



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2022

Evolutionary-functional genomics for an enhanced resolution of arthropod gene function

Ruzzante Livio

Ruzzante Livio, 2022, Evolutionary-functional genomics for an enhanced resolution of arthropod gene function

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB_26D5A182653A3

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département d'Ecologie et Evolution

Evolutionary-functional genomics for an enhanced resolution of arthropod gene function

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine

de l'Université de Lausanne

par

Livio RUZZANTE

Master en Science en biologie de l'Université de Neuchâtel.

Jury

Prof. Andrea Superti-Furga, Président

Prof. Robert Waterhouse, Directeur de thèse

Prof. Nicolas Salamin, Co-directeur de thèse

Prof. Christophe Dessimoz, Expert interne

Prof. George Christophides, Expert externe

Lausanne

2022



UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole Doctorale

Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Andrea	Superti-Furga
Directeur·trice de thèse	Monsieur	Prof.	Robert	Waterhouse
Co-directeur·trice	Monsieur	Prof.	Nicolas	Salamin
Expert·e·s	Monsieur	Prof.	Christophe	Dessimoz
	Monsieur	Prof.	Georges K.	Christophides

le Conseil de Faculté autorise l'impression de la thèse de

Livio Ruzzante

Master en biologie, Université de Neuchâtel, Suisse

intitulée

**Evolutionary-Functional Genomics for an
Enhanced Resolution of Arthropod Gene Function**

Lausanne, le 3 février 2023

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Andrea Superti-Furga

Acknowledgments

This academic journey has been, for me, first and foremost an experience of incalculable human value and personal growth as well as maturity, awareness, and responsibility. I have been fortunate that all this has been supported by an exceptional, sensitive, patient, attentive, and precise mentor; Rob Waterhouse. I have been doubly fortunate because this support has been accompanied by his very high level of scientific expertise, diplomatic and dissemination skills, allowing me to be inspired and grow rapidly in the academic and professional spheres as well. I have always been impressed by Rob's planning skills and ability to take a long-term global view of matters. He always seems to be a step ahead of everyone by a few moves, like a chess grandmaster. I really could not have imagined a better guide than Rob, and these years will be truly unforgettable for me and an example of how to face the inevitable difficulties with a light mind, dedication, focus, and hard work through a step-by-step approach.

I cannot fail to give special thanks to my thesis partner, Romain Feron, who helped me enormously in the development of the bioinformatics workflow and to whom I acknowledge great seriousness, competence, and well-placed irony. Romain has brought a sense of freshness and new competencies to the lab that I could not have imagined before meeting him in person, challenging me to structurally change my approach to this research work.

I would further like to thank Maarten Reijnders for helping me take my first steps in Python scripting, for developing GO-Figure! and for providing GO annotations with CrowdGO, essential for generating the results of the second chapter of this thesis. I would also like to thank the other members of the Waterhouse group who provided me with valuable feedback and suggestions both during the research work and during the writing of the thesis: Giulia Campli, Harald Detering, Sagane Dind, and Antonin Thiébaud. I would like to thank the organisations and institutes that enabled and ensured the smooth running of this academic research project: the Swiss National Science Foundation, the Department of Ecology and Evolution of the University of

Lausanne, and the Swiss Institute of Bioinformatics. Thanks are also due to Anne Cuendet and Corinne Dentan, key people for the administrative management and bureaucratic support.

I would additionally like to thank all those people involved in the external support during these years. First and foremost, my beautiful family in Italy, who saw me leave the country when I was very young and watched me grow up at a distance, not without difficulty, but always supported me with unwavering trust and love. Mother Grazia, father Gian Luigi, and sister Elisa. They are joined by the friends I've met along the way, especially: Lucio, Chiara, Bianca, Sebi, Manu, Patrick, Didace, the Rondini, the Lake aux Rats, the F.C. Phenix brothers, and the carefree Fantacontea. Finally, a very special thanks to Pilar, who met me at a difficult time many years ago, and helped us build something important and irreplaceable together.

This work is dedicated to the memory of my proud and kind grandmothers Concetta and Luisa, my grandfathers Ivo and Antonino, my uncle Claudio, and my sweet, brave aunts Annalisa and Fiorella.

Table of Contents

Acknowledgments	2
Table of Contents	4
(EN) Summary	6
(FR) Résumé	8
Introduction	10
Chapter 1: <i>Evol-Feat</i>, a Toolkit for Quantifying Gene Evolutionary Features	16
Summary	16
Theoretical Background on Gene Evolutionary Features	17
Arthropod Gene Evolutionary Feature Characterisation	25
Design and Development of the Evolutionary Features	26
Orthology-Only Features	30
Phylogenomics Features	31
Gene Turnover Features	33
Synteny-Conservation Features	34
Additional User-Input Features	36
<i>Evol-Feat</i> Bioinformatics Workflow	36
Exploratory Assessments of the Feature Distributions	38
Interpreting the Evolutionary Features through Dimensionality Reduction Techniques	45
Concluding Remarks	52
Chapter 2: Characterisation of Evolutionary-Functional Correspondences	54
Summary	54
Theoretical Background on Evolutionary Profile Clustering and Functional Inference	55
Orthologous Group Gene Ontology Annotations	58
Functional Enrichments of Individual Evolutionary Features	58
An Overview of Orthologous Group Clustering	66
Self-Organising Map: Clusters of Evolutionary Profiles	68
Self-Organising Map: Superclusters of Evolutionary Profiles	73
Hierarchical Clustering of SOM Superclusters	77
Functional Annotations of the SOM Clusters	84
Functional Annotations of the SOM Superclusters	87
Lineage-Specific Evolutionary Frameworks	98
Concluding Remarks	104

Chapter 3: Functional Constraints on Insect Immune System Components Govern Their Evolutionary Trajectories	108
Summary	108
Appended Publication	110
Chapter 4: Scientific Collaborations and Additional Resources	112
Summary	112
Of Genes and Genomes: Mosquito Evolution and Diversity	113
Evolutionary Superscaffolding and Chromosome Anchoring to Improve <i>Anopheles</i> Genome Assemblies	113
Genome Sequence of the Wheat Stem Sawfly, <i>Cephus cinctus</i> , Representing an Early-Branching Lineage of the Hymenoptera, Illuminates Evolution of Hymenopteran Chemoreceptors	114
Draft Genome Assembly and Population Genetics of an Agricultural Pollinator, the Solitary Alkali Bee (Halictidae: <i>Nomia melanderi</i>)	115
Genus-Wide Characterization of Bumblebee Genomes Provides Insights into Their Evolution and Variation in Ecological and Behavioral Traits	115
<i>Anopheles</i> Mosquitoes Reveal New Principles of 3D Genome Organization in Insects	116
The <i>Orthophile</i> Workflow and the Phylogenetic Reconstruction of the Arthropoda Phylum	117
Chapter 5: Conclusions and Perspectives	120
Summary of the Principal Conclusions	122
Implications, Perspectives, and Future Outlook	127
Bibliography	136
Appendix 1: Immunity Case-Study	150
Appendix 2: Species Tree	172

(EN) Summary

The fast accumulating data on genetic sequences from all branches of the tree of life, supported by recent advances in genome sequencing technologies, has enabled rapid scientific progress in the field of evolutionary biology and in our understanding of gene evolution through natural selection. Such progress, however, requires the support of dedicated bioinformatics software to efficiently aid hypothesis formulation on gene function and evolution through comparative genomics approaches. These must furthermore be designed to be freely accessible and able to analyse an ever increasing amount of genomes from public repositories. Specifically, detailed experimental characterisation of the biological processes and functions associated with the increasing numbers of newly sequenced genes from less-known species is not scalable. Computational approaches are therefore required to build on existing knowledge from better-studied species and make predictions or inferences that are supported by a cross-species evolutionary framework. This thesis showcases new approaches aiming to support such research efforts through automated comparative genomics tools, by characterising the evolutionary trajectories of genes from hundreds of species and associating them with putative functional roles.

Arthropods are invertebrate animals, comprising almost 80% of the total described animal diversity. Including crustaceans, arachnids, myriapods, and insects, among others, they constitute an ideal study system for characterising gene evolutionary-functional correspondences, given their vast range of physiological and ecological adaptations over more than 600 million years of evolution. Furthermore, they comprise key species of fundamental epidemiological and agroecological interest, and a better understanding of their genetic sequences is fueled by global societal challenges, including the spread of devastating diseases, agricultural pest control, and biodiversity conservation. This thesis work thus aims to specifically improve the resolution of evolutionary-functional correspondences across arthropod species with available genomic data, and provide a readily-available resource to explore evolutionarily-informed putative gene function predictions of previously undescribed genes. These include the generation of statistically supported

hypotheses associating the evolutionary histories of gene families with biological processes such as organismal development, cellular organisation, immune response mechanisms, chemosensation, and insecticide resistance among others.

The first chapter presents the *Evol-Feat* toolkit, a scalable bioinformatics workflow to characterise the evolutionary trajectories of genes from multiple species. Used to compare the genes from 170 arthropod genome sequences, the first chapter focuses on the design and computation of “evolutionary features”: metrics designed to quantify changes in genetic sequences and gene repertoires. The second chapter focuses on clustering methods applied to the distributions of arthropod gene evolutionary features, to identify subsets of genes that show similar evolutionary profiles, and presents methodologies for associating these sets of genes with putative functions. The third chapter presents the application of the *Evol-Feat* toolkit to investigate the evolutionary-functional relationships between immune-related genes, in a proof-of-concept case-study focussed on the African malaria mosquito, *Anopheles gambiae*. The fourth and last chapter showcases the collaborative scientific work which accompanied the design and development of the *Evol-Feat* workflow through specific research studies including phylogenomic reconstructions of arthropod lineages, the exploration and description of available genomic resources, the investigation of physiological and ecological adaptations in hymenopterans, and bioinformatics software development.

(FR) Résumé

L'accumulation rapide de données sur les séquences génétiques de toutes les branches de l'arbre du vivant, soutenue par les récentes améliorations des technologies de séquençage du génome, a permis des progrès scientifiques rapides dans le domaine de la biologie évolutive et dans notre compréhension de l'évolution des gènes par la sélection naturelle. Ces progrès nécessitent toutefois l'utilisation de logiciels bioinformatiques spécialisés pour faciliter la formulation d'hypothèses sur la fonction et l'évolution des gènes par des approches de génomique comparative. Ces logiciels doivent en outre être développés pour être librement accessibles et capables d'analyser un nombre toujours croissant de génomes provenant de banques de données publiques. Plus précisément, la caractérisation expérimentale détaillée des processus et fonctions biologiques associés au nombre croissant de gènes nouvellement séquencés provenant d'espèces moins connues n'est pas extensible. Des approches computationnelles sont donc nécessaires pour s'appuyer sur les connaissances existantes des espèces mieux étudiées et faire des prédictions ou des déductions soutenues par un cadre évolutif inter-espèces. Cette thèse présente de nouvelles approches visant à accompagner de tels efforts de recherche par le biais d'outils automatisés de génomique comparative, en caractérisant les trajectoires évolutives de gènes de plusieurs centaines d'espèces ainsi qu'en leur associant des rôles fonctionnels hypothétiques.

Les arthropodes sont des animaux invertébrés, qui représentent près de 80 % de la diversité animale totale actuellement décrite. Comprenant entre autres des crustacés, des arachnides, des myriapodes et des insectes, ils constituent un système d'étude idéal pour caractériser les correspondances évolution-fonction des gènes, étant donné leur vaste gamme d'adaptations physiologiques et écologiques sur plus de 600 millions d'années d'évolution. En outre, ils comprennent des espèces clés d'un point de vue épidémiologique et agroécologique, et une meilleure compréhension de leurs séquences génétiques est nécessaire au vu des défis sociétaux globaux rencontrés, notamment la propagation de maladies dévastatrices, le contrôle des parasites agricoles et la conservation de la biodiversité. Ce travail de thèse a donc pour but d'améliorer la

résolution des correspondances évolution-fonction chez les espèces d'arthropodes ayant des données génomiques disponibles, et de fournir une ressource facilement accessible pour explorer les prédictions de fonctions génétiques informées par l'évolution de gènes non décrits auparavant. Ces prédictions comprennent la génération d'hypothèses soutenues statistiquement, associant l'histoire évolutive des familles de gènes à des processus biologiques tels que le développement de l'organisme, l'organisation cellulaire, les mécanismes de réponse immunitaire, la chimiodétection et la résistance aux insecticides, entre autres.

Le premier chapitre présente la boîte à outils *Evol-Feat*, un workflow bioinformatique permettant de caractériser les trajectoires évolutives des gènes de plusieurs espèces. Après la description théorique et méthodologique d'*Evol-Feat*, utilisé pour comparer les gènes de 170 génomes d'arthropodes, le premier chapitre se concentre sur la conception et le calcul des "caractéristiques évolutives" : des métriques conçues pour quantifier les changements dans les séquences génétiques et les répertoires de gènes. Le deuxième chapitre se concentre sur les méthodes de regroupement appliquées aux distributions des caractéristiques évolutives des gènes d'arthropodes, afin d'identifier des sous-ensembles de gènes qui présentent des profils évolutifs similaires, et présente des méthodologies pour associer ces ensembles de gènes à des fonctions putatives. Le troisième chapitre présente l'utilisation d'*Evol-Feat* pour étudier les relations entre évolution et fonctions chez les gènes liés à l'immunité, dans le cadre d'une étude de validation de concept axée sur le moustique africain de la malaria, *Anopheles gambiae*. Le quatrième et dernier chapitre présente le travail scientifique collaboratif qui a accompagné la conception et le développement d'*Evol-Feat* par le biais d'études de recherche spécifiques, notamment des reconstructions phylogénomiques de lignées d'arthropodes, l'exploration et la description des ressources génomiques disponibles, l'étude des adaptations physiologiques et écologiques chez les hyménoptères et le développement de logiciels bioinformatiques.

Introduction

The increasingly thorough sampling of the tree of life through genomic sequencing offers new opportunities to explore the links between gene evolution and gene function (Lemmon and Lemmon 2013; Waterhouse 2015). Due to their incredible biological diversity, arthropods present an ideal study system, considering their vast range of physiological, behavioural, and ecological adaptations (Giribet and Edgecombe 2019). However, this diversity, coupled with rapidly growing amounts of genomic data, presents a challenge to understanding the underlying biology because detailed functional genetics research requires significant investments in time, expertise, and resources. Comparative genomics approaches must therefore continue to assist the interpretation of such increasingly accessible genomic data (Thomas et al. 2020; Kapli, Yang, and Telford 2020). This thesis presents an evolutionary biology framework for comparative genomics analyses, providing biologists with an exploratory resource to capture the evolutionary histories of gene sequences and map them to their putative functional roles and constraints. It proposes a novel approach to characterise arthropod genomic diversity, shaped through hundreds of millions of years of evolution and at the origin of the extraordinarily diverse range of organismal functions. Deepening our understanding of such diverse arthropod evolutionary-functional correspondences cannot be sustained without an increased access to automated and scalable bioinformatics tools (Richards, Childers, and Childers 2018; Li et al. 2019). Such approaches will therefore be crucial, not only for scientific progress but also to inform strategies facing several of the global challenges posed to human societies by an increasing jeopardisation of arthropod populations, exacerbated by rapid climate change.

Arthropods (from the Ancient Greek words for “jointed-feet”) are invertebrate animals with an exoskeleton, segmented body, and paired jointed appendages. Comprising not only insects, arthropods group every species of chelicerates (spiders, scorpions, mites, horse-shoe crabs, and others) and mandibulates (crustaceans, hexapods, centipedes, millipedes, and others) (E. Clark, Peel, and Akam 2019). Together they form the phylum Arthropoda, which vastly surpasses the total number of species in all other phyla combined, and

with an estimated count of 7 million species, it comprises some of the largest animal taxonomic classes, orders, and families: 5.5 million insects species including 1.5 million beetles, and 1.3 million non-insect terrestrial arthropods, including 1 million mite species (Stork 2018). Arthropods could be viewed as the most representative phylum for metazoan comparative analyses simply for their sheer diversity and dominant presence within animal taxonomy. They are estimated to represent more than 80% of the total described animal diversity. As an example, Curculionidae, the true weevil family of beetles, is the largest animal family by species count, with currently 62'000 described species and 220'000 estimated species (Oberprieler, Marvaldi, and Anderson 2007). Moreover, insects have been proposed as a flagship for global biodiversity and conservation, due to their capacity to act as powerful indicators for environmental quality and threats, and the effectiveness of conservation programmes (Wilson and Fox 2021). Such an enormous taxonomic span mirrors the arthropods' astounding range of occupied habitats and ecological functions, such as pollination, seed dispersion, food supply, predation, and decomposition .

Their ecological diversity includes large and tiny herbivores (beetles and aphids), flying predators (dragonflies and robber flies), highly complex eusocial colonies (ants, termites, and social hymenopterans), and solitary individuals (solitary bees and spiders). They evolved to be plant pollinators (hymenopterans and lepidopterans), obligatory blood-feeders (ticks and lice), or both (mosquitoes and true flies). Arthropods have colonised every continent and can be found in most habitats of the world, from sea bottoms to deserts. As key ecological actors of so many diverse landscapes, this incredible arthropod biodiversity must be better understood and protected against threats of anthropogenic disturbances, including global warming, habitat disruption, pesticide overuse, and invasive species transportation (Harvey et al. 2020; Wilson and Fox 2021). Nevertheless, arthropods can hugely impact human societies with agroecological, forestry conservation, and epidemiological challenges as they are also associated with the spread of devastating diseases, crop failures, and ecosystem disruptions. Crop-destructing pest species like the fall armyworm or disease vectors like the African malaria mosquito constitute a threat to millions of people worldwide, particularly in the poorest regions of the world (Neafsey et al. 2015; Sinka et al. 2020; Harrison et al. 2019).

Entomologists have described much of this ecological diversity shaped over ~600 million years of evolution, and during the last two decades, genomics technologies have begun to catalogue and characterise the correspondingly large genetic and genomic diversity (Misof et al. 2014; Schwentner et al. 2017; Tihelka et al. 2021). The number of arthropod genomes in publicly available repositories has grown steadily thanks to decreasing sequencing costs and the relative ease of genome assembly when compared to often larger genomes of vertebrates (Gregory et al. 2007). Since the first arthropod genome in 2000, when the genome of the common fruit fly *Drosophila melanogaster* was published (Adams 2000), the National Center for Biotechnology Information (NCBI) databases have seen the number of hosted arthropod genomes rising to represent ~2'000 species (Schoch et al. 2020; Feron and Waterhouse 2022a). The constant and fast progress in the development of genome sequencing and assembly technologies will help increase this number further and to an extent unimaginable only a decade ago (Hotaling et al. 2021). As a consequence, increasingly comprehensive molecular phylogenomics studies are already continuously challenging our knowledge on arthropod evolution and development (Neafsey et al. 2015; E. Clark, Peel, and Akam 2019; Thomas et al. 2020; Sun et al. 2021). To be able to take advantage of these data to help understand arthropod biology, it is thus essential to develop innovative interpretative informatics tools powered by comparative genomics analyses (Feron and Waterhouse 2022b).

To bridge the gap between raw genomic data and biological insights, the scientific community is facing two main challenges: on the one hand, the large-scale genome data interpretability to associate genetic signatures with biological or ecological properties (Nagy et al. 2020). On the other hand, these evolutionary-functional correspondences are still fundamentally poorly understood. As such, evolutionarily-informed gene function predictions can help fill the knowledge gap linking gene sequences to biological functions with statistically supported hypotheses (Gabaldón and Koonin 2013). Investigating the relationships between gene evolution and genetic functional diversity might shed light on how natural selection forces shape these correspondences (Tatusov, Koonin, and Lipman 1997). Such forces are defined as the combination

of functional constraints hypothesised to regulate the evolution, survival, and adaptability of populations through selection-driven processes spanning millions of years (Koonin 2003; Koonin and Wolf 2010; Koonin 2011). Specifically, gene families are characterised by quantifiable characteristics such as differing levels of sequence conservation and evolutionary rates, or the emergence of lineage-specific gene losses and gene family expansions (Krylov 2003; Wolf, Carmel, and Koonin 2006; R. M. Waterhouse et al. 2007; R. M. Waterhouse, Zdobnov, and Kriventseva 2011).

The thesis focuses on the principal building blocks of the central dogma of genotype-to-phenotype mechanisms, that is, genes, and specifically protein-coding genes, assuming that the potentiality of function can be determined at the gene level. As with any model, this approach abstracts some of the realities of the biological complexities determined by preceding and subsequent molecular mechanisms of the gene-to-protein chain of events. It will not investigate gene regulatory networks of gene activation and expression through, e.g. intron sequences, regulatory elements, transcription factor binding sites, differential splicing, transcript inhibition and degradation, as well as different modes of gene sequence transmission, including horizontal gene transfer events, introgressions, virus-like and retrovirus insertions. Addressing such complications would require further analyses largely beyond the scope of this thesis. The approach and methodology defined hereafter are nevertheless conceptual, and aspects of it could be applied to investigate the evolutionary-functional correspondences of families and classes of genetic elements other than protein-coding genes.

Ultimately, the output-oriented goals of this thesis include: 1) providing experimental biologists with a readily-available resource where thousands of arthropod genes from model and non-model species alike can be explored in the context of their evolutionary-functional correspondences; and 2) providing computational biologists with a tool to capture and compare the evolutionary features of genes to support evolutionary-functional hypotheses on custom data. Gene evolutionary features include measurements of phylogenetic age and species-span, conservation and synteny scores, copy-number variation, sequence evolutionary rates, quantifications of gene family losses and expansion

events. Mapping similarly behaving genes (in terms of their evolutionary features) across a phylogeny of arthropod species then allows the exploration of associations with putative broad biological functions using Gene Ontology enrichment analysis. Such systematic approaches are necessary, and increasingly so, in this era of rapidly decreasing genome sequencing costs, characterised by fast-accumulating data but not as fast-developing interpretative capabilities. This work therefore aims to benefit the arthropod biology scientific community by providing a readily-available tool for exploring arthropod gene evolution and function while showcasing the leverageable information gathering potential of large-scale comparative genomics studies.

The first chapter of this thesis focuses on the theory and the definition of novel evolutionary features, their exploratory analyses, pairwise comparisons, their distributions across the genomic space, initial interpretative results, and a description of the bioinformatics workflow to compute these features, *Evol-Feat*. The second chapter then expands on the clustering techniques applied to the evolutionarily-defined profiles of arthropod orthologous genes and how they can be partitioned into separate modules of evolutionary trajectories. These modules are then assessed from a functional perspective, to characterise evolutionary-functional correspondences that help build informed hypotheses on arthropod gene functions. The third chapter presents a case-study application of the evolutionary features conceptual framework on the immune gene repertoire of the African malaria mosquito, *Anopheles gambiae*. The fourth and final chapter summarises the additional scientific collaborative research efforts and outputs that helped drive progress and shape the practical implementation of the *Evol-Feat* bioinformatics workflow and its theoretical conceptualisation.

Chapter 1: *Evol-Feat*, a Toolkit for Quantifying Gene Evolutionary Features

Summary

This chapter focuses on the reasoning behind the planning and implementation of the *Evol-Feat* bioinformatics workflow, starting with an overall introduction to the concept of evolutionary features and modules, and the biological motivations to investigate the correspondences between gene evolution and gene function. This is followed by the descriptions of the online resources providing the data at the origin of the thesis work. After describing the choice and nature of the databases, reporting then focuses on the evolutionary features grouped by data source type while describing their computational methodology and biological relevance. The chapter then expands on descriptions dedicated to the *Evol-Feat* workflow usage, reproducibility, and scalability. The last section focuses on preliminary and exploratory statistics, including pairwise comparisons, correlations, and principal component analysis of the feature distributions.

Theoretical Background on Gene Evolutionary Features

Comparative genomics offers opportunities for improving the understanding of animal biology through integrative and evolutionarily informed approaches to elucidating the putative functions of thousands of genes from hundreds of genomes. By leveraging existing resources and propagating the knowledge from better to less studied systems, comparative approaches can be used to generate well-supported hypotheses on gene function (Koonin 2005). This requires developing cross-species comparisons to quantify patterns of change and explore the relations between how genes evolve and the biological roles they perform: gene evolutionary-functional correspondences. Such correspondences, within the context of comparative evolutionary analyses, are identified as originating from a specific theoretical framework of analysis, and cornerstone of describing evolutionary relationships between genes: sequence homology, and more specifically, orthology.

Homology has been used by evolutionary biologists throughout history as a term describing common evolutionary ancestry between anatomical structures across different species. Usually supported by similarities in morphology and function, two anatomical structures would nevertheless not be defined homologous when not originating from the same common ancestor species. For centuries, comparing fossil records to hypothesise the evolution of homologous anatomical structures fueled the scientific description of species taxonomy and evolutionary relationships. With the advent of molecular biology, sequence homology became the equivalent concept applied to DNA, RNA, and protein sequences, defining two or more sequences homologous strictly when sharing common molecular ancestry (Fitch 1970). High percent sequence similarity and identity of aligned amino acids and nucleotides are good starting measures to infer homology across sequences; however these are not definitive, as they may arise from convergent evolution or by chance, especially in shorter sequences. High confidence inferences of sequence homology must thus be supported by appropriate statistical testing, such as the bit score and E-value from the NCBI BLAST (Kerfeld and Scott 2011), determining the likelihood of sequence similarities arising by chance.

Orthology extends the concept of homology by defining as orthologs those genes in different species derived by vertical descent from a single gene in their last common ancestor species (Koonin 2005): i.e. homologous genes are orthologs when emerging through speciation events. Sequence homology can, however, emerge through other mechanisms too: paralogy indicates homology via within-species gene duplication events, and xenology via horizontal gene transfer. While xenologs are considerably less widespread in eukaryotes than prokaryotes (Krylov 2003), paralogs are common, and their true identity is often challenged by imperfect automated genome annotations, where undetected genes in other species could grant them ortholog status. Adding to the complexity presented by such intricate evolutionary scenarios, the identification and definition of homologous genes are further complicated by the inclusion of more than one pair of genes and species, and the occurrence of gene losses and domain rearrangements such as gene fusions and fissions. Additionally, different sequence evolutionary rates can contribute to an increased sequence divergence through point mutations, insertions, and deletions. As well as affecting gene function to varying degrees, these possible changes are also likely to influence orthology delineation and subsequent functional inferences guided by orthology data. The full range of evolutionary scenarios must thus be considered when using comparative approaches to investigate putative gene functions. The consequences of such diversifications may result in different functional outputs through pseudo- or neo-functionalisation, non-functionalisation, sub-functionalisation, or gene dosage increase (Zhang 2003; Micheli and Camilloni 2022).

Nevertheless, orthology is widely considered a powerful approach to infer the biological functions of uncharacterised genes in newly sequenced genomes from experimentally validated gene functions of model organisms. At the origin of this often implicit procedure stands the ortholog conjecture, stating that orthologs carry out biologically equivalent functions in different organisms; and that the functions of paralogs typically diverge after duplication (Koonin 2005). More importantly, it is generally recognised that physiologically essential functions (where the corresponding gene knock-outs are lethal to the organism) are likely to be conserved across species through conserved orthologous genes

(Altenhoff et al. 2012; Gabaldón and Koonin 2013). Additionally, the ortholog conjecture stands firmer concerning single-copy orthologs, i.e. orthologous genes with no detected paralogs, and likely indicators of highly conserved sequences bound to be maintained so as not to disrupt the organism's fitness. Such observations could mechanistically be explained, for example, by the fact that the oldest genes are, to some extent, interlocked in more extensive functional networks, wholly inherited during the evolution of extant lineages (Schlitt et al. 2003; Domazet-Lošo, Brajković, and Tautz 2007).

As orthologous genes must refer to their last common ancestor, orthology delineation is, by definition, hierarchical and relative to the selected set of species. Different methodologies have been employed to capture orthologous genes, for example, by reconstructing homologous gene trees followed by a reconciliation with the species tree (Emms and Kelly 2019). *OrthoDB* - the hierarchical catalogue of orthologs - (Waterhouse et al. 2013) employs a different procedure, progressively clustering all-against-all pairwise sequence comparisons. The resulting hierarchical clusters of orthologs are subsequently expanded with all closely related in-paralogs (within-species duplicated genes emerging after a particular speciation event). This procedural definition of orthology unequivocally defines the multi-species hierarchical relationships of genes, including orthologs, co-orthologs (multi-copy orthologs), and paralogs. The hierarchical clusters of orthologs are named *orthologous groups* and refer to sets of all homologous genes evolving from a single ancestral gene after a reconstructed speciation event. The earliest phylogenetic classification efforts to define hierarchical clusters of orthologs provided their first definitions, including the microbial clusters of orthologous groups of proteins (COGs, (Tatusov et al. 2000)) and the eukaryotic orthologous groups (KOGs, (Koonin et al. 2004)). However, only the *OrthoDB* orthologous groups, as defined by (Kriventseva et al. 2019), will serve as the supporting data for the orthology delineation of arthropod genes at the basis of this thesis work.

Orthology delineation, specifically the KOGs, supported the first systematic explorations of the correspondences between gene evolution and function. Multi-species comparisons to explore gene evolutionary features were pioneered by Koonin and colleagues in (Krylov 2003), identifying the propensity

for gene loss (PGL) and sequence divergence as complementary measures of the conservation of a gene. Analysing seven eukaryotic genomes, the PGL, a measure of gene loss frequency across eukaryotic lineages, was positively correlated with the amount of amino-acid substitutions in the corresponding protein sequences. Low PGL scores were further associated with high gene essentiality (measured as the lethality from gene knockout effects), high expression rates, and a higher number of protein-protein interactions. Nevertheless, fewer amino-acid substitutions were not found to be significantly associated with the indispensability of the biological function, leaving the hypothesis that essential genes tend to evolve slower than non-essential ones (the “knockout-rate” prediction, (Jordan et al. 2002)) to be further tested. These exploratory results and first quantification efforts of evolutionary features highlighted gene loss as a strong evolutionary driving force defining the gene repertoires of eukaryotes through functional constraints. The functional adaptation potential of eukaryotic genes through higher sequence mutation frequencies, as captured by the KOGs’ evolutionary rates, seems to be restricted by the size of the protein’s interaction network, expression levels and physiological viability.

Initially subject to correlational analyses only, these first evolutionary features were extended and further investigated in (Wolf, Carmel, and Koonin 2006) with a Principal Component Analysis (PCA). The resulting first three main axes were found to be related to 1) the gene’s *status*: positive contributions from expression levels, number of protein-protein interactions, number of paralogs (copy-number), and knockout lethality (essentiality), in parallel with negative contributions from sequence evolutionary rates and gene losses; 2) the gene’s *adaptability*: enhanced by an increased number of gene copies, a higher number of protein-protein interactions, and lower knockout lethality; 3) the gene’s *reactivity*: positively driven by gene losses, expression levels, and copy-number while negatively driven by the number of protein-protein interactions. The first axis was interpreted as the gene’s overall essentiality, while the second and third axes as a reflection of the role of the gene in the organism’s functional and evolutionary plasticity. Functional annotations of extreme values highlighted the associations of high PC1 (status) scoring genes with the translation system and cytoskeletal proteins. High PC2 (adaptability)

scoring genes were instead associated with cellular processes and signalling genes, while high PC3 (reactivity) scores with metabolic enzymes.

