550 species genome sequenced (of which >360 are publicly available and $\sim$200 are in process of release)—many at the chromosome-level. This rich dataset allows comparative investigation of gene family evolution, genome organization, and adaptation at both macro- and microevolutionary scales (Kim et al. 2021; Kim et al. 2024; Darwin Tree of Life; Darwin Tree of Life Project 2022). A significant proportion of high-quality assemblies come from the recent work of Bernard Kim and colleagues, who have generated genome sequences for over 200 Drosophilidae species (Kim et al. 2024). These genomes were predominantly assembled using a hybrid approach that combines Oxford Nanopore Technologies (ONT) long reads and Illumina short reads (Espinosa et al. 2024). Most genomes were sequenced using ONT R9.4.1 (87 species) or the more accurate R10.4.1 flow cells (80 species), often with supplementary Illumina polishing. This single-fly sequencing strategy has proven highly effective, regularly yielding assemblies with >1 Mb contig N50, >98% BUSCO completeness, and genome-wide consensus accuracies exceeding QV40 for R10.4.1 genomes (Kim et al. 2024). Another major source of high-contiguity assemblies is the Darwin Tree of Life (DToL) project, which aims to generate reference-quality genomes for all eukaryotic species in the United Kingdom. As part of this initiative, so far 9 drosophilid species have been sequenced using PacBio HiFi technology, which provides highly accurate long reads with low error rates. The resulting assemblies are typically scaffolded with high-throughput chromatin conformation capture (Hi-C), yielding chromosome-level assemblies. DToL genomes are particularly valuable for studying genome organization, transposable elements, and synteny due to their high structural resolution. In addition, genome sequencing efforts by individual research labs contributed to the dataset (e.g., Vicoso and Bachtrog 2015; Mahajan et al. 2018; Bracewell et al. 2019; Ellison and Bachtrog 2019a; Ellison and Bachtrog 2019b; Bronski et al. 2020; Conner et al. 2021; Suvorov et al. 2022). Together, these efforts have produced a diverse collection of genome assemblies spanning most major drosophilid

lineages. Contiguity metrics vary depending on sequencing technology and assembly pipeline, but many recent assemblies achieve contig N50 values >1 Mb and scaffold N50 values approaching chromosomal scale. BUSCO completeness for these assemblies is typically >95%, and in many cases exceeds 98%, reflecting near-complete coverage of the expected gene space.

In addition to genome assemblies, transcriptomic data are increasingly available for Drosophilidae. Public RNAseq datasets span a wide range of species, tissues, developmental stages, and environmental conditions—including infection challenge experiments, stress responses, and ecological adaptations. These data are essential for annotating protein-coding genes, identifying alternative isoforms, and studying the evolution of gene expression and regulation. Several studies, including those focused on immune response, behaviour, and development, have contributed high-quality transcriptomes for both model and non-model species (Troha et al. 2018; Yang et al. 2018; Nozawa et al. 2021; Li et al. 2022; Church et al. 2023).

However, despite the growing availability of genome and transcriptome data, comprehensive and standardized gene annotations remain sparse. Only a fraction of species with genome assemblies have publicly available annotations, and among those, many rely on automated pipelines without cross-species consistency. RefSeq provides curated annotations for several species, but mainly from the melanogaster species group. Only a few species distant from the melanogaster species group have annotations and these are annotated by GenBank and other repositories, which lacks high-quality, consistent gene models (Prieto-Banos et al. 2025). Because of the long processing times for genome annotations by RefSeq, GenBank and Ensembl, often several years, labs that assemble genomes usually prefer also to annotate them. This lack of uniform annotations impedes comparative analyses of gene structure, orthology, and evolutionary rate.

1.2 Protein-coding gene annotation

Gene annotation is the process of identifying gene structures within a genome, including the definition of coding sequences, intron-exon boundaries, and alternative splice variants. The annotation methods can be categorised into: *ab initio* prediction (intron-exon structure prediction using statistical models) and sequence-alignment based approaches, which aligns RNAseq, ESTs (Expressed Sequence Tag), cDNA (complementary DNA) and protein sequences onto genome assemblies to predict transcripts. Most annotation pipelines combine both sources of transcript to generate final gene sets (Freedman and Sackton 2024). Because

gene annotations provide the foundation for nearly all functional and evolutionary analyses, the quality and consistency of gene annotations is paramount. Particularly in comparative genomics, annotation quality directly affects all downstream analyses, including orthology inference, gene family reconstruction, molecular evolution studies, and functional genomics (Wu et al. 2013; Prieto-Banos et al. 2025). Despite the increasing ease of genome assembly through advances in long-read sequencing technologies, gene annotation remains a major challenge, especially in non-model species, (Freedman and Sackton 2024).

1.2.1 Factors affecting protein-coding gene annotation

Protein-coding genes in eukaryotic genomes are typically organized into multi-exonic structures beginning with a 5′ untranslated region (5′-UTR) in the first exon and ending with a 3′-UTR in the terminal exon. The protein-coding portion of the gene (coding sequence, CDS) lies between canonical translation start and stop codons and is often interrupted by introns, which can account for the majority of transcribed sequence. Intron lengths in insect genomes vary widely, but a significant proportion are relatively short, while others can be quite long, even reaching tens of kilobases (Presgraves 2006). For example, in *D. melanogaster*, over a third of introns are between 60 and 70 nucleotides long, with a tail of longer introns extending into the kilobase range (Pai et al. 2017). Exon lengths, on the other hand, are somewhat less variable (Deutsch and Long 1999; Suetsugu et al. 2013). A major source of complexity in genomes arises from the heterogeneity in exon-intron structure, where exons are interspersed between introns of variable length, and genes may occur densely packed, overlapping, or interleaved with antisense transcripts. Further complications come from exceptional cases, including dicistronic and polycistronic transcripts, noncanonical splice sites, trans-splicing, alternative translation initiation, and stop-codon readthroughs, as well as rare cases such as ribosomal frameshifting or HAC1-type intron processing (Matthews et al. 2015). Each of these features challenges the accuracy of proteins generated by automated pipelines.

The presence of transposable elements (TEs) in the genome adds another layer of difficulty. These repetitive, mobile DNA sequences can occupy large portions of the genome, ranging from $\sim$2.7% to $\sim$25% of total genome size in *Drosophila* (Mérel et al. 2020). During genome assembly, TE-rich regions are particularly prone to collapse or misassembly, leading to fragmented gene models or entirely missing loci, especially in regions with low-coverage (Ou et al. 2020). Moreover, many TEs (with the exception of Short Interspersed Nuclear Elements, SINEs) can carry protein-coding regions or pseudogenes that can closely resemble bona fide

host genes. These TE-derived sequences are not functional host genes (but see Lippman et al. 2004; Mallet et al. 2004), yet they are often mistakenly annotated as such, inflating gene counts and complicating orthology inference. Filtering out these false-positive gene predictions remains a key challenge for producing reliable, comparable annotations across species.

These structural challenges are compounded by the dynamic evolutionary processes that shape gene content. Insects, like other metazoans, undergo frequent gene duplication and loss, which generates gene families that can vary in size and composition across lineages. Insect genomes typically encode 10,000 to 20,000 protein-coding genes, and many of these belong to gene families that have undergone lineage-specific expansions (Waterhouse 2015). The evolutionary fate of duplicated genes can vary, they may acquire new functions (neo-functionalization), partition existing functions (sub-functionalization), or become pseudogenes through non-functionalization. These processes contribute to the rapid turnover of gene families observed in many insect clades, particularly in genes involved in chemosensation, immunity, and reproduction (Hahn et al. 2007; Rondón et al. 2022). Annotation pipelines must therefore contend with short exons, long and variable introns, densely packed or overlapping gene structures, pervasive TEs, and the evolutionary fluidity of gene families (Ejigu and Jung 2020). Without careful integration of transcriptomic and homology-based evidence, annotation tools risk fragmenting genes, misidentifying pseudogenes, or omitting lineage-specific genes altogether. In the following sections, I review the main categories of annotation methods—including *ab initio* prediction, evidence-based approaches, and comparative annotation and evaluate their performance and limitations in the context of comparative genomics.

1.2.2 Overview of protein-coding gene annotation methods

The core algorithms used to predict protein-coding genes in the eukaryotic genomes have remained conceptually stable for over two decades. Most genome annotation pipelines continue to rely on statistical models such as generalized hidden Markov models (GHMMs), which were first implemented in widely used tools such as GeneMark (Besemer and Borodovsky 2005; Lomsadze et al. 2005), and Augustus (Stanke et al. 2004; Stanke et al. 2006). These *ab initio* methods are trained on known gene structures and predict coding regions based solely on the genomic sequence. They can identify gene features including intron–exon boundaries and coding sequence (CDS) regions by recognizing sequence composition, splice motifs, and codon usage patterns. The predictive power of *ab initio* models, however, can be significantly enhanced by incorporating external evidence (“hints”). These hints are typically obtained from aligning transcriptomic (RNAseq, ESTs) or proteomic data (or known proteins from related species) to the target genome (Gabriel et al. 2024). This approach improves gene

model accuracy by constraining predictions to biologically supported regions and improving the placement of splice junctions, start and stop codons, and UTRs. Two of the most widely used annotation programs are MAKER (Cantarel et al. 2008) and BRAKER (Gabriel et al. 2024). These pipelines integrate both *ab initio* predictors (AUGUSTUS and/or GeneMark-ES/ET) with external evidence, and allow for iterative refinement. However, the use of RNAseq data alone in these pipelines has limitations: genes with low or tissue-specific expression may be missed entirely, and it is more effective when derived from very closely related species (preferably same species or individuals). Protein evidence, in contrast, is more conserved across evolutionary distances, making it especially useful for annotating genomes of non-model species with no transcriptomic data. However, protein alignments alone contain no information at the UTRs or isoform level, and may miss lineage-specific genes or recent duplications. Therefore, combining transcriptomic and protein evidence increases sensitivity and specificity, identifying both conserved genes and lineage- or species-specific genes (Freedman and Sackton 2024).

Comparative annotation methods

The recent explosion in genome sequencing across Drosophilidae has created both new opportunities and practical challenges for genome annotation. While over 550 drosophilid species now have genome assemblies available, only approximately 1/3rd have transcriptomic data. This disparity highlights a major limitation of evidence-based annotations that rely on species-specific RNAseq; for most newly sequenced species, such data simply do not exist. Many genes and isoforms are expressed only in specific tissues, at specific development stages, or in response to a environment condition. Obtaining such data for non-model species to fully annotate their genome requires access to fresh or flash-frozen tissue samples, which are not always readily available. Transcript projection methods bypass this by making use of rich resources in well-studied organisms. These methods can be combined with extrinsic "hints" from either the target genome or related genomes to inform gene models.

One of the earliest strategies for transcript projection used pairwise genome alignment, where a well-annotated reference genome is aligned to a target genome to infer orthologous gene structures. This method works best for closely related species, where synteny and sequence conservation enable high-confidence mapping. Early tools like Projector implemented this strategy by aligning exons from a reference genome to a target using splice-aware alignment (Meyer and Durbin 2004). A more recent and widely used tool, Liftoff, refines this approach by incorporating both sequence similarity and local synteny, allowing it to handle complex structural variation and even partial gene mapping (Shumate and Salzberg 2021). These tools are especially valuable in clades, where close relatives of a species with a reference annotation

are well represented, and annotation projection can then provide a strong first-pass gene set. However, pairwise approaches are inherently limited by their one-to-one structure. They only leverage the relationship between two genomes at a time and do not exploit the evolutionary context provided by multiple species. As a result, they may miss lineage-specific novel genes, struggle with tandemly duplicated regions, or fail to distinguish conserved coding elements from conserved non-coding elements (König et al. 2018). These limitations become more pronounced when annotating genomes that are more distantly related to the reference, or when the reference itself has gaps or annotation errors.

Transition to multiple species alignment and comparative gene prediction

To overcome these challenges, comparative gene prediction approaches have shifted toward using multiple genome alignments, which provide broader evolutionary context and allow for the identification of conserved gene structures across an entire clade (König et al. 2018). Insects, and Drosophilidae in particular, offer ideal systems for such methods due to the availability of large numbers of genome assemblies (Kim et al. 2024). One foundational tool in this field is transMap, which projects reference transcript models from one genome onto one or more other genomes (Stanke et al. 2008). However, in cases where gene structures have diverged substantially or where alignment quality varies, additional refinement is needed to convert these projections into high-confidence annotations. This refinement can be provided by AugustusCGP, an extension of the *ab initio* gene prediction tool AUGUSTUS that supports comparative gene prediction across a set of aligned genomes (Konig et al. 2016). AugustusCGP integrates cross-species sequence conservation, synteny, and when available, species-specific evidence such as RNAseq or protein hints. The method favours gene structures that are conserved across the alignment, but also allows for true structural differences, such as exon loss, gene duplication, or alternative splicing. By doing so, it avoids overfitting to the reference genome and can identify novel genes or lineage-specific structural variants. These tools are integrated into the Comparative Annotation Toolkit (CAT; Fiddes et al. 2018), a framework for clade wide annotation that combines transMap projections with AugustusCGP refinements. CAT uses a multiple genome alignments produced by tools such as progressiveCactus (Armstrong et al. 2020), and a combination of transcript projections and *de novo* refinement to annotate hundreds of genomes simultaneously, producing gene sets that are structurally consistent and orthology-aware.

Simultaneous multiple genome annotation has several distinct advantages over *ab initio* or pairwise genome annotations methods. First, it improves consistency across species, which is essential for comparative studies. Second, it reduces false positives by anchoring predictions in both sequence conservation and known gene models. Third, it increases sensitivity to lineage- or clade-specific events (clade wide gene gain or loss, horizontal transfer events), identifying gene that might be missed by other methods (König et al. 2018; Armstrong et al. 2019). For example, annotations generated using a comparative method AugustusCGP have been shown to substantially improve exon-level consistency between closely related species like *D. melanogaster* and *D. simulans*. In one study, over 89% of *D. simulans* exons predicted using comparative annotation matched orthologous exons in *D. melanogaster*, compared to ∼81% using single-species *ab initio* prediction (König et al. 2018). CAT has also been successfully applied to comparatively annotate 13 mammalian genomes using the mouse (mm10) GENCODE VM15 as the reference annotation, recovering over 97% of BUSCO genes in all genomes. Even in distantly related species such as humans (most recent common ancestor; mrca ∼90 mya), CAT generated transcripts showed high concordance with well curated reference (>91% CAT introns and >75% of CAT protein coding isoforms exactly match GENCODE annotation), demonstrating its robustness across large phylogenetic distances (Fiddes et al. 2018).

1.3 Innate immunity in *Drosophila*

Like all animals, insects live in environments full of diverse microorganisms, ranging from benign symbionts to lethal pathogens. In species like *Drosophila*, which lay eggs and develop in fermenting or decaying plants, fungi, and fruits, microbial contact is unavoidable. During feeding, microbes enter the digestive tract and may colonize the gut epithelium. Some of these microorganisms establish themselves as part of the commensal flora, while others trigger immune responses if they breach epithelial barriers or are recognized as harmful (Gupta and Nair 2020). Systemic infections may also arise through septic injuries—such as parasitic nematode infections, parasitoid wasps depositing their eggs into fly larvae and wounding by mites (Subasi et al. 2024)—introducing microbes directly into the haemocoel (the general body cavity). The immune system of *Drosophila*, like that of all organisms except vertebrates, is entirely innate. It relies on evolutionarily conserved and rapidly inducible mechanisms to detect and eliminate pathogens. The epithelial surfaces of the body acts as first-line of defence against microbes. The epithelial cells lining the tracts of digestive (the gut), genital system, tracheal (respiratory) system and the Malpighian tubules, all produce antimicrobial peptides (AMPs) to inhibit microbial growth (Ferrandon et al. 1998; Tzou et al. 2000). Microbes that

succeed in entering the haemocoel are countered by cellular and humoral immune responses, each involving activation of multiple specialized pathways (Hoffmann and Reichhart 2002; Yu et al. 2022). These systemic responses are tailored to the nature of the invading microbe—be it bacterial, fungal, viral, or eukaryotic organisms (trypanosomatids and parasitoid wasps) and together ensure robust defence.

1.3.1 Cellular immunity

Cellular defence to counter systemic infection involves the wide-ranging action of haemocytes, which are considered analogous to blood cells of vertebrates. These cells originate from the haematopoietic organ known as the lymph gland, which contains progenitor cells that are released into the haemolymph during larval development upon dispersal of the gland (Crozatier and Meister 2007). There are three main types of differentiated haemocytes in *D. melanogaster*, (i) plasmatocytes (macrophage/monocyte-like cells), which are involved in phagocytosis of apoptotic cells and invading microbes; (ii) crystal cells, which are required for melanization reaction; and (iii) lamellocytes, large flat cells that encapsulate parasitoid wasp eggs and other foreign bodies too large to be phagocytosed. In the following sections, I will summarize the cell mediated immunity in *Drosophila* (Figure 1.2).

Phagocytosis

Phagocytosis is one of the most powerful and rapid ways to eliminate microbial invaders. This process is primarily mediated by plasmatocytes, the main phagocytic blood cell type in *Drosophila* (95% of hemocytes; Yu et al. 2022). Pathogen recognition is achieved through receptor-ligand interactions and opsonization. In *Drosophila*, the thioester-containing protein (TEP) family acts as opsonins, promoting the recognition and engulfment of microbes by plasmatocytes (Haller et al. 2018). Molecules on the surface of various particles, including bacterial peptidoglycans, lipopolysaccharides (LPSs), fungal β-1,3-glucans, and phosphatidylserine ("eat-me" signal on surface of apoptotic cells) can be recognised by phagocytic cells. Apoptotic cells are recognised by evolutionary conserved receptors Croquemort (CRO, CD36 homologue), Draper, Six-microns-under (Simu), and Santa-maria (Franc et al. 1996; Manaka et al. 2004; Hilu-Dadia et al. 2025). Other phagocytic receptors include scavenger receptor family Peste and dSR-C1 (Rämet et al. 2001; Philips et al. 2005), members of the Nimrod C-type family such as NimC1 and Eater (Kocks et al. 2005; Kurucz et al. 2007), immunoglobulin superfamily protein Dscam (Watson et al. 2005), and integrin subunits like Integrin αPS3/βv (Nonaka et al. 2013), which all are essential for pathogen recognition and internalization. Following internalization, engulfed particles are enclosed within phagosomes, which undergo a

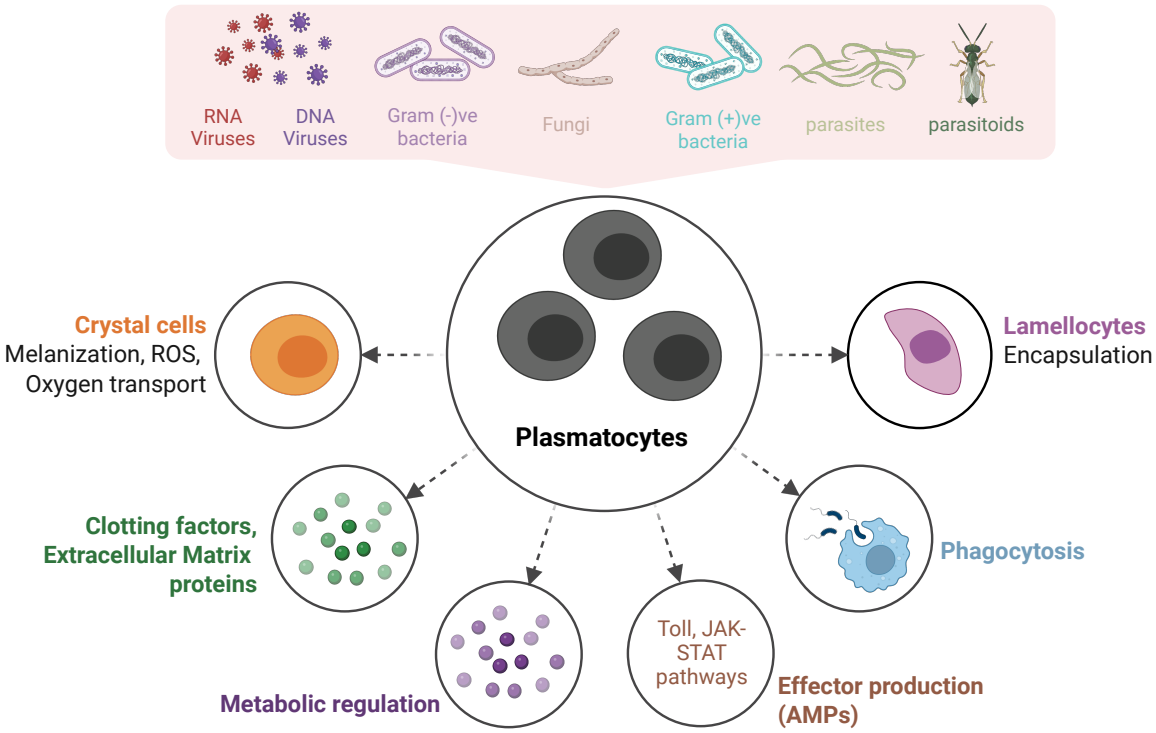

Figure 1.2: Cellular immune responses in *Drosophila*
Plasmatocytes, the predominant haemocyte type, perform diverse functions including phagocytosis of microbes and apoptotic cells, secretion of clotting factors, extracellular matrix proteins, antimicrobial peptides (AMPs), and cytokines that regulate signalling pathways such as Toll and JAK-STAT. They can transdifferentiate into crystal cells or lamellocytes. Crystal cells are involved in melanization and reactive oxygen species (ROS) production, while lamellocytes are specialized for encapsulation of large pathogens such as parasitoid wasp eggs. Haemocyte-derived signals also contribute to systemic immune regulation, metabolic homeostasis, and tissue repair. The illustration summarizes the roles of each haemocyte type in response to bacterial, fungal, and viral infections, as well as parasitoid attack. The figure was created with BioRender.com.

maturation process involving sequential fusion with early and late endosomes. This endows the phagosome with antimicrobial properties through acidification and the acquisition of enzymes. Ultimately, fusion with lysosomes forms phagolysosomes, where hydrolytic enzymes degrade the internalized material, completing the clearance of pathogens or apoptotic cells (Melcarne et al. 2019).

Melanization

Melanization is one of the earliest and most acute immune responses in insects, rapidly activated upon pathogen entry through the cuticle or septic injury. This reaction, visible as darkening at wound or infection sites, is involved in blood coagulation, wound healing, and pathogen encapsulation. Melanin is produced through the oxidation of phenols to quinones, catalysed by phenoloxidase (PO), which is activated by proteolytic cleavage of its precursor, prophenoloxidase (PPO; Tang et al. 2006). In *Drosophila*, PPO is released into the haemolymph by the rupture of crystal cells, a process triggered by pathogen recognition and tissue damage via serine protease (SP) cascades—linking melanization to both humoral and cellular immunity. Crystal cell rupture in *Drosophila* has recently been likened to pyroptosis-like, a caspase-dependent, inflammatory form of programmed cell death reminiscent of mammalian pyroptosis (Dziedziech and Theopold 2022). This process is driven by JNK signalling and reactive oxygen species (ROS), ultimately leading to cell swelling and PPO release (Bidla et al. 2007; Myers et al. 2018).

Three PPO genes are present in the *Drosophila*: PPO1 and PPO2, expressed in crystal cells, and PPO3, expressed in lamellocytes. Uniquely, PPO3 can trigger spontaneous melanization without infection, suggesting intrinsic enzymatic activity independent of SP cleavage. PPO3 is also essential in mutant backgrounds with increased lamellocyte population, such as hop$^{tum\text{-}l}$, implicating it in lamellocyte mediated melanization. Several SPs have been identified as PPO activators: Hayan (activates PPO1 and PPO2), MP1 (required against bacterial and fungi infections), and MP2 (also called Sp7 or PAE1), which specifically activates PPO1 during fungal infection (Castillejo-López and Häcker 2005; Nam et al. 2012). These pathways show specificity, Hayan is important for wound-induced melanization, while MP2 plays a prominent role in pathogen-induced melanization, particularly against *Staphylococcus aureus* and fungi (Dudzic et al. 2019). Because excessive melanization is detrimental to host survival, *Drosophila* employs serine protease inhibitors (serpins) to regulate the reaction. Spn27A, Spn28D, and Spn77Ba inhibit melanization under homeostatic conditions (De Gregorio et al. 2002;

Scherfer et al. 2008; Tang et al. 2008). Upon infection, Spn27A is depleted, allowing PO activation. Spn77Ba inhibits melanization in the trachea by suppressing MP1 and MP2 activity. Remarkably, the parasitoid wasp *Leptopilina boulardi* secretes its own serpin (LbSPNy) to block melanization and evade host immunity (Colinet et al. 2009).

Encapsulation

Encapsulation is a specialized immune response in *Drosophila* larvae, particularly against parasitoid wasp eggs, which are too large to be eliminated by phagocytosis. Instead, the immune system mounts a cellular response led by lamellocytes that surround and isolate the wasp eggs. This response is triggered by activation of key immune pathways, notably Toll and JAK-STAT, which stimulate lamellocyte production both from progenitor differentiation in the lymph gland and from the transdifferentiation of circulating plasmatocytes (Anderl et al. 2016; Banerjee et al. 2019). Upon infection, plasmatocytes are rapidly recruited to the wasp egg surface, where they initiate a response and transdifferentiate into lamellocytes (Russo et al. 1996; Anderl et al. 2016). These lamellocytes then form multilayered capsules around the egg, physically sequestering it. Integrin-β localization to the lamellocyte membrane is crucial for proper capsule formation (Irving et al. 2005; Xavier and Williams 2011). The capsule is subsequently melanized, a process that involves both lamellocytes and crystal cells, contributing to the elimination of the parasite. Recognition of the wasp egg is not yet fully understood, although C-type lectins like DL2 and DL3 have been implicated in similar encapsulation responses (Ao et al. 2007). Additionally, a distinct form of encapsulation, phagocytic encapsulation has been described, wherein enlarged plasmatocytes engulf clusters of pathogens. This process is activated by p38 MAPK signalling and can significantly improve host survival in the later stages of infection (Shinzawa et al. 2009).

1.3.2 Humoral immunity

A defining feature of the humoral immune response in *Drosophila* is the systemic production of antimicrobial peptides (AMPs). The expression of AMPs is controlled by NF-κB signalling pathways, which are activated upon recognition of pathogen-associated molecular patterns (PAMPs) by host pattern recognition receptors (PRRs). Once triggered, these pathways orchestrate a coordinated immune response, primarily mediated by the fat body (analogue of the vertebrate liver) and supported by circulating haemocytes. This response includes discharge of

massive array of AMPs and other effectors into the haemolymph, where they inhibit microbial growth and contribute to systemic metabolic regulation. In the following sections, I summarize the two major NF-κB signalling pathways that regulate humoral immunity in Drosophila: Toll and Imd (Figure 1.3).

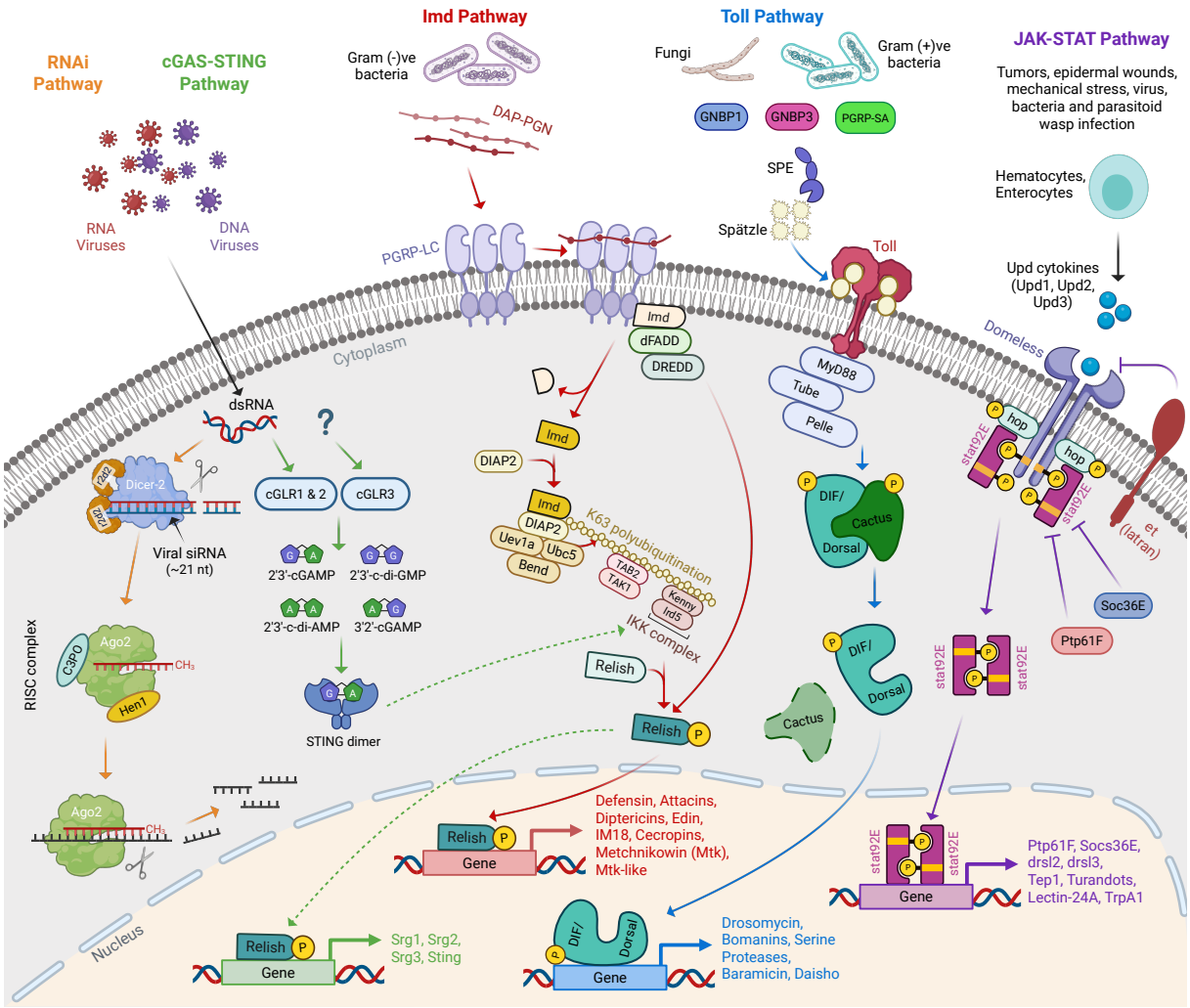

Figure 1.3: Schematic representation of innate immune signalling pathways in *Drosophila*. This diagram illustrates five major immune signalling pathways, Toll (cyan), Imd (red), RNA interference (orange), cGAS–STING (green), and JAK/STAT (purple)—each shown with color coded arrows. Arrows indicate activation and blunt-ended bars (“T”) indicate inhibition. Dashed arrows represent crosstalk between pathways. The figure is created with BioRender.com.

The Toll pathway

In *Drosophila*, the Toll pathway plays a central role in orchestrating the humoral immune response, particularly against Gram-positive bacteria and fungi. Unlike mammalian Toll-like receptors, which directly bind microbial patterns, *Drosophila* Toll is activated indirectly through a host-derived ligand, the Nerve Growth Factor-like cytokine Spätzle (Spz; Weber et al. 2003). Upon infection, microbial components such as lysine-type peptidoglycan (PGNs) from Gram-positive bacteria or β-glucans from fungi are detected by pattern recognition receptors such as PGRP-SA, PGRP-SD (recognises Gram-negative bacteria; Leone et al. 2008), GNBP1, and GNBP3 (Michel et al. 2001; Gobert et al. 2003; Pili-Floury et al. 2004; Gottar et al. 2006). These recognition events trigger a serine protease cascade in the haemolymph, involving serine proteases (SPs), including modSP and grass (Kambris et al. 2006; Buchon et al. 2009). This cascade culminates in the cleavage and activation of pro-Spätzle into its mature, dimeric form, by Spz-processing enzyme (SPE; Jang et al. 2006). The Spz dimers then binds to one Toll receptor molecule. This triggers the conformational changes and leads to dimerisation of Toll receptors and subsequently activation of downstream signalling. Intracellular signalling cascade in the Toll pathways initiates by interaction of C-terminus of Toll dimer with adaptor protein MyD88, which forms complex with Tube, and the kinase Pelle (Moncrieffe et al. 2008). This leads to phosphorylation and subsequent degradation of Cactus, an IκB-like inhibitor that normally retains the NF-κB transcription factors Dorsal and Dif in the cytoplasm (Lemaitre et al. 1996; Wu and Anderson 1998). With Cactus degraded, Dorsal and Dif translocate to the nucleus, where they drive the transcription of immune effector genes, particularly antimicrobial peptides (AMPs) such as Drosomycin, Bomanins, and Baramicins, which are especially active against fungal and Gram-positive infections (Figure 1.3). Beyond AMP induction, Toll signalling promotes haemocyte proliferation, lamellocyte differentiation, dispersion of lymph gland and effective encapsulation of parasitoid wasp eggs (Ligoxygakis et al. 2002; Dudzic et al. 2019). Toll activation also intersects with melanization pathways and wound repair, with serine proteases like Hayan contributing to both Toll signalling and PPO1 activation (Dudzic et al. 2019).

The Imd pathway

The Imd (immune deficiency) pathway is primarily activated by diaminopimelic acid (DAP)-type PGNs on Gram-negative bacteria and a few Gram-positive bacteria (such as *Bacillus spp.*, *Listeria innocua*). Recognition is mediated by pattern recognition receptors (PGRPs), briefly the transmembrane receptor PGRP-LC, extracellular receptor PGRP-SD and the cytosolic receptor PGRP-LE (Leulier et al. 2003; Kaneko et al. 2004; Kaneko and Silverman 2005;

Iatsenko et al. 2016). The short, secreted form of PGRP-LE facilitates extracellular PGN binding and delivery to PGRP-LC, while the full-length cytoplasmic form detects monomeric PGN fragments (or TCTs) that enter the cell and can independently activate the pathway (Takehana et al. 2002; Kaneko et al. 2006; Lim et al. 2006; Neyen et al. 2012). PGN recognition initiates a receptor-proximal signalling complex involving Imd, Fadd, and the caspase Dredd (Georgel et al. 2001; Leulier et al. 2002). Activated Dredd cleaves Imd, triggering its K63 ubiquitination via the E3 ligase Diap2, which recruits Tak1 and Tab2, leading to phosphorylation of the IKK complex (Ird5 and Kenny; Silverman et al. 2000; Kleino et al. 2005). This ultimately results in phosphorylation and cleavage of Relish, a NF-κB transcription factor (Silverman et al. 2000; Stöven et al. 2003). The active Rel fragment translocates to the nucleus to induce antimicrobial peptide (AMP) gene expression, such as Defensin, Diptericin, Attacin, Metchnikowin and Cecropin (Figure 1.3). Imd signalling also plays a role in gut homeostasis—maintaining a basal level of immune activation in response to commensal microbiota. This basal activity is essential for pathogen resistance but is tightly regulated by negative feedback via amidase PGRPs (PGRP-LB and PGRP-SC1/2) and transcriptional repressors like Caudal, which prevent excessive immune activation and maintains microbial balance in the gut (Ryu et al. 2008; Kleino and Silverman 2014).

1.3.3 Antiviral immunity

RNAi antiviral pathway

RNA interference (RNAi) is a major antiviral mechanism in *Drosophila*. Unlike the previously described Toll and Imd pathways, RNAi acts through sequence-specific gene silencing, targeting viral RNAs for degradation. There are three RNAi pathways, the microRNA pathway (miRNA), which is responsible for endogenous gene expression, the piwi-interacting RNA pathway (piRNA), which protects germline genomes from transposable elements, and the small-interfering RNA pathway (siRNA), which protects against viruses and transposable elements. The siRNA is a cell-autonomous mechanism (but see Saleh et al. 2006; Saleh et al. 2009; Tassetto et al. 2017), activated directly by the presence of viral double-stranded RNA (dsRNA), a replication intermediate of RNA viruses and produced by DNA viruses through various mechanisms, such as convergent transcription (where two strands of DNA are transcribed in opposite directions). Dicer-2 (Dcr-2), an RNase III family enzyme, recognizes and cleaves long dsRNA molecules into $\sim$21 nucleotide siRNAs (Aliyari and Ding 2009). These virus-derived

siRNAs are then loaded into the Argonaute-2 (Ago2) protein, the core component of the RNA-induced silencing complex (RISC). The RISC complex uses the guide strand of the siRNA to base-pair with complementary viral RNA transcripts, leading to their endonucleolytic cleavage and degradation (Ding 2010).

cGAS-STING antiviral pathway

The cyclic GMP-AMP (cGAMP) synthase-stimulator of interferon genes (cGAS–STING) pathway, long recognized as a key antiviral response in vertebrates, has an evolutionarily conserved counterpart in *Drosophila*, where it plays a pivotal role in defence against both RNA and DNA viruses (reviewed in Cai et al. 2022). In contrast to the RNAi pathway, which directly targets viral replication, the STING pathway triggers NF-κB-dependent transcriptional responses upon sensing foreign nucleic acids. In *D. melanogaster*, this pathway is initiated by two cGAS-like receptors (cGLR1 and cGLR2), which detect cytosolic viral dsRNA. Upon recognition, these cGLRs activates cyclic dinucleotides (CDNs) such as 2′3′-cGAMP, 3′2′-cGAMP, 2′3′-c-di-AMP, and 2′3′-c-diGMP (Holleufer et al. 2021; Slavik et al. 2021; Cai et al. 2023). These CDNs activate STING, a ER membrane-associated adaptor protein that activates signalling leading to Relish dependent transcription of antiviral genes (Martin et al. 2018). Recent work also suggests that STING-mediated responses extend beyond transcriptional regulation. For example, during Zika virus infection, STING promotes the expression of autophagy related genes (*Atg8-II*), contributing to viral restriction in neural tissues (Liu et al. 2018). This autophagy induction appears to be Relish-dependent, suggesting crosstalk between STING signalling, NF-κB activation, and cellular stress pathways (Liu et al. 2018; Hu et al. 2024). Although the molecular details of STING pathway activation and effector recruitment in insects are still emerging, current evidence places this pathway as a central component of *Drosophila*'s antiviral defence, acting in parallel to the RNAi pathway.

1.3.4 Auxiliary immune signalling pathways

In addition to the canonical Toll, Imd, and antiviral pathways, a range of other evolutionarily conserved signalling cascades modulate immune responses in *Drosophila*. These include the JAK-STAT pathway, which regulates stress response, hemocyte proliferation and lamellocyte differentiation in response to parasitoid wasp infection and tissue damage. In addition, stress-activated kinase cascades such as the Jun N-terminal kinase (JNK) and p38 MAPK pathways

mediate immune responses to wounding and oxidative stress, and intersect with Toll and Imd signalling (Yu et al. 2022). Although these pathways are not solely immune-specific, they provide essential regulatory inputs that shape the magnitude and context of the immune response.

JAK-STAT pathway

The JAK-STAT pathway in *Drosophila* is a highly conserved signalling cascade involved in immunity, stress responses, and developmental processes. Unlike the Toll and Imd pathways, which are primarily triggered by microbial pattern recognition, JAK-STAT mainly activated by cytokines. It responds to tumours, epidermal wounds, mechanical stress, parasitoid wasp infection. The pathway is initiated by the secretion of Unpaired (Upd) family cytokines (Upd1, Upd2, and Upd3; Hombría et al. 2005; Wright et al. 2011; Myllymäki and Rämet 2014), which bind to the transmembrane receptor Domeless (Dome; Brown et al. 2001). This interaction activates the Janus kinase of *Drosophila*, Hopscotch (Hop), which phosphorylates the intracellular domain of Dome and creates docking sites for the STAT92E transcription factor. Once recruited, STAT92E is phosphorylated, dimerizes, and translocates into the nucleus to activate transcription of immune related genes (Figure 1.3; Yan et al. 1996; Brown et al. 2003; Myllymäki and Rämet 2014).

Upd2 and Upd3 (Upd3 in particular) are upregulated following septic injury, parasitic wasp infection, and viral exposure, acting as a damage-associated molecular pattern (DAMP; Agaisse et al. 2003). In the context of immunity, JAK-STAT signalling promotes several critical responses. It induces the expression of TotA (Turandot A) and Tep1 (thioester-containing protein 1) genes involved in systemic immune responses and stress resilience (Ekengren and Hultmark 2001; Dostalova et al. 2017). In cellular immunity, it regulates the differentiation and proliferation of plasmotocytes and lamellocytes (Makki et al. 2010). Activation of JAK-STAT in the lymph gland niche, the Posterior signalling Center (PSC), is essential for initiating lamellocyte production and dissociation of the hematopoietic organ during wasp infection (Krzemień et al. 2007; Gao et al. 2009). Additionally, JAK-STAT signalling contributes to epithelial defence and regeneration, especially in the gut, where it maintains barrier integrity following infection (by promoting proliferation of ISCs) or oxidative stress (Chakrabarti et al. 2016).

1.4 Evolution of immune genes in *Drosophila*

The evolutionary dynamics of immune genes are shaped by the relentless selective pressure exerted by pathogens, which represent one of the most persistent and potent selective forces encountered by animals and plants (Waterhouse et al. 2007; Han 2019; Barrat-Charlaix and Neher 2024). These interactions drive rapid genetic change, particularly in genes that mediate host defence, leading to the repeated observation that immune-related genes rank among the fastest-evolving portions of genomes (Sackton et al. 2007; Obbard et al. 2009a; Singh et al. 2012; Shultz and Sackton 2019; Vinkler et al. 2023). This rapid divergence reflects the evolutionary arms race between host defences and the diverse and fast-evolving microbial threats they encounter.

Much work concerning the evolution of genes of the innate immune system has focused on *D. melanogaster* and its close relatives. Comparative studies in *Drosophila* have revealed stronger signals of adaptive evolution of genes involved in pathogen recognition such as *PGRPs*, *Tep*, *Nimrod* receptors; signalling such as *relish*, *ird5*, *key*, *Dredd*; and antiviral defence such as *Ago2*, *R2D2*, *Dcr2* (Jiggins and Kim 2006; Sackton et al. 2007; Obbard et al. 2009a). Interestingly, while some immune components evolve rapidly, others, particularly core signalling genes and AMPs tend to be more conserved. This pattern of AMP evolution was unexpected, given their central role in the immune response and the frequent signatures of positive selection observed for AMPs in other insects (Viljakainen and Pamilo 2008; Bulmer et al. 2010) and vertebrates (Hollox and Armour 2008; Tennessen and Blouin 2008). One explanation for this pattern is that AMP variation may be maintained through balancing selection rather than continuous directional change. For instance, polymorphism in *Diptericin A* locus has been linked to functional variation in pathogen resistance, suggesting that trade-offs between protection, fitness costs, and microbial tolerance shape their evolution (Unckless and Lazzaro 2016; Chapman et al. 2019).

Gene duplication, loss, and neofunctionalization also play central roles in the diversification of immune responses (Levine et al. 2016; Sackton et al. 2017; Attah et al. 2024; Gao and Zhu 2024). Many AMP families (e.g., Diptericins, Cecropins, Attacins) and recognition molecules (e.g., Nimrod receptors) belong to rapidly evolving, lineage-specific multigene families. These families exhibit high turnover rates, with frequent gains and losses contributing to idiosyncratic immune gene repertoires across species. For example, the AMP Drosomycin is restricted to the *Sophophora* subgenus, while *Diptericin C* (*DptC*) is found only in species of the *Drosophila* subgenus (Hanson et al. 2016). A recent study showed that *DptB*, an AMP that evolved to

provide defence against *Acetobacter* (a prevalent bacterium in fruit-feeding sites) in fruit-feeding *Drosophila* species (Hanson et al. 2023). Remarkably, species that have shifted to feeding on mushrooms or parasitic plant ecologies have repeatedly lost *DptB*, consistent with ecological divergence shaping immune gene evolution (Hanson et al. 2023).

While the rapid evolution of immune genes is a general pattern, the specific genes and pathways under selection often vary between species, reflecting differences in pathogen exposure and ecological context. For example, in *Drosophila innubila*, a mushroom-feeding species from the quinaria group, genomic analyses revealed rapid evolution in components of the Toll pathway but not in antiviral RNAi genes—contrasting with what was found for *D. melanogaster*, where RNAi pathway genes are among the fastest evolving components of the immune system (Hill et al. 2019). This suggests that immune system architecture and the predominant pathogens in a species' environment jointly determine which components of the immune repertoire are most dynamic. While deep phylogenetic divergence can obscure orthology among fast-evolving genes, especially AMPs, improved genome assemblies and annotation methods are gradually overcoming these challenges. Expanding beyond model organisms has revealed both conserved patterns, such as widespread positive selection in recognition and some signalling genes and lineage-specific phenomena, such as novel AMP expansions or immune gene losses correlated with ecological shifts (Sackton et al. 2007; Sackton and Clark 2009; Hanson et al. 2016).

Together, these findings illustrate that immune gene evolution in *Drosophila* is governed by a combination of mechanisms: pervasive positive selection in genes mediating host-pathogen interactions; gene family turnover shaped by ecological specialization; and balancing selection maintaining polymorphism at effector loci. As the number of sequenced *Drosophila* genomes continues to grow, and as functional studies extend beyond *D. melanogaster*, our understanding of how immune systems evolve across diverse evolutionary and ecological landscapes is likely to deepen considerably.

1.5 Research objectives

This thesis investigates the evolution and functional diversity of immune-related genes across the diverse family Drosophilidae. The overall aim of this thesis is to understand how immune system vary across Drosophilidae in terms of gene content, sequence evolution, copy number and expression. To address this, I pursue the following objectives:

Chapter 2

- **Generate high-quality genome annotations across Drosophilidae species**
  I aim to generate protein-coding gene annotations, combining de novo gene prediction with transcriptomic and protein-based evidence, enabling the systematic annotation of over 300 Drosophila genomes. This effort provides the foundational gene models required for downstream comparative and functional analyses.
- **Assess annotation quality**
  I evaluate annotation quality using phylogenetic generalized linear mixed models and demonstrate the use of gene annotation by studying codon usage bias across Drosophilidae.

Chapter 3

- **Characterize patterns of immune gene turnover and sequence evolution**
  Using the annotated gene sets, I quantify gene gains and losses across immune gene classes or pathways, and estimate rate of protein sequence evolution, testing whether immune genes show signatures of elevated adaptative evolution relative to non-immune genes.
- **Identify predictors of immune gene evolutionary dynamics**
  I test whether gene-level and structural features, such as expression level, genetic/protein interactions, relative solvent accessibility, or gene length predict evolutionary rates.

Chapter 4

- **Assess transcriptional response to infection in diverged non-model Drosophilidae species**
  I compare the transcriptional response to gram-negative bacteria infection and evaluate bioinformatic recovery of immune genes using pathogen-challenged RNAseq data.
- **Discover novel antimicrobial peptides in non-model species**
  I aim to identify novel AMP candidates or other effectors using pathogen-challenged RNAseq data, focusing on non-model Drosophilidae species where immune repertoires are currently poorly characterized.

Chapter 2

Comparative gene annotation of 304 species of Drosophilidae

The text in this chapter is from bioRxiv preprint: Dhakad P, Kim B, Petrov D, Obbard DJ (bioRxiv) **"Comparative gene annotation of 304 species of Drosophilidae"** [DOI: 10.1101/2025.04.14.648771]

I wrote this chapter with comments and textual edits from Prof. Darren Obbard. Thanks to Dr. Bernard Kim and Prof. Dmitri Petrov for making the Drosophila genomes data available to us.

2.1 Abstract

High-quality genome annotations are essential if we are to address central questions in comparative genomics, such as the origin of new genes, the drivers of genome size variation, and the evolutionary forces shaping gene content and structure. Here, we present protein-coding gene annotations for 304 species of the family Drosophilidae, generated using the Comparative Annotation Toolkit (CAT) and BRAKER3, and incorporating available RNAseq and protein evidence. We take a comparative phylogenetic approach to annotation, with the aim of improving consistency and accuracy, and to generate a robust set of gene annotations and orthology assignments. We analyse our annotations using a phylogenetic mixed-model approach and find that gene number and CDS length exhibit moderate phylogenetic heritability (43.3% and 12.3%, respectively). This suggests that while evolutionary history contributes to variation in these traits, species-specific factors—including assembly error—play a substantial role in shaping observed differences. To illustrate the utility of our annotations for comparative analyses, we investigate codon usage bias and amino acid composition across Drosophilidae. We find that codon usage is correlated with overall GC content and evolves slowly, but that

it is also strongly shaped by selection—such that, in general, species with the strongest selection on synonymous codon usage show the lowest GC bias in third codon positions. This comparative annotation dataset forms part of an on-going collaborative project to sequence and annotate all species of Drosophilidae, with data and annotations being made rapidly and freely available on an on-going basis. We hope that this effort will serve as a foundation for studies in evolutionary and functional genomics and comparative biology across Drosophilidae.

2.2 Introduction

Fundamental questions in comparative genomics include the origin of new genes, the causes of genome size variation, and the factors that shape rates of genome evolution. However, addressing these questions requires not just the genome sequence, but also accurate identification and characterization of genomic features such as coding DNA sequence. High-quality genome annotations are thus essential for the identification of loci that underpin phenomena such as adaptation and speciation (Ejigu and Jung 2020). However, the annotation of genomes remains challenging, due to complex gene structures, long non-coding regions, and species-specific features (Ejigu and Jung 2020; Kwon et al. 2023; Vuruputoor et al. 2023). Automated annotation methods, while efficient, often produce artifacts that can mislead interpretations of gene function and evolutionary relationships (Promponas et al. 2015; Scalzitti et al. 2020; Mathe and Dunand 2021). For example, an early annotation of the *Daphnia pulex* genome may have over-estimated the number of genes and over-predicted paralogs of genes involved in environmental responsiveness, potentially leading to initial misinterpretations of the basis of its adaptive capabilities (Colbourne et al. 2011; Ye et al. 2017).

Over the past decade, advances in long-read sequencing technologies and scaffolding methods, together with the declining cost of sequencing, have led to a dramatic increase in the number and quality of genome assemblies across the tree of life (Dijk et al. 2023; Espinosa et al. 2024). This surge is exemplified by large-scale initiatives such as the Darwin Tree of Life project (Darwin Tree of Life Project 2022), which aims to sequence around 70,000 species in the UK and Ireland, the Vertebrate Genome Project (Rhie et al. 2021), which has the goal of sequencing all extant vertebrate species, and the African BioGenome Project (Sharaf et al. 2023), which seeks to sequence more than 105 thousand species in Africa. However, despite advances in sequencing and assembly, genome annotation remains a major bottleneck (Freedman and Sackton 2024). Most genomes from large-scale sequencing projects are annotated independently using automated pipelines such as BRAKER, Augustus, and MAKER (Cantarel et al. 2008; Stanke et al. 2008; Gabriel et al. 2024). While these

annotations can incorporate evidence from reference protein databases (e.g., OrthoDB), they typically don't take advantage of closely related species or conserved gene order to improve accuracy and consistency across genomes (Venkatraman et al. 2021; Nachtweide et al. 2024). As a result, different pipelines can identify slightly different sets of genes due to variation in prediction algorithms, parameter settings, and assumptions about gene structure (Weisman et al. 2022). This inconsistency means that genes missed by one pipeline might be detected by another, while certain gene models may only partially align between annotations (Freedman and Sackton 2024). Consequently, such independent annotations make it difficult to achieve a standardized gene set for comparative analyses (Colbourne et al. 2011; Ye et al. 2017; Weisman et al. 2022). Comparative annotation, on the other hand, aims to address these issues by using alignment with well-annotated reference genomes to help guide predictions, reducing discrepancies and aligning gene models more consistently across closely related species (Fiddes et al. 2018; König et al. 2018). This approach can not only increase the accuracy of gene predictions, but can also ensure a more robust, comparable gene set for evolutionary studies. The Comparative Annotation Toolkit (CAT) is one such method, designed to annotate genomes by projecting known gene models from a reference genome onto target genomes within a phylogenetic framework (Fiddes et al. 2018). CAT integrates evidence from such a 'lift-over' with short and long read RNAseq data, Iso-seq data , and protein alignments to refine gene model predictions, weighting features that are shared among close relatives more heavily.

For over a century, *Drosophila melanogaster* and its relatives have been at the forefront of genetics, genomics, and evolutionary research, leading to influential discoveries that have shaped these fields (Beller and Oliver 2006). High-quality genome assemblies and annotations have been developed for several key species, beginning with the pioneering sequencing of the *Drosophila melanogaster* genome, which served as a reference for subsequent genomic studies (Adams et al. 2000; Richards et al. 2005). The Drosophila 12 Genomes Project expanded this foundation, offering comparative insights across multiple species, while initiatives such as the ModENCODE project further enriched our knowledge with detailed transcriptomic and epigenomic data (Clark et al. 2007; Roy et al. 2010). Individual research groups have continued to sequence target species (Mahajan et al. 2018; Puerma et al. 2018; Bronski et al. 2020; Li et al. 2022; such as *D. miranda*, *D. guanche* etc), making possible resources such as DrosOMA (drosoma.dcsr.unil.ch)—providing genus-wide orthology information for 36 Drosophila species (Thiebaut et al. 2023). This progress has now culminated in the on-going community effort to achieve a comprehensive genomic study of the entire family Drosophilidae (Suvorov et al. 2022; Kim et al. 2024). This effort includes the de novo sequencing of new species, scaffolding and improvement of existing genomes, and the generation of new tran-

scriptomic data (Kim et al. 2024). As of April 2024, around 360 different drosophilid species had been sequenced to varying levels of completeness—some fragmentary, from short-read data alone (e.g. *Drosophila setifemur* and *Drosophila ironensis*; Li et al. 2022), but many to chromosome-level assemblies, using long-read data and/or scaffolding information from HiC (e.g. *Chmomyza fuscimana*; Obbard et al. 2023a).