Soon after, macroevolutionary adaptations were characterised by more comprehensive phylogenies associated with *D. melanogaster* gene evolutionary histories, an approach the authors called *genomic phylostratigraphy* (Domazet-Lošo, Brajković, and Tautz 2007). Without the direct use of orthology delineations, the analysis framework included *D. melanogaster* phylogenomic strata and embryo gene expression data. The phylogenomic strata were defined from sets of homologous genes obtained with NCBI BLAST and associated with different phylogenetic ages of *D. melanogaster* genes. The results of the phylostratigraphic map indicated epoch-specific emergences of protein families and that some genes retain ancient signals of their evolutionary histories: ancestral genes become interlocked into pathways during the evolution of lineage-specific adaptations. With the increasing availability of genome data, genomic phylostratigraphy may be applied to uncover broad evolutionary processes in *Drosophila* and other lineages.

These rather theoretical observations came in parallel with a series of more applied studies which examined additional evolutionary features and expanded the exploration of evolutionary-functional correspondences. The acquisition of the second mosquito genome, *Aedes aegypti*, enabled the detailed characterisation of the immune-related genes and pathways by sequence divergence, copy-number, and species span (Waterhouse et al. 2007), in comparison with the already available genome sequences of *A. gambiae* and *D. melanogaster*. Distinct evolutionary trajectories were associated with different immunity-related functional modules: recognition receptors of bacteria and fungi are characterised by universal and copy-number expanded genes; adaptations of ancient recognition domains drive neo-functionalizations such as malaria parasite recognition; immune signal modulators show a range of dynamics reflecting the species-specific adaptations of protein networks' assemblies; signal transduction pathways are universally constrained, and their genes evolve in concert, likely interlocked in protein-protein interactions; effector mechanisms are characterised by diverging evolutionary trajectories, depending on their effector activity.

Evolutionary features of essentiality, gene copy-number, taxonomic span and sequence divergence were further investigated with the categorisation of orthologous genes from 40 vertebrates, 23 arthropods, and 32 fungi genome sequences (Waterhouse, Zdobnov, and Kriventseva 2011). Supported by a vastly increased genome availability, sequence evolutionary rates were found to be significantly more constrained in single-copy orthologs and in orthologous groups containing physiologically essential genes, confirming the “knockout-rate” prediction previously undetected in (Krylov 2003). The study further highlighted two main evolutionary trajectories within gene repertoire diversity, the first characterised by young, lineage-specific, and evolutionarily dynamic sequences, and the second by old, universal, and evolutionarily stable sequences. The “single-copy control” versus “multicopy licence” is hypothesised to play an underappreciated role in the forces driving the expansion of the evolutionary landscape of gene families.

Shortly after, the *evolutionary rate covariation* (ERC), defined as the sequence covariation of a pair of proteins over evolutionary time, was identified as a promising phylogenetic signature in 18 budding yeast species; it was associated with proteins’ physical interactions and shared function (N. L. Clark, Alani, and Aquadro 2012). Notwithstanding assumptions of orthology, ERC may reflect the co-evolution of the molecular interfaces between interacting proteins but has nevertheless been demonstrated to also emerge across non-interacting but co-functional enzymes. Multi-species ERC characterisations of entire gene repertoires could serve as a powerful addition to the suite of evolutionary features for the automated functional group assignments of uncharacterised proteins.

Others too have recognised the potential of using approaches to quantify gene evolutionary trajectories. Several bioinformatics tools were developed to define and measure gene evolutionary features and then explore how these relate to biological pathways and functions. Evolutionary barCode (EvoluCode, (Linard et al. 2012)) used metrics of sequence and domain conservation, orthology, synteny, and phyletic distributions to barcode the human proteome on a vertebrate evolutionary timescale. Resulting clusters of evolutionary barcodes,

consisting of protein subsets sharing similar evolutionary histories, were successfully enriched with biological functions. The automated workflow was successfully implemented to predict the functional roles of the proof-of-concept methionine salvage sub-pathway, and could be extended to all pathways and higher evolutionary scales, although with particular care to the high sensitivity derived from highly correlated parameters.

CLustering by Inferred Models of Evolution (CLIME, (Li et al. 2014)) used phylogenetic profiles within a species tree, a homology matrix, and pathways of interest to partition gene sets into evolutionary models. CLIME expands the pathways of interest with newly scanned genome components by partitioning the gene sets into evolutionary modules defined by the homology- and phylogeny-inferred evolutionary histories. Applied to ~1'000 annotated human pathways and the proteomes of yeast, red algae, and the malaria parasite, it revealed that half of the resulting modules contained genes with no shared sequence similarity, and increased the modularity of traditionally well-studied pathways. While limitations arose from the poorer resolution of the homology matrix compared to orthology delineations, these observations further highlight the need for alternative evolutionary features in the quest for putative gene function prediction. Similarly to CLIME, Protein Phylogenetic profiling (ProtPhylo, (Cheng and Perocchi 2015)) used co-evolution profiles constructed from orthology delineations and phyletic profiles corresponding to presence/absence in the species subset to identify functionally-linked proteins in 2'048 sequenced organisms. ProtPhylo ultimately provides functional annotations for all the considered proteins, including subcellular localisations, the presence of transmembrane helices, protein domain families, and protein-protein interactions.

In summary, as a result of the increased genomic sampling and new bioinformatics tools for comparative genomics studies, several efforts have quantified changes in gene sequences and copy-numbers across hundreds of species, associating them with the emergences and losses of gene families, phenotypical, and physiological adaptations (Thomas et al. 2020; Fernández and Gabaldón 2020; Guijarro-Clarke, Holland, and Paps 2020). These initial exploratory studies and subsequent detailed comparative genomics efforts

provide us with an overview of the expectations and potential progress in advancing our understanding of evolutionary-functional correspondences of arthropod genes. Different categories of genes have originated through and can be identified by different evolutionary trajectories, here defined in terms of rates of repertoire and sequence changes. Defining and computing different measures of gene evolution can place genes along a spectrum of evolutionary features such as slow- to fast-evolving, or stable to dynamic copy-numbers. Some of these properties are correlated, some are anti-correlated, and others seem unique or follow more subtle underlying interactions. Subsets of genes can be characterised by extreme values of certain features while presenting average values of others.

These observations suggest that unique evolutionary features, or combinations of features, can capture specific functional properties while others, acting in concert, are likely reflections of the evolution of genes in interlocked pathways bound by physiological constraints. Characterising such evolutionary trajectories can thus inform the prediction of generic functional properties such as essentiality, expression levels, or interactions; detailed and lineage-specific functional adaptations; or broader functional complexes, modules, and pathways. Rather than seeing the complexity of gene families' evolution as a problem, this thesis work aims to leverage the full spectrum of possible evolutionary trajectories of arthropod genes, harnessing the existing knowledge to build accordingly complex models and enhance the resolution of hypotheses on gene function.

Current limitations of such studies include a low descriptive resolution originating from either the inclusion of only a few measured evolutionary features or from a small and imbalanced selection of genomes, especially in older studies. Methods to quantify evolutionary features may not be easily reproducible as they are often conceived and developed for single applications, especially in mainly theoretical studies. Most studies highlight genomic variations across vast taxonomic ranges, describing patterns of evolution across metazoan and eukaryotic evolutionary timeframes. These likely overlook more subtle lineage-specific evolutionary trajectories associated with more recent arthropod radiations of great societal interest, such as mosquitoes and pollinators.

Moreover, orthology delineation may prove challenging, especially for large evolutionary timescales or with a limited number of considered genomes. These limitations are often identified and commented on by the authors, who point out how such coarse measurements and analyses will become more refined with the increasing availability of sequenced eukaryotic genomes (Krylov 2003; Domazet-Lošo, Brajković, and Tautz 2007; Li et al. 2014). Ultimately, this thesis takes advantage of growing genomic resources for diverse arthropod species by taking inspiration from these initial explorations on quantifying gene evolutionary features and relating them to putative biological roles and functional modules. It first defines a higher-resolution evolutionary framework of comparative analyses to characterise and partition arthropod genes and secondly provides a customisable workflow for exploring evolutionary-functional correspondences and using them to inform hypotheses on gene function, particularly for less studied organisms.

Arthropod Gene Evolutionary Feature Characterisation

The major part of the thesis work aimed to 1) take advantage of improving taxonomic resolution with the increased collection of available arthropod genomes; 2) define a more comprehensive suite of quantifiable multi-species evolutionary features; and 3) build a deployable, reproducible and scalable workflow to measure the evolutionary features on custom datasets. The thesis's scientific novelty consists of defining a suite of 16 evolutionary features and computing them across genomic and phylogenomic data spanning 170 arthropod species. The species selection includes all high-quality arthropod genomes available in *OrthoDB*, the hierarchical catalogue of orthologs, version 10.1 (Kriventseva et al. 2019). It spans 56 dipterans, 40 hymenopterans, 16 lepidopterans, 16 hemipterans, 10 arachnids, nine coleopterans, six crustaceans and 17 other arthropods comprising two odonates, two springtails and one representative caddisfly, strepsipteran, cockroach, mayfly, thrip, orthopteran, ice crawler, body louse, termite, bristletail, dipluran, horseshoe crab, and centipede. The complete list of species is available at <https://www.orthodb.org> and from the species tree in Appendix 2. In the context of this thesis, the evolutionary

features are defined as metrics capturing particular characteristics of the evolutionary trajectories of arthropod genes and were computed at their orthologous group level, meaning that individual scores were assigned to each arthropod orthologous group (sometimes referred to as gene family). Such features can be grouped into four main subgroups: 1) orthology metrics computed from *OrthoDB's* orthology relationships, 2) ancestral state metrics computed from gene family evolution analyses, 3) phylogenomics relationships, and 4) sequence and structural evolutionary metrics computed from protein-protein alignments and genome annotations.

Design and Development of the Evolutionary Features

Features were quantified as a suite of 16 orthology-based evolutionary metrics per orthologous group that included: universality (UNI) computed as the proportion of the total species present; duplicability (DUP) computed as the proportion of species present with multi-copy orthologs; average ortholog copy-number (ACN); copy-number variation (CNV) computed as the standard deviation of ortholog counts per species present divided by the ACN; evolutionary age (AGE) of the last common ancestor in terms of millions of years since divergence from the ultrametric species phylogeny; relative universality (RUN) computed as the universality score divided by the number of species emerged from the orthologous group's last common ancestor species. Gene turnover was estimated using the Computational Analysis of gene Family Evolution (*CAFE5*, (Mendes et al. 2020)) software to quantify proportions of gene gains (expansions, EXP), gene losses (contractions, CON), or no copy-number changes (stable, STA). Additional *CAFE5*-derived metrics were computed by dividing gene gains, losses and no copy-number changes by the orthologous group's last common ancestor clade species-span, i.e. relative expansions (REX), relative contractions (RCO) and relative stability (RST). Orthology data combined with genomic location data were used to quantify average synteny conservation (SYN) as the proportion of orthologs that maintain their orthologous neighbours in the genomes of the other species. The synteny conservation metric was followed by the addition of the maximum cross-species synteny conservation

score (MSY), and the synteny conservation score relative to the size of the orthologous group's last common ancestor clade species-span (RSY). Finally, each orthologous group's evolutionary rate (EVR) corresponds to the average rate of protein sequence divergence normalised by the distance between each pair of species as computed by *OrthoDB* (Waterhouse et al. 2013).

Table 1: Evolutionary feature descriptions. For each evolutionary feature, the feature name, acronym, description, and source data are presented. Acronyms: CAFE5, Computational Analysis of gene Family Evolution version 5; OG, Orthologous Group; GFF, General Feature Format file.

Evolutionary Feature	Acronym	Description	Data Source
Taxonomic Age	AGE	Age of the last common ancestor of species in an OG, in terms of millions of years since divergence, computed from the ultrametric species phylogeny	170-arthropod orthology; 170-arthropod phylogeny
Universality	UNI	The proportion of the total species present in an OG (all species, UNI=1)	170-arthropod orthology;
Duplicability	DUP	The proportion of species present in an OG that have multi-copy orthologs	170-arthropod orthology;
Average Copy Number	ACN	The average (mean) ortholog copy number across all species present in an OG	170-arthropod orthology;
Copy Number Variation	CNV	The standard deviation of ortholog counts per species present in an OG divided by the ACN	170-arthropod orthology;
Expansions	EXP	CAFE5 quantified proportions of gene gain nodes for an OG	170-arthropod orthology; 170-arthropod phylogeny
Contractions	CON	CAFE5 quantified proportions of gene loss nodes for an OG	170-arthropod orthology; 170-arthropod phylogeny
Stability	STA	CAFE5 quantified proportions of no copy-number change nodes for an OG	170-arthropod orthology; 170-arthropod phylogeny
Synteny	SYN	The species-averaged proportion of orthologs in an OG that maintain their orthologous neighbours in the genomes of the other species	170-arthropod orthology; 159-arthropod GFF
Evolutionary Rate	EVR	The average rate of protein sequence divergence normalised by the distance (% identity) between each pair of species as computed by <i>OrthoDB</i>	170-arthropod orthology;
Relative Universality	RUN	UNI divided by the number of nodes derived from OG's last common ancestor	170-arthropod orthology; 170-arthropod phylogeny
Relative Expansions	REX	EXP divided by the number of nodes derived from OG's last common ancestor	170-arthropod orthology; 170-arthropod phylogeny
Relative Contractions	RCO	CON divided by the number of nodes derived from OG's last common ancestor	170-arthropod orthology; 170-arthropod phylogeny
Relative Stability	RST	STA divided by the number of nodes derived from OG's last common ancestor	170-arthropod orthology; 170-arthropod phylogeny
Relative Synteny	RSY	SYN divided by the number of nodes derived from OG's last common ancestor	170-arthropod orthology; 170-arthropod phylogeny;

			159-arthropod GFF
Maximum Synteny	MSY	The species-maximum proportion of orthologs in an OG that maintain their orthologous neighbours in the genomes of the other species	170-arthropod orthology; 159-arthropod GFF

Orthology-Only Features

Genes of any two or more different species can be defined as orthologous when they descend from a single gene present in the last common ancestor of the compared species. Orthologous groups are sets of genes across multiple species found to be orthologous amongst themselves. Orthology is a powerful evolutionary concept for gene function prediction if we assume that physiologically essential functions are likely to be conserved across species through orthologous genes. In this context, our first analyses are based on a dataset of orthology information from the genesets of 170 arthropod species downloaded from *OrthoDB* v10 (Kriventseva et al. 2019). The orthology delineation was obtained from all 170 arthropod species, comprising 2'206'003 genes assigned to 82'474 orthologous groups using pair-wise assessments of protein sequence homology between complete genomes as described in (Zdobnov et al. 2021). Orthology-based features were designed to quantitatively describe gene family evolutionary trajectory properties capturing arthropod gene essentiality, phylogenetic span, age of emergence, and copy-number variations.

Four evolutionary features were computed directly from gene copy counts by processing the orthology delineation table (relating gene identifiers to orthologous group memberships). Universality (UNI) was defined as the number of arthropod species represented in each orthologous group divided by the total number of considered species (170 arthropods). For example, an orthologous group that included genes from only two species would be considered very specific, scoring a total of $2/170$ UNI, or approximately 0.012 out of 1. In contrast, a score of 1 would indicate an orthologous group recovering at least one gene for all 170 species and defining it as universal.

Duplicability (DUP) was defined as the proportion of arthropod species in an orthologous group with more than one gene copy over the total number of species with at least one gene copy. The duplicability feature was designed to capture the propensity of orthologous groups to evolve towards either generalised multi-copyness or single-copyness. This feature and the following ones describing gene copy-number propensities and variations are deemed

relevant in characterising evolutionary-functional correspondences, as gene copy number expansions are supposedly a critical evolutionary strategy likely driving functional diversity, specifically gene neo- or sub-functionalisation. A duplicability score of zero indicates that the orthologous group is fully represented by single-copy genes, whereas a score of one indicates that the orthologous group is granted a multi-copy licence, as described in (Waterhouse, Zdobnov, and Kriventseva 2011).

Average copy number (ACN) was defined as the mean average number of gene copies across the orthologous group's representative species. Copy-number variation (CNV) was defined as the standard deviation of ortholog counts per species in any given orthologous group divided by the ACN. These features were designed to capture the orthologous group's trend across the species phylogeny to settle the gene family expansion towards a particular number of copies. Gene families can differ significantly in how much freedom functional constraints allow for expansions, ranging from just a couple of gene copies to several dozens. ACN scores range from 1, indicating that the orthologous group will likely maintain only one gene copy, to the maximum computed score of 73 mean gene copies per species. A CNV score of 0 indicates that the orthologous group will likely maintain the same number of copies across all represented species. In contrast, an increase will indicate a larger variability of gene copy counts, up to the maximum computed score of 6.

Phylogenomics Features

The time of emergence of each orthologous group's last common ancestor gene could be estimated using a molecular phylogenomic reconstruction of the 170 arthropod species. The species phylogeny was reconstructed using orthologous groups, which showed at least 90% single-copyness (meaning that a minimum of 90% of the orthologous group's represented species assigned one and only one gene copy to the group) and 90% universality (meaning that a minimum of 90% of the species assigned at least one gene copy to the group). This resulted in a selection of 1'363 arthropod orthologous groups. More

stringent selection parameters did not yield sufficient orthologous group numbers as the occurrence of fully universal single-copy genes inversely decreases with the considered evolutionary timescale, being absent in the *OrthoDB* version 10's Arthropoda-level delineated orthology. Multi-copy genes, when occurring, were processed by selecting the longest sequence; missing genes were represented with '-' gaps in the concatenated super-alignment.

The sequence retrieval, processing, trimming, alignment, concatenation and phylogenetic reconstruction was computed with *Orthophile*, a side-project explicitly developed for this purpose, and with additional manual adjustments to reflect the topology from the most recent literature. Obtaining the best possible phylogenetic reconstruction plays an essential role in the validity of the overall *Evol-Feat* workflow, as it constitutes the preliminary data from which most features are subsequently computed, and the correct placement of arthropod taxa had to be carefully assessed. More details on the *Orthophile* workflow and usage, as well as the detailed arthropod phylogenetic reconstruction methodology, can be found in the specific section in Chapter 4. The resulting species tree can be found in Appendix 2.

The corresponding taxonomic age (AGE) dates each arthropod orthologous group with an evolutionary feature score ranging from 602 million years (myr), indicating orthologous groups with representative genes from the last common ancestor species of all considered arthropods; to 7.5 myr, representing orthologous groups with the only representative genes from the two most closely-related species, *Drosophila yakuba* and *Drosophila erecta*. The relative universality (RUN) feature builds on the species phylogeny to compute a score representing the species span within the specific clade the orthologous group has emerged from. That is, the number of species represented in the orthologous group divided by the number of species emerging from the orthologous group's last common ancestor. This feature was designed to capture universality within smaller clades, particularly relevant for younger orthologous groups.

Gene Turnover Features

Gene turnover was estimated with computational analysis of gene family evolution using *CAFE5* (Mendes et al. 2020). Outputs of *CAFE5* include orthologous group-specific trees with annotations of estimated gene copy counts on the internal nodes of the species phylogeny. From the extant species' gene copy counts, *CAFE5* modelled each orthologous group tree with a *lambda* parameter, the rate of change of evolution (gene turnover rate, also called birth-death parameter). Maximum likelihood scores were then used to select among the different distributions and gamma rates to define the final estimated number of gene copies per ancestral node. The analyses included the addition of an error model estimation, compensating for assembly errors (e.g. missed detection of genes during orthology delineation). Computational efforts did not produce results unless the 84 largest orthologous groups were removed from the dataset, as the extreme variation in copy numbers would impede the algorithm's convergence to estimate the global *lambda*. The first *CAFE5* run, only including the orthologous groups going up to the root of the species tree and without the largest orthologous groups, estimated a global *lambda* ($\cong 0.0017$) as a full-phylogeny baseline value to compute the orthologous group-specific gene turnover rates. The second *CAFE5* run refined the estimates with the specification of the previously obtained *lambda* and computed estimates for all converging orthologous groups (including the non-root groups). A recursive custom script was then employed to gradually re-include the 84 largest orthologous groups into *lambda* convergence with subsequent *CAFE5* runs, merging the results at each iteration. Finally, from the 82'474 orthologous groups only 18 could not be assigned estimated gene turnover rates and ancestral gene copy counts, for the reconstruction of their ancestral dynamic state phylogeny never reached convergence.

For each resulting orthologous group tree, summary counts of gene copy number expansion and contraction events were inferred. Gene expansion events (EXP) were detected when children nodes showed higher copy counts than the parent node with custom regular expression formulas recursively scanning each orthologous group tree. Gene contraction (CON) events, or gene losses, were

detected when children nodes showed lower copy counts than the parent node, and stable (STA) gene copy counts were detected when no change appeared between parent node counts and children node counts. Lineage-specific gene turnover features were defined by dividing the previously defined *CAFE5* event counts by the total number of nodes emerging from the orthologous group's common ancestor species. This would result in three additional features which captured clade-specific dynamics rather than generic absolute counts: relative expansions (REX), relative contractions (RCO), and relative stability (RST). The inclusion of gene turnover features was fundamental as lineage-specific gene losses and expansions have been consistently identified as major contributors to eukaryotic functional adaptations driven by selective pressure (Krylov 2003), and as main evolutionary forces determining gene sequence divergence and conservation.

Synteny-Conservation Features

The role of synteny in determining or keeping a functional genetic repertoire is being increasingly explored and understood. It is thought that metazoans show signs of ancestral syntenic blocks, which could have formed a constrained framework of an essential genetic toolbox fundamental for animal chromosome evolution (Simakov et al., 2020), and synteny consideration has been proven to increase overall genome assembly quality (Anselmetti et al. 2015; Vakirlis, Carvunis, and McLysaght 2020; Waterhouse et al. 2020). Additionally, the insect immune system exploratory study results showed how important such a measurement of synteny can be by adding a unique evolutionary signature capable of capturing a yet unexplored corner of the evolutionary genetic map (Ruzzante et al. 2022), as detailed in Chapter 3.

Most genome annotations were retrieved from several online genome repositories, including the AntGenome and BeeBase portals from the Hymenoptera Genome Database (Elsik et al. 2018), the AphidBase portal from the Bioinformatics Platform for Agroecosystem Arthropods (Legeai et al. 2010), ButterflyBase (Papanicolaou et al. 2008), FlyBase (Thurmond et al. 2019),

EnsemblGenomes (Yates et al. 2022), the i5k Initiative (i5K Consortium 2013), VectorBase (Giraldo-Calderón et al. 2015), InsectBase (Mei et al. 2022), MidgeBase (Yoshida et al. 2022), GenBank (Sayers et al. 2019), and RefSeq (O’Leary et al. 2016). The corresponding gene identifiers were mapped to the *OrthoDB* custom gene identifications with custom Python scripts employing regular expression functions. Of the 170 species, only eleven General Feature Format (GFF) files could not be obtained, including the ones from: *Drosophila busckii*, *Bombyx mori*, *Camponotus floridanus*, *Ceratosolen solmsi marchali*, *Galloisiana yuasai*, *Harpegnathos saltator*, *Lasioglossum albipes*, *Melipona quadrifasciata*, *Mengenilla moldrzyki*, *Ooceraea biroi*, and *Solenopsis invicta*.

The GFF files served as a source to compute orthologous group synteny conservation scores by extracting the orthologous genes’ chromosomal position and producing an ordered list of genes per species. This approach does not capture genes’ proximity other than along the DNA sequence (primary structure). The ordered gene lists were then cross-compared to detect whether neighbouring genes in a given species belonged to the same orthologous groups as the neighbouring genes of another species. Local duplications of genes (i.e. multiple neighbouring genes mapping to the same orthologous group) might have incorrectly prevented the identification of syntenic blocks of genes. The synteny conservation detection algorithm was hence instructed to evaluate the first next neighbouring gene with different orthologous group membership. Synteny conservation scores were computed by assigning 0.5 points per left-side conserved synteny and 0.5 points per right-side conserved synteny at each species-pair comparison. One point was granted to the orthologous group if two species showed both-sides conserved gene synteny. Adding up the all-versus-all species synteny conservation points per orthologous group generated three separate features: average synteny (SYN) when divided by the total number of species with an annotated genome file; maximum synteny (MSY), the maximum possible cross-species synteny-conservation score; and relative synteny (RSY) when divided by the number of species arising from the orthologous group’s last common ancestor. Synteny conservation scores could not be computed for 4’319 orthologous groups, i.e. groups assigned solely to genes for which the respective gene identifier could not be traced back from the GFF files to the *OrthoDB* nomenclature. Although generically referred to as synteny, these evolutionary

features more precisely capture ortholog-level microsynteny, as only the orthologous groups from the immediate neighbouring genes were considered for the measurement of synteny conservation scores along the organisation of protein-coding genes.

Additional User-Input Features

The last evolutionary feature was defined as the average rate of protein sequence divergence normalised by the distance between each pair of species and aims to capture the orthologous group's evolutionary rate (EVR). The uniform distribution of the EVR scores correspond to previously reported accounts from the literature (Waterhouse, Zdobnov, and Kriventseva 2011), with most orthologous groups showing low to intermediate evolutionary rates, and a far-right stretch capturing fewer orthologous groups with high EVR scores. Although the EVR was extracted from *OrthoDB* data repository, it can be extracted from protein sequences' pairwise comparison scores as computed by *OrthoFinder*. Other user-provided features could be e.g. dN, dS, dN/dS (computed themselves per OG), and in principle, others too. Population-level amino-acid variation metrics, as well as multi-species whole-genome alignment conservation metrics were successfully employed in Chapter 3, when investigating the evolutionary features of *A. gambiae* immune-related genes.

Evol-Feat Bioinformatics Workflow

The computation of the evolutionary features has been automated with an open-source, scalable, and reproducible bioinformatics workflow. Contrary to previous exploratory efforts of evolutionary-functional correspondences, this newly defined framework of gene metric characterisation has been designed to be fully accessible and easily usable by biologists to interpret their data from orthology delineation of gene sets, species phylogenies, and, optionally, gene annotations. This enables the automated exploration of different eukaryotic species and lineages at different evolutionary timescales. The workflow was

implemented as R and Python scripts within a Snakemake (Köster and Rahmann 2012) framework, an increasingly popular workflow management tool. Snakemake enables the definition of controlled input to output rules, avoiding unnecessarily repeated computing, and allows for traceability and automated logging, automated management of resources with an optimised allocation of memory and computational time requirements. Reproducibility of the workflow was granted by defining dedicated Conda environments for each set of computational steps requiring dependencies in the form of external R and Python packages.

Evol-Feat is provided with a user guide specifying input requirements and formatting, definitions of computed features, config file parameters, and descriptions of the output. The minimal input requirements include an orthology delineation table and a species phylogeny, allowing for the computation of most of the evolutionary features. Orthology delineation tables can be easily downloaded from existing online sources such as *OrthoDB*, or computed with available tools such as *OrthoFinder*. The latter also provides a species phylogeny reconstruction inferred from single-copy orthologous genes. If the species of interest are included in *OrthoDB*, a species phylogeny can be further obtained with only a list of species names with the *Orthophile* workflow, described in Chapter 4.

Computation of synteny-related features requires the additional optional input of genome annotation data. Detailed definitions of the features and the data used for their computation are summarised in Table 1. *Evol-Feat* outputs include a table of the evolutionary feature scores per orthologous group, and the corresponding clustering results. Summary results of dimensionality reductions, clustering visualisations, and orthologous group-to-cluster memberships are automatically saved into tabular output files. If the user additionally provides gene ontology annotations, the workflow will run a functional enrichment sub-workflow for the evolutionarily similar clusters of orthologous groups. Details of this procedure and interpretations are discussed in Chapter 2. The workflow further allows the user to define species lists, enabling the extraction of lineage-specific evolutionary feature scores and clustering. The increased resolution of evolutionary profiles characterisation and clustering corresponding

to single species or subsets of species allows for more precise formulations of the evolutionary-functional correspondences in specific sublineages. An example is provided at the end of Chapter 2 with the extraction of *D. melanogaster*-specific evolutionary profiles. Finally, adding a table of user-defined gene sets allows a hierarchical clustering of the gene sets of interest by their average evolutionary feature scores, as shown in Figures 8 and 9.

Exploratory Assessments of the Feature Distributions

Once the evolutionary features were defined and computed, even though each captured and represented specific features by itself, it was crucial to integrate them into a framework of comparable values that could be used to assess high and low feature scores in a dimensionless system. This nondimensionalisation, or scaling, which is a necessary step for the following statistical analyses, aimed to centre and scale the distribution of each feature while also removing the features' dimensional units. With the dimensionless features, scaled and centred, an evolutionary profile (the vector of relative evolutionary features scores) was assigned to each Arthropoda orthologous group. The scaling of units was performed with the base-R function *scale*, centring the feature distributions by subtracting the means and scaling them by dividing them by their standard deviations. From the initial 82'474 *OrthoDB* orthologous groups, 4'337 contained missing values for either *CAFE5* analyses or synteny conservation scores and were removed, resulting in $N = 78'137$ distinct scaled evolutionary profiles.

The first step to understanding how the previously defined features shape the evolutionary trajectories of arthropod genes was to examine each distribution, check for pairwise comparisons, and test for cross-feature correlations. An initial overview of each feature definition's uniqueness and interdependencies confirmed already known relationships while highlighting undescribed patterns of evolutionary feature correspondences. The pairwise comparison plots, feature distributions and Pearson's correlations are presented

in Figure 1 and were produced with the *ggpairs* version 2.1.2 R package. The statistically significant correlation p-values were computed with the *cor.test* function, performing a t-distribution test on the null hypothesis, assuming no correlation exists between paired-sample distributions (Pearson's correlation coefficient $\rho = 0$). Shapiro-Wilk, Anderson–Darling, and one–sample Kolmogorov–Smirnov tests for normality, together with Bartlett's test of homogeneity of variances, rejected the normality-distribution and homoscedasticity assumptions for each of the 16 scaled features, indicating that parametric statistical testing and modelling need to be carefully considered. For this and other reasons, comparisons of median average values of the feature distributions were generally preferred to the comparisons of arithmetic means, as exemplified and further detailed in the clustering sections of Chapters 2 and 3.

Most orthologous groups appear to represent few species, with the universality distribution inversely exponential, but with an increase in orthologous group counts towards complete universality. Considering the large number of diverse species (170) and the long evolutionary timespan (600 million years), it is expected that fully universal orthologous groups make up only a small fraction of the complete set of all orthologous groups. Approximately 44% of orthologous groups are represented by two species only ($UNI \approx 0.012$), while 0.5% are universal ($UNI = 1$, not discernible from the zoomed-out representation in Figure 1). The duplicability feature highlights the tendency of most orthologous groups to be single-copy ($DUP = 0$); at the same time, smaller peaks can be found at intermediate duplicability levels, corresponding to orthologous groups which expanded in specific clades only (e.g. lineage-specific maintained duplications), and at extreme levels ($DUP = 1$), capturing fully multi-copy orthologous groups. These observations confirm similar gene universality and duplicability patterns highlighted across twelve drosophilids comparative genomics analyses in (Waterhouse, 2015), indicating that arthropod genes are also attracted to opposite poles of single/multi-copy and clade-specificity/universality. Most genes are single-copy and either widespread or sparse; few have intermediate universality scores, and virtually no multi-copy genes are intermediately spread.

The taxonomic age (AGE) feature distribution is skewed, showing different peaks across the feature's range. Specifically, the highest peak is represented by orthologous groups dating 602 million years, about 21% of all orthologous groups. These numbers seem to contradict the UNI distribution; they indicate, however, that a significant number of orthologous groups is shared across two species only, no matter their evolutionary divergence time, and that for a significant part of them, the two species are a chelicerate and a mandibulate (i.e. the first branch separation in the phylogeny). Several smaller peaks appear in the distribution, at ~550 myr and corresponding to the initial radiation of chelicerates, at ~300-420 myr and corresponding to the radiation of hexapods, at 100-180 myr and corresponding to the first radiation of mosquitoes and a last one at 10-70 myr, corresponding to the genus-specific radiations of the over-represented drosophilids and anophelines. The clear partitioning between older genes and younger genes is further highlighted by the universality (UNI) to relative universality (RUN) comparison. Figure 1 indicates that with an increase in RUN, a measure of lineage-specific taxonomic spread, the corresponding UNI distinguishes between two increasingly separated evolutionary trajectories: the first capturing fully universal genes present across the whole arthropod phylogeny, and the second capturing highly lineage-specific genes with low overall UNI scores. The orthologous groups characterised by strongly matching UNI and RUN scores indicate a robust linear relationship between the two evolutionary features, and the corresponding evolutionary trajectory likely captures essential genes which cannot be lost even when characterised by intermediate levels of taxonomic span across the whole phylogeny.

Among the highest positively-correlated features, we can find UNI and stability (STA), indicating that the more species an orthologous group is represented by, the more likely it will show stable gene copy numbers across the phylogeny, rather than expansion or loss events. The relationship is supported by a strong positive correlation (0.983, $p < 0.001$) in a linearly increasing fashion. This pattern confirms the hypothesis that the oldest and universally present genes are constrained to stable gene-copy counts, which are in turn partly associated with lower sequence evolutionary rates, largely confirming the "knockout-rate prediction". Similarly, STA also positively correlates with RUN, although with a bifurcation creating two distinct trajectories. The first is

characterised by an increasing relative universality corresponding to an increasing stability in gene copy-counts. Contrastingly, the second trajectory associates an increase in relative universality with constantly low copy-number stability scores. Genes universally present within specific lineages are likely representing essential and relatively old lineage-specific adaptations. The bifurcation between the RUN and STA features in Figure 1 precisely showcases how such genes fall into one of two clearly separated groups: either characterised by very stable gene copy counts, or by fast gene turnover. Furthermore, the trajectory representing the highly stable and relatively universal orthologous groups indicates a positive linear relationship between the two features. Such a clear-cut is distinctive to the RUN-STA comparison, while only faintly recovered by the UNI-STA comparison.

The higher taxonomic resolution provided by RUN indicates how the known relationships between universality and copy-number stability are not fully explanatory with respect to the evolution of lineage-specific gene families, for which a two-way mode of operation emerges. The first mode captures the “knockout-rate” prediction already observed at larger evolutionary timescales for fully universal orthologous groups. The second mode captures an additional evolutionary trajectory, suggesting that relatively universal genes, likely representative of essential lineage-specific adaptations, can further be characterised by fast gene turnover, represented by low scores in gene copy-number stability. Such relationships are nevertheless not clearly observed between RUN and the relative stability (RST), and high RST scores are associated with younger genes. These observations may hence indicate the recruitment of old and overall stable genes during the evolution of lineage-specific adaptations, for which different copy-numbers emerge across different arthropod subclades. The overall gene copy-number stability scores of such genes are hence inevitably affected by lineage-specific spikes of copy-number expansions and contractions. Ultimately, this characterisation may enable the distinction between old, universal and stable genes, and the old, relatively universal but unstable genes recruited during the evolution of lineage-specific essential adaptations. This interpretation is in line with the conclusions from the genomic phylostratigraphic map of *D. melanogaster* macroevolutionary adaptations in (Domazet-Lošo, Brajković, and Tautz 2007):

ancestral genes are observed to be interlocked into pathways during the evolution of lineage-specific adaptations while retaining signals of their evolutionary histories.

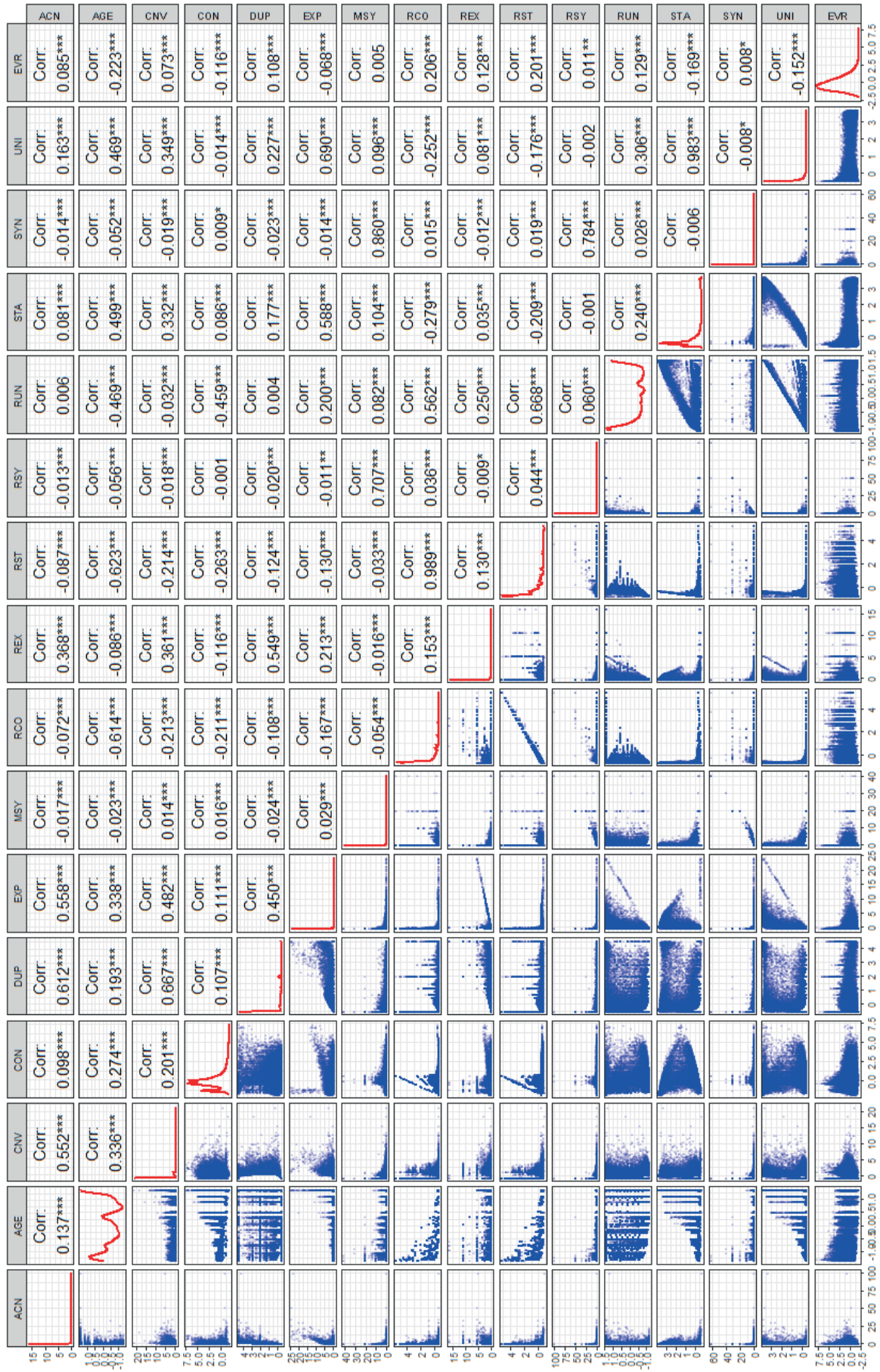


Figure 1: Distributions and pairwise comparisons of the 16 scaled and centred evolutionary features. The plots below the diagonal show each pairwise-feature comparison in blue, where the horizontal axis represents the right panel's feature scaled values and the vertical axis the top panel's feature scaled values. The diagonal shows each feature's distribution in red. The boxes above the diagonal report the pairwise-feature Pearson's correlation coefficient, ranging from -1 for total negative correlation, 0 for no correlation, and 1 for total positive correlation. The significance level of the paired-samples correlation is represented with *** if the p-value is < 0.001, ** if < 0.01, * if < 0.05, and nothing otherwise.

Interpreting the Evolutionary Features through Dimensionality Reduction Techniques

Dimensionality reduction is a crucial process when analysing high-dimensional data. It enables the first visualisation of distinctive groups of samples when projecting the coordinates in a two-dimensional plot and reduces the computational requirements of clustering algorithms by reducing the number of considered variables. Two different non-linear dimensionality reduction algorithms were initially explored: t-Distributed Stochastic Neighbour Embedding (t-SNE, R package *Rtsne*) and Uniform Manifold Approximation and Projection (UMAP, R package *uwot*). Both techniques are routinely used in bioinformatics analyses when using genomic datasets with high dimensionality. They usually clearly separate sample clusters by artificially exaggerating the distances across data points. Figure 2 presents the two-dimensional distributions of all Arthropoda evolutionary profiles (scaled vectors of evolutionary features) by both tSNE and UMAP dimensionality reduction techniques. The respective best clustering algorithm partitioned the coordinates to visualise the separation of sample groups. tSNE coordinates (Figure 2A) were assigned cluster memberships by the Ordering Points To Identify Cluster Structure (OPTICS, R package *dbscan*) algorithm. UMAP (Figure 2B) coordinates were assigned cluster memberships by the Density-Based Spatial Clustering of Applications with Noise (DBSCAN, R package *dbscan*).

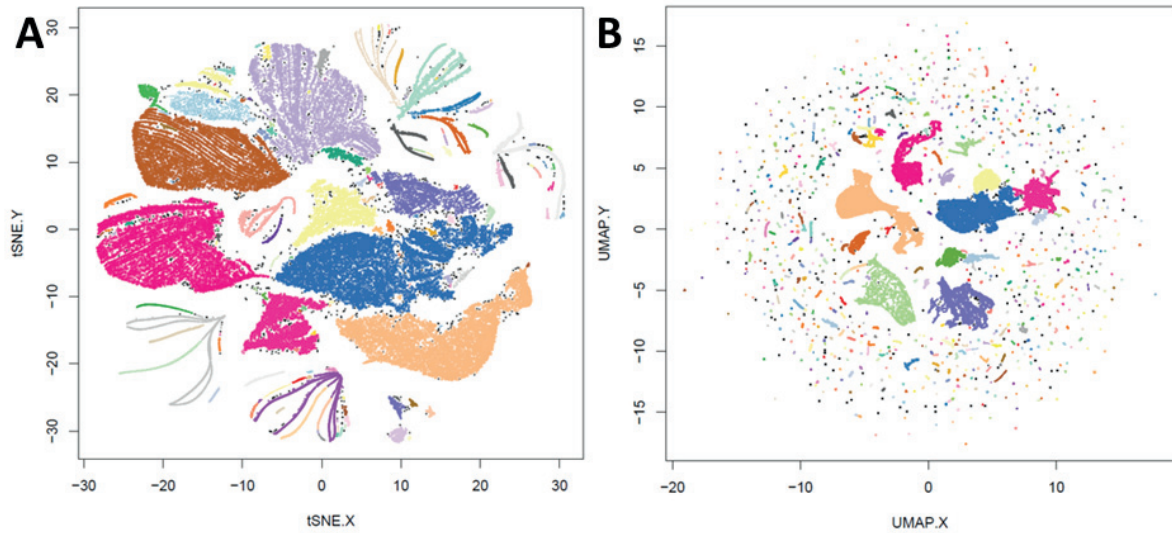


Figure 2: Distributions of Arthropoda evolutionary profiles following dimensionality reduction. The first two tSNE coordinates (Figure 2A, with the perplexity parameter set to the square root of the number of samples) were plotted and, for the sole purpose of visualisation, colour-coded by their corresponding cluster memberships computed with the OPTICS algorithm. Figure 2B shows the first two UMAP coordinates (default parameters) plotted and colour-coded by their corresponding cluster memberships computed with the DBSCAN algorithm.

The dimensionality reductions and subsequent clustering of the evolutionary profiles highlight a few large clusters of orthologous groups with similar evolutionary trajectories, constituting the core of the data, as opposed to the periphery consisting of several scattered small clusters of outlying evolutionary profiles. Nevertheless, neither tSNE nor UMAP coordinates were investigated further as the biological interpretation of the resulting non-linearly transformed space cannot be associated with the individual feature contributions. Principal Component Analysis (PCA) was therefore preferred as the default dimensionality reduction technique, where its feature contributions can unambiguously describe each coordinate axis (principal component). The arguments favouring the use of PCA coordinates also stand out when performing more complex clustering of the evolutionary profiles, as will be further discussed in Chapter 2. Specifically, using principal component coordinates automatically solves the bias of partitioning the data by several hyper-correlated variables, defined by arbitrary design choices (the individual PCA coordinates being projections of the data over orthogonal eigenvectors, hence pulling the hyper-correlated features towards a single dimension). Additionally, using PCA coordinates rather than tSNE or UMAP resulted in a more homogenous partitioning of the evolutionary profiles across the Self-Organising Map, the main clustering algorithm used in Chapter 2.

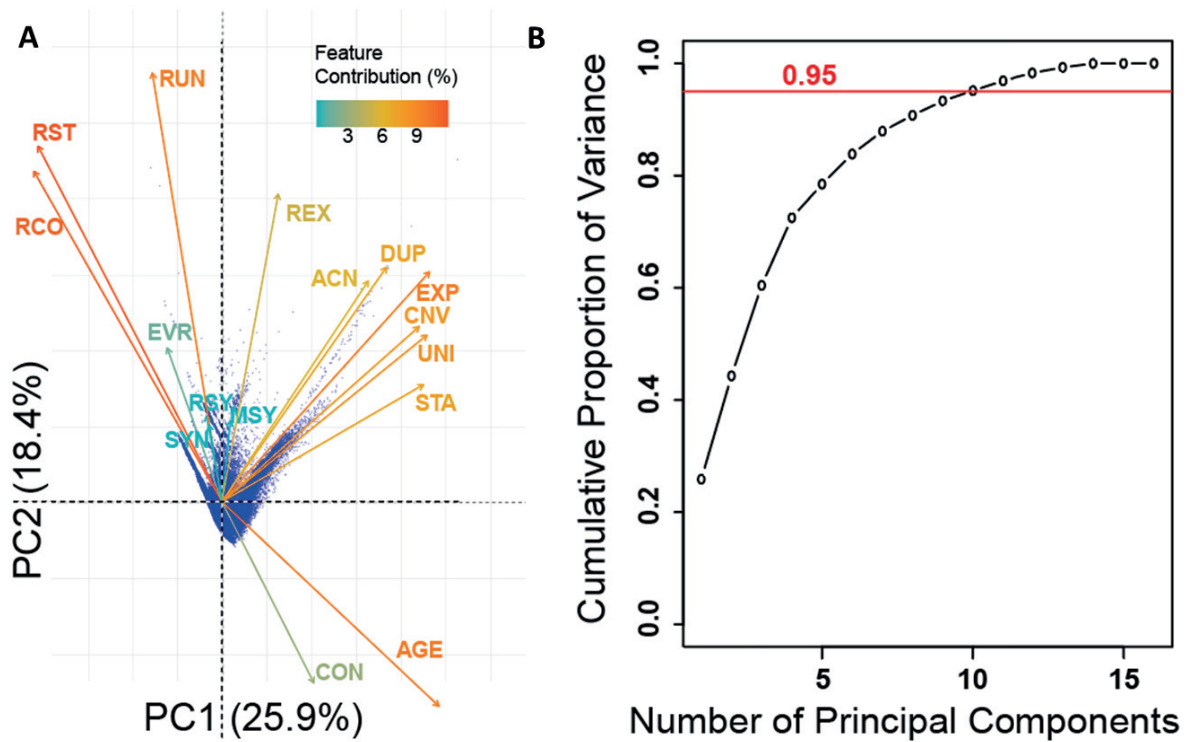


Figure 3: Principal Component Analysis. Figure 3A represents the evolutionary profiles projected onto the first two principal components space. Principal Component 1 (PC1) contributes to 25.9% of the dataset’s explained variance, PC2 to an additional 18.4%. The arrows show the directions and magnitudes (colour-coded percentages) of the specific feature contributions to each PC. Figure 3B shows the cumulative proportion of explained variance increasing with each additional principal component. The red line indicates 95% of explained variance, crossed at the inclusion of the first 10 principal components.

Principal Component Analysis was performed with the *stats* R package function *princomp*. Cumulatively, the first 10 principal components correspond to 95% of the explained variance. As highlighted in Figure 3A, the first two components show initial groupings of evolutionary features capturing different directionalities of the first two principal components. RST and RCO's similar direction and magnitudes might indicate that orthologous groups which underwent gene loss events in specific lineages of the phylogeny also tend to maintain stable copy counts where gene losses did not occur. Other features with aligned directionalities are the contributions of copy-number features (DUP, ACN, CNV) as well as universality (UNI), stability (STA) and expansions (EXP). Within this group, whose contributions are all rather closely aligned, nuances of the differences across features are lost. Still, it seems natural that expansion events are linked with copy-number features while universality is linked with stability, confirming the patterns shown in the pairwise comparisons of features. Pointing to their specific directions are the contributions of the lineage-specific features: relative expansions (REX) and relative universality (RUN), as well as taxonomic age (AGE) and gene losses (CON). The sequence evolutionary rate (EVR) and synteny-conservation features (SYN, MSY, RSY) are not well represented by PC1 and PC2; they nevertheless contribute strongly to PC6 and PC3, respectively. Table 2 reports the individual feature contributions to the first 10 principal components, whose coordinates will constitute the starting data for the unsupervised clustering and evolutionary profile partitioning of Chapter 2.

Table 2. Feature percentage contributions to the Principal Components. Each principal component is summarised by its percentage explanation of total variance, its main feature contributions in units of PC coordinates' space (reporting the ones with an absolute percentage contribution greater than 6%), and a qualitative description of the corresponding interpreted evolutionary driving forces. Strong positively contributing evolutionary features are represented in black (driven by high feature scores), and strong negatively contributing features are highlighted in red (driven by low feature scores).

Principal Component	Explained Variance (%)	Main Feature Contributions (> 6%)	Evolutionary Feature Description
1	25.87	AGE (0.367), EXP (0.351), UNI (0.347), STA (0.341), CNV (0.334), RCO (-0.318), RST (-0.311), DUP (0.279)	Age, universality, widespread and stable gene duplications opposing lineage-specific stability and gene losses.
2	18.42	RUN (0.462), RST (0.383), RCO (0.356), REX(0.332), DUP(0.253)	Lineage-specific dynamics and duplicability.
3	16.13	SYN (-0.579), MSY (-0.569), RSY (-0.540)	Synteny conservation.
4	12.07	STA (-0.445), UNI (-0.438), DUP (0.357), ACN (0.342), RUN (-0.305), REX (0.273)	Lineage-specific gene expansions opposing universality and stability.
5	5.99	CON (0.750), EVR (-0.358), RCO (0.351), RST (0.313)	Gene losses opposing sequence evolutionary rate.
6	5.36	EVR (-0.870), REX (0.331), CON (-0.267)	Lineage-specific expansions opposing sequence evolutionary rate and gene losses.
7	4.07	ACN (0.583), REX (-0.565), EXP (0.395), CON (-0.289)	Widespread expansions opposing widespread gene losses and lineage-specific expansions.
8	2.83	CNV (-0.532), AGE (-0.413), REX (0.382), CON (0.336), EXP (0.284), RCO (-0.253)	Expansions and widespread losses opposing age, copy-number variations and lineage-specific losses.

9	2.57	AGE (0.601), CNV (-0.486), REX (0.286), RUN (-0.281), RCO (0.269)	Age, lineage-specific expansions and losses opposing copy-number variations and relative universality.
10	1.84	RSY (-0.705), MSY (0.481), DUP (-0.315)	Maximum syntenic conservation opposing relative syntenic conservation and duplicability.

Concluding Remarks

The principal component analysis of 16 gene evolutionary features from 170 arthropod genomes identified four main axes of evolutionary forces driving the diversification of the evolutionary trajectories of arthropod genes. Using a different suite of evolutionary features and benefitting from a higher computational resolution, the results of this chapter confirm and expand on the conclusions from early exploratory studies on gene evolutionary features.

The first axis of arthropod gene evolution is characterised by the gene's age, universality, widespread and stable gene duplications on one end of the spectrum, and by lineage-specific stability and gene losses on the opposing end. In line with Koonin's definition of gene *status* or *importance* (Jordan et al. 2002), the first axis likely captures old, universal, highly conserved and physiologically essential genes. The second axis is characterised by the gene's lineage-specific dynamics and duplications, including gene copy-number expansion and contraction events. In line with Koonin's definition of gene *adaptability*, the second axis likely captures lineage-specific adaptations through copy-number increases and subsequent negative purifying selections.

The third axis of arthropod gene evolution is instead characterised by the absence of microsynteny. Emerging as a major contributor to the diversity of arthropod evolutionary trajectories, synteny conservation may play a role in particular metabolic pathways by affecting the functionality of synteny blocks, as in the conserved Osiris and TipE-like genes of flies (Li, Waterhouse, and Zdobnov 2011; Shah et al. 2012). Higher scores of conserved gene synteny can be associated with increased protein connectivity. Therefore, the third axis partially captures Koonin's definition of *reactivity*, also expressed in terms of low numbers of protein-protein interactions. The remaining partial definition of Koonin's definition of *reactivity*, expressed in terms of gene copy-numbers and losses, is readily recovered by the fourth axis, characterised by lineage-specific gene expansions and duplicability, together with the minor fifth axis, mostly defined by gene losses.

In line with the phylostratigraphic map of *D. melanogaster* genes (Domazet-Lošo, Brajković, and Tautz 2007), lineage-specific adaptations are confirmed to also recruit ancestral genes retaining signals of their evolutionary history and characterised by dynamic gene turnover. Widespread, universal, and old genes are instead characterised by stable copy-number counts (or slow gene turnover), and are distinctly separated from younger, lineage-specific genes. These are further associated with either high or low potential for duplicability, in line with and expanding the “single-copy control” versus “multicopy licence” descriptions from (Waterhouse, Zdobnov, and Kriventseva 2011), and likely indicators of essential, ancestral, and housekeeping functions versus more recent lineage-specific adaptations. The distribution of sequence evolutionary rates seems to follow more subtle evolutionary trajectories than previously observed in (Waterhouse, Zdobnov, and Kriventseva 2011) and in more lineage-specific or case-study analyses, as in immunity-related mosquito genes (Ruzzante et al. 2022). This observation further identifies the change in sequence evolutionary rates as a strong predictor driving the diversification of the genetic repertoire in more recent and lineage-specific adaptations, while losing its explanatory power when considering larger evolutionary timescales. At the broader Arthropoda scale, fast and slow gene evolution is found to be less tightly associated with age and universality, and more with the distribution of gene losses and family expansions.

The confirmation of several hypotheses on evolutionary feature distributions and relationships from previous studies, coupled with the uncovering of novel observations and trajectories, supports the translatability of evolutionary hypotheses based on few eukaryotic species to large-scale arthropod genome studies. Understanding the patterns of eukaryotic gene evolution requires fine-tuning of the definitions and range of evolutionary features. In parallel with a higher taxonomic resolution, this work will provide a fundamental theoretical basis for future arthropod comparative genomics research studies.

Chapter 2: Characterisation of Evolutionary-Functional Correspondences

Summary

This chapter first provides a theoretical background for partitioning genes into evolutionary modules and describes the reasoning behind the automated generation of gene function from high-throughput genomic data. This introductory section is followed by an overview of the methodology enabling functional inference and an initial exploration and verification step of the functional associations of extreme values of the evolutionary feature distributions. The focus then switches to cluster analysis, detailing the techniques that resulted in the most meaningful and optimised partitioning of arthropod gene evolutionary features. After describing the main types of evolutionary trajectories adopted by clusters of arthropod genes, the discussion then highlights the emerging functional properties of such evolutionary modules assigned through gene ontology enrichment analysis. The chapter concludes by showcasing an additional analysis framework, leveraging the full arthropod-scale evolutionary-functional correspondences to generate more detailed and lineage-specific hypotheses and concluding remarks.

Theoretical Background on Evolutionary Profile Clustering and Functional Inference

In Chapter 1, 82'474 orthologous groups delineated from over 2.2 million arthropod genes were assigned individual evolutionary profiles defined by their phylogenetic age, taxonomic span, synteny conservation, sequence evolutionary rate, copy-number dynamics and turnover, for a total of 16 distinct evolutionary features. By extending the concept of inferring gene function by bringing together genes through sequence homology and orthology delineation, Chapter 2 leverages the increased evolutionary resolution to cluster arthropod genes into sets of similar evolutionary trajectories and assign them to functional modules and putative biological roles. Previous attempts to partition the evolutionary histories of genes include a phylogenetic inference-based approach optimising the selection of clustering methods for sequence alignments (Gori et al. 2016), and the annotations of human biological pathways to evolutionary modules inferred from species trees and sequence homology matrices (Li et al. 2014). The most comprehensive high dimensionality description and visualisation of complex evolutionary histories, enriched with detailed molecular functions, was deployed by EvoluCode (Linard et al. 2012). The study partitioned the human proteome into evolutionary profiles defined by metrics of sequence identity and protein domains, among others. Inspired by these attempts, this chapter showcases the exploration of different clustering techniques and challenges to obtain an unprecedented higher-resolution evolutionary-functional map of biologically meaningful clusters of arthropod genes.

Cluster analysis is the approach of grouping samples with similar quantifiable properties using statistical methods, creating labels for groups of samples that are more similar to each than to the rest of the samples. Cluster analysis - or clustering - falls into the category of unsupervised machine learning algorithms, intending to assign samples to clusters without a priori knowledge of the cluster sizes nor assisted by a validation learning process to guide the cluster membership assignments. Clustering techniques can vary greatly; therefore, care and consideration need to be put into a cluster analysis as results will likely be variable depending on the selected algorithms, parameters, and thresholds.

As there is no universally recognised best technique of clustering for every type and size of data, clustering techniques may be chosen simply because a specific approach has been successfully deployed in the past with similar types of data or by carefully evaluating the algorithm's statistical convergence and optimal trade-offs. Examples include the elbow method and the silhouette analysis, used after k-means clustering to determine the optimal number of clusters and maximise the explained variation (model's fit) to the number of clusters ratio, thus limiting the possibilities of overfitting the model to the data. Ultimately, the goal of cluster analysis applied to the evolutionary features was to partition arthropod orthologous groups to verify whether (and if so, which) functional biological properties are associated with evolutionarily similar clusters of orthologous genes, i.e. sharing similar cross-species evolutionary trajectories. In the case of emerging evolution-to-function patterns, the resulting evolutionary modules could serve as models for theoretical broad-scale (considering all genes from all species) or fine-scale (a set of genes from one or few closely related species) evolution-to-function predictions.

Building predictive automated functional annotations of sets of genes is required when working at scale, where comparative genomics can provide solutions when large-scale experimental elucidation efforts are not conceivable. Following a similar approach to EvoluCode (Linard et al. 2012), the *Evol-Feat* workflow assigns statistically supported putative gene functions to evolutionarily-defined modules using Gene Ontology enrichment analyses. The Gene Ontology (GO) database is the most comprehensive resource detailing gene function (Ashburner et al. 2000; Gene Ontology Consortium 2019), allowing for a hierarchical characterisation of a gene's biological processes, cellular compartmentalisation, and molecular functions. It provides a structured representation of the current scientific knowledge that allows contrasts to determine which biological processes, functions, and cellular locations are significantly over- or under-represented in groups of genes from high-throughput studies (Yon Rhee et al. 2008; Robertson et al. 2018; Sun et al. 2021). Inherent biases from the GO are likely to be present and will inevitably affect the quality of the functional annotations. With respect particularly to arthropod genes, these may include over-representations of model organism species, particularly *Drosophila melanogaster*, and research topics including immunity, development

or insecticide resistance experiments. Related enriched GO terms will most likely emerge from the scientific descriptions of biological mechanisms most investigated in arthropod biology research.

Enrichment Analysis for Gene Ontology is the process of quantifying and validating with statistical testing the overrepresentation of annotated functions (GO terms) for a given list of genes (foreground) compared to the annotations of all other genes (background). Commonly used statistical tests include Fisher's exact and Kolmogorov-Smirnov (KS) tests, providing comparable p-values for statistical support and hypothesis testing for each GO term. Randomly generated lists of genes statistically tested against the full spectrum of the GO database are unlikely to generate meaningful and well-supported functional enrichments. Accordingly, one of the goals of this chapter is to determine whether lists of genes that share similar evolutionary profiles can generate statistically significant GO term enrichment results. Functional annotations emerging from sets of orthologous groups partitioned by evolutionary features will indicate whether gene functional constraints can determine the course of gene family evolution across the arthropod phylogeny, allowing for the exploration and evolutionarily informed predictions of biological processes of uncharacterised genes.

If genes with similar or analogous functional roles are constrained to follow similar evolutionary trajectories, then clustering genes using their evolutionary profiles should bring together genes with similar or analogous functional roles. Clusters showing statistically supported enrichments for certain biological functions (using the GO as the means of annotating functions) likely indicate that modules of evolutionary trajectories effectively bring together genes with similar or analogous functions. Such observations also confirm that a gene's functional role defines and constrains the evolutionary path it can follow. The principal expectation driving the work presented in this chapter is that starting from a higher-dimensional description of arthropod evolutionary trajectories, it is possible to infer an equivalent higher resolution partitioning and description of the associated functional roles and constraints.

Orthologous Group Gene Ontology Annotations

Function prediction of evolutionary-defined modules was performed with enrichment analysis for gene ontologies on clusters of orthologous groups with consensus GO term annotations obtained from deep learning-, sequence similarity- and protein domain-based approaches. In order to obtain a consistent and comprehensive catalogue of gene functional annotations, GO terms were assigned to the 2.2 million arthropod protein-coding genes using CrowdGO (Reijnders and Waterhouse 2021b). CrowdGO provides machine learning and semantic similarity-guided consensus GO annotations from the results of four sequence-based functional annotation tools: DeepGOPlus (Kulmanov and Hoehndorf 2020), Wei2GO (Reijnders 2022); InterProScan (Jones et al. 2014); and FunFam (Scheibenreif et al. 2019). In order to assign function predictions to genes, DeepGOPlus uses a deep learning model to detect protein motifs (short conserved sequence patterns associated with distinct functions); Wei2GO uses sequence homology; while InterProScan and FunFams use domain homology (groups of folded three-dimensional protein structures with distinct molecular functions). CrowdGO then re-evaluates each gene-term annotation, and a consensus dataset is produced with high-scoring confident annotations and low-scoring rejected annotations. Only the Biological Process (BP) GO terms with CrowdGO's default 0.5 cut-off were retained for downstream analyses. Orthologous group GO terms were thus assigned by merging all of the GO terms from each gene in the orthologous group and removing the terms represented by a single gene only. This filtering step was applied in order to exclude possibly erroneous or highly specific GO terms, aiming to avoid single-gene-based GO term predictions propagating to the whole group.