Here we contribute to this continuing community effort by providing a comparative coding-sequence annotation for 304 of the highest quality drosophilid genomes. We do this using a combination of CAT and BRAKER3, combining publicly available RNAseq data and previous RefSeq reference genomes (Fiddes et al. 2018; Gabriel et al. 2024). We use phylogenetic linear mixed models to assess and compare the annotations, and as an example of the utility of our comprehensive annotation we analyse codon and amino-acid usage bias across the family. To facilitate its use in both single-gene and genome-wide studies, the annotation is made freely available in the form of genome annotation files and also as aligned (and optionally masked) orthology groups, with annotations linked to gene orthology . In the future, we plan to continue updating this resource with regular new releases as new genomes become available. We hope that this will be a key resource, enabling gene-based analyses of evolution within this important model system.

2.3 Materials and Methods

2.3.1 Genome assemblies

We selected an initial candidate set of genome assemblies by supplementing those of Kim et al. (2024) with all other publicly available drosophilid genomes available as of February 2024. This included genomes available in the RefSeq database (Release 222; O'Leary et al. 2016), those generated by the Darwin Tree of Life project (Darwin Tree of Life Project 2022), and many assemblies generated by individual labs (Zhou and Bachtrog 2012; Sanchez-Flores et al. 2016; Renschler et al. 2019; Wei et al. 2022). For each species, we short-listed genome assemblies that had an N50 greater than 50 Kbp and a BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness score of over 90% (Simao et al. 2015), selecting the assembly with the highest N50 where multiple assemblies were available. To identify and mask repetitive elements, we used RepeatMasker v4.1.2 (Smit Hubley R & Green P n.d.) and Dfam release 3.7 repeat library (Storer et al. 2021). These soft-masked genomes were used for all subsequent analyses.

2.3.2 RNAseq and protein Data

To annotate the genomes we used the Comparative Annotation Toolkit (CAT; Fiddes et al. 2018), a pipeline that leverages external evidence ('hints') combining data such as RNAseq, Iso-seq, proteins, and attempted lift-over from align ed references. For each species, we first gathered available RNAseq data to provide transcript evidence. We identified suitable RNAseq datasets using the ENA Portal API (www.ebi.ac.uk), selecting up to 10 paired-end RNAseq datasets and prioritizing those with Poly-A selection to enrich mRNA (Yuan et al. 2024). Where available, we included data from up to 10 tissue types, including whole body, carcass, thorax, brain, testes and ovaries, as well as different developmental stages and both sexes (selected SRA numbers for each species are in Supplementary file A.1). The chosen RNAseq reads were then down-sampled and normalized to 100x coverage using BBNorm of BBMap v38.95 (Bushnell 2014). BBNorm normalizes RNAseq reads by down-sampling high-coverage areas to achieve a more uniform coverage distribution, which reduces data file size and accelerates downstream analyses. Reads in regions with low coverage were kept as is, ensuring that these areas were not underrepresented in the normalized dataset. The normalized RNAseq reads were aligned to their respective species genomes using the STAR v2.7.9a with default parameters (Dobin et al. 2013).

To provide protein 'hints', we extracted predicted protein sequences from Arthropoda using the OrthoDB v10 protein database (orthodb.org; Kriventseva et al. 2019). Such protein sequences provide evolutionary conserved evidence that complements RNAseq data, particularly for genes that may be underrepresented or absent in the RNAseq datasets. We aligned these protein sequences to the genomes using miniprot v0.12 (Li 2023) with parameters "*-ut8 –gtf genome_file*", which are optimized for mapping proteins to genomic sequences. The alignment files generated by miniprot were then converted into hints files using the "*aln2hints.pl*" script from the GALBA toolkit (Bruna et al. 2023).

2.3.3 Reference species and cactus alignment

In addition to RNAseq and protein hints, the CAT pipeline attempts a lift-over of annotations (Fiddes et al. 2018). This uses genome-scale alignments in the hierarchical alignment format (Hickey et al. 2013), each comprising a single reference species and several target species. To define reference clades for annotation, we first generated a preliminary species tree of the 304 drosophilid species (plus seven outgroup species) using 1824 single-copy BUSCO loci (Simao et al. 2015). Nucleotide sequences from each locus were aligned separately using MAFFT v7.520 (Katoh and Standley 2013) and used to infer a maximum likelihood (ML) gene

tree using IQ-TREE v2.2.6 (Minh et al. 2020) under a GTR+I+G4 substitution model. These ML gene trees were then combined to infer a species tree using ASTRAL-III v5.15.5 (Zhang 2024), which aims to resolve gene-tree species tree incongruence under a model of incomplete lineage sorting.

We selected 37 'reference' species for lift-over annotations based on the completeness and quality of their genomes, as indicated by RefSeq annotations (O'Leary et al. 2016). Using the ETE 3 python package (Huerta-Cepas et al. 2016), we applied a pre-order tree traversal strategy to identify subclades that contained at least one reference species and included the most distantly related leaf within a predefined phylogenetic distance (measured as the expected number of substitutions per site). We varied the phylogenetic distance threshold between 0.005 and 0.35 to ensure each subclade included 3 to 15 species, with lower thresholds for densely populated regions of Drosophilidae tree and higher thresholds to include more distantly related species. From these, we then chose the reference species with gene counts most similar to those of the best-annotated drosophilid model, Drosophila melanogaster, as its annotation is likely to be the most complete and accurate within the family, having benefited from extensive manual curation and detailed transcriptomic and functional data (Matthews et al. 2015). Each selected reference was then used to define a clade for subsequent genome alignment and lift-over processes. For species located on very long branches (i.e. divergence >0.35) and for subclades lacking any reference-species annotations, we attempted a lift-over directly from Drosophila melanogaster.

We then used these 'lift-over subclades' as guide trees to generate multiple whole-genome alignments with ProgressiveCactus (Armstrong et al. 2020). This approach ensured that the alignments were computationally feasible, and that closely-related genomes aligned together. Finally, we employed the Comparative Annotation Toolkit (Fiddes et al. 2018) to annotate multiple target genomes simultaneously, using a lift-over of the selected reference annotation for each subclade.

2.3.4 Running CAT

To perform the genome annotations, we first prepared the reference annotations and extrinsic 'hints' for use in CAT (Fiddes et al. 2018). RefSeq annotation files were converted using the "convert_ncbi_gff3" script provided by CAT, and the resulting GFF3 files were validated with the "validate_gff3" script to ensure compatibility. We then employed three modes of AUGUSTUS (Stanke et al. 2008) in CAT: two based on transMap projections (AugustusTM/R) that project annotations from reference genomes onto target genomes, and one using ab-initio and

comparative gene predictions (AugustusCGP) guided by extrinsic hints (Konig et al. 2016). We used Comparative Gene Prediction (CGP) parameters trained on 12 well annotated Drosophila species from the Drosophila 12 Genomes Project, based on exon and intron scoring (bioinf.uni-greifswald.de/augustus/datasets/; Konig et al. 2016).

2.3.5 Complementation with BRAKER3

To complement the comparative annotations generated by CAT, and to reduce potential reference bias, we additionally incorporated de novo non-comparative CDS predictions made by BRAKER3 (Gabriel et al. 2024). BRAKER3 is an automated gene prediction pipeline that integrates RNAseq data with gene prediction algorithms to generate gene models without reference to a reference annotation, improving the identification of novel genes and gene duplicates that may be absent from the reference. We provided RNAseq reads and Diptera protein sequences as external source of hints. However, among the annotations generated by BRAKER3 we identified a number of transposable elements (TEs) which are not the focus of our study. Therefore, to simplify the annotation, we removed TEs annotated using EarlGrey v4.1.0 (Baril et al. 2024). We applied TEstrainer (github.com/jamesdgalbraith/TEstrainer) to retain recently duplicated non-TE genes that were identified as TEs by EarlGrey. Finally, to combine the BRAKER3 annotations with those from CAT we compared the coding sequences (CDS) of overlapping genes between the two annotation sets. For one-to-one overlapping genes, we selected the annotation with the longest CDS. In cases of one-to-many or many-to-one overlaps, we preferred the CDS annotations from CAT. Additionally, we retained all non-overlapping genes with a CDS length greater than 150 nucleotides.

2.3.6 Annotation quality assessment

To assess the annotation quality, we used OMArk and BUSCO for assessing completeness based on evolutionary informed expectations of gene content (Simao et al. 2015; Nevers et al. 2025). BUSCO was run in protein mode using the Diptera OrthoDB dataset to estimate the proportion of conserved single-copy orthologs recovered in each annotation (Kuznetsov et al. 2023). For OMArk, we first generated omamer search databases for each species using the LUCA.h5 orthology database (omabrowser.org/oma/current/). We then ran OMArk with taxon ID 7214 (Drosophilidae), assigning proteins to orthologous groups based on phylogenetically informed gene family classifications.

2.3.7 Orthogroup assignment and CDS alignment

We identified coding sequence homology across the 304 Drosophila species and one outgroup species (*Musca domestica*) using OrthoFinder v2.5.5 (Emms and Kelly 2019). OrthoFinder first identifies homology using an all-vs-all blast similarity search and then clusters sequences using a Markov clustering algorithm (MCL inflation parameter of 1.5). Subsequently, it can then identify "Hierarchical Orthologous Groups" (HOGs) that comprise the genes descended from a common ancestral gene at a specific taxonomic level—with HOGs defined for different clades nested within each other along the species phylogeny. We extracted HOGs at the level of Drosophilidae (i.e. sets of homologous sequences that have their MRCA at the base of Drosophilidae, or later) and retained for further analysis those HOGs that included at least two species and contained more than three sequences.

The sequences for each of the chosen HOGs were aligned using the MACSE v2 (Ranwez et al. 2018) and MAFFT v7.520 (Katoh and Standley 2013) aligners, an approach intended to minimize the impact of any frameshifts and in-frame stop codon errors that might arise from sequencing, assembly, or annotation problems, while maximizing codon-aligned sequence length. MACSE v2 incorporates several steps to improve alignment quality, including a prefilter to remove long non-homologous insertions that may result from incorrect annotations, such as intron inclusions or alternative splicing. It then uses HMMCleaner to mask residues that appear misaligned and applies post-processing filters to mask isolated codons and patchy sequences, removing sequences if more than 80% of the residues are masked (Di Franco et al. 2019). Finally, the alignments were trimmed at both ends until a nucleotide position represented by at least 70% of the sequences was reached, ensuring a high-quality alignment for downstream analyses of protein coding sequence. From the 304 annotated drosophilid genomes, we generated 35,836 HOGs and 23,150 high-quality alignments. These HOGs form the basis of subsequent analyses of evolutionary relationships and functional conservation of genes across family Drosophilidae, and all masked and unmasked aligned sequences are made available at 10.5281/zenodo.15016917.

2.3.8 Phylogenetic generalized linear mixed model analyses

We used Generalized Linear Mixed Model (GLMM) analyses of gene number and CDS length to identify clade-to-clade variation and outlier genomes across the 304 species. Such variation may result from true evolutionary divergence, or from reference bias and the availability (or otherwise) of RNAseq data, or from systematic errors in assembly or annotation quality. The relative impact of such factors is naturally addressed in a linear mixed model framework,

treating species as a random effect and phylogenetic distance from reference species, genome size, assembly contiguity (N50), and the availability of RNAseq as fixed-effect predictors. Because related species exhibit correlated traits (leading to pseudo-replication, if not accounted for; Freckleton 2009), and because the phylogenetic correlation among related species (e.g. 'phylogenetic inertia' or 'phylogenetic heritability') may be of direct interest—reflecting clade-to-clade variation in gene content or the efficacy of selection—we employed a Phylogenetic mixed model approach (Hadfield and Nakagawa 2010), implemented in the R package MCMCglmm (Hadfield 2010). This incorporates phylogenetic relationships to model the covariance among species, while evaluating the influence of fixed predictors such as phylogenetic distance from the lift-over reference and RNAseq availability.

To do this, we first generated a revised species-tree topology of the 304 drosophilid species using 251 single-copy HOGs employing IQTREE2 and ASTRAL-III, as for BUSCO genes above. To infer relative branch lengths in approximate time, we randomly selected 10,000 amino acid sites from the HOGs, and (conditioning on the IQTREE/ASTRAL topology) we used BEAST (Suchard et al. 2018) to re-infer branch lengths on the fixed ASTRAL tree topology under a LG+G+I model with 7 gamma categories and an uncorrelated relaxed log-normal clock (Drummond et al. 2006). For the tree prior, we used birth-death process model (Gernhard 2008), setting the fully-informative prior for the MRCA of the subgenera *Drosophila* and *Sophophora* to 47 mya (95% prior density 42-52 mya; Suvorov et al. 2022), a uniform step prior between 0 and 1 on the birth-death growth rate, and the remaining priors to their default values. We ran the MCMC chain for 109 generations sampling every 20,000 steps, and stationarity and mixing were assessed from visual inspection of the MCMC chain and Effective Sample Size (ESS) in Tracer (Rambaut et al. 2018). After discarding 20% of the sampled estimates as burn-in, we report divergence times as median node height for each of the clades in the summary tree.

To assess the factors influencing gene number and CDS length across drosophilid species, we fitted a multivariate phylogenetic mixed model using MCMCglmm (Hadfield 2010). Our model included gene number and mean CDS length as response variables, allowing us to analyse their (co-)variation with respect to predictors such as distance from reference, RNAseq availability, status as a liftover reference, assembled genome size, and assembly scaffold N50. This phylogenetic approach can help to disentangle the effects of biological and technical factors on genome annotation metrics, while accounting for phylogenetic relatedness among species. We inferred statistical 'significance' on the basis of 95% highest posterior density (HPD) credibility intervals. The model was specified as follows (MCMCglmm syntax):

```
prior <- list(B = list(mu = rep(0, 12), V = diag(12) * 1e+10),
    G = list(G1 = list(V=diag(2), nu=2, alpha.mu=rep(0,2),
    alpha.V=diag(2)*1000)),
    R = list(V=diag(2), nu=2.002))

model <- MCMCglmm(cbind(mean_len, genes) ~ trait - 1 +
    trait:distances + trait:RNA_seq + trait:ref +
    trait:genome_size + trait:N50, random = ~us(trait):Phylo,
    rcov = ~us(trait):units, family = rep("gaussian", 2),
    ginverse = list(Phylo = InverseTree),
    prior = prior, data = annot_stat,
    nitt = 1000000, burnin = 100000,
    thin = 1000, pr = TRUE)
```

Briefly, trait is a reserved variable that indexes columns of the response matrix in multi-response models, with -1 removing the global intercept so that each trait has its own baseline estimate. Fixed effects included phylogenetic distance from the reference species (*trait:distances*), availability of RNAseq data (*trait:RNA_seq*), whether the species was itself a lift-over reference (*trait:ref*), assembled genome size (*trait:genome_size*), assembly contiguity (*trait:N50*). The random effects term *us(trait):Phylo* describes the phylogenetic (co)variance matrix between gene length and gene number, and the residual variance term *us(trait):units* describes the residual covariance matrix. The argument ginverse fits a covariance structure among species to model non-independence due to common ancestry (Hadfield 2010). Variance in CDS length and gene number were treated as Gaussian.

2.3.9 Evolution of GC, codon, and amino acid composition across Drosophilidae

Genome-wide GC content, codon usage, and amino-acid composition are shaped by a combination of mutational bias and natural selection (Hershberg and Petrov 2008; Kokate et al. 2021). To illustrate the utility of our coding sequence annotations and alignments for large-scale evolutionary sequence analyses, we examined the evolution of these coding-sequence traits across the 304 drosophilid species. We used the R package ‘cubar’ (Zhang 2024) to calculate the overall GC content of coding sequences (‘GC’), and GC content at third codon positions (GC3). Additionally, we estimated the whole-genome GC content and the GC content of non-coding sequences using ‘geecee’ (Blankenberg et al. 2007). We used ‘cusp’ tool from

EMBOSS (Rice et al. 2000) to calculate amino acid frequencies in each species, and the nitrogen to carbon ratio (N/C) for protein sequences was calculated as the weighted average of the N/C ratios of individual amino acids, with weights corresponding to the proportion of each amino acid in the sequence. We used a PCA analysis of amino acids frequencies to reduce the dimensionality.

We estimated the strength of selection on codon usage bias (S) using the approach of Reis and Wernisch (2009), which compares codon frequencies in highly expressed versus reference gene sets. We ranked Drosophila melanogaster genes according to their overall expression level (Supplementary file A.2; expression data obtained from FlyBase: flybase.org) and analysed the HOGs that contained these genes, assuming that the globally most highly-expressed genes in Drosophila melanogaster are also highly expressed in other species. As expected, these genes were dominated by those encoding ribosomal proteins, yolk proteins, salivary gland secretions and elongation factors whose high expression is likely to be conserved. HOGs were then ranked in order of Drosophila melanogaster expression level and binned in 20 expression categories of ca. 600 genes in each category (Supplementary file A.3). S was then estimated as the log-odds ratio of optimal to non-optimal codon frequencies for two-fold degenerate codons, where the preferred codon was identified from the most highly expressed gene category (Eyre-Walker and Bulmer 1995; Reis and Wernisch 2009). Note that, to distinguish selection from mutational bias, this method assumes the reference and highly expressed gene sets have similar mutational patterns. Finally, we obtained bootstrap confidence intervals for S by resampling genes within expression categories.

As above, we used a multivariate Phylogenetic Generalized Linear Mixed Model (PGLMM) implemented in MCMCglmm to assess the among-species phylogenetic (co)variance in GC3, non-coding GC, estimated strength of selection on codon usage bias (S), observed frequencies of amino acids and N/C ratio, while accounting for phylogenetic effects.

2.3.10 Data availability

All code used for data processing, generating figures, and statistical analyses are available through GitHub (github.com/DhakadPankaj/Fly-annotation). Drosophila annotation data is publicly available at 10.5281/zenodo.15016917.

2.4 Results and Discussion

2.4.1 Gene annotation of 304 species

We selected 17 NCBI 'RefSeq' annotations for use as potential lift-over references and combined this information with RNAseq data from 91 species, protein hints (from mapping of OrthoDB proteins) from all species, and de novo prediction to annotate the remaining genomes using CAT and Braker3. On average, we identified 14,543 genes in each genome, with a mean CDS length of 1.6 Kbp. This is very similar to the gold-standard reference, *Drosophila melanogaster*, which currently has 13,904 protein-coding genes of mean CDS length 1.54 Kbp (Genome assembly Release 6.53; GCF_000001215.4). However, there was substantial variation, reflecting both evolutionary variation among species and potential variation in genome assembly quality and RNAseq availability (Figure 2.1).

Most species (291 of 304) were annotated with between 12,500 and 17,000 protein-coding genes, but with notable outliers. These included *Drosophila vulcana* (Bronski et al. 2020), *Drosophila miranda* (Bachtrog et al. 2019) and *Drosophila punjabiensis* (Bronski et al. 2020), which each appeared to possess more than 20,000 genes (Figure 2.1 and Supplementary file A.4). Previous studies have suggested that gene gain on *Drosophila miranda*'s neo-Y chromosome may account for the elevated gene count in this species (Bachtrog et al. 2019), and similar major structural changes could extend to other species with unexpectedly high gene numbers. However, it could also be that the miss-assembly or annotation errors led to unusually high numbers of genes (Torresen et al. 2019).

The mean CDS length for most species (296 of the 304 species) ranged between 1.42 and 1.74 Kbp. The outliers included *Drosophila pseudoobscura ssp. pseudoobscura* (GenBank accession: GCA_000001765.3), which exhibits unusually short CDSs (mean length of 1.15 Kbp) but a total gene count of 14,546 close to the family median of 14,249 genes (Figure 2.1 and Supplementary file A.4). This observation raises the possibility that many gene models in such assemblies may be fragmented, incomplete, or represent short repetitive elements,

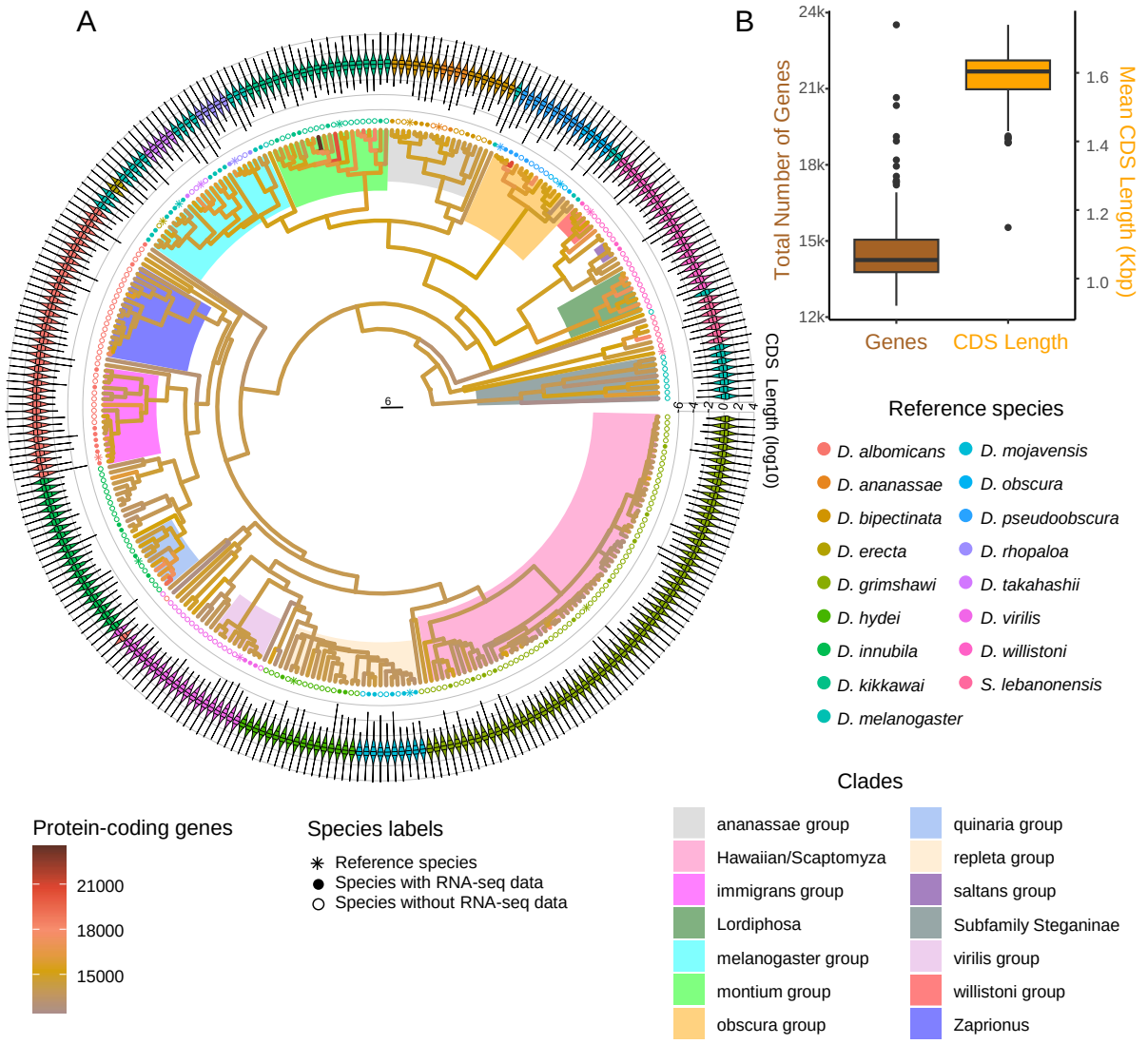

Figure 2.1: Overview of 304 Drosophila genome annotations.
(A) Phylogenetic tree with ancestral reconstruction of the number of protein-coding genes mapped onto branches. Tip labels are coloured according to the reference species for the clade, stars indicate the reference species, and filled versus open circles indicate the availability of RNAseq data for that species. Violin plots (outermost layer) represent the distribution of coding sequence (CDS) lengths on a logarithmic scale. (B) The box plot represents the range in gene number and mean CDS length across family Drosophilidae. An alternative version of the tree, with taxon labels, is provided in Supplementary file A.7.

possibly reflecting issues with assembly quality or annotation (Torresen et al. 2019; Ko et al. 2022). Another possibility is that these observed variations in annotated gene number and CDS length might arise due to duplication/loss of gene families in the ancestors of group of species or at tips.

Establishing whether this variation reflects differences in annotation or assembly quality, or true evolutionary processes, is necessarily challenging in the absence of ground truth data for comparison. We first used BUSCO to assess the genome completeness at both the genome and annotated-protein levels. A close agreement between them suggests that the new annotations do not lack significant portion of BUSCO genes detectable at the genome level (Figure A.1). Only *Drosophila pseudoobscura ssp. pseudoobscura* and *Drosophila americana* had difference of more than 10% between genome and protein levels BUSCO completeness. We also used OMArk to further assess and compare the quality of protein-coding genes. OMArk estimates the completeness, consistency, fragmentation and contamination of gene-repertoire in a species by comparison with conserved orthologous groups (HOGs). We observed high levels of HOG completeness across most species. However, *Drosophila recens* and *Drosophila miranda* showed a high number of duplications (Supplementary file A.5). For *Drosophila miranda* it has previously been shown that there are occurrences of gene gain on its neo-Y chromosome, but the source of duplicates in *Drosophila recens* remains unclear and not previously reported. Interestingly, we also found a high number of duplications in *Drosophila recens*' close relatives *Drosophila subquinaria* and *Drosophila suboccidentalis* (Supplementary file A.5). Additionally, OMArk identified a significant level of contamination in the genome of *Drosophila vulcana* (5,235 genes identified as contaminant), a genome that is also flagged as 'contaminated' in NCBI database (Bronski et al. 2020). Other genomes with apparently high levels of contamination include *Drosophila punjabiensis* (3,131 genes), *Drosophila prolaticilia* (3,811 genes) and *Drosophila nannoptera* (2,006 genes).

2.4.2 Orthology inference

We used OrthoFinder (Emms and Kelly 2019) to infer CDS orthology across 304 species of Drosophilidae, using *Musca domestica* as an outgroup. OrthoFinder assigned 98.9% of predicted proteins to orthogroups (OGs), with 95.5% further classified into Hierarchical Orthologous Groups (HOGs). We identified a total of approximately 35 thousand HOGs across the 304 drosophila species genomes (Table 2.1). More than 90% of genes in each species were assigned to HOGs, although some species—such as *Drosophila pseudoobscura ssp. pseudoobscura*, *Leucophenga varia*, *Drosophila vulcana*, *Drosophila quasianomalipes*, *Drosophila americana*, and *Drosophila differens*—had lower assignment rates (Figure A.2).

Feature	Count
Number of species	304
Total number of proteins	4438506
Number of Orthogroups (OGs)	38692
Number of proteins in OGs	4391361 (98.9%)
Number of root-level HOGs	35836
Number of proteins in HOGs	4240044 (95.5%)
Number of universal single-copy HOGs ($\geq$300 species)	251
Number of HOGs with all species present	775
Number of HOGs with $\sim$99% of species present ($\geq$301)	4994
Number of HOGs containing Dmel genes	12151
Number of ancient HOGs (mrca $\geq$50 mya)	15974

Table 2.1: Summary statistics of HOGs.

To interpret patterns of gene conservation and turnover, we classified HOGs into two broad categories: widely conserved HOGs, which include genes shared across most Drosophilidae species, and species- or clade-specific HOGs, which appear to be restricted in their distribution. The widely conserved HOGs were further divided into 'universal' HOGs, present in nearly all species ($\geq$99%), and ancient HOGs, shared by species with MRCA of at least 50 mya. Universal HOGs likely represent genes experiencing strong evolutionary constraint, encoding core biological functions necessary across all species. Ancient HOGs, while also conserved, may include genes that have been differentially retained or lost in some major lineages. In contrast, species- and clade-specific HOGs could represent recent gene family expansions or lineage-specific adaptations. Where restricted HOGs encode functional proteins, they may contribute to species-specific traits; however, they also likely arise from methodological artifacts, such as genome annotation errors or orthology misassignment.

Over half of all the predicted protein-coding genes fell into universal HOGs, reinforcing the idea that most genes are highly conserved. However, a substantial fraction of the HOGs ($\sim$20,000) contained genes from a small number of species (<30), suggesting either recent evolutionary gains or problems with orthology inference. Interestingly, some HOGs classified as ancient were present in only a few species, and it seems probable that many of these 'sparse' HOGs reflect fragmented assemblies or annotation inconsistencies, rather than true biological patterns. This highlights the importance of caution when interpreting or analysing low-representation HOGs in comparative studies.

By classifying HOGs based on their evolutionary conservation and phylogenetic distribution, we provide a framework for future studies of gene conservation and turnover in Drosophilidae. Universal and ancient HOGs likely represent functionally essential genes, while restricted HOGs may point to lineage-specific innovations or methodological challenges. This classification allows us to distinguish between broad evolutionary patterns and potential technical noise, improving the reliability of our comparative analyses.

To obtain a well-supported set of orthogroups, and to mitigate potential errors from sparse HOGs, we examined HOGs that contain at least one *Drosophila melanogaster* gene in more detail. Given that *Drosophila melanogaster* has one of the most complete and thoroughly verified gene sets among multicellular eukaryotes, its representation in a HOG provides additional confidence that the group represents a biologically meaningful gene family rather than an artifact. We identified 12,151 such HOGs, which were broadly shared across the majority of Drosophilidae species (Figure A.3). This approach provides greater confidence when analysing conserved gene families and allows us to assess how widely well-characterized genes are distributed across phylogeny. Our analysis substantially expands the most recently-available gene-orthology set for Drosophilidae from 36 species (Thiebaut et al. 2023) to 304 species, offering a more comprehensive understanding of gene families across the entire clade, and we hope that this dataset will be valuable for future studies focusing on comparative genomics, evolutionary biology, and functional genomics within Drosophilidae.

2.4.3 Phylogenetic inference using BUSCO and HOG genes

Comparative analyses require an ultrametric phylogenetic tree describing relationships, and thus the expected covariance in traits, among species (Hadfield and Nakagawa 2010). The most comprehensive molecular phylogeny of Drosophilidae to date encompasses 704 species, but is based on only 17 reference genes and thus has many deep branches that are not resolved with high confidence (Finet et al. 2021). More recently, genome-sequencing of 360 species has enabled a BUSCO-gene tree based on one thousand loci—greatly improving confidence in deeper relationships (Kim et al. 2024). However, branch lengths were inferred using 4-fold degenerate sites, which will tend to underestimate deep branch lengths due to substitution saturation. While both BUSCO genes and single-copy HOGs are conserved across Drosophilidae species and thus likely to experience greater purifying selection, BUSCO genes are identified based on their universal presence across a broader evolutionary scale (i.e., Diptera) whereas HOGs are defined within Drosophilidae. Here we infer species trees using two alternative gene-sets, one using 251 single-copy HOGs, and one using 1,824 BUSCO genes. We found that the HOG tree and the BUSCO tree were highly concordant and showed

highly supported relationships for all but five species (Figure A.4). For example, in the HOG tree *Drosophila fuyamai* was positioned as a sister to *Drosophila carrolli*, *Drosophila rhopaloa*, and *Drosophila prolongata*, whereas the BUSCO tree also included *Drosophila kurseongensis* in this clade. Other conflicting relationships can be found in the Figure A.4. Most internal branches were well supported in both trees, but in some places, the HOG tree exhibited slightly lower local posterior probabilities, particularly for short branches (Supplementary file A.8 vs Supplementary file A.9). These discrepancies likely reflect increased discordance due to incomplete lineage sorting (Suvorov et al. 2022), as resolving such branches requires a larger number of gene trees.

2.4.4 Factors affecting annotated gene number and CDS Length

To better understand apparent variation in gene number and CDS length across Drosophilidae, we fitted a phylogenetic generalized linear mixed model (Hadfield 2010) to assess the impact of 'reference' genome status, phylogenetic distance from the reference, genome size, assembly quality (N50), and the availability of RNAseq data on these two traits.

Our analysis found no significant differences in gene number or CDS length between our annotations and the established reference annotations, indicating that our annotations are of comparable quality (gene number: p=0.7, 95% HPD CI [-565, 402]; CDS length: p=0.6, 95% HPD CI [-0.04, 0.005]). However, species lost an average of ca.14 genes for each extra million years of divergence from their liftover reference (p<0.001, CI [-22, -6.2]), while on average CDS length increased by just one nucleotide per million years divergence from reference (p=0.002; CI [0.03, 1.45]; Figure 2.2). This reflects the increased challenges associated with lift-over between more divergent genomes, but—likely because lift-over was only one source of data among many—the effect is relatively small. Increased assembly contiguity (higher N50) in these generally highly contiguous genomes (minimum N50 for inclusion 50 Kbp) was unexpectedly associated with an average of ca. 13 fewer predicted genes. This may be consistent with a reduction in fragmented gene models (p=0.004; CI [-22, -3.5]; Figure 2.2). The inclusion of RNAseq data also unexpectedly reduced gene number (by ca. 400 genes; p=0.002; CI [-692, -143])—without affecting CDS length (p=0.2, CI [-42, 8.26]). This may reflect a reduction in the overall false-positive rate, or a reduction in annotated pseudogenes, or the joining of disjunct exons. We found that ca. 13 genes were on average gained with each additional 1Mbp of genome assembly size (p<0.001; CI [10, 16.5]), but mean CDS length decreased by less than 1 bp (p<0.001; CI [-0.9, -0.4]; Figure 2.2)—that is, larger genome assemblies contained more genes, but these genes were shorter on average. This pattern

could reflect biological processes, such as gene duplications or the expansion of non-coding regions. Alternatively, it might result from the annotation of transposable elements (TEs) as genes, or more fragmented assemblies (perhaps as a result of more repetitive sequence) and thus more fragmented gene models (Konstantinidis and Tiedje 2004).

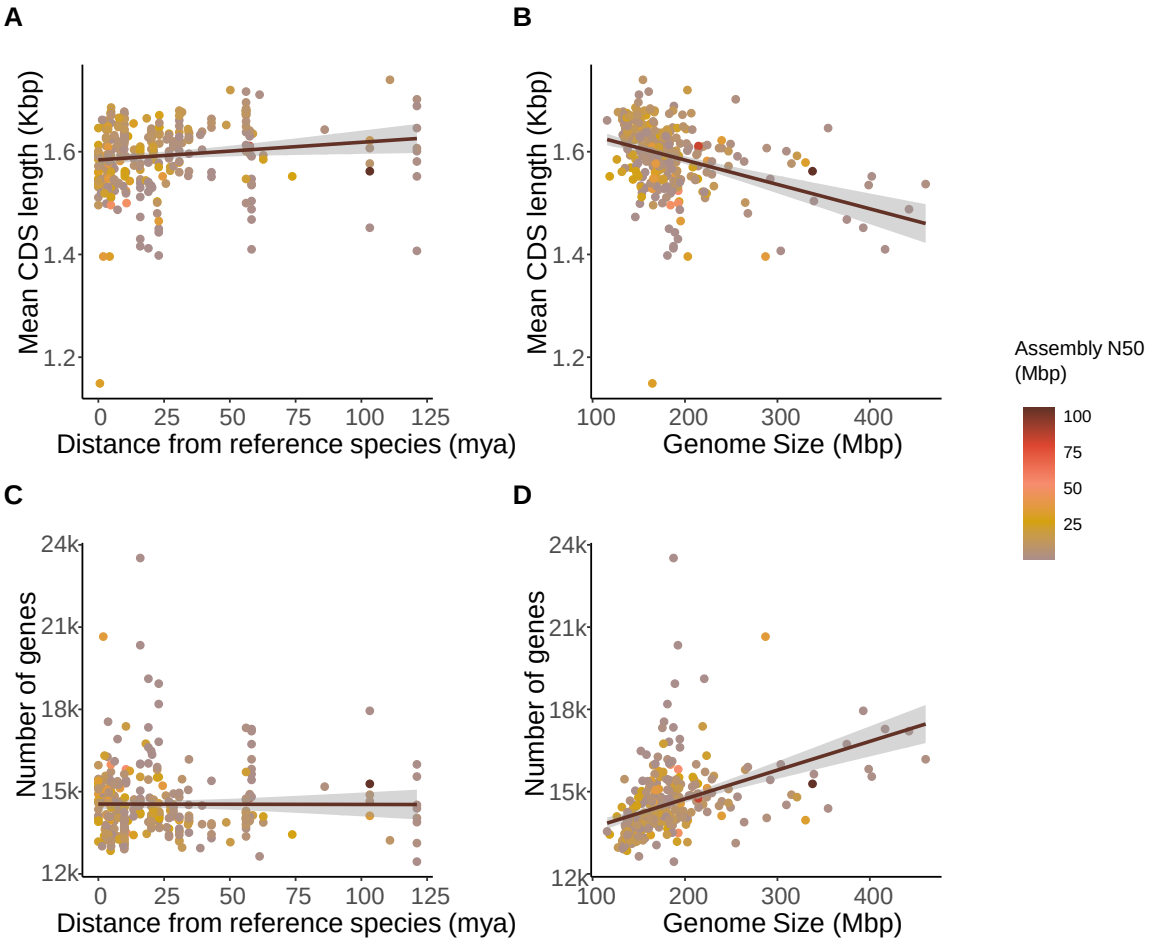

Figure 2.2: Variation in gene number and CDS length across Drosophilidae.
(A) Coding sequence (CDS) length increases with phylogenetic distance from the reference species used for lift-over annotation. (B) CDS length decreases as genome size increases. (C) Gene number remains largely unaffected by phylogenetic distance from the reference. (D) Gene number increases with genome size, suggesting a relationship between genome expansion and gene content. In all plots the points are coloured by genome contiguity (N50). Fitted lines and 95% confidence windows are derived from a non-phylogenetic linear model and are for illustration only; see main text for phylogenetic mixed model analyses.

After accounting for these fixed effects, we found little evidence for major differences in gene number or CDS length among Drosophila clades. Posterior estimates of differences among internal nodes generally had confidence intervals overlapping zero, suggesting that number of genes and CDS length have remained relatively stable across lineages (Figure A.5). Correspondingly, phylogenetic heritability took intermediate values for both gene number (43.3%; CI [23.8, 64.3]) and mean CDS length (12.3%; CI [4.2, 23.4]), indicating that while evolutionary history plays a role, species-specific factors contribute substantially to variation in these traits. The negative covariance between gene number and CDS length (-0.42; CI [-0.77, -0.12]) suggests species with more genes tend to have shorter CDS lengths, possibly reflecting differences in annotation stringency or genome assembly quality. This suggests that large-scale shifts in these genomic traits are rare, and variation in gene number and length are more likely driven by annotation artifacts or species-specific factors rather than broad evolutionary trends. Nevertheless, the montium group had a generally higher inferred gene number and shorter CDS lengths compared to other clades (Figure A.5), which may reflect lineage-specific effects—although this group includes *Drosophila vulcana* and *Drosophila punjabiensis*, which both showed high bacterial contamination in the OMArk analysis. Taken together, these results indicate that while gene number and CDS length variation occur at the species level, they do not seem to have strong phylogenetic structuring at deeper evolutionary timescales, perhaps predominantly reflecting differences in assembly and annotation rather than long-term evolutionary trends.

2.4.5 GC composition and Codon Usage Bias in Drosophilidae

To illustrate the potential utility of our new annotations, we analysed variation in GC content and its relationship with codon usage bias (CUB) across Drosophilidae (Behura and Severson 2012; Kokate et al. 2021). Genomic GC content ranged from 21% in *Drosophila neohypocausta* to 49% in *Drosophila nannoptera* (Supplementary file A.6), with coding regions, as expected, showing higher GC content (range: 41–57% GC) than the genome-wide average. GC content at third codon positions (GC3) is a widely used proxy for codon bias, and reflects a balance between mutational pressure and selection acting on synonymous mutations (Behura and Severson 2012; Kokate et al. 2021). In our analysis, GC3 was highly correlated between related species, with an estimated phylogenetic heritability of 1 (i.e. no residual variance that is not captured by the phylogenetic effect), indicating strong conservation within clades and little variation among closely related species. GC3 was also positively correlated with non-coding GC content (Figure 2.3 and Figure 2.4; phylogenetic correlation from the PGLMM: 0.52; $p<0.001$; CI [0.41, 0.59]), suggesting that genome-wide mutational biases contribute to

both coding and non-coding base composition. Nevertheless, we observed substantial clade-specific deviations; notably the willistoni and saltans groups, along with subfamily *Steganinae*, had much lower GC3, whereas the genus *Zaprionus* and the melanogaster, montium, obscura, ananassae, repleta, and virilis species groups exhibited elevated GC3 (Figure A.6). These differences mirror a recent analysis of 29 Drosophila species, in which subgenera *Sophophora* and *Drosophila* exhibited distinct codon preferences (Kokate et al. 2021). Such lineage-specific differences could reflect factors beyond mutation bias or overall GC composition, potentially including selection for translational efficiency or efficacy (Heger and Ponting 2007).

To quantify the role of selection in determining codon usage, we estimated the strength of selection on two-fold codons (quantified by the 'S' statistic of Reis and Wernisch 2009). Estimates of S ranged from 0.24 in *Drosophila pachea* (95% bootstrap interval across genes [0.22, 0.29]) up to 1.08 [0.96, 1.20] in *Drosophila takahashii* (Figure 2.3). As expected, the melanogaster, montium, and ananassae species groups showed elevated GC3 and S, confirming stronger selection in favour of GC-ending codons in these groups (Heger and Ponting 2007; Vicario et al. 2007; Kokate et al. 2021). However, the willistoni and saltans groups—which display low GC3—also showed relatively high S, confirming that the AT-bias seen in *Drosophila willistoni* is (at least in part) a result of selection (Powell et al. 2003; Singh et al. 2006; Heger and Ponting 2007). A similar pattern was also seen in the subfamily *Steganinae* and overall GC3 and S are negatively correlated across Drosophilidae—indicating that species with higher GC3 are, on average, experiencing weaker selection on codon usage (Figure 2.4; phylogenetic correlation from the PGLMM: -0.2; $p<0.001$; CI [-0.33, -0.10]). Interestingly, a weak positive correlation between S and genome size (Figure 2.4; phylogenetic correlation from the PGLMM: 0.15; $p=0.02$; CI: [0.03, 0.32]) suggests that species with larger genomes tend to experience slightly stronger selection on codon usage—in contrast to what might be expected under relaxed constraint in species with small effective population size (Petit and Barbadilla 2009).

2.4.6 Amino Acid Composition

To investigate whether the variation in codon usage is associated with variation in amino acid composition, we analysed the relative proportions of all 20 amino acids across the annotated proteins of Drosophilidae. In general, it is thought that amino acid usage is influenced by a combination of mutational biases, translational selection, and functional constraints—but genome-wide nucleotide composition has been shown to play a significant role in shaping amino acid frequencies (Behura and Severson 2012; Williford and Demuth 2012). Our principal component analysis (PCA) revealed that more closely related species share more similar

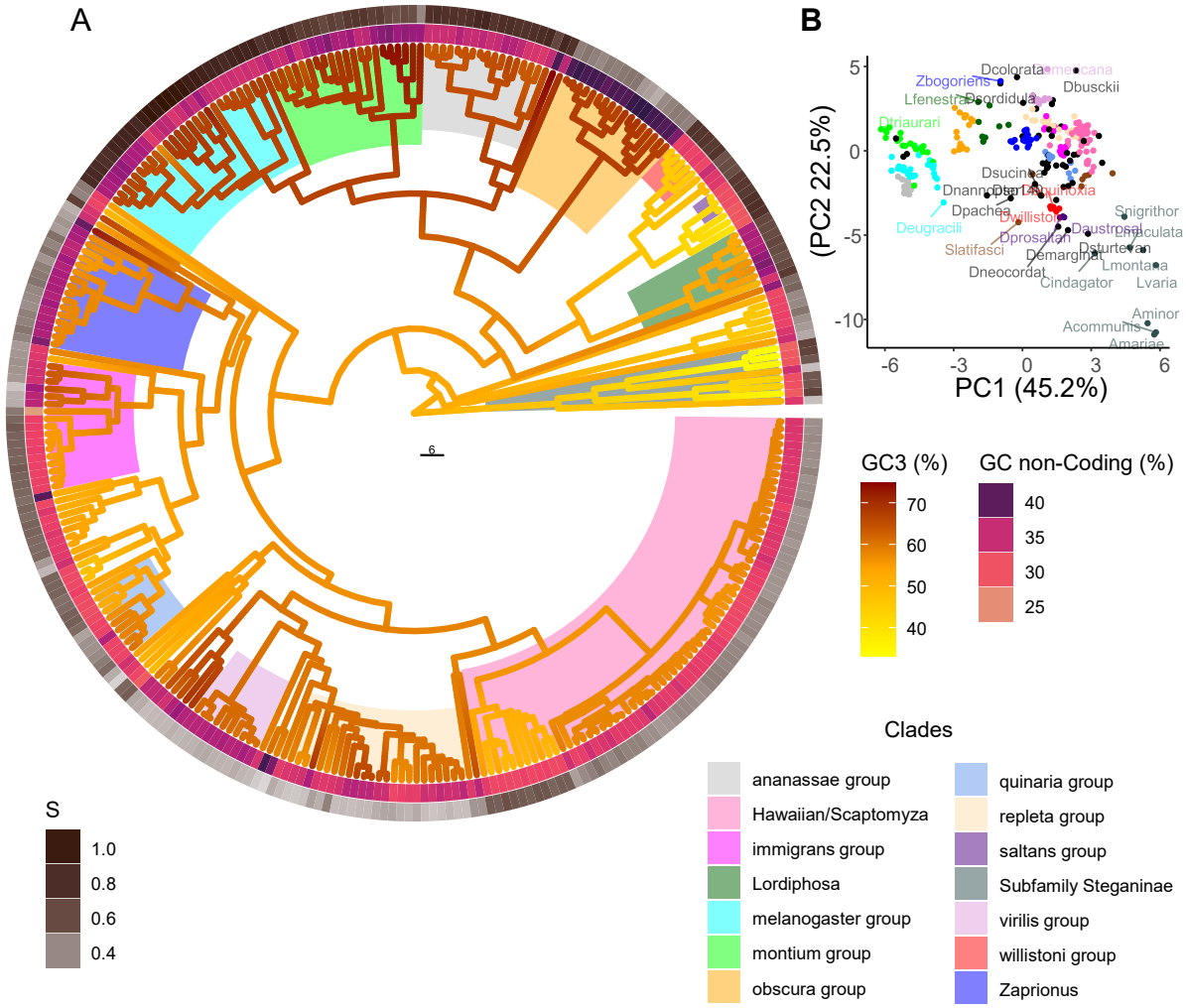

Figure 2.3: Codon usage, amino acid composition and selection on codon usage across Drosophilidae.

(A) Phylogenetic distribution of GC content at third codon positions (GC3), non-coding GC content, and strength of selection on codon usage bias (S) across Drosophilidae. The tree is color-coded by clades, and branches are coloured according to GC3. Inner ring shows GC content in non-coding regions and outer ring shows strength of selection on codon usage. (B) Principal component analysis (PCA) of amino acid usage across species, showing that closely related species exhibit similar amino acid composition patterns.

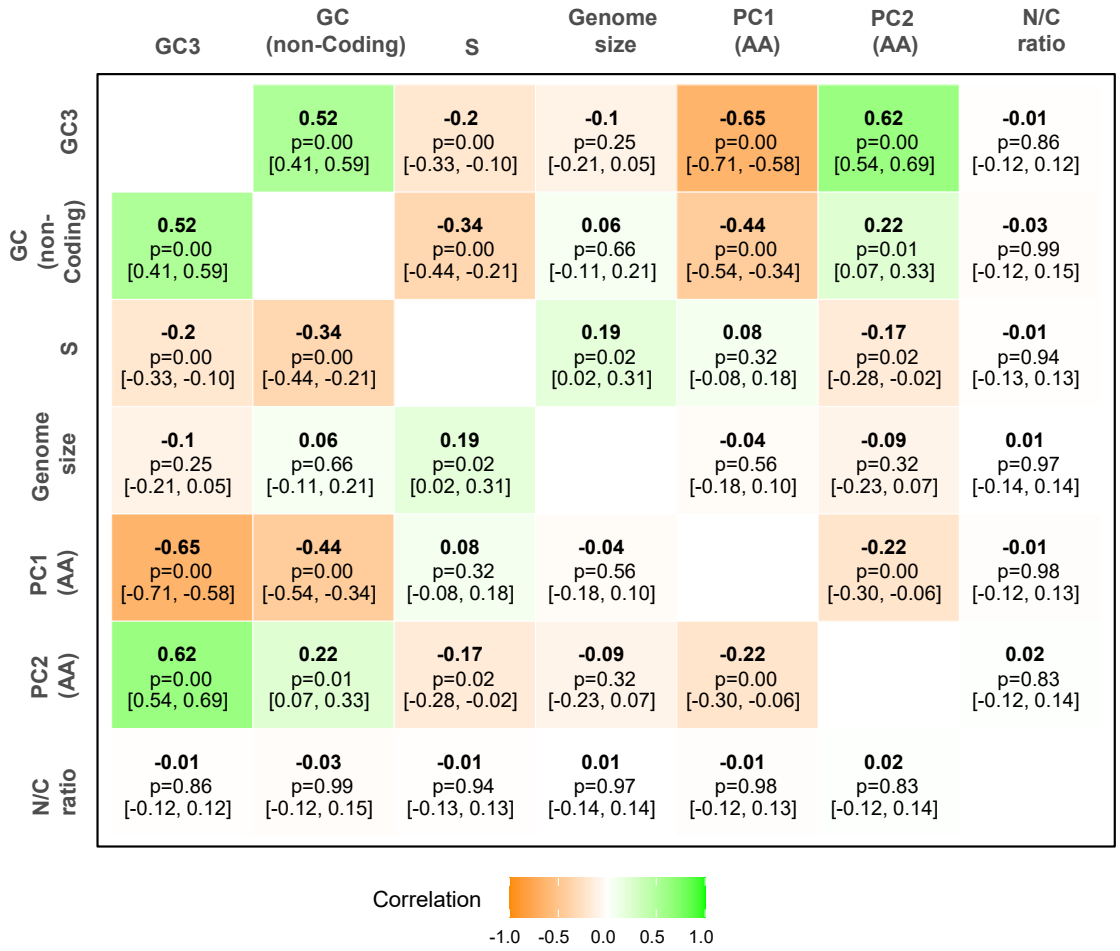

Figure 2.4: Correlation matrix of codon usage, genome features, and amino acid composition. Pairwise phylogenetic correlations between GC content at third codon positions (GC3), GC content in non-coding regions, strength of selection on codon usage bias (S), genome size, principal components PC1 and PC2 of amino acid usage, and nitrogen-to-carbon (N/C) ratio of amino acids. Values represent correlation coefficients with 95% confidence intervals and MCMC p-values. Strong correlations (positive or negative) are highlighted in green and orange, respectively, indicating relationships between nucleotide composition, codon usage, and amino acid preferences.

amino acid usage patterns (Figure 2.3B). PC1 primarily separated species based on GC content on the codons, with high-GC3 genomes enriched GC-rich amino acids (Pro, Gly, Ala, Arg) and low-GC3 genomes enriched AT-rich amino acids (Asn, Tyr, Ile), see Figure 2.5. To assess whether the patterns were linked to biochemical properties of the amino-acids, such as nitrogen-to-carbon (N/C) ratio or amino acid essentiality (measured in Drosophila melanogaster; Croset et al. 2016; Park and Carlson 2018), we examined the remaining PCA loadings. However, we found no clear patterns, suggesting that other factors, such as protein structure or functional constraints, may play a small role in shaping variation in amino acid use among species of Drosophilidae.

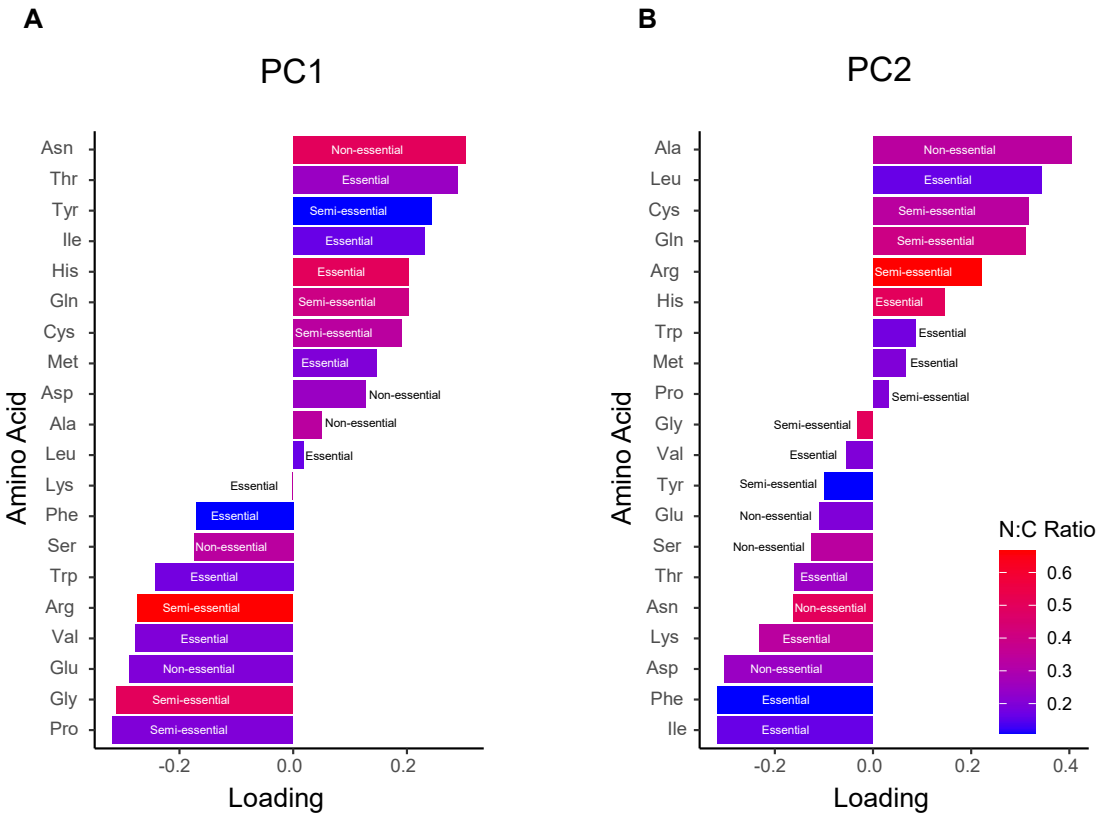

Figure 2.5: Principal component analysis (PCA) loadings of amino acid usage.
Bar plots showing the loadings of individual amino acids on the first two principal components: (A) PC1 and (B) PC2. Amino acids are coloured according to their nitrogen to carbon (N:C) ratio, with higher ratios in red and lower ratios in blue. Essentiality categories (essential, semi-essential, non-essential) are indicated alongside each bar. Positive and negative loadings reflect the relative contribution of each amino acid to the corresponding principal component.