Functional Enrichments of Individual Evolutionary Features

Before exploring the distributions and functional annotations of the evolutionary profiles, the extreme values of the evolutionary features were investigated with

enrichment analysis for gene ontologies. This first exploratory step and sanity check informs and tests some of the hypotheses regarding the correspondences between evolutionary features and gene function introduced in Chapter 1 and the current literature (Krylov 2003; Waterhouse, Zdobnov, and Kriventseva 2011), for example, the functional relationships between gene stability and duplicability, universality and essentiality. Additionally, this approach validates whether similar functional properties characterise correlated features. The enrichment analyses were performed on the top and bottom percentiles of each of the 16 evolutionary feature scores with the R package topGO (Alexa and Rahnenfuhrer 2020). The distributions of orthologous group evolutionary feature scores were assessed for enriched GO terms using the Kolmogorov-Smirnov test with the *weight01* algorithm and a *nodeSize* parameter of 10, pruning the GO hierarchy from the terms which have fewer than 10 annotated orthologous groups. The resulting statistically significant enrichments were then processed with GO-Figure! (Reijnders and Waterhouse 2021a), a software to summarise and visualise GO enrichment analysis results, which can be repetitive/redundant given the hierarchical structure of the GO graph. For simplicity, Table 3 reports only the statistically significantly enriched GO term descriptors for the extreme values (top and bottom) of each evolutionary feature.

Examples of GO-Figure! processed enrichment analyses for ontologies are provided in Figure 4 for the top orthologous group expansions (EXP, left), and relative expansions (REX, right) feature scores. The sanity check here is to ask whether the enrichment analysis has identified biological processes associated with gene families known to show frequent expansions. Among the wide range of GO term descriptors, known lineage-specific and universal expanded biological processes are nevertheless recovered, including digestive system development and metabolic processing of organic substances (gene family expansions likely linked to dietary adaptations); signalling, phagocytosis, and immune response to bacteria (gene family expansions likely linked to immune-related pathways). Other more generic and essential biological processes might have been captured from ancestral and conserved expansion of genes for a variety of biological purposes, ranging from the emergence of novel signalling pathways and enzymatic functions to gene dosage increases linked to the regulation of

transcription and generic cellular processes, as well as organelle organisation and cell-type differentiation.

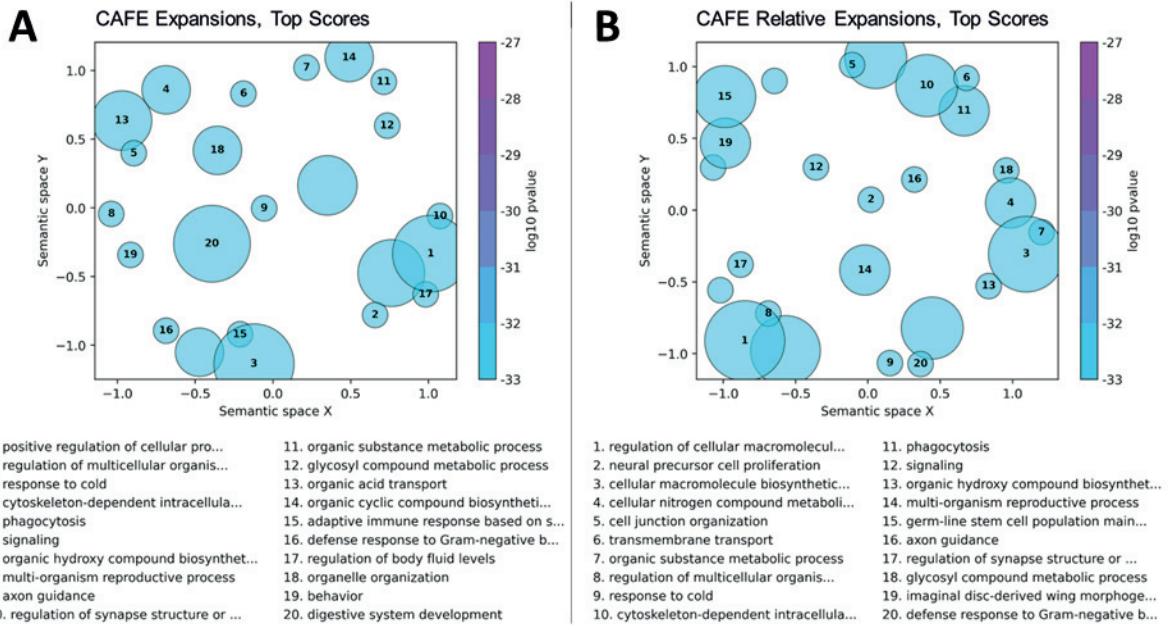


Figure 4: GO-Figure! representation of significantly enriched GO terms from the top expanded (EXP, Panel A) and relatively expanded (lineage-span scaled, REX, Panel B) orthologous groups. GO terms are grouped by their semantic similarity as computed in GO-Figure! and scaled (circle size) by their number of represented terms. The log₁₀ p-value scale indicates that all p-values are < 10⁻³³, due to the few but diverse included groups from the skewed distributions, compared to the much larger and uniform background.

Enrichment analysis results of large lists of generic but diverse GO terms are difficult to interpret. For this reason, Table 3 mainly serves as a checking point, overall validating the hypotheses relating expected functional constraints and characteristics to evolutionary trajectories. Patterns confirming expected associations can nevertheless be found and interpreted. Conservative copy-number variations are linked to essential and generic genetic transcription, anatomical development and regulation of metabolic processes. The youngest orthologous groups are enriched with immune-related functions, including cytolysis and disruption of cells from foreign organisms, indicating the dynamic and neo-functionalisation characteristic of parts of the immune response, as it may result from molecular arms races against pathogens. Orthologous groups characterised by the least amount of gene family losses (CON) are associated with old essential and conserved cellular mechanisms such as phagocytosis, the highly conserved TOR and Notch signalling pathways, mitosis and DNA damage control processes, and mitochondrial activity.

In contrast, groups with the most gene copy losses are enriched with more specific compound metabolism, sensory perception processes, and immune response regulation. It seems likely that processes associated with large gene family expansions are also expected to be lost along the arthropod phylogeny. Similarly, enriched bottom percentiles of features associated with highly conserved and essential mechanisms are represented by low scores of duplicability (DUP), evolutionary rates (EVR) and expansions (EXP). These regulate metabolic activity and transcription, anatomical and post-embryonic development, and cell structure organisation. Although linked to several generic biological processes, orthologous groups with the highest evolutionary rates (EVR) are also functionally enriched with all the expected expanded and ecologically diversified processes, including sensory perception, receptor signalling, chemosensation, the immune response through pathogen interaction and disruption. These biological processes are shared with the top scoring values of relative contractions (RCO) and low values of stability (STA, specifically chemosensation, notoriously characterised by large gene family expansions) and universality (UNI), likely representing orthologous groups capturing highly dynamic gene families, lineage-specific and with high gene turnover rates. More generic GO term enrichments are found across top scores of EVR and low

synteny conservation (SYN, MSY, RSY), reflecting the heavy skewness of their respective distributions.

Table 3: Most distinctive GO enrichment analyses of evolutionary features' extreme values.

Evolutionary Feature	Direction	Descriptors of Significantly Enriched GO Terms (p-value < 0.05)
ACN & CNV	<i>bottom</i>	positive regulation of metabolic process; anatomical structure development; RNA splicing, via transesterification reactions; post-embryonic development; nucleoside phosphate metabolic process
AGE	<i>bottom</i>	killing of cells of other organism; disruption of cells of other organism; cytolysis
CON	<i>bottom</i>	phagocytosis; positive regulation of cellular process; cytosolic transport; protein ubiquitination; small molecule metabolic process; neurogenesis; mitochondrion organization; oogenesis; ubiquitin-dependent protein catabolic process; protein targeting; transmembrane receptor protein serine/threonine kinase binding; snRNA 3'-end processing; positive regulation of Notch signaling; ribosome biogenesis; sensory perception of pain; nuclear division; mitotic G2 DNA damage checkpoint; mitochondrial translation; lateral inhibition; synapse organization; vacuolar transport; TOR signaling; vesicle-mediated transport to the plasma membrane; DNA repair; dorsal closure; carbohydrate metabolic process; neural precursor cell proliferation
	<i>top</i>	glycosyl compound metabolic process; nuclear division; ammonium ion metabolic process; behavior; regulation of transmembrane receptor protein serine/threonine kinase signaling pathway; detection of stimulus involved in sensory perception; regulation of synapse structure or activity; epithelium migration; negative regulation of cytokine process; leukocyte migration
DUP	<i>bottom</i>	cellular macromolecule metabolic process; positive regulation of metabolic process; positive regulation of nucleic acid-templated transcription; anatomical structure development; RNA splicing, via transesterification; regulation of cellular metabolic process; cell differentiation; post-embryonic development; nucleoside phosphate metabolic process
EVR	<i>bottom</i>	regulation of cellular localization; protein-containing complex subunit organization; behavior; microtubule-based movement; epithelium migration; cell junction organization
	<i>top</i>	sensory perception; cilium movement; cytoskeleton-dependent intracellular transport; mitotic nuclear envelope disassembly; negative regulation of endopeptidase; cell projection morphogenesis; telomere maintenance; nucleic acid metabolic process; microtubule cytoskeleton organization; killing of cells of other organism; disruption of cells of other organism; protein-DNA complex assembly; detection of stimulus involved in sensory perception; ionotropic glutamate receptor signaling; cytolysis; negative regulation of cytokine production; chemosensory behavior; double-strand break repair; male meiotic nuclear division; innate immune response; transcription by RNA polymerase III; regulation of mitotic metaphase/anaphase
EXP & REX	<i>bottom</i>	positive regulation of metabolic process; anatomical structure development; post-embryonic development; nucleoside phosphate metabolic process
MSY	<i>bottom</i>	post-embryonic development; anatomical structure development; peptidoglycan metabolic process; potassium ion transport; behavior; cell wall macromolecule metabolic process; cellular calcium ion homeostasis; signaling; cell junction organization; nucleoside phosphate metabolic process; regulation of cellular response to growth factor stimulus
RCO	<i>top</i>	killing of cells of other organism; disruption of cells of other organism; detection of stimulus involved in sensory perception; cytolysis

RST	<i>top</i>	regulation of cellular macromolecule biosynthetic process; nuclear division; positive regulation of developmental process; multicellular organism; aging; determination of adult lifespan; mitochondrion organization; positive regulation of transcription; cytosolic transport; RNA splicing, via transesterification; lateral inhibition; regulation of synapse structure or activity; anterior/posterior axis specification; Golgi organization; dorsal closure; carbohydrate metabolic process; ubiquitin-dependent protein catabolic process
RSY	<i>bottom</i>	organic substance metabolic process; post-embryonic development; anatomical structure development; regulation of cellular macromolecule biosynthetic process; regulation of macromolecule metabolic process; potassium ion transport; peptidoglycan metabolic process; behavior; cell wall macromolecule metabolic process; cellular calcium ion homeostasis; cell junction organization; signaling; regulation of transmembrane receptor protein serine/threonine kinase signaling pathway; nucleoside phosphate metabolic process; microtubule-based movement; response to stimulus
RUN	<i>bottom</i>	reproduction; epithelium development; post-embryonic development; nuclear division; protein ubiquitination; neurogenesis; endocytosis; cellular calcium ion homeostasis; cytoskeleton-dependent intracellular transport; cell junction organization; RNA splicing, via transesterification; regulation of transmembrane receptor protein serine/threonine kinase signaling pathway; glycosyl compound metabolic process; potassium ion transport
	<i>top</i>	protein ubiquitination; regulation of catabolic process; RNA splicing, via transesterification; imaginal disc-derived wing morphogenesis; Golgi organization; carbohydrate metabolic process; germ-line stem cell population maintenance
STA	<i>bottom</i>	chemosensory behavior
SYN	<i>bottom</i>	post-embryonic development; anatomical structure development; peptidoglycan metabolic process; potassium ion transport; behavior; cell wall macromolecule metabolic process; cellular calcium ion homeostasis; signaling; cell junction organization; nucleoside phosphate metabolic process; regulation of cellular response to growth factor stimulus
UNI	<i>bottom</i>	microtubule-based movement; killing of cells of other organism; disruption of cells of other organism

Filtering and practical considerations: Given the skewed distributions of some of the evolutionary features (see Figure 1), some enrichment analyses resulted in extensive lists of statistically significantly enriched GO terms. When most orthologous groups are pulled to one side, the few but diverse orthologous groups on the opposite side percentile produce strongly statistically significant enriched terms as they are compared to the much larger noisy background of the GO universe (the list of all orthologous groups' GO terms). Percentile distributions of features which resulted in too large lists with extreme statistically significant GO term enrichments (p-value < 1×10^{-30} for 50 terms or more) are thus not shown in Table 3 but still considered for further interpretation. Such GO terms likely obtain high statistical support thanks to their high-frequency annotations in gene lists, are very generic and do not necessarily contribute to the biological interpretation of the functional annotations. They included descriptors such as: "cellular process", "metabolic process", "nucleobase-containing compound metabolic process", "cellular metabolic process", "protein metabolic process", and "transcription, DNA-templated". Functional enrichments of bottom percentiles of Average Copy-Number and Copy-Number Variation, as well as *CAFE5* Expansions and Relative *CAFE* Expansions, resulted in the same enriched terms and were thus combined.

An Overview of Orthologous Group Clustering

A number of clustering techniques were tested to analyse the arthropod evolutionary features and allow for a diversity of interpretations and alternatives to a single clustering method. This provides opportunities to compare results based on different methodologies while simultaneously serving as an example of evolutionary feature data-tailored clustering strategies that should serve as a baseline and eventually be expanded in future work. Exploration of initial clustering results from some of the most common unsupervised clustering algorithms currently used in various applications, such as pattern recognition or genomic data grouping (Borkowska et al. 2014; Manduchi et al. 2021), enabled an informed selection of the most suitable techniques. Distribution-based algorithms were discarded, given the skewness and unknown distributions of the evolutionary features, as shown in Chapter 1, hampering the meaningfulness of

distribution-specific properties such as medians and standard deviations. Hierarchical-, centroid- and density-based clustering techniques were thus preferred, including Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Hierarchical-DBSCAN (HDBSCAN), Ordering Points To Identify the Clustering Structure (OPTICS), k-means, hierarchical clustering with a cut-off and Self-Organising Maps (SOM). The resulting clusters had to be comparatively similar in size, as subsequent functional annotation of lists with tens of thousands of orthologous groups would produce vague and non-significant function predictions, de facto removing most of the data from meaningful assessments, as exemplified by the imbalanced distributions of evolutionary feature scores shown in Figure 1.

For this reason, the clustering techniques were deemed valid when producing clusters of evolutionary profiles with non-extreme differences in cluster sizes. K-means, OPTICS and DBSCAN were discarded as they produced a handful of clusters regrouping most of the orthologous groups in central clusters and a constellation of small-sized peripheral clusters. Additionally, the k-means algorithm requires a user-specified number of resulting clusters, which can be estimated with several techniques, such as the elbow method or the silhouette analysis. However, the optimal number of clusters was always deemed too low to be helpful for this case. While statistically solid and correctly representing the evolutionary profiles' unbalanced distribution, it could not serve the purpose of discriminating a number of clusters to be manually interpreted and evaluated. The trade-off lies between clustering accuracy and the number of obtained clusters, and no balance could be found. HDBSCAN, a hierarchical decision tree approach to DBSCAN, was discarded as too many orthologous groups were considered "unclustered", i.e. they could not statistically be grouped to any final DBSCAN cluster. Hierarchical clustering of the ~80'000 evolutionary profiles showed to be too computationally demanding (with a time complexity of $\mathcal{O}(n^3)$ requiring $\Omega(n^2)$ memory), with the added difficulty of cut-off parameter selection at which cutting the dendrogram for cluster membership designation. Finally, the Self-Organising Map was identified as the most appropriate clustering and visualisation algorithm, homogeneously distributing orthologous groups across a size-predetermined grid of multi-dimensional predicted ranges of evolutionary profiles.

Self-Organising Map: Clusters of Evolutionary Profiles

Traditional clustering methods often perform poorly on high-dimensional datasets, given the unoptimised similarity measures and underlying complexity. Dimensionality reduction techniques are successfully employed to increase the clustering efficacy through low-dimensional data representations. Nevertheless, combining standard dimensionality reduction techniques with traditional clustering methods does not yield results that can be easily represented in human-interpretable dimensional spaces, and more sophisticated approaches are required (Manduchi et al. 2021). Contrastingly, the self-organising map (SOM), or Kohonen network (Kohonen 1982; Wehrens and Kruisselbrink 2018), is a computational method for clustering and visualisation of high-dimensional data which provides interpretable and low-dimensional representations of complex input data. High-dimensionality datasets are ultimately projected in a two-dimensional grid of nodes where spatially closer clusters are enriched with similar input feature values. The resulting grid thus conserves the underlying structure of the input data space.

The SOM algorithm builds an unsupervised learning model and is best suited for cases when it is crucial to maintain the topology between input and output spaces. Preserving the topology was an essential aspect of clustering the evolutionary profiles, given the different shapes and skewness of the feature distributions. The algorithm works by initialising a list of predefined $i * j$ nodes positioned in a grid where all possible values of the model parameters are inferred from the range of the input feature values. Each sample is assigned to a node with a competitive learning training process, randomly initialising samples to node associations and updating them through successive retraining iterations to obtain the best matching unit via euclidean distance measurements. The node weights are updated to match the input vectors, and the final models, corresponding to the grid clusters, are called codebook vectors and are represented by hexagonal cells in the grid visualisation. The resulting codebook vectors correspond to gene models defined by hypothetical evolutionary profile scores built from the full range of the input evolutionary feature scores. These evolutionary “archetypes” are conceptually equivalent to the eigengenes of gene

expression profiles from gene co-expression network studies (Langfelder and Horvath 2007). The hexagonal shape of the cells better represents cell proximities to other cells when compared to squared shapes.

SOM clustering was performed on the full range of arthropod evolutionary profiles. The top 10 principal components' scores were preferred to the evolutionary feature scores to avoid clustering biases guided by unbalanced groups of strongly correlated features. Using the first $N * M$ matrix ($N = 78'474$ orthologous groups; $M = 10$ principal components) as input data, the SOM was built with the *supersom* function from the R package *kohonen* version 3.0 (Wehrens and Kruisselbrink 2018). The number of learning process iterations was set at $r_{len} = 500$, and the learning rates were set at the default *alpha* values, linearly decreasing from 0.05 to 0.01. The number of learning process iterations matches the times the complete dataset is presented to the network. The training process could be halted when improvements in mean euclidean distances from the dataset to the network nodes did not substantially change with additional iterations, while maintaining a homogeneous distribution of the orthologous groups across SOM cell coordinates (i, j) . Examples of such assessment procedures are shown in Figure 5.

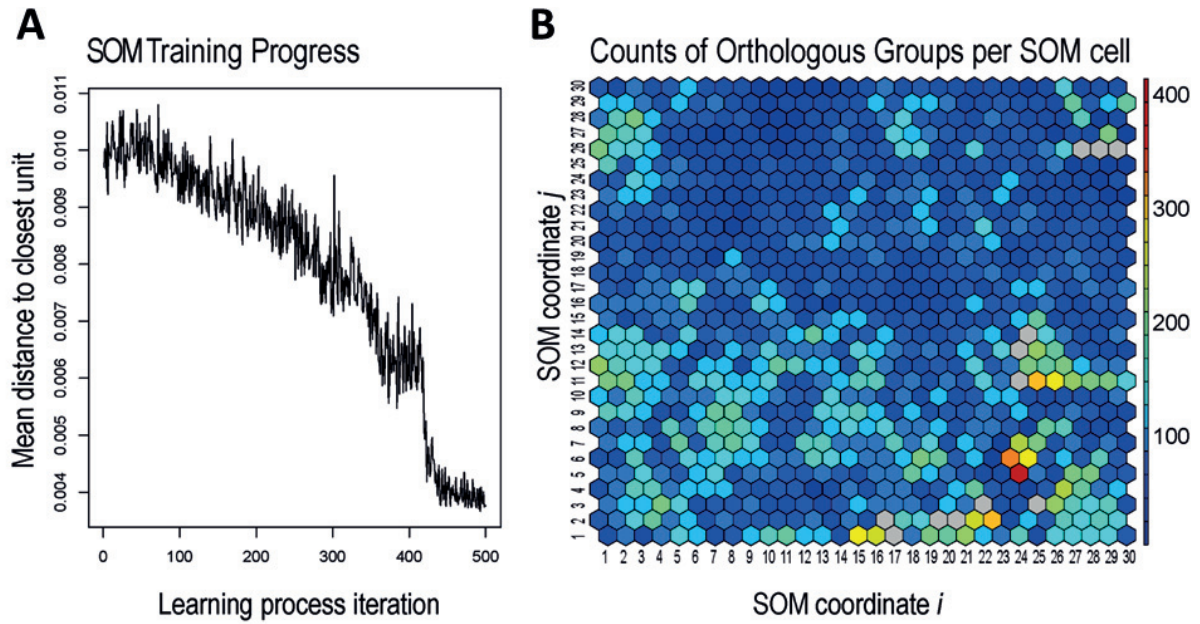


Figure 5: Examples of SOM training process and orthologous groups partitioning. In Panel A, the average of the euclidean distances from input vectors to the corresponding closest network units is computed after each learning process iteration. The distances are minimised when approaching 500 iterations, with a mean value of ~ 0.004 . In Panel B, colour-coded counts of orthologous groups mapped to SOM nodes, with a mean of 87.5 groups (medium blue) per cell, a median of 78, a maximum of 414 (dark red), a minimum of 4 (dark blue), and a standard deviation of 47.4. Cells without orthologous groups assigned to them are coloured in grey.

The dimensions of the SOM grid were set at 30 * 30 cells for a total of 900 distinct SOM cells, or cluster archetypes, each corresponding to a particular eigengene (or codebook vector), and assigned with matching orthologous groups. The grid dimensions were chosen to partition between 80 to 100 orthologous groups per cell, a range corresponding to the maximum size of orthologous group clusters that could capture meaningful functional annotations. The topology of the SOM brings together cells with similar evolutionary profiles while opposing sides of the grid correspond to the most dissimilar ones. It is important to note that each randomly initialised implementation of the SOM can vary greatly depending on the grid dimensions, expected cluster sizes and homogeneity, input features and learning rates and iterations. The implementations presented in this thesis are snapshots of possible ways to represent the full arthropod evolutionary features data. However, new exploratory analyses (either on lineage-specific subsets of this data or new data) must be carefully defined to reflect the input data dimensions and variation.

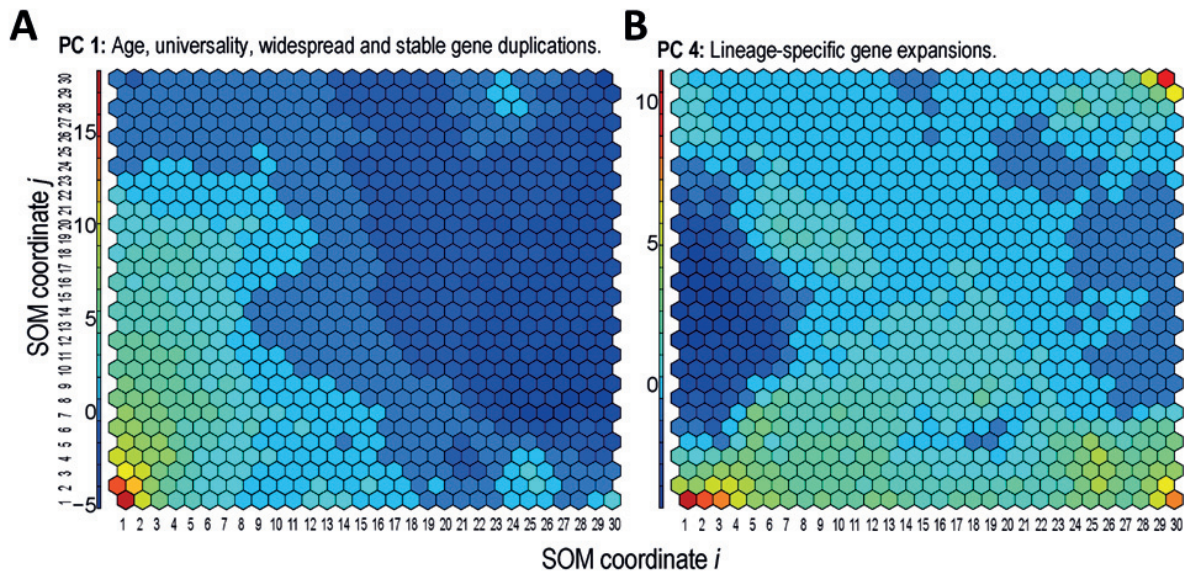


Figure 6: Examples of SOM codebook vectors highlighting specific feature weights of the model. In Panel A, eigengenes, or SOM codebook vectors, are coloured by their corresponding values of PC1 (high dark red, low dark blue). PC1 is associated with orthologous groups' universality, stability and widespread gene duplications. In Panel B, codebook vectors are coloured by their values of PC4, associated with lineage-specific gene expansions.

Eigengenes (SOM cell models, or codebook vectors) highlighted by their feature weights can be visualised for each input feature. In Figure 6 are represented examples corresponding to PC1 and PC4, fully described in Table 2: the first associated with widespread stable, universal and duplicated orthologous groups; the second associated with orthologous groups characterised by lineage-specific expansions. These highlights suggest intricate distributions of gene evolutionary features, not necessarily following one-directional gradients. The bottom left corner of PC1 in Figure 6A captures patterns of old taxonomic age coupled with high taxonomic spans and widespread stable gene duplications, including lineage-specific expansions. On the other hand, the opposing corner of the map captures orthologous groups with similar lineage-specific expansion scores but characterised by young age, low universality, and unstable copy-numbers. Enhancing individual feature contributions across the SOM highlights the complexity of capturing clusters of high-dimensional evolutionary trajectories: the presence of contrasting peaks and valleys confirms the necessity of describing novel evolutionary features. Rather than focusing on one-dimensional feature distributions, such complex multi-dimensional analyses allow for the uncovering and characterising of undescribed landscapes of genes' evolutionary features and histories.

Self-Organising Map: Superclusters of Evolutionary Profiles

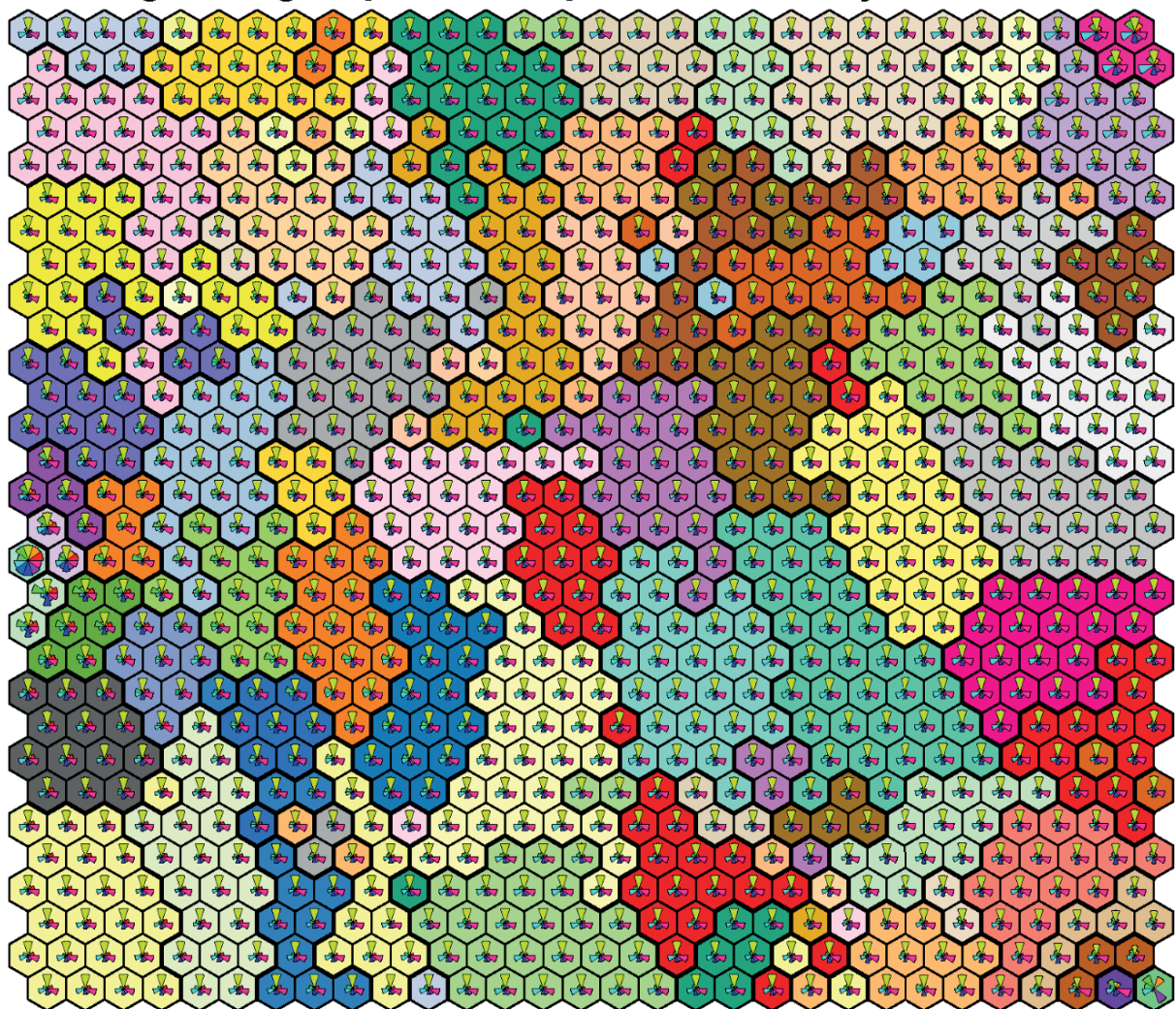
The dimensions of the arthropod SOM were tailored to optimise the distribution of orthologous groups across the grid and allow specific and meaningful functional annotations. The resulting extensive dataset required summary interpretations and visualisation techniques, not for specific predictions of evolutionary modules but for obtaining a more comprehensive interpretative framework for relating patterns of evolution to functional constraints. A more exhaustive exploration of the varieties of evolutionary trajectories and their associations with correspondingly broader functional categories required additional clustering methods on top of the SOM. The 900 codebook vectors

were further clustered using the k-means algorithm, with the *kmeans* function from the R *stats* package.

The k-means algorithm partitions the input data into the user-defined size of k clusters in which each observation is assigned to the cluster with the minimised euclidean distance to the iteratively updated cluster centre. The choice of the algorithm reflected the necessity of preemptively selecting a reasonable amount of k-means superclusters. The SOM cells were assigned to a total of 60 k-means superclusters, twice the lateral size of the SOM, with an expected average distribution of 15 cells per supercluster. Figure 7 displays the full arthropod SOM highlighting each cell's codebook vector feature contributions with pie charts and the colour-coded background representing the k-means supercluster memberships. The k-means superclusters are not always contiguous across the SOM topology, given the different nature of the two clustering algorithms when solving multi-dimensional proximity in a two-dimensional grid. Nevertheless, the superclusters show overall high levels of spatial consistency, further confirmed by the hierarchical clustering discussed in the next section.

As evolutionary profiles are distributed across the grid, islands of the most distinctive ones can be recognised at the SOM edges. On the top right corner are represented young, lineage-specific, slow-evolving orthologous groups with lineage-specific expansions. Instead, the bottom right corner is represented by groups with old, slow-evolving gene family expansions with maintained copy-numbers, high synteny conservation scores and occasional gene losses. In the immediate left are recovered orthologous groups with similarly high average synteny conservation scores but lacking the lineage-specific expansions and, interestingly, low lineage-specific synteny conservation. Towards the SOM's mid-left edge, slow-evolving orthologous groups are strongly characterised by fast-evolving lineage-specific and unstable expansions. Further confirming the observations from Table 2, describing the principal components, high sequence evolutionary rates seem to contribute to clustering driving forces only when coupled with counts of gene losses and expansions.

Self-Organising Map of Arthropod Evolutionary Profiles



Feature Contributions of SOM cells

- ▲ PC1: old, universal, stable expansions
- ▲ PC2: lineage-specific dynamics
- ▲ PC3: low synteny conservation
- ▲ PC4: lineage-specific unstable expansions
- ▲ PC5: gene losses with low evolutionary rates
- ▲ PC6: slow-evolving lineage-specific expansions
- ▲ PC7: widespread expansions
- ▲ PC8: young expansions with widespread losses
- ▲ PC9: old, lineage-specific expansions and losses with stable gene copy-numbers
- ▲ PC10: low relative synteny conservation and low gene duplicability

K-means Superclusters

■ 1	■ 11	■ 21	■ 31	■ 41	■ 51
■ 2	■ 12	■ 22	■ 32	■ 42	■ 52
■ 3	■ 13	■ 23	■ 33	■ 43	■ 53
■ 4	■ 14	■ 24	■ 34	■ 44	■ 54
■ 5	■ 15	■ 25	■ 35	■ 45	■ 55
■ 6	■ 16	■ 26	■ 36	■ 46	■ 56
■ 7	■ 17	■ 27	■ 37	■ 47	■ 57
■ 8	■ 18	■ 28	■ 38	■ 48	■ 58
■ 9	■ 19	■ 29	■ 39	■ 49	■ 59
■ 10	■ 20	■ 30	■ 40	■ 50	■ 60

Figure 7: Self-organising map of arthropod evolutionary profiles with k-means superclusters. The top 10 principal components of ~80'000 arthropod orthologous groups' evolutionary profiles were assigned to 900 cells of the self-organising map. Cell-to-cell proximity indicates similarity in multi-dimensional evolutionary trajectories. Within cells, pie charts represent the magnitudes of the node's feature contributions, where larger sectors correspond to stronger evolutionary driving forces. Each cell is assigned to one of the 60 k-means superclusters, represented by an automatic selection of the most distinctive pastel colour codes.

Hierarchical Clustering of SOM Superclusters

While SOM cells and k-means superclusters are associated with specific evolutionary profiles; these correspond to representative archetypes (eigengenes) obtained from the algorithm's learning model built from the input data used for clustering optimisation and visualisation purposes. With the goals of 1) highlighting patterns emerging from the raw data instead of modelled values and 2) more concisely describing broader groups of evolutionary trajectories, a hierarchical clustering algorithm was applied to the lists of orthologous groups associated with the k-means superclusters. Employing the SOM algorithm was necessary to spatially distribute and enrich clusters with evolutionarily-similar orthologous groups in a lower-dimensional space that could be more easily interpreted while maintaining the underlying topology of the dataset. While such an approach is optimised for a detailed and precise exploration of the evolutionary space of arthropod genes, it was crucial to further summarise and characterise broader patterns of evolutionary trajectories for the descriptive purposes of investigating more generic evolutionary-functional correspondences.

Generating the SOM k-means superclusters summarised and vastly reduced the size of the dataset, and hierarchical clustering was finally deemed suitable to effectively discriminate the already sufficiently diverse superclusters' average evolutionary feature scores. The hierarchical clustering algorithm was therefore applied to 60 evolutionary profiles only, instead of the original ~80'000 from the initial dataset, overcoming the computational limitations described at the beginning of this chapter. This additional clustering step would generate hierarchical relationships, displayed as dendrograms in Figure 8, of both superclusters of orthologous groups and their corresponding evolutionary features. Relationships built on the evolutionary features, as described in Chapter 1, rather than on principal components, enables an intuitive association of evolutionary profile types with more easily interpretable and quantifiable properties. Both feature and supercluster dendrograms were combined in Figure 9, displaying a heatmap highlighting comparable scaled feature values from each of the 60 superclusters.

Profiles built from the 16 quantified evolutionary features successfully delineate key similarities and differences amongst the 60 SOM superclusters. Contrasting the profile of a given supercluster against the profiles of all superclusters reveals the evolutionary features that most clearly distinguish each supercluster. Several bootstrap-supported groupings of superclusters and subsets of features are revealed when hierarchical clustering is applied to the matrix of evolutionary feature profiles of all arthropod orthologous groups. Delineation of the hierarchical similarities amongst superclusters and features enables the identification of subsets of features that vary in concert, and broad groups of evolutionarily similar superclusters. The orthologous groups were averaged by median values for each supercluster and employed to build a dissimilarity matrix with Pearson's correlation distances. Performing bootstrapped clustering with the average linkage method resulted in several well-supported subsets and groupings. Using Pearson's correlation distances for clustering aimed to give weight to the features' directions rather than their magnitudes or ranks (Kassambara 2017), on the lines of the principal components approach employed for the SOM feature clustering.

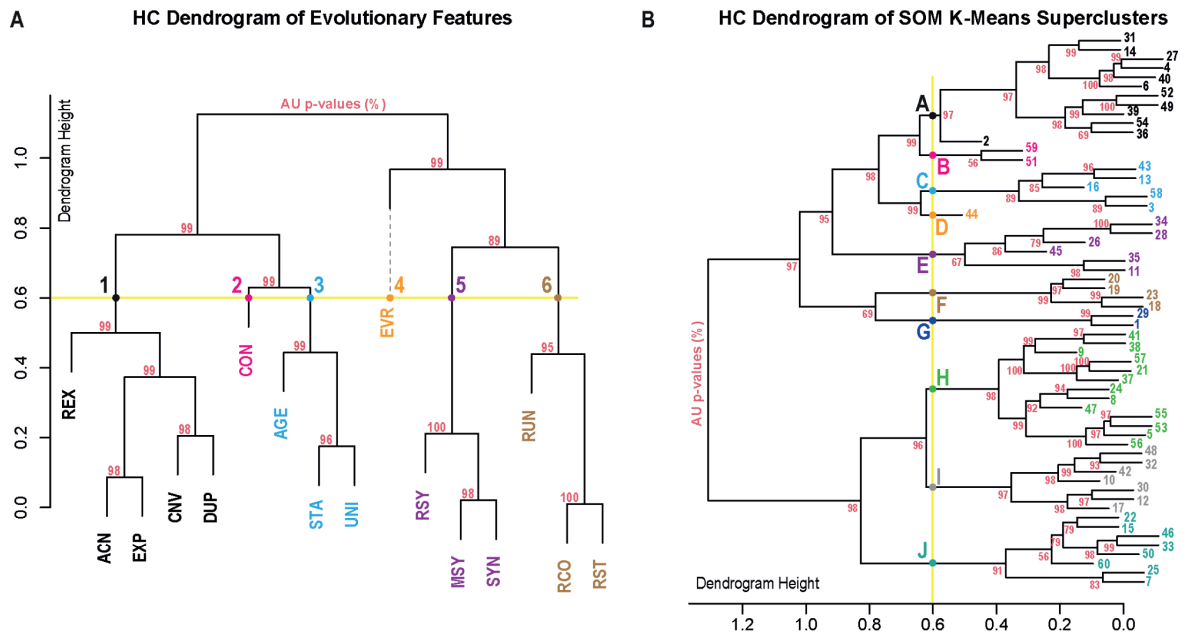


Figure 8: Hierarchical clustering relationships of evolutionary features and SOM superclusters. Panel A represents the dendrogram of the hierarchical relationships of the 16 evolutionary features. Panel B represents the dendrogram of the hierarchical relationships of the 60 SOM k-means superclusters of orthologous groups. Both dendrograms were cut at an arbitrary Pearson’s correlation distance of 0.6 (indicated with yellow lines), obtaining six feature groups in A (1-6) and 10 evolutionary profile types in B (A-J). Confidence measures for each node are represented by percentage AU p-values in red (100% for maximum support and 0% for no support).

The distance matrix was computed from the family medians using the *dist* and *cor* functions from the *stats* package in R, with distances defined as *1-correlation* for matrices computed with *cor*. Hierarchical clustering was performed using R's *hclust* function from the *stats* package, and bootstrap support of the resulting hierarchies (dendrograms) was obtained using R's *pvclust* package with 10'000 bootstrap replicates and a fixed seed of 12345, which calculates p-values for hierarchical clustering via multiscale bootstrap resampling (Suzuki and Shimodaira 2006). The selection of clustering algorithm method (average-linking), distance method (Pearson's correlation dissimilarity matrix) and profiles' averaging function (median) reflects the methodology described in Chapter 3 (Ruzzante et al. 2022). It corresponds to the best-identified clustering algorithm combination revealed by similar analyses performed on the evolutionary features of *A. gambiae* immune-related gene families, producing the most reproducible distribution of approximately unbiased (AU) support values across 10'000 hierarchical clustering bootstrap replicates obtained with *pvclust*. The approximately unbiased (AU) p-values provide a confidence measure for each node of the cluster dendrograms of families and evolutionary features. The resulting heatmap displayed in Figure 9 was obtained with the ComplexHeatmap R package (Gu, Eils, and Schlesner 2016).

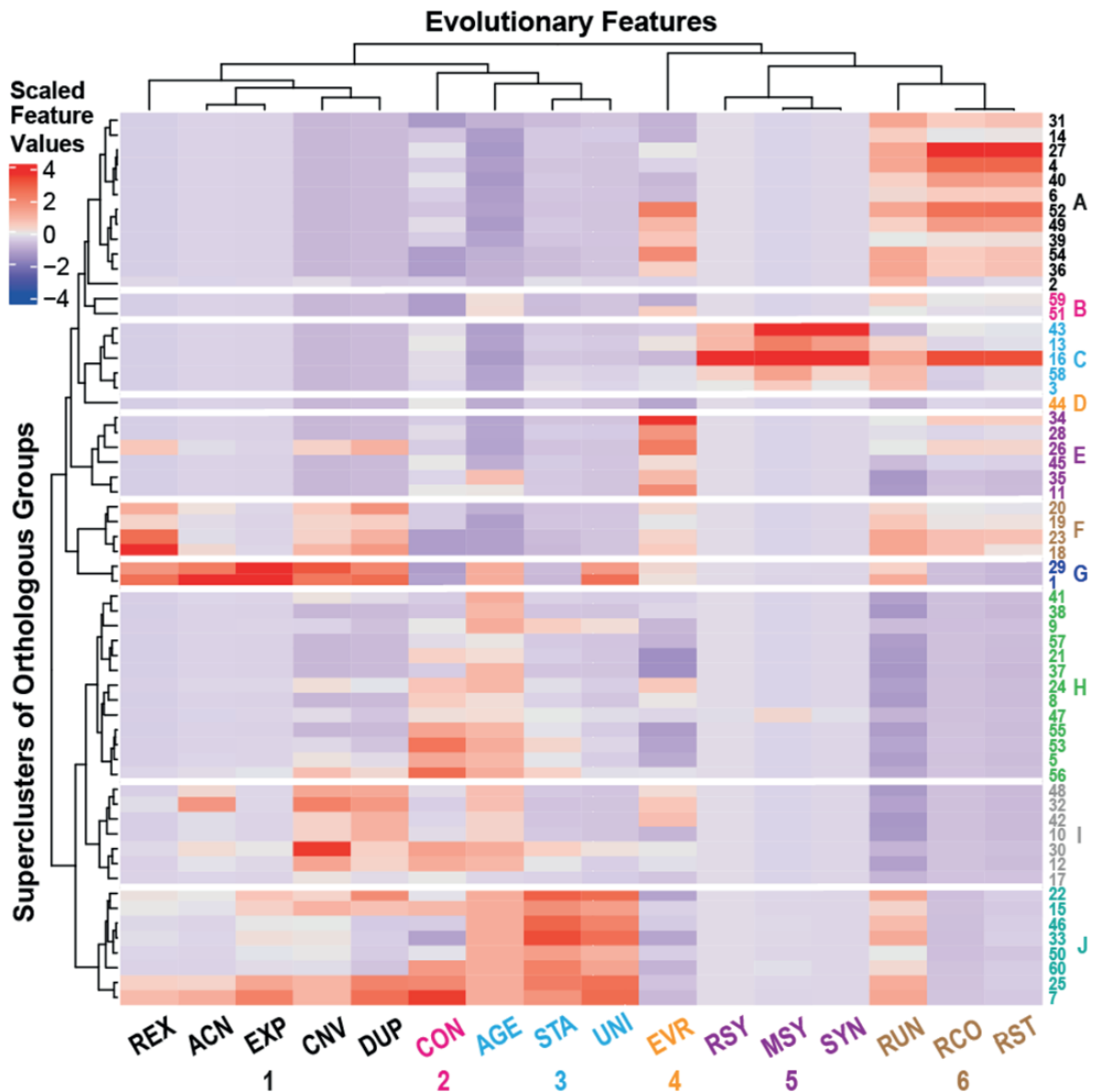


Figure 9: Heatmap from the hierarchical clustering of SOM superclusters and evolutionary features. On the horizontal axis are represented the 16 evolutionary features, and on the vertical axis are represented the 60 SOM k-means superclusters. Opposing the labels are shown the corresponding dendrograms representing the hierarchical relationships. Both features (1-6) and superclusters types (A-J) are colour coded to indicate the cutoff partitioning from Figure 8. Heatmap squares are coloured from dark blue to dark red to indicate the scaled values of the orthologous groups' superclusters averaged by the statistical median.

Figure 8A once again confirms some of the expected relationships between evolutionary features: copy-number, duplicability and gene family expansion features come together in group 1; group 2 is solely constituted by widespread gene losses; group 3 captures taxonomy spread, age and stability features; group 4 recovers sequence evolutionary rate; group 6 recovers the three synteny-conservation features; and group 7 represents features of lineage-specific dynamics, including relative universality, gene family losses and stability. The robustness of the hierarchical clusters is indicated by the node support values, consistently high throughout the dendrograms of features and, to a lesser extent, superclusters. The description of distinct evolutionary feature profiles associated with groups of superclusters in Figure 9 allows for a precise characterisation of the 10 types of evolutionary trajectories of arthropod orthologous groups. This step created a profiling framework that eased the description of the correspondences between the 60 superclusters and their respective functional annotation summaries, as presented in Table 2.

Type A describes 12 superclusters and 23'892 orthologous groups, defined by high lineage-specific universality and gene turnover dynamics with variable sequence evolutionary rates and young taxonomic age. **Type B** describes two superclusters and 1'936 orthologous groups, defined by high lineage-specific taxonomic spread, medium age, no gene losses and low duplicability. **Type C** describes five superclusters and 1'428 orthologous groups, defined by the highest synteny conservation scores, young age, high lineage-specific taxonomic spread and low duplicability. **Type D** describes one supercluster and 2'787 orthologous groups, defined by overall medium values with slightly lower age and sequence evolutionary rates. **Type E** describes six superclusters and 6'667 orthologous groups, defined by the fastest-evolving genes with medium to low universality, stability, age and lineage-specific spread. **Type F** describes four superclusters and 3'678 orthologous groups, defined by the highest counts of lineage-specific expansions, consequential high duplicability, few gene losses, young age, and mid-high evolutionary rates. **Type G** describes two superclusters and 77 orthologous groups, defined by the highest widespread expansions and copy-numbers of old and universally spread gene families, with few gene losses. **Type H** describes 13 superclusters and 23'820 orthologous groups, defined by old age, low lineage-specific universality and

dynamics, mid-low sequence evolutionary age and frequent losses. **Type I** describes seven superclusters and 6'603 orthologous groups, defined by the highest gene copy-number variation, high duplicability, old age, variable evolutionary rates and low lineage-specific spread. **Type J** describes 12 superclusters and 7'249 orthologous groups, defined by the oldest age, most stable and universally spread gene families with low sequence evolutionary rates, mid-high expansions and variable loss frequencies.

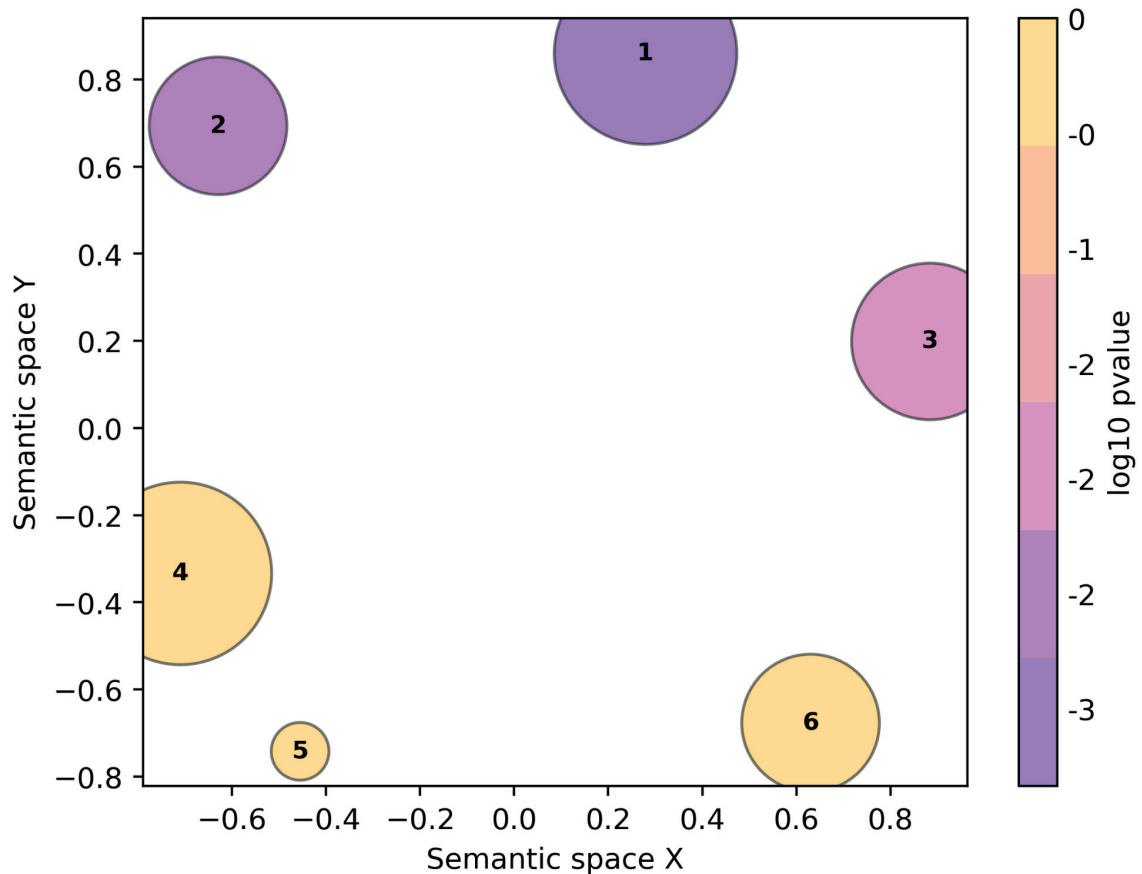
Radically different evolutionary trajectory types can be found on the opposing sides of the heatmap. At the top, type *A* captures the youngest, lineage-specific and fast-evolving genes in striking contrast to the oldest, universally maintained, slow-evolving and stable gene families of the bottom type *J*. Interestingly, type *J* is also characterised by highly diverging counts of gene losses (superclusters 7 and 33), potentially indicating variable functional essentiality across different arthropod clades, for which the genetic house-keeping properties are not universally maintained. More likely, these high loss counts are a direct consequence of the widespread and lineage-specific expansions, and plausibly not all gene copies might have been retained over such large evolutionary time scales. Although, whenever expansion events have not been detected, these widespread, old, slow-evolving genes are most likely maintained in single or few gene copy counts, as indicated by the low copy-number features and duplicability. Types *F* and *G* share large lineage-specific expansions and medium to high evolutionary rates. Type *F*, though, solely describes young and clade-specific genes, possibly indicating frequent gene duplications acting as recent genetic adaptation mechanisms. Alternatively, type *G* seems to capture those ancient, widespread, fundamental gene family expansions shared and maintained across the entire arthropod phylogeny.

Functional Annotations of the SOM Clusters

Functional enrichment of the 900 SOM cells was performed with the same approach described in the enrichment of extreme evolutionary feature values at the beginning of this chapter. Instead of feature score ranks evaluated by the Kolmogorov-Smirnov test, lists of orthologous groups were tested with Fisher's exact test using *topGO's* "weight01" algorithm. Of the 893 SOM cells successfully associated with lists of evolutionarily-similar orthologous groups, 675 were assigned statistically significant GO terms enrichments (p -value < 0.05), with no cell showing patterns of extreme statistically significant terms as previously observed in some of the most skewed extreme feature values. The 900 SOM cell GO term enrichments were summarised and visualised with GO-Figure!. Given the SOM's relatively homogenous distribution and limited cluster sizes, functional enrichments of its cells provide specific and easily interpretable functional information on biological processes.

As an example, Figure 10 shows the GO-Figure! visualisations of the GO terms enrichment of the peculiar SOM cell (30, 1), identifiable at the bottom-right corner of the full arthropod SOM in Figure 7. The cell brings together 14 orthologous groups, evolutionarily characterised by lineage-specific dynamics and, in particular: old, slow-evolving gene family expansions with maintained copy-numbers; high synteny conservation scores and occasional gene losses. Querying the orthologous groups on the *OrthoDB* website reveals how most groups are shared by several fruit fly species, with exceptions including a few species of crustaceans, mosquitoes and lepidopterans. Functional annotation of the cell pinpoints biological processes related to auditory behaviour and sensory perception of sound. The potassium ion transport term may be associated with the mechanical induction of the auditory neural circuit. The distant semantic similarity between the first two terms and the 6th term, although all related to the auditory system, are likely distanced by GO-Figure! architecture, which keeps separated related processes when related by the "is a process of" ontological connection. Interestingly, one additional significantly enriched term ($p=0.00213$) not reported by GO-Figure!'s summarization algorithm was "adult walking behaviour" grouped within the 6th term "response

to auditory stimulus", assuming that walking could be semantically described as a response to sound stimuli in flies. Such results are nevertheless not meant to produce precise molecular functional annotations across millions of arthropod genes, for which detailed functional genetic research is still necessary, but rather provide a range of automatically computed predictions for exploratory analyses.



- | | |
|--------------------------------|---|
| 1. auditory behavior | 4. nucleobase-containing compound meta... |
| 2. sensory perception of sound | 5. cellular macromolecule metabolic pr... |
| 3. potassium ion transport | 6. response to auditory stimulus |

Figure 10: Summary visualisation of SOM cell (30, 1) functional enrichment. The statistically significant (1: $p=0.00046$; 2: $p=0.00594$; 3: $p=0.01666$) groups of enriched GO terms can be found from mid to top of the vertical axis; where the top edge corresponds to the semantically similar terms relating to sound perception and auditory behaviour, while the right group corresponds to a generic potassium transport term.

Functional Annotations of the SOM Superclusters

Table 4 shows the k-means superclusters' functional enrichments, ordered by the evolutionary profile types characterised in Figure 9. The statistical enrichment was performed as previously described for the 900 SOM cells. Several superclusters, mostly of type *H* and *I*, recovered large lists of statistically significant (p -value < 0.05) GO terms due to capturing extensive lists of orthologous groups associated with broad ranges of biological functional properties. Of the 60 superclusters, 10 did not show statistically significant functional enrichments for any term, particularly within type *A* and *E*, and a few showed precise biological process enrichments, likely clustering orthologous groups characterised by extremely peculiar evolutionary profiles. These observations indicate that it is possible to recover similarly evolving genes from comparative genomics analyses and successfully associate them with either broad or specific functional categories. It is important to note that variations of evolutionary trajectories are still present within evolutionary profile types and SOM superclusters are here summarised by averages to ease the description of generic patterns of evolution-to-function characterisations.

It is likely that different but fundamental biological processes, imposing similar functional constraints shaped through individual organism fitness and survivability or selective pressures on whole populations, adopt the same evolutionary trajectories. These may not be distinguished by employing evolutionary characterisation and would produce large clusters of evolutionarily-described orthologous groups with a high variation in functional annotations. Finally, the resolution at which such correspondences are inferred cannot satisfy the characterisation of all arthropod gene families, for which additional evolutionary properties may be needed to be defined and clustering techniques refined and specifically tailored.

The following paragraphs discuss the identification and interpretation of generic patterns of evolution-to-function correspondences while keeping in mind that more specific annotations are gathered from the full SOM's higher resolution but cannot be individually discussed here. Ultimately, the 900 cell functional

annotation plots will be made available for interactive exploration through a Dash web application, an open-source framework for building data visualisation interfaces for Python, currently in development. The web app will include a browsable look-up table for the complete list of arthropod gene identifiers, the corresponding associated orthologous group evolutionary feature scores and the cell's GO-Figure! plots of summarised functional enrichments.

Table 4. Functional annotations of superclusters of orthologous groups. The 60 SOM k-means superclusters of orthologous groups (OG) are described by their size, evolutionary profile type and top 30 statistically significant GO term enrichments (p-value < 0.05, *n.s.* otherwise). Extremely significant terms (p-value < 10⁻³³) emerge from over-represented terms such as '*cellular process*' or '*metabolic process*' and were filtered out. The largest lists of GO terms were manually curated and semantically summarised to represent the most distinguishable biological processes.

Super-Cluster ID	Counts of OG	Evolut. Profile Type	Curated Summaries of the Top 30 Significantly Enriched GO terms
31	2'892	A	phosphorus metabolism
14	3'855	A	neuromuscular junction development; compound eye morphogenesis; cell polarity; amnioserosa formation; posterior Malpighian tubule development
27	789	A	hemolysis in other organism; disruption of cells of other organism
4	2'288	A	<i>n.s.</i>
40	2'541	A	<i>n.s.</i>
6	2'671	A	transport vesicle recycling; protein localization to Golgi apparatus
52	712	A	transcription, RNA-templated
49	1'074	A	<i>n.s.</i>
39	2'044	A	<i>n.s.</i>
54	1'128	A	positive regulation of antibacterial peptide biosynthesis
36	2'013	A	<i>n.s.</i>
2	1'885	A	cell adhesion; embryonic plasmatocyte differentiation; leg morphogenesis; cell shape regulation; glial cell development; trehalose biosynthesis; negative regulation of post-mating female receptivity; mitotic spindle organization
59	1'146	B	positive regulation of nucleic acid-templated transcription
51	790	B	chloride transport
43	48	C	chaeta development
13	68	C	protein insertion into membrane
16	14	C	auditory behavior; sensory perception of sound; potassium ion transport
58	268	C	meiotic DNA double-strand break processing involved in reciprocal meiotic recombination; mitochondrial membrane organization
3	1'030	C	chaeta development; heterochromatin assembly; cellular response to X-ray; mevalonate pathway; response to symbiotic bacterium; SAGA complex assembly; tricarboxylic acid cycle; ecdysone biosynthesis; reciprocal meiotic recombination; apoptosis involved in tissue homeostasis; cohesin loading; proteasome assembly; spermatogenesis; protein secretion; citrate metabolism; digestive tract development; response to nicotine; peptidoglycan recognition protein signaling pathway
44	2'787	D	response to gibberellin; chromate transport
34	260	E	<i>n.s.</i>

28	1'024	E	protein localization to kinetochore; kinetochore assembly; chromosome separation; aromatic amino acid transport; regulation of chromosome segregation
26	403	E	n.s.
45	2'431	E	n.s.
35	1'798	E	nuclear pore complex assembly; L-phenylalanine biosynthesis
11	751	E	n.s.
20	367	F	branching involved in ureteric bud morphogenesis; regulation of hemocyte proliferation; sensory perception of bitter taste
19	1'893	F	urea transmembrane transport; peptidyl-proline hydroxylation to 4-hydroxy-L-proline; 4-hydroxyproline metabolism; regulation of voltage-gated potassium channel activity
23	1'162	F	n.s.
18	256	F	chemosensory behavior; hemolysis in other organism; disruption of cells of other organism
29	71	G	reproduction; nuclear division; leukocyte migration; glycosyl compound metabolism; regulation of body fluid levels; flavonoid glucuronidation; gene expression
1	6	G	DNA metabolism; lipid hydroxylation; lauric acid metabolism; amacrine cell differentiation; insecticide catabolism; nuclear division; hemolymph coagulation; neural crest cell migration; detection of chemical stimulus involved in sensory perception of smell; cellularization; hyperosmotic response; epithelium development; regulation of tumor necrosis factor production; post-embryonic development; monosaccharide biosynthesis; phenol-containing compound biosynthesis; cell motility; gene silencing
41	1'642	H	anatomical structure development; transmembrane receptor protein serine/threonine kinase signaling pathway; nuclear division; signaling; histone acetylation; response to insulin
38	2'856	H	microtubule-based movement; social behavior; erythrocyte homeostasis; basophil differentiation; benzoate transport; gastrulation with mouth forming second; dosage compensation
9	1'090	H	RNA splicing; nuclear division; neurogenesis; epithelium development; skeletal system morphogenesis; histone acetylation
57	3'003	H	growth; histone ubiquitination; axon extension; female receptivity; microtubule-based movement
21	3'071	H	post-embryonic development; microtubule-based movement; reproduction; calcium homeostasis; anatomical structure development; cell junction organization; cellular localization; unidirectional conjugation; ethanolamine transport; nicotinamide riboside transport; polyamine transmembrane transport
37	2'957	H	development; microtubule-based movement; positive regulation of gene expression; hatching behavior; postsynaptic density organization
24	1'228	H	nuclear division; kidney epithelium development; signaling; sensory perception; reproduction; poly-hydroxybutyrate biosynthesis; regulation of trehalose metabolism; RNA-dependent DNA biosynthesis; astral microtubule organization
8	2'121	H	oocyte microtubule cytoskeleton organization; cilium movement; DNA methylation on adenine; axonemal dynein complex assembly
47	693	H	neurogenesis; sensory perception; reproduction; nuclear division; DNA transposition; L-cystine transport; negative phototaxis;

			protein localization to nuclear pore; sensory organ morphogenesis
55	1'743	H	post-embryonic development; cellular calcium ion homeostasis; reproduction; epithelium development; cell junction organization; chloride transport; glycerol-3-phosphate metabolism; glycosyl compound metabolism
53	1'020	H	cellular calcium ion homeostasis; reproduction; epithelium development; cell junction organization; regulation of transmembrane receptor protein serine/threonine kinase signaling pathway; growth; post-embryonic development; Notch signaling involved in heart development; L-alanine metabolism; apical protein localization; regulation of synapse structure or activity
5	1'801	H	post-embryonic development; cellular calcium ion homeostasis; epithelium development; microtubule-based movement; monosaccharide biosynthesis; glycosyl compound metabolism; reproduction; thermotaxis; cell-cell adhesion; response to insecticide
56	595	H	cytokine production; glycosyl compound metabolism; T cell activation; leukocyte cell-cell adhesion; leukocyte migration; sensory perception; pyridine nucleotide biosynthesis; plasmid maintenance
48	783	I	cell wall catabolism; oligosaccharide catabolism; response to steroid hormone; substituted mannan metabolism; cGMP-mediated signaling; evenomation of other organism; defense response to Gram-negative bacterium; endopeptidase activity; detoxification of cadmium ion; ganglioside catabolism; chorion-containing eggshell pattern formation; self proteolysis; carotenoid biosynthesis; inner ear development; chloride transport; metal ion homeostasis; piRNA metabolism
32	101	I	chorion-containing eggshell formation; columnar/cuboidal epithelial cell development; hemolysis in other organism; disruption of cells of other organism; telomere maintenance
42	1'568	I	cellular response to methanol; negative regulation of inflammatory response; cell wall organization; response to paraquat
10	1'721	I	chloride transport; paracrine signaling; imidazole-containing compound catabolism; negative regulation of endopeptidase activity; response to auditory stimulus; positive regulation of cellular response to oxidative stress; fasciculation of motor neuron axon; mediolateral intercalation; cardioblast differentiation; regulation of Roundabout signaling pathway; protein arginylation; deoxycytidine catabolic process; neuroblast differentiation; inflammatory response
30	117	I	reproduction; response to cold; adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains; cellular calcium ion homeostasis; cellular response to acid chemical; post-embryonic development; negative regulation of cytokine production; ionotropic glutamate receptor signaling pathway; bacteriocin transport; detection of chemical stimulus involved in sensory perception of pain; negative regulation of neuron migration
12	913	I	response to cold; sensory perception; peptidoglycan biosynthesis; lipopolysaccharide biosynthesis; pentacyclic triterpenoid biosynthesis; phosphorylation of STAT protein; induction of bacterial agglutination; pyridine nucleotide biosynthesis; fructose 6-phosphate metabolism; sex determination, primary response to X:A ratio
17	1'400	I	microtubule depolymerization; pronuclear migration; pole cell formation; sensory organ morphogenesis; pronuclear fusion; female meiotic nuclear division; filopodium assembly; cellular component organization; secondary branching, open tracheal system; regulation of asymmetric cell division; Malpighian tubule tip cell differentiation; regulation of kainate selective glutamate receptor activity

22	901	J	determination of adult lifespan; glycosyl compound metabolism
15	291	J	reproduction; epithelium development; nuclear division; leukocyte differentiation; response to cold; calcium homeostasis; post-embryonic development; organic hydroxy compound biosynthesis; cellular component organization; cellular response to acid chemical; regulation of terminal button organization; cytoskeleton-dependent intracellular transport
46	1'651	J	epithelium development; post-embryonic development; cytoskeleton-dependent intracellular transport; cilium movement
33	2'822	J	ubiquitin-dependent protein catabolism
50	876	J	nuclear division; reproduction; epithelium development; cellular component organization; cytoskeleton-dependent intracellular transport; post-embryonic development; skeletal system morphogenesis
60	316	J	response to stimulus; epithelium development; cytochrome complex assembly; regulation of Rho protein signal transduction; signaling; central nervous system neuron axonogenesis; calcium homeostasis; reproduction
25	300	J	post-embryonic development; response to peptide hormone; glycolipid metabolism; leukocyte differentiation; regulation of body fluid levels; ionotropic glutamate receptor signaling pathway
7	92	J	sleep; reproduction; response to peptide hormone; regulation of body fluid levels; development; cytokine production; leukocyte differentiation; defense response to Gram-positive bacteria; cuticle hydrocarbon biosynthesis

Type *A*, defined by high lineage-specific gene losses and species-span with low copy-numbers can be subdivided in two groups of sequence evolutionary rate. Slow-evolving orthologous groups of type *A* (31-6 and 2) recover genes involved in insect anatomical organisation, including embryonic, neuromuscular and Malpighian tubule (arthropods' osmoregulatory and secretion organs) development, compound eye and leg morphogenesis. Significant behavioural associations include post-mating female receptivity, trehalose biosynthesis (involved in freezing and dehydration survival), hemolysis (involved in blood-feeding), and disruption of other organisms' cells (related to stinging/biting insects). Fast-evolving type *A* genes are only significantly enriched for the immunity-related regulation of antibacterial response, likely reflecting evolutionary adaptations resulting from host-pathogen arms races. Five superclusters show non-significant functional enrichments, evenly shared across subtypes, indicating type *A*'s capacity to retrieve at the same time highly specific and nonspecific functional annotations.