2.5 Conclusions

This work, to generate standardized, simultaneous multi-species coding DNA sequence annotations across 304 species of Drosophilidae, forms part of an ongoing community effort working toward a comprehensive genomic study of the entire family (Kim et al. 2024). We envisage that these new annotations, orthology assignments, and multiple sequence alignments will provide a valuable resource for both single-gene and genome-wide evolutionary studies. And, along with future updates as new genomes are sequenced, this resource will support future research in studies of adaptation and functional genomics within this key model clade.

2.6 Acknowledgements

We wish to thank members of Institute for Ecology and Evolution at the University of Edinburgh for the collaborative provision of shared computational resources, James Galbraith for help with EarlGrey and TEstrainer, and Eric Lai and Garima Setia for feedback on missing gene models. We would also like to acknowledge the contributions of the broader Drosophila research community, whose collaborative efforts in genome sequencing have made this work possible.

Chapter 3

Predictors of sequence divergence and gene turnover in the Drosophila immune system

The text in this chapter is from manuscript (in preparation): Dhakad P, Obbard DJ **"Predictors of sequence divergence and gene turnover in the Drosophila immune system"**

I wrote this chapter with comments and textual edits from Prof. Darren Obbard. Thanks to Dr. Jarrod Hadfield for help and advice on the use of MCMCglmm. Thanks to Dr. Bernard Kim and Prof. Dmitri Petrov for making the Drosophila genomes data available to us.

3.1 Abstract

The evolutionary dynamics of immune genes are shaped by diverse selective pressures, yet the relative roles of gene-level traits, functional specialization, and pathway context remain poorly understood. Here, we applied Bayesian multivariate models to quantify how gene length, expression level, genetic/protein interactions, and structural features such as relative solvent accessibility (RSA) predict rates of protein sequence divergence (dN/dS) and gene turnover (λ). We find immune genes evolved significantly faster at the protein sequence level than non-immune genes, but contrary to expectations—exhibited lower gene turnover rates. Sequence evolution was strongly and positively associated with RSA, and negatively with gene length, expression, and genetic/protein-protein interactions, while gene turnover rate was largely unaffected by these factors. Functional and pathway-level analyses revealed evidence for accelerated evolution of effectors, receptors, and antiviral genes, with cGAS–STING and Toll pathways showing the highest dN/dS. Gene turnover rate was elevated only in effectors,

whereas cellular defence genes were particularly conserved. These findings highlight how immune diversification in Drosophila arises from multiple, partly independent evolutionary axes, shaped jointly by molecular constraints, functional roles, and lineage-specific pathogen pressures.

3.2 Introduction

The evolution of immune genes is thought to be shaped by a persistent arms race between hosts and their pathogens. Across animals, immune-related genes are generally reported to be among the most rapidly evolving components of the genome, exhibiting elevated rates of non-synonymous substitution, gene duplication and loss, and structural innovation that often exceed genomic background levels (Nielsen et al. 2005; Sackton et al. 2007; Obbard et al. 2009b; Shultz and Sackton 2019; Vinkler et al. 2023). This rapid evolution is often interpreted as a hallmark of coevolution, where pathogens exert recurrent selective pressure, driving adaptive changes in recognition, signalling, and effector components of the immune system (Sironi et al. 2015; Świderská et al. 2018; Velová et al. 2018; Shultz and Sackton 2019; Lazzaro et al. 2020; Davies et al. 2021).

In *Drosophila melanogaster*, the innate immune system is well characterized and has served as a foundational model for understanding both the mechanisms and evolution of immunity (Lemaitre and Hoffmann 2007), especially of invertebrates. The *Drosophila* immune system is composed of both cellular and humoral immune responses acting against pathogens. The cellular components are primarily mediated by hematocytes, and their differentiated populations are responsible for phagocytosis of microbes by plasmatocytes, and encapsulation and melanisation of larger parasites by lamellocytes and crystal cells respectively (Honti et al. 2014; Balog et al. 2021)—although there is substantial variation among species (Salazar-Jaramillo et al. 2014; Cinege et al. 2024). The humoral component consists of the Toll and Imd NF-κB pathways, while additional components include the JAK-STAT cytokine pathway, JNK pathway, as well as RNAi and cGAS–STING antiviral systems (reviewed in Westlake et al. 2024). Broadly, these different pathways respond to distinct pathogen classes. Toll signalling primarily targets Gram-positive bacteria and fungi, Imd targets Gram-negative bacteria, RNAi and cGAS-STING act against viral nucleic acids, and JAK-STAT is activated by cellular stress or wounding (Gottar et al. 2002; Wang et al. 2006; Myllymäki and Rämet 2014; Tafesh-Edwards and Eleftherianos 2020; Cai et al. 2023; Huang et al. 2023). However, the role of these pathways is not absolute. Mounting evidence points to crosstalk between immune pathways and overlapping roles in pathogen defence (Tanji et al. 2007; Nishide et al. 2019). For example,

Some AMPs (such as *Drosomycin*) are activated by both Toll and Imd pathways (Valanne et al. 2010). Moreover, infections with gram-positive and gram-negative bacteria can co-activate both pathways in a synergistic manner (Tanji et al. 2007). Beyond their classical roles, both the Toll and Imd pathways have also been implicated in antiviral immunity, Toll signalling mediated AMPs upregulated against Drosophila X virus (DXV), while Imd mediates responses to Sindbis and Cricket Paralysis viruses (Costa et al. 2009; Sabin et al. 2010).

In all pathways, several genes have been found to be rapidly evolving, potentially as a consequence of host-parasite arms races (Schlenke and Begun 2003; Obbard et al. 2006; Jiggins and Kim 2007; Sackton et al. 2007; Obbard et al. 2009a; Palmer et al. 2018). Genes acting at different steps in these pathways—including extracellular recognition proteins (such as PGRPs, GNBPs), intracellular signalling molecules (such as Relish, Dorsal), and effector genes (such as *Attacin*, *Defensin*)—differ markedly in their evolutionary dynamics. While receptor genes are undergoing rapid evolution, the evolutionary forces acting on signalling and effector genes remains more debated (Sackton et al. 2007; Hanson et al. 2016). Sackton et al. (2007) showed that signalling genes with modulatory functions are subject to positive or 'diversifying' selection (i.e. selection driving fixed differences among species), whereas effector genes tend to turnover in gene number faster. However, subsequent studies, including those in other organisms, have found that effector genes evolve rapidly by both gene duplication and positive selection (Tennessen 2005; Hollox and Armour 2008; Unckless and Lazzaro 2016; Hanson and Lemaitre 2020; Hanson et al. 2023). In fact, orthologs of some effector genes in these pathways can be difficult to identify due to their high rates of sequence divergence (Sackton and Clark 2009; Hanson et al. 2016).

Such analyses suggest that antiviral defence mechanisms may be especially prone to experiencing strong selection, as seen in vertebrates (e.g., Enard et al. 2016; Ito et al. 2020; Scheben et al. 2023). Notably, RNAi pathway components are among the most rapidly evolving genes in *D. melanogaster* and closely related species, particularly those involved in small RNA biogenesis and antiviral defence (Obbard et al. 2006; Obbard et al. 2009b; Palmer et al. 2018; but compare Hill et al. 2019). Similar patterns have recently emerged for the cGAS-STING pathway in Drosophila, where species encode variable numbers of cGLRs (cGAMP-like receptors), often exhibiting species-specific expansions and functional divergence (Cai et al. 2020; Cai et al. 2023). For example, the antiviral potency of different cGLR-generated cyclic dinucleotides (Such as $2'3'$-cGAMP, $3'2'$-cGAMP, $2'3'$-c-di-AMP) varies across species in their ability to inhibit Drosophila C virus (DCV), suggesting a lineage-specific functional diversification (Cai et al. 2023).

However, the majority of insights to date have come from a handful of species and/or immune genes, and their generality remains unclear. For example, comparisons limited to *D. melanogaster* and its close relatives have likely captured only a subset of immune adaptation, leaving broader patterns unexplored. Such comparative analyses across more distantly related clades remain scarce, but a recent study found rapid evolution of Toll signalling genes in *D. innubila*, contrasting with what was found for *D. melanogaster*, where RNAi genes show the fastest rates of evolution (Hill et al. 2019). Broad phylogenetic sampling across the Drosophilidae would offer a unique opportunity to quantify how immune gene families evolve over tens of millions of years of divergence, spanning diverse ecological niches, microbial exposures, and life histories (Kim et al. 2024; Dhakad et al. 2025a). Early analyses of a small number of immune genes already suggested lineage-specific adaptation between the melanogaster and virilis groups, suggesting that different pathogen pressures may shape distinct evolutionary trajectories (Morales-Hojas et al. 2009). Expanding both the number of species and the immune gene repertoire should improve power to detect lineage-specific adaptations and can reveal macroevolutionary trends obscured in narrow comparisons (Lažetić and Troemel 2021).

Importantly, although numerous studies have identified high rates of nonsynonymous substitution or gene family turnover in immune genes, relatively few have simultaneously or systematically examined molecular or functional features that might explain variation in evolutionary trajectories. The evolutionary rate of a proteins may be influenced only weakly by the functional role of the protein; other factors, such as gene expression level, protein structure, gene length, intron number, recombination rate and protein-protein interaction may dominate, when they are taken in account (Drummond et al. 2005; Larracuente et al. 2008; Zhang and Yang 2015; Hagai et al. 2018; Moutinho et al. 2019; Zhong et al. 2021). For example, genes with higher baseline expression levels tend to evolve more slowly, likely due to pleiotropic constraints and costs of misfolded proteins (The expression level-evolutionary rate anticorrelation; Drummond et al. 2005; Hagai et al. 2018; Zhong et al. 2021). Similarly, genes encoding structurally 'buried' or interaction-rich proteins often show stronger purifying selection due to higher functional constraint (Moutinho et al. 2019; Chaurasia and Dutheil 2022). By contrast, proteins with high relative solvent accessibility (RSA)—that is, those with many surface-exposed residues—may offer more opportunities for adaptive substitutions, particularly in immune receptors and effectors interacting directly with pathogens (Moutinho et al. 2019). A gene's number of physical or genetic interactions may constrain its evolutionary flexibility, as highly connected genes can have wider systemic effects (Pang et al. 2010; Papakostas et al. 2014; Zhang and Yang 2015). Likewise, gene length may affect both mutation target size and potential for functional modularity, which could differentially shape immune versus non-immune gene evolution (Lar-

racuente et al. 2008; Zhang and Yang 2015; Moutinho et al. 2019). Integrating these potentially important, but often ignored, predictors into models of immune gene evolution could provide a more mechanistic understanding of why some genes or sites evolve adaptively, while others do not.

To address this, we examine the evolutionary dynamics of immune gene families across 304 species of Drosophilidae. Using a curated set of immune-related genes and length- and location-matched non-immune genes ('controls'), we first estimate relative rates of protein sequence divergence (dN/dS), the proportion of sites under diversifying selection, test for evidence of positive selection, and quantify gene family turnover (λ). Then, using a mixed-model approach, we quantify the role of structural and regulatory gene features—such as relative solvent accessibility (RSA), baseline gene expression, number of gene interactions, and gene length—in predicting variation in evolutionary outcomes, regardless of immune function. At the same time, we test whether immune genes differ systematically from non-immune genes in their rates of evolution, while accounting for the gene-level predictors. Finally, we ask whether different immune classes and pathways—such as 'recognition', 'signalling', and 'effector' genes, or Toll, Imd, and RNAi pathways—exhibit distinct patterns of sequence divergence and gene turnover. Together, this study aims to reveal the general determinants of rapid protein evolution and assess how functional roles shape the tempo and mode of immune gene diversification.

3.3 Materials and Methods

3.3.1 Gene selection and orthology assignment

To investigate the evolutionary dynamics of immune gene families, we curated a list of well characterized immune-related genes in *Drosophila melanogaster* from literature searches (De Gregorio et al. 2001; De Gregorio 2002; Lindmo et al. 2008; Early et al. 2017; Troha et al. 2018; Cai et al. 2020), including the recent "The *Drosophila* immunity handbook" (Westlake et al. 2024). Where possible, we assigned each gene to a known functional class ('recognition', 'signalling', 'effector', 'antiviral') and immune pathway ('Toll', 'Imd', 'RNAi', 'cGAS-STING', 'JAK-STAT', 'MAPK'). Where membership was unclear, or not unique, we assigned immune genes to a 'Multiple' or 'Unclassified' category (Table 3.1). For each immune gene, we then identified up to four non-immune ('control') genes that were approximately matched for size and genome location. To help mitigate the impact of local genomic features (e.g. recombination rate,

chromatin accessibility—at least to the extent that synteny is maintained with *D. melanogaster* (Felsenstein 1974; Comeron et al. 2008; Charlesworth et al. 2009; Cherry 2010; Soni and Eyre-Walker 2022), the 'control' genes were required to be protein-coding, located within ±50 kb of the immune gene in the *D. melanogaster* genome, and between 0.5-2 times its length.

Pathway/Class	Receptor	Signalling	Effector	Antiviral	Unclassified	Multiple
Cellular defense	28	125	47	19	83	2
IMD	12	44	9	0	0	1
Toll	4	44	17	0	0	3
RNAi	0	1	0	9	0	0
cGAS-STING	0	0	0	0	0	0
JAK-STAT	1	18	1	1	0	0
JNK	1	1	1	0	0	1
MAPK	0	13	0	0	0	0
Multiple	0	25	6	0	2	31

Table 3.1: Immune gene classification based on immune pathways and functional classes. Total 615 immune genes grouped into 566 unique HOGs.

Using the *D. melanogaster* references, immune and 'control' gene orthogroups were then identified from our recent comparative annotation of 304 Drosophilidae species (Table 3.1; Dhakad et al. 2025a). These Hierarchical Orthologous Groups (HOGs) allowed us to compare gene family evolution across deeply diverged lineages in the family Drosophilidae. Some previous studies have chosen to limit their analyses to a subset of closely-related taxa, thereby avoiding a high proportion of noisy, potentially saturated, long branches (e.g. avoiding branches longer than ∼25 million years; Sackton et al. (2007) used just 6 species in the melanogaster group). However, as almost no lineages in the present dataset lack sampled close relatives (i.e. only 9 of 606 branches longer than ∼25 million years, mean branch length 3.7 million years), such saturation is unlikely to prove problematic.

3.3.2 Estimating rates of sequence evolution

To quantify protein-coding sequence evolution, we used the 'BUSTED' model from the HyPhy package (Murrell et al. 2015). Conditional on gene tree topology, BUSTED tests for 'episodic' diversifying selection at any site on any branch of a phylogeny (i.e. selection in favour of amino-acid change on some, but not all, branches). BUSTED returns an overall estimate of the ratio of non-synonymous to synonymous substitutions (dN/dS) per gene (Hierarchical Orthogroup; HOG) and reports whether there is evidence for a response to positive selection anywhere in the gene tree (Murrell et al. 2015). Codon alignments for each gene family (HOG) were

generated using MACSE (Ranwez et al. 2018), a codon-aware aligner specifically designed to handle the frameshifts and sequencing errors common in large, diverse datasets. To mitigate the risk of misalignment and subsequent detection of false positives in selection analyses, we applied a series of quality controls in the MACSE OMM pipeline, which uses MAFFT as an aligner to handle larger datasets (Katoh and Standley 2013). First, a pre-filter step removes long, non-homologous regions that may result from mis-annotation (such as retained introns or gene fusions). Second, HmmCleaner is used to identify and mask residues that appear misaligned at the amino acid level (Di Franco et al. 2019). These masked positions were then mapped back to the nucleotide alignment. A third post-processing filtering step was applied to eliminate patchy and isolated codons (sequences with >80% of codons masked were excluded entirely). Finally, we trimmed the extremities of the alignments, discarding poorly aligned ends until a site with at least 70% of nucleotides is reached. These steps collectively aim to reduce alignment artifacts that could otherwise inflate dN/dS estimates or lead to spurious detection of selection. However, this may come at the cost of excluding rapidly diverging regions and thus reducing the power to detect strong selection. Gene trees were reconstructed for each HOG using IQTREE2 (Minh et al. 2020), with the best-fit substitution model identified by ModelFinder (Kalyaanamoorthy et al. 2017) under the Bayesian Information Criterion. We assessed branch support using 1,000 ultrafast bootstrap replicates. These phylogenies and alignments were then supplied to BUSTED for likelihood-based detection of positive selection. BUSTED was run in MPI-parallelized mode with default settings, specifying the entire tree as foreground (to test the entire phylogeny for positive selection) for both immune and non-immune genes. We extracted the gene-wide mean dN/dS values, the proportion of sites detected to evolve under diversifying selection, and a binary indicator as to whether a gene showed evidence of episodic diversifying selection (p-value threshold ≤ 0.001). For such 'selected' HOGs, specific sites undergoing episodic diversifying selection were subsequently identified by HyPhy models 'MEME' and 'FEL' (Kosakovsky Pond and Frost 2005; Murrell et al. 2012).

3.3.3 Estimating gene turnover rate

To estimate the rate of gene turnover (gain and loss per million years, 'λ') for immune and non-immune gene families, we used CAFE5 (Computational Analysis of Gene Family Evolution; Mendes et al. 2021). CAFE models gene family evolution under stochastic birth-death process. However, observed gene copy numbers can be affected by non-biological factors such as assembly errors, variation in genome completeness, or gene family annotation artefacts. To help mitigate this, we first ran the CAFE base model to estimate an error distribution. These error estimates were then incorporated into the subsequent CAFE run to help minimise technical

noise prior to estimating ancestral family sizes and λ values (Mendes et al. 2021). CAFE requires gene families to be present at the root of the species tree. As such, families absent at the root ("orphan" or lineage-specific families) are forced to have an ancestral copy number of at least one. This constraint leads to spurious inferences of gene loss along branches where the family is truly absent, thereby distorting turnover estimates. In addition, we limited the dataset to families with a maximum observed copy number change of 10 genes across all species, as families with extreme variation in size tend to violate model assumptions and often yield unstable or non-converging likelihoods. These filtering steps reduced the dataset to 1,761 'gene families' (i.e. HOGs; 489 immune and 1,272 non-immune HOGs). While these filters were necessary for model accuracy and convergence, they likely biased the analysis toward more conserved families. Specifically, the exclusion of non-rooted families disproportionately removes fast-evolving gene families that arose de novo more recently, or underwent lineage-specific expansions (or families that are lost from a entire ancient clade). As a result, our λ estimates reflect evolutionary patterns among a relatively stable subset of gene families and likely underestimate turnover rates in more dynamic gene categories. Finally, CAFE was run under gamma model (k=2) with a Poisson distribution for gene family counts at the root to estimate the λ per family. Because CAFE5 does not explicitly return λ = 0, we also recorded a binary variable indicating whether λ was substantially greater than zero (values $\leq$3x10^{-7} were treated as effectively invariant). This indicator was included in subsequent linear models to distinguish stable gene families from those exhibiting measurable turnover.

3.3.4 Mixed-model analyses of sequence evolution and gene turnover

To evaluate how structure, expression level, gene length, number of genetic/protein interactions, and functional properties of genes predict their evolutionary dynamics, we took a multivariate 'meta-analytic' approach using the Bayesian mixed-model R package MCMCglmm (Hadfield 2010). Our goal was to jointly estimate how gene-level traits such as expression level, relative solvent accessibility (RSA), number of genetic/protein interactions, gene length, and gene function predict the aspects of molecular evolution inferred using HyPhy and CAFE (above).

We fitted 7 multivariate models (provided in Table 3.2) that included five response variables for each HOG: (i) the log-transformed dN/dS, (ii) the log-transformed proportion of sites under diversifying selection, (iii) a binary indicator of whether a gene showed evidence of episodic diversifying selection, (iv) the log-transformed rate of gene family turnover (λ), and (v) a binary indicator of whether the estimated turnover rate was clearly distinguishable from zero. These response variables reflect distinct but potentially interrelated evolutionary processes—namely,

sequence-level adaptation, copy number evolution, and episodic selection pressure. Modelling these traits jointly allowed us to quantify covariation between them and assess whether shared gene properties predict their evolution in parallel. We included fixed effects representing gene-level properties expected to influence evolutionary dynamics: gene length, baseline expression level in *D. melanogaster*, predicted RSA, and number of reported protein and genetic interactions in *D. melanogaster*. Predictors were included in the model as trait-specific fixed effects using the 'trait:' syntax in MCMCglmm, allowing each explanatory variable to have a separate effect on each evolutionary response (model syntax in Supplementary file B.5).

Category	Model	Response Variables	Focal Predictor	Family (link)
Sequence evolution	M1	log(dN/dS), log(prop)*, episodic selection (binary)	Gene type (Immune vs Non-immune)	Gaussian, Gaussian, Threshold (probit)
	M2	log(dN/dS), log(prop)*, episodic selection (binary)	Immune Class	Gaussian, Gaussian, Threshold (probit)
	M3	log(dN/dS), log(prop)*, episodic selection (binary)	Immune Pathway	Gaussian, Gaussian, Threshold (probit)
Gene turnover	M4	log(λ), non-zero λ (binary)	Gene Type (Immune vs Non-immune)	Gaussian, Threshold (probit)
	M5	log(λ), non-zero λ (binary)	Immune Class	Gaussian, Threshold (probit)
	M6	log(λ), non-zero λ (binary)	Immune Pathway	Gaussian, Threshold (probit)
For correlation	M7	log(dN/dS), log(prop)*, episodic selection (binary), log(λ), non-zero λ (binary)	Gene type (Immune vs Non-immune)	Gaussian, Gaussian, Threshold (probit), Gaussian, Threshold (probit)

Table 3.2: Summary of multivariate Bayesian models used to investigate evolutionary dynamics of immune gene families.
All models included fixed effects for gene length, expression level (FPKM), relative solvent accessibility (RSA), and interaction count; a random effect for positional ID (immune HOG) linking each immune gene with its matched non-immune control (unstructured covariance); and unstructured residual covariance. *Proportion of sites under diversifying selection (on log scale).

RSA was estimated using predicted protein structures of *D. melanogaster* genes obtained from AlphaFold, SWISS-MODEL (Expasy), and ModPipe (Pieper et al. 2014; Waterhouse et al. 2018; Abramson et al. 2024). Solvent accessible surface area (SASA) was computed from the predicted PDB structures using DSSP (https://github.com/PDB-REDO/dssp.git; Kabsch and Sander 1983), and normalized by dividing each residue's SASA by its reference maximum accessible surface area for a fully exposed residue of that amino acid type (Tien et al. 2013). Our script for RSA estimation was adapted from Sydykova et al. (2018). For each HOG, we calculated the mean RSA across all residues in the corresponding *D. melanogaster* protein. Gene expression was summarized as the mean FPKM (Fragments Per Kilobase of exon per Million mapped fragments) value of each *D. melanogaster* gene across developmental stages and tissues, using expression data obtained from FlyBase (https://flybase.org/). We ranked the HOGs according to mean FPKM values, resulting in 40 expression categories. To quantify protein-protein and genetic interactions, we obtained both protein-protein and genetic interactions for *D. melanogaster* genes from the Drosophila Interactions Database (DroID v2018_08; Murali et al. 2011). The median number of interactions across all *D. melanogaster* genes in a given HOG was used as the interaction count for that HOG.

To reduce the impact of any variation associated with genomic location, immune-related genes (i.e. HOGs) were assigned to 'gene groups' along with their neighbouring 'control' genes, and gene-group ID was fitted as a random effect with an unstructured variance-covariance matrix across traits (analogous to a 'paired' test). Residual variances were also modelled with an unstructured covariance matrix to capture correlations among traits not explained by the predictors. Our Bayesian priors for fixed effects were chosen to be weakly informative, and a weakly informative parameter-expanded prior was used for the random effects (Supplementary file B.6; Hadfield 2010).

To investigate whether immune genes differed from non-immune genes after accounting for structural and regulatory gene-level features, we fitted models (M1 and M4; Table 3.2) that included gene type (immune or non-immune) as a fixed effect. And, to assess variation within immune genes, we fitted additional models (M2, M3, M5 and M6; Table 3.2) that included either immune functional class (effector, signalling, recognition, antiviral) or pathway (Toll, Imd, JAK-STAT, RNAi, cGAS-STING) as categorical predictors. All of these models (M1-M7) retained the same structural and regulatory covariates as fixed effects and were fitted separately for sequence evolution statistics (log(dN/dS), log proportion of sites under selection, binary episodic selection) and for gene turnover statistics (log(λ) and binary non-zero λ; Table 3.2). In all models, the fixed effects were allowed to vary across traits, enabling us to quantify how immune category membership influences distinct evolutionary processes while adjusting for

gene length, expression, RSA, and interactions. All models were run for 2.1 million MCMC steps, with a 100,000 step burn-in and thinning interval of 100, resulting in a posterior sample of 20,000 steps. Posterior convergence was assessed visually and through effective sample sizes. Summaries of all models' variables and predictors are presented in Table 3.2, and output from each model can be found in Supplementary file B.6.

Pairwise contrasts among immune classes or pathways were conducted by subtracting the posterior sample for one level from that of the other, calculating a credible interval for the difference, and estimating 'pMCMC' values from the resulting contrasts (i.e., the fraction of the posterior density in the smaller tail overlapping zero). All statistical analysis analyses were performed using R Statistical Software (v4.5; R Core Team 2025) and figures were generated using ggplot2 (Wickham 2016).

3.3.5 Data availability

All code used for MCMC model fitting, statistical analyses, and figure generation are available at github.com/DhakadPankaj/Gene_Family_Evolution. Sequence alignments and gene trees used in this study can be accessed at 10.5281/zenodo.15016917.

3.4 Results and Discussion

To investigate the evolutionary dynamics of immune gene families across Drosophilidae, we analysed a curated set of *D. melanogaster* immune genes and their homologs in 304 species, alongside size- and location-matched non-immune genes ('controls'). We estimated rates of protein sequence evolution using codon-based alignments and gene trees to compute gene-wide dN/dS values and to detect signals of episodic diversifying selection (i.e. positive selection driving divergence among species) using the approach implemented in HyPhy 'BUSTED' (Murrell et al. 2015). We estimated gene family turnover rates (duplication and loss per million years) using CAFE5 (Mendes et al. 2021), under a stochastic birth-death model calibrated to an approximately-dated species phylogeny.

To identify the factors shaping immune gene evolution, we fitted a series of Bayesian multivariate generalized linear mixed models using the MCMCglmm R package (Hadfield 2010). These models simultaneously inferred the predictors of five evolutionary responses: the log-transformed dN/dS ratio, log-transformed gene turnover rate, $\log(\lambda)$, a binary indicator of whether a gene family exhibits non-zero turnover, the proportion of sites (codons) inferred to evolve under diversifying selection, and the presence or absence of statistically 'significant'

evidence for episodic selection (see Materials and Methods). We included four fixed effect predictors based on previously reported determinants of the rate of adaptive evolution, including relative solvent accessibility (RSA), baseline gene expression, number of genetic/protein interactions, and gene length (Larracuente et al. 2008; Zhang and Yang 2015; Moutinho et al. 2019; Chaurasia and Dutheil 2022). This integrated framework allowed us to test which gene features best predict patterns of sequence evolution and gene turnover, and whether immune genes exhibit distinct evolutionary dynamics relative to non-immune 'control' genes, after accounting for gene-level predictors. We further examined how these patterns differ across different immune functional classes and immune pathways, after statistically controlling for gene-level structural and regulatory constraints.

3.4.1 Structural and gene-level features predict patterns of molecular evolution and turnover

Regardless of immune function, understanding how gene-level features and protein structure constrain or facilitate molecular evolution is essential to interpret patterns of sequence divergence and gene family dynamics (Duret and Mouchiroud 2000; Jordan et al. 2005; Ingvarsson 2006; Larracuente et al. 2008; Zhang and Yang 2015; DuBose and Roode 2024). We evaluated the predictive role of four features—gene length, expression level, number of protein/gene interactions, and RSA—on five evolutionary responses: global dN/dS, gene turnover rate, probability of family size variation, the proportion of sites inferred to evolve under diversifying selection, and the probability that a gene is inferred to evolve under 'episodic' selection.

Gene length is well known to correlate with rates of evolution, with longer genes evolving more slowly than shorter ones (Yang and Gaut 2011; Soni and Eyre-Walker 2022). Several mechanisms may underlie this pattern, including more potential sites for deleterious mutations to occur, increased interference from linked selection constraining the efficacy of natural selection (Hill-Robertson effects), greater functional constraint due to more interaction partners or regulatory complexity in longer genes, and the potential for faster protein synthesis in shorter genes (Loewe and Charlesworth 2007; Larracuente et al. 2008; Zhang and Yang 2015). Consistent with this, we found an increase in 1kb of gene length is associated with a 4.9% reduction in dN/dS (95% HPD CI [-7.1, -2.7], $p<0.001$) and a $\sim$9% decrease in proportion of sites under diversifying selection (CI [-12.7, -5.2], $p<0.001$; see Supplementary file 1 for model summary). Interestingly, despite showing reduced dN/dS and fewer positively selected sites, longer genes were more likely to be inferred as evolving under episodic diversifying selection (as assessed by the BUSTED model), with each 1 kb increase in gene length associated with a 16% increase in the probability of detecting such selection ($p<0.001$). This association may

reflect the increased statistical power of selection tests in longer genes, which provide more codon sites where diversifying selection could be detected, even if most of the gene evolves under purifying or neutral selection (Yang and Dos Reis 2011; Murrell et al. 2015). In contrast, gene length was not significantly associated with the rate of gene turnover (λ) among genes showing any variation in family size (p = 0.57; Supplementary file 1). However, longer genes were significantly less likely to exhibit any family size variation at all, with a 1.9% decrease in the probability of observing any turnover, per additional kilobase of coding sequence (p = 0.002). Together, these results suggest that longer genes are not only more constrained at the sequence level, but also less prone to gene duplication and loss, reinforcing the notion that gene length imposes broad constraints on both molecular and genomic evolution.

Gene expression level, widely considered a major constraint on protein evolution (Drummond et al. 2005), was associated with measures of sequence evolution but not copy number variation. As expected, genes with higher baseline expression in *D. melanogaster* showed a 1.3% decrease in dN/dS per expression level (CI [-1.6, -0.1], p<0.001) and a 1.4% reduction in the proportion of sites under diversifying selection per expression level (CI [-2.0, -0.08], p<0.001). These findings are consistent with previous observations in which highly expressed genes experience stronger purifying selection, likely due to constraints imposed by protein misfolding, translational error sensitivity and high pleiotropic effects (Drummond et al. 2005; Zhang and Yang 2015; Bédard et al. 2022). However, we also detected a modest positive association between expression level and the probability of detecting episodic selection, with 0.4% increase per expression level (p<0.001). This suggests that even highly expressed genes can be targets of occasional bursts of adaptive evolution, potentially reflecting context-dependent pressures—for example, tissue-specific or inducible expression under stress or infection (Gu and Su 2007; Larracuente et al. 2008; Kryuchkova-Mostacci and Robinson-Rechavi 2015; Stanley and Kulathinal 2016).

In our analysis, relative solvent accessibility (RSA) emerged as the strongest overall predictor of molecular adaptation. Genes encoding proteins with more surface-exposed residues (i.e., higher RSA) showed markedly elevated rates of sequence evolution: dN/dS increased by 17.2% per standard deviation in RSA (CI [13.0, 21.0], p<0.001), and the proportion of sites under diversifying selection increased by 24.2% (CI [16.5, 32.1], p<0.001). RSA also predicted the probability of detecting episodic selection, with a 3.4% increase in probability per standard deviation (CI [1.7, 5.0], p < 0.001). These findings align with studies that integrated structural and evolutionary analysis, showing that surface-exposed residues evolve faster due to weaker structural constraint and increased functional accessibility (Moutinho et al. 2019; Chaurasia and Dutheil 2022). In contrast, RSA was not significantly associated with gene turnover rate (λ)

among those families that showed some copy number variation (p = 0.35). However, RSA was negatively associated with the probability of any gene family size variation, showing a 3.6% decrease per standard deviation (CI [-5.2, –2.0], p<0.001). This suggests a potential trade-off; while surface-exposed residues are more likely to undergo adaptive sequence change, the genes encoding such proteins are less prone to duplication or loss, possibly due to dosage sensitivity and functional pleiotropy.

As expected, genes with a higher number of genetic/protein interaction partners were under stronger evolutionary constraint. Specifically, each additional interaction was associated with a 0.41% decrease in dN/dS (CI [-0.51, -0.31], p<0.001) and a 0.44% reduction in the proportion of sites under diversifying selection (CI [-0.62, -0.26], p<0.001). These findings are consistent with the 'centrality–lethality' hypothesis, which posits that highly connected proteins (network hubs) evolve more slowly due to their essential roles and pleiotropic effects on multiple cellular processes (Fraser et al. 2002; Hahn and Kern 2005; but see Jordan et al. 2003 and Mekic et al. 2024). In contrast, interaction count had no significant effect on gene turnover rates among variable families (p = 0.27), nor on the likelihood of family size variation (p = 0.87).

3.4.2 Immune genes exhibit elevated sequence divergence, but lower turnover compared to non-immune genes

After accounting for gene length, baseline expression, number of genetic/protein interactions, and relative solvent accessibility (RSA), we found that immune genes evolve significantly faster at protein sequence level than non-immune genes, but exhibit lower rates of gene turnover (Figure 3.1). Immune genes had a mean dN/dS of 0.10, compared to 0.07 for non-immune genes—both slightly higher than estimates from *D. melanogaster* and its closest relatives reported by Sackton et al. (2007). Including gene-level predictors strengthened this immune/non-immune contrast, suggesting that part of the immune-specific signal was previously masked by confounding factors such as expression level or structural constraint. The 23.3% higher dN/dS of immune genes (95% HPD CI [13.8, 32.7], p<0.001; Figure 3.1; Supplementary file B.1) was mirrored by elevated signatures of positive selection. Immune genes had a 26.3% higher proportion of sites under diversifying selection (CI [8.1, 44]; p=0.001; Figure 3.1) and an 8.8% increase in the probability of detecting episodic selection (CI [3.2, 14.6], p <0.001), indicating rapid adaptive evolution. These results are consistent with previous findings that immune genes are among the most rapidly evolving in many taxa (Viljakainen et al. 2009; Harpur and Zayed 2013; Scheben et al. 2023), including Drosophila (Sackton et al. 2007; Obbard et al. 2009a; Shultz and Sackton 2019).

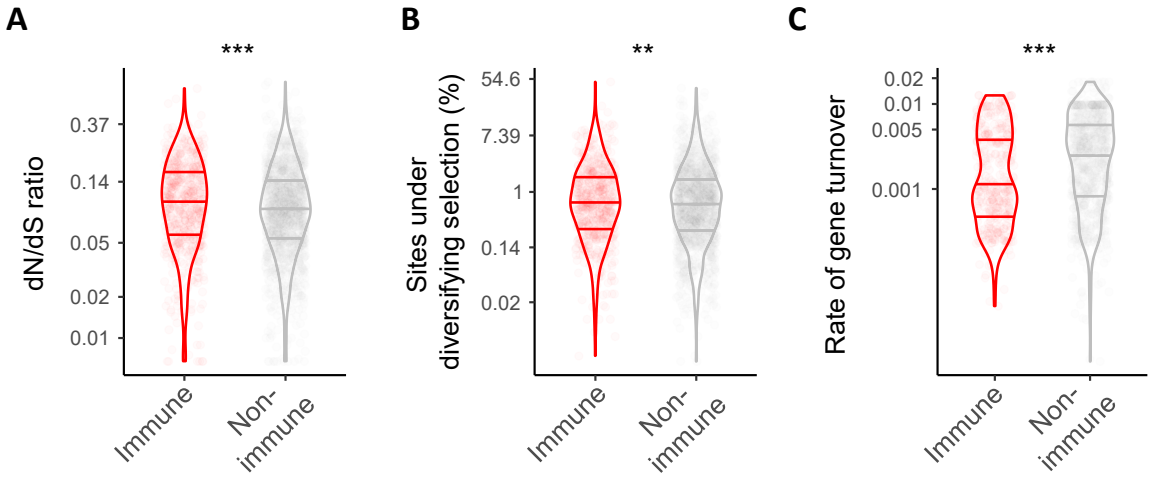

Figure 3.1: Comparative evolutionary rates of immune and non-immune genes across species of Drosophilidae.
Violin plots show (A) the non-synonymous to synonymous substitution rate ratio (dN/dS), (B) the proportion of sites (codons) under episodic diversifying selection (as inferred from BUSTED), and (C) the estimated rate of gene turnover (λ, births and deaths per million years). The y-axes are plotted on a log scale for visualization but labelled with untransformed values. Immune genes (red) show significantly higher dN/dS and a greater proportion of sites under positive selection, but lower turnover rates, compared to position- and size-matched non-immune controls (grey). The plotted points depict raw estimates for each HOG, not GLMM model fits, but 'significance' levels are derived from the Bayesian MCMCglmm models: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

Across the Drosophilidae as a whole, immune genes were also more likely to exhibit detectable variation in gene family size, with a 6.2% increase in the probability detecting some variation in copy number (CI [2.4, 10], p = 0.001; Supplementary file B.1). However, among those genes with some variation in gene family size, immune genes had a significantly lower estimated turnover rate (λ = 0.003) compared to non-immune genes (λ = 0.005), a 44.2% reduction (CI [-55.2, -32.5]; p<0.001; Figure 3.1; Supplementary file B.1). This finding is somewhat unexpected, as many previous studies have emphasized the role of gene duplication and loss in immune gene evolution (Salazar-Jaramillo et al. 2014; Levine et al. 2016; Crysnanto and Obbard 2019). However, these studies were restricted to a small subset of Drosophila species, where recent duplications and losses are more visible, whereas over deeper evolutionary timescales immune gene families may be relatively stable, punctuated by lineage-specific bursts of expansion and contraction.

3.4.3 The role of functional class in immune gene evolution

To examine how evolutionary dynamics vary across immune gene functions, we grouped immune genes into six functional classes: 'receptors', 'effectors', 'signalling', 'antiviral', genes involved in 'multiple' roles, and other 'unclassified' immune-related genes. We found substantial heterogeneity in evolutionary rates among these functional classes. Compared to non-immune genes, several immune classes exhibited elevated dN/dS values. Receptors, effectors, antiviral, and unclassified genes all showed significantly higher dN/dS than non-immune genes (mean dN/dS = 0.07 vs mean dN/dS: receptor = 0.11, effector = 0.12, antiviral = 0.13, all comparisons p<0.001 ; Figure 3.2; Supplementary file B.2). Among immune classes, antiviral proteins evolved most rapidly, with significantly higher dN/dS than signalling proteins (mean dN/dS = 0.13 vs 0.07, p<0.001), consistent with earlier studies highlighting the rapid evolution of antiviral pathways genes (Figure 3.2; Obbard et al. 2006; Obbard et al. 2009b; Palmer et al. 2018). There was also significant variation among immune classes, for example effectors (including AMPs) and receptors also had higher dN/dS than signalling proteins (mean difference from 'signalling' for both = 0.04, p < 0.001), in line with their direct roles at the host–pathogen interface resulting in high substitution rates (Figure 3.2; Lazzaro 2005; Sackton et al. 2007; Unckless and Lazzaro 2016). In contrast, dN/dS of signalling proteins were lowest among immune classes and indistinguishable from non-immune proteins (mean dN/dS = 0.073 vs 0.074, p = 0.85), likely reflecting their involvement in conserved pathways with pleiotropic functions. Patterns of the proportion of sites under detectable 'diversifying' (i.e. positive) selection mirrored some, but not all of these trends in dN/dS. Effectors, and unclassified genes had significantly higher proportions of sites under diversifying selection compared to non-immune

genes (Figure 3.2). Surprisingly, antiviral genes did not differ from non-immune genes in the proportion of positively selected sites, despite their high overall dN/dS (Supplementary file B.2). This pattern may reflect weaker purifying constraint on antiviral genes, inflating overall dN/dS without producing strong, recurrent signals of site-specific adaptation.

Gene turnover dynamics also varied across functional classes, though in less clear-cut ways. Relative to non-immune genes, all immune genes classes except effectors and genes with 'multiple' roles had significantly lower rates of gene turnover (Figure 3.2, Supplementary file B.2). Among immune genes, effectors exhibited the highest turnover rates, being duplicated or lost faster than signalling (mean λ = 0.006 vs 0.004, p = 0.001), receptor (mean λ = 0.006 vs 0.004, p = 0.006), or antiviral genes (mean λ = 0.006 vs 0.004, p = 0.026). These results highlight different axes of gene evolution in functional classes of immune genes; effector genes appear to diversify through both duplication and coding sequence change, antiviral and receptor genes have predominantly adapted via coding sequence change only and signalling genes are substantially more conserved across both dimensions.

3.4.4 The role of pathways in immune gene evolution

To determine whether immune pathways (as opposed to classes, above) have experienced distinct evolutionary pressures across Drosophilidae, we compared the dN/dS ratio, the proportion of sites under detectable episodic diversifying selection, and gene turnover rates across immune pathways, using a GLMM approach as above. We found substantial variation in evolutionary rate among immune pathways (Figure 3.3, Supplementary file B.3). The cGAS-STING pathway genes had the highest dN/dS, more than twice that of non-immune genes (mean dN/dS = 0.2 vs 0.074, p = 0.02) and significantly higher than all other immune pathways except Toll, RNAi, and Imd (mean dN/dS: Toll = 0.12, RNAi = 0.1, Imd = 0.08, all $p < 0.05$; Figure 3.3; Supplementary file B.3). Surprisingly, Toll pathway genes—often reported to be more conserved component of insect immunity—had marginally higher dN/dS than the Imd pathway (mean diff. = 0.03, CI [0.007, 0.06], p = 0.04) and slightly higher (but not significantly) than RNAi genes (p = 0.65). This contrasts with previous reports from *D. melanogaster*-*D. simulans* lineages, which emphasized rapid evolution in Imd and RNAi (Sackton et al. 2007; Obbard et al. 2009b), but is consistent with broader phylogenetic survey (e.g. quinaria group) showing accelerated Toll evolution in mushroom-breeding species (Hill et al. 2019). These patterns suggest that the relative rate of pathway evolution may be shaped by lineage-specific pathogen pressures. Pathways involved in broader cellular signalling, such as MAPK and genes classified in "multiple" immune pathways, had the lowest dN/dS (0.06 and 0.07, respect-

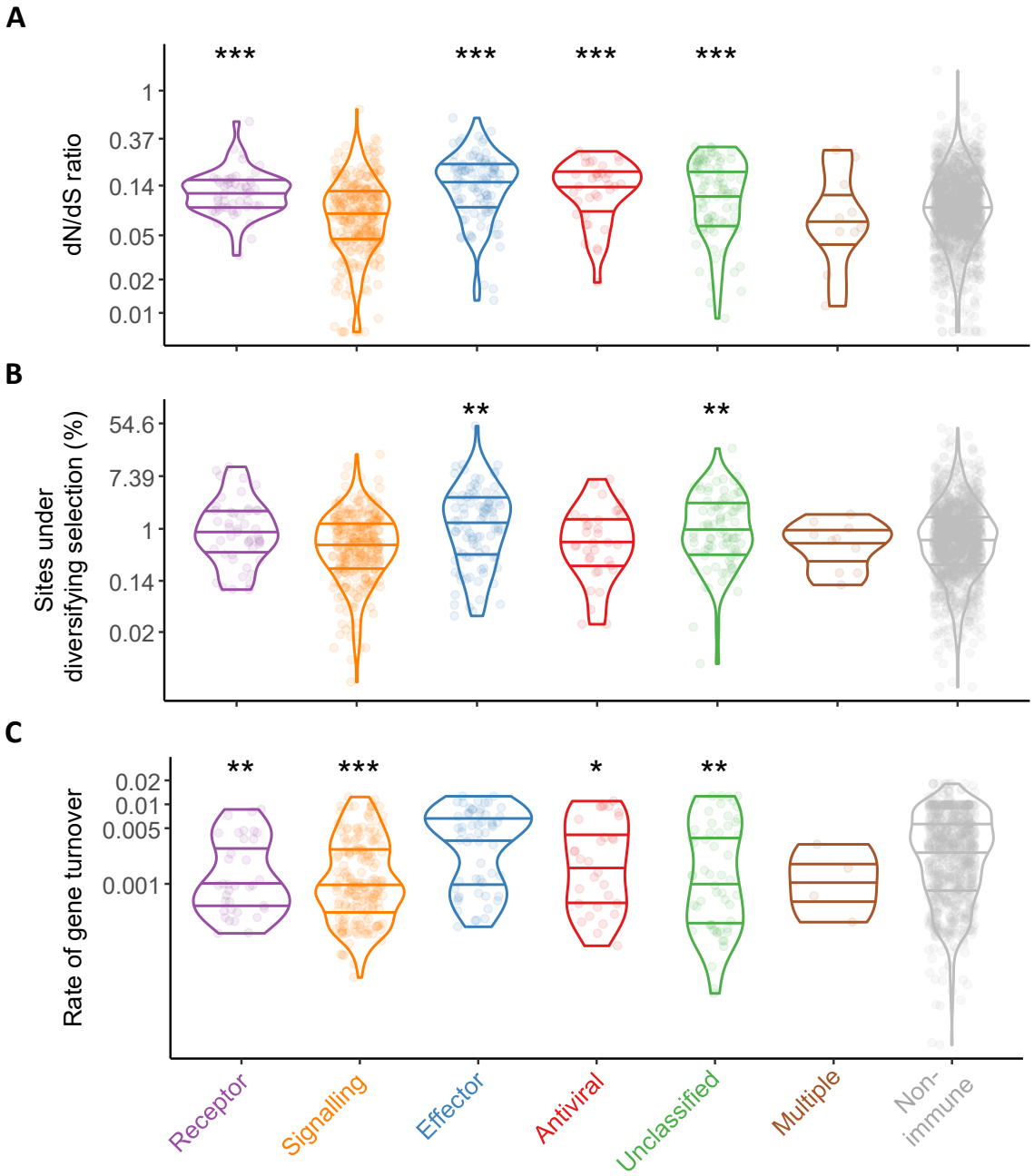

Figure 3.2: Variation in evolutionary rates among immune functional classes.
Violin plots show (A) dN/dS, (B) proportion of sites (codons) under episodic diversifying selection, and (C) gene turnover rate (λ, per million years) for receptors (purple), signalling components (orange), effectors (blue), antiviral genes (red), unclassified immune genes (green), and genes assigned to multiple functional categories (brown), compared to non-immune controls (grey). All y-axes are plotted on a log scale but labelled with untransformed values. The plotted points depict raw estimates for each HOG, not GLMM model fits, but 'significance' levels are derived from the Bayesian MCMCglmm models: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***). The p-values are reported with respect to non-immune genes.

ively), consistent with their additional functions in conserved developmental and metabolic processes (Shilo 2014). Despite these differences in overall dN/dS, the proportion of codon sites evolving under detectable episodic diversifying selection did not vary markedly among pathways (Figure 3.3, Supplementary file B.3).

Gene turnover rates were also similar among pathways, except for cellular defence genes, which had significantly lower turnover rate than non-immune genes (mean λ = 0.002 vs 0.005, $p < 0.001$; Figure 3.3, Supplementary file B.3). This stability may reflect functional constraints on phagocytic and encapsulation-related genes, where dosage balance and structural complexity could limit the retention of duplicates. Supporting this idea, a comparative study of haemopoiesis pathway genes and those differentially expressed during the encapsulation response found that haemopoiesis-associated genes are highly conserved and present in Drosophila species, regardless of their resistance phenotype (Salazar-Jaramillo et al. 2014). Such conservation suggests that the core machinery of cellular immunity evolves under strong stabilizing selection, even across lineages facing diverse pathogen pressures, in contrast to recognition and effector roles, where copy number changes are more frequent.

3.4.5 Some individual genes may be hotspots of rapid adaptive evolution

Regardless of pathway or role, some immune genes may be targets of recurrent strong selection, i.e. potential 'coevolutionary hotspots' (Jiggins and Kim 2007). Interestingly, dN/dS and gene turnover were uncorrelated after accounting for gene-level predictors (Figure B.1), suggesting that protein sequence divergence and copy number dynamics represent largely independent axes of immune gene evolution. Thus, to investigate whether specific immune genes individually exhibit distinct evolutionary trajectories—favouring sequence-level adaptation, gene copy number change, or both—irrespective of their functional class or pathway, we examined pairwise relationships among dN/dS, gene turnover rate (λ), and the proportion of sites (codons) under episodic diversifying selection (Figure 3.4). Genes in the top 2.5% for each metric were considered, with exceptionally high dN/dS (>0.25), rate of gene turnover (>0.107 gain/loss per million years per gene), or widespread site-level adaptation (proportion of sites >6.7%). Only two genes, *Srg1* (Sting-regulated gene 1) and *sid* (Stress induced DNase) ranked in top 2.5% for both dN/dS and λ. Three other genes, *bam* (bag of marbles), *CG14957*, and *CG6357* ranked in top 2.5% for both dN/dS and the proportion of positively selected sites. The *CG9733* was the only gene in the top 2.5% for both λ and proportion of selected sites (Figure 3.4). This suggest that concurrent sequence evolution and copy number change is rare in immune genes. Under all three metrics, genes involved in cellular defence

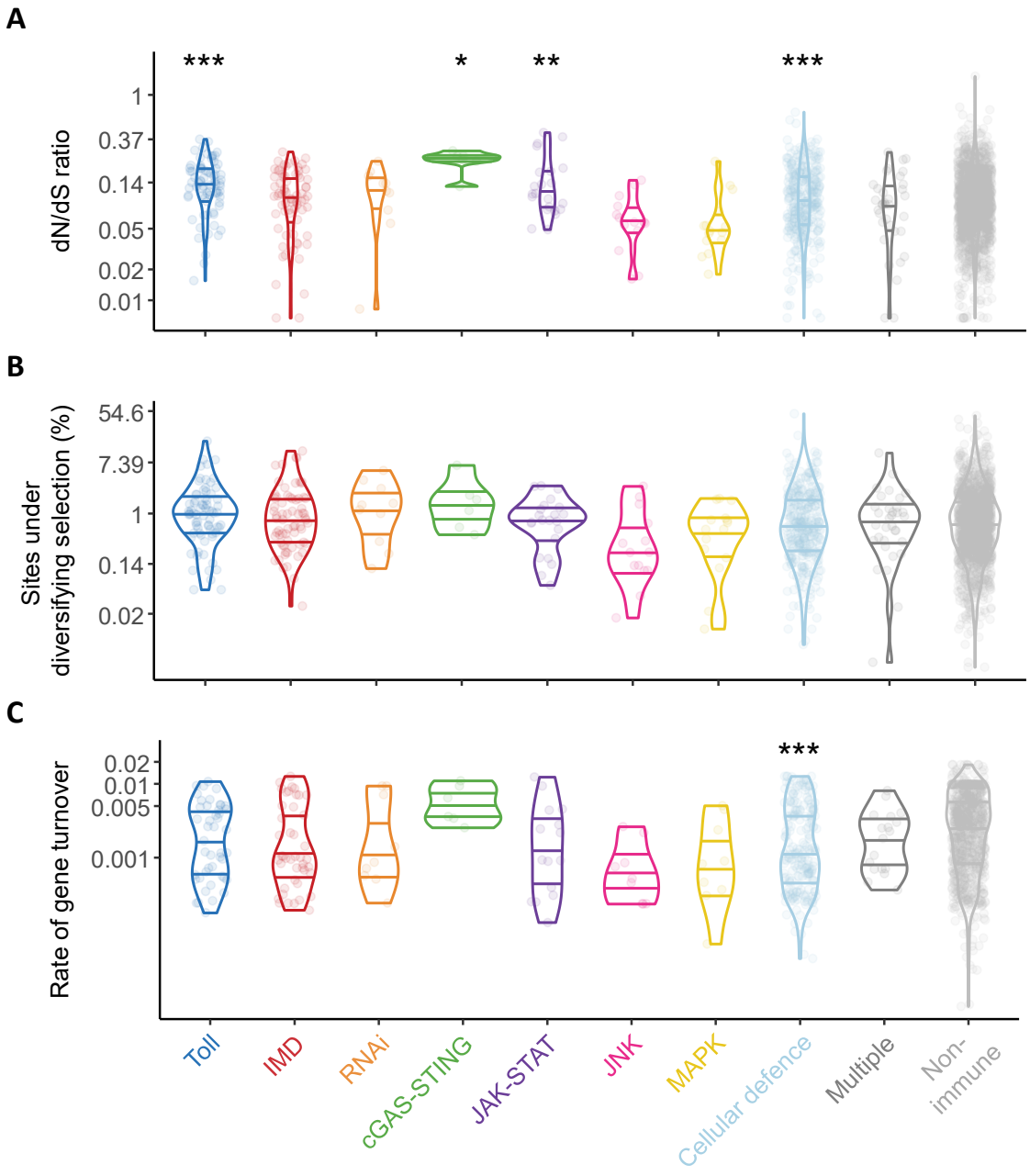

Figure 3.3: Variation in evolutionary rates among immune pathways.
Violin plots show (A) dN/dS, (B) proportion of sites (codons) under episodic diversifying selection, and (C) gene turnover rate (λ, per million years) for genes assigned to major immune pathways, compared to position- and size-matched non-immune controls. All y-axes are plotted on a log scale but labelled with untransformed values. The plotted points depict raw estimates for each HOG, not GLMM model fits, but 'significance' levels are derived from the Bayesian MCMCglmm models: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***). The p-values are reported with respect to non-immune genes.

were over-represented (20/32), followed by those in the Imd (5), Toll (4), JAK–STAT (2), and cGAS–STING (1) pathways. Strikingly, RNAi pathway genes (*Ago2*, *R2D2*, and *Dcr2*) did not rank in the top 2.5% of any metric, despite previous reports of rapid RNAi evolution in *D. melanogaster* and close relatives (Figure 3.4; Supplementary file B.4; Obbard et al. 2006; Obbard et al. 2009b; Palmer et al. 2018). This suggests that such acceleration may be lineage-specific, rather than pervasive across Drosophilidae (Hill et al. 2019). Likewise, all functional immune classes were equally represented among fastest evolving immune genes, except for 'antiviral' class (only one gene). This suggests that while antiviral pathway genes do show elevated rates of evolution on average, they do not individually stand out as the fastest evolving elements of the immune system across the whole of the family Drosophilidae (Hill et al. 2019).