The only significant associations of type *B*, grouping those genes characterised mainly by high levels of conservation and low copy-numbers, are with chloride ion transportation and the positive regulation of transcription mechanisms. With average values on all other evolutionary features, type *B* likely recovers a wide selection of mediumly old, stably maintained and physiologically important gene functions. Type *B* genes are not universally widespread but seem to have higher-than-average relative universality scores, possibly indicating lineage-specific essential adaptations. No particular functional module or biological property emerged from the few statistically significant GO term enrichments.

Functional annotation of superclusters profiled by type *C*, the most synteny-conserving genes, young and with low copy numbers, strongly relate to the development of chaetae. Chaetae are functionally diverse chitinous bristles involved in various systems, including sensory reception and stridulation. Emerging consequent functions include auditory behaviour and sound perception from supercluster 16, capturing the most distinctive orthologous groups defined by lineage-specific all-or-nothing gene losses, likely linked to functional mechanisms tied to whole syntenic blocks of genes. Such syntenic mechanisms

of all-or-nothing inheritance inevitably draw attention to mating-related speciation mechanisms in dipterans (through cross-species radiation of wing-vibration patterns) and orthopterans (through cross-species radiation of stridulation patterns). Although not specifically retrieved from type *C* functional annotations (let alone the X-ray irradiation response term), chitinous structures are known to be involved in iridescence through light diffraction, notoriously in the diversification of lepidopteran wing scales. These are also associated with mating behaviour and possibly further support the syntenic-dependent inheritance of type *C* genes, associating the heritability of highly syntenic blocks of genes with chitinous structures and sexually-selected arthropod species radiations. Additional interesting annotations specifically from supercluster 3 include responses to plant-defence mechanisms, including terpenes (mevalonate pathway), citric acid (tricarboxylic acid cycle and citrate metabolism) and nicotine. Again, these biological processes may drive speciation through selective pressures in herbivore insects, as recently described in coleopterans (Seppey et al. 2019).

Supercluster 44 is the only representative of the evolutionary type *D*, grouping a large set of genes mainly defined by young age, low sequence evolutionary rates and copy numbers. The resulting functional enrichment solely points to chromate transport and response to gibberellin, a class of phytohormones regulating various plant developmental processes. These genes are likely brought up by herbivore insects, with chromate transport potentially linked to a series of chromate-based insecticides. As such products are selected for their high efficacy and durability, it is reasonable to assume that they target conserved lineage-specific metabolic pathways governed by slow-evolving genes, reducing the emergence potential of insecticide resistance-driving adaptations.

Type *E*, although mostly grouping together non-significantly enriched superclusters, seems to mainly point to nuclear arrangement mechanisms involved in cell division, represented by chromosome segregation, kinetochore and nuclear pore complex assembly. These conserved and fundamental functions are unintuitively associated with fast-evolving young genes. Nevertheless, given the overall low statistical significance of type *E* enrichments and orthologous groups' median universality, it is possible that the emerging nuclear division

functions are supported by a portion of genes within a broader functional diversity and may not be adequately represented by the supercluster's evolutionary feature median values.

Type *F* orthologous groups are characterised by lineage-specific gene expansions, high potential for duplication, high lineage-specific spread, young age and relatively high sequence evolutionary rates. Similarly to type *A*, Type *F* superclusters also bring together genes involved in Malpighian tubule (the functional analogues to mammalian kidneys) development and activity, including ureteric bud morphogenesis and urea transmembrane transport. More specific than type *A* development-related functional enrichments, developmental associations with type *F* are specific to Malpighian tubule formation and mechanisms, further supported by the enrichment of hemocyte proliferation, related to renal tubule morphogenesis in *Drosophila* (Bunt et al. 2010). Hydroxyproline metabolism-related enrichments point to plant defence mechanisms against pathogens and herbivore insects (Kite et al. 1995; Bhattacharya et al. 2013), possibly linking Malpighian tubule-related mechanisms of detoxification with selective pressures driving gene family expansions in specific lineages. Such evolutionary-functional correspondences confirm previous findings, associating lineage-specific expansions with response to pathogens and environmental stress (Lespinet et al. 2002) across larger evolutionary timescales. Supercluster 18 adds more generic enrichments to type *F* for the perception of bitter taste, chemosensory behaviour and hemolysis.

Functional annotations recovered from enrichments of type *G* superclusters emerge from a relatively low number of orthologous groups (77 in total) compared to the much larger sizes of other evolutionary profile types. Such groups are characterised by the most taxonomically widespread and old gene family expansions, both universal and lineage-specific. These old groups with few losses, consequently high copy-numbers and copy-number variations, show average values of evolutionary rates. Seemingly, this peculiar evolutionary type, bringing together only two superclusters, is adopted by a wide range of biological functions, including reproduction, immunity, homeostasis, retinal neuron differentiation and generic neuronal development, chemosensory perception of smell, tumour necrosis and insecticide catabolism. Represented by

such broad annotations, type *G* orthologous groups' functional correspondences could be better informed by the higher resolution functional enrichments of specific SOM cells.

Superclusters of type *H* constitute one of the largest clusters of orthologous groups. With similarly large sizes compared to type *A*, they are nevertheless contrastingly significantly enriched for a wide variety of biological functions. More frequently than the rest, however, emerge enrichments related to organs' developmental processes and microtubule-based movement (associated with the movement of organelles or other cellular components), with the latter specifically enriched in 7 out of 13 superclusters. Contrary to type *G* superclusters, bringing together only a few orthologous groups, the breath of type *H* functional annotations could nevertheless similarly benefit from higher-resolution GO term enrichments of its specific SOM cells, leaving room for additional extensive explorations. Unsurprisingly, development and cellular structure-related functions correspond to old genes with averagely low sequence evolutionary rates, copy numbers and expansion potential. They are, however, not universally widespread across the arthropod phylogeny, possibly bringing together genes that may have been functionally replaced in specific sub-lineages by other genes with different evolutionary types. It is also likely that such old genes may not have been entirely captured and correctly annotated or orthologically-delineated across the full breadth of the 170 arthropod genomes, de facto biasing the computation of taxonomy-sensitive evolutionary features.

Evolutionarily similar to type *H*, type *I* also brings together a broad range of diverse biological functions. Feature-wise, the main difference distinguishing type *I* from type *H* is constituted by the higher scores from copy-number related features, indicating functional constraints requiring higher potential for gene duplicability. Within the large lists of significantly enriched GO terms, rather unsurprisingly emerges the "*response*" descriptor, associated with terms relating to the response to cold and different chemicals (methanol, acids and oxidative stress), auditory stimulus, signalling pathways, and the inflammatory response. More specific immune response mechanisms include cytokine production, recombination of receptors from immunoglobulin domains and response to

bacteria. Specific to superclusters 48 and 32 are worthy of note the enrichments for eggshell formation.

Finally, type *J* superclusters are mainly enriched for developmental processes. Similarly to type *I*, additional significant annotations include functions associated with the immune response (including haemocyte differentiation, cytokine production and defence response to bacteria) and various responses to external stimuli. Supercluster 33 seems to be specifically enriched for the ubiquitin-dependent catabolism of proteins across 2'822 orthologous groups. Moreover, particularly highlighted for their higher-than-average gene expansions and copy-numbers, superclusters 25 and 7 (the dendrogram's sister groups to the rest of type *J* superclusters) are additionally enriched for post-embryonic development and cuticle biosynthesis. Such mechanisms are likely associated with wide variations of cuticle proteins in arthropod developmental stages and body regions (Charles 2010), regulating moulting and metamorphosis. Overall, type *J* recruits genes from old orthologous groups defined by strikingly high stability of ancestral state copy-numbers and widespread universality, including widespread representations in specific arthropod sublineages with either low or high gene-copy numbers. Low sequence evolutionary rates further highlight type *J*'s essential, ancient and universally maintained functional properties. When detected, type *J* universal gene losses are strictly associated with higher-than-average gene expansions with maintained universality and stability. Such patterns are likely a result of gene copy-number decreases following large gene family expansions, possibly by mechanisms of purifying selection, and are not indicators of extensive gene losses associated with fundamental biological functions related to developmental and immune-response processes.

Lineage-Specific Evolutionary Frameworks

Alternative analysis frameworks capturing different evolutionary timescales can be investigated by computing evolutionary features on different user-input orthology delineation tables and phylogenies, from distant eukaryotic radiations to more recent specific arthropod clades. Moreover, the *Evol-Feat* workflow allows the user to specify a species list from the input data, enabling the automated single-species or lineage-specific extraction of orthologous groups and consequent clustering of evolutionary profiles and functional annotations. Such approaches can focus the exploration of evolutionary-functional correspondences on specific clades of interest with an increased clustering resolution. By comparing the evolutionary modules of genes across different clades and timeframes of functional constraints, such analyses may further highlight gene evolution patterns unique to more recent adaptive radiations if supported by the corresponding functional enrichments.

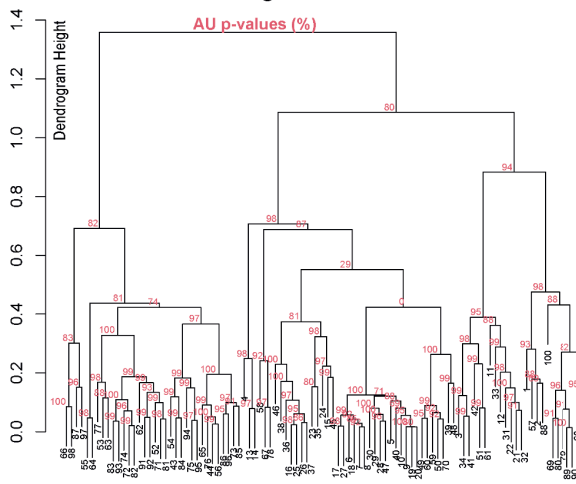
Extraction of lineage-specific orthologous groups was performed twice. Chapter 3 of this thesis discusses in detail the evolutionary profiles of immune-related genes of the African malaria mosquito, *Anopheles gambiae*, further enriched with additional population-level evolutionary features and compared with specific gene expression data. Here briefly showcases an example of a species-specific analysis of the genes of the common fruit fly, *Drosophila melanogaster*. The choice of species was supported by the extensiveness of *D. melanogaster* genome annotations and related increased accessibility to alternative gene functional information compared to other arthropod species, potentially enabling further functional validation methods in the future steps of this area of research. Alternative resources to validate gene function include FlyBase gene groups (Thurmond et al. 2019), KEGG pathways (Ogata et al. 1999) and gene expression databases such as Bgee (Bastian et al. 2008).

From the same Arthropoda orthology and phylogeny input data, the 13'193 *D. melanogaster* genes assigned to 9'642 orthologous group evolutionary profiles were extracted for clustering and functional enrichment analysis, following the methodology described previously in this chapter. Principal

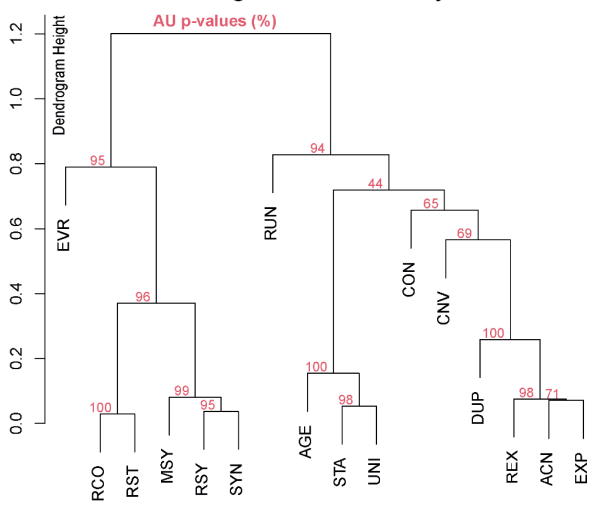
component analysis reached a 95% combined explained variation with the first eight principal components, as opposed to the 10 components from the complete arthropod data. Similarly to the principal components analysis in Chapter 1, but more distinctly defined here, four main groups of evolutionary features leading to the distribution of the evolutionary profiles can be identified. These corresponded to 1) universality, stability, and age features; 2) copy-number features; 3) gene losses and; 4) lineage-specific, synteny-conservation, and evolutionary rate features.

Clustering of the evolutionary profiles was performed with a SOM grid of 10x10 dimension, so as to distribute on average ~90 orthologous groups per cell over a total of 100 SOM cells, following the same parameter selection used in Chapter 1 for the full arthropod SOM. Hierarchical clustering analysis was then performed on the 100 median evolutionary profiles extracted from the average scores of orthologous group features per SOM cell. The robustness of clusters was assessed with AU P-values, resulting in highly supported nodes across most bifurcations of the hierarchical clusters of SOM cells' orthologous groups and evolutionary features. The corresponding statistically supported dendrograms and heatmap are shown in Figure 11.

A HC Dendrogram of SOM Clusters



B HC Dendrogram of Evolutionary Features



C Evolutionary Features

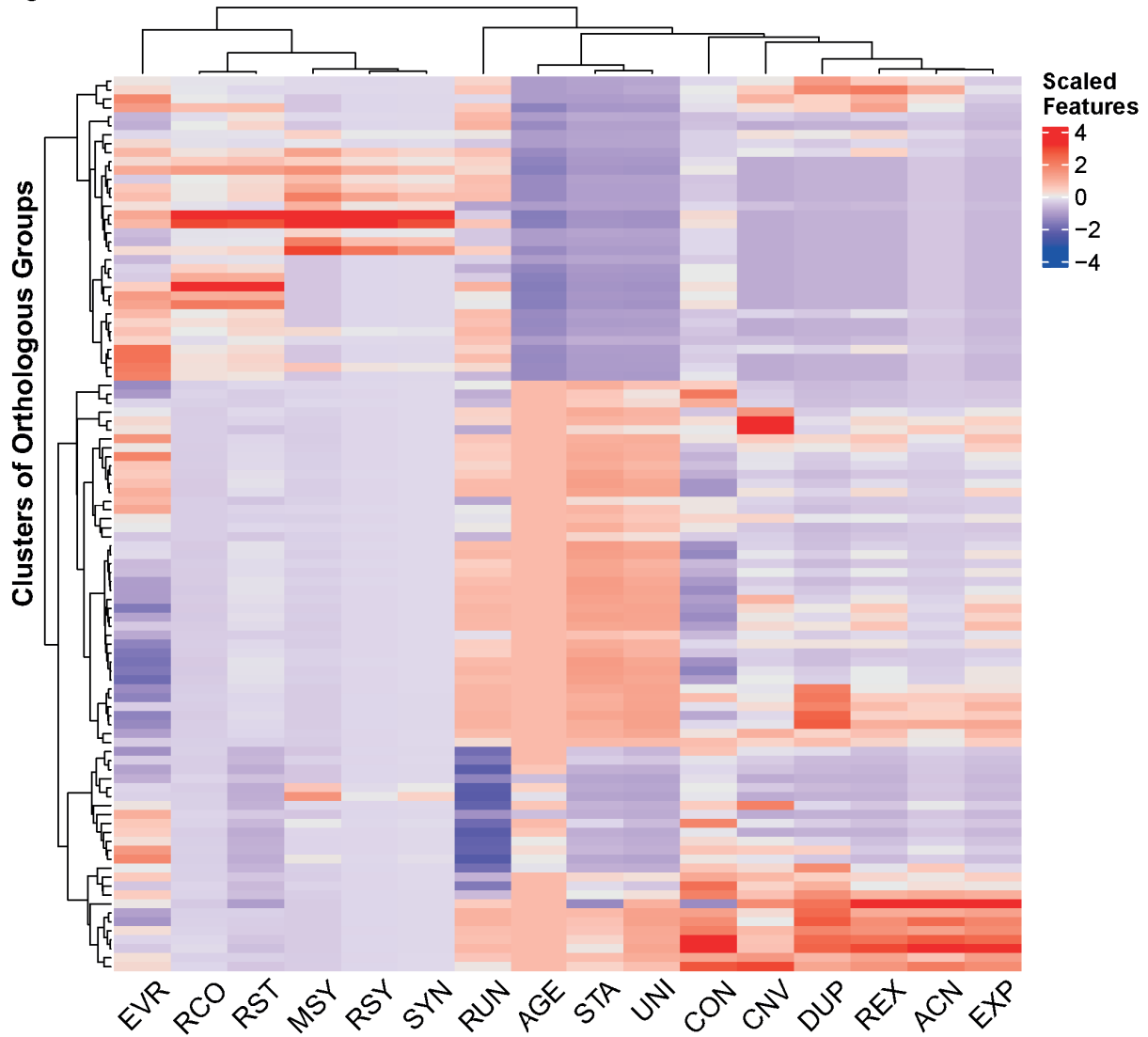


Figure 11: Hierarchical clustering relationships of *D. melanogaster* evolutionary features and SOM clusters. Panel A represents the dendrogram of the hierarchical relationships of the 100 SOM clusters of orthologous groups. Panel B represents the dendrogram of the hierarchical relationships of the 16 evolutionary features. Confidence measures for each node are represented by percentage AU p-values in red (100% for maximum support and 0% for no support). Panel C shows the corresponding heatmap regrouping the hierarchical clusters in A and B, with colour-coded scaled scores of the individual evolutionary features, averaged by median value across the orthologous groups assigned to the SOM clusters.

Compared with the complete arthropod heatmap, the *D. melanogaster* orthologous groups are characterised by similarly behaving groups of evolutionary features, with copy-number and turnover features clustering together, evolutionary rate acting in concert with synteny-conservation features and taxonomic age associated with universality and turnover stability. Differences emerge at the universal gene loss feature (CON) placement, grouping with copy-number and expansion features instead of age and stability. The statistical support is nevertheless low in this case, and overall feature relationships are widely recovered. More notable differences are identified within the median score representations of the orthologous group's feature scores. Compared to the complete arthropod map from Chapter 2, larger portions of the *D. melanogaster*-specific heatmap are occupied by relatively older, universal and stable orthologous groups and those characterised by recent and syntenic gene losses, likely mirroring more recent diversifications in the evolution of drosophilids. These observations highlight how species- or lineage-specific evolutionary maps, even if built from the same evolutionary timescale of measurements, can directly increase the resolution of evolutionary trajectory partitioning and add interpretative nuances to the description of arthropod evolution.

Gene Ontology Enrichment Analysis was performed on the 100 SOM clusters of orthologous groups, following the methodology described earlier in this chapter. Statically significant functional enrichments were obtained for 18 of the 20 SOM k-means superclusters and 91 of the 100 SOM clusters of the orthologous groups containing *D. melanogaster* genes. Thanks to increased access to functional annotations of *D. melanogaster* genes compared to all other arthropod species, it is possible to validate the putative GO function prediction with alternative resources. One of the possible methods included large-scale comparisons with *Drosophila* gene groups, as defined in FlyBase (Thurmond et al. 2019). However, the results are not reported here, as the comparison was challenging due to a greater specificity within molecular function characterisation of FlyBase's gene groups when compared to the broader biological properties emerging from the GO functional enrichment of the SOM clusters. Alternative complementary functional validation strategies are thereby left for future work, including comparisons with KEGG pathway memberships (Ogata et al. 1999),

expression profiles and co-expression networks from transcriptomics analyses (Kuang et al. 2022).

Concluding Remarks

The results from Chapters 1 and 2 show how functional annotations can be associated with sets of genes defined by the evolutionary trajectories of arthropod orthologous groups. Both specific and broader types of evolutionary-functional correspondences can be successfully identified with statistically significant functional enrichments of most clusters of evolutionary trajectories, capturing the evolutionary histories of arthropod genes and linking them to putative biological roles. The resolution of this novel comparative genomics toolkit can be further increased with lineage-specific profiling and clustering of evolutionary features. The resulting arthropod evolutionary-functional map constitutes a readily available resource for the arthropod biology community, allowing for the exploration of more than two million genes from 170 species, characterised by their evolutionary histories and putative functional predictions. Moreover, the *Evol-Feat* bioinformatics workflow enables the investigation of evolutionary-functional correspondences from custom metrics, orthology, and phylogeny datasets. Gene Ontology enrichments of 900 clusters of evolutionary profiles provided statistically supported annotations of biological processes across the high-resolution evolutionary map of arthropod genes. Patterns of evolutionary-functional correspondences were recovered with hierarchical clustering and functional enrichments of the 60 most distinct evolutionary trajectories. These were further summarised to identify 10 different evolutionary strategies observed from the evolution of arthropod gene families, characterised with evolutionary-functional correspondences by associating different evolutionary histories with functional modules of biological properties.

Universally widespread orthologous groups with lineage-specific gene losses and low copy-numbers are partitioned into two sets: slow-evolving genes are enriched with functions related to organism development and anatomical organisation, while the fast-evolving ones are associated with the immunity-related regulation of antibacterial response. Highly syntenic and young orthologous groups with low copy-numbers and lineage-specific all-or-nothing gene losses are associated with the development of chaetae, auditory behaviour,

sound perception, and response to plant-defence mechanisms. These seem to recover lineage-specific adaptations requiring the diversification of, e.g. strongly sexually selected traits such as sensory reception in fly mating behaviour, mediated through the chitinous bristles known as chaetae. Similar evolutionary trajectories, requiring highly syntenic genes to be inherited in genomic blocks, are recovered from other lineage-specific adaptations too, including the response to plant toxins in herbivorous arthropods.

Response to classes of insecticides and phytohormones are also associated with young, slow-evolving orthologous groups with low copy-numbers. These correspondences are likely driven by genes in specific clades of herbivorous arthropods linked to essential but lineage-specific functions, representing the ideal target for such classes of agrochemical products. Young, fast-evolving orthologous groups defined by lineage-specific gene expansions and high duplicability are associated with Malpighian tubule (the main osmoregulatory and excretory organs of arthropods) development and activity, and the metabolic detoxification of plant defence mechanisms. These evolutionary trajectories likely recover young and lineage-specific adaptations for detoxification mechanisms driven by gene family expansions in response to selective pressures from environmental stress. Similarly, more specific functional enrichments point to chemoreception and chemosensation. Contrastingly, old and widespread orthologous groups defined by expansions and high copy-numbers with intermediate evolutionary rates recover a vast range of biological functions, from reproduction to immunity, neuronal development, and insecticide catabolism. Characterising only relatively few orthologous groups, such evolutionary trajectories are likely associated with specific and precise functional modules, but similarly adopted in different organismal biological processes.

Developmental processes and housekeeping functions such as the maintenance of cellular structure and components are associated with old, slow-evolving orthologous groups with low copy-numbers and low potential for gene family expansions. These groups bring together the largest numbers of arthropod genes and more detailed functional annotations could be obtained from the higher-resolution enrichments of the Self-Organising Map cells. Similar evolutionary trajectories but with high copy-number counts and duplicability

potential are enriched for various mechanisms of response to different sources including: cold, chemicals, oxidative stress, auditory stimulus, bacteria, and inflammatory response in general. More specifically, cytokine production, immunoglobulin assembly, and response to bacteria. Haemocyte differentiation, cytokine production, and response to bacteria are further enriched in old orthologous groups characterised by the highest gene stability scores, widespread taxonomic representation, and low sequence evolutionary rates. These are further enriched with a range of developmental processes related to the regulation of arthropod moulting and metamorphosis. Such evolutionary trajectories recover even more ancestral and conserved biological functions, where immunity-related and development-related genes likely point to known pleiotropic properties.

In conclusion, the framework of analyses presented in this thesis showcases a fully automated, reproducible, scalable, and novel approach to explore the evolutionary-functional properties of uncharacterised genes from model and non-model organisms. This framework is deployed in specific test-case studies focusing on subsets of arthropod genes presented in Chapters 3 and 4. The assessments of function here rely on the Gene Ontology, i.e. knowledge-based functional classifications. Alternative functional classifications such as gene-expression-based clustering, as explored in the manuscript presented in Chapter 3, could offer interesting new perspectives. The scientific relevance and novelty of the work presented in this thesis originates from developing and applying the *Evol-Feat* toolkit together with the necessary conceptual framework. The outputs contribute to the goals of building bioinformatics comparative genomics solutions to tackle the challenges of comprehensively profiling and functionally interpreting the ever-growing amount of genomic data from a comparative and evolutionarily-informed perspective.

Chapter 3: Functional Constraints on Insect Immune System Components Govern Their Evolutionary Trajectories

Summary

The main research output of this doctoral thesis work was recently completed through the publication entitled "*Functional Constraints on Insect Immune System Components Govern Their Evolutionary Trajectories*" (Ruzzante et al. 2022), in the journal *Molecular Biology and Evolution*. The overarching goals of my thesis work centre on the question of whether quantitative characterisation of gene evolutionary histories can define distinct dynamics that are associated with different functional roles in biological systems. Using insect immunity as my test case system, I applied methods I developed as part of my main doctoral project to compute metrics that quantify gene evolutionary histories. These were employed to characterise evolutionary features of immune gene repertoires, and then to explore relationships between gene family evolutionary profiles and their roles in immunity to understand how different constraints may relate to distinct dynamics.

The multispecies comparative analyses identified evolutionary-functional correspondences suggesting that constraints on genes with similar or analogous functions govern their evolutionary trajectories. I identified three main axes of evolutionary trajectories characterised by gene duplication and synteny, maintenance/stability and sequence conservation, and loss and sequence divergence, highlighting similar and contrasting patterns across these axes amongst subsets of immune genes. The results provide the first multi-dimensional quantitative characterisation of gene evolutionary histories that support the conclusion that where and how different genes participate in immune defence responses limit the range of possible evolutionary scenarios that are tolerated by natural selection.

On starting this thesis work, the field as a whole was using only a handful of metrics to quantify gene evolutionary change dynamics: protein-level sequence divergence, DNA-level sequence divergence, and occasionally copy-number variation and population-level polymorphism (if data were available), and a framework for combined analyses was lacking. Through the extensive methods development part of my thesis work, I built a comprehensive analysis workflow to quantify a suite of metrics comprising 18 different evolutionary features, which I applied to this insect immunity study. Importantly, this further involved building analysis workflows to run and compare results from different clustering algorithms, including building creative data visualisation solutions, to be able to confirm the robustness of the results. The test case study system of insect immunity highlights the potential of applying the comparative genomics approaches that were developed to characterise how functional constraints on different components of other biological systems govern their evolutionary trajectories. The published article is accessible from the link on the next page and is attached in Appendix 1.

Appended Publication

Functional Constraints on Insect Immune System Components Govern Their Evolutionary Trajectories

Livio Ruzzante, Romain Feron, Maarten J.M.F. Reijnders, Antonin Thiebaut, and Robert M. Waterhouse *

Department of Ecology and Evolution, Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland

*Corresponding author: E-mail: robert.waterhouse@unil.ch.

Associate editor: Thomas Leitner

Published manuscript: <https://doi.org/10.1093/molbev/msab352>

Also available in the Appendix 1 section of this thesis.