In contrast to some previous studies in *Drosophila*, we found evidence for adaptive sequence evolution in AMPs. For example, *Defensin* and *IM18* (Paillotin) carried seven and three positively selected sites, respectively, most of which were in the mature functional peptide (5 out of 7 for *Defensin* and all sites for *IM18*, Figure B.2). This supports growing evidence that some AMPs in Drosophila are also experiencing positive selection (Unckless and Lazzaro 2016, also in other insects: Viljakainen and Pamilo 2008; Harpur and Zayed 2013; Erler et al. 2014). Other AMPs, such as *cecropins* and *lysozymes*, instead showed high turnover rates. The *cecropins* in particular have undergone multiple independent expansions and losses across the Drosophila (Ramos-Onsins and Aguadé 1998; Sackton et al. 2007). Interestingly, this AMP family is ancient in Diptera but appears to have been lost entirely from the subfamily Steganinae, with only truncated copies recovered in two *Amiota* species (Figure B.3; Hultmark 1993).

Several signalling genes also showed evidence for rapid evolution. *Bam*, *serpins*, *Pten*, and *et* (eyetransformer) had high dN/dS, with *bam* and *grnd* also ranking in the top 2.5% for the proportion of positively selected sites (Figure 3.4; Supplementary file B.4). The gene *bam* is essential for germline stem cell differentiation and gametogenesis and has previously been shown to evolve adaptively in *D. melanogaster* group species, likely driven by *Wolbachia* infections (Flores et al. 2015; Bubnell et al. 2022). Our analysis revealed 13 positively selected codons across at least 65 branches of the *bam* gene tree, confirming repeated protein diversification across Drosophilidae (Figure B.2; Bubnell et al. 2022).

Receptors were also well-represented among the most rapidly evolving genes, including *PGRP-LF* and *PGRP-SB2*, as well as four phagocytic receptors—*Tep*, *Sr-C*, *NimC1*, and *NimB3*. The high representation of phagocytic receptors suggests that their microbial targets may be structurally more variable than the ligands of *PGRPs* or *GNBPs*, driving repeated host–pathogen co-evolution. Among these, two families stood out for their particularly strong adaptive sig-

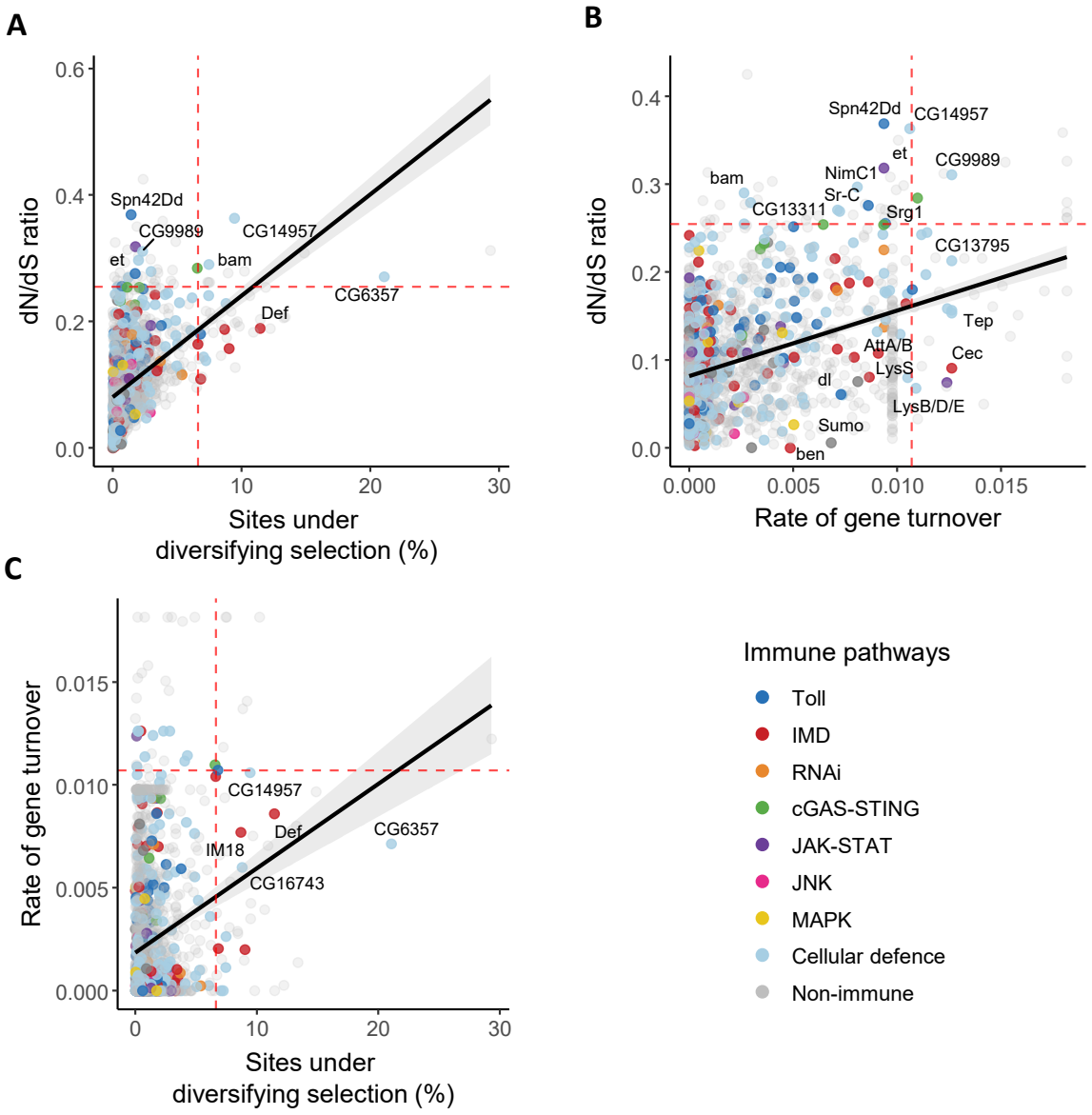

Figure 3.4: Pairwise relationships among evolutionary metrics for immune genes.
(A) dN/dS versus proportion of sites under episodic diversifying selection, (B) dN/dS versus gene turnover rate (λ), and (C) proportion of sites versus gene turnover rate. Red horizontal and vertical lines indicate the top 2.5% thresholds for each metric. Each point represents an immune gene, coloured according to its assigned immune pathway.

natures: thioester-containing proteins (*Tep*) and scavenger receptors of class C (*Sr-C*). *Tep* genes (*Tep1* and *Tep2*) harboured 148 codon sites under episodic diversifying selection, while *Sr-C* genes (*Sr-CI*, *Sr-CII*, *Sr-CIII* and *Sr-CIV*) had 111 such sites. In Tep proteins, selection sites clustered around conserved MG2/3, TED, and A2M domains (Figure 3.5; Sackton et al. 2007), which mediate pathogen opsonization and binding to pathogen surface, facilitating their phagocytosis (Williams and Baxter 2014; Shokal and Eleftherianos 2017). In Sr-C proteins, selection sites clustered in Sushi (also known as Complement control protein), MAM and Somatomedin B domains (Figure 3.5; Lazzaro 2005), which are critical for microbial recognition and endocytosis in plasmatocytes (Zani et al. 2015). Notably, *Tep* genes had markedly higher gene turnover rates (λ = 0.012) than most immune families, reflecting repeated duplication and loss events across the phylogeny. For example, the *D. melanogaster* paralogs *Tep1* and *Tep2* originated before the *melanogaster-ananassae* split, whereas other lineages—such as the *obscura*, *willistoni–saltans*, *repleta–virilis*, and Hawaiian and *Scaptomyza* clades—either retained a single ancestral copy or experienced independent expansions (Figure 3.5). By contrast, Sr-C receptors showed relatively stable copy numbers (λ = 0.005) but higher average dN/dS (0.27), consistent with strong, ongoing amino acid divergence in pathogen-binding domains (Figure 3.5). These patterns likely reflect differences in the functional and structural constraints of the two receptor types. Tep's are soluble opsonins in the haemolymph, where variation in pathogen communities could perhaps favour the gain or loss of paralogs, expanding the repertoire of pathogen-binding specificities. In contrast, Sr-C proteins (*Sr-CI* and *Sr-CII*) are membrane-bound phagocytic receptors whose turnover may be more constrained by the need to maintain conserved transmembrane and cytosolic domains, that are essential for signalling and endocytosis (Sojo et al. 2016). As a result, adaptive change in *Sr-Cs* is more likely to occur through amino acid substitutions at extracellular binding domains rather than through changes in copy number. Together, these gene-level case studies illustrate how immune gene families can adapt through partially overlapping, but distinct, combinations of sequence level adaptation and gene turnover even within a shared functional context.

3.5 Conclusions

Our comparative analysis of 304 Drosophilidae genomes reveals that different axes of immune gene evolution, protein sequence divergence and gene turnover are shaped by overlapping, but partly independent, sets of predictors. Relative solvent accessibility (RSA) emerged as the strongest positive predictor of sequence evolution, consistent with the idea that surface-exposed residues, particularly at functional interfaces, are hotspots for adaptive change. In

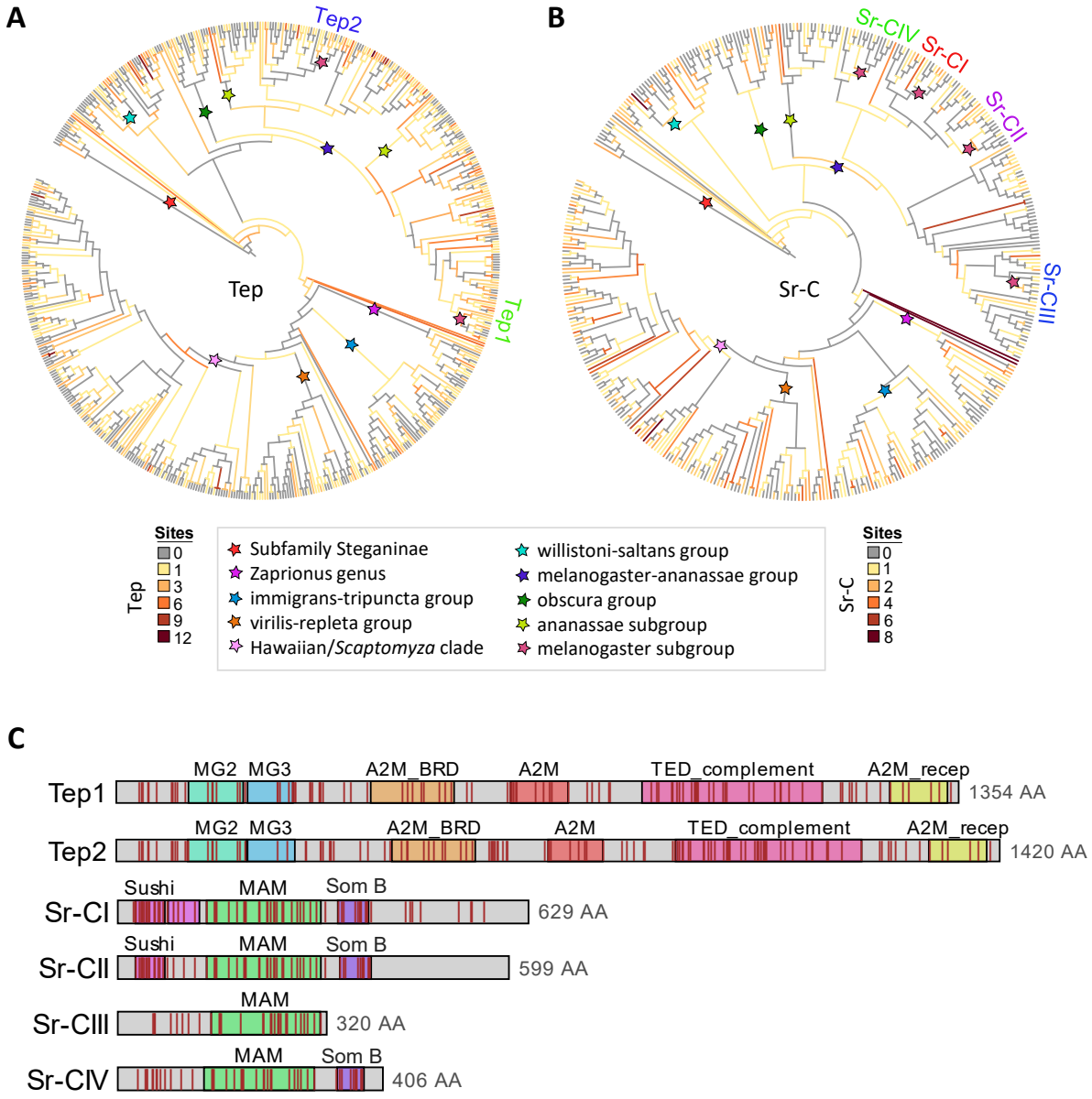

Figure 3.5: Evolutionary patterns of the thioester-containing protein (Tep) and class C scavenger receptor (Sr-C) families in Drosophilidae.
(A, B) Maximum-likelihood gene trees (topology only) of Tep (A) and Sr-C (B) families, rooted by the subfamily Steganinae. Branch colors indicate the number of sites (codons) under episodic diversifying selection (as estimated by MEME). Major Drosophila group radiations are marked with stars; *D. melanogaster* paralogs are labelled at the tips. (C) Schematic gene structures of Tep and Sr-C showing conserved functional domains (not to scale) and approximate positions of positively selected sites (red bars). Positively selected sites were identified with MEME.

contrast, sequence divergence was negatively associated with gene expression level, gene length, and the number of genetic/protein interactions, patterns that match observations across plants, vertebrates, and insects (Duret and Mouchiroud 2000; Zhang and Yang 2015; Moutinho et al. 2019). These relationships point to constraints imposed by structural and regulatory complexity—indicating that immune gene evolution is shaped not only by host–pathogen arms races but also by general molecular and genomic features. The rate of gene turnover, however, largely unaffected by these gene- and protein-level predictors.

After accounting for gene length, expression level, number of genetic/protein interactions, and RSA, we found that immune proteins evolved faster at the sequence level than non-immune genes but exhibited lower overall rates of gene turnover. This finding is surprising, given that many lineage-specific studies in Drosophila and other taxa have reported extensive duplication and loss in immune-related families (Sackton et al. 2007; Han 2019; Khan et al. 2019; Domazet-Lošo et al. 2024; Manousi et al. 2025). However, other studies highlight that the contribution of these birth-death events to the diversification of multi-gene families in long-term evolution seems to be minor (Nei and Rooney 2005). Also, the gene families with high gene turnover identified in these studies are those that produce a variety of gene products (some effectors and receptors families), whereas the majority of immune gene families have conserved, specialized functions that are less likely to tolerate dosage imbalance or loss of essential components.

While these patterns are robust, several methodological and biological factors should be considered when interpreting them. Our turnover estimates from CAFE5 were restricted to gene families present at the origin of Drosophilidae, which may have excluded the most volatile multi-copy families. This conservative approach likely underestimates turnover for certain effector gene families, especially those with lineage-specific origins or extreme expansions. In addition, our immune gene set contains a relatively high proportion of conserved signalling genes ($\sim$51%) compared to rapidly evolving effectors ($\sim$16%) and receptors ($\sim$8.8%); this composition could reduce the average differences between immune and non-immune categories for both dN/dS and λ.

Overall, our findings reveal that Drosophila immune genes diversify along multiple evolutionary axes, with the relative importance of sequence adaptation and gene turnover varying by functional class, pathway, and gene-level and structural constraints. The contrasting patterns in Tep's, Sr-Cs, and AMPs illustrate how immune defence systems can adapt through distinct

combinations of evolutionary processes, often tailored to a gene's biochemical role and interaction with pathogens. These results underscore the value of integrating genomic, structural, and functional perspectives to understand immune system evolution and provide a comparative framework for exploring the evolution of immunity in other taxa.

3.6 Acknowledgements

We wish to thank Bernard Kim and Dmitri Petrov for making Drosophila genome data available to us in advance of their publication. We also would like to thank Jarrod Hadfield for help and advice on the use of MCMCglmm.

Chapter 4

Transcriptomic analysis of three non-model Drosophilidae reveals novel AMP candidates

The text in this chapter is from bioRxiv preprint: Dhakad P, Newman D, Obbard DJ (bioRxiv) **"Transcriptomic analysis of non-model Drosophilidae reveals novel AMP candidates"** [DOI: 10.1101/2025.06.06.658223]

I wrote this chapter with comments and textual edits from Prof. Darren Obbard. Dhobasheni Newman helped in setting up the fly infection experiment. Thanks to Dr. Bernard Kim and Prof. Dmitri Petrov for making the Drosophila genomes data available to us.

4.1 Abstract

Drosophila melanogaster has been a valuable model for dissecting the molecular architecture of innate immunity. However, the family Drosophilidae encompasses over 4000 species, spanning deep evolutionary divergences and diverse ecologies. Here, we use immune challenge with the gram-negative pathogen *Providencia rettgeri* to investigate the conservation and evolution of immune responses in three non-model drosophilid species that diverged from *D. melanogaster* over 45 million years ago—*Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa*. We find that all three species retain a core set of immune signalling and recognition genes, but exhibit substantial variation in effector gene content and inducibility. In particular, *Scaptodrosophila deflexa* lacks orthologs of multiple antimicrobial peptides (AMPs) known from *D. melanogaster*, including *DptA*, *AttA*, and *AttC*, and shows little transcriptional response to bacterial-challenge with *Providencia rettgeri*. In contrast, both of the *Hirtodrosophila* species exhibit substantial transcriptional responses, including strong

induction of canonical Imd pathway genes. Microbiome profiling of our samples revealed higher *Providencia* abundance in *H. cameraria*, and high levels of the defensive symbiont *Spiroplasma* in *S. deflexa*—potentially explaining differences in infection outcome. Our combined annotation and expression analysis of these species also allowed us to identify 20 novel AMP-like candidates, many with structural features like known AMPs. Our study demonstrates the feasibility of functional immune analyses in non-model Drosophila species and reveals striking lineage-specific differences in immune gene repertoire and expression. These findings highlight the importance of non-model, wild-derived taxa for uncovering novel immune effectors and understanding evolutionary forces shaping insect immunity.

4.2 Background

In all organisms, the innate immune system forms the first line of defence against pathogens and parasites (Nurnberger et al. 2004; Buchon et al. 2014; Silva and Gomes 2024). By rapidly recognizing and responding to infections, it reduces pathogen survival and replication—decreasing pathogen fitness to the benefit of the host. These antagonistic interactions often lead to coevolutionary dynamics, where immune-related genes—particularly those involved in pathogen recognition and effector functions—evolve at significantly higher rates compared to the genome-wide background (e.g. Hughes and Nei 1988; Sackton et al. 2007; Waterhouse et al. 2007; Obbard et al. 2009a; Singh et al. 2012). The fruit fly *Drosophila melanogaster* has long served as a model for dissecting innate immunity in insects and helped in understanding key pathways such as Toll, Imd, JAK/STAT, and RNA interference (RNAi), which orchestrate defence responses against bacteria, fungi, and viruses (e.g. Lemaitre and Hoffmann 2007; Buchon et al. 2014).

As in vertebrates, *D. melanogaster* mounts both humoral and cellular innate immune responses to combat pathogen infection. In Drosophila, cellular immunity primarily involves phagocytosis by plasmatocytes and encapsulation by lamellocytes, targeting invading microbes and larger parasites, respectively (Tepass et al. 1994; Evans et al. 2003). In contrast, the humoral response mainly leads to the rapid production and systemic release of antimicrobial peptides (AMPs), which are secreted into the haemolymph to directly kill pathogens. Two major signalling pathways, Toll and Imd, regulate the majority of immune genes in Drosophila, including the production of AMPs. The Toll pathway is predominantly activated by lysine (Lys)-type peptidoglycan found in Gram-positive bacteria, as well as fungal β-glucans and circulating persephone proteins (Issa et al. 2018). The Imd pathway is activated through the detection of diaminopimelic acid (DAP)-type peptidoglycan from Gram-negative and certain Gram-positive

bacteria (Buchon et al. 2014). In addition to Toll and IMD, the JAK/STAT pathway plays a modulatory role in immune defence. While its primarily role is in development, stress response, and stem cell maintenance, it also contributes to immunity by regulating the expression of genes such as those encoding thioester-containing proteins (TEPs) and Turandot stress proteins (Lagueux et al. 2000; Agaisse et al. 2003; Dostalova et al. 2017).

The completion of 12 Drosophila genomes in 2007 opened the door to evolutionary analyses across multiple species, enabling researchers to investigate gene copy number variation, patterns of positive selection, and lineage-specific immune responses (Clark et al. 2007). One striking pattern that emerged was an apparent dichotomy in the evolutionary dynamics of Drosophila immune genes; while core signalling components of immune pathways are often deeply conserved, the evolution of recognition and effector genes can be highly dynamic (Jiggins and Kim 2005; Sackton et al. 2007; Hanson et al. 2016). Upstream signalling molecules such as *Relish* and *Dif* family members typically persist as 1:1 orthologs, even in distantly related insects (Viljakainen 2015), and have detectable sequence homology in mammals (e.g., NF-κB), highlighting the deep conservation of immune signalling genes (Silverman and Maniatis 2001). In contrast, AMP gene families frequently undergo duplication, pseudogenization, and loss, and exhibit high copy number variation between species (Jiggins and Kim 2005; Quesada et al. 2005; Sackton et al. 2007; Hanson et al. 2019). In some species, AMPs such as *drosocin*, *drosomycin*, *turandot*, and *metchnikowin* are either completely absent or show such a high sequence divergence that they are difficult to identify through standard homology searches (Sackton and Clark 2009; Salazar-Jaramillo et al. 2014; Hanson et al. 2019).

This contrast raises an important but hard to answer question: what drives the dynamic evolutionary patterns of antimicrobial peptides (AMPs)—including gene duplications, losses, and lineage-specific expression? This question is difficult to answer, in part, because our experimental understanding of the drosophilid antibacterial response comes largely from a few of the more experimentally tractable species, predominantly *D. melanogaster* and its close relatives within the subgenus *Sophophora* (Westlake et al. 2024)—with only handful of species outside of this subgenus (Sackton and Clark 2009; Salazar-Jaramillo et al. 2014; Hanson et al. 2016; Hanson et al. 2019; Hanson et al. 2023). Nevertheless, comparative studies have revealed striking lineage-specific variation in effector gene repertoires.

The antimicrobial peptide *diptericin* offers a particularly well-characterized example of this variation. *Diptericin* genes are key effectors of the Imd pathway and show differential inducibility by Gram-negative bacteria across Drosophila species (Hanson et al. 2019). A population genetics study found that serine/arginine polymorphism at 69th residue of *DptA* significantly

alters susceptibility to *Providencia rettgeri* bacterial infection in *D. melanogaster* and *D. simulans*–an association interpreted as a signature of balancing selection (Unckless and Lazzaro 2016). More recent work has shown that the selective landscape acting on *diptericin* may be highly context-dependent: interactions between host genotype, sex, environmental stress (e.g., starvation), and pathogen exposure can shape the evolutionary trajectories of AMP alleles, potentially maintaining diversity over time (Mullinax et al. 2025).

At a broader phylogenetic scale, AMP families exhibit even more extreme evolutionary dynamics. For example, while *DptA* has been lost or pseudogenized in some non-melanogaster species, others possess divergent paralogs (e.g., *DptC* in the subgenus *Drosophila*) that are syntenic but highly diverged at the sequence level (Hanson et al. 2016). Consistent with this, AMP duplicates can undergo neofunctionalization: in *D. virilis*, a *defensin* paralog has evolved from a role in bacterial killing to one in toxin neutralization (Gao and Zhu 2024). Together, these findings suggest that while the basic architecture of innate immunity is ancient and broadly conserved, the downstream effector repertoire is evolutionarily labile and shaped by selection from species-specific microbial exposure, life-history, and ecological pressure.

Now, with the recent availability of nearly 400 drosophilid genomes (Kim et al. 2024)—over 300 of which are annotated (Dhakad et al. 2025a)—there is an opportunity to explore the extent to which canonical immune responses are conserved across the drosophilid phylogeny. For example, we can ask if deeply diverged lineages harbor novel immune effectors, or if more distantly diverged non-model species—with their unique ecological niches and evolutionary histories—possess a distinct immune repertoire. However, one major challenge remains; most functional studies have focused on species amenable to long-term laboratory culture, but these represent only a small fraction of drosophilid diversity (Kim et al. 2024). Understanding gene function in species that are not easily cultured in large numbers remains challenging.

In this study, we demonstrate the feasibility of characterizing immune responses to bacterial-challenge in non-model, less easily cultured, species of Drosophilidae by performing comparative transcriptomic analysis on individual first-generation wild-derived flies. We selected three common European species for analysis—*Hirtodrosophila cameraria*, *Hirtodrosophila confusa*, and *Scaptodrosophila deflexa*—each of which is highly divergent from *D. melanogaster* and other well-studied taxa (>45 MYA; Figure 4.1; Suvorov et al. 2022). These species are not only genetically distant, but also ecologically distinct. *Hirtodrosophila confusa* is a relatively large drosophilid and a fungal specialist that thrives in cool temperate environments and is frequently found in association with large fungal fruiting bodies such as dryad's saddle (*Polyporus squamosus*). *Hirtodosophila cameraria* is also a specialist fungus breeder, moderately abundant in the UK on basidiomycete fungi such as *Phallus impudicus* and *Lactarius*

quietus (Grimaldi and Richenbacher 2023; Obbard et al. 2023b). *Scaptodrosophila deflexa*, in contrast, is thought to lay its eggs in yeast-rich sap fluxes. However, despite their broad distribution and ecological interest, these species remain very poorly studied; for example, the genera *Hirtodrosophila* and *Scaptodrosophila* have both been shown to be polyphyletic (*Hirtodrosophila* partly within the paraphyletic *Drosophila*; Finet et al. 2021; Kim et al. 2024). Despite the availability of high-quality genome sequences (Obbard et al. 2023b; Kim et al. 2024), they are yet to receive any attention in functional or comparative genomics. To our knowledge, only a single report addresses egg viability in *H. confusa* (David et al. 2005), and no studies to date have explored their immunogenomics or transcriptomic responses to infection. By analysing these lineages, we aim to expand our understanding of immune system diversity and evolution within Drosophilidae, moving beyond the traditional model organisms and exploring a broader phylogenetic landscape.

Here we compare the transcriptomes of unchallenged flies with those of flies we experimentally challenged with the Gram-negative bacterial pathogen *Providencia rettgeri*. Our goals are to: (i) evaluate whether pathogen-challenged transcriptomes improve the annotation of immune genes in species lacking reference-quality genomes; (ii) assess the feasibility of differential expression analyses in these non-model species; and (iii) identify conserved and novel immune-related genes, including potential AMPs. Our findings suggest that despite their deep evolutionary divergence, bacterial infection induces measurable and broadly similar transcriptional responses in *H. cameraria* and *H. confusa* species but not in the more distant *S. deflexa*—a pattern that could potentially be shaped by differences in microbiota, immune strategy, or symbiont-mediated protection.

4.3 Materials and methods

4.3.1 Fly collection and infection

We collected wild-mated female *Hirtodrosophila confusa* (n = 2), *H. cameraria* (n = 3 females), and *Scaptodrosophila deflexa* (n=1) from the Hermitage of Braid (Edinburgh, UK, all within 500m of 55.9 N, 3.2 W) on 4th August 2023. Wild-caught female flies were housed separately under a 12:12-hour light-dark cycle and allowed to lay eggs. Wild caught *H. confusa* and *H. cameraria* were maintained on whole mushrooms (*Agaricus bisporus*), and *S. deflexa* on fermentative substrate intended to mimic sap flux conditions: Drosophila medium supplemented with tiny slices of ripe banana and cotton plug heads moistened with cider and maple syrup. Emerging first-generation female flies were collected and housed in same-sex vials

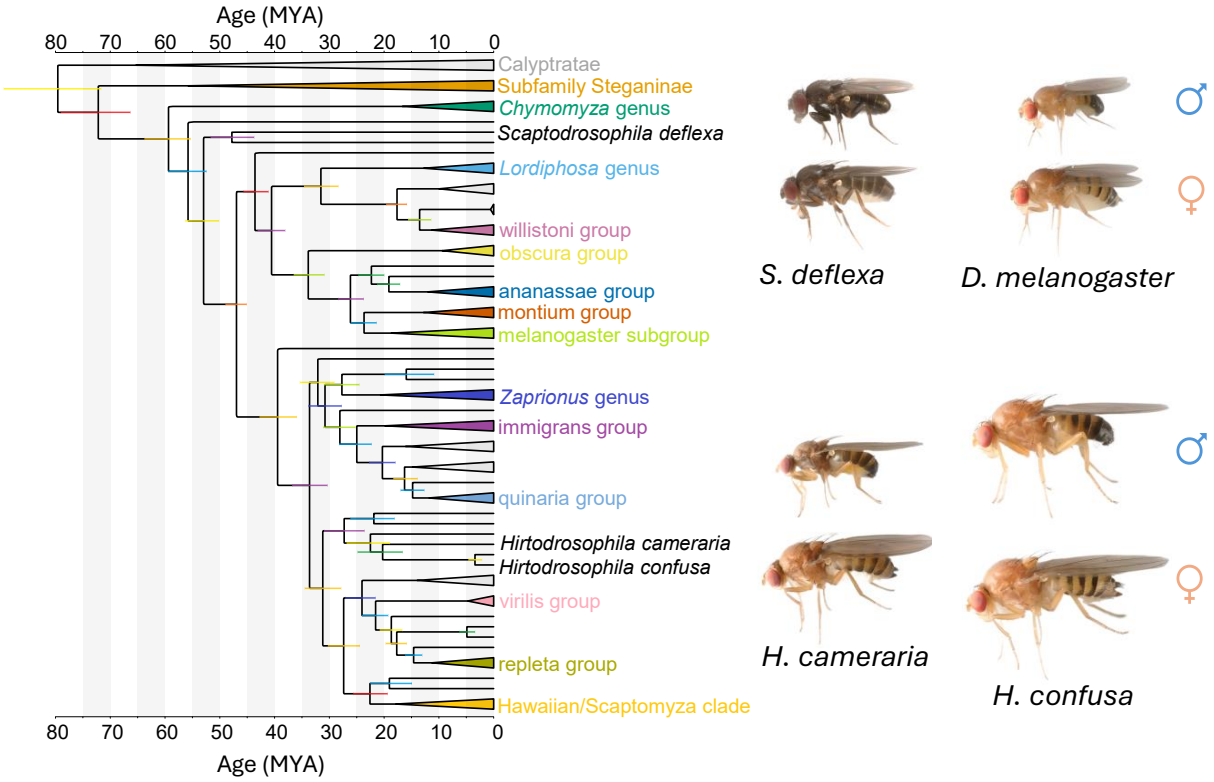

Figure 4.1: The phylogenetic position of the study species within Drosophilidae.
An approximate time-calibrated phylogeny showing the relationships of major lineages of Drosophilidae, including the placement of the three non-model species used in this study—*Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa*. The *Hirtodrosophila* species belong to a diverse and understudied cluster within the subgenus Drosophila, while *S. deflexa* branches deep within the subfamily Drosophilinae, representing one of the earliest branching lineages relative to *D. melanogaster* (mrca >50 MYA). Divergence times are shown in millions of years ago (MYA) along the x-axis. Taxonomic groups are color-coded by genus/subgenus or species group, and collapsed clades indicate major drosophilid lineages for clarity. Images to the right right show adult male (♂) and female (♀) flies of each of the three focal species, plus *D. melanogaster* at the same scale for comparison. The tree figure was generated as described in Dhakad et al. (2025a), based on 285 single-copy BUSCO orthologs.

containing standard Lewis medium (*Bloomington Drosophila Stock Center. Indiana University Bloomington.* N.d.) for 3–5 days prior to bacterial challenge. As a result, all flies were between 3 and 6 days old at the time of infection. For bacterial challenge, we used the 'Dmel' strain of *Providencia rettgeri* originally isolated from wild *Drosophila melanogaster* by Brian Lazzaro (Juneja and Lazzaro 2009), and provided to us by Pedro Vale (University of Edinburgh). *Providencia rettgeri* cultures were initiated from a single colony and grown overnight in 10ml LB broth at 37°C with shaking. The bacterial culture was centrifuged at 5000 rpm for 5 min at 4°C and the supernatant was discarded. Bacteria were diluted to an optical density of OD_{600} = 0.1 in sterile phosphate-buffered saline (PBS) prior to infection (Chambers et al. 2019). Flies were anesthetized on CO_2, and bacterial challenge was administered by puncturing the thorax with a 0.14 mm diameter stainless steel pin dipped in the bacterial suspension. Five females per species (four females and one male for *S. deflexa*) were challenged in this way. Control flies were anesthetized but otherwise left unmanipulated, and thus this experiment did not control for wounding effects. Both challenged and unchallenged flies were maintained at room temperature for 16 hours post-infection, then flash-frozen in Trizol reagent (Invitrogen) and stored at -80°C until RNA extraction. RNA extraction and sequencing were carried out by Novogene (www.novogene.com) using a standard TRIzol-based extraction protocol. Sequencing libraries were prepared using an rRNA depletion approach (without poly-A selection) with TruSeq stranded total RNA library kit and sequenced on an Illumina NovaSeq platform to generate 150 bp paired-end reads.

4.3.2 Quality control and mapping

On average, samples contained 58.35 million raw reads per fly. We preprocessed raw reads for quality control using fastp v0.24.0, with the "*-c*" option enabled for overlap base correction and "*-y -Y 20*" for low-complexity filtering (Chen 2023). All other parameters were kept at their default settings. Reads were mapped to their respective genomes using STAR RNAseq aligner v2.7.10b, generating sorted BAM files as output (Dobin et al. 2013). On average, 83.98%, 66.55%, and 82.02% of reads per library mapped uniquely to the *H. confusa* (assembly accession: GCA_035043065.1), *H. cameraria* (GCA_949708635.1), and *S. deflexa* genomes, respectively (Supplementary file C.4). The *Scaptodrosophila deflexa* genome was sequenced by Bernard Kim (Princeton) and Dmitri Petrov (Stanford), using ONT R10.4.1 sequencing technology and assembled with hifiasm (Cheng et al. 2021), and was generously made available to us in advance of publication. BUSCO completeness and N50 for all the genomes used in this study can be found in Supplementary file C.4.

4.3.3 Gene annotation

To assess the impact of pathogen challenged RNAseq on our ability to recover immune genes, as compared with unchallenged RNAseq, we generated 3 independent genome annotations for each species using: (1) RNAseq data from pathogen-challenged individuals, (2) RNAseq data from unchallenged (naïve) individuals, and (3) combined RNAseq data. Genome annotations were generated using BRAKER3 in ETP mode, with extrinsic protein hints provided from *D. melanogaster* RefSeq proteins (Gabriel et al. 2024). To assess gene orthology and recover immune-related orthogroups, we extracted the longest isoform per gene from each annotation set and ran OrthoFinder v2.5.5 using the predicted proteomes of all species and annotation sets, along with the *D. melanogaster* proteome (Emms and Kelly 2019). To generate hierarchical orthogroups (HOGs), gene trees were inferred by OrthoFinder using alignments generated with MAFFT (Katoh and Standley 2013) and maximum-likelihood inference performed with IQ-TREE2 (Minh et al. 2020).

4.3.4 Differential gene expression analysis and functional annotation

Read counts per gene were quantified using "featureCounts" (Liao et al. 2014). Genes with fewer than 10 total counts across all samples were excluded from downstream analysis. Differential expression analysis was conducted using the DESeq2 package in R (Love et al. 2014). We performed principal component analysis (PCA) using the *plotPCA()* function on regularized log-transformed (rlog) counts, with $\sim$*treatment* specified as the design formula. The top 500 most variable genes were used to calculate principal components. Genes were considered significantly differentially expressed if they had an adjusted p-value < 0.05 and a $|\log_2$ fold change$| \geq 1$. Heatmaps of expression patterns were generated using the pheatmap package (Kolde 2025).

Functional annotation of genes was performed using eggNOG-mapper v2.1.12 with Diptera HMM database using HMMER searches (*-m hmmer*) and additional filters for hits with e-value ≤ 0.05, bit score ≥ 60, and percent identity ≥ 40 (Johnson et al. 2010; Cantalapiedra et al. 2021). GO terms were restricted to non-electronic evidence codes, and PFAM domains were realigned (*–pfam_realign realign*). Gene Ontology (GO) enrichment analysis was performed using "topGO" package in R by applying Fisher's exact test and the "*weight01*" algorithm, which accounts for the hierarchical structure of GO terms (Alexa and Rahnenfuhrer 2016). The p-values from GO analyses were corrected using the Benjamini and Hochberg procedure with the FDR threshold set to 0.05 (Benjamini and Hochberg 1995).

4.3.5 Metagenomic analysis of unmapped reads

To quantify microbial abundance in flies and to assess whether pathogen-induced expression differences could be affected by microbial load, we performed denovo metatranscriptomic analysis on unmapped reads from each sample. Paired-end unmapped reads were extracted from STAR-mapped BAM files using samtools v1.13 with the flags "*-f 12 -F 256*", to retain only unmapped read pairs (Danecek et al. 2021). These reads were assembled de novo using rnaSPAdes v4.1.0 (Bushmanova et al. 2019), with default parameters. Open reading frames (ORFs) were predicted from each transcriptome assembly using EMBOSS "getorf", retaining protein length $\geq$200 amino acids. The resulting proteins were queried against the NCBI non-redundant (nr) protein database using "diamond blastp" v2.1.10 (*–evalue 1e-20, –outfmt 6*) (Buchfink et al. 2021) and taxonomic lineages were assigned to diamond hits using TaxonKit (Shen and Ren 2021). To estimate relative microbial load, total reads mapped to each organism were normalized to host-mapped read counts for each sample to control for sequencing depth. Alpha diversity (Shannon index) was calculated from genus-level profiles using the vegan and phyloseq packages in R (McMurdie and Holmes 2013; J 2022). Statistical comparisons between pathogen-challenged and unchallenged groups were performed using the Wilcoxon rank-sum test.

4.3.6 Prediction of novel AMPs

To identify potential novel immune effectors, we screened all significantly upregulated genes that had no detectable homologs in *D. melanogaster* and a predicted peptide length $\leq$200 amino acids. These candidates were assessed using two AMP prediction tools: (1) AMP Scanner v2, a deep neural network-based classifier (Veltri et al. 2018), and (2) amPEPpy, a Python implementation of the amPEP random forest classifier (Lawrence et al. 2021). The random forest model was trained using 712 known AMPs from the APD3 database and 712 matched non-AMP sequences (see original publication for dataset details: Veltri et al. 2018). All candidate peptides were also analysed for the presence of N-terminal signal peptides using SignalP 6.0 webserver (Teufel et al. 2022). Physiochemical properties of candidates were obtained using AMP predictor tool of APD3 (Wang et al. 2016) and ExPASy "ProtParam" (Wilkins et al. 1999). Finally, 3-D structures of AMP candidates were generated using the AlphaFold 3 server (Abramson et al. 2024).

4.3.7 Availability of data and materials

All RNAseq reads data generated for this study have been deposited in the SRA database, and can be found under project ID PRJNA1270041.

4.4 Results

4.4.1 Pathogen challenge does not substantially improve annotation

To assess the impact of the availability of pathogen-challenged RNAseq on gene predictions, we generated three separate genome annotations for each species using BRAKER3 (Gabriel et al. 2024). The three annotations were based on (i) RNAseq from pathogen-challenged flies, (ii) RNAseq from unchallenged flies, and (iii) combined data. All three annotations recovered the same set of core genes in each species, and most gene models were shared between the challenged and unchallenged annotations, with only a small number unique to one set or the other (Table 4.1; Figure 4.2). The combined RNAseq data produced slightly more gene models compared to either pathogen-challenged or unchallenged datasets alone for *H. cameraria* and *S. deflexa*, whereas the pathogen-challenged dataset yielded slightly more genes for *H. confusa* (Table 4.1). To assess the global similarity between genome annotations generated from pathogen-challenged and unchallenged RNAseq datasets, we counted the shared (i.e. overlapping coordinates) and unique gene models, recording how many were assignable to an orthogroup (Figure 4.2). As expected, most genes were shared between pathogen-challenged and unchallenged annotations (category C; Figure 4.2), suggesting consistent annotation of core gene sets across datasets. Only a small number of genes were uniquely recovered in either pathogen-challenged or unchallenged annotations (categories A, B, D, and E; Figure 4.2), and many of these lacked orthogroup assignment, implying that they may be annotation errors, or potentially novel genes. To examine whether pathogen-challenged RNAseq improved recovery of immune-related genes, we compared immune gene orthogroup ('Hierarchical OrthoGroup'; HOG; Schreiber and Sonnhammer 2013) representation between pathogen-challenged and unchallenged annotations for each species. Small differences were observed, but most immune genes were either recovered in both annotations or missing from both, with few genes uniquely recovered by pathogen-challenged RNAseq (Table 4.1). To complement this analysis, we performed Gene Ontology (GO; Ashburner et al. 2000) enrichment analysis on genes uniquely annotated from pathogen-challenged or unchallenged

RNAseq datasets. However, no significant enrichment for immune-related terms was detected among genes uniquely recovered in pathogen-challenged annotations (Figure C.1). Overall, these results indicate that pathogen-challenged RNAseq did not substantially improve overall immune gene discovery or annotation completeness compared to unchallenged RNAseq.

Species	Genes	Mean CDS length (Kbp)	Immune OGs	Immune genes	Immune CDS length (Kbp)
D. melanogaster	13986	1.54	568	638	1.74
H. cameraria (unchallenged)	14407	1.46	515	545	1.88
H. cameraria (challenged)	14203	1.47	520	551	1.86
H. cameraria (combined)	14503	1.47	539	573	1.87
H. confusa (unchallenged)	14200	1.39	490	519	1.72
H. confusa (challenged)	14469	1.39	500	530	1.78
H. confusa (combined)	14300	1.41	534	566	1.76
S. deflexa (unchallenged)	13017	1.50	490	519	1.95
S. deflexa (challenged)	12991	1.48	485	514	1.93
S. deflexa (combined)	13148	1.50	522	553	1.95

Table 4.1: Gene annotation statistics and immune gene recovery from pathogen-challenged, unchallenged, and combined RNAseq data for each species.

4.4.2 The detectable immune repertoire differs between species

To further assess how immune gene recovery varied across the three divergent drosophilid lineages, we examined the presence/absence of genes in *H. cameraria*, *H. confusa*, and *S. deflexa* that have homology with a curated set of 638 well-characterized *D. melanogaster* immune-related genes (Westlake et al. 2024). These 638 immune genes were clustered into 568 HOGs. Of these, *H. cameraria* recovered genes in 539 HOGs, *H. confusa* in 534, and *S. deflexa* in 522 (Table 4.1; Supplementary file C.1). In total, 497 HOGs had immune genes recovered across all three species, while 11 HOGs contained only *D. melanogaster* genes. These uniquely missing HOGs included several AMPs such as *drosocin* (*Dro*), all members of the *turandot* family, and *drosomycins* (*Drs* and *Drsl*1–6; Supplementary file C.1). These apparent losses support previous observations that the *drosomycin* and *turandot* AMP families are largely restricted to *D. melanogaster* and closely related species in the subgenus *Sophophora*

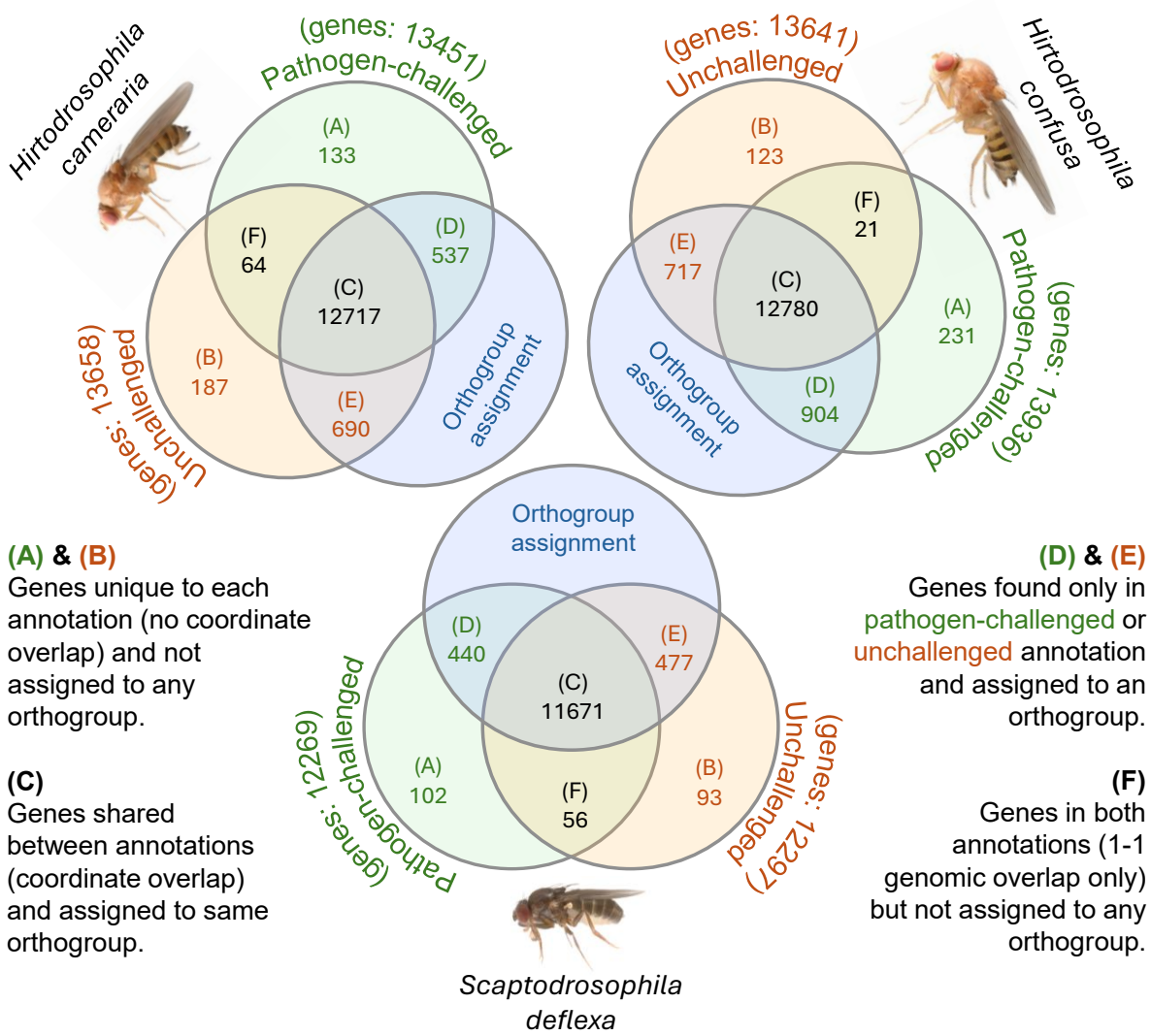

Figure 4.2: Comparison of gene sets annotated using RNAseq from pathogen-challenged and unchallenged samples.
Venn diagrams illustrating the overlap of gene models annotated using RNAseq from pathogen-challenged (green) and unchallenged (orange) samples of Hirtodrosophila cameraria, H. confusa, and Scaptodrosophila deflexa. Gene sets were compared based on (i) genomic coordinate (1:1 overlap), and (ii) orthogroup assignment using OrthoFinder. Numbers inside Venn diagrams (A to F) represent gene model counts. The total number of predicted genes per condition is shown in parentheses next to each label.

(Sackton and Clark 2009; Hanson et al. 2016). However, we found homologs of a *drosocin*-like gene in all three species, and in *H. confusa* and *H. cameraria* this gene encodes multiple tandem repeats of the peptide domain—as previously reported in *D. neotestacea* and other species in the subgenus *Drosophila*. In total, only two recognition proteins (*PGRP-SB2* and *CG12780*) and six signaling genes (*serpin 42Dd*, *MstProx/Toll-3*, *Toll-4*, *sphinx1*, *sphinx2*, and *amnesiac*) were missing from all three species. Interestingly, *S. deflexa* lacked orthologs of *diptericin A* (*DptA*), *attacin C* (*AttC*) and *attacin D* (*AttD*), suggesting lineage-specific loss of these AMPs.

To test whether the species and immune-gene functional categories differed in the probability of each 'gene' (i.e. orthogroup) being recovered, we fitted a Bayesian binomial linear mixed-effects model using MCMCglmm. Posterior predictions revealed a significant species effect, with *H. cameraria* (posterior mean = 99.88%; 95% Credible Interval: 99.61-99.98) and *H. confusa* (99.79%; CI: 99.35-99.97) showing similar probabilities of immune gene recovery, while *S. deflexa* recovered significantly fewer immune genes (99.58%; CI: 98.80-99.93). This may partly reflect differences in genome and annotation quality, but also likely reflects genuine differences in immune gene conservation and loss. We also detected a marginal effect of gene category, such that canonical signaling genes (99.97%; CI: 99.91-99.99) were more likely to be recovered than effector genes (99.46%; CI: 98.23-99.94), although receptor and unknown categories were not significantly different to effectors. Among interaction terms, only the combination of *S. deflexa* and signaling genes showed a significant positive deviation from expectation (99.97%, CI: 99.89–99.99), indicating that signaling genes were relatively well retained even in *S. deflexa*. Other species and category interactions terms were not significantly different from additive expectations.

To illustrate the evolutionary lability of effector gene families, we chose to examine the *diptericin* and *attacin* genes in more detail. Phylogenetic analysis of *diptericins* (Figure 4.3 A) confirmed three distinct clades: *DptA*, *DptB*, and *DptC*, the latter being restricted to the subgenus *Drosophila* (Hanson et al. 2016). *Scaptodrosophila deflexa* encoded only a single *DptB*-like gene, clustering with *orthologs* from *H. confusa*, *H. cameraria*, and *D. melanogaster*, but lacked any detectable *DptA* and *DptC* homologs. In contrast, multiple *DptA* and *DptC* paralogs were recovered in both *Hirtodrosophila* species, suggesting lineage-specific expansions (Figure 4.3 B). Similarly, the *attacin* gene tree (Figure C.2) showed that *H. cameraria* and *H. confusa* have multiple copies of *AttC*, and *AttA/B*, while *S. deflexa* possesses only a single *AttA/B*-like gene with no detectable orthologs of *AttC* and *AttD* (Figure 4.3 C; Figure C.2). The clear pair of *AttA* and *AttB* genes in most species (Figure C.2) might superficially suggest that each species has experienced a recent duplication independently. However, it is more likely that

these patterns reflect concerted evolution in *diptericin* and *attacin* gene families (Jiggins and Kim 2005; Cortazar-Chinarro et al. 2020), highlighting the challenges of inferring orthologs of AMPs across divergent genomes. Taken together, these results show the rapid and idiosyncratic evolution of effector gene repertoires across drosophilid species, and illustrate how their annotation is particularly sensitive to both sequence divergence and genome assembly quality.

4.4.3 Immune challenge triggers a conserved immune response in *Hirtodrosophila* but not in *S. deflexa*

To identify genes that are transcribed in response to bacterial infection, we performed differential expression analysis between pathogen-challenged and unchallenged individuals using DESeq2 (Love et al. 2014). We identified 363 significantly upregulated genes (adj. $p < 0.05$, $\log_2$ fold change ≥ 1) in *H. cameraria*, 149 genes in *H. confusa* and only 34 genes in *S. deflexa* after bacterial challenge (Supplementary file C.2). In addition, we identified 230 significantly downregulated genes (adj. $p < 0.05$, $\log_2$ fold change ≤ -1) in *H. cameraria*, 82 genes in *H. confusa* and only 7 genes in *S. deflexa* (Supplementary file C.2).

To visually assess the consistency of the response across individuals, we performed principal component analysis (PCA) of normalized expression data. In both *H. cameraria* and *H. confusa*, PCA analysis revealed clear separation between pathogen-challenged and unchallenged samples along PC1 (explaining 63% and 56% variance, respectively), confirming a strong and coordinated response to infection (Figure 4.4 A). In contrast, samples from *S. deflexa* showed no clear separation by treatment, indicating limited transcriptional response (Figure 4.4 A). The well-fitted dispersion estimates further suggest that this pattern reflects a weak response rather than elevated inter-individual variability (Figure C.3). Heatmaps of significantly differentially expressed genes corroborated these patterns, further highlighting the strong transcriptional induction in the two *Hirtodrosophila* species and the lack of a detectable response in *S. deflexa* (Figure 4.4 B).

As expected from Gram-negative bacterial challenge, the most strongly upregulated genes in *H. cameraria* and *H. confusa* were homologs of canonical AMPs—such as *diptericins*, *attacins*, and *cecropins*—that are downstream targets of the Imd pathway in *Drosophila melanogaster* (Figure 4.5 A and B). Interestingly, *bomanins*—targets of Toll pathway—were also upregulated, although upregulation was not to the level of Gram-negative specific AMPs. We also observed significant upregulation of peptidoglycan receptors, such as homologs of *PGRP-SB1*, *PGRP-LB*, *PGRP-SD*, *PGRP-SA*, *PGRP-LF*, *PGRP-LC*, and *PGRP-SC2* and as well as serine proteases (e.g. *sp7*), which have established roles in immune activation and microbial killing

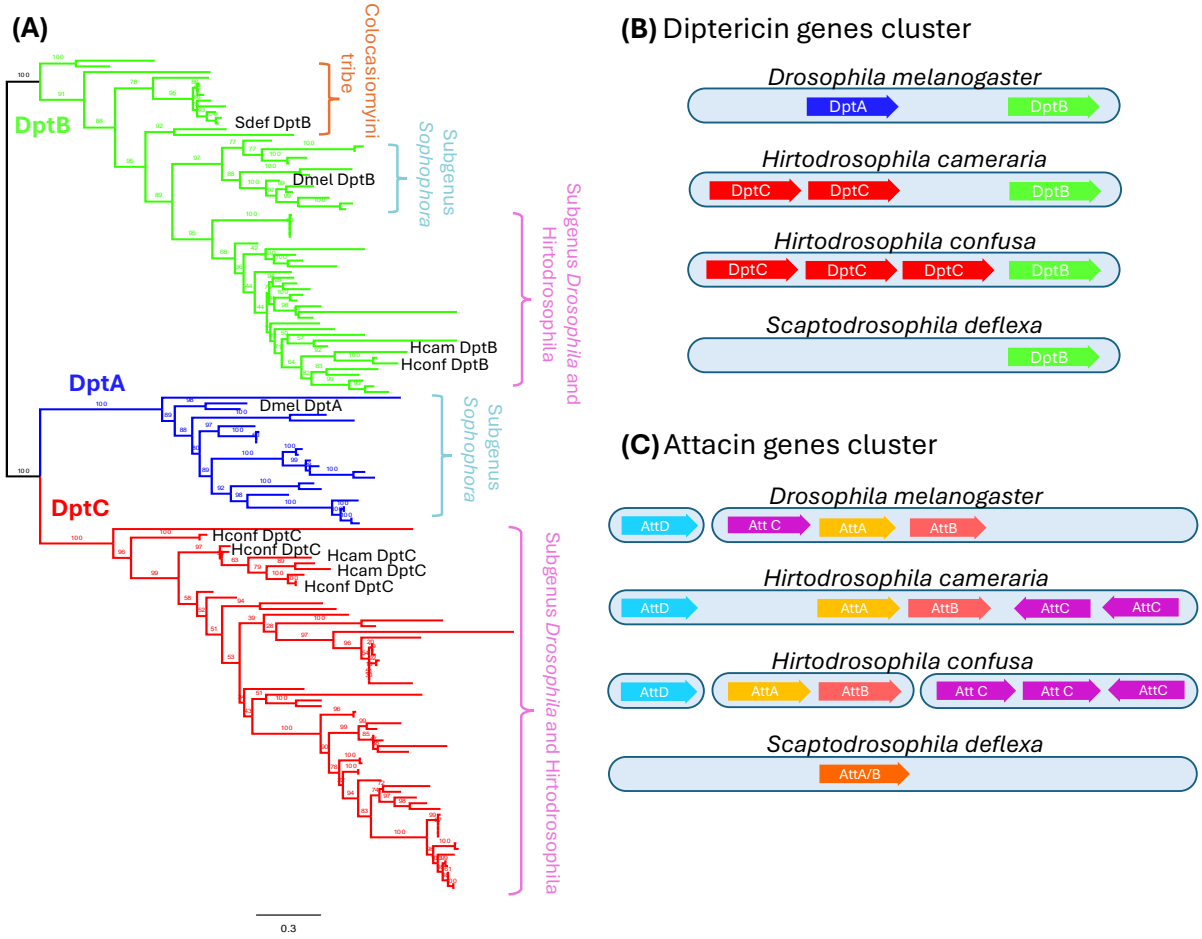

Figure 4.3: Divergence and lineage-specific organization of *diptericin* and *attacin* gene families in drosophilids.

(A) Maximum-likelihood gene tree of diptericin genes from 51 drosophilid species, including *Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa*, generated using IQ-TREE2 based on aligned amino acid sequences. The gene tree highlights three distinct clades corresponding to *DptA* (blue), *DptB* (green), and *DptC* (red). *DptC* is restricted to the subgenus *Drosophila*, including the *Hirtodrosophila* species, and is absent from both *D. melanogaster* and *S. deflexa*. (B) Synteny of *diptericin* gene clusters in *D. melanogaster*, *H. cameraria*, *H. confusa*, and *S. deflexa*. In *D. melanogaster*, *DptA* is upstream of *DptB*. In contrast, the *Hirtodrosophila* species contain multiple tandem repeats of *DptC* genes upstream of *DptB* inplace of *DptA*. *Scatopdrosophila deflexa* has only a single *DptB*-like gene and lacks both *DptA* and *DptC*, possibly indicating lineage-specific gene loss or pseudogenization (independently confirmed with tblastn). (C) Synteny of *attacin* gene clusters in the same four species. In *D. melanogaster*, *attacin* genes are on two different chromosomes, *AttD* on 3R and *AttC*, *AttA*, and *AttB* on 2R. *Hirtodrosophila cameraria* exhibits all four *attacin* genes on same contig, with two *AttC* duplicates positioned downstream of *AttA* and *AttB*. *Hirtodrosophila confusa* also retains *AttD* and *AttC* (3 copies), although on different contigs. In contrast, *S. deflexa* has only a single *AttA/B*–like gene, with no identifiable orthologs of *AttC* or *AttD* (independently confirmed with tblastn). Note: Gene cluster diagrams in panels B and C are schematic and not drawn to scale; gene sizes and intergenic distances are illustrative only.

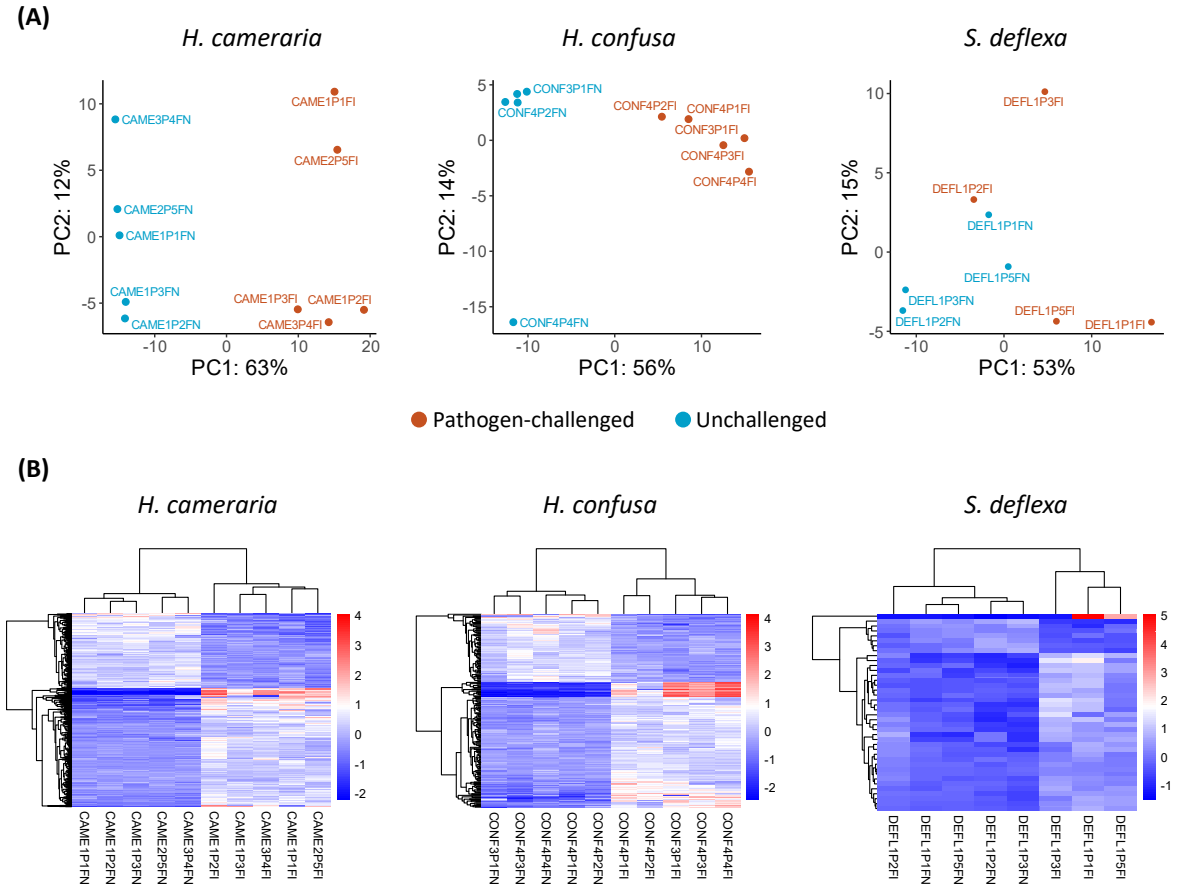

Figure 4.4: Principal component analysis (PCA) and correlation heatmaps of pathogen-challenged and unchallenges samples.
(A) Principal component analysis (PCA) of normalized gene expression data shows clear separation between pathogen-challenged and unchallenged individuals in *H. cameraria* and *H. confusa*, but not in *S. deflexa*, indicating weaker transcriptional response in the latter. (B) Heatmaps of significantly differentially expressed genes (adj. $p < 0.05$, $|\log_2 \mathrm{FC}| \geq 1$).

(Figure 4.5 B; Supplementary file C.2). Apart from immune effectors, signaling genes such as *Rel* and *pirk* that regulate the Imd pathway were also upregulated. In contrast, *S. deflexa* only showed detectable induction of three PGRP receptors among known immune genes (*PGRP-LB*, *PGRP-SD*, and *PGRP-SC2*), consistent with a weak or absent transcriptional immune response to *Providencia rettgeri* (Figure 4.5 B; Supplementary file C.2). Despite the overall similarity of the transcriptional response to infection in *H. confusa* and *H. cameraria*, notable differences may nevertheless exist. For example, *defensin* (*Def*) and *IM33* were only detectably induced in *H. confusa*, and *transferrin 1* and *listericin* only detectably induced in *H. cameraria* (Figure 4.5 B; Supplementary file C.2). Additionally, in all species, we identified several strongly induced genes without clear homologs in *D. melanogaster*, suggesting the induction of novel or species-specific immune genes, many of which might be novel AMPs (below).

We carried out Gene Ontology analyses of the differentially expressed genes to help identify functional classes of genes whose expression is induced or repressed upon pathogen challenge. In both *H. cameraria* and *H. confusa*, upregulated genes were significantly enriched for terms such as “defense response to Gram-negative/Gram-positive bacteria”, “antibacterial humoral response” and “response to fungus” (Figure C.4). These enrichments confirm that infection induced a coordinated immune response. In contrast, only a few immune-related GO terms (“defense response to Gram-negative bacteria” and “antibacterial humoral response”) were significantly enriched among the few upregulated genes in *S. deflexa*, consistent with its subdued transcriptional response. Downregulated genes across all species were enriched for terms related to metabolism and structural processes.

4.4.4 Microbiome variation could underlie immune response heterogeneity

All flies were first-generation lab-reared individuals derived from wild-caught parents, and were reared on non-standard media. This increases the likelihood that the flies carry diverse (and divergent) microbial communities. Such microbiota can influence baseline immune status, alter pathogen susceptibility, or modulate the host’s immune response to bacterial-challenge (Lhocine et al. 2008; Paredes et al. 2011; Bosco-Drayon et al. 2012; Blum et al. 2013). Thus, to explore whether microbiome composition varies, and might therefore have contributed to the apparent variation in immune responses, particularly the muted transcriptional induction observed in *S. deflexa*, we profiled the taxonomic abundance using the unmapped RNAseq reads.

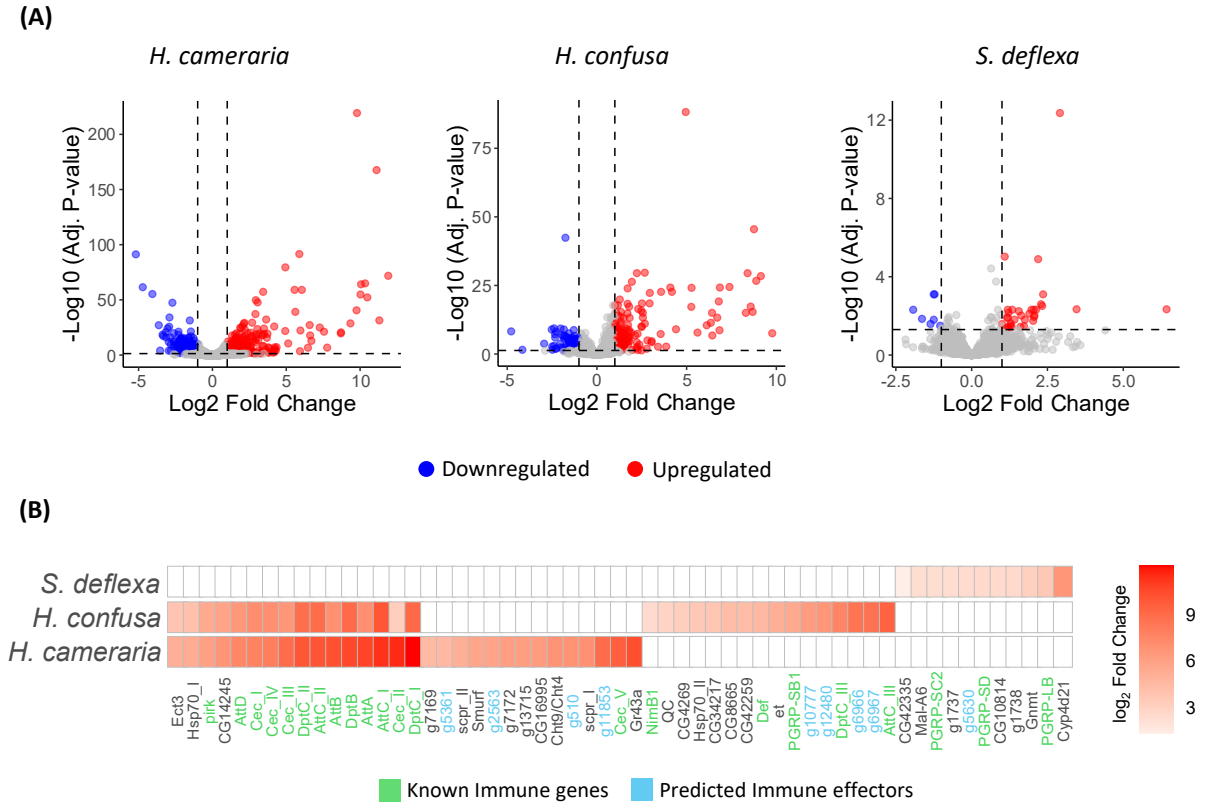

Figure 4.5: Pathogen-challenge with *Providencia rettgeri* induces immune responses in *Hirtodrosophila* species but not in *S. deflexa*.

(A) Volcano plots displaying $\log_2$ fold change versus -$\log_{10}$ adjusted p-values for all expressed genes, highlighting significantly upregulated genes (red) and significantly downregulated genes (blue). (B) Heatmap showing $\log_2$ fold change of selected top genes (based on $\log_2$ fold change and adj. p value) following pathogen-challenge across all three species. Canonical Imd-pathway target AMPs such as *diptericin*, *attacin*, and *cecropin* are strongly induced in *H. cameraria* and *H. confusa*, but not in *S. deflexa*. Known immune effectors are highlighted in green and novel immune effectors predicted in this study are highlighted in blue.

We first constructed a de novo assembly of unmapped reads for each library, followed by 'diamond blastp' (Buchfink et al. 2021) searches against the NCBI 'nr' protein database for taxonomic assignment. The assembly of unmapped reads produced between 15,439 and 45,168 contigs per sample, and taxonomic assignment revealed that bacterial and viral taxa dominated the microbial profiles, with occasional hits to fungal and other lineages (Figure 4.6 A and B; Figure C.4). Bacterial composition included common Drosophila gut-associated taxa such as *Citrobacter*, *Fructilactobacillus*, and *Pseudomonas*. In addition, we found reads assigned to genera such as *Klebsiella*, *Escherichia*, *Streptococcus*, and *Salmonella*, which are rarely reported in Drosophila natural microbiome (Figure 4.6 A and B; Chandler et al. 2011; McMullen et al. 2021). Their presence here may reflect transient acquisition from rearing media, or lineage-specific associations unique to wild flies. Overall, these broad taxonomic patterns were generally similar between pathogen-challenged and unchallenged individuals within each species. However, more fine-scale differences emerged when we quantified within-sample microbial diversity (Figure 4.7). Notably, *H. cameraria* exhibited a significant increase in alpha diversity following pathogen challenge (Wilcoxon rank-sum test, $p = 0.01$), suggesting that infection alters the richness or evenness of microbial communities in this species (Figure 4.7). In contrast, microbial diversity remained stable across treatments in *H. confusa* and *S. deflexa* (Figure 4.7). This increase in microbial diversity in *H. cameraria* may be associated with susceptibility to *Providencia* infection. Indeed, *Providencia* reads were more abundant in pathogen-challenged *H. cameraria* individuals, suggesting successful bacterial replication. In *H. confusa*, *Providencia* was detected at lower levels and in fewer individuals, and *S. deflexa* showed little evidence of *Providencia* infection, with only two challenged individuals containing detectable levels (Figure 4.6 A). A few unchallenged individuals also showed low *Providencia* abundance, potentially due to environmental exposure or read contamination.

Interestingly, we detected high levels of *Spiroplasma* in all *S. deflexa* individuals. This vertically transmitted bacterial endosymbiont is known to protect some species of Drosophila against parasitoids, nematodes, and bacterial pathogens (Xie et al. 2011; Hamilton et al. 2016; Ballinger and Perlman 2017; Hrdina et al. 2024). This includes protection of *D. melanogaster* against *Providencia alcalifaciens* through host iron sequestration and enhanced melanization (Hrdina et al. 2024). Crucially, these defence strategies can operate independently of canonical Toll and Imd pathway AMP gene upregulation while still providing effective physiological protection, and this may not result in (host) transcriptional signatures of a response to infection. *Wolbachia* was also detected at moderate levels in *S. deflexa*, though its effects are known to be more context dependent (Wong et al. 2011; Gupta et al. 2017; Perlmutter et al. 2025).

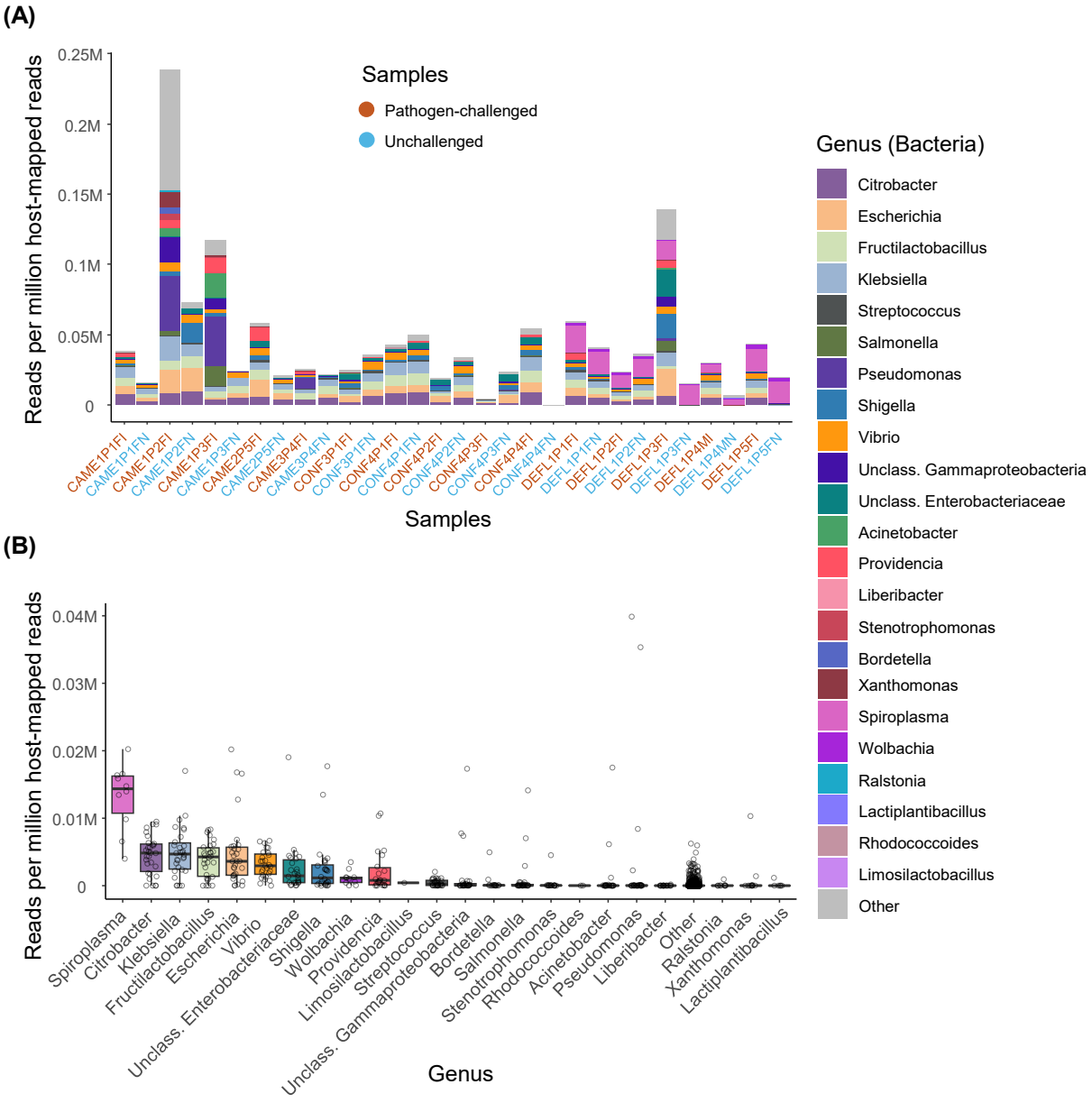

Figure 4.6: Microbiome composition in pathogen-challenged and unchallenged samples across three drosophilid species.

(A) Stacked bar plot showing the relative abundance of the top 23 bacterial genera (measured as reads per million host-mapped reads) across all samples. The 23 genera represent the union of the top 10 most abundant genera from each sample. Samples are ordered by species (*H. cameraria*, *H. confusa*, *S. deflexa*) and are colour-coded by treatment status: pathogen-challenged (red) and unchallenged (blue). (B) Boxplot summarizing the relative abundance of each bacterial genus across all samples. Each box represents the distribution of abundance for a given genus, with individual data points indicating values from individual samples. Genera are ordered by median abundance. *Spiroplasma* shows the highest median abundance overall, and it's only found in *S. deflexa*.

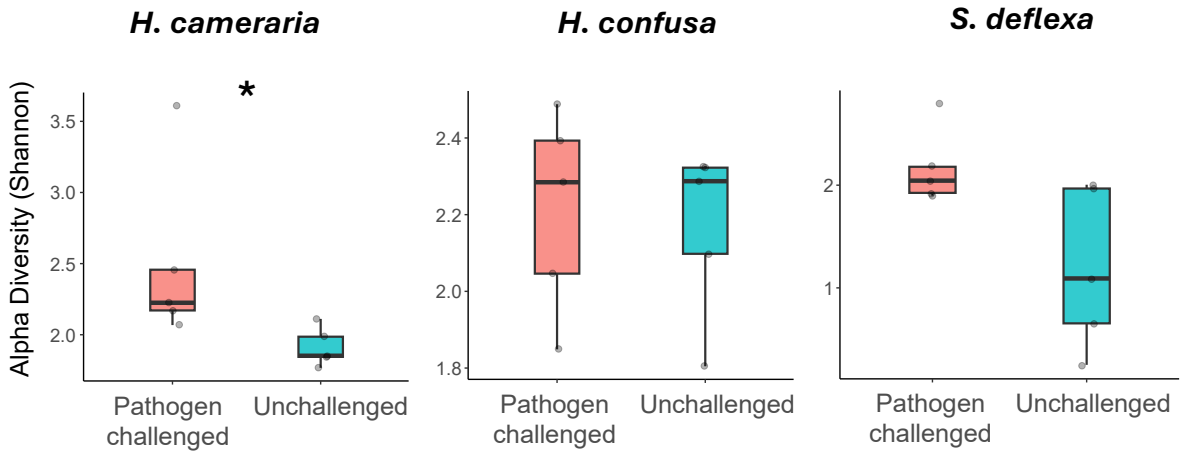

Figure 4.7: Alpha diversity in pathogen-challenged and unchallenged samples across three drosophilid species.
Boxplots of alpha diversity (Shannon index) comparing pathogen-challenged and unchallenged samples within each species. A significant increase in diversity is observed in *H. cameraria* upon infection (Wilcoxon rank-sum test, p value = 0.01), suggesting infection-induced shifts in microbial richness and/or evenness. No significant difference in alpha diversity was observed in *H. confusa* or *S. deflexa*.

4.4.5 Divergent species encode novel candidate AMP-like proteins

Despite the broadly conserved immune repertoire and similar transcriptional responses observed in *H. cameraria* and *H. confusa*, our analyses also revealed considerable species-specific differences—particularly in the apparently muted immune response of *S. deflexa*. These findings highlight the possibility that deeply diverged drosophilid species can encode lineage-specific immune effectors that are highly divergent in sequence and thus undetectable through simple homology searches against *D. melanogaster*. We hypothesized that such genes may include novel antimicrobial peptides, which are often short, secreted, cationic proteins with low levels of sequence conservation that makes them hard to detect (Moretta et al. 2020; Hanson and Hedelin 2025). To identify potential novel AMP-like candidates, we focused on genes differentially expressed in response to pathogen challenge (adj. $p < 0.05$, $\log_2$ FC ≥ 1) that lacked detectable orthologs in *D. melanogaster*, and encoded short peptides of the length expected for known AMPs (i.e. $\leq$200 amino acids). Across the three species, this approach yielded a total of 41 candidates, 22 from *H. cameraria*, 13 from *H. confusa*, and 6 from *S. deflexa* (Supplementary file C.3). Fourteen of these candidates were found

in multispecies hierarchical orthogroups (HOGs), suggesting that a subset may represent conserved but previously unannotated drosophilid AMP families not found in *D. melanogaster*. The remaining candidates appeared to be species-specific or present in low-copy, potentially orphan gene families.

To evaluate their potential to encode AMPs, we screened all 41 candidates using two AMP prediction tools, AMP Scanner v2 (Veltri et al. 2018) and amPEPpy (Lawrence et al. 2021), as well as SignalP 6.0 (Teufel et al. 2022) for signal peptide prediction. Thirty-three of the 41 candidates were predicted as AMPs by at least one method, and 20 of these also had predicted N-terminal signal peptides (Supplementary file C.3), indicating likely secretion. Based on physicochemical criteria, we further categorized the predicted AMPs into two groups: (i) Strong candidates (14) that encode positively charged (net charge +1 to +10), secreted peptides often enriched in glycine, proline, or cationic residues; and (ii) Likely candidates (6) that encode secreted peptides with weakly positive or negative net charge (–2 to +1), but are strongly induced by infection (Supplementary file C.3). Among the strong candidates, we identified five in *H. cameraria*, seven in *H. confusa*, and two in *S. deflexa*. Expression levels of these genes were generally high ($\log_2$ FC ranging from 1.08 to 8.30), and most were exclusively induced in pathogen-challenged individuals, supporting a role in infection response. Structural predictions using AlphaFold 3 (Abramson et al. 2024) revealed that many of these strong candidates adopt conformations typical of known AMPs, including amphipathic α-helices and $\alpha\beta$-sheet-rich peptides (Figure 4.8).

Several candidates exhibited striking similarity to well-characterized AMP classes. For example, the gene *Hconf/g6966* ($\sim$20% proline) resembles the proline-rich *apidaecins* of bees ($\sim$30% proline; Li et al. 2006), featuring four tandem repeats of potential mature peptides flanked by Furin cleavage motifs (RXRR). The gene *Hcam/g510* adopts a compact α-helical structure reminiscent of *IM18* (*Paillotin*; Tian et al. 2025) from *D. melanogaster*, while *Sdef/g5630* ($\sim$30% glycine) shows structural similarity to the glycine-rich AMP *holotricin* from *Aedes aegypti* ($\sim$49% glycine; Saucereau et al. 2022). Additional candidates included *Hconf/g6967*, a paralog of *g6966*, also similar to *apidaecins*; *Hcam/g5361*, with $\sim$64% identity to *D. grimshawi* CecC; and *Hcam/g2563*, which shares $\sim$46% identity with *D. grimshawi lysozyme P*. Notably, both *Hcam/g4126* and *Hconf/g9391* have the potential to form *lysozyme*-like structures (Figure 4.8). An especially intriguing example is *Hcam/g11853*, which was not predicted as an AMP by either amPEPpy (Lawrence et al. 2021) or AMP Scanner v2 (Veltri et al. 2018), yet was among the most strongly induced genes following infection ($\log_2$ FC = 8.70; Table 2). Its homolog *Hconf/g10777* was similarly upregulated ($\log_2$ FC = 6.12) and clustered in the same HOG (Supplementary file C.3). While blastp searches with this gene

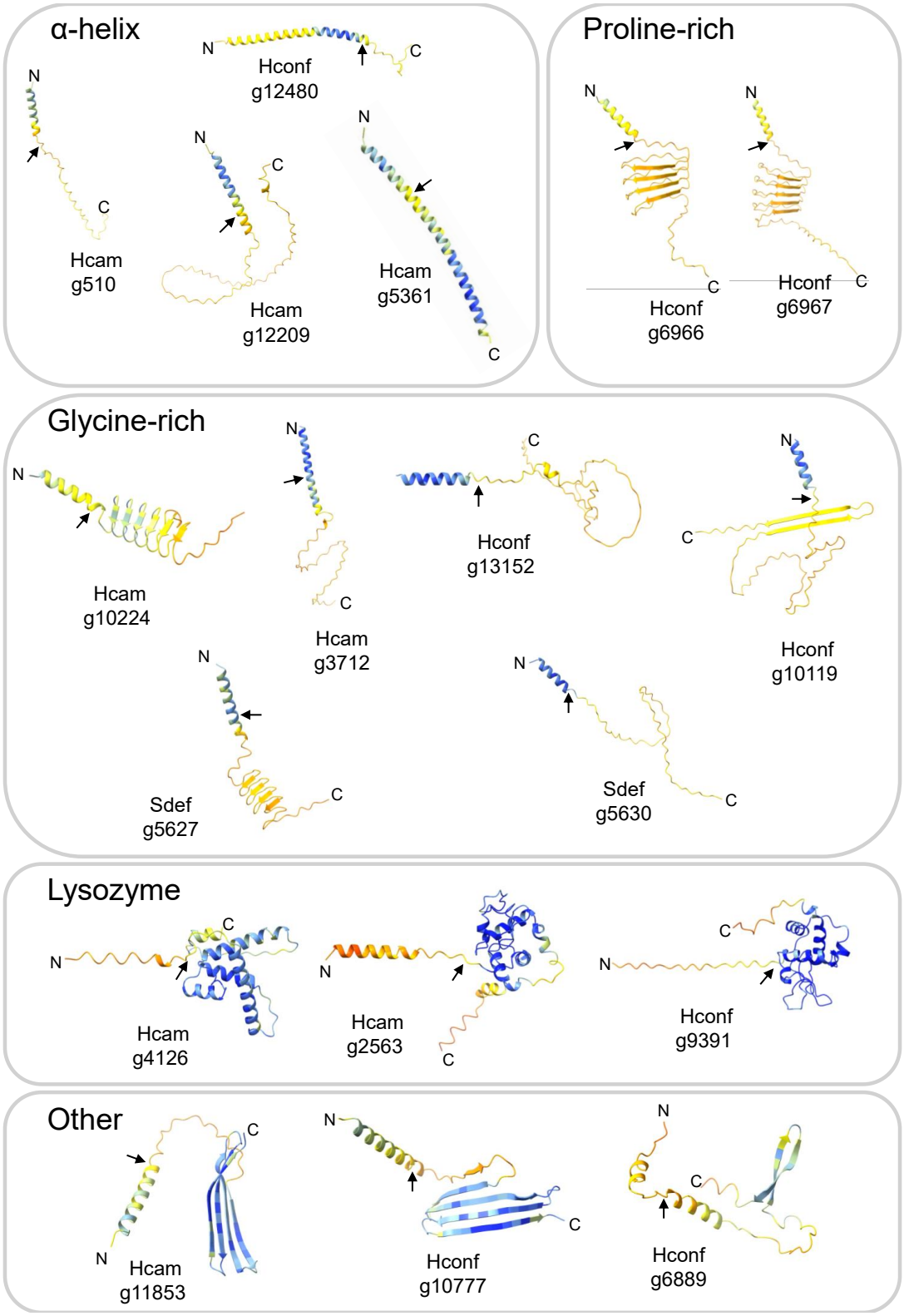

Figure 4.8: Predicted structures of novel AMP candidates.
AlphaFold3 predicted protein structures, N- and C-termini are labelled, and black arrows denote the predicted signal peptide cleavage sites based on SignalP6.0. The structures are coloured by pLDDT confidence score (blue = high accuracy, light blue = backbone expected to model well, yellow/orange = lower confidence, red = very low confidence).

returned only uncharacterised Drosophila proteins, weak hits to the *edin* ('elevated during infection') gene family suggest that these genes may represent highly divergent *edin*-like effectors that escape homology detection—underscoring the evolutionary lability of this gene family in Drosophilidae. Predicted 3-D structures for all strong and likely AMP candidates are presented in Figure 4.8. Taken together, their predicted secretion, AMP-like physicochemical properties, infection-induced expression, and structural similarity to known AMPs suggest that these genes likely encode novel immune effectors.

4.5 Discussion

In this study, we provide a comparative analysis of transcriptomic responses following bacterial-challenge with Gram-negative pathogen *Providencia rettgeri* in three wild-derived, non-model drosophilid species—*H. cameraria*, *H. confusa*, and *S. deflexa*. These species, which diverged from *D. melanogaster* over 45 million years ago, represent largely uncharacterised and ecologically diverse branches of the drosophilid phylogeny. Our goals were to demonstrate the challenges of differential expression analyses in near-wild flies, to assess whether canonical immune responses are conserved across these lineages, and to identify novel candidate immune effectors in taxa previously inaccessible to functional genomic studies. Through immune gene annotation, differential gene expression, microbiome profiling, and AMP prediction, our results reveal both conserved and lineage-specific features of insect immunity and highlight the value of including non-model species in comparative immune transcriptomic research.

Despite the intuitive appeal of using pathogen-stimulated transcriptomes to improve immune gene prediction in non-model species, we found that pathogen-challenged RNAseq did not substantially improve annotation completeness or immune gene recovery. BRAKER3 based annotations derived from pathogen-challenged, unchallenged, and combined RNAseq datasets yielded similar numbers of genes, with most orthologous genes consistently annotated across all datasets. While a few genes were unique to a single dataset, these rarely encoded immune-related functions and were often unassigned to orthogroups—suggesting lower confidence or assembly artefacts. These observations suggest that most immune genes have sufficient constitutive baseline expression to provide informative hints for de-novo gene prediction tools. Although RNAseq also indirectly informs gene annotation by revealing new transcript isoforms and splicing events, and increases confidence, it is thus not essential for the identification of novel genes.

By identifying homologs of a curated set of 638 immune-related *D. melanogaster* genes, we found that all three species shared a conserved core of immune signalling and recognition genes. However, as in previous studies, we observed extensive lineage-specific variation among effector genes, particularly antimicrobial peptides, which were highly variable in presence, copy number, and inducibility (Sackton et al. 2007; Sackton and Clark 2009; Hanson et al. 2016). Notably, *Scaptodrosophila deflexa* exhibited the lowest recovery of immune genes overall and lacked orthologs of several canonical AMPs, including *DptA*, *AttC*, and *AttD*. This pattern echoes previous finding in *Scaptodrosophila lebanonensis*, where a gene syntenic to *DptA* was identified but exhibited very little sequence similarity to any *diptericin* and in subgenus Drosophila is replaced by *DptC* (Hanson et al. 2016). It is therefore plausible that *DptA* has either been lost or has diverged beyond recognition in *S. deflexa*, potentially reflecting strong diversifying selection or relaxed constraint. In contrast, *H. cameraria* and *H. confusa* both harboured multiple paralogs of *DptC*, and *AttC*, suggesting lineage-specific expansions. These patterns of gain and loss were supported quantitatively by our Bayesian mixed-effects model, which showed that both species identity and gene functional category significantly predicted immune gene recovery. Effector genes were recovered with slightly, but significantly, lower probabilities than signalling genes, reinforcing the view that AMP gene families are particularly prone to evolutionary turnover. Among the species, *S. deflexa* showed the lowest predicted probability of immune gene recovery, again driven in part by the absence of some canonical AMP orthologs. These findings highlight the evolutionary lability of AMP gene families in Drosophilidae, characterized by frequent duplication, pseudogenization, and loss—hallmarks of strong, dynamic selective pressures likely imposed by pathogen diversity and host ecology. They also underscore the challenge of identifying fast-evolving immune effectors in non-model organisms, where high divergence can obscure orthology relationships and thereby our inability to functionally annotate these genes.

Interestingly, while both *H. cameraria* and *H. confusa* mounted robust transcriptional responses to *Providencia rettgeri* infection—upregulating canonical AMPs (such as *diptericins*, *attacins*, and *cecropins*), PGRP receptors, serine proteases, and transcriptional regulators (such as *Relish* and *pirk*)—*S. deflexa* appeared to exhibit minimal transcriptional changes, with only two PGRPs and no known AMPs upregulated. Assuming this was not simply lower power in our experiment, there are two likely explanations. First, *S. deflexa* individuals showed lower *Providencia* load, suggesting reduced infection burden or greater resistance. Second, all individuals harboured *Spiroplasma*, an endosymbiont known to protect flies from pathogenic bacteria (including *Providencia*) via mechanisms such as iron sequestration and melanization—defence strategies that bypass transcriptional AMP induction (Hrdina et al. 2024). The concurrent presence of *Wolbachia* might further perturb the host immune responses, poten-

tially by masking or dampening canonical transcriptional responses. In mosquitoes, experimental infection with *Wolbachia* protects against broad-spectrum anti-microbe and parasite by upregulating immune effector molecules and those involved in antimicrobial pathway (Kambris et al. 2009; Moreira et al. 2009; Hughes et al. 2011). Together, these findings suggest that both reduced *Providencia* infection success and the presence of defensive endosymbionts—especially *Spiroplasma*—may explain the weak immune transcriptional response observed in *S. deflexa*. More broadly, our results underscore the importance of profiling the microbiome when interpreting host transcriptional responses in non-model, wild-derived insects, where natural symbionts and background microbial variation may obscure or reshape immune phenotypes.

Finally, our identification of 41 novel AMP-like genes—many of which are not widely conserved across Drosophilidae, and are short, secreted, and structurally similar to known AMPs—suggests that AMP evolution may be even more dynamic than previously appreciated. Notably, many candidates shared structural features with known AMP families such as *apidaecins*, *holotricins*, *cecropins*, and *lysozymes*, but lacked clear sequence homology to any annotated *D. melanogaster* genes. Other candidates were not predicted to be AMPs, but were highly induced—including *Hcam/g11853* and its homolog *Hconf/g10777*, which could represent highly divergent homologs of *edin*. These results build on previous reports that lineage-specific AMPs are common in Drosophilidae (Sackton and Clark 2009; Hanson et al. 2016; Hanson et al. 2023) and emphasize that reliance on model species alone likely underestimates the diversity of immune effectors in nature.

4.6 Conclusions

This study expands the scope of functional immune genomics into non-model drosophilids and highlights both the conserved and lineage-specific components of insect immunity, particularly we report extreme divergence and presence of novel AMPs in distant lineages of Drosophilidae. Our findings advocate for a broader phylogenetic sampling in immunological research and demonstrate the feasibility of high-resolution transcriptomics in species that are not amenable to laboratory domestication. In the future, the functional validation of novel immune effectors, through antimicrobial assays or in vivo perturbations will be critical to understand their roles in host defence. Additionally, as genome and transcriptome data continue to accumulate across Drosophilidae, studies like this will be essential for understanding how immune systems evolve, diversify, and interact with microbial environments across evolutionary time.

4.7 Acknowledgements

We would like to thank the Friends of the Hermitage of Braid and City of Edinburgh Forestry and Natural Heritage for permission to collect flies. We also would like to thank Bernard Kim and Dmitri Petrov for the pre-publication Scaptodrosophila genome, and Katy Monteith and Pedro Vale for providing the bacterial aliquots.

Chapter 5

General Discussion

I wrote this chapter with minor comments from Prof. Darren Obbard. In this chapter, I have highlighted major thesis aims and results, with discussion on limitations and implications of the studies.

Across more than 400 million years of insect evolution, the innate immune system has been shaped by persistent antagonistic interactions with pathogens, resulting in some of the most rapidly evolving genes in animal genomes. Yet most work on immune gene evolution in Drosophila has focused on a small set of model species, particularly *D. melanogaster* and its close relatives. This bias leaves large swathes of unexplored drosophilid diversity comprising thousands of ecologically distinct species. This thesis aimed to address this gap in three complementary ways. First, I generated consistent protein-coding gene annotations across 304 drosophilid genomes, enabling robust comparative analyses at an unprecedented phylogenetic scale. Second, I used these resources to investigate the evolutionary dynamics of immune gene families, assessing both protein sequence divergence and gene turnover, and exploring how these dynamics vary across immune functional classes and pathways. Third, I generated pathogen-challenged RNAseq data for diverged non-model drosophilids, providing a critical resource to improve the annotation of rapidly evolving and lineage-specific immune genes. Finally, I applied comparative transcriptomics to investigate the transcriptional response to bacterial infection in three non-model drosophilids, identifying both conserved and lineage-specific immune responses.

5.1 Comparative gene annotations of 304 genomes

I generated standardized protein-coding annotations for 304 drosophilid genomes using a combination of the comparative annotation toolkit (CAT) and BRAKER3. The pipeline was designed to benefit from conserved gene structure between species while allowing recovery of lineage-specific transcripts. Annotation evidence included reference gene sets, alignments of conserved genomic regions, and external hints from RNAseq and protein data. This ensured that all annotations were produced under a single, internally consistent framework, suitable for downstream comparative analyses. To my knowledge, this represents the first individual effort to produce gene annotations at such a broad phylogenetic scale in insects. The resulting dataset spans all major drosophilid lineages and is explicitly built for cross-species comparability, enabling the application of phylogenetic models that account for shared ancestry and variation in assembly or annotation quality. Quality assessments revealed no substantial variation in overall gene number or coding sequence (CDS) length across most Drosophila species. However, a few species emerged as outliers. For instance, *D. vulcana* and *D. punjabiensis* exhibited unusually high gene counts, likely due to bacterial contamination in their assemblies, whereas *D. miranda*'s elevated gene numbers can be attributed to gene duplications associated with the evolution of its neo-Y chromosome (Bachtrog et al. 2019). Such deviations could also reflect annotation artefacts or assembly quality differences. Because multiple genomes were annotated simultaneously using the same pipeline, systematic annotation failures are improbable; such errors would have affected many species. While contig N50 is expected to have limited influence on gene content—given that all assemblies had N50 values >50 kb and the average *Drosophila* gene length is $\sim$2.5 kb. Current state-of-the-art pipelines cannot merge gene fragments located on different scaffolds or contigs, meaning that species with more fragmented assemblies could lead to broken gene models (see Figure 2.2; Mariene and Wasmuth 2025).

The resulting gene annotations dataset provide a valuable resource for comparative genomics across Drosophilidae. First, consistent annotations reduces technical noise when estimating orthology and gene family sizes, which is particularly important for dynamic gene families such as those involved in immunity (Fiddes et al. 2018). Second, functional annotations from *D. melanogaster* can be reliably transferred across species, facilitating evolutionary and functional analyses of poorly characterised genes. Third, the dataset fills a critical gap for clade-wide comparative genomics, enabling integrative analyses of gene sequence evolution, gene family turnover, and gene expression patterns in a phylogenetically informed framework.

Looking forward, the long-term value of this dataset will depend on how it is used, maintained, and extended. At present, these annotations offer a unique foundation for hypothesis driven analyses of gene family evolution and functional diversification across drosophilids. As sequencing technologies advance and assemblies improve to chromosome-level contiguity, many gene models will need updating. Short-read RNAseq, while useful for validation, cannot capture full-length isoforms. Moreover, Short-read RNAseq struggles with lowly expressed genes because short fragments require assembly into full transcripts, which is difficult for low-abundance transcripts due to potential ambiguity and computational limitations. These limitations raises the risk that parts of the immune repertoire or other rapidly evolving families remain systematically under-annotated. Future iterations could benefit from integrating long-read transcriptomics, pan-genome alignments, or even experimental validation of predicted transcripts. In retrospect, a useful next step would be to develop a community-driven, versioned annotation resource for drosophilids where new genomes and transcriptomes can be integrated under a unified framework. Such an approach would not only improve accuracy but also ensure that this dataset continues to grow as a shared resource for the field.

5.2 Evolution of immune gene families

In Chapter 3, I investigated the evolutionary dynamics of immune-related gene families across Drosophilidae, integrating measures of protein sequence divergence (dN/dS), sites under diversifying selection, and gene turnover rates. Across the family Drosophilidae, immune genes as a whole exhibited elevated dN/dS and a higher proportion of sites under diversifying selection compared to non-immune genes, consistent with the long-standing view that host–pathogen interactions impose unusually strong selective pressures. Yet the picture was more complex than a simple narrative of an ongoing “arms race”. While immune genes were more likely than non-immune genes to undergo some family size variation, their estimated turnover rates were lower on average. This suggests that immune gene repertoires are often stable across deep evolutionary time, punctuated by relatively rare expansions or losses concentrated in particular lineages or gene families.

One of the clearest patterns was the heterogeneity among functional classes and pathways. Core signalling components were highly conserved, as expected given their pleiotropic roles and essential functions in developmental and metabolic pathways. By contrast, receptor and effector genes—those that most directly interact with pathogens—showed some of the strongest signals of rapid evolution. Among effectors, antimicrobial peptides (AMPs) stood out. Some AMP families, such as *Defensin* and *IM18*, were stable across the family, but others, including

Cecropins and *Attacins*, showed striking lineage-specific expansions and losses. The *Diptericin* locus provides a particularly interesting case study: *DptB* has been gained or lost in association with ecological shifts, especially mushroom-feeding, where the absence of *Acetobacter* appears to relax selective pressure (Hanson et al. 2023). Whether this kind of ecology-driven filtering is general across AMPs remains an open question. The higher turnover rates I observe in effector families are consistent with such a process, but functional validation remains sparse outside a few well-studied cases. Perhaps the most surprising finding was the contrast between antiviral pathways. Genes in the cGAS–STING pathway showed the strongest signal of rapid protein evolution, whereas genes in the RNAi pathway were far more conserved. Both contribute to antiviral defence, yet their selective regimes appear starkly different. This raises several possibilities: that cGAS–STING is under more direct and dynamic interaction with viral antagonists, whereas RNAi may represent a more deeply conserved and less easily evaded defence; or alternatively, that RNAi functions are constrained by pleiotropy, limiting their capacity to diversify. Distinguishing between these explanations will require more functional and ecological work.

Overall, takeaways from these results: first, they challenge the temptation to treat “immune genes” as a single evolutionary category. The diversity of patterns—conservation in signalling genes, lineage-specific expansions in effectors, paradoxical stability in RNAi—suggests that selective regimes vary at fine functional and ecological scales. Second, they illustrate that the notion of a ubiquitous host–pathogen arms race is too simplistic. While arms-race dynamics may operate in some receptor/effector–pathogen interactions, much of immune gene evolution seems better described as ecology-dependent filtering, constraint, or lineage-specific bursts of innovations. Third, they highlight the difficulty, and perhaps futility, of averaging evolutionary rates across tens of millions of years. General trends are informative, but the most biologically meaningful insights often come from examining individual gene families in their ecological and phylogenetic context. Averaging across the entire Drosophilidae inevitably masks recent or lineage-specific adaptations. Species-level population data, rather than family-wide divergence estimates, are often better suited to detecting balancing selection or selective sweeps. Moreover, functional annotation of immune genes remains incomplete, and our reliance on *D. melanogaster* orthology may bias inferences, particularly for rapidly evolving families. A natural next step would be to integrate population-level data (e.g., McDonald–Kreitman test, compares variation within species vs divergence between species) with functional assays, ideally across species occupying different ecological niches.

5.3 Transcriptional response to bacterial infection in non-model drosophilids

In Chapter 4, first, I tested whether pathogen-challenged RNAseq could improve the annotation of immune genes in divergent, non-model drosophilids. In principle, such data should help recover highly divergent transcripts, particularly in gene families such as antimicrobial peptides (AMPs) that often escape homology-based annotation. However, in practice, the utility of infection-derived RNAseq for annotation was limited. While expression data can confirm existing annotations and occasionally highlight novel transcripts, my analyses showed that RNAseq alone is insufficient to systematically recover missing or highly diverged immune genes. This finding underscores the challenges of annotating fast-evolving immune families, and the limitations of short-read transcriptomics in identifying novel transcripts without complementary approaches, such as long-read isoform sequencing (Iso-Seq) or targeted homology searches guided by structural features (Ruperti et al. 2023; Zhao et al. 2024). Second, I applied comparative transcriptomics to investigate the immune response to bacterial infection in three non-model drosophilid species. Surprisingly, canonical immune pathways (Toll and Imd) were strongly induced mainly in two *Hirtodrosophila* species, but *Scaptodrosophila deflexa* showed muted response to infection. This difference suggests that while the core immune signalling modules remain conserved, the magnitude and breadth of transcriptional responses can diverge substantially across lineages. Such divergence may reflect ecological specialization, differences in constitutive immunity (relative role of cellular and humoral immunity), or shifts in the cost–benefit balance of mounting strong transcriptional responses.

Several important limitations of this study must be acknowledged. A key caveat is that the flies were not reared on standardized media; instead, they were maintained under conditions tailored to their natural diets. While this was necessary because of the difficulty of keeping these species in the laboratory, it raises the possibility that some of the observed transcriptional differences may reflect uncontrolled factors—such as baseline microbiome composition or background bacterial load—rather than host genotype alone. In addition, the infection experiments lacked wounding controls (i.e., sterile needle pricks), making it difficult to disentangle transcriptional responses to bacterial infection from responses to injury. Sample size was also a limitation, particularly in *S. deflexa*, which was derived from a single founder female. Low replication reduces power to detect consistent transcriptional signals and increases the risk of confounding by individual variation (Schurch et al. 2016; Degen and Medo 2025). Furthermore, my analysis focused solely on gene expression changes, without parallel measurements of

host survival, bacterial load, or physiological stress. These phenotypic assays would help connect transcriptional responses to the actual efficacy of the immune defence. Finally, while pathogen-challenged RNAseq highlighted potential novel AMPs, their functional relevance remains untested and requires experimental validation.

Despite these challenges, the study makes several important contributions. First, it provides new comparative transcriptomic resources for non-model drosophilids. Second, it demonstrates that while the broad architecture of the innate immune response is conserved (as highlighted in two *Hirtodrosophila* species), transcriptional divergence is common and may itself represent a target of selection. Third, it highlights the limits of short-read RNAseq for immune gene annotation, clarifying where future annotation efforts should focus. In summary, this chapter shows that while pathogen-challenged RNAseq does not improve discovery of lineage-specific immune genes, it provides critical insights into transcriptional divergence across drosophilids, complementing the sequence-level and gene turnover analyses of Chapter 3.

5.4 Future directions

This thesis establishes a comparative framework for studying the evolution of immune genes across an unprecedented breadth of drosophilid diversity, combining large-scale genome annotations, evolutionary analyses of immune gene families, and comparative transcriptomics in non-model species. While these data have provided new insights into how immune systems evolve across deep evolutionary timescales, they also raise as many questions as they answer. Below I outline several avenues for future work, organized around the major gaps and uncertainties that remain.

Linking immune gene evolution to ecological and microbial contexts

One of the most striking findings from this thesis is that immune gene evolution is highly heterogeneous—some families remain conserved while others undergo duplication or loss, and/or show evidences for rapid adaptive evolution. Yet, beyond specific examples such as *Diptericin B* and *Acetobacter* associations, we know only very little about the ecological forces that drive these lineage-specific patterns. The next step is to directly integrate ecological and microbiome data, including pathogen communities, symbionts, microbiome composition, and host diet. This integration could test whether immune gene losses (e.g., *DptB* in mushroom-feeding flies) consistently correspond to ecological shifts in microbial exposure, or whether

other factors such as population size, host life history, or developmental niche also play roles. The big unanswered question here is: Are patterns of immune gene diversification largely predictable from ecology, or do they reflect idiosyncratic evolutionary histories? Resolving this would move us from cataloguing immune diversity to explaining it.

Bridging the gap between genotype and phenotype

While this thesis documents sequence evolution, gene turnover, and transcriptional responses, it stops short of linking genetic variation to measurable effects on host immunity and fitness. This is a central challenge: sequence change is easy to detect, but understanding its functional consequences is far harder. For example, do lineage-specific AMPs or divergent regulatory responses translate into differences in survival, pathogen clearance, or competitive fitness? Addressing this requires experimental work in non-model species, including in vitro antimicrobial assays, heterologous expression, and gene knockouts/knockdowns. These studies would provide the crucial link between observed evolutionary change and actual adaptive benefits.

Evolution across timescales: from within-species variation to deep phylogeny

The comparative framework developed here primarily captures long-term evolutionary patterns across tens of millions of years. However, strong pathogen selection also operates at the population level, leaving signatures of local adaptation and balancing selection. For example, in *D. melanogaster* and *D. simulans* amino acid polymorphism S69R of *DptA* relates to host defence against *Providencia rettgeri* (Unckless et al. 2016; Mullinax et al. 2025), but we do not know how general this is across the clade. Extending analyses to population genomic data in non-model species could reveal whether similar trade-offs between resistance, tolerance, and fitness costs recur across lineages. A major unanswered question is whether the same genes are repeatedly targeted by selection, or whether local adaptation and long-term divergence largely involve distinct subsets of the immune system. Bridging lineage-specific and deep evolutionary dynamics is essential to understand whether we are truly observing an “arms race” or a more complex mosaic of selection pressures.