Author Contributions

R.M.W. conceived the study. L.R., R.F., M.J.M.F.R., A.T., and R.M.W. designed the analyses. L.R. analyzed orthology data, quantified evolutionary features, and performed clustering analyses and statistical testing. R.F. built whole genome alignments, and analyzed alignment and variation data. M.R. analyzed variation data. A.T. curated data and assisted with gene expression data analysis. L.R. and R.M.W. wrote the manuscript with input from all authors. All authors read and approved the manuscript.

Chapter 4: Scientific Collaborations and Additional Resources

Summary

This chapter describes several collaborative scientific efforts, most of which led to successful publications in peer-reviewed journals or are currently being developed. Each section briefly introduces the research output, describes my specific contributions and highlights how such efforts strategically assisted the design and implementation of the *Evol-Feat* analysis workflow and the characterisation of arthropod evolutionary-functional correspondences. The research output includes scientific publications investigating the genome structures and the corresponding ecological adaptations of various arthropod lineages, from mosquitoes to hymenopterans. The technical contributions include the successful application of precursor versions of the *Evol-Feat* workflow and phylogeny reconstructions of arthropod clades. The last section of this chapter is dedicated to a detailed explanation of the phylogeny reconstruction of the 170 arthropod species used for the analyses in Chapters 1 and 2, supported by the development of the *Orthophile* workflow, a tool for constructing species phylogenies from thousands of orthologous genes.

Of Genes and Genomes: Mosquito Evolution and Diversity

Published in Trends in Parasitology (Ruzzante, Reijnders, and Waterhouse 2019), the review article explored and presented the available resources and applications supporting the research on mosquito evolution and diversity through comparative and functional genomics approaches. This work enabled me to familiarise myself with the “*mozomics*” (mosquito genomics) environment of species diversity, databases and comparative genomics software, a crucial starting point for the whole thesis work, especially for Chapter 3, which focuses on the mosquito immune repertoire. Among the diversity of tools and resources, the exploration of the NCBI Taxonomy (Schoch et al. 2020), the Mosquito Taxonomic Inventory (Harbach 2013), VectorBase (Giraldo-Calderón et al. 2015) and various mosquito genomics and transcriptomics data and data visualisation techniques. It represents a comprehensive overview of the available bioinformatics resources to help guide the research plans of biologists interested in the study of mosquito species, and more generally in the use of comparative genomics tools.

DOI: <https://doi.org/10.1016/j.pt.2018.10.003>

Evolutionary Superscaffolding and Chromosome Anchoring to Improve *Anopheles* Genome Assemblies

This collaboration, published in BMC Biology (Waterhouse et al. 2020), showcased the improvement of 21 *Anopheles* mosquito genome assemblies using consensus sets from 3 gene synteny-detection methods to increase the accuracy of scaffold adjacencies. With the abundance of available online genomic data, this work illustrates how better quality genome assemblies can be obtained with evolutionarily-informed techniques, avoiding additional sequencing costs. My contribution resulted in the development of custom programming scripts to process and measure synteny-conservation scores across mosquito genome

assembly files, which contributed to the development of the synteny-conservation features in the *Evol-Feat* workflow described in Chapter 1. This research collaboration helped me become familiar with genome assemblies and their structure, as well as with the General Feature Format (GFF) used to detail gene structural annotations. It constituted my first scripting effort using several programming languages including R, Python, and Bash to process and analyse genomic data, and was crucial for my understanding of gene sets and synteny-detection algorithms.

DOI: <https://doi.org/10.1186/s12915-019-0728-3>

Genome Sequence of the Wheat Stem Sawfly, *Cephus cinctus*, Representing an Early-Branching Lineage of the Hymenoptera, Illuminates Evolution of Hymenopteran Chemoreceptors

The publishing of the genome sequence of *Cephus cinctus* (Robertson et al. 2018) added a representative species from the Cephoidea superfamily of sawflies to the collection of high-quality hymenopteran genomes. Part of an early-branching lineage of Hymenoptera, the *C. cinctus* genome assembly allowed for a better understanding of the evolution of hymenopteran chemoreceptors through comparative genomics analyses. The results suggest that *C. cinctus* presents small lineage-specific chemoreceptor gene family expansions, possibly involved in molecular adaptations to agricultural wheat, for which it represents a major agricultural pest. The research article included my reconstruction of 17 arthropod species phylogeny and orthology profiling of the corresponding gene sets. This scientific collaboration was fundamental for my understanding and further development of phylogenetic reconstruction and orthology profiling techniques, crucial for the first part of Chapter 1 of this thesis work, including the computation of orthology- and phylogeny-related evolutionary features.

DOI: <https://doi.org/10.1093/gbe/evy232>

Draft Genome Assembly and Population Genetics of an Agricultural Pollinator, the Solitary Alkali Bee (Halictidae: *Nomia melanderi*)

The publishing of the high-quality draft genome assembly of the solitary bee *Nomia Melanderi* (Kapheim et al. 2019) constituted an opportunity to become further accustomed to whole-genome data and associated comparative genomics analyses. The research article uncovered previously uncharacterised transposable elements as the most abundant type of the bee's repetitive DNA. Findings of recent population bottlenecks and slower-evolving genes might serve as a basis for further understanding the evolutionary origins of eusociality in bees. My collaboration included the computation and presentation of different genome assembly quality statistics, including comparative tables reporting scaffold N50, genome lengths and BUSCO (Simão et al. 2015) scores for several hymenopteran species. It additionally included a phylogeny reconstruction for 15 arthropod species and measuring the single-copyness and taxonomic span of the corresponding gene sets. The algorithms for the computation of genome assembly statistics were collected in *Infoseq*, a publicly available repository at <https://github.com/laruzzante/infoseq>. Getting familiarised with such analyses, including implementing the quantification of taxonomic span and orthology profiling, resulted in a necessary further training for the phylogenetic reconstruction and orthology copy-number detection algorithms for the future development of the *Evol-Feat* workflow presented in Chapter 1.

DOI: <https://doi.org/10.1534/g3.118.200865>

Genus-Wide Characterization of Bumblebee Genomes Provides Insights into Their Evolution and Variation in Ecological and Behavioral Traits

The de-novo sequencing and assemblies of 17 species of bumblebees, representing all of the 15 subgenera, enabled the first genus-wide quantification and characterisation of *Bombus* diversity (Sun et al. 2021). Patterns of genomic

variation were associated with key ecological and behavioural traits of these global pollinators, including positive selection for foraging, diet, immunity and detoxification, lineage-specific social parasitism, and altitude adaptations. My contributions consisted of computing a subset of the evolutionary features described in Chapter 1 to identify the most dynamically evolving and conserved genes together with their corresponding functional enrichments, as described in the first sections of Chapter 2. This represents the first research application of the beta-version of the *Evol-Feat* workflow. Specific results derived from my analyses indicate that bumblebee orthologous groups with the broadest species representation are functionally enriched for core biological processes, while lineage-specific or sparse taxonomic spans are enriched for adaptive functions including olfactory and gustatory perception and detoxification. Lineage-specificity is additionally associated with higher sequence evolutionary rates and turnover dynamicity. These patterns largely confirm the observations described in Chapter 2's functional assessment of evolutionary trajectories.

DOI: <https://doi.org/10.1093/molbev/msaa240>

Anopheles Mosquitoes Reveal New Principles of 3D Genome Organization in Insects

Five *Anopheles* mosquito genomes were profiled to investigate the influence of chromosome organisation mechanisms on gene function and evolutionary dynamics (Lukyanchikova et al. 2022). Patterns observed from genome-wide chromatin interactions revealed associations with cytological structures, epigenetic profiles, and gene expression levels. My contribution consisted of computing and providing a phylogenetic species reconstruction of the *Anopheles* genus. This enabled me to further train my ability to build phylogenetic reconstructions and test different phylogenetic reconstruction software, with the ultimate goal of producing a significantly more challenging phylogeny of 170 arthropod species.

DOI: <https://doi.org/10.1038/s41467-022-29599-5>

The *Orthophile* Workflow and the Phylogenetic Reconstruction of the Arthropoda Phylum

The phylogenetic reconstruction of the 170 arthropod species, supporting the computation of most of Chapter 1's evolutionary features, required particular care and significant time investment. The main challenges at the origin of the phylogenetic reconstruction included a wide selection of substitution models and other parameters as well as the organisational and computational strategies and requirements to analyse a large dataset comprised of 231'710 gene sequences (from 1'363 universal orthologous genes per species). These were coupled with an overall large evolutionary timeframe and broad cross-species evolutionary distances and rate variations, stemming from a biased species sampling overrepresenting the *Anopheles* and *Drosophila* genera above all and complicating the inference of a correct phylogenetic reconstruction. Retrieving, pre-processing, aligning, and comparing the different outputs of various phylogenetic reconstruction approaches for such a dataset required the implementation of a specific workflow. I therefore developed the *Orthophile* workflow, a necessary step to advance the thesis work and its continued development constitutes a valuable research tool for computational biologists to easily obtain comprehensive phylogenetic reconstructions from thousands of orthologous genes. The phylogenetic reconstruction of the Arthropoda phylum is attached in Appendix 2, branch lengths are shown provided in units of million years along the branches, while bootstrap values from 0 (minimum support) to 1 (maximum support) are shown after node bifurcations.

The Snakemake workflow consists of a collection of Python scripts and Bash commands only requiring the user to input a list of species names or taxonomic identifiers (as defined in NCBI Taxonomy). The workflow then proceeds to send Application Programming Interface (API) queries to the *OrthoDB* server and fetches the sequences of orthologous genes in a user-specified output folder. The sequences are then automatically organised in specific subfolders, aligned, trimmed, and concatenated. Different parameters for the selection of orthologous groups can be tuned, including *OrthoDB*'s orthology delineation level (e.g. Eukaryota, Metazoa, Vertebrata), the

percentage single-copyness, and percentage species-span. The resulting phylogenetic reconstruction is then computed by either RAxML (for increased precision) or FastTree (for increased performance), but additional software can be included by tweaking the workflow's rules and Conda environments. Furthermore, parts of the workflow can be launched without satisfying prior input requirements, as long as the name formatting is consistent with the workflow's rules. For example, custom gene sequences can be inserted into the user-specified *OrthoDB* fetching folder, and the workflow will include those in the phylogenetic reconstruction analysis. This specific case is a valid workaround for when species of interest are not included in *OrthoDB*, but the user is still interested in maintaining the overall framework of analyses provided by *Orthophile*. The following paragraphs will detail the reconstruction process to obtain the arthropod species phylogeny used in Chapters 1 and 2 by describing the initial output of *Orthophile* and the subsequent necessary manual curation.

The protein sequences from 170 arthropod genomes were aligned with *muscle* version 3.8 (Edgar 2004) and trimmed with *trimAl*'s version 1.4 (Capella-Gutierrez, Silla-Martinez, and Gabaldon 2009) *-strictplus* option (optimised for Neighbour Joining phylogenetic tree reconstruction). A first phylogenetic species tree was reconstructed with *IQ-TREE* (Minh et al. 2020) and the *msub -nuclear* option, which selected the *LG+F+R10* (*LG* = nuclear general matrix (Le and Gascuel 2008), *+F* = empirical amino acid frequencies from the data, *+R10* = *FreeRate* model (Yang 1995; Soubrier et al. 2012) that generalises the *+G* model by relaxing the assumption of Gamma-distributed rates. The *FreeRate* model typically fits data better than the *+G* model and is recommended for the analysis of large datasets) as the best possible model among the generic amino acid substitution models. This combination of software and parameter selection provided the best possible reconstruction compared with the newest arthropod phylogeny literature, as other approaches failed to either group non-insect hexapods in a monophyletic group or to place daphnia among crustaceans. Nevertheless, although the overall placement of species seemed largely correct, a few species, such as the body louse and thrips, were not correctly placed. As many evolutionary features would subsequently be computed on reconstructed events and evolutionary distances inferred from the phylogeny, it was crucial to obtain its most accurate reconstruction.

Lastly, this was achieved by manually adjusting the root placement of the branches and following the desired topology with *TreeView* (Page 2003), creating a subsequent alignment constraint file with a custom script provided by *FastTree*, and re-running the phylogeny reconstruction using *FastTree* version 2.1 (Price, Dehal, and Arkin 2010), which allows for an input tree and a constraint file to search for possible phylogenetic reconstruction reconciliations. The resulting tree correctly placed all the species and major clades except for the missing separation of diplurans and springtails. The molecular tree file was re-rooted at the Chelicerata outgroup with the *Newick Utilities* (Junier and Zdobnov 2010) and time-calibrated with the *chronos* function from the R package *ape* version 5.0 (Paradis and Schliep 2019). The time calibration, which allows for distributing the sequence evolution across a specified time range, effectively converting a molecular tree into a time tree (in units of millions of years), was obtained by specifying 11 evolutionary distances ranging from more closely related species (e.g. the honeybee *Apis mellifera* and the common eastern bumblebee *Bombus impatiens*) to more distant ones (e.g. the centipede *Strigamia maritima* and the common fruit fly *Drosophila melanogaster*). The cross-species divergence times were obtained from the *TimeTree* (Kumar et al. 2017) online knowledge base.

Chapter 5: Conclusions and Perspectives

Early research in the field preceding this thesis work highlighted several principles of how patterns of gene and gene family evolution relate to their functional properties (Jordan et al. 2002; Krylov 2003; Wolf, Carmel, and Koonin 2006). Subsequent studies pioneered approaches to further quantify and explore these patterns using phylostratigraphy-based metrics (Domazet-Lošo, Brajković, and Tautz 2007), measures of gene copy-number dynamics and protein sequence divergence (Waterhouse, Zdobnov, and Kriventseva 2011), also adding synteny and protein domain compositions (Linard et al. 2012), or focusing only on covariations of protein sequence divergence (Clark, Alani, and Aquadro 2012), or employing presence-absence matrices with tree-based (Li et al. 2014) or non-tree-based (Cheng and Perocchi 2015) profiling. These studies built a foundation but remained limited in several key aspects: they usually examined only a few evolutionary features at a time; they often focused on pairwise correlation analyses; they usually explored a single dataset with a fixed evolutionary span (e.g. all available eukaryotes); the early studies especially were limited in the number of species they could use; their use of multidimensional data analysis and clustering approaches was seemingly *ad hoc*, they did not provide tools and/or detailed-enough methods descriptions to be able to reproduce their results or redeploy their methods on new datasets; those using presence-absence matrices ignored signals from gene duplication patterns; and they often did not fully incorporate the species phylogeny, i.e. phylogenetic relatedness, into their analyses.

The work undertaken in this thesis was therefore driven by the motivations to answer four main questions: 1) Can we build a suite of evolutionary feature quantification methods that harmonises and extends the previous work in the field, with clearly defined metrics that can take advantage of using a variety of types of data? 2) Can we comprehensively examine the main multidimensional data analysis and clustering techniques to determine the most robust approaches to use for exploring the patterns and profiles derived from the quantified gene and gene family evolutionary features? 3) Can we build an analysis framework that allows others to apply these evolutionary feature

quantifications to their own datasets and to the downstream steps of clustering and dataset exploration? And finally, 4) can we combine the results from evolutionary feature quantification and profiling with gene and gene family functional information to investigate evolutionary-functional correspondences with an enhanced resolution, using the diverse phylum of Arthropoda as our study system? Practical outputs thus focused on: 1) providing experimental biologists with a readily-available resource where thousands of arthropod genes from model and non-model species alike can be explored in the context of their evolutionary-functional correspondences; and 2) providing computational biologists with a tool to capture and compare the evolutionary features of genes to support evolutionary-functional hypotheses on custom data.

This chapter presents a brief summary of the principal conclusions from the thesis work in the context of the current data and analysis framework. It then explores the potential implications for the field, summarising the perspectives gained and future outlook in the context of using multi-species genomic datasets to further our understanding of gene evolutionary-functional correspondences.

Summary of the Principal Conclusions

The work presented in Chapter 1 firstly demonstrates how comparative genomics datasets can be assessed using the *Evol-Feat* workflow to quantify a suite of metrics that capture different gene evolutionary characteristics from large collections of cross-species repertoires of orthologous genes. Comprehensive analysis of feature correlations, pairwise and multidimensional, confirm previous trends and expand on the conclusions from early exploratory studies on gene evolutionary features. Multivariate analysis of the evolutionary features dataset from 170 arthropod species identifies a first axis of gene metric distributions capturing old, universal, highly conserved and physiologically essential genes. The second axis captures lineage-specific adaptations through copy-number increases and subsequent negative purifying selection. The third axis of arthropod gene evolution is driven by synteny conservation scores. The minor fourth and fifth axes capture, respectively, lineage-specific gene expansions and gene losses.

Lineage-specific gene family expansions, likely associated with relatively recent adaptations, are observed to recruit young genes as well as ancestral genes characterised by dynamic gene turnover, in line with previous observations from phylostratigraphy analyses on the evolutionary histories of *D. melanogaster* genes. Widespread, universal, and old genes are instead characterised by slow gene turnover and are distinctly separated from younger, lineage-specific genes. These are further associated with either high or low potential for duplicability, in line with and expanding previous descriptions of “single-copy control” versus “multicopy licence” modes of gene evolution and likely indicators of essential, ancestral, and housekeeping functions versus more recent lineage-specific adaptations. The distributions of sequence evolutionary rates follow more subtle evolutionary trajectories than previously observed in lineage-specific or case-study investigations, e.g. mosquito immunity genes. Multidimensional analysis of the evolutionary features highlighted how the orthologous group-averaged evolutionary rates of protein sequence divergence is a much stronger driver of genetic repertoire diversity in the *Anopheles* immunity case-study when compared to the results from the the entire 170 arthropod

species dataset. Sequence evolutionary rate is therefore considered a more relevant variable for comparative analyses within specific lineages and recent evolutionary timescales, while losing its applicability amid the variability of larger, complete gene repertoires, and more distantly-related species. At the broader Arthropoda scale, fast versus slow gene sequence evolution is found to be less tightly associated with age and universality but more with the distribution of gene losses and family expansions. The confirmation of several hypotheses on evolutionary feature distributions and relationships from previous studies that employed considerably smaller datasets, coupled with the uncovering of novel observations and trajectories, supports the translatability of evolutionary hypotheses based on a few eukaryotic species to large-scale multi-species genome studies, here using arthropods as the study system. In parallel with a greater taxonomic resolution, Chapter 1 brings a theoretical basis for quantification in arthropod comparative genomics research studies, while providing an open, modifiable, and scalable bioinformatics workflow enabling further explorations of evolutionary-functional correspondences on custom data.

Chapter 2 takes this further to firstly demonstrate that different evolutionary features are associated with different functional annotation enrichments. Using the suites of computed features, each arthropod orthologous group was associated with a specific evolutionary profile. Clustering algorithms were successively implemented to bring together and partition genes into sets of distinct evolutionary trajectories, at high- and low- resolution scales, capturing both broad and specific patterns of evolutionary-functional correspondences. Coupling the clusters of evolutionary profiles with corresponding Gene Ontology (GO) annotations, Chapter 2 describes the statistically significant functional enrichments associating particular evolutionary trajectories (or strategies) with identified sets of biological processes. Results from this chapter show how evolutionarily-similar genes can exhibit similar or analogous functions, supporting the hypothesis that functional constraints impact the range of possible arthropod gene evolutionary trajectories. Among the fine- and broad-scale described evolutionary correspondences, examples include universally widespread genes with lineage-specific losses and low copy-numbers being partitioned into two categories: slow-evolving genes are enriched with functions related to organism development and anatomical organisation, while

the fast-evolving ones are associated with the immunity-related regulation of antibacterial response. Highly syntenic and young orthologous groups with low copy-numbers and lineage-specific all-or-nothing gene losses are associated with the development of chaetae, auditory behaviour, sound perception, and response to plant-defence mechanisms.

Response to classes of insecticides and phytohormones are associated with young, slow-evolving orthologous groups with low copy-numbers. Young, fast-evolving orthologous groups defined by lineage-specific gene expansions and high duplicability are associated with Malpighian tubule development and activity, and the metabolic detoxification of plant defence mechanisms. Functional enrichments of higher-resolution subsets of similar evolutionary profiles point to chemoreception and chemosensation. Contrastingly, old and widespread orthologous groups defined by expansions and high copy-numbers with intermediate evolutionary rates recover a range of biological functions, from reproduction to immunity, neuronal development, and insecticide catabolism. Characterising only relatively few orthologous groups, such evolutionary trajectories are likely associated with specific functional modules, adopted across several organismal biological processes and subjected to similar selection pressures. Developmental processes and housekeeping functions such as the maintenance of cellular structure and components are associated with old, slow-evolving orthologous groups with low copy-numbers and low potential for gene family expansions. Similar evolutionary trajectories but with high copy-number counts and duplicability potential are enriched for various mechanisms of reaction to different environmental pressures including: cold, chemicals, oxidative stress, auditory stimulus, bacteria, and inflammatory response in general. Haemocyte differentiation, cytokine production, and response to bacteria are further enriched in old orthologous groups characterised by the highest gene stability scores, widespread taxonomic representation, and low sequence evolutionary rates. These are further enriched with a range of developmental processes related to the regulation of arthropod moulting and metamorphosis. Such evolutionary trajectories recover even more ancestral and conserved biological functions, where immunity-related and development-related genes likely point to known pleiotropic properties.

The framework of analyses presented in this thesis showcases a fully automated, reproducible, scalable, and novel approach to explore the evolutionary-functional properties of uncharacterised genes from model and non-model organisms. This framework has been successfully deployed in specific test-case studies focusing on subsets of arthropod genes presented in Chapters 3 and 4. The research output of these chapters includes scientific publications investigating the genome structures and the corresponding ecological adaptations of several arthropod lineages. Describing the evolutionary-functional correspondences enabled the formulation of hypotheses on gene family evolution and adaptation in various organisms, from mosquitoes to hymenopterans.

In Chapter 3, focussing on the insect immune repertoire, three main axes of evolutionary trajectories are identified, these are driven by gene duplication and synteny, maintenance/stability and sequence conservation, and loss and sequence divergence, highlighting similar and contrasting patterns across these axes amongst subsets of immune genes. Using a comprehensive taxonomic selection of insect genomes for the quantification of the evolutionary features, the observations are detailed, as a proof-of-concept, for the two species with the most experimentally described immune gene functional information (i.e. the most experimental work implicating genes in different immune responses): the African malaria mosquito *Anopheles gambiae*, and the common fruit fly *Drosophila melanogaster*. Characterising patterns of genomic change in species where putative functions and interactions of system components are relatively well described allowed us to explore whether genes with similar roles exhibit similar evolutionary trajectories. The results suggest that where and how genes participate in immune responses limit the range of possible evolutionary scenarios they exhibit, associating known functional constraints with the diversification of the insect immune-related gene families. The results from our test-case study system of insect immunity highlight the applicability of comparative genomics approaches for characterising how functional constraints on different components of biological systems might govern their evolutionary trajectories.

Chapter 4 describes several successful collaborative scientific efforts using the *Evol-Feat* framework of analysis for a detailed description of arthropod genes

characterised by their evolutionary features and putative functional roles associated with them. Employed on genes annotated from the genome assembly of the wheat stem sawfly *Cephus cinctus*, the *Evol-Feat* framework supported the comparative genomics analyses for a better understanding of the evolution of hymenopteran chemoreceptors. The results show that *C. cinctus* has representatives for most conserved and expanded chemosensory gene lineages amongst bees, wasps, and ants (Apocrita). It also maintains several lineages that have been lost from the Apocrita, most notably the carbon dioxide receptor subfamily. The family analyses also show that *C. cinctus* presents small lineage-specific chemoreceptor gene family expansions that might be involved in adaptations to grasses including wheat. The *Evol-Feat* workflow was also employed in a research study comparing bumblebee genomes representing all 15 subgenera of *Bombus*. The bumblebee genes from orthologous groups with the broadest species representation are found to be functionally enriched for core biological processes, while lineage-specific or sparse taxonomic spans are enriched for adaptive functions including olfactory and gustatory perception and detoxification. Lineage-specificity is additionally associated with higher sequence evolutionary rates and turnover dynamicity. Other interesting patterns included gene evolutionary features hinting at adaptations to recognise flowers, to better cope with life at high altitudes, and to being able to forage scarce food sources over long distances, as well as revealing that bumblebees have fewer genes involved in detoxification of pesticides or defence against pathogens in comparison to many other insects. These and other studies presented in Chapter 4 served as opportunities for the practical development of components of the *Evol-Feat* workflow as well as for building an understanding of the needs of biologists interested in incorporating evolutionary analyses into their research.

Implications, Perspectives, and Future Outlook

As the field of genomics continues to advance, the amount of data available to researchers is rapidly increasing. This influx has the potential to greatly improve our understanding of gene function and underlying mechanisms of evolution and adaptation. One of the key benefits of accessible genomics data is the ability to use large-scale datasets to identify genetic associations with specific traits including morphological, physiological, behavioural, and ecological adaptations (Nagy et al. 2020). This can be accomplished through the use of computational tools that can analyse large amounts of genomic data to identify patterns, associations, and correlations. For example, illuminating genetic variations that are linked to disease susceptibility or adaptations to specific niches. Another benefit of data from new genomics technologies is the ability to use functional genomics approaches to improve understanding of gene function (Feder and Mitchell-Olds 2003; Walton, Sheehan, and Toth 2020). By combining available functional genomics data with comparative genomics data (e.g. orthology), researchers can develop computational models that enhance the transfer of gene functional information from well-studied organisms to poorly characterised species, providing a valuable tool for propagating knowledge and investigating biological processes across the tree of life (Gabaldón and Koonin 2013). Furthermore, accumulating genomics data can be used to improve our understanding of evolutionary processes at increasingly detailed levels of resolution. By studying large-scale genomic datasets, researchers can identify the genetic changes that have occurred over time, providing valuable insights into the mechanisms of genome evolution (Koonin 2011) and the role of genetic variation in shaping the diversity of life on Earth.

The research reported in this thesis leverages this potential of increasingly accessible data by building a comparative genomics framework to quantify genetic changes over evolutionary time in order to improve understanding of gene function and evolution. A major implication for the field as a whole is how the *Evol-Feat* framework offers new possibilities to refine how the community grapples with the task of accurately transferring functional annotations from well-studied species to others. The multiple metrics offer a multidimensional

quantification of gene family evolutionary rates, delineating spectra that identify “slow and stable” orthologous groups where we can more confidently transfer annotations, and at the same time helping to distinguish “fast and dynamic” orthologous groups for which the transfer of functional information needs to be more tentative. As recognised by Eugene V. Koonin, “*The huge majority of genes in the sequenced genomes will never be studied experimentally, so for most genomes transfer of functional information between orthologs is the only means of detailed functional characterization.*” (Koonin 2005). The process of transfer is by no means standardised across the field, and there is no obvious one-size-fits-all solution, however, the quantifications provided by the *Evol-Feat* framework could be employed to bring a level of objectivity to future standardisation efforts.

A further important implication for the field is that this framework allows for the building of hypotheses on putative gene functional roles in a manner that does not rely on sequence homology alone. Demonstrating that genes exhibiting similar evolutionary profiles can be linked to similar or analogous functions means that clustering genes by their quantified evolutionary dynamics instead of their sequence similarities can provide an alternative means for the tentative transfer of functional information. This concept is similar to the approaches of using presence/absence genomic phylostratigraphy to identify genes that themselves are not homologous but which function together as members of physically interacting protein complexes. Instead of only considering presence/absence matrices, the evolutionary profiles can be built using a suite of features quantified by *Evol-Feat* and which are thus able to capture co-evolutionary patterns that could point to potential functional dependencies, similarities, or analogies.

An additional key implication, especially for researchers using functional genomics to identify candidate gene sets of interest, is how the *Evol-Feat* framework allows for the investigation of the evolutionary features of sets of candidate genes. For example, gene lists from studying organismal responses to various treatments and/or from examining life cycle progressions are typically interrogated using GO term enrichment analysis to understand what is functionally “special” about a particular set of genes. Going beyond functional

properties, the suite of evolutionary features quantified for each gene provides possibilities for researchers to also ask what is evolutionarily “special” about their set of genes. For example, exploratory analyses with honeybee viral infection data indicated that early-response genes were significantly enriched for genes from “fast and dynamic” orthologous groups while late-response up-regulated genes were evolutionarily “slow and stable”.

Beyond the domain of comparative genomics for investigating gene evolution and function, there are broader implications for enhancing the ability to use genomic data to better understand organismal biology, here focussing on arthropods. Understanding and cataloguing the evolutionary landscapes of gene families can assist, for example, in more precisely predicting the effectiveness of agro-ecological phytotherapeutics, the immune response to pathogens across species, or the ecological adaptation possibilities in species threatened by rapid climate change. More specifically, rapidly expanding research and industrial applications such as artificial gene-drive technologies, very promising in the fields of pest control for disease spread limitation and crop protection, can greatly benefit from complementary evolutionary-informed analyses. With the purpose of disrupting populations of disease vectors and pests, gene-drive models may be assisted in the quest for identifying ideal gene candidates by detailed knowledge of specific gene repertoire’s evolutionary features and putative functions. Gene essentiality and turnover, sequence conservation, evolutionary rates, population-level sequence variations, clade-specificities, and copy-number counts, coupled with hypotheses on the biological processes associated with their respective evolutionary modules, can expand the search to as-yet uncharacterised genes, while directing it towards ranges of required evolutionary features. These include, for example, targeting genes with low potential to evolve and adapt, involved in physiologically essential processes with low/high lineage-specificity, high resistance to sequence change, or varying degrees of taxonomic reach, all of which can increase the efficiency of gene-drive model designs and aid in predicting and controlling their impact on animal populations.

Moreover, facing global threats on ecosystems and biodiversity caused by rapid climate change and habitat disruption from human interference requires

understanding the genetic mechanisms and properties underlying the potential for ecological adaptations across a wide range of species, especially non-model organisms. The mapping across taxa, through comparative genomics, of genetic modules associated with known adaptive traits can reveal or generate hypotheses about the genetic plasticity and limitations of the response of understudied organisms to shifting climatic and environmental conditions. Species dispersion and habitat occupation models needed to guide biodiversity conservation policies can greatly benefit from such complementary evolutionarily-informed hypotheses. These include ranges of tolerance to changing temperatures, resistance to altitude, air salinity, drought, seasonality disruption, and the immune response adaptability to new pathogen exposures. Such ecological-functional properties, even when lacking precise molecular laboratory work within the species of interest, may be predicted via the propagation of functional annotations from existing studies in well-described species to under-studied species and lineages, through delineated modules of gene evolutionary histories and features.

In summary, the increasing accessibility of genomics data can greatly enhance our understanding of gene function and the underlying mechanisms of evolutionary adaptation. By using the latest computational tools and approaches, researchers can thus harness the power of large-scale genomic datasets to improve our understanding of the genetic basis of biological processes and develop new strategies for a wide range of animal and crop disease prevention, pest control, and biodiversity conservation policy making. Nevertheless, gathering functional information on uncharacterised genes, especially from non-model organisms often at the centre of agroecological and epidemiological societal challenges, still requires costly detailed laboratory work. As a result, when investigating under-studied organisms, gene functional information is usually transferred onto undescribed genes from sequence homology with genes of well-described model organisms. Although successful in specific cases of highly conserved gene sequences and functions, this approach leads to discrepancies and disadvantages when describing the functions of gene repertoires from under-represented or distant taxonomic clades. Therefore, efforts to better understand the biological processes associated with the astounding diversity of arthropod adaptations and the evolution of their genetic

sequences lack specific resources and tools. The main purpose of the *Evol-Feat* framework of analysis developed in this thesis work is to serve as a practical example on how to face the complex challenge of interpreting the rapidly accumulating amount of genomic data. In this case, the workflow is conceived to exploit the, so-far relatively untapped, higher-resolution evolutionary-functional correspondences harnessed with large-scale comparative-evolutionary analyses from already available data collections, through automated and scalable bioinformatics solutions.