The problem of generalization in immune evolution

A recurring theme in this thesis is that while immune genes are on average among the fastest evolving in the genome, averaging across genes and timescales may be misleading. Some genes evolve rapidly in one lineage but are conserved in another, and even within a single pathway, receptors and effectors can behave very differently from signalling molecules. This raises a critical question for the field: Are we overstating the universality of the host–pathogen arms race? Current comparative genomics often assumes that rapid evolution is a hallmark of immune genes, but this thesis highlights that such patterns are highly uneven and context-dependent. A more productive approach may be to shift from genome-wide averages to case studies that combine evolutionary analyses with ecological and functional data, allowing us to identify the specific conditions under which arms race dynamics truly apply.

Methodological advances for immune gene annotation and analysis

A limitation of this thesis and of most comparative immunogenomic studies is the difficulty of annotating highly divergent immune genes. Short-read RNAseq proved insufficient for systematically recovering missing AMPs, while homology-based methods often fail due to extreme sequence divergence. Future work should leverage long-read transcriptomics (e.g., Iso-Seq, Nanopore cDNA sequencing) to resolve full-length immune transcripts, and integrate structural features to identify highly divergent and rapidly evolving proteins that defy homology-based searches.

Broadening the comparative framework beyond *Drosophila*

Finally, while drosophilids provide a uniquely tractable system, they represent only one insect lineage. Many of the questions raised here—about the predictability of immune gene turnover, the balance between conservation and innovation, and the ecological correlates of immune repertoire evolution—remain unanswered across insects more broadly. Extending these approaches to other Diptera, or even to non-dipteran insects, will reveal whether the patterns observed in drosophilids are general principles or lineage-specific peculiarities. Comparative analyses that bridge across insects and even into vertebrates could address the fundamental question: Are there universal rules governing immune gene evolution, or is the immune system's evolutionary trajectory always contingent on lineage-specific history and ecology?

In summary, this thesis provides a first step toward a comparative framework for understanding immune gene evolution across a large, phylogenetically diverse family Drosophilidae. But the larger questions remain open: What determines why some immune genes race ahead while others remain conserved? How do ecological interactions shape immune diversity? And to

what extent can we ever generalize about “immune gene evolution” when the selective landscape is so fragmented across time, space, and taxa? Tackling them will require integrating genomics, ecology, and functional biology in a way that moves beyond descriptive patterns toward mechanistic explanations.

Bibliography

Morgan, T. H. (1910). 'Sex limited inheritance in drosophila'. In: *Science* 32.812, pp. 120–122. DOI: 10.1126/SCIENCE.32.812.120.

Morgan, T. H. (1915). *The mechanism of Mendelian heredity*. New York: Holt.

Muller, H. J. (1932). *Further studies on the nature and causes of gene mutations.* Vol. 6. Menasha: International Congress of Genetics.

Mukherjee, A. S. and W. Beermann (1965). 'Synthesis of ribonucleic acid by the x-chromosomes of Drosophila melanogaster and the problem of dosage compensation [55]'. In: *Nature* 207.4998, pp. 785–786. DOI: 10.1038/207785A0.

Felsenstein, J. (1974). 'The evolutionary advantage of recombination'. In: *Genetics* 78.2, pp. 737–756. DOI: 10.1093/genetics/78.2.737.

Illmensee, K. and A. P. Mahowald (1974). 'Transplantation of posterior polar plasm in Drosophila. Induction of germ cells at the anterior pole of the egg.' In: *Proceedings of the National Academy of Sciences of the United States of America* 71.4, pp. 1016–1020. DOI: 10.1073/PNAS.71.4.1016.

Nüsslein-volhard, C. and E. Wieschaus (1980). 'Mutations affecting segment number and polarity in drosophila'. In: *Nature* 287.5785, pp. 795–801. DOI: 10.1038/287795A0.

Ashburner, M. (1981). 'Entomophagous and other bizarre Drosophilidae'. In: *The Genetics and Biology of Drosophila*. Vol. 3a. New York: Academic Press, pp. 375–421.

Cline, T. W. (1983). 'The interaction between daughterless and sex-lethal in triploids: A lethal sex-transforming maternal effect linking sex determination and dosage compensation in Drosophila melanogaster'. In: *Developmental Biology* 95.2, pp. 260–274. DOI: 10.1016/0012-1606(83)90027-1.

Kabsch, W. and C. Sander (1983). 'Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features'. In: *Biopolymers* 22.12, pp. 2577–2637. DOI: 10.1002/bip.360221211.

Bargiello, T. A. and M. W. Young (1984). 'Molecular genetics of a biological clock in Drosophila'. In: *Proceedings of the National Academy of Sciences* 81.7, pp. 2142–2146. DOI: 10.1073/PNAS.81.7.2142.

Bargiello, T. A., F. R. Jackson and M. W. Young (1984). 'Restoration of circadian behavioural rhythms by gene transfer in Drosophila'. In: *Nature* 312.5996, pp. 752–754. DOI: 10.1038/312752A0.

Reddy, P. et al. (1984). 'Molecular analysis of the period locus in Drosophila melanogaster and identification of a transcript involved in biological rhythms'. In: *Cell* 38.3, pp. 701–710. DOI: 10.1016/0092-8674(84)90265-4.

Zehring, W. A. et al. (1984). 'P-element transformation with period locus DNA restores rhythmicity to mutant, arrhythmic drosophila melanogaster'. In: *Cell* 39.2, pp. 369–376. DOI: 10.1016/0092-8674(84)90015-1.

Gehring, W. J. and Y. Hiromi (1986). 'Homeotic genes and the homeobox.' In: *Annual review of genetics* 20, pp. 147–173. DOI: 10.1146/ANNUREV.GE.20.120186.001051.

Louis J. (1986). 'Ecological specialization in the Drosophila melanogaster species subgroup: A case study of D. sechellia'. In: *Ecol. Genet.* 7, pp. 215–229.

Salz, H. K., T. W. Cline and P. Schedl (1987). 'Functional changes associated with structural alterations induced by mobilization of a P element inserted in the Sex-lethal gene of Drosophila.' In: *Genetics* 117.2, pp. 221–231. DOI: 10.1093/GENETICS/117.2.221.

Bryan, G. J. et al. (1988). 'Hawaiian courtship songs: Evolutionary innovation in communicaton signals of Drosophila'. In: *Science* 240.4849, pp. 217–219. DOI: 10.1126/SCIENCE.3127882.

Hughes, A. L. and M. Nei (1988). 'Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection'. In: *Nature* 335.6186, pp. 167–170. DOI: 10.1038/335167a0.

Salz, H. K. et al. (1989). 'The Drosophila female-specific sex-determination gene, Sex-lethal, has stage-, tissue-, and sex-specific RNAs suggesting multiple modes of regulation.' In: *Genes & development* 3.5, pp. 708–719. DOI: 10.1101/GAD.3.5.708.

Desalle, R. (1992). 'The phylogenetic relationships of flies in the family drosophilidae deduced from mtDNA sequences'. In: *Molecular Phylogenetics and Evolution* 1.1, pp. 31–40. DOI: 10.1016/1055-7903(92)90033-D.

Hultmark, D. (1993). 'Immune reactions in Drosophila and other insects: a model for innate immunity'. In: *Trends in Genetics* 9.5, pp. 178–183. DOI: 10.1016/0168-9525(93)90165-E.

Tepass, U., L. I. Fessler, A. Aziz and V. Hartenstein (1994). 'Embryonic origin of hemocytes and their relationship to cell death in Drosophila'. In: *Development* 120.7, pp. 1829–1837. DOI: 10.1242/dev.120.7.1829.

Benjamini, Y. and Y. Hochberg (1995). 'Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing'. English. In: *Journal of the Royal Statistical Society Series B-Statistical Methodology* 57.1, pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Eyre-Walker, A. and M. Bulmer (1995). 'Synonymous substitution rates in enterobacteria'. In: *Genetics* 140.4, pp. 1407–1412. DOI: 10.1093/genetics/140.4.1407.

Hoffmann, J. A. (1995). 'Innate immunity of insects'. In: *Current Opinion in Immunology* 7.1, pp. 4–10. DOI: 10.1016/0952-7915(95)80022-0.

Russo, C. A., N. Takezaki and M. Nei (1995). 'Molecular phylogeny and divergence times of drosophilid species.' In: *Molecular Biology and Evolution* 12.3, pp. 391–404. DOI: 10.1093/OXFORDJOURNALS.MOLBEV.A040214.

Franc, N. C., J. L. Dimarcq, M. Lagueux, J. Hoffmann and R. A. B. Ezekowitz (1996). 'Croquemort, a novel drosophila hemocyte/macrophage receptor that recognizes apoptotic cells'. In: *Immunity* 4.5, pp. 431–443. DOI: 10.1016/S1074-7613(00)80410-0.

Lemaitre, B., E. Nicolas, L. Michaut, J. M. Reichhart and J. A. Hoffmann (1996). 'The dorsoventral regulatory gene cassette spatzle/Toll/Cactus controls the potent antifungal response in Drosophila adults'. In: *Cell* 86.6, pp. 973–983. DOI: 10.1016/S0092-8674(00)80172-5.

Russo, J., S. Dupas, F. Frey, Y. Carton and A. B R E H E L I N (1996). 'Insect immunity: early events in the encapsulation process of parasitoid (Leptopilina boulardi) eggs in resistant and susceptible strains of Drosophila'. In: *Parasitology* 112.1, pp. 135–142. DOI: 10.1017/S0031182000065173.

Yan, R., S. Small, C. Desplan, C. R. Dearolf and J. E. Darnell (1996). 'Identification of a Stat gene that functions in Drosophila development'. In: *Cell* 84.3, pp. 421–430. DOI: 10.1016/S0092-8674(00)81287-8.

Campos-Ortega, J. A. (1998). 'The genetics of the Drosophila achaete-scute gene complex: a historical appraisal'. In: *Int. J. Dev. Biol* 42, pp. 291–297.

Ferrandon, D. et al. (1998). 'A drosomycin-GFP reporter transgene reveals a local immune response in Drosophila that is not dependent on the Toll pathway'. In: *EMBO Journal* 17.5, pp. 1217–1227. DOI: 10.1093/EMBOJ/17.5.1217.

Ramos-Onsins, S. and M. Aguadé (1998). 'Molecular Evolution of the Cecropin Multigene Family in Drosophila: Functional Genes vs. Pseudogenes'. In: *Genetics* 150.1, pp. 157–171. DOI: 10.1093/genetics/150.1.157.

Rogina, B., J. W. Vaupel, L. Partridge and S. L. Helfand (1998). 'Regulation of gene expression is preserved in aging Drosophila melanogaster'. In: *Current Biology* 8.8, pp. 475–478. DOI: 10.1016/S0960-9822(98)70184-8.

Sequencing Consortium, C. elegans (1998). 'Genome sequence of the nematode C. elegans: A platform for investigating biology'. In: *Science* 282.5396, pp. 2012–2018. DOI: 10.1126/SCIENCE.282.5396.2012.

Wu, L. P. and K. V. Anderson (1998). 'Regulated nuclear import of Rel proteins in the Drosophila immune response'. In: *Nature* 392.6671, pp. 93–97. DOI: 10.1038/32195.

Deutsch, M. and M. Long (1999). 'Intron—exon structures of eukaryotic model organisms'. In: *Nucleic Acids Research* 27.15, pp. 3219–3228. DOI: 10.1093/NAR/27.15.3219.

Wilkins, M. R. et al. (1999). 'Protein Identification and Analysis Tools in the ExPASy Server'. In: *2-D Proteome Analysis Protocols*. Humana Press, pp. 531–552. DOI: 10.1385/1-59259-584-7:531.

Adams, M. D. et al. (2000). 'The genome sequence of Drosophila melanogaster'. In: *Science* 287.5461, pp. 2185–2195. DOI: 10.1126/science.287.5461.2185.

Ashburner, M. et al. (2000). 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium'. In: *Nat Genet* 25.1, pp. 25–29. DOI: 10.1038/75556.

Duret, L. and D. Mouchiroud (2000). 'Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate'. In: *Molecular Biology and Evolution* 17.1, pp. 68–70. DOI: 10.1093/oxfordjournals.molbev.a026239.

Lagueux, M., E. Perrodou, E. A. Levashina, M. Capovilla and J. A. Hoffmann (2000). 'Constitutive expression of a complement-like protein in toll and JAK gain-of-function mutants of Drosophila'. In: *Proc Natl Acad Sci U S A* 97.21, pp. 11427–11432. DOI: 10.1073/pnas.97.21.11427.

Rice, P., I. Longden and A. Bleasby (2000). 'EMBOSS: the European Molecular Biology Open Software Suite'. In: *Trends Genet* 16.6, pp. 276–277. DOI: 10.1016/s0168-9525(00)02024-2.

Silverman, N. et al. (2000). 'A Drosophila IκB kinase complex required for Relish cleavage and antibacterial immunity'. In: *Genes & Development* 14.19, pp. 2461–2471. DOI: 10.1101/GAD.817800.

Tzou, P. et al. (2000). 'Tissue-specific inducible expression of antimicrobial peptide genes in Drosophila surface epithelia'. In: *Immunity* 13.5, pp. 737–748. DOI: 10.1016/S1074-7613(00)00072-8.

Brown, S., N. Hu and J. C. G. Hombría (2001). 'Identification of the first invertebrate interleukin JAK/STAT receptor, the Drosophila gene domeless'. In: *Current Biology* 11.21, pp. 1700–1705. DOI: 10.1016/S0960-9822(01)00524-3.

De Gregorio, E., P. T. Spellman, G. M. Rubin and B. Lemaitre (2001). 'Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays'. In: *Proc Natl Acad Sci U S A* 98.22, pp. 12590–12595. DOI: 10.1073/pnas.221458698.

Ekengren, S. and D. Hultmark (2001). 'A Family of Turandot-Related Genes in the Humoral Stress Response of Drosophila'. In: *Biochemical and Biophysical Research Communications* 284.4, pp. 998–1003. DOI: 10.1006/BBRC.2001.5067.

Georgel, P. et al. (2001). 'Drosophila Immune Deficiency (IMD) Is a Death Domain Protein that Activates Antibacterial Defense and Can Promote Apoptosis'. In: *Developmental Cell* 1.4, pp. 503–514. DOI: 10.1016/S1534-5807(01)00059-4.

Michel, T., J. M. Relchhart, J. A. Hoffmann and J. Royet (2001). 'Drosophila Toll is activated by Gram-positive bacteria through a circulating peptidoglycan recognition protein'. In: *Nature* 414.6865, pp. 756–759. DOI: 10.1038/414756A.

Rämet, M. et al. (2001). 'Drosophila Scavenger Receptor CI Is a Pattern Recognition Receptor for Bacteria'. In: *Immunity* 15.6, pp. 1027–1038. DOI: 10.1016/S1074-7613(01)00249-7.

Silverman, N. and T. Maniatis (2001). 'NF-kappaB signaling pathways in mammalian and insect innate immunity'. In: *Genes Dev* 15.18, pp. 2321–2342. DOI: 10.1101/gad.909001.

Bergman, C. M. et al. (2002). 'Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophilagenome'. In: *Genome Biology* 3.12. DOI: 10.1186/GB-2002-3-12-RESEARCH0086.

De Gregorio, E. (2002). 'The Toll and Imd pathways are the major regulators of the immune response in Drosophila'. In: *The EMBO Journal* 21.11, pp. 2568–2579. DOI: 10.1093/emboj/21.11.2568.

De Gregorio, E. et al. (2002). 'An immune-responsive Serpin regulates the melanization cascade in Drosophila'. In: *Developmental Cell* 3.4, pp. 581–592. DOI: 10.1016/S1534-5807(02)00267-8.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe and M. W. Feldman (2002). 'Evolutionary Rate in the Protein Interaction Network'. In: *Science* 296.5568, pp. 750–752. DOI: 10.1126/science.1068696.

Gottar, M. et al. (2002). 'The Drosophila immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein'. In: *Nature* 416.6881, pp. 640–644. DOI: 10.1038/nature734.

Hoffmann, J. A. and J. M. Reichhart (2002). 'Drosophila innate immunity: An evolutionary perspective'. In: *Nature Immunology* 3.2, pp. 121–126. DOI: 10.1038/NI0202-121.

Leulier, F., S. Vidal, K. Saigo, R. Ueda and B. Lemaitre (2002). 'Inducible expression of double-stranded RNA reveals a role for dFADD in the regulation of the antibacterial response in Drosophila adults'. In: *Current Biology* 12.12, pp. 996–1000. DOI: 10.1016/S0960-9822(02)00873-4.

Ligoxygakis, P. et al. (2002). 'A serpin mutant links Toll activation to melanization in the host defence of Drosophila'. In: *The EMBO Journal* 21.23, pp. 6330–6337. DOI: 10.1093/EMBOJ/CDF661.

Misra, S. et al. (2002). 'Annotation of the Drosophila melanogaster euchromatic genome: a systematic review'. In: *Genome Biology* 3.12, pp. 1–22. DOI: 10.1186/GB-2002-3-12-RESEARCH0083.

Takehana, A. et al. (2002). 'Overexpression of a pattern-recognition receptor, peptidoglycan-recognition protein-LE, activates imd/relish-mediated antibacterial defense and the prophenoloxidase cascade in Drosophila larvae'. In: *Proceedings of the National Academy of Sciences of the United States of America* 99.21, pp. 13705–13710. DOI: 10.1073/PNAS.212301199.

Agaisse, H., U. M. Petersen, M. Boutros, B. Mathey-Prevot and N. Perrimon (2003). 'Signaling role of hemocytes in Drosophila JAK/STAT-dependent response to septic injury'. In: *Dev Cell* 5.3, pp. 441–450. DOI: 10.1016/s1534-5807(03)00244-2.

Brown, S., N. Hu and J. C. G. Hombria (2003). 'Novel level of signalling control in the JAK/STAT pathway revealed by in situ visualisation of protein-protein interaction during Drosophiladevelopment'. In: *Development* 130.14, pp. 3077–3084. DOI: 10.1242/DEV.00535.

Evans, C. J., V. Hartenstein and U. Banerjee (2003). 'Thicker than blood: conserved mechanisms in Drosophila and vertebrate hematopoiesis'. In: *Dev Cell* 5.5, pp. 673–690. DOI: 10.1016/s1534-5807(03)00335-6.

Gobert, V. et al. (2003). 'Dual Activation of the Drosophila Toll Pathway by Two Pattern Recognition Receptors'. In: *Science* 302.5653, pp. 2126–2130. DOI: 10.1126/SCIENCE.1085432.

Jordan, I. K., Y. I. Wolf and E. V. Koonin (2003). 'No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly'. In: *BMC Evolutionary Biology* 3.1, p. 1. DOI: 10.1186/1471-2148-3-1.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren and E. S. Lander (2003). 'Sequencing and comparison of yeast species to identify genes and regulatory elements'. In: *Nature* 423.6937, pp. 241–254. DOI: 10.1038/NATURE01644.

Leulier, F. et al. (2003). 'The Drosophila immune system detects bacteria through specific peptidoglycan recognition'. In: *Nature Immunology* 4.5, pp. 478–484. DOI: 10.1038/NI922.

Powell, J. R., E. Sezzi, E. N. Moriyama, J. M. Gleason and A. Caccone (2003). 'Analysis of a shift in codon usage in Drosophila'. In: *J Mol Evol* 57 Suppl 1.0, pp. 214–25. DOI: 10.1007/s00239-003-0030-3.

Schlenke, T. A. and D. J. Begun (2003). 'Natural Selection Drives Drosophila Immune System Evolution'. In: *Genetics.* 164.4, pp. 1471–1480. DOI: 10.1093/genetics/164.4.1471.

Stöven, S. et al. (2003). 'Caspase-mediated processing of the drosophila NF-κB factor relish'. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.10, pp. 5991–5996. DOI: 10.1073/PNAS.1035902100.

Weber, A. N. et al. (2003). 'Binding of the Drosophila cytokine Spätzle to Toll is direct and establishes signaling'. In: *Nature Immunology* 4.8, pp. 794–800. DOI: 10.1038/NI955.

Kaneko, T. et al. (2004). 'Monomeric and polymeric gram-negative peptidoglycan but not purified LPS stimulate the Drosophila IMD pathway'. In: *Immunity* 20.5, pp. 637–649. DOI: 10.1016/S1074-7613(04)00104-9.

Konstantinidis, K. T. and J. M. Tiedje (2004). 'Trends between gene content and genome size in prokaryotic species with larger genomes'. In: *Proc Natl Acad Sci U S A* 101.9, pp. 3160–3165. DOI: 10.1073/pnas.0308653100.

Lippman, Z. et al. (2004). 'Role of transposable elements in heterochromatin and epigenetic control'. In: *Nature* 430.6998, pp. 471–476. DOI: 10.1038/NATURE02651.

Mallet, F. et al. (2004). 'The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology'. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.6, pp. 1731–1736. DOI: 10.1073/PNAS.0305763101.

Manaka, J. et al. (2004). 'Draper-mediated and phosphatidylserine-independent phagocytosis of apoptotic cells by Drosophila hemocytes/macrophages'. In: *Journal of Biological Chemistry* 279.46, pp. 48466–48476. DOI: 10.1074/jbc.M408597200.

Meyer, I. M. and R. Durbin (2004). 'Gene structure conservation aids similarity based gene prediction'. In: *Nucleic Acids Research* 32.2, pp. 776–783. DOI: 10.1093/NAR/GKH211.

Nurnberger, T., F. Brunner, B. Kemmerling and L. Piater (2004). 'Innate immunity in plants and animals: striking similarities and obvious differences'. In: *Immunol Rev* 198.1, pp. 249–266. DOI: 10.1111/j.0105-2896.2004.0119.x.

Pili-Floury, S. et al. (2004). 'In Vivo RNA Interference Analysis Reveals an Unexpected Role for GNBP1 in the Defense against Gram-positive Bacterial Infection in Drosophila Adults'. In: *Journal of Biological Chemistry* 279.13, pp. 12848–12853. DOI: 10.1074/jbc.M313324200.

Stanke, M., R. Steinkamp, S. Waack and B. Morgenstern (2004). 'AUGUSTUS: a web server for gene finding in eukaryotes'. In: *Nucleic Acids Research* 32.Web Server issue, W309. DOI: 10.1093/NAR/GKH379.

Tamura, K., S. Subramanian and S. Kumar (2004). 'Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks'. In: *Molecular Biology and Evolution* 21.1, pp. 36–44. DOI: 10.1093/MOLBEV/MSG236.

Besemer, J. and M. Borodovsky (2005). 'GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses'. In: *Nucleic Acids Research* 33.Web Server issue, W451. DOI: 10.1093/NAR/GKI487.

Castillejo-López, C. and U. Häcker (2005). 'The serine protease Sp7 is expressed in blood cells and regulates the melanization reaction in Drosophila'. In: *Biochemical and Biophysical Research Communications* 338.2, pp. 1075–1082. DOI: 10.1016/J.BBRC.2005.10.042.

David, J. R. et al. (2005). 'Male sterility at extreme temperatures: a significant but neglected phenomenon for understanding Drosophila climatic adaptations'. In: *J Evol Biol* 18.4, pp. 838–846. DOI: 10.1111/j.1420-9101.2005.00914.x.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke and F. H. Arnold (2005). 'Why highly expressed proteins evolve slowly'. In: *Proceedings of the National Academy of Sciences* 102.40, pp. 14338–14343. DOI: 10.1073/pnas.0504070102.

Hahn, M. W. and A. D. Kern (2005). 'Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks'. In: *Molecular Biology and Evolution* 22.4, pp. 803–806. DOI: 10.1093/molbev/msi072.

Hombría, J. C. G., S. Brown, S. Häder and M. P. Zeidler (2005). 'Characterisation of Upd2, a Drosophila JAK/STAT pathway ligand'. In: *Developmental Biology* 288.2, pp. 420–433. DOI: 10.1016/J.YDBIO.2005.09.040.

Irving, P. et al. (2005). 'New insights into Drosophila larval haemocyte functions through genome-wide analysis'. In: *Cellular Microbiology* 7.3, pp. 335–350. DOI: 10.1111/J.1462-5822.2004.00462.X.

Jiggins, F. M. and K. W. Kim (2005). 'The evolution of antifungal peptides in Drosophila'. In: *Genetics* 171.4, pp. 1847–1859. DOI: 10.1534/genetics.105.045435.

Jordan, I. K., L. Mariño-Ramírez and E. V. Koonin (2005). 'Evolutionary significance of gene expression divergence'. In: *Gene* 345.1, pp. 119–126. DOI: 10.1016/j.gene.2004.11.034.

Kaneko, T. and N. Silverman (2005). 'Bacterial recognition and signalling by the Drosophila IMD pathway'. In: *Cellular Microbiology* 7.4, pp. 461–469. DOI: 10.1111/J.1462-5822.2005.00504.X.

Kleino, A. et al. (2005). 'Inhibitor of apoptosis 2 and TAK1-binding protein are components of the Drosophila Imd pathway'. In: *EMBO Journal* 24.19, pp. 3423–3434. DOI: 10.1038/SJ.EMBOJ.7600807.

Kocks, C. et al. (2005). 'Eater, a Transmembrane Protein Mediating Phagocytosis of Bacterial Pathogens in Drosophila'. In: *Cell* 123.2, pp. 335–346. DOI: 10.1016/J.CELL.2005.08.034.

Kosakovsky Pond, S. L. and S. D. W. Frost (2005). 'Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection'. In: *Molecular Biology and Evolution* 22.5, pp. 1208–1222. DOI: 10.1093/molbev/msi105.

Lazzaro, B. P. (2005). 'Elevated Polymorphism and Divergence in the Class C Scavenger Receptors of Drosophila melanogaster and D. simulansSequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AY865019, AY865135.' In: *Genetics* 169.4, pp. 2023–2034. DOI: 10.1534/genetics.104.034249.

Lomsadze, A., V. Ter-Hovhannisyan, Y. O. Chernoff and M. Borodovsky (2005). 'Gene identification in novel eukaryotic genomes by self-training algorithm'. In: *Nucleic Acids Research* 33.20, pp. 6494–6506. DOI: 10.1093/NAR/GKI937.

Nei, M. and A. P. Rooney (2005). 'Concerted and Birth-and-Death Evolution of Multigene Families'. In: *Annual Review of Genetics* 39.1, pp. 121–152. DOI: 10.1146/annurev.genet.39.073003.112240.

Nielsen, R. et al. (2005). 'A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees'. In: *PLOS Biology* 3.6, e170. DOI: 10.1371/journal.pbio.0030170.

Philips, J. A., E. J. Rubin and N. Perrimon (2005). 'Drosophila RNAi screen reveals CD36 family member required for mycobacterial infection'. In: *Science* 309.5738, pp. 1251–1253. DOI: 10.1126/SCIENCE.1116006.

Quesada, H., S. E. Ramos-Onsins and M. Aguade (2005). 'Birth-and-death evolution of the Cecropin multigene family in Drosophila'. In: *J Mol Evol* 60.1, pp. 1–11. DOI: 10.1007/s00239-004-0053-4.

Richards, S. et al. (2005). 'Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution'. In: *Genome Res* 15.1, pp. 1–18. DOI: 10.1101/gr.3059305.

Tennessen, J. A. (2005). 'Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection'. In: *Journal of Evolutionary Biology* 18.6, pp. 1387–1394. DOI: 10.1111/j.1420-9101.2005.00925.x.

Watson, F. L. et al. (2005). 'Immunology: Extensive diversity of Ig-superfamily proteins in the immune system of insects'. In: *Science* 309.5742, pp. 1874–1878. DOI: 10.1126/SCIENCE.1116887.

Beller, M. and B. Oliver (2006). 'One hundred years of high-throughput Drosophila research'. In: *Chromosome Res* 14.4, pp. 349–362. DOI: 10.1007/s10577-006-1065-2.

Drummond, A. J., S. Y. Ho, M. J. Phillips and A. Rambaut (2006). 'Relaxed phylogenetics and dating with confidence'. In: *PLoS Biol* 4.5, e88. DOI: 10.1371/journal.pbio.0040088.

Gottar, M. et al. (2006). 'Dual Detection of Fungal Infections in Drosophila via Recognition of Glucans and Sensing of Virulence Factors'. In: *Cell* 127.7, pp. 1425–1437. DOI: 10.1016/j.cell.2006.10.046.

Ingvarsson, P. K. (2006). 'Gene Expression and Protein Length Influence Codon Usage and Rates of Sequence Evolution in Populus tremula'. In: *Molecular Biology and Evolution* 24.3, pp. 836–844. DOI: 10.1093/molbev/msl212.

Jang, I. H. et al. (2006). 'A Spätzle-processing enzyme required for toll signaling activation in drosophila innate immunity'. In: *Developmental Cell* 10.1, pp. 45–55. DOI: 10.1016/J.DEVCEL.2005.11.013.

Jiggins, F. M. and K. W. Kim (2006). 'Contrasting evolutionary patterns in Drosophila immune receptors'. In: *Journal of Molecular Evolution* 63.6, pp. 769–780. DOI: 10.1007/S00239-006-0005-2.

Kambris, Z. et al. (2006). 'Drosophila Immunity: A Large-Scale In Vivo RNAi Screen Identifies Five Serine Proteases Required for Toll Activation'. In: *Current Biology* 16.8, pp. 808–813. DOI: 10.1016/j.cub.2006.03.020.

Kaneko, T. et al. (2006). 'PGRP-LC and PGRP-LE have essential yet distinct functions in the drosophila immune response to monomeric DAP-type peptidoglycan'. In: *Nature Immunology* 7.7, pp. 715–723. DOI: 10.1038/NI1356.

Li, W. F., G. X. Ma and X. X. Zhou (2006). 'Apidaecin-type peptides: biodiversity, structure-function relationships and mode of action'. In: *Peptides* 27.9, pp. 2350–2359. DOI: 10.1016/j.peptides.2006.03.016.

Lim, J. H. et al. (2006). 'Structural Basis for Preferential Recognition of Diaminopimelic Acid-type Peptidoglycan by a Subset of Peptidoglycan Recognition Proteins'. In: *Journal of Biological Chemistry* 281.12, pp. 8286–8295. DOI: 10.1074/JBC.M513030200.

Obbard, D. J., F. M. Jiggins, D. L. Halligan and T. J. Little (2006). 'Natural Selection Drives Extremely Rapid Evolution in Antiviral RNAi Genes'. In: *Current Biology* 16.6, pp. 580–585. DOI: 10.1016/j.cub.2006.01.065.

Presgraves, D. C. (2006). 'Intron Length Evolution in Drosophila'. In: *Molecular Biology and Evolution* 23.11, pp. 2203–2213. DOI: 10.1093/MOLBEV/MSL094.

Saleh, M. C. et al. (2006). 'The endocytic pathway mediates cell entry of dsRNA to induce RNAi silencing'. In: *Nature Cell Biology* 8.8, pp. 793–802. DOI: 10.1038/NCB1439.

Singh, N. D., P. F. Arndt and D. A. Petrov (2006). 'Minor shift in background substitutional patterns in the Drosophila saltans and willistoni lineages is insufficient to explain GC content of coding sequences'. In: *BMC Biol* 4.1, p. 37. DOI: 10.1186/1741-7007-4-37.

Stanke, M. et al. (2006). 'AUGUSTUS: ab initio prediction of alternative transcripts'. In: *Nucleic Acids Research* 34, W435–W439. DOI: 10.1093/NAR/GKL200.

Tang, H., Z. Kambris, B. Lemaitre and C. Hashimoto (2006). 'Two Proteases Defining a Melanization Cascade in the Immune System of Drosophila'. In: *Journal of Biological Chemistry* 281.38, pp. 28097–28104. DOI: 10.1074/JBC.M601642200.

Wang, X.-H. et al. (2006). 'RNA Interference Directs Innate Immunity Against Viruses in Adult Drosophila'. In: *Science* 312.5772, pp. 452–454. DOI: 10.1126/science.1125694.

Ao, J., E. Ling and X. Q. Yu (2007). 'Drosophila C-type lectins enhance cellular encapsulation'. In: *Molecular Immunology* 44.10, pp. 2541–2548. DOI: 10.1016/J.MOLIMM.2006.12.024.

Bidla, G., M. S. Dushay and U. Theopold (2007). 'Crystal cell rupture after injury in Drosophila requires the JNK pathway, small GTPases and the TNF homolog Eiger'. In: *Journal of Cell Science* 120.7, pp. 1209–1215. DOI: 10.1242/JCS.03420.

Blankenberg, D. et al. (2007). 'A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly'. In: *Genome Res* 17.6, pp. 960–964. DOI: 10.1101/gr.5578007.

Clark, A. G. et al. (2007). 'Evolution of genes and genomes on the Drosophila phylogeny'. In: *Nature* 450.7167, pp. 203–218. DOI: 10.1038/nature06341.

Crozatier, M. and M. Meister (2007). 'Drosophila haematopoiesis'. In: *Cellular Microbiology* 9.5, pp. 1117–1126. DOI: 10.1111/J.1462-5822.2007.00930.X.

Edwards, K. A., L. T. Doescher, K. Y. Kaneshiro and D. Yamamoto (2007). 'A Database of Wing Diversity in the Hawaiian Drosophila'. In: *PLOS ONE* 2.5, e487. DOI: 10.1371/JOURNAL.PONE.0000487.

Gu, X. and Z. Su (2007). 'Tissue-driven hypothesis of genomic evolution and sequence-expression correlations'. In: *Proceedings of the National Academy of Sciences* 104.8, pp. 2779–2784. DOI: 10.1073/pnas.0610797104.

Hahn, M. W., M. V. Han and S. G. Han (2007). 'Gene Family Evolution across 12 Drosophila Genomes'. In: *PLOS Genetics* 3.11, e197. DOI: 10.1371/JOURNAL.PGEN.0030197.

Heger, A. and C. P. Ponting (2007). 'Variable strength of translational selection among 12 Drosophila species'. In: *Genetics* 177.3, pp. 1337–1348. DOI: 10.1534/genetics.107.070466.

Jiggins, F. M. and K. W. Kim (2007). 'A screen for immunity genes evolving under positive selection in Drosophila'. In: *Journal of Evolutionary Biology* 20.3, pp. 965–970. DOI: 10.1111/j.1420-9101.2007.01305.x.

Krzemień, J. et al. (2007). 'Control of blood cell homeostasis in Drosophila larvae by the posterior signalling centre'. In: *Nature* 446.7133, pp. 325–328. DOI: 10.1038/NATURE05650.

Kurucz, É. et al. (2007). 'Nimrod, a Putative Phagocytosis Receptor with EGF Repeats in Drosophila Plasmatocytes'. In: *Current Biology* 17.7, pp. 649–654. DOI: 10.1016/J.CUB.2007.02.041.

Lemaitre, B. and J. Hoffmann (2007). 'The host defense of Drosophila melanogaster'. In: *Annu Rev Immunol* 25.1, pp. 697–743. DOI: 10.1146/annurev.immunol.25.022106.141615.

Lin, M. F. et al. (2007). 'Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes'. In: *Genome Research* 17.12, pp. 1823–1836. DOI: 10.1101/GR.6679507.

Loewe, L. and B. Charlesworth (2007). 'Background Selection in Single Genes May Explain Patterns of Codon Bias'. In: *Genetics* 175.3, pp. 1381–1393. DOI: 10.1534/genetics.106.065557.

Sackton, T. B. et al. (2007). 'Dynamic evolution of the innate immune system in Drosophila'. In: *Nat Genet* 39.12, pp. 1461–1468. DOI: 10.1038/ng.2007.60.

Stark, A. et al. (2007). 'Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures'. In: *Nature* 450.7167, pp. 219–232. DOI: 10.1038/NATURE06340.

Tanji, T., X. Hu, A. N. R. Weber and Y. T. Ip (2007). 'Toll and IMD Pathways Synergistically Activate an Innate Immune Response in Drosophila melanogaster'. In: *Molecular and Cellular Biology* 27.12. DOI: 10.1128/MCB.01814-06.

Vicario, S., E. N. Moriyama and J. R. Powell (2007). 'Codon usage in twelve species of Drosophila'. In: *BMC Evol Biol* 7.1, p. 226. DOI: 10.1186/1471-2148-7-226.

Waterhouse, R. M. et al. (2007). 'Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes'. In: *Science* 316.5832, pp. 1738–1743. DOI: 10.1126/science.1139862.

Cantarel, B. L. et al. (2008). 'MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes'. In: *Genome Res* 18.1, pp. 188–196. DOI: 10.1101/gr.6743907.

Comeron, J. M., A. Williford and R. M. Kliman (2008). 'The Hill–Robertson effect: evolutionary consequences of weak selection and linkage in finite populations'. In: *Heredity* 100.1, pp. 19–31. DOI: 10.1038/sj.hdy.6801059.

Dai, H. et al. (2008). 'The evolution of courtship behaviors through the origination of a new gene in Drosophila'. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.21, pp. 7478–7483. DOI: 10.1073/PNAS.0800693105.

Gernhard, T. (2008). 'The conditioned reconstructed process'. In: *J Theor Biol* 253.4, pp. 769–778. DOI: 10.1016/j.jtbi.2008.04.005.

Hershberg, R. and D. A. Petrov (2008). 'Selection on codon bias'. In: *Annu Rev Genet* 42.Volume 42, 2008, pp. 287–299. DOI: 10.1146/annurev.genet.42.110807.091442.

Hollox, E. J. and J. A. Armour (2008). 'Directional and balancing selection in human beta-defensins'. In: *BMC Evolutionary Biology* 8.1, pp. 1–14. DOI: 10.1186/1471-2148-8-113.

Larracuente, A. M. et al. (2008). 'Evolution of protein-coding genes in Drosophila'. In: *Trends in Genetics* 24.3, pp. 114–123. DOI: 10.1016/j.tig.2007.12.001.

Leone, P. et al. (2008). 'Crystal structure of Drosophila PGRP-SD suggests binding to DAP-type but not lysine-type peptidoglycan'. In: *Molecular Immunology* 45.9, pp. 2521–2530. DOI: 10.1016/J.MOLIMM.2008.01.015.

Lhocine, N. et al. (2008). 'PIMS modulates immune tolerance by negatively regulating Drosophila innate immune signaling'. In: *Cell Host Microbe* 4.2, pp. 147–158. DOI: 10.1016/j.chom.2008.07.004.

Lindmo, K. et al. (2008). 'The PI 3-kinase regulator Vps15 is required for autophagic clearance of protein aggregates'. In: *Autophagy* 4.4, pp. 500–506. DOI: 10.4161/auto.5829.

Markow, T. A. and P. O'Grady (2008). 'Reproductive Ecology of Drosophila'. In: *Functional Ecology* 22.5, pp. 747–759.

Moncrieffe, M. C., J. G. Grossmann and N. J. Gay (2008). 'Assembly of Oligomeric Death Domain Complexes during Toll Receptor Signaling'. In: *Journal of Biological Chemistry* 283.48, pp. 33447–33454. DOI: 10.1074/JBC.M805427200.

Ryu, J. H. et al. (2008). 'Innate immune homeostasis by the homeobox gene Caudal and commensal-gut mutualism in Drosophila'. In: *Science* 319.5864, pp. 777–782. DOI: 10.1126/SCIENCE.1149357.

Scherfer, C. et al. (2008). 'Drosophila Serpin-28D regulates hemolymph phenoloxidase activity and adult pigmentation'. In: *Developmental Biology* 323.2, pp. 189–196. DOI: 10.1016/J.YDBIO.2008.08.030.

Stanke, M., M. Diekhans, R. Baertsch and D. Haussler (2008). 'Using native and syntenically mapped cDNA alignments to improve de novo gene finding'. In: *Bioinformatics* 24.5, pp. 637–644. DOI: 10.1093/bioinformatics/btn013.

Tang, H., Z. Kambris, B. Lemaitre and C. Hashimoto (2008). 'A Serpin that Regulates Immune Melanization in the Respiratory System of Drosophila'. In: *Developmental Cell* 15.4, pp. 617–626. DOI: 10.1016/j.devcel.2008.08.017.

Tennessen, J. A. and M. S. Blouin (2008). 'Balancing Selection at a Frog Antimicrobial Peptide Locus: Fluctuating Immune Effector Alleles?' In: *Molecular Biology and Evolution* 25.12, pp. 2669–2680. DOI: 10.1093/MOLBEV/MSN208.

Van Der Linde, K. et al. (2008). 'A supertree analysis and literature review of the genus Drosophila and closely related genera (Diptera, Drosophilidae)'. In: *Insect Systematics and Evolution* 39.3, pp. 241–267. DOI: 10.1163/187631208788784237.

Viljakainen, L. and P. Pamilo (2008). 'Selection on an antimicrobial peptide defensin in ants'. In: *Journal of Molecular Evolution* 67.6, pp. 643–652. DOI: 10.1007/S00239-008-9173-6.

Aliyari, R. and S. W. Ding (2009). 'RNA-based viral immunity initiated by the Dicer family of host immune receptors'. In: *Immunological Reviews* 227.1, pp. 176–188. DOI: 10.1111/J.1600-065X.2008.00722.X.

Buchon, N. et al. (2009). 'A single modular serine protease integrates signals from pattern-recognition receptors upstream of the Drosophila Toll pathway'. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.30, pp. 12442–12447. DOI: 10.1073/PNAS.0901924106.

Charlesworth, B., A. J. Betancourt, V. B. Kaiser and I. Gordo (2009). 'Genetic Recombination and Molecular Evolution'. In: *Cold Spring Harbor Symposia on Quantitative Biology* 74.0, pp. 177–186. DOI: 10.1101/sqb.2009.74.015.

Colinet, D. et al. (2009). 'A serpin from the parasitoid wasp Leptopilina boulardi targets the Drosophila phenoloxidase cascade'. In: *Developmental & Comparative Immunology* 33.5, pp. 681–689. DOI: 10.1016/J.DCI.2008.11.013.

Costa, A., E. Jan, P. Sarnow and D. Schneider (2009). 'The Imd Pathway Is Involved in Antiviral Immune Responses in Drosophila'. In: *PLoS One* 4.10, e7436. DOI: 10.1371/journal.pone.0007436.

Dworkin, I. and C. D. Jones (2009). 'Genetic Changes Accompanying the Evolution of Host Specialization in Drosophila sechellia'. In: *Genetics* 181.2, p. 721. DOI: 10.1534/GENETICS.108.093419.

Freckleton, R. P. (2009). 'The seven deadly sins of comparative analysis'. In: *J Evol Biol* 22.7, pp. 1367–1375. DOI: 10.1111/j.1420-9101.2009.01757.x.

Gao, H., X. Wu and N. Fossett (2009). ' Upregulation of the Drosophila Friend of GATA Gene u-shaped by JAK/STAT Signaling Maintains Lymph Gland Prohemocyte Potency '. In: *Molecular and Cellular Biology* 29.22, pp. 6086–6096. DOI: 10.1128/MCB.00244-09.

Juneja, P. and B. P. Lazzaro (2009). 'Providencia sneebia sp. nov. and Providencia burhodogranariea sp. nov., isolated from wild Drosophila melanogaster'. In: *Int J Syst Evol Microbiol* 59.Pt 5, pp. 1108–1111. DOI: 10.1099/ijs.0.000117-0.

Kambris, Z., P. E. Cook, H. K. Phuc and S. P. Sinkins (2009). 'Immune activation by life-shortening Wolbachia and reduced filarial competence in mosquitoes'. In: *Science* 326.5949, pp. 134–136. DOI: 10.1126/science.1177531.

Morales-Hojas, R., C. P. Vieira, M. Reis and J. Vieira (2009). 'Comparative analysis of five immunity-related genes reveals different levels of adaptive evolution in the virilis and melanogaster groups of Drosophila'. In: *Heredity (Edinb)* 102.6, pp. 573–578. DOI: 10.1038/hdy.2009.11.

Moreira, L. A. et al. (2009). 'A Wolbachia symbiont in Aedes aegypti limits infection with dengue, Chikungunya, and Plasmodium'. In: *Cell* 139.7, pp. 1268–1278. DOI: 10.1016/j.cell.2009.11.042.

Obbard, D. J., J. J. Welch, K. W. Kim and F. M. Jiggins (2009a). 'Quantifying adaptive evolution in the Drosophila immune system'. In: *PLoS Genet* 5.10, e1000698. DOI: 10.1371/journal.pgen.1000698.

Obbard, D. J., K. H. J. Gordon, A. H. Buck and F. M. Jiggins (2009b). 'The evolution of RNAi as a defence against viruses and transposable elements'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1513, pp. 99–115. DOI: 10.1098/rstb.2008.0168.

Petit, N. and A. Barbadilla (2009). 'Selection efficiency and effective population size in Drosophila species'. In: *J Evol Biol* 22.3, pp. 515–526. DOI: 10.1111/j.1420-9101.2008.01672.x.

Reis, M. dos and L. Wernisch (2009). 'Estimating translational selection in eukaryotic genomes'. In: *Mol Biol Evol* 26.2, pp. 451–461. DOI: 10.1093/molbev/msn272.

Sackton, T. B. and A. G. Clark (2009). 'Comparative profiling of the transcriptional response to infection in two species of Drosophila by short-read cDNA sequencing'. In: *BMC Genomics* 10.1, p. 259. DOI: 10.1186/1471-2164-10-259.

Saleh, M. C. et al. (2009). 'Antiviral immunity in Drosophila requires systemic RNA interference spread'. In: *Nature* 458.7236, pp. 346–350. DOI: 10.1038/NATURE07712.

Shinzawa, N. et al. (2009). 'p38 MAPK-Dependent Phagocytic Encapsulation Confers Infection Tolerance in Drosophila'. In: *Cell Host and Microbe* 6.3, pp. 244–252. DOI: 10.1016/j.chom.2009.07.010.

Viljakainen, L. et al. (2009). 'Rapid Evolution of Immune Proteins in Social Insects'. In: *Molecular Biology and Evolution* 26.8, pp. 1791–1801. DOI: 10.1093/molbev/msp086.

Bulmer, M. S., F. Lay and C. Hamilton (2010). 'Adaptive evolution in subterranean termite antifungal peptides'. In: *Insect Molecular Biology* 19.5, pp. 669–674. DOI: 10.1111/J.1365-2583.2010.01023.X.

Cherry, J. L. (2010). 'Expression Level, Evolutionary Rate, and the Cost of Expression'. In: *Genome Biology and Evolution* 2.0, pp. 757–769. DOI: 10.1093/gbe/evq059.

Ding, S. W. (2010). 'RNA-based antiviral immunity'. In: *Nature Reviews Immunology* 10.9, pp. 632–644. DOI: 10.1038/NRI2824.

Hadfield, J. D. (2010). 'MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package'. English. In: *Journal of Statistical Software* 33.2, pp. 1–22. DOI: DOI10.18637/jss.v033.i02.

Hadfield, J. D. and S. Nakagawa (2010). 'General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters'. In: *J Evol Biol* 23.3, pp. 494–508. DOI: 10.1111/j.1420-9101.2009.01915.x.

Johnson, L. S., S. R. Eddy and E. Portugaly (2010). 'Hidden Markov model speed heuristic and iterative HMM search procedure'. In: *BMC Bioinformatics* 11.1, p. 431. DOI: 10.1186/1471-2105-11-431.

Makki, R. et al. (2010). 'A Short Receptor Downregulates JAK/STAT Signalling to Control the Drosophila Cellular Immune Response'. In: *PLOS Biology* 8.8, e1000441. DOI: 10.1371/JOURNAL.PBIO.1000441.

Pang, K., C. Cheng, Z. Xuan, H. Sheng and X. Ma (2010). 'Understanding protein evolutionary rate by integrating gene co-expression with protein interactions'. In: *BMC Systems Biology* 4.1, p. 179. DOI: 10.1186/1752-0509-4-179.

Roy, S. et al. (2010). 'Identification of functional elements and regulatory circuits by Drosophila modENCODE'. In: *Science* 330.6012, pp. 1787–1797. DOI: 10.1126/science.1198374.

Sabin, L. R., S. L. Hanna and S. Cherry (2010). 'Innate antiviral immunity in Drosophila'. In: *Current Opinion in Immunology* 22.1. DOI: 10.1016/j.coi.2010.01.007.

Valanne, S. et al. (2010). 'Genome-Wide RNA Interference in Drosophila Cells Identifies G Protein-Coupled Receptor Kinase 2 as a Conserved Regulator of NF-κB Signaling'. In: *The Journal of Immunology* 184.11, pp. 6188–6198. DOI: 10.4049/jimmunol.1000261.

Van Der Linde, K., D. Houle, G. S. Spicer and S. J. Steppan (2010). 'A supermatrix-based molecular phylogeny of the family Drosophilidae'. In: *Genetics Research* 92.1, pp. 25–38. DOI: 10.1017/S001667231000008X.

Chandler, J. A., J. M. Lang, S. Bhatnagar, J. A. Eisen and A. Kopp (2011). 'Bacterial communities of diverse Drosophila species: ecological context of a host-microbe model system'. In: *PLoS Genet* 7.9, e1002272. DOI: 10.1371/journal.pgen.1002272.

Colbourne, J. K. et al. (2011). 'The ecoresponsive genome of Daphnia pulex'. In: *Science* 331.6017, pp. 555–561. DOI: 10.1126/science.1197761.

Hughes, G. L., R. Koga, P. Xue, T. Fukatsu and J. L. Rasgon (2011). 'Wolbachia infections are virulent and inhibit the human malaria parasite Plasmodium falciparum in Anopheles gambiae'. In: *PLoS Pathog* 7.5, e1002043. DOI: 10.1371/journal.ppat.1002043.

Murali, T. et al. (2011). 'DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila'. In: *Nucleic Acids Research* 39.suppl_1, pp. D736–D743. DOI: 10.1093/nar/gkq1092.

Négre, N. et al. (2011). 'A cis-regulatory map of the Drosophila genome'. In: *Nature* 471.7339, pp. 527–531. DOI: 10.1038/NATURE09990.

Paredes, J. C., D. P. Welchman, M. Poidevin and B. Lemaitre (2011). 'Negative regulation by amidase PGRPs shapes the Drosophila antibacterial response and protects the fly from innocuous infection'. In: *Immunity* 35.5, pp. 770–779. DOI: 10.1016/j.immuni.2011.09.018.

Wong, Z. S., L. M. Hedges, J. C. Brownlie and K. N. Johnson (2011). 'Wolbachia-mediated antibacterial protection and immune gene regulation in Drosophila'. In: *PLoS ONE* 6.9, e25430. DOI: 10.1371/journal.pone.0025430.

Wright, V. M., K. L. Vogt, E. Smythe and M. P. Zeidler (2011). 'Differential activities of the Drosophila JAK/STAT pathway ligands Upd, Upd2 and Upd3'. In: *Cellular Signalling* 23.5, pp. 920–927. DOI: 10.1016/J.CELLSIG.2011.01.020.

Xavier, M. J. and M. J. Williams (2011). 'The Rho-Family GTPase Rac1 Regulates Integrin Localization in Drosophila Immunosurveillance Cells'. In: *PLOS ONE* 6.5, e19504. DOI: 10.1371/JOURNAL.PONE.0019504.

Xie, J., B. Tiner, I. Vilchez and M. Mateos (2011). 'Effect of the Drosophila endosymbiont Spiroplasma on parasitoid wasp development and on the reproductive fitness of wasp-attacked fly survivors'. In: *Evol Ecol* 53.5, pp. 1065–1079. DOI: 10.1007/s10682-010-9453-7.

Yang, L. and B. S. Gaut (2011). 'Factors that Contribute to Variation in Evolutionary Rate among Arabidopsis Genes'. In: *Molecular Biology and Evolution* 28.8, pp. 2359–2369. DOI: 10.1093/molbev/msr058.

Yang, Z. and M. Dos Reis (2011). 'Statistical Properties of the Branch-Site Test of Positive Selection'. In: *Molecular Biology and Evolution* 28.3, pp. 1217–1228. DOI: 10.1093/molbev/msq303.

Behura, S. K. and D. W. Severson (2012). 'Comparative analysis of codon usage bias and codon context patterns between dipteran and hymenopteran sequenced genomes'. In: *PLoS ONE* 7.8, e43111. DOI: 10.1371/journal.pone.0043111.

Bosco-Drayon, V. et al. (2012). 'Peptidoglycan sensing by the receptor PGRP-LE in the Drosophila gut induces immune responses to infectious bacteria and tolerance to microbiota'. In: *Cell Host Microbe* 12.2, pp. 153–165. DOI: 10.1016/j.chom.2012.06.002.

Murrell, B. et al. (2012). 'Detecting Individual Sites Subject to Episodic Diversifying Selection'. In: *PLoS Genetics* 8.7, e1002764. DOI: 10.1371/journal.pgen.1002764.

Nam, H. J., I. H. Jang, H. You, K. A. Lee and W. J. Lee (2012). 'Genetic evidence of a redox-dependent systemic wound response via Hayan protease-phenoloxidase system in Drosophila'. In: *EMBO Journal* 31.5, pp. 1253–1265. DOI: 10.1038/EMBOJ.2011.476.

Neyen, C., M. Poidevin, A. Roussel and B. Lemaitre (2012). 'Tissue- and Ligand-Specific Sensing of Gram-Negative Infection in Drosophila by PGRP-LC Isoforms and PGRP-LE'. In: *The Journal of Immunology* 189.4, pp. 1886–1897. DOI: 10.4049/JIMMUNOL.1201022.

Singh, R. S., J. Xu, R. J. Kulathinal, B. P. Lazzaro and A. G. Clark (2012). *Rapid evolution of innate immune response genes*. Oxford : Oxford University Press, pp. 203–210. DOI: 10.1093/acprof:oso/9780199642274.003.0020.

Williford, A. and J. P. Demuth (2012). 'Gene expression levels are correlated with synonymous codon usage, amino acid composition, and gene architecture in the red flour beetle, Tribolium castaneum'. In: *Mol Biol Evol* 29.12, pp. 3755–3766. DOI: 10.1093/molbev/mss184.

Zhou, Q. and D. Bachtrog (2012). 'Sex-specific adaptation drives early sex chromosome evolution in Drosophila'. In: *Science* 337.6092, pp. 341–345. DOI: 10.1126/science.1225385.

Blum, J. E., C. N. Fischer, J. Miles and J. Handelsman (2013). 'Frequent replenishment sustains the beneficial microbiome of Drosophila melanogaster'. In: *mBio* 4.6, pp. 00860–13. DOI: 10.1128/mBio.00860-13.

Dobin, A. et al. (2013). 'STAR: ultrafast universal RNA-seq aligner'. In: *Bioinformatics* 29.1, pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

Harpur, B. A. and A. Zayed (2013). 'Accelerated Evolution of Innate Immunity Proteins in Social Insects: Adaptive Evolution or Relaxed Constraint?' In: *Molecular Biology and Evolution* 30.7, pp. 1665–1674. DOI: 10.1093/molbev/mst061.

Hickey, G., B. Paten, D. Earl, D. Zerbino and D. Haussler (2013). 'HAL: a hierarchical format for storing and analyzing multiple genome alignments'. In: *Bioinformatics* 29.10, pp. 1341–1342. DOI: 10.1093/bioinformatics/btt128.

Katoh, K. and D. M. Standley (2013). 'MAFFT multiple sequence alignment software version 7: improvements in performance and usability'. In: *Mol Biol Evol* 30.4, pp. 772–780. DOI: 10.1093/molbev/mst010.

McMurdie, P. J. and S. Holmes (2013). 'phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data'. In: *PLoS ONE* 8.4, e61217. DOI: 10.1371/journal.pone.0061217.

Nonaka, S., K. Nagaosa, T. Mori, A. Shiratsuchi and Y. Nakanishi (2013). 'Integrin αpS3/β-mediated Phagocytosis of Apoptotic Cells and Bacteria in Drosophila'. In: *Journal of Biological Chemistry* 288.15, pp. 10374–10380. DOI: 10.1074/jbc.M113.451427.

Schreiber, F. and E. L. L. Sonnhammer (2013). 'Hieranoid: hierarchical orthology inference'. In: *J Mol Biol* 425.11, pp. 2072–2081. DOI: 10.1016/j.jmb.2013.02.018.

Suetsugu, Y. et al. (2013). 'Large scale full-length cDNA sequencing reveals a unique genomic landscape in a lepidopteran model insect, Bombyx mori.' In: *G3 (Bethesda, Md.)* 3.9, pp. 1481–1492. DOI: 10.1534/G3.113.006239/-/DC1.

Tien, M. Z., A. G. Meyer, D. K. Sydykova, S. J. Spielman and C. O. Wilke (2013). 'Maximum Allowed Solvent Accessibilites of Residues in Proteins'. In: *PLoS One* 8.11, e80635. DOI: 10.1371/journal.pone.0080635.

Wu, P. Y., J. H. Phan and M. D. Wang (2013). 'Assessing the impact of human genome annotation choice on RNA-seq expression estimates'. In: *BMC Bioinformatics* 14.Suppl 11, S8. DOI: 10.1186/1471-2105-14-S11-S8.

Yassin, A. (2013). 'Phylogenetic classification of the Drosophilidae Rondani (Diptera): The role of morphology in the postgenomic era'. In: *Systematic Entomology* 38.2, pp. 349–364. DOI: 10.1111/J.1365-3113.2012.00665.X.

Buchon, N., N. Silverman and S. Cherry (2014). 'Immunity in Drosophila melanogaster–from microbial recognition to whole-organism physiology'. In: *Nat Rev Immunol* 14.12, pp. 796–810. DOI: 10.1038/nri3763.

Bushnell, B. (2014). *BBMap: A Fast, Accurate, Splice-Aware Aligner*.

Erler, S., P. Lhomme, P. Rasmont and H. M. G. Lattorff (2014). 'Rapid evolution of antimicrobial peptide genes in an insect host–social parasite system'. In: *Infection, Genetics and Evolution* 23, pp. 129–137. DOI: 10.1016/J.MEEGID.2014.02.002.

Honti, V., G. Csordás, É. Kurucz, R. Márkus and I. Andó (2014). 'The cell-mediated immunity of Drosophila melanogaster: Hemocyte lineages, immune compartments, microanatomy and regulation'. In: *Developmental & Comparative Immunology* 42.1, pp. 47–56. DOI: 10.1016/j.dci.2013.06.005.

Kleino, A. and N. Silverman (2014). 'The Drosophila IMD pathway in the activation of the humoral immune response'. In: *Developmental & Comparative Immunology* 42.1, pp. 25–35. DOI: 10.1016/J.DCI.2013.05.014.

Liao, Y., G. K. Smyth and W. Shi (2014). 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features'. In: *Bioinformatics* 30.7, pp. 923–930. DOI: 10.1093/bioinformatics/btt656.

Love, M. I., W. Huber and S. Anders (2014). 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2'. In: *Genome Biol* 15.12, p. 550. DOI: 10.1186/s13059-014-0550-8.

Myllymäki, H. and M. Rämet (2014). 'JAK/STAT Pathway in Drosophila Immunity'. In: *Scandinavian Journal of Immunology* 79.6, pp. 377–385. DOI: 10.1111/SJI.12170.

Papakostas, S. et al. (2014). 'Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions'. In: *Nature Communications* 5.1. DOI: 10.1038/ncomms5071.

Pieper, U. et al. (2014). 'ModBase, a database of annotated comparative protein structure models and associated resources'. In: *Nucleic Acids Research* 42.D1, pp. D336–D346. DOI: 10.1093/nar/gkt1144.

Salazar-Jaramillo, L. et al. (2014). 'Evolution of a cellular immune response in Drosophila: a phenotypic and genomic comparative analysis'. In: *Genome Biol Evol* 6.2, pp. 273–289. DOI: 10.1093/gbe/evu012.

Shilo, B. Z. (2014). 'The regulation and functions of MAPK pathways in Drosophila'. In: *Methods* 68.1, pp. 151–159. DOI: 10.1016/J.YMETH.2014.01.020.

Williams, M. and R. Baxter (2014). 'The structure and function of thioester-containing proteins in arthropods'. In: *Biophysical Reviews* 6.3-4, pp. 261–272. DOI: 10.1007/s12551-014-0142-6.

Flores, H. A., J. E. Bubnell, C. F. Aquadro and D. A. Barbash (2015). 'The Drosophila bag of marbles Gene Interacts Genetically with Wolbachia and Shows Female-Specific Effects of Divergence'. In: *PLoS Genetics* 11.8, e1005453. DOI: 10.1371/journal.pgen.1005453.

Kryuchkova-Mostacci, N. and M. Robinson-Rechavi (2015). 'Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse'. In: *PLoS One* 10.6, e0131673. DOI: 10.1371/journal.pone.0131673.

Magnacca, K. N. and D. K. Price (2015). 'Rapid adaptive radiation and host plant conservation in the Hawaiian picture wing Drosophila (Diptera: Drosophilidae)'. In: *Molecular Phylogenetics and Evolution* 92, pp. 226–242. DOI: 10.1016/J.YMPEV.2015.06.014.

Matthews, B. B. et al. (2015). 'Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data'. In: *G3 (Bethesda)* 5.8, pp. 1721–1736. DOI: 10.1534/g3.115.018929.

Murrell, B. et al. (2015). 'Gene-wide identification of episodic selection'. In: *Mol Biol Evol* 32.5, pp. 1365–1371. DOI: 10.1093/molbev/msv035.

Promponas, V. J., I. Iliopoulos and C. A. Ouzounis (2015). 'Annotation inconsistencies beyond sequence similarity-based function prediction - phylogeny and genome structure'. In: *Stand Genomic Sci* 10.1, p. 108. DOI: 10.1186/s40793-015-0101-2.

Simao, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva and E. M. Zdobnov (2015). 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs'. In: *Bioinformatics* 31.19, pp. 3210–3212. DOI: 10.1093/bioinformatics/btv351.

Sironi, M., R. Cagliani, D. Forni and M. Clerici (2015). 'Evolutionary insights into host–pathogen interactions from mammalian sequence data'. In: *Nature Reviews Genetics* 16.4, pp. 224–236. DOI: 10.1038/nrg3905.

Vicoso, B. and D. Bachtrog (2015). 'Numerous Transitions of Sex Chromosomes in Diptera'. In: *PLOS Biology* 13.4, e1002078. DOI: 10.1371/JOURNAL.PBIO.1002078.

Viljakainen, L. (2015). 'Evolutionary genetics of insect innate immunity'. In: *Brief Funct Genomics* 14.6, pp. 407–412. DOI: 10.1093/bfgp/elv002.

Waterhouse, R. M. (2015). 'A maturing understanding of the composition of the insect gene repertoire'. In: *Curr Opin Insect Sci* 7, pp. 15–23. DOI: 10.1016/j.cois.2015.01.004.

Zani, I. et al. (2015). 'Scavenger Receptor Structure and Function in Health and Disease'. In: *Cells* 4.2, pp. 178–201. DOI: 10.3390/cells4020178.

Zhang, J. and J.-R. Yang (2015). 'Determinants of the rate of protein sequence evolution'. In: *Nature Reviews Genetics* 16.7, pp. 409–420. DOI: 10.1038/nrg3950.

Alexa, A. and J. Rahnenfuhrer (2016). 'topGO: Enrichment analysis for Gene Ontology. R package version 2.28. 0'. In: *Cranio*.

Anderl, I. et al. (2016). 'Transdifferentiation and Proliferation in Two Distinct Hemocyte Lineages in Drosophila melanogaster Larvae after Wasp Infection'. In: *PLOS Pathogens* 12.7, e1005746. DOI: 10.1371/JOURNAL.PPAT.1005746.

Chakrabarti, S. et al. (2016). 'Remote Control of Intestinal Stem Cell Activity by Haemocytes in Drosophila'. In: *PLOS Genetics* 12.5, e1006089. DOI: 10.1371/JOURNAL.PGEN.1006089.

Croset, V., M. Schleyer, J. R. Arguello, B. Gerber and R. Benton (2016). 'A molecular and neuronal basis for amino acid sensing in the Drosophila larva'. In: *Sci Rep* 6.1, p. 34871. DOI: 10.1038/srep34871.

Enard, D., L. Cai, C. Gwennap and D. A. Petrov (2016). 'Viruses are a dominant driver of protein adaptation in mammals'. In: *eLife* 5. Ed. by G. McVean, e12469. DOI: 10.7554/eLife.12469.

Hamilton, P. T., F. Peng, M. J. Boulanger and S. J. Perlman (2016). 'A ribosome-inactivating protein in a Drosophila defensive symbiont'. In: *Proc Natl Acad Sci U S A* 113.2, pp. 350–355. DOI: 10.1073/pnas.1518648113.

Hanson, M. A., P. T. Hamilton and S. J. Perlman (2016). 'Immune genes and divergent antimicrobial peptides in flies of the subgenus Drosophila'. In: *BMC Evol Biol* 16.1, p. 228. DOI: 10.1186/s12862-016-0805-y.

Huerta-Cepas, J., F. Serra and P. Bork (2016). 'ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data'. In: *Mol Biol Evol* 33.6, pp. 1635–1638. DOI: 10.1093/molbev/msw046.

Iatsenko, I., S. Kondo, D. Mengin-Lecreulx and B. Lemaitre (2016). 'PGRP-SD, an Extracellular Pattern-Recognition Receptor, Enhances Peptidoglycan-Mediated Activation of the Drosophila Imd Pathway'. In: *Immunity* 45.5, pp. 1013–1023. DOI: 10.1016/j.immuni.2016.10.029.

Izumitani, H. F., Y. Kusaka, S. Koshikawa, M. J. Toda and T. Katoh (2016). 'Phylogeography of the Subgenus Drosophila (Diptera: Drosophilidae): Evolutionary History of Faunal Divergence between the Old and the New Worlds'. In: *PLOS ONE* 11.7, e0160051. DOI: 10.1371/JOURNAL.PONE.0160051.

Konig, S., L. W. Romoth, L. Gerischer and M. Stanke (2016). 'Simultaneous gene finding in multiple genomes'. In: *Bioinformatics* 32.22, pp. 3388–3395. DOI: 10.1093/bioinformatics/btw494.

Levine, M. T., H. M. Vander Wende, E. Hsieh, E. C. P. Baker and H. S. Malik (2016). 'Recurrent Gene Duplication Diversifies Genome Defense Repertoire in Drosophila'. In: *Molecular Biology and Evolution* 33.7, pp. 1641–1653. DOI: 10.1093/MOLBEV/MSW053.

Massey, J. H. and P. J. Wittkopp (2016). 'The genetic basis of pigmentation differences within and between Drosophila species'. In: *Current topics in developmental biology* 119, p. 27. DOI: 10.1016/BS.CTDB.2016.03.004.

O'Leary, N. A. et al. (2016). 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation'. In: *Nucleic Acids Res* 44.D1, pp. 733–45. DOI: 10.1093/nar/gkv1189.

Sanchez-Flores, A. et al. (2016). 'Genome Evolution in Three Species of Cactophilic Drosophila'. In: *G3 (Bethesda)* 6.10, pp. 3097–3105. DOI: 10.1534/g3.116.033779.

Schurch, N. J. et al. (2016). 'How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?' In: *RNA* 22.6, pp. 839–851. DOI: 10.1261/RNA.053959.115/-/DC1.

Sojo, V., C. Dessimoz, A. Pomiankowski and N. Lane (2016). 'Membrane Proteins Are Dramatically Less Conserved than Water-Soluble Proteins across the Tree of Life'. In: *Molecular Biology and Evolution* 33.11, pp. 2874–2884. DOI: 10.1093/MOLBEV/MSW164.

Stanley, C. E. and R. J. Kulathinal (2016). 'flyDIVaS: A Comparative Genomics Resource for Drosophila Divergence and Selection'. In: *G3 Genes|Genomes|Genetics* 6.8, pp. 2355–2363. DOI: 10.1534/g3.116.031138.

Unckless, R. L. and B. P. Lazzaro (2016). 'The potential for adaptive maintenance of diversity in insect antimicrobial peptides'. In: *Philos Trans R Soc Lond B Biol Sci* 371.1695, p. 20150291. DOI: 10.1098/rstb.2015.0291.

Unckless, R. L., V. M. Howick and B. P. Lazzaro (2016). 'Convergent Balancing Selection on an Antimicrobial Peptide in Drosophila'. In: *Current Biology* 26.2, pp. 257–262. DOI: 10.1016/j.cub.2015.11.063.

Wang, G., X. Li and Z. Wang (2016). 'APD3: the antimicrobial peptide database as a tool for research and education'. In: *Nucleic Acids Res* 44.D1, pp. 1087–93. DOI: 10.1093/nar/gkv1278.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Ballinger, M. J. and S. J. Perlman (2017). 'Generality of toxins in defensive symbiosis: Ribosome-inactivating proteins and defense against parasitic wasps in Drosophila'. In: *PLoS Pathog* 13.7, e1006431. DOI: 10.1371/journal.ppat.1006431.

Bergman, P., S. Seyedoleslami Esfahani and Y. Engström (2017). 'Drosophila as a Model for Human Diseases—Focus on Innate Immunity in Barrier Epithelia'. In: *Current Topics in Developmental Biology* 121, pp. 29–81. DOI: 10.1016/BS.CTDB.2016.07.002.

Dostalova, A., S. Rommelaere, M. Poidevin and B. Lemaitre (2017). 'Thioester-containing proteins regulate the Toll pathway and play a role in Drosophila defence against microbial pathogens and parasitoid wasps'. In: *BMC Biol* 15.1, p. 79. DOI: 10.1186/s12915-017-0408-0.

Early, A. M. et al. (2017). 'Survey of Global Genetic Diversity Within the Drosophila Immune System'. In: *Genetics* 205.1, pp. 353–366. DOI: 10.1534/genetics.116.195016.

Gupta, V. et al. (2017). 'The route of infection determines Wolbachia antibacterial protection in Drosophila'. In: *Proc Biol Sci* 284.1856, p. 20170809. DOI: 10.1098/rspb.2017.0809.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. Von Haeseler and L. S. Jermiin (2017). 'ModelFinder: fast model selection for accurate phylogenetic estimates'. In: *Nature Methods* 14.6, pp. 587–589. DOI: 10.1038/nmeth.4285.

Pai, A. A. et al. (2017). 'The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture'. In: *eLife* 6. DOI: 10.7554/ELIFE.32537.

Sackton, T. B., B. P. Lazzaro, A. G. Clark and P. Wittkopp (2017). 'Rapid Expansion of Immune-Related Gene Families in the House Fly, Musca domestica'. In: *Molecular Biology and Evolution* 34.4, p. 857. DOI: 10.1093/MOLBEV/MSW285.

Shokal, U. and I. Eleftherianos (2017). 'Evolution and Function of Thioester-Containing Proteins and the Complement System in the Innate Immune Response'. In: *Frontiers in Immunology* 8. DOI: 10.3389/fimmu.2017.00759.

Tassetto, M., M. Kunitomi and R. Andino (2017). 'Circulating Immune Cells Mediate a Systemic RNAi-Based Adaptive Antiviral Response in Drosophila'. In: *Cell* 169.2, pp. 314–325. DOI: 10.1016/j.cell.2017.03.033.

Ye, Z. et al. (2017). 'A New Reference Genome Assembly for the Microcrustacean Daphnia pulex'. In: *G3 (Bethesda)* 7.5, pp. 1405–1416. DOI: 10.1534/g3.116.038638.

Fiddes, I. T. et al. (2018). 'Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome annotation'. In: *Genome Res* 28.7, pp. 1029–1038. DOI: 10.1101/gr.233460.117.

Hagai, T. et al. (2018). 'Gene expression variability across cells and species shapes innate immunity'. In: *Nature* 563.7730, pp. 197–202. DOI: 10.1038/s41586-018-0657-2.

Haller, S. et al. (2018). ' Quorum-sensing regulator RhlR but not its autoinducer RhlI enables Pseudomonas to evade opsonization '. In: *EMBO reports* 19.5. DOI: 10.15252/EMBR.201744880.

Issa, N. et al. (2018). 'The Circulating Protease Persephone Is an Immune Sensor for Microbial Proteolytic Activities Upstream of the Drosophila Toll Pathway'. In: *Molecular Cell* 69.4, pp. 539–550. DOI: 10.1016/J.MOLCEL.2018.01.029.

König, S., L. Romoth and M. Stanke (2018). 'Comparative Genome Annotation'. In: *Methods in Molecular Biology*. Springer New York, pp. 189–212. DOI: 10.1007/978-1-4939-7463-4{_}6.

Liu, Y. et al. (2018). 'Inflammation-induced STING-dependent autophagy restricts Zika virus infection in the Drosophila brain'. In: *Cell host & microbe* 24.1, p. 57. DOI: 10.1016/J.CHOM.2018.05.022.

Mahajan, S., K. H. Wei, M. J. Nalley, L. Gibilisco and D. Bachtrog (2018). 'De novo assembly of a young Drosophila Y chromosome using single-molecule sequencing and chromatin conformation capture'. In: *PLoS Biol* 16.7, e2006348. DOI: 10.1371/journal.pbio.2006348.

Martin, M., A. Hiroyasu, R. M. Guzman, S. A. Roberts and A. G. Goodman (2018). 'Analysis of Drosophila STING Reveals an Evolutionarily Conserved Antimicrobial Function'. In: *Cell reports* 23.12, p. 3537. DOI: 10.1016/J.CELREP.2018.05.029.

Myers, A. L., C. M. Harris, K. M. Choe and C. A. Brennan (2018). 'Inflammatory production of reactive oxygen species by Drosophila hemocytes activates cellular immune defenses'. In: *Biochemical and Biophysical Research Communications* 505.3, pp. 726–732. DOI: 10.1016/J.BBRC.2018.09.126.

Palmer, W. H., J. D. Hadfield and D. J. Obbard (2018). 'RNA-Interference Pathways Display High Rates of Adaptive Protein Evolution in Multiple Invertebrates'. In: *Genetics* 208.4, pp. 1585–1599. DOI: 10.1534/genetics.117.300567.

Park, J. and J. R. Carlson (2018). 'Physiological responses of the Drosophila labellum to amino acids'. In: *J Neurogenet* 32.1, pp. 27–36. DOI: 10.1080/01677063.2017.1406934.

Puerma, E. et al. (2018). 'The High-Quality Genome Sequence of the Oceanic Island Endemic Species Drosophila guanche Reveals Signals of Adaptive Evolution in Genes Related to Flight and Genome Stability'. In: *Genome Biol Evol* 10.8, pp. 1956–1969. DOI: 10.1093/gbe/evy135.

Rambaut, A., A. J. Drummond, D. Xie, G. Baele and M. A. Suchard (2018). 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7'. In: *Syst Biol* 67.5, pp. 901–904. DOI: 10.1093/sysbio/syy032.

Ranwez, V., E. J. P. Douzery, C. Cambon, N. Chantret and F. Delsuc (2018). 'MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons'. In: *Mol Biol Evol* 35.10, pp. 2582–2584. DOI: 10.1093/molbev/msy159.

Suchard, M. A. et al. (2018). 'Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10'. In: *Virus Evol* 4.1, vey016. DOI: 10.1093/ve/vey016.

Świderská, Z. et al. (2018). 'Avian Toll-like receptor allelic diversity far exceeds human polymorphism: an insight from domestic chicken breeds'. In: *Scientific Reports* 8.1. DOI: 10.1038/s41598-018-36226-1.

Sydykova, D. K., B. R. Jack, S. J. Spielman and C. O. Wilke (2018). 'Measuring evolutionary rates of proteins in a structural context'. In: *F1000Research* 6, p. 1845. DOI: 10.12688/f1000research.12874.2.

Troha, K., J. H. Im, J. Revah, B. P. Lazzaro and N. Buchon (2018). 'Comparative transcriptomics reveals CrebA as a novel regulator of infection tolerance in D. melanogaster'. In: *PLoS Pathog* 14.2, e1006847. DOI: 10.1371/journal.ppat.1006847.

Velová, H., M. W. Gutowska-Ding, D. W. Burt and M. Vinkler (2018). 'Toll-Like Receptor Evolution in Birds: Gene Duplication, Pseudogenization, and Diversifying Selection'. In: *Molecular Biology and Evolution* 35.9, pp. 2170–2184. DOI: 10.1093/molbev/msy119.

Veltri, D., U. Kamath and A. Shehu (2018). 'Deep learning improves antimicrobial peptide recognition'. In: *Bioinformatics* 34.16, pp. 2740–2747. DOI: 10.1093/bioinformatics/bty179.

Waterhouse, A. et al. (2018). 'SWISS-MODEL: homology modelling of protein structures and complexes'. In: *Nucleic Acids Research* 46.W1, W296–W303. DOI: 10.1093/nar/gky427.

Yang, H. et al. (2018). 'Re-annotation of eight Drosophila genomes'. In: *Life Science Alliance* 1.6, e201800156. DOI: 10.26508/LSA.201800156.

Ahmed, O. M. et al. (2019). 'Evolution of Mechanisms that Control Mating in Drosophila Males'. In: *Cell Reports* 27.9, pp. 2527–2536. DOI: 10.1016/J.CELREP.2019.04.104.

Armstrong, J., I. T. Fiddes, M. Diekhans and B. Paten (2019). 'Whole-Genome Alignment and Comparative Annotation'. In: *Annual Review of Animal Biosciences* 7.Volume 7, 2019, pp. 41–64. DOI: 10.1146/ANNUREV-ANIMAL-020518-115005/CITE/REFWORKS.

Bachtrog, D., S. Mahajan and R. Bracewell (2019). 'Massive gene amplification on a recently formed Drosophila Y chromosome'. In: *Nat Ecol Evol* 3.11, pp. 1587–1597. DOI: 10.1038/s41559-019-1009-9.

Banerjee, U., J. R. Girard, L. M. Goins and C. M. Spratford (2019). 'Drosophila as a Genetic Model for Hematopoiesis'. In: *Genetics* 211.2, pp. 367–417. DOI: 10.1534/GENETICS.118.300223.

Bracewell, R., K. Chatla, M. J. Nalley and D. Bachtrog (2019). 'Dynamic turnover of centromeres drives karyotype evolution in drosophila'. In: *eLife* 8. DOI: 10.7554/ELIFE.49002.

Bushmanova, E., D. Antipov, A. Lapidus and A. D. Prjibelski (2019). 'rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data'. In: *GigaScience* 8.9. DOI: 10.1093/gigascience/giz100.

Chambers, M. C., E. Jacobson, S. Khalil and B. P. Lazzaro (2019). 'Consequences of chronic bacterial infection in Drosophila melanogaster'. In: *PLoS ONE* 14.10, e0224440. DOI: 10.1371/journal.pone.0224440.

Chapman, J. R., T. Hill, R. L. Unckless and M. Wayne (2019). 'Balancing Selection Drives the Maintenance of Genetic Variation in Drosophila Antimicrobial Peptides'. In: *Genome Biology and Evolution* 11.9, p. 2691. DOI: 10.1093/GBE/EVZ191.

Crysnanto, D. and D. J. Obbard (2019). 'Widespread gene duplication and adaptive evolution in the RNA interference pathways of the Drosophila obscura group'. In: *BMC Evolutionary Biology* 19.1. DOI: 10.1186/s12862-019-1425-0.

Di Franco, A., R. Poujol, D. Baurain and H. Philippe (2019). 'Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences'. In: *BMC Evol Biol* 19.1, p. 21. DOI: 10.1186/s12862-019-1350-2.

Dudzic, J. P., M. A. Hanson, I. Iatsenko, S. Kondo and B. Lemaitre (2019). 'More Than Black or White: Melanization and Toll Share Regulatory Serine Proteases in Drosophila'. In: *Cell Reports* 27.4, pp. 1050–1061. DOI: 10.1016/J.CELREP.2019.03.101.

Ellison, C. and D. Bachtrog (2019a). 'Contingency in the convergent evolution of a regulatory network: Dosage compensation in Drosophila'. In: *PLOS Biology* 17.2, e3000094. DOI: 10.1371/JOURNAL.PBIO.3000094.

Ellison, C. and D. Bachtrog (2019b). 'Recurrent gene co-amplification on Drosophila X and Y chromosomes'. In: *PLOS Genetics* 15.7, e1008251. DOI: 10.1371/JOURNAL.PGEN.1008251.

Emms, D. M. and S. Kelly (2019). 'OrthoFinder: phylogenetic orthology inference for comparative genomics'. In: *Genome Biol* 20.1, p. 238. DOI: 10.1186/s13059-019-1832-y.

Han, G. Z. (2019). 'Origin and evolution of the plant immune system'. In: *New Phytologist* 222.1, pp. 70–83. DOI: 10.1111/NPH.15596.

Hanson, M. A., B. Lemaitre and R. L. Unckless (2019). 'Dynamic Evolution of Antimicrobial Peptides Underscores Trade-Offs Between Immunity and Ecological Fitness'. In: *Front Immunol* 10, p. 2620. DOI: 10.3389/fimmu.2019.02620.

Hill, T., B. S. Koseva and R. L. Unckless (2019). 'The Genome of Drosophila innubila Reveals Lineage-Specific Patterns of Selection in Immune Genes'. In: *Molecular Biology and Evolution* 36.7, pp. 1405–1417. DOI: 10.1093/MOLBEV/MSZ059.

Khan, I. et al. (2019). 'The Vertebrate TLR Supergene Family Evolved Dynamically by Gene Gain/Loss and Positive Selection Revealing a Host–Pathogen Arms Race in Birds'. In: *Diversity* 11.8, p. 131. DOI: 10.3390/d11080131.

Kriventseva, E. V. et al. (2019). 'OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs'. In: *Nucleic Acids Res* 47.D1, pp. D807–D811. DOI: 10.1093/nar/gky1053.

Melcarne, C., B. Lemaitre and E. Kurant (2019). 'Phagocytosis in Drosophila: From molecules and cellular machinery to physiology'. In: *Insect Biochemistry and Molecular Biology* 109, pp. 1–12. DOI: 10.1016/J.IBMB.2019.04.002.

Moutinho, A. F., F. F. Trancoso and J. Y. Dutheil (2019). 'The Impact of Protein Architecture on Adaptive Evolution'. In: *Molecular Biology and Evolution* 36.9, pp. 2013–2028. DOI: 10.1093/molbev/msz134.

Nishide, Y. et al. (2019). 'Functional crosstalk across IMD and Toll pathways: insight into the evolution of incomplete immune cascades'. In: *Proceedings of the Royal Society B: Biological Sciences* 286.1897. DOI: 10.1098/rspb.2018.2207.

Renschler, G. et al. (2019). 'Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling'. In: *Genes Dev* 33.21-22, pp. 1591–1612. DOI: 10.1101/gad.328971.119.

Shultz, A. J. and T. B. Sackton (2019). 'Immune genes are hotspots of shared positive selection across birds and mammals'. In: *eLife* 8. DOI: 10.7554/ELIFE.41815.

Torresen, O. K. et al. (2019). 'Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases'. In: *Nucleic Acids Res* 47.21, pp. 10994–11006. DOI: 10.1093/nar/gkz841.

Armstrong, J. et al. (2020). 'Progressive Cactus is a multiple-genome aligner for the thousand-genome era'. In: *Nature* 587.7833, pp. 246–251. DOI: 10.1038/s41586-020-2871-y.

Bronski, M. J., C. C. Martinez, H. A. Weld and M. B. Eisen (2020). 'Whole Genome Sequences of 23 Species from the Drosophila montium Species Group (Diptera: Drosophilidae): A Resource for Testing Evolutionary Hypotheses'. In: *G3 (Bethesda)* 10.5, pp. 1443–1455. DOI: 10.1534/g3.119.400959.

Cai, H. et al. (2020). '2′3′-cGAMP triggers a STING- And NF-κB-dependent broad antiviral response in Drosophila'. In: *Science Signaling* 13.660, p. 4537. DOI: 10.1126/SCISIGNAL.ABC4537.

Cortazar-Chinarro, M. et al. (2020). 'Antimicrobial peptide and sequence variation along a latitudinal gradient in two anurans'. In: *BMC Genet* 21.1, p. 38. DOI: 10.1186/s12863-020-00839-1.

Ejigu, G. F. and J. Jung (2020). 'Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing'. In: *Biology (Basel)* 9.9, p. 295. DOI: 10.3390/biology9090295.

Gupta, A. and S. Nair (2020). 'Dynamics of Insect–Microbiome Interaction Influence Host and Microbial Symbiont'. In: *Frontiers in Microbiology* 11, p. 545024. DOI: 10.3389/FMICB.2020.01357/XML.

Hanson, M. A. and B. Lemaitre (2020). 'New insights on Drosophila antimicrobial peptide function in host defense and beyond'. In: *Curr Opin Immunol* 62, pp. 22–30. DOI: 10.1016/j.coi.2019.11.008.

Ito, J., R. J. Gifford and K. Sato (2020). 'Retroviruses drive the rapid evolution of mammalian APOBEC3 genes'. In: *Proceedings of the National Academy of Sciences.* 117.1, pp. 610–618. DOI: 10.1073/pnas.1914183116.

Lazzaro, B. P., M. Zasloff and J. Rolff (2020). 'Antimicrobial peptides: Application informed by evolution'. In: *Science* 368.6490, eaau5480. DOI: 10.1126/science.aau5480.

Mérel, V., M. Boulesteix, M. Fablet and C. Vieira (2020). 'Transposable elements in Drosophila'. In: *Mobile DNA 2020 11:1* 11.1, pp. 1–20. DOI: 10.1186/S13100-020-00213-Z.

Minh, B. Q. et al. (2020). 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era'. In: *Mol Biol Evol* 37.5, pp. 1530–1534. DOI: 10.1093/molbev/msaa015.

Moretta, A. et al. (2020). 'A bioinformatic study of antimicrobial peptides identified in the Black Soldier Fly (BSF) Hermetia illucens (Diptera: Stratiomyidae)'. In: *Sci Rep* 10.1, p. 16875. DOI: 10.1038/s41598-020-74017-9.

Ou, S. et al. (2020). 'Effect of sequence depth and length in long-read assembly of the maize inbred NC358'. In: *Nature Communications 2020 11:1* 11.1, pp. 1–10. DOI: 10.1038/s41467-020-16037-7.

Scalzitti, N., A. Jeannin-Girardon, P. Collet, O. Poch and J. D. Thompson (2020). 'A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms'. In: *BMC Genomics* 21.1, p. 293. DOI: 10.1186/s12864-020-6707-9.

Tafesh-Edwards, G. and I. Eleftherianos (2020). 'JNK signaling in Drosophila immunity and homeostasis'. In: *Immunology Letters* 226, pp. 7–11. DOI: 10.1016/j.imlet.2020.06.017.

Balog, J. Á. et al. (2021). 'Immunoprofiling of Drosophila Hemocytes by Single-cell Mass Cytometry'. In: *Genomics, Proteomics & Bioinformatics* 19.2, pp. 243–252. DOI: 10.1016/j.gpb.2020.06.022.

Buchfink, B., K. Reuter and H. G. Drost (2021). 'Sensitive protein alignments at tree-of-life scale using DIAMOND'. In: *Nat Methods* 18.4, pp. 366–368. DOI: 10.1038/s41592-021-01101-x.

Cantalapiedra, C. P., A. Hernandez-Plaza, I. Letunic, P. Bork and J. Huerta-Cepas (2021). 'eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale'. In: *Mol Biol Evol* 38.12, pp. 5825–5829. DOI: 10.1093/molbev/msab293.

Cheng, H., G. T. Concepcion, X. Feng, H. Zhang and H. Li (2021). 'Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm'. In: *Nat Methods* 18.2, pp. 170–175. DOI: 10.1038/s41592-020-01056-5.

Conner, W. R. et al. (2021). 'A phylogeny for the Drosophila montium species group: A model clade for comparative analyses'. In: *Molecular Phylogenetics and Evolution* 158, p. 107061. DOI: 10.1016/J.YMPEV.2020.107061.

Danecek, P. et al. (2021). 'Twelve years of SAMtools and BCFtools'. In: *GigaScience* 10.2. DOI: 10.1093/gigascience/giab008.

Davies, C. S. et al. (2021). 'Contemporary evolution of the innate immune receptor gene TLR3 in an isolated vertebrate population'. In: *Molecular Ecology* 30.11, pp. 2528–2542. DOI: 10.1111/mec.15914.

Finet, C. et al. (2021). 'DrosoPhyla: Resources for Drosophilid Phylogeny and Systematics'. In: *Genome Biol Evol* 13.8. DOI: 10.1093/gbe/evab179.

Holleufer, A. et al. (2021). 'Two cGAS-like receptors induce antiviral immunity in Drosophila'. In: *Nature* 597.7874, pp. 114–118. DOI: 10.1038/S41586-021-03800-Z.

Kim, B. Y. et al. (2021). 'Highly contiguous assemblies of 101 drosophilid genomes'. In: *eLife* 10. DOI: 10.7554/eLife.66405.

Kokate, P. P., S. M. Techtmann and T. Werner (2021). 'Codon usage bias and dinucleotide preference in 29 Drosophila species'. In: *G3 (Bethesda)* 11.8. DOI: 10.1093/g3journal/jkab191.

Lawrence, T. J. et al. (2021). 'amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool'. In: *Bioinformatics* 37.14, pp. 2058–2060. DOI: 10.1093/bioinformatics/btaa917.

Lažetić, V. and E. R. Troemel (2021). 'Conservation lost: host-pathogen battles drive diversification and expansion of gene families'. In: *The FEBS Journal* 288.18, pp. 5289–5299. DOI: 10.1111/febs.15627.

Mathe, C. and C. Dunand (2021). 'Automatic Prediction and Annotation: There Are Strong Biases for Multigenic Families'. In: *Front Genet* 12, p. 697477. DOI: 10.3389/fgene.2021.697477.

McMullen, J. G., E. Bueno, F. Blow and A. E. Douglas (2021). 'Genome-Inferred Correspondence between Phylogeny and Metabolic Traits in the Wild Drosophila Gut Microbiome'. In: *Genome Biol Evol* 13.8. DOI: 10.1093/gbe/evab127.

Mendes, F. K., D. Vanderpool, B. Fulton and M. W. Hahn (2021). 'CAFE 5 models variation in evolutionary rates among gene families'. In: *Bioinformatics* 36.22-23, pp. 5516–5518. DOI: 10.1093/bioinformatics/btaa1022.

Nozawa, M. et al. (2021). 'Shared evolutionary trajectories of three independent neo-sex chromosomes in Drosophila'. In: *Genome Research* 31.11, pp. 2069–2079. DOI: 10.1101/GR.275503.121/-/DC1.

Rhie, A. et al. (2021). 'Towards complete and error-free genome assemblies of all vertebrate species'. In: *Nature* 592.7856, pp. 737–746. DOI: 10.1038/s41586-021-03451-0.

Shen, W. and H. Ren (2021). 'TaxonKit: A practical and efficient NCBI taxonomy toolkit'. In: *J Genet Genomics* 48.9, pp. 844–850. DOI: 10.1016/j.jgg.2021.03.006.

Shumate, A. and S. L. Salzberg (2021). 'Liftoff: accurate mapping of gene annotations'. In: *Bioinformatics* 37.12, pp. 1639–1643. DOI: 10.1093/BIOINFORMATICS/BTAA1016.

Slavik, K. M. et al. (2021). 'cGAS-like receptors sense RNA and control 3′2′-cGAMP signalling in Drosophila'. In: *Nature* 597.7874, pp. 109–113. DOI: 10.1038/S41586-021-03743-5.

Storer, J., R. Hubley, J. Rosen, T. J. Wheeler and A. F. Smit (2021). 'The Dfam community resource of transposable element families, sequence models, and genome annotations'. In: *Mob DNA* 12.1, p. 2. DOI: 10.1186/s13100-020-00230-y.

Venkatraman, M., R. C. Fleischer and M. T. N. Tsuchiya (2021). 'Comparative Analysis of Annotation Pipelines Using the First Japanese White-Eye (Zosterops japonicus) Genome'. In: *Genome Biol Evol* 13.5. DOI: 10.1093/gbe/evab063.

Zhong, X., M. Lundberg and L. Råberg (2021). 'Divergence in Coding Sequence and Expression of Different Functional Categories of Immune Genes between Two Wild Rodent Species'. In: *Genome Biology and Evolution* 13.3. DOI: 10.1093/gbe/evab023.

Bédard, C., A. F. Cisneros, D. Jordan and C. R. Landry (2022). 'Correlation between protein abundance and sequence conservation: what do recent experiments say?' In: *Current Opinion in Genetics & Development* 77, p. 101984. DOI: 10.1016/J.GDE.2022.101984.

Bubnell, J. E., C. K. S. Ulbing, P. Fernandez Begne and C. F. Aquadro (2022). 'Functional Divergence of the bag-of-marbles Gene in the Drosophila melanogaster Species Group'. In: *Molecular Biology and Evolution* 39.7. DOI: 10.1093/molbev/msac137.

Cai, H., C. Meignin and J. L. Imler (2022). 'cGAS-like receptor-mediated immunity: the insect perspective'. In: *Current Opinion in Immunology* 74, pp. 183–189. DOI: 10.1016/J.COI.2022.01.005.

Chaurasia, S. and J. Y. Dutheil (2022). 'The Structural Determinants of Intra-Protein Compensatory Substitutions'. In: *Molecular Biology and Evolution* 39.4. DOI: 10.1093/molbev/msac063.

Darwin Tree of Life Project, C. (2022). 'Sequence locally, think globally: The Darwin Tree of Life Project'. In: *Proc Natl Acad Sci U S A* 119.4, e2115642118. DOI: 10.1073/pnas.2115642118.

Dziedziech, A. and U. Theopold (2022). 'Proto-pyroptosis: An Ancestral Origin for Mammalian Inflammatory Cell Death Mechanism in Drosophila melanogaster'. In: *Journal of Molecular Biology* 434.4, p. 167333. DOI: 10.1016/J.JMB.2021.167333.

Irion, U. and C. Nusslein-Volhard (2022). 'Developmental genetics with model organisms'. In: *Proceedings of the National Academy of Sciences of the United States of America* 119.30, e2122148119. DOI: 10.1073/PNAS.2122148119.

J, O. (2022). *vegan: Community Ecology Package. R package version 2.6–4*.

Ko, B. J. et al. (2022). 'Widespread false gene gains caused by duplication errors in genome assemblies'. In: *Genome Biol* 23.1, p. 205. DOI: 10.1186/s13059-022-02764-1.

Li, F. et al. (2022). 'Phylogenomic analyses of the genus Drosophila reveals genomic signals of climate adaptation'. In: *Mol Ecol Resour* 22.4, pp. 1559–1581. DOI: 10.1111/1755-0998.13561.

Rondón, J. J. et al. (2022). 'Evolution of the odorant-binding protein gene family in Drosophila'. In: *Frontiers in Ecology and Evolution* 10, p. 957247. DOI: 10.3389/FEVO.2022.957247/XML.

Saucereau, Y. et al. (2022). 'Structure and dynamics of Toll immunoreceptor activation in the mosquito Aedes aegypti'. In: *Nat Commun* 13.1, p. 5110. DOI: 10.1038/s41467-022-32690-6.

Soni, V. and A. Eyre-Walker (2022). 'Factors That Affect the Rates of Adaptive and Nonadaptive Evolution at the Gene Level in Humans and Chimpanzees'. In: *Genome Biology and Evolution* 14.2. DOI: 10.1093/gbe/evac028.

Suvorov, A. et al. (2022). 'Widespread introgression across a phylogeny of 155 Drosophila genomes'. In: *Curr Biol* 32.1, pp. 111–123. DOI: 10.1016/j.cub.2021.10.052.

Teufel, F. et al. (2022). 'SignalP 6.0 predicts all five types of signal peptides using protein language models'. In: *Nat Biotechnol* 40.7, pp. 1023–1025. DOI: 10.1038/s41587-021-01156-3.

Wei, K. H., D. Mai, K. Chatla and D. Bachtrog (2022). 'Dynamics and Impacts of Transposable Element Proliferation in the Drosophila nasuta Species Group Radiation'. In: *Mol Biol Evol* 39.5. DOI: 10.1093/molbev/msac080.

Weisman, C. M., A. W. Murray and S. R. Eddy (2022). 'Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes'. In: *Curr Biol* 32.12, pp. 2632–2639. DOI: 10.1016/j.cub.2022.04.085.

Yu, S., F. Luo, Y. Xu, Y. Zhang and L. H. Jin (2022). 'Drosophila Innate Immunity Involves Multiple Signaling Pathways and Coordinated Communication Between Different Tissues'. In: *Front Immunol* 13, p. 905370. DOI: 10.3389/fimmu.2022.905370.

Bruna, T. et al. (2023). 'Galba: genome annotation with miniprot and AUGUSTUS'. In: *BMC Bioinformatics* 24.1, p. 327. DOI: 10.1186/s12859-023-05449-z.

Cai, H. et al. (2023). 'The virus-induced cyclic dinucleotide 2′3′-c-di-GMP mediates STING-dependent antiviral immunity in Drosophila'. In: *Immunity* 56.9, pp. 1991–2005. DOI: 10.1016/J.IMMUNI.2023.08.006.

Chen, S. (2023). 'Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp'. In: *iMeta* 2.2, e107. DOI: 10.1002/imt2.107.

Church, S. H., C. Munro, C. W. Dunn and C. G. Extavour (2023). 'The evolution of ovary-biased gene expression in Hawaiian Drosophila'. In: *PLOS Genetics* 19.1, e1010607. DOI: 10.1371/JOURNAL.PGEN.1010607.

Dijk, E. L. van et al. (2023). 'Genomics in the long-read sequencing era'. In: *Trends Genet* 39.9, pp. 649–671. DOI: 10.1016/j.tig.2023.04.006.

Grimaldi, D. A. and C. Richenbacher (2023). 'Exceptional Species Diversity of Drosophilidae (Diptera) in a Neotropical Forest'. English. In: *American Museum Novitates* 2023.3997. DOI: 10.1206/3997.1.

Hanson, M. A., L. Grollmus and B. Lemaitre (2023). 'Ecology-relevant bacteria drive the evolution of host antimicrobial peptides in Drosophila'. In: *Science* 381.6655, eadg5725. DOI: 10.1126/science.adg5725.

Huang, J. et al. (2023). 'A Toll pathway effector protects Drosophila specifically from distinct toxins secreted by a fungus or a bacterium'. In: *Proceedings of the National Academy of Sciences* 120.12. DOI: 10.1073/pnas.2205140120.

Kuznetsov, D. et al. (2023). 'OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity'. In: *Nucleic Acids Res* 51.D1, pp. D445–D451. DOI: 10.1093/nar/gkac998.

Kwon, T., E. R. Hanschen and B. T. Hovde (2023). 'Addressing the pervasive scarcity of structural annotation in eukaryotic algae'. In: *Sci Rep* 13.1, p. 1687. DOI: 10.1038/s41598-023-27881-0.

Li, H. (2023). 'Protein-to-genome alignment with miniprot'. In: *Bioinformatics* 39.1. DOI: 10.1093/bioinformatics/btad014.

Obbard, D. J., p. Wellcome Sanger Institute Tree of Life, D. N. A. P. c. Wellcome Sanger Institute Scientific Operations, c. Tree of Life Core Informatics and C. Darwin Tree of Life (2023a). 'The genome sequence of a drosophilid fruit fly, Chymomyza fuscimana (Drosophilidae) (Zetterstedt, 1838)'. In: *Wellcome Open Res* 8, p. 477. DOI: 10.12688/wellcomeopenres.20122.1.

– (2023b). 'The genome sequence of a drosophilid fruit fly, Hirtodrosophila cameraria (Haliday, 1833)'. In: *Wellcome Open Res* 8.361, p. 361. DOI: 10.12688/wellcomeopenres.19850.1.

Ruperti, F. et al. (2023). 'Cross-phyla protein annotation by structural prediction and alignment'. In: *Genome Biology* 24.1, pp. 1–21. DOI: 10.1186/S13059-023-02942-9.

Scheben, A. et al. (2023). 'Long-Read Sequencing Reveals Rapid Evolution of Immunity- and Cancer-Related Genes in Bats'. In: *Genome Biology and Evolution* 15.9. DOI: 10.1093/gbe/evad148.

Sharaf, A. et al. (2023). 'Bridging the gap in African biodiversity genomics and bioinformatics'. In: *Nat Biotechnol* 41.9, pp. 1348–1354. DOI: 10.1038/s41587-023-01933-2.

Thiebaut, A. et al. (2023). 'DrosOMA: the Drosophila Orthologous Matrix browser'. In: *F1000Res* 12, p. 936. DOI: 10.12688/f1000research.135250.2.

Vinkler, M. et al. (2023). 'Understanding the evolution of immune genes in jawed vertebrates'. In: *Journal of Evolutionary Biology* 36.6, pp. 847–873. DOI: 10.1111/JEB.14181.

Vuruputoor, V. S. et al. (2023). 'Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes'. In: *Appl Plant Sci* 11.4, e11533. DOI: 10.1002/aps3.11533.

Abramson, J. et al. (2024). 'Accurate structure prediction of biomolecular interactions with AlphaFold 3'. In: *Nature* 630.8016, pp. 493–500. DOI: 10.1038/s41586-024-07487-w.

Attah, V. et al. (2024). 'Duplication and neofunctionalization of a horizontally transferred xyloglucanase as a facet of the Red Queen coevolutionary dynamic'. In: *Proceedings of the National Academy of Sciences of the United States of America* 121.24, e2218927121. DOI: 10.1073/PNAS.2218927121.

Baril, T., J. Galbraith and A. Hayward (2024). 'Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline'. In: *Mol Biol Evol* 41.4. DOI: 10.1093/molbev/msae068.

Barrat-Charlaix, P. and R. A. Neher (2024). 'Eco-evolutionary dynamics of adapting pathogens and host immunity'. In: *eLife* 13. DOI: 10.7554/ELIFE.97350.1.

Cinege, G. et al. (2024). 'Cellular Immunity of Drosophila willistoni Reveals Novel Complexity in Insect Anti-Parasitoid Defense'. In: *Cells* 13.7, p. 593. DOI: 10.3390/cells13070593.

Domazet-Lošo, M., T. Široki, K. Šimičević and T. Domazet-Lošo (2024). 'Macroevolutionary dynamics of gene family gain and loss along multicellular eukaryotic lineages'. In: *Nature Communications* 15.1. DOI: 10.1038/s41467-024-47017-w.

DuBose, J. G. and J. C. de Roode (2024). 'The link between gene duplication and divergent patterns of gene expression across a complex life cycle'. In: *Evolution letters.* 8.5, pp. 726–734. DOI: 10.1093/evlett/qrae028.

Espinosa, E., R. Bautista, R. Larrosa and O. Plata (2024). 'Advancements in long-read genome sequencing technologies and algorithms'. In: *Genomics* 116.3, p. 110842. DOI: 10.1016/j.ygeno.2024.110842.

Freedman, A. H. and T. B. Sackton (2024). 'Building better genome annotations across the tree of life'. DOI: 10.1101/2024.04.12.589245.

Gabriel, L. et al. (2024). 'BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA'. In: *Genome Res* 34.5, pp. 769–777. DOI: 10.1101/gr.278090.123.

Gao, B. and S. Zhu (2024). 'The evolutionary novelty of insect defensins: from bacterial killing to toxin neutralization'. In: *Cell Mol Life Sci* 81.1, p. 230. DOI: 10.1007/s00018-024-05273-5.

Hrdina, A., M. Serra Canales, A. Arias-Rojas, D. Frahm and I. Iatsenko (2024). 'The endosymbiont Spiroplasma poulsonii increases Drosophila melanogaster resistance to pathogens by enhancing iron sequestration and melanization'. In: *mBio* 15.8, e0093624. DOI: 10.1128/mbio.00936-24.

Hu, Y., R. Ye, J. Su, Y. Rui and X. F. Yu (2024). 'cGAS–STING-mediated novel nonclassic antiviral activities'. In: *Journal of Medical Virology* 96.2, e29403. DOI: 10.1002/JMV.29403.

Kim, B. Y. et al. (2024). 'Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life'. In: *PLoS Biol* 22.7, e3002697. DOI: 10.1371/journal.pbio.3002697.

Mekic, R. et al. (2024). 'Number of human protein interactions correlates with structural, but not regulatory conservation of the respective genes'. In: *Frontiers in Genetics* 15. DOI: 10.3389/fgene.2024.1472638.

Nachtweide, S., L. Romoth and M. Stanke (2024). 'Comparative Genome Annotation'. In: *Methods in Molecular Biology*. Springer US, pp. 165–187. DOI: 10.1007/978-1-0716-3838-5{_}7.

Silva, R. and F. M. Gomes (2024). 'Evolution of the Major Components of Innate Immunity in Animals'. In: *J Mol Evol* 92.1, pp. 3–20. DOI: 10.1007/s00239-024-10155-2.

Subasi, B. S., V. Grabe, M. Kaltenpoth, J. Rolff and S. A. Armitage (2024). 'How frequently are insects wounded in the wild? A case study using Drosophila melanogaster'. In: *Royal Society Open Science* 11.6. DOI: 10.1098/RSOS.240256.

Westlake, H., M. A. Hanson and B. Lemaitre (2024). *The Drosophila immunity handbook*. EPFL Press, p. 248.

Yuan, D. et al. (2024). 'The European Nucleotide Archive in 2023'. In: *Nucleic Acids Res* 52.D1, pp. D92–D97. DOI: 10.1093/nar/gkad1067.

Zhang, H. (2024). *cubar: Codon Usage Bias Analysis. R package*.

Zhao, Y. et al. (2024). 'Integrating Iso-seq and RNA-seq data for the reannotation of the greater amberjack genome'. In: *Scientific Data* 11.1, pp. 1–10. DOI: 10.1038/S41597-024-03495-7.

Degen, P. M. and M. Medo (2025). 'Replicability of bulk RNA-Seq differential expression and enrichment analysis results for small cohort sizes'. In: *PLOS Computational Biology* 21.5, e1011630. DOI: 10.1371/JOURNAL.PCBI.1011630.

Detcharoen, M., P. Pramual and A. Nilsai (2025). 'Phylogenomic Analysis Reveals Evolutionary Relationships of Tropical Drosophilidae: From Drosophila to Scaptodrosophila'. In: *Ecology and Evolution* 15.3, e71100. DOI: 10.1002/ECE3.71100.

Dhakad, P., B. Kim, D. Petrov and D. J. Obbard (2025a). 'Comparative gene annotation of 304 species of Drosophilidae'. In: *bioRxiv*, p. 2025.04.14.648771. DOI: 10.1101/2025.04.14.648771.

Dhakad, P., D. Newman and D. J. Obbard (2025b). 'Transcriptomic analysis of non-model Drosophilidae reveals novel AMP candidates'. In: *bioRxiv*, p. 2025.06.06.658223. DOI: 10.1101/2025.06.06.658223.

Hanson, M. A. and L. Hedelin (2025). *Humoral immunity in insects: Antimicrobial peptides and other host defense peptides*. [Place of publication not identified] : Elsevier. DOI: 10.1016/B978-0-323-95424-2.00002-9.

Hilu-Dadia, R. et al. (2025). 'Santa-maria is a glial phagocytic receptor that acts with SIMU to recognize and engulf apoptotic neurons'. In: *Cell Reports* 44.1, p. 115201. DOI: 10.1016/J.CELREP.2024.115201.

Kolde, R. (2025). *pheatmap: Pretty Heatmaps. R package version 1.0.13*.

Manousi, D., S. Naseer, S. A. M. Martin, S. R. Sandve and M. Saitou (2025). 'Gene gain and loss drive the diversification of gig immune genes in teleosts: structural and regulatory insights from Atlantic salmon'. DOI: 10.1101/2025.07.01.662619.

Mariene, G. M. and J. D. Wasmuth (2025). 'Genome assembly variation and its implications for gene discovery in nematodes'. In: *International Journal for Parasitology* 55.5, pp. 239–252. DOI: 10.1016/J.IJPARA.2025.01.004.

Mullinax, S. R. et al. (2025). 'A suite of selective pressures supports the maintenance of alleles of a Drosophila immune peptide'. In: *eLife* 12. DOI: 10.7554/eLife.90638.

Nevers, Y. et al. (2025). 'Quality assessment of gene repertoire annotations with OMArk'. In: *Nat Biotechnol* 43.1, pp. 124–133. DOI: 10.1038/s41587-024-02147-w.

Perlmutter, J. I., A. Atadurdyyeva, M. E. Schedl and R. L. Unckless (2025). 'Wolbachia enhances the survival of Drosophila infected with fungal pathogens'. In: *BMC Biol* 23.1, p. 42. DOI: 10.1186/s12915-025-02130-0.

Prieto-Banos, S. et al. (2025). 'Annotation matters: the effect of structural gene annotation on orthology inference'. In: *Bioinformatics* 41.7. Ed. by P. Robinson. DOI: 10.1093/BIOINFORMATICS/BTAF365.

Tian, Y., X. Yue, R. Jiao, M. A. Hanson and B. Lemaitre (2025). 'Functional Characterization of Paillotin: An Immune Peptide Regulated by the Imd Pathway with Pathogen-Specific Roles inDrosophilaImmunity'. DOI: 10.1101/2025.05.12.653313.

Bloomington Drosophila Stock Center. Indiana University Bloomington. (N.d.).

Smit Hubley R & Green P, A. F. A. (n.d.). *RepeatMasker Open-4.0*.

Appendix A

Supplementary materials for Chapter 2

Supplementary tables/files can be found here: 10.5281/zenodo.16904124

Supplementary file A.1: SRA accession numbers for RNAseq datasets used in genome annotation.

Supplementary file A.2: *Drosophila melanogaster* genes ranked by overall expression level (FPKM).

Supplementary file A.3: Expression category assignments for HOGs based on *D. melanogaster* gene expression ranks.

Supplementary file A.4: Summary statistics for genome annotations and orthology assignments.

Supplementary file A.5: BUSCO and OMArk completeness assessments for genome annotations.

Supplementary file A.6: Estimates of codon usage bias metrics (GC3, GC content in non-coding regions and selection strength S) across species.

Supplementary file A.7: Time calibrated species tree inferred from HOG gene sets.

Supplementary file A.8: Species tree generated using HOG gene sets.

Supplementary file A.9: Species tree generated using BUSCO gene sets.

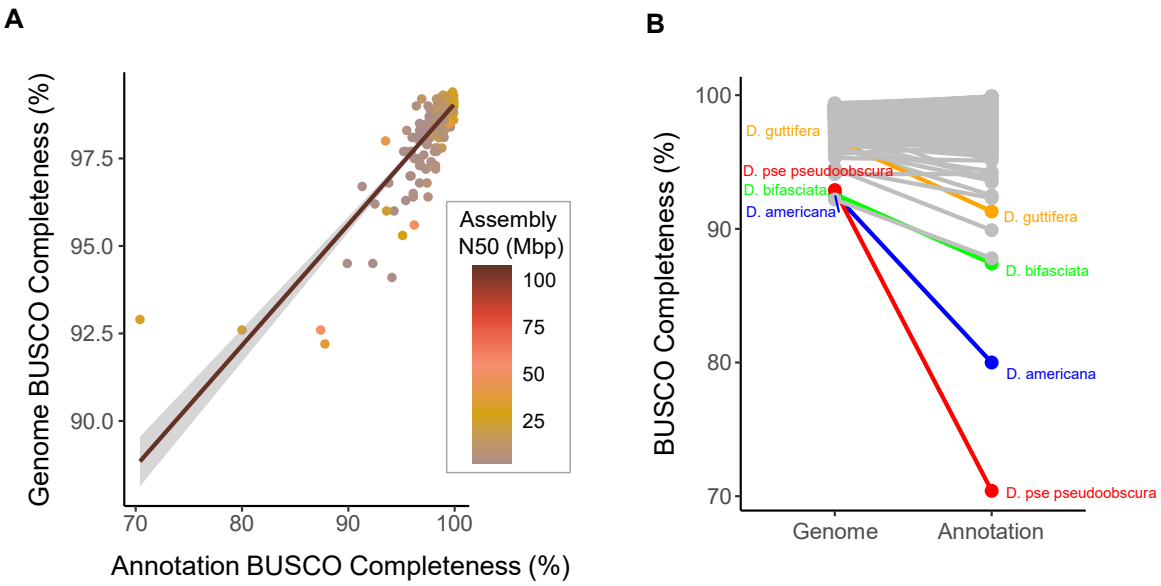

Figure A.1: Assessment of BUSCO completeness scores of genome assemblies and annotated protein sets.
(A) As expected, annotation BUSCO score increased with genome BUSCO scores. (B) Species with low genome BUSCO scores produces spurious gene sets.

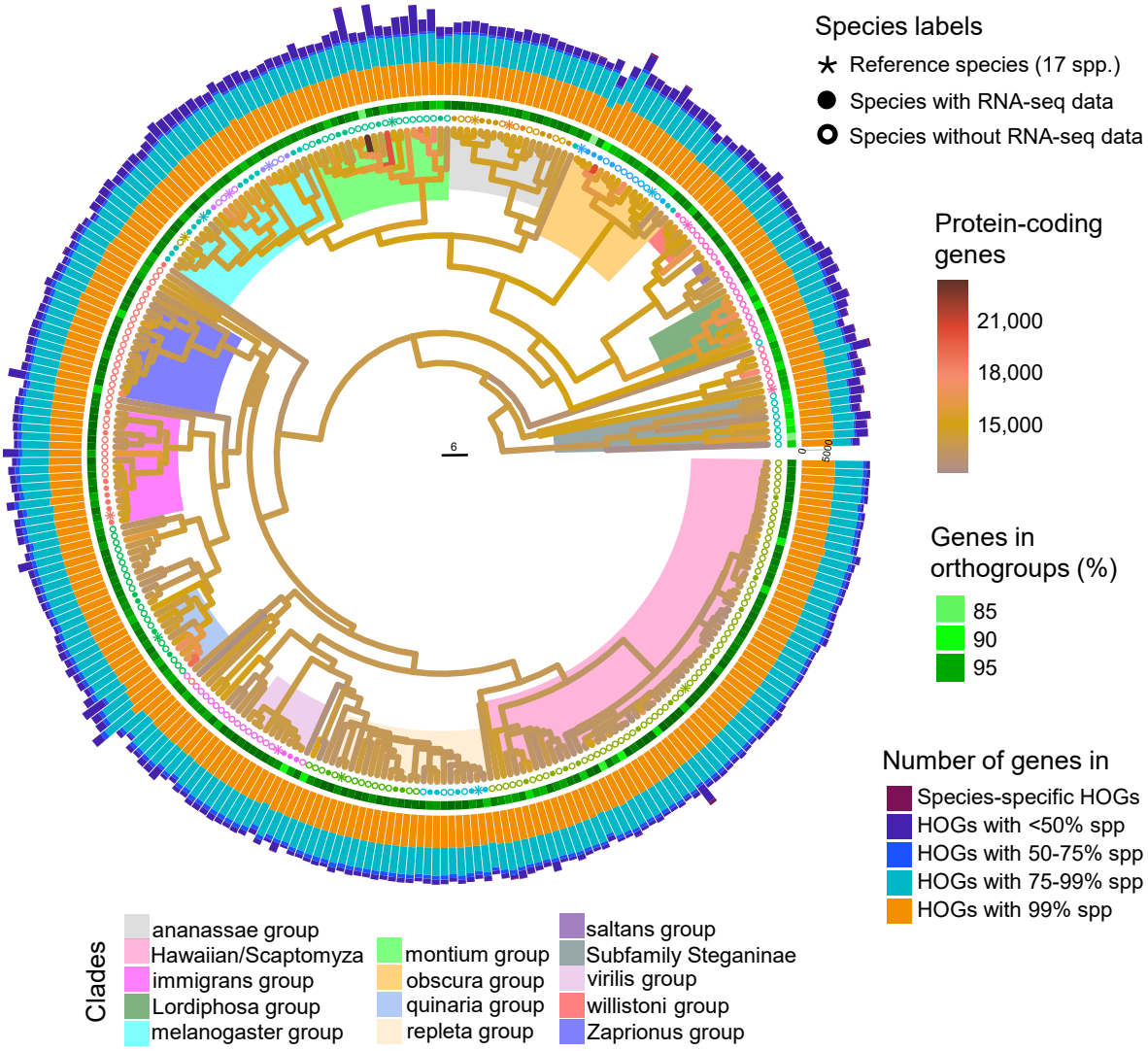

Figure A.2: Overview of orthology assignments across 304 Drosophilidae species.
Phylogenetic tree with ancestral reconstruction of the number of protein-coding genes mapped onto branches. Tip labels are coloured according to the reference species for the clade, stars indicate the reference species, and filled versus open circles indicate the availability of RNAseq data for that species. Inner tiles (green) indicates the percentage of genes in orthogroups for each species. Outermost layer (stacked barplot) indicates the number of genes HOG categories (HOGs that contains proportion of species or species-specific HOGs).

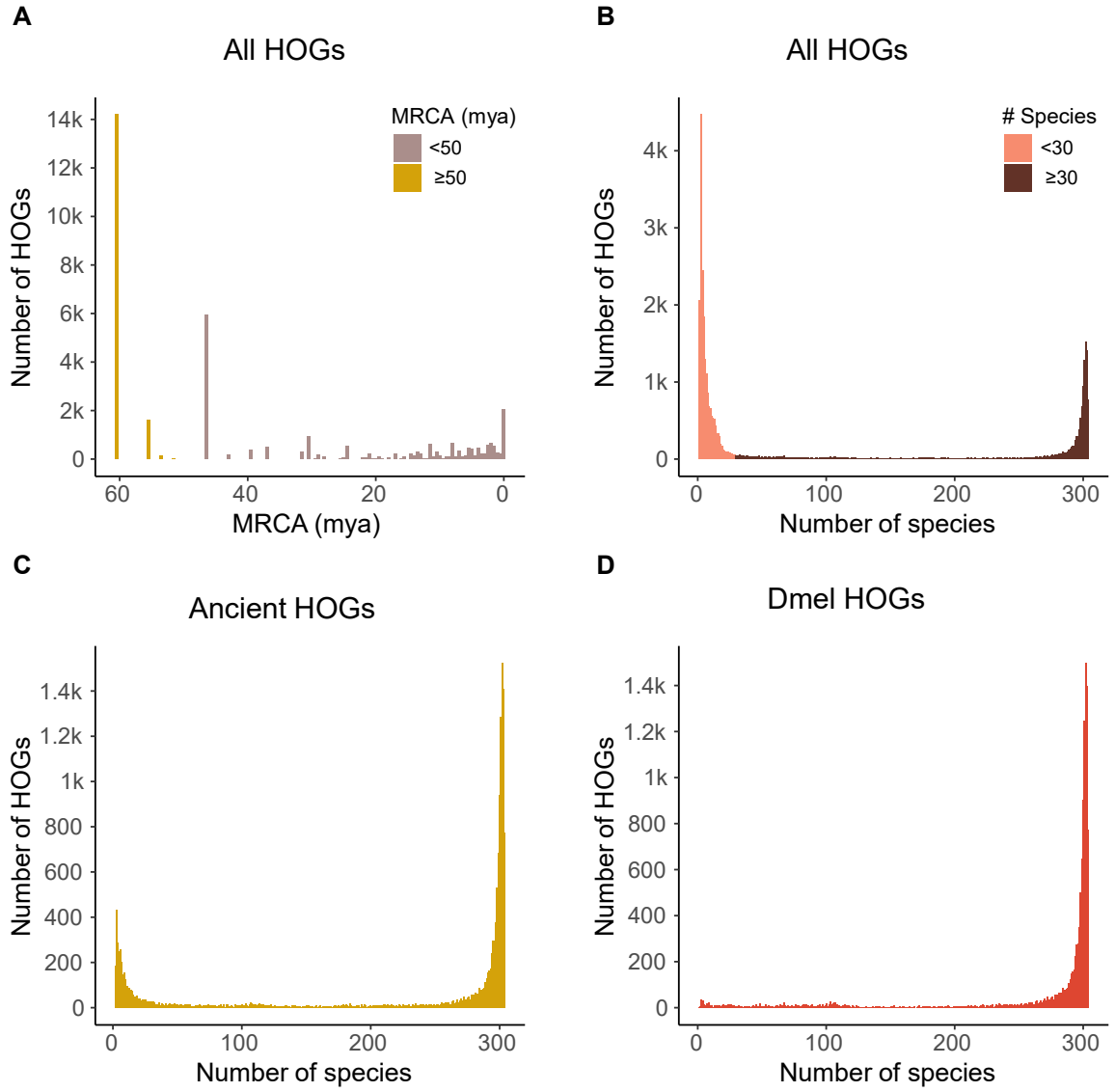

Figure A.3: Classification of Hierarchical Orthologous Groups (HOGs).
(A) Distribution of most recent common ancestor (MRCA, million years) of species in each HOGs. (B) Distrubution of number of species in each HOG. (C) Distribution of number of species in ancient HOGs (MRCA $\geq$50). (D) Distribution of number of species in HOGs that contains at least one *D. melanogaster* gene.

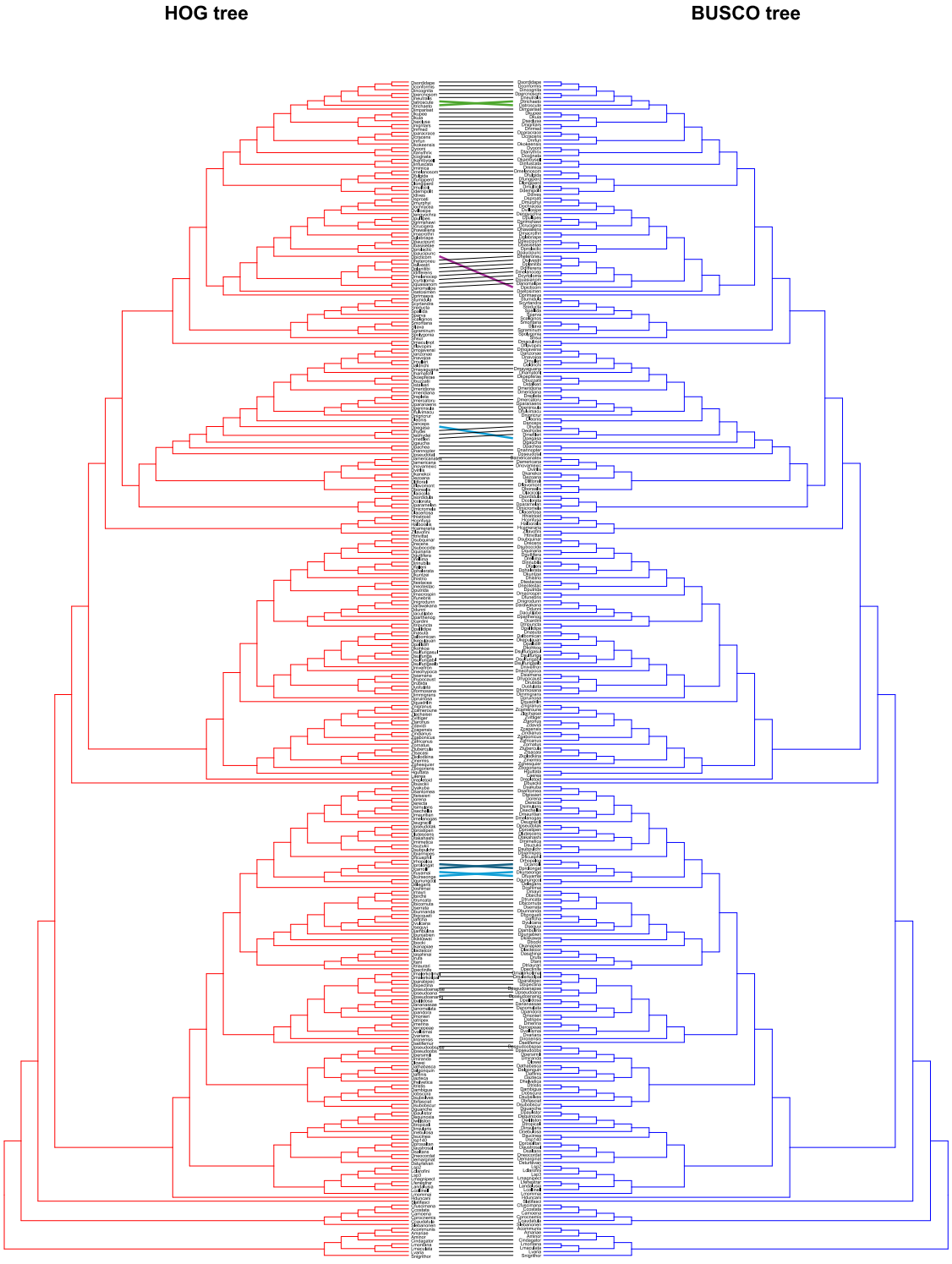

Figure A.4: Tanglegram of species trees constructed using HOGs and BUSCO genes.

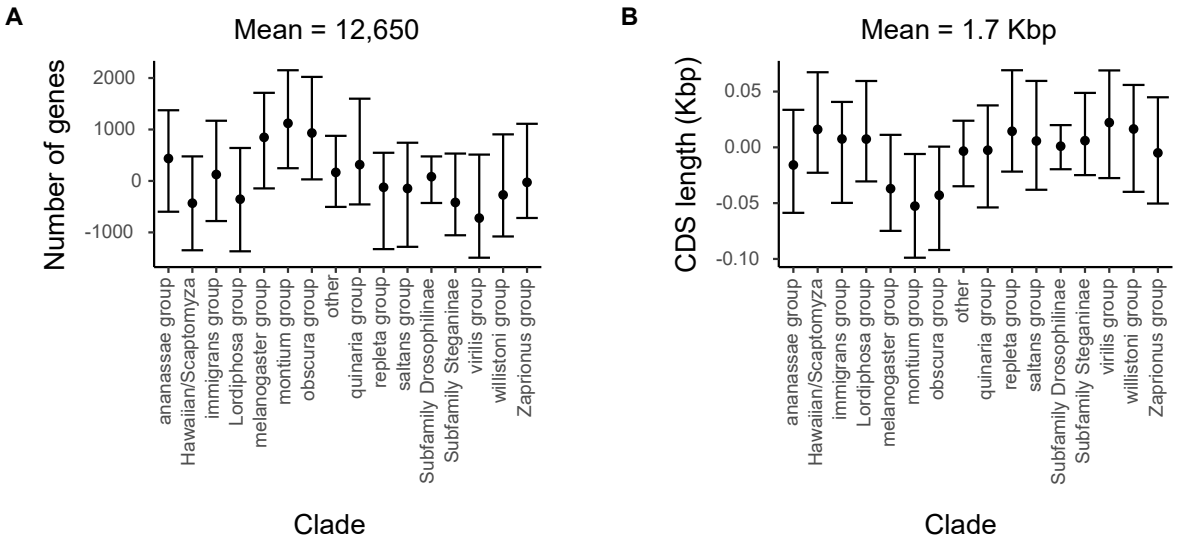

Figure A.5: Posterior estimates of gene number and mean CDS length reconstructed at internal nodes of the species phylogeny.

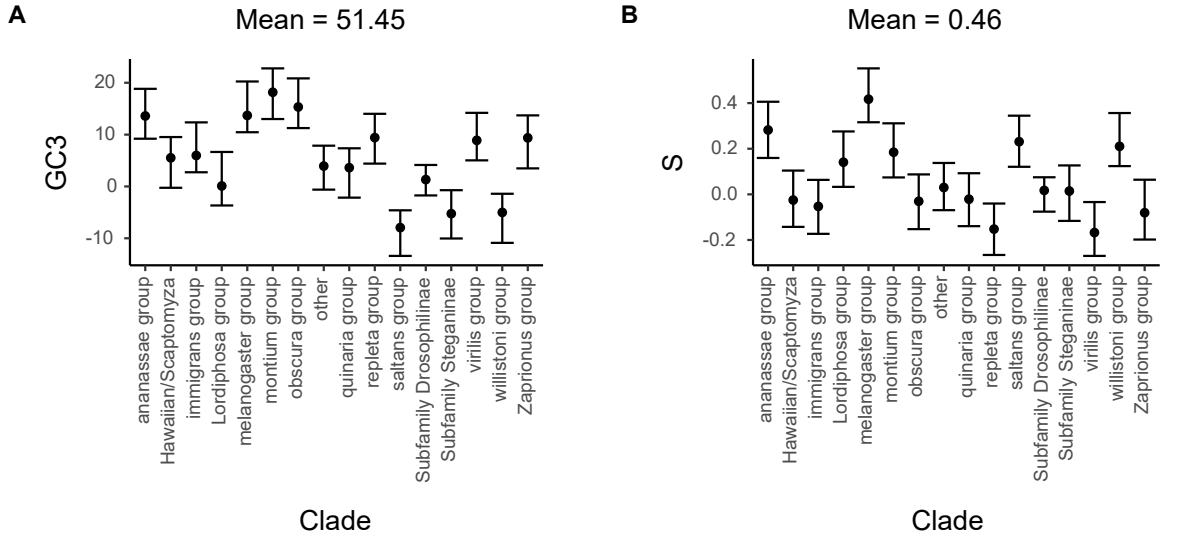

Figure A.6: Posterior estimates of GC3 content and strength of selection (S) on codon usage bias at internal nodes of the species phylogeny.

Appendix B

Supplementary materials for Chapter 3

Supplementary tables/files can be found here: 10.5281/zenodo.16904124

Supplementary file B.1: Model M1 and M4 result summaries.

Supplementary file B.2: Model M2 and M5 result summaries.

Supplementary file B.3: Model M3 and M6 result summaries.

Supplementary file B.4: List of genes ranking in top2.5% for dN/dS ratio, proportion of sites under diversifying selection and gene turnover.

Supplementary file B.5: Model syntax and priors.

Supplementary file B.6: Model M1-M7 full summaries.

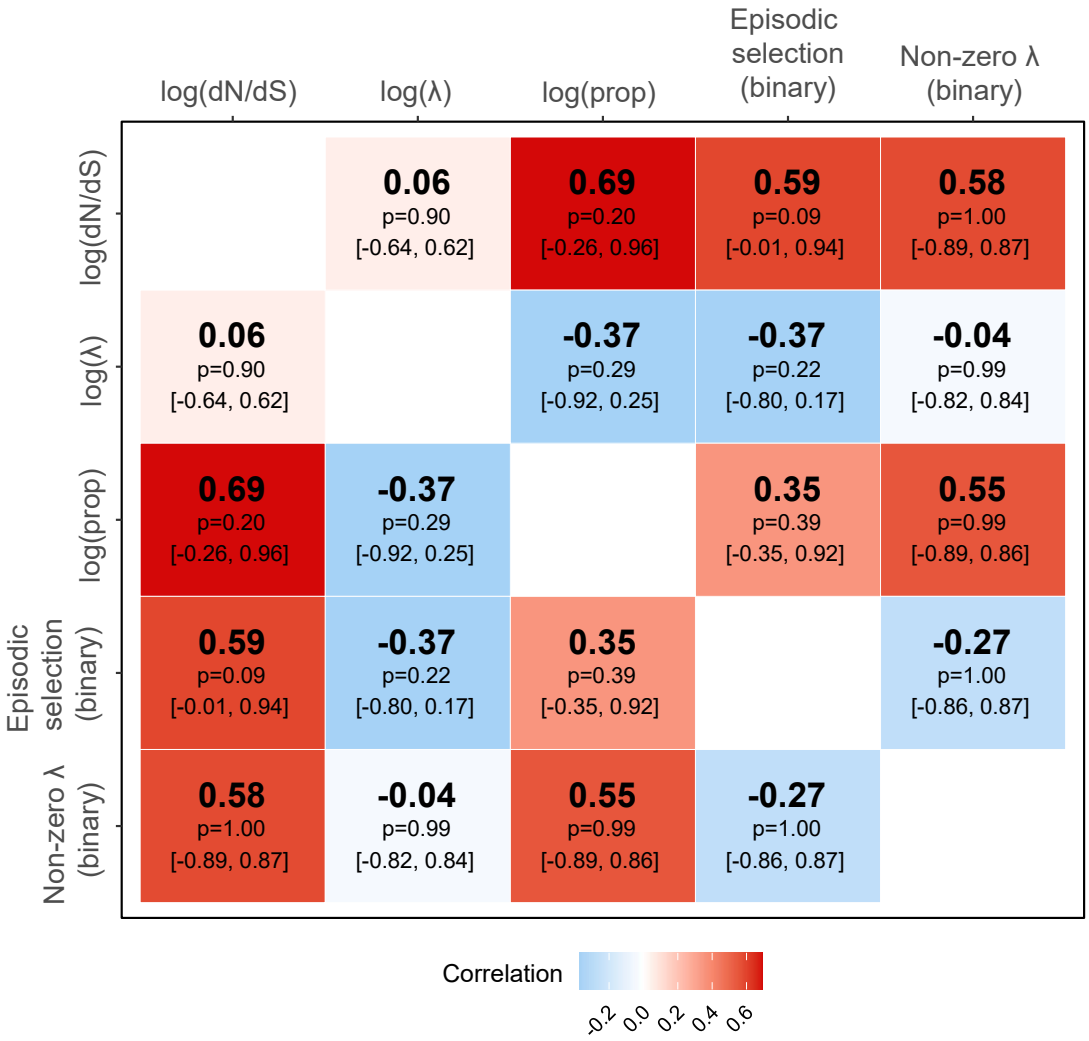

Figure B.1: Correlation matrix of estimates of sequence evolution and gene turnover (λ). Pairwise correlations between dN/dS ratio, proportion of sites under diversifying selection, and λ. Values represent correlation coefficients with 95% confidence intervals and MCMC p-values.

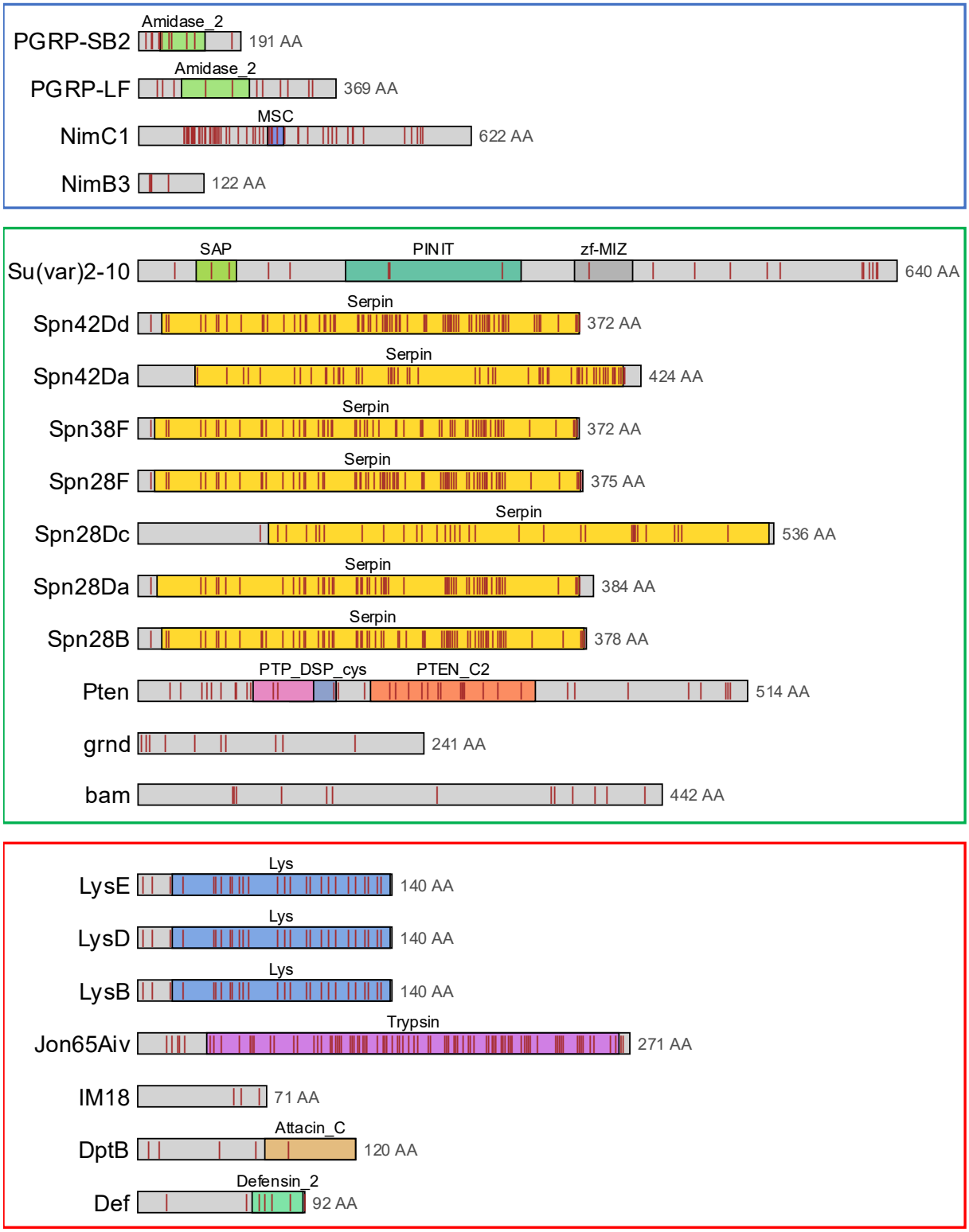

Figure B.2: Distribution of sites under diversifying selection in genes encoding receptors, signalling and effectors proteins.
Schematic gene structures of receptors (blue box), signalling (green box), and effectors (red box), showing conserved functional domain structures (not to scale) and approximate positions of positively selected sites (red bars). Positively selected sites were identified with MEME. Only genes identified in top 2.5% (excluding Tep and Sr-C receptors) of any evolutionary metrics represented here.

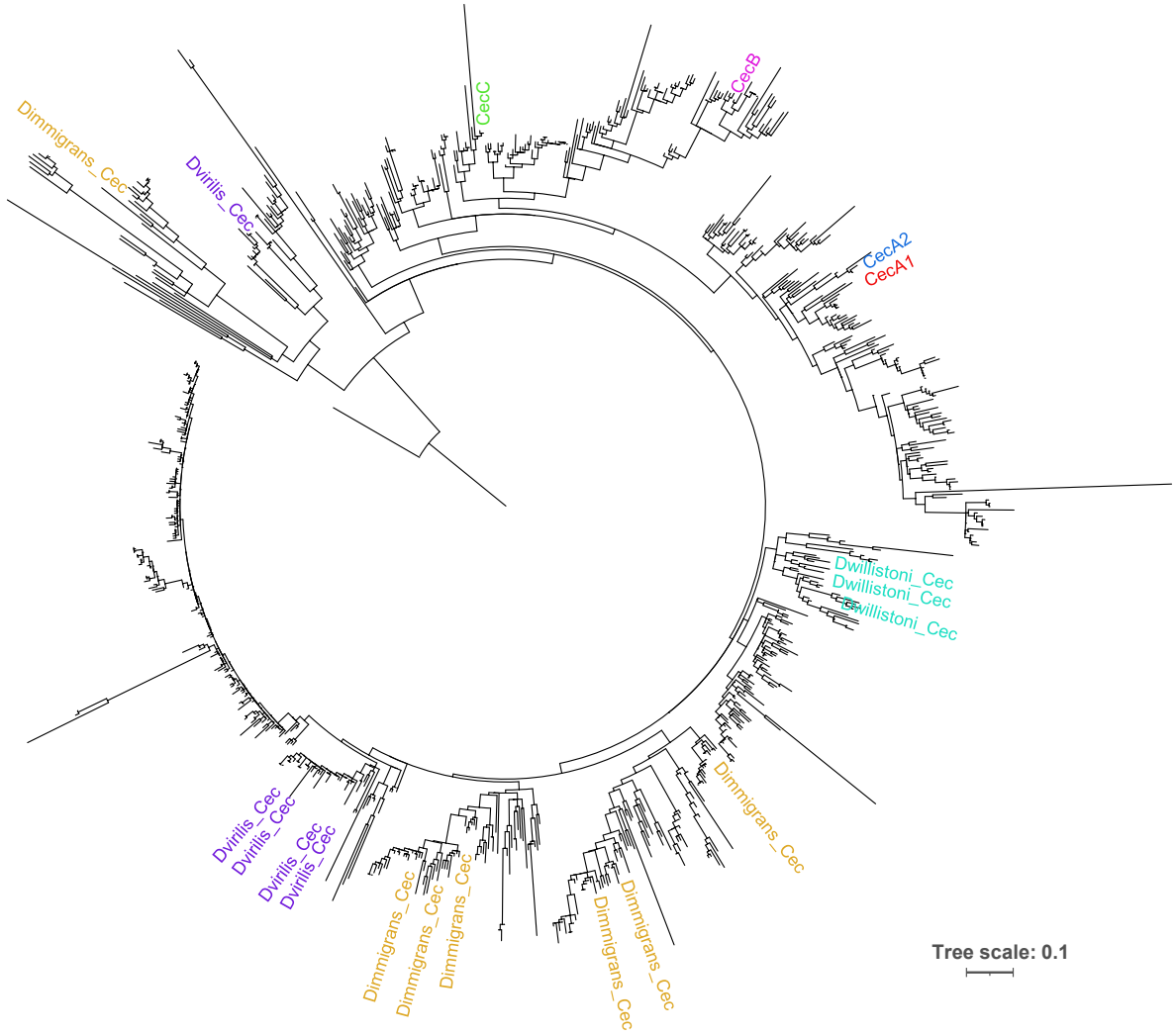

Figure B.3: Phylogeny of *Cecropin* gene family.
Maximum-likelihood gene tree of *Cecropin* genes from 304 Drosophilidae species. Cecropin paralogs (*CecA1*, *CecA2*, *CecB* and *CecC*) of *D. melanogaster* originated after the split of willistoni group within subgenus *Sophophora*. There were multiple independent duplications of *Cecropins* in other species groups. Tree rooted with *Scaptodrosophila latifasciaeformis* cecropin.

Appendix C

Supplementary materials for Chapter 4

Supplementary tables/files can be found here: 10.5281/zenodo.16904124

Supplementary file C.1: List of immune genes recovered (annotated) from *Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa* genomes.

Supplementary file C.2: List of differentially expressed genes between pathogen-challenged and unchallenged samples from *Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa*.

Supplementary file C.3: List of novel AMP candidates identified in this study.

Supplementary file C.4: Samples metadata and summary of reads mapped to the respective genomes.

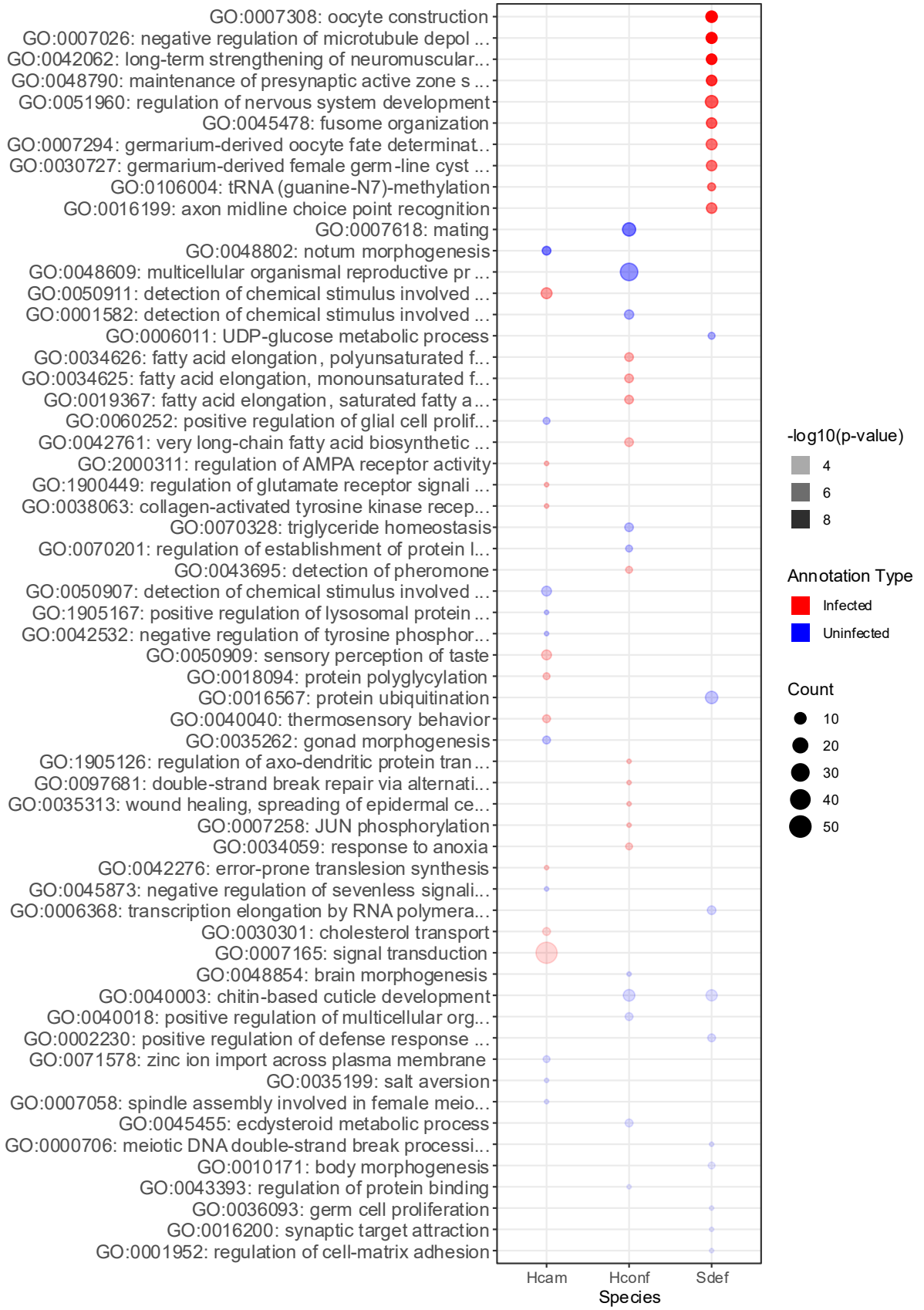

Figure C.1: GO term enrichment in unique genes recovered from *Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa* annotations.

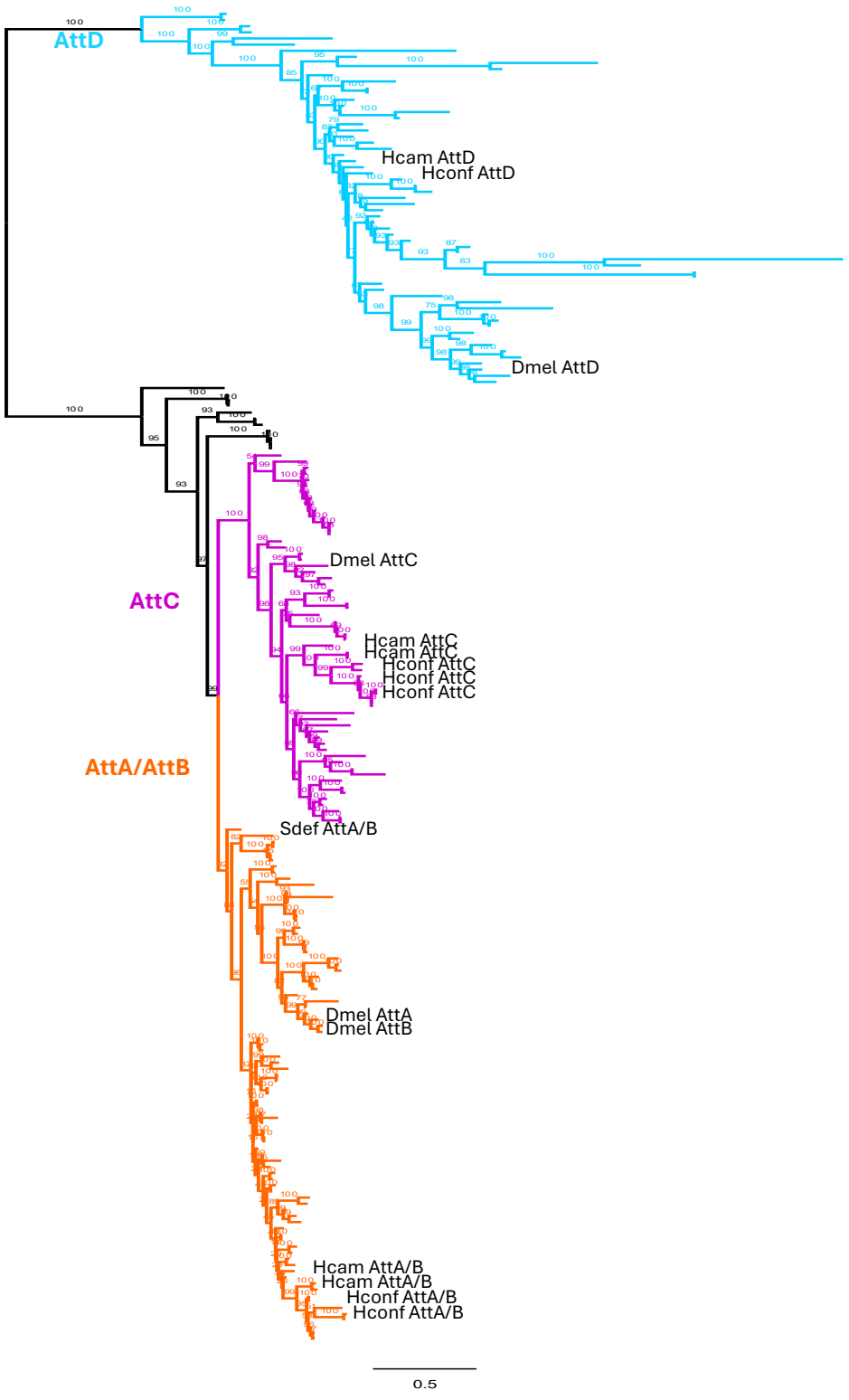

Figure C.2: Phylogeny of *Attacin* gene family.
Maximum-likelihood gene tree of *Attacin genes* from 51 drosophilid species, including *Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa*, generated using IQ-TREE2 based on aligned amino acid sequences. The gene tree highlights three distinct clades corresponding to *AttA* and *AttB* (orange), *AttC* (purple), and *AttD* (cyan). There were multiple copies of *AttC* in two *Hirtodrosophila* species. *AttC* and *AttD* were missing from *S. deflexa*, while it contains one copy of *AttA/B* like gene.

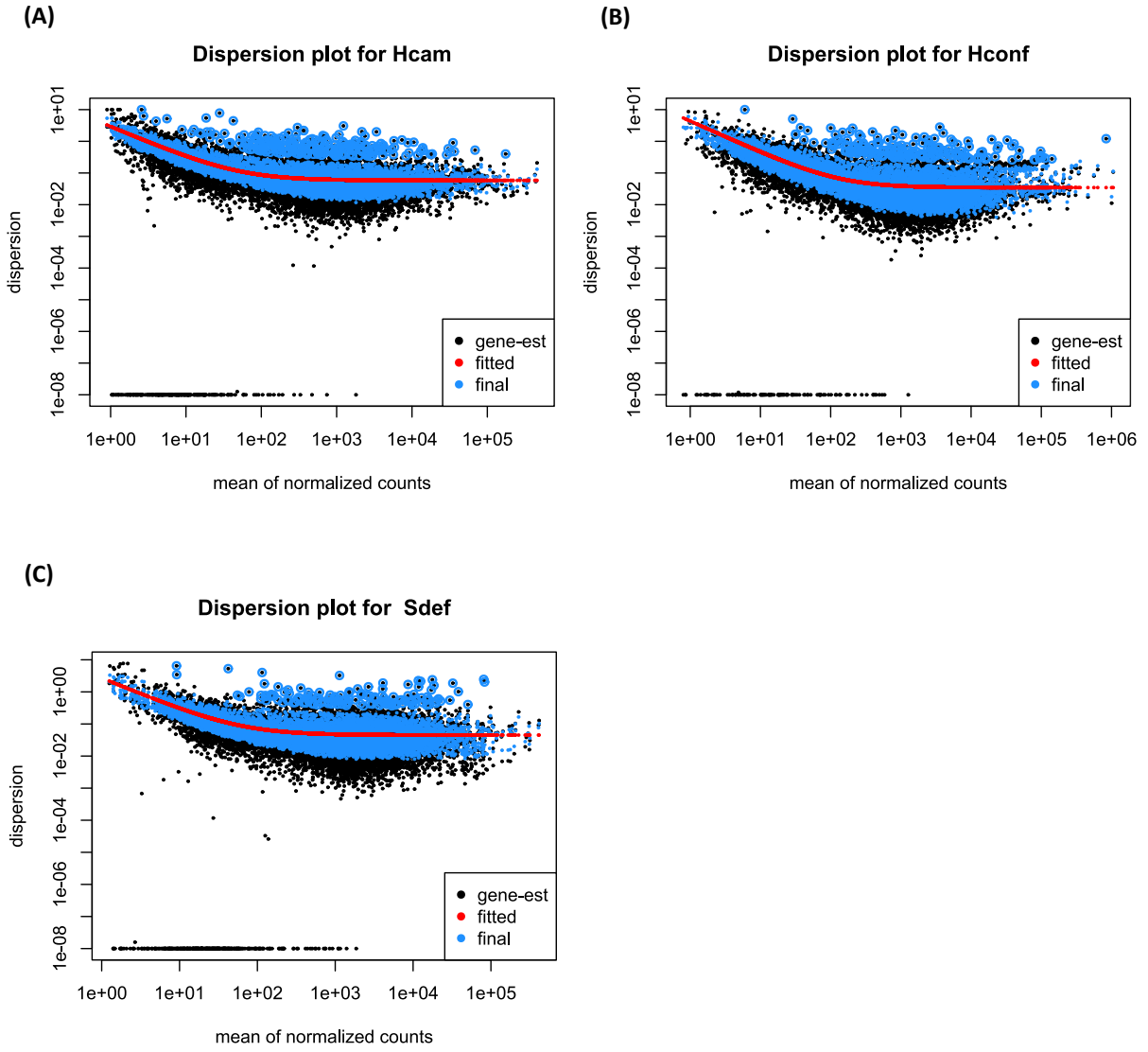

Figure C.3: Gene-wise dispersion estimates.

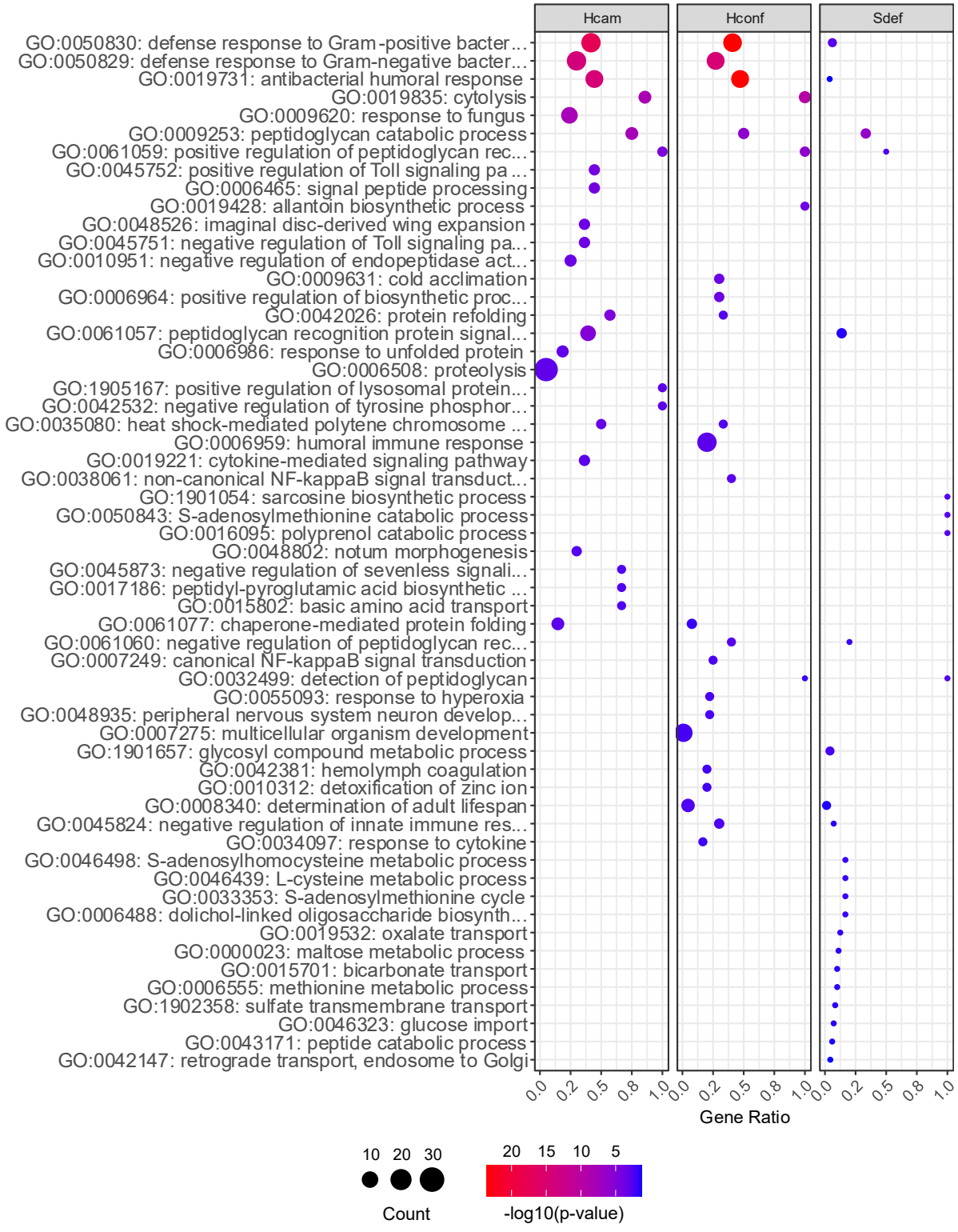

Figure C.4: GO term enrichment in differentially expressed genes between pathogen-challenged and unchallenged samples from *Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa*.

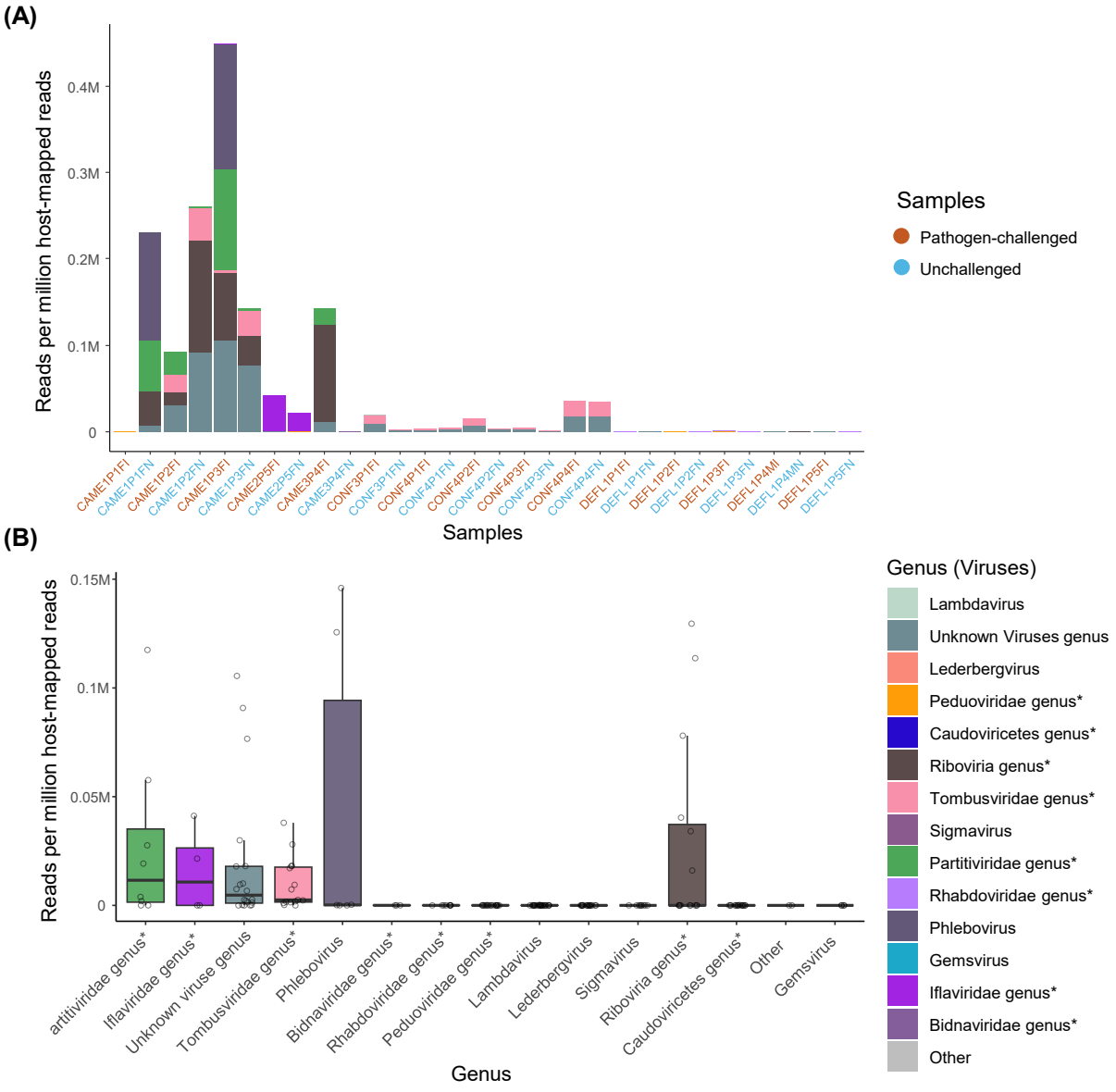

Figure C.5: Microbiome composition in pathogen-challenged and unchallenged samples from *Hirtodrosophila cameraria*, *H. confusa*, and *Scaptodrosophila deflexa*.
(A) Stacked bar plot showing the relative abundance of the top 14 virus genera (measured as reads per million host-mapped reads) across all samples. The 14 genera represent the union of the top 10 most abundant genera from each sample. Samples are ordered by species (*H. cameraria*, *H. confusa*, *S. deflexa*) and are color-coded by treatment status: pathogen-challenged (red) and unchallenged (blue). (B) Boxplot summarizing the relative abundance of each virus genus across all samples. Each box represents the distribution of abundance for a given genus, with individual data points indicating values from individual samples. Genera are ordered by median abundance. *Unclassified genus.