The development of the *Evol-Feat* framework through a modifiable, open-source, and scalable bioinformatics workflow, means that these new approaches can be further used for exploring evolutionary-functional correspondences beyond arthropods. Such applications were beyond the scope of the thesis work, but investing time and effort into the workflow design means that the resulting tools offer the possibility to quantify evolutionary trajectories and functional enrichments of evolutionarily-similar clusters of genes from user-defined custom datasets. Throughout the development of the analysis tools required to carry out the objectives of the thesis a large amount of exploratory work was required to investigate the appropriateness of various approaches to quantify evolutionary features, build reliable profiles, sanity-check the results, and investigate evolutionary-functional correspondences. The major perspectives gained from these efforts can be summarised by three main recurring issues: constraints imposed by the reliance on a species phylogeny; the inherent limitations of understanding lineage-specific patterns in lineages with little or no functional data; and challenges to understand clustering robustness. These recurring issues are not specific to this thesis work, and they remain challenges for the field as a whole to consider.

With respect to species phylogenies, the rapidly accumulating genomics data mean that the achievable species resolution is also increasing dramatically. However, this usually makes the reconstruction of robust species phylogenies even more challenging, with many uncertain branching patterns and conflicts with described taxonomies. Many of the evolutionary features we wish to quantify inherently rely on a species tree, e.g. the age of an orthologous group, and in particular the quantifications that require ancestral state reconstructions

across a phylogeny, e.g. gene gains and/or losses. It is therefore increasingly clear that for progress to continue to be made, the field as a whole needs to consider the development and implementation of methods that are less reliant on a fixed species phylogeny but instead can operate by taking species and/or gene tree uncertainty into account.

Another recurring issue that highlights a general perspective in comparative genomics applications is the observation that lineage-specific evolutionary patterns are usually challenging or near-impossible to relate to functional properties. This is because the majority of knowledge on gene functions comes from detailed studies in a few species, meaning that for many lineages there are no well-studied representative species from which to derive some functional hints. This outlook is changing as genome resources and functional genomics experiments are expanding possibilities for developing new arthropod model systems (Feron and Waterhouse 2022b). Nevertheless, obtaining relevant lineage-specific functional data will continue to lag behind the generation of reference genomes and therefore these limitations will persist. Importantly, delineating lineage-specific evolutionary patterns highlights trends that could help prioritise functional investigations, and in so doing the identification of gaps serves to start addressing those gaps.

Regarding clustering and its robustness, the work for this thesis required the exploration of an array of different approaches to assess their appropriateness and effectiveness. Because the quantified evolutionary features produce results that exhibit highly heterogeneous and non-normal distributions, assumptions for most standard clustering techniques are violated. Combining dimensionality reduction techniques with traditional clustering methods did not yield results that could be easily represented in human-interpretable dimensional spaces. Instead, profile clustering onto eigengenes emerged as the most effective solution. The eigengenes represent an ensemble of possible evolutionary feature profiles onto which each orthologous group can be mapped based on the similarity to its own profile. This proved not only effective but also flexible in terms of being able to apply this approach with variable numbers of evolutionary features and orthologous groups. This therefore offers a new perspective for the field to consider further as we aim to explore multiple

dimensions of gene evolutionary histories, with different densities of species sampling, and across longer or shorter evolutionary timescales.

Recognising the principal implications for the field and the key perspectives gained from the thesis work, the future outlook can be summarised with respect to the main questions that guided the thesis objectives. Firstly, the developed framework harmonised and extended previous work in the field to define a suite of evolutionary feature metrics: it is clear that some metrics are more challenging to obtain than others (e.g. constraint from whole genome alignments, or nucleotide polymorphisms from population genomics samples), and that additional new metrics per gene or per orthologous group could be defined (e.g. protein domain content/complexity, or orthologous group homology uniqueness). While steps for obtaining gain/loss metrics were fully integrated into *Evol-Feat*, other more challenging metrics rely on obtaining additional data or using tools/workflows developed by others. Notably, the framework was designed to incorporate additional user-provided metrics, and it provides the exploratory tools needed to examine how such additional metrics contribute to the evolutionary profiling, therefore such extensions in the future should be entirely feasible. Secondly, although the exploration of options for multidimensional data analysis and clustering techniques was extensive, new or different combinations of methods might be useful to examine in the future. Profile mapping onto eigengenes proved successful and it seems to be a suitably extensible solution to use for exploring the patterns and profiles derived from the quantified gene and gene family evolutionary features. Nevertheless, if new metrics are added and/or very different combinations of metrics are employed then this approach would need to be carefully re-evaluated. Thirdly, the future applications of the framework to different taxonomic datasets or other biological case studies (development, insecticide resistance, detoxification, etc.) would serve to further test its generalisability. This is made possible by the fact that the analysis framework was designed and built in a way that allows others to apply it to their own datasets and to downstream dataset exploration and clustering steps. The design also means that future applications can employ the minimal-input orthology data, or additional inputs for feature quantifications, according to the data available to the user. This will greatly facilitate the future use of and possible development/extension of the *Evol-Feat* workflow with a

wide range of possible applications. Fourthly, initial sanity checks and subsequent investigations of evolutionary-functional correspondences focused on combining the results from evolutionary feature quantification and profiling with gene functional information in the form of GO term annotations. In the mosquito immunity case study this was extended to examine correspondences using functional categorisations based instead on gene co-expression analyses to identify immune families that function in concert. For other taxa with accumulating transcriptomics data, future studies will be able to use the mosquito case study as a template for the co-interrogation of gene evolution and function with an enhanced resolution. Finally, in terms of the practical output objectives: the prototyped browsable interactive tools that would provide experimental biologists with the ability to explore thousands of arthropod genes, the context of their evolutionary-functional correspondences could be further developed in the future. The tools for computational biologists to quantify and analyse gene evolutionary features and functional properties on custom data will greatly facilitate the use of multi-species genomic datasets to further our understanding of gene evolutionary-functional correspondences.

Bibliography

- Adams, M. D. 2000. 'The Genome Sequence of *Drosophila Melanogaster*'. *Science* 287 (5461): 2185–95. <https://doi.org/10.1126/science.287.5461.2185>.
- Alexa, Adrian, and Jorg Rahnenfuhrer. 2020. 'TopGO: Enrichment Analysis for Gene Ontology'. Bioconductor. <https://doi.org/10.18129/B9.BIOC.TOPGO>.
- Altenhoff, Adrian M., Romain A. Studer, Marc Robinson-Rechavi, and Christophe Dessimoz. 2012. 'Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs'. *PLOS Computational Biology* 8 (5): e1002514. <https://doi.org/10.1371/journal.pcbi.1002514>.
- Anselmetti, Yoann, Vincent Berry, Cedric Chauve, Annie Chateau, Eric Tannier, and Sèverine Bérard. 2015. 'Ancestral Gene Synteny Reconstruction Improves Extant Species Scaffolding'. *BMC Genomics* 16 (Suppl 10): S11. <https://doi.org/10.1186/1471-2164-16-S10-S11>.
- Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. 'Gene Ontology: Tool for the Unification of Biology'. *Nature Genetics* 25 (1): 25–29. <https://doi.org/10.1038/75556>.
- Bastian, Frederic, Gilles Parmentier, Julien Roux, Sebastien Moretti, Vincent Laudet, and Marc Robinson-Rechavi. 2008. 'Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species'. In *Data Integration in the Life Sciences*, edited by Amos Bairoch, Sarah Cohen-Boulakia, and Christine Froidevaux, 124–31. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-69828-9_12.
- Bhattacharya, Ramcharan, Murali Krishna Koramutla, Manisha Negi, Gregory Pearce, and Clarence A. Ryan. 2013. 'Hydroxyproline-Rich Glycopeptide Signals in Potato Elicit Signalling Associated with Defense against Insects and Pathogens'. *Plant Science: An International Journal of Experimental Plant Biology* 207 (June): 88–97. <https://doi.org/10.1016/j.plantsci.2013.03.002>.
- Borkowska, Edyta M, Andrzej Kruk, Adam Jedrzejczyk, Marek Rozniecki,

- Zbigniew Jablonowski, Magdalena Traczyk, Maria Constantinou, et al. 2014. 'Molecular Subtyping of Bladder Cancer Using Kohonen Self-Organizing Maps'. *Cancer Medicine* 3 (5): 1225–34. <https://doi.org/10.1002/cam4.217>.
- Bunt, Stephanie, Clare Hooley, Nan Hu, Catherine Scahill, Helen Weavers, and Helen Skaer. 2010. 'Hemocyte-Secreted Type IV Collagen Enhances BMP Signaling to Guide Renal Tubule Morphogenesis in *Drosophila*'. *Developmental Cell* 19 (2): 296–306. <https://doi.org/10.1016/j.devcel.2010.07.019>.
- Capella-Gutierrez, S., J. M. Silla-Martinez, and T. Gabaldon. 2009. 'TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses'. *Bioinformatics* 25 (15): 1972–73. <https://doi.org/10.1093/bioinformatics/btp348>.
- Charles, J. P. 2010. 'The Regulation of Expression of Insect Cuticle Protein Genes'. *Insect Biochemistry and Molecular Biology, Insect Cuticle*, 40 (3): 205–13. <https://doi.org/10.1016/j.ibmb.2009.12.005>.
- Cheng, Yiming, and Fabiana Perocchi. 2015. 'ProtPhylo: Identification of Protein–Phenotype and Protein–Protein Functional Associations via Phylogenetic Profiling'. *Nucleic Acids Research* 43 (W1): W160–68. <https://doi.org/10.1093/nar/gkv455>.
- Clark, Erik, Andrew D. Peel, and Michael Akam. 2019. 'Arthropod Segmentation'. *Development* 146 (18): dev170480. <https://doi.org/10.1242/dev.170480>.
- Clark, Nathan L., Eric Alani, and Charles F. Aquadro. 2012. 'Evolutionary Rate Covariation Reveals Shared Functionality and Coexpression of Genes'. *Genome Research* 22 (4): 714–20. <https://doi.org/10.1101/gr.132647.111>.
- Domazet-Lošo, Tomislav, Josip Brajković, and Diethard Tautz. 2007. 'A Phylostratigraphy Approach to Uncover the Genomic History of Major Adaptations in Metazoan Lineages'. *Trends in Genetics* 23 (11): 533–39. <https://doi.org/10.1016/j.tig.2007.08.014>.
- Edgar, Robert C. 2004. 'MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput'. *Nucleic Acids Research* 32 (5): 1792–97. <https://doi.org/10.1093/nar/gkh340>.
- Elsik, Christine G., Aditi Tayal, Deepak R. Unni, Gregory W. Burns, and Darren E. Hagen. 2018. 'Hymenoptera Genome Database: Using HymenopteraMine

- to Enhance Genomic Studies of Hymenopteran Insects'. In *Eukaryotic Genomic Databases*, edited by Martin Kollmar, 1757:513–56. *Methods in Molecular Biology*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-7737-6_17.
- Emms, David M., and Steven Kelly. 2019. 'OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics'. *Genome Biology* 20 (1): 238. <https://doi.org/10.1186/s13059-019-1832-y>.
- Feder, Martin E., and Thomas Mitchell-Olds. 2003. 'Evolutionary and Ecological Functional Genomics'. *Nature Reviews Genetics* 4 (8): 649–55. <https://doi.org/10.1038/nrg1128>.
- Feron, Romain, and Robert M Waterhouse. 2022a. 'Assessing Species Coverage and Assembly Quality of Rapidly Accumulating Sequenced Genomes'. *GigaScience* 11 (January): giac006. <https://doi.org/10.1093/gigascience/giac006>.
- Feron, Romain, and Robert M. Waterhouse. 2022b. 'Exploring New Genomic Territories with Emerging Model Insects'. *Current Opinion in Insect Science* 51 (June): 100902. <https://doi.org/10.1016/j.cois.2022.100902>.
- Fitch, Walter M. 1970. 'Distinguishing Homologous from Analogous Proteins'. *Systematic Biology* 19 (2): 99–113. <https://doi.org/10.2307/2412448>.
- Gabaldón, Toni, and Eugene V. Koonin. 2013. 'Functional and Evolutionary Implications of Gene Orthology'. *Nature Reviews Genetics* 14 (5): 360–66. <https://doi.org/10.1038/nrg3456>.
- Gene Ontology Consortium, The. 2019. 'The Gene Ontology Resource: 20 Years and Still GOing Strong'. *Nucleic Acids Research* 47 (D1): D330–38. <https://doi.org/10.1093/nar/gky1055>.
- Giraldo-Calderón, Gloria I., Scott J. Emrich, Robert M. MacCallum, Gareth Maslen, Emmanuel Dialynas, Pantelis Topalis, Nicholas Ho, et al. 2015. 'VectorBase: An Updated Bioinformatics Resource for Invertebrate Vectors and Other Organisms Related with Human Diseases'. *Nucleic Acids Research* 43 (D1): D707–13. <https://doi.org/10.1093/nar/gku1117>.
- Giribet, Gonzalo, and Gregory D. Edgecombe. 2019. 'The Phylogeny and Evolutionary History of Arthropods'. *Current Biology* 29 (12): R592–602. <https://doi.org/10.1016/j.cub.2019.04.057>.
- Gori, Kevin, Tomasz Suchan, Nadir Alvarez, Nick Goldman, and Christophe Dessimoz. 2016. 'Clustering Genes of Common Evolutionary History'.

- Molecular Biology and Evolution* 33 (6): 1590–1605.
<https://doi.org/10.1093/molbev/msw038>.
- Gregory, T. Ryan, James A. Nicol, Heidi Tamm, Bellis Kullman, Kaur Kullman, Ilia J. Leitch, Brian G. Murray, Donald F. Kapraun, Johann Greilhuber, and Michael D. Bennett. 2007. 'Eukaryotic Genome Size Databases'. *Nucleic Acids Research* 35 (suppl_1): D332–38.
<https://doi.org/10.1093/nar/gkl828>.
- Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. 'Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data'. *Bioinformatics* 32 (18): 2847–49.
<https://doi.org/10.1093/bioinformatics/btw313>.
- Harbach, R. E. 2013. 'Mosquito Taxonomic Inventory'.
- Harrison, Rhett D., Christian Thierfelder, Frédéric Baudron, Peter Chinwada, Charles Midega, Urs Schaffner, and Johnnie van den Berg. 2019. 'Agro-Ecological Options for Fall Armyworm (*Spodoptera Frugiperda* JE Smith) Management: Providing Low-Cost, Smallholder Friendly Solutions to an Invasive Pest'. *Journal of Environmental Management* 243 (August): 318–30. <https://doi.org/10.1016/j.jenvman.2019.05.011>.
- Harvey, Jeffrey A., Robin Heinen, Inge Armbrrecht, Yves Basset, James H. Baxter-Gilbert, T. Martijn Bezemer, Monika Böhm, et al. 2020. 'International Scientists Formulate a Roadmap for Insect Conservation and Recovery'. *Nature Ecology & Evolution* 4 (2): 174–76.
<https://doi.org/10.1038/s41559-019-1079-8>.
- Hotaling, Scott, John S Sproul, Jacqueline Heckenhauer, Ashlyn Powell, Amanda M Larracuenta, Steffen U Pauls, Joanna L Kelley, and Paul B Frandsen. 2021. 'Long-Reads Are Revolutionizing 20 Years of Insect Genome Sequencing'. Edited by Federico Hoffmann. *Genome Biology and Evolution*, June, evab138. <https://doi.org/10.1093/gbe/evab138>.
- i5K Consortium. 2013. 'The I5K Initiative: Advancing Arthropod Genomics for Knowledge, Human Health, Agriculture, and the Environment'. *Journal of Heredity* 104 (5): 595–600. <https://doi.org/10.1093/jhered/est050>.
- Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, et al. 2014. 'InterProScan 5: Genome-Scale Protein Function Classification'. *Bioinformatics* 30 (9): 1236–40.
<https://doi.org/10.1093/bioinformatics/btu031>.

- Jordan, I. King, Igor B. Rogozin, Yuri I. Wolf, and Eugene V. Koonin. 2002. 'Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria'. *Genome Research* 12 (6): 962–68. <https://doi.org/10.1101/gr.87702>.
- Junier, Thomas, and Evgeny M. Zdobnov. 2010. 'The Newick Utilities: High-Throughput Phylogenetic Tree Processing in the Unix Shell'. *Bioinformatics* 26 (13): 1669–70. <https://doi.org/10.1093/bioinformatics/btq243>.
- Kapheim, Karen M., Hailin Pan, Cai Li, Charles Blatti, Brock A. Harpur, Panagiotis Ioannidis, Beryl M. Jones, et al. 2019. 'Draft Genome Assembly and Population Genetics of an Agricultural Pollinator, the Solitary Alkali Bee (Halictidae: *Nomia Melanderi*)'. *G3 Genes|Genomes|Genetics*, January, g3.200865.2018. <https://doi.org/10.1534/g3.118.200865>.
- Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. 2020. 'Phylogenetic Tree Building in the Genomic Age'. *Nature Reviews Genetics* 21 (7): 428–44. <https://doi.org/10.1038/s41576-020-0233-0>.
- Kassambara, Alboukadel. 2017. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. Edition 1. Multivariate Analysis 1. France: STHDA.
- Kerfeld, Cheryl A., and Kathleen M. Scott. 2011. 'Using BLAST to Teach "E-Value-Tionary" Concepts'. *PLOS Biology* 9 (2): e1001014. <https://doi.org/10.1371/journal.pbio.1001014>.
- Kite, Geoffrey C., Alison C. Plant, Andrew Burke, Monique J. S. Simmonds, Walter M. Blaney, and Linda E. Fellows. 1995. 'Accumulation of Trans-3-Hydroxy-L-Proline by Seeds and Leaves of the Edible Madagascan Legume *Lemuropisum Edule* H. Perrier'. *Kew Bulletin* 50 (3): 585–90. <https://doi.org/10.2307/4110329>.
- Kohonen, Teuvo. 1982. 'Self-Organized Formation of Topologically Correct Feature Maps'. *Biological Cybernetics* 43 (1): 59–69. <https://doi.org/10.1007/BF00337288>.
- Koonin, Eugene V. 2003. 'Comparative Genomics, Minimal Gene-Sets and the Last Universal Common Ancestor'. *Nature Reviews Microbiology* 1 (2): 127–36. <https://doi.org/10.1038/nrmicro751>.
- Koonin, Eugene V. 2005. 'Orthologs, Paralogs, and Evolutionary Genomics'. *Annual Review of Genetics* 39 (1): 309–38.

- <https://doi.org/10.1146/annurev.genet.39.073003.114725>.
- Koonin, Eugene V. 2011. 'Are There Laws of Genome Evolution?' Edited by Philip E. Bourne. *PLoS Computational Biology* 7 (8): e1002173. <https://doi.org/10.1371/journal.pcbi.1002173>.
- Koonin, Eugene V., Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Dmitri M. Krylov, Kira S. Makarova, Raja Mazumder, et al. 2004. 'A Comprehensive Evolutionary Classification of Proteins Encoded in Complete Eukaryotic Genomes'. *Genome Biology* 5 (2): R7. <https://doi.org/10.1186/gb-2004-5-2-r7>.
- Koonin, Eugene V., and Yuri I. Wolf. 2010. 'Constraints and Plasticity in Genome and Molecular-Phenome Evolution'. *Nature Reviews Genetics* 11 (7): 487–98. <https://doi.org/10.1038/nrg2810>.
- Köster, Johannes, and Sven Rahmann. 2012. 'Snakemake—a Scalable Bioinformatics Workflow Engine'. *Bioinformatics* 28 (19): 2520–22. <https://doi.org/10.1093/bioinformatics/bts480>.
- Kriventseva, Evgenia V., Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Renata Dias, Felipe A. Simão, and Evgeny M. Zdobnov. 2019. 'OrthoDB V10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs'. *Nucleic Acids Research* 47 (D1): D807–11. <https://doi.org/10.1093/nar/gky1053>.
- Krylov, D. M. 2003. 'Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution'. *Genome Research* 13 (10): 2229–35. <https://doi.org/10.1101/gr.1589103>.
- Kuang, Junyao, Nicolas Buchon, Kristin Michel, and Caterina Scoglio. 2022. 'A Global Anopheles Gambiae Gene Co-Expression Network Constructed from Hundreds of Experimental Conditions with Missing Values'. *BMC Bioinformatics* 23 (1): 170. <https://doi.org/10.1186/s12859-022-04697-9>.
- Kulmanov, Maxat, and Robert Hoehndorf. 2020. 'DeepGOPlus: Improved Protein Function Prediction from Sequence'. Edited by Lenore Cowen. *Bioinformatics* 36 (2): 422–29. <https://doi.org/10.1093/bioinformatics/btz595>.
- Kumar, Sudhir, Glen Stecher, Michael Suleski, and S. Blair Hedges. 2017. 'TimeTree: A Resource for Timelines, Timetrees, and Divergence Times'.

- Molecular Biology and Evolution* 34 (7): 1812–19.
<https://doi.org/10.1093/molbev/msx116>.
- Langfelder, Peter, and Steve Horvath. 2007. 'Eigengene Networks for Studying the Relationships between Co-Expression Modules'. *BMC Systems Biology* 1 (1): 54. <https://doi.org/10.1186/1752-0509-1-54>.
- Le, Si Quang, and Olivier Gascuel. 2008. 'An Improved General Amino Acid Replacement Matrix'. *Molecular Biology and Evolution* 25 (7): 1307–20. <https://doi.org/10.1093/molbev/msn067>.
- Legeai, F., S. Shigenobu, J.-P. Gauthier, J. Colbourne, C. Rispe, O. Collin, S. Richards, A. C. C. Wilson, T. Murphy, and D. Tagu. 2010. 'AphidBase: A Centralized Bioinformatic Resource for Annotation of the Pea Aphid Genome'. *Insect Molecular Biology* 19 (s2): 5–12. <https://doi.org/10.1111/j.1365-2583.2009.00930.x>.
- Lemmon, Emily Moriarty, and Alan R. Lemmon. 2013. 'High-Throughput Genomic Data in Systematics and Phylogenetics'. *Annual Review of Ecology, Evolution, and Systematics* 44 (1): 99–121. <https://doi.org/10.1146/annurev-ecolsys-110512-135822>.
- Lespinet, Olivier, Yuri I. Wolf, Eugene V. Koonin, and L. Aravind. 2002. 'The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes'. *Genome Research* 12 (7): 1048–59. <https://doi.org/10.1101/gr.174302>.
- Li, F., X. Zhao, M. Li, K. He, C. Huang, Y. Zhou, Z. Li, and J. R. Walters. 2019. 'Insect Genomes: Progress and Challenges'. *Insect Molecular Biology* 28 (6): 739–58. <https://doi.org/10.1111/imb.12599>.
- Li, Jia, Robert M. Waterhouse, and Evgeny M. Zdobnov. 2011. 'A Remarkably Stable TipE Gene Cluster: Evolution of Insect Para Sodium Channel Auxiliary Subunits'. *BMC Evolutionary Biology* 11 (1): 337. <https://doi.org/10.1186/1471-2148-11-337>.
- Li, Yang, Sarah E. Calvo, Roe Gutman, Jun S. Liu, and Vamsi K. Mootha. 2014. 'Expansion of Biological Pathways Based on Evolutionary Inference'. *Cell* 158 (1): 213–25. <https://doi.org/10.1016/j.cell.2014.05.034>.
- Linard, Benjamin, Ngoc Hoan Nguyen, Francisco Prosdocimi, Olivier Poch, and Julie D. Thompson. 2012. 'EvoluCode: Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data'. *Evolutionary Bioinformatics* 8 (January): EBO.S8814. <https://doi.org/10.4137/EBO.S8814>.
- Lukyanchikova, Varvara, Miroslav Nuriddinov, Polina Belokopytova, Alena

- Taskina, Jiangtao Liang, Maarten J. M. F. Reijnders, Livio Ruzzante, et al. 2022. 'Anopheles Mosquitoes Reveal New Principles of 3D Genome Organization in Insects'. *Nature Communications* 13 (1): 1960. <https://doi.org/10.1038/s41467-022-29599-5>.
- Manduchi, Laura, Matthias Hüser, Martin Faltys, Julia Vogt, Gunnar Rätsch, and Vincent Fortuin. 2021. 'T-DPSOM: An Interpretable Clustering Method for Unsupervised Learning of Patient Health States'. In *Proceedings of the Conference on Health, Inference, and Learning*, 236–45. CHIL '21. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3450439.3451872>.
- Mei, Yang, Dong Jing, Shenyang Tang, Xi Chen, Hao Chen, Haonan Duanmu, Yuyang Cong, et al. 2022. 'InsectBase 2.0: A Comprehensive Gene Resource for Insects'. *Nucleic Acids Research* 50 (D1): D1040–45. <https://doi.org/10.1093/nar/gkab1090>.
- Mendes, Fábio K, Dan Vanderpool, Ben Fulton, and Matthew W Hahn. 2020. 'CAFE 5 Models Variation in Evolutionary Rates among Gene Families'. *Bioinformatics* 36 (22–23): 5516–18. <https://doi.org/10.1093/bioinformatics/btaa1022>.
- Micheli, Gioacchino, and Giorgio Camilloni. 2022. 'Can Introns Stabilize Gene Duplication?' *Biology* 11 (6): 941. <https://doi.org/10.3390/biology11060941>.
- Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era'. *Molecular Biology and Evolution* 37 (5): 1530–34. <https://doi.org/10.1093/molbev/msaa015>.
- Misof, B., S. Liu, K. Meusemann, R. S. Peters, A. Donath, C. Mayer, P. B. Frandsen, et al. 2014. 'Phylogenomics Resolves the Timing and Pattern of Insect Evolution'. *Science* 346 (6210): 763–67. <https://doi.org/10.1126/science.1257570>.
- Nagy, László G, Zsolt Merényi, Botond Hegedüs, and Balázs Bálint. 2020. 'Novel Phylogenetic Methods Are Needed for Understanding Gene Function in the Era of Mega-Scale Genome Sequencing'. *Nucleic Acids Research* 48 (5): 2209–19. <https://doi.org/10.1093/nar/gkz1241>.
- Neafsey, Daniel E., Robert M. Waterhouse, Mohammad R. Abai, Sergey S.

- Aganezov, Max A. Alekseyev, James E. Allen, James Amon, et al. 2015. 'Highly Evolvable Malaria Vectors: The Genomes of 16 Anopheles Mosquitoes'. *Science* 347 (6217): 1258522. <https://doi.org/10.1126/science.1258522>.
- Oberprieler, Rolf G., Adriana E. Marvaldi, and Robert S. Anderson. 2007. 'Weevils, Weevils, Weevils Everywhere*'. *Zootaxa* 1668 (1): 491–520. <https://doi.org/10.11646/zootaxa.1668.1.24>.
- Ogata, Hiroyuki, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. 1999. 'KEGG: Kyoto Encyclopedia of Genes and Genomes'. *Nucleic Acids Research* 27 (1): 29–34. <https://doi.org/10.1093/nar/27.1.29>.
- O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciuffo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. 'Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation'. *Nucleic Acids Research* 44 (D1): D733–45. <https://doi.org/10.1093/nar/gkv1189>.
- Page, Roderic D.M. 2003. 'Visualizing Phylogenetic Trees Using TreeView'. *Current Protocols in Bioinformatics* 00 (1): 6.2.1-6.2.15. <https://doi.org/10.1002/0471250953.bi0602s01>.
- Papanicolaou, Alexie, Steffi Gebauer-Jung, Mark L. Blaxter, W. Owen McMillan, and Chris D. Jiggins. 2008. 'ButterflyBase: A Platform for Lepidopteran Genomics'. *Nucleic Acids Research* 36 (suppl_1): D582–87. <https://doi.org/10.1093/nar/gkm853>.
- Paradis, Emmanuel, and Klaus Schliep. 2019. 'Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R'. Edited by Russell Schwartz. *Bioinformatics* 35 (3): 526–28. <https://doi.org/10.1093/bioinformatics/bty633>.
- Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. 'FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments'. *PLOS ONE* 5 (3): e9490. <https://doi.org/10.1371/journal.pone.0009490>.
- Reijnders, Maarten J. M. F. 2022. 'Wei2GO: Weighted Sequence Similarity-Based Protein Function Prediction'. *PeerJ* 10 (February): e12931. <https://doi.org/10.7717/peerj.12931>.
- Reijnders, Maarten J. M. F., and Robert M. Waterhouse. 2021a. 'Summary Visualizations of Gene Ontology Terms With GO-Figure!' *Frontiers in*

Bioinformatics 1 (April): 638255.
<https://doi.org/10.3389/fbinf.2021.638255>.

Reijnders, Maarten J.M.F., and Robert M. Waterhouse. 2021b. 'CrowdGO: Machine Learning and Semantic Similarity Guided Consensus Gene Ontology Annotation'. Preprint. *Bioinformatics*.
<https://doi.org/10.1101/731596>.

Richards, Stephen, Anna Childers, and Christopher Childers. 2018. 'Editorial Overview: Insect Genomics: Arthropod Genomic Resources for the 21st Century: It Only Counts If It's in the Database!' *Current Opinion in Insect Science*, Insect genomics * Development and regulation, 25 (February): iv–vii. <https://doi.org/10.1016/j.cois.2018.02.015>.

Robertson, Hugh M, Robert M Waterhouse, Kimberly K O Walden, Livio Ruzzante, Maarten J M F Reijnders, Brad S Coates, Fabrice Legeai, et al. 2018. 'Genome Sequence of the Wheat Stem Sawfly, *Cephus cinctus*, Representing an Early-Branching Lineage of the Hymenoptera, Illuminates Evolution of Hymenopteran Chemoreceptors'. *Genome Biology and Evolution*, October. <https://doi.org/10.1093/gbe/evy232>.

Ruzzante, Livio, Romain Feron, Maarten J M F Reijnders, Antonin Thiébaud, and Robert M Waterhouse. 2022. 'Functional Constraints on Insect Immune System Components Govern Their Evolutionary Trajectories'. *Molecular Biology and Evolution* 39 (1): msab352.
<https://doi.org/10.1093/molbev/msab352>.

Ruzzante, Livio, Maarten J. M. F. Reijnders, and Robert M. Waterhouse. 2019. 'Of Genes and Genomes: Mosquito Evolution and Diversity'. *Trends in Parasitology* 35 (1): 32–51. <https://doi.org/10.1016/j.pt.2018.10.003>.

Sayers, Eric W, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, and Ilene Karsch-Mizrachi. 2019. 'GenBank'. *Nucleic Acids Research* 47 (D1): D94–99. <https://doi.org/10.1093/nar/gky989>.

Scheibenreif, Linus, Maria Littmann, Christine Orengo, and Burkhard Rost. 2019. 'FunFam Protein Families Improve Residue Level Molecular Function Prediction'. *BMC Bioinformatics* 20 (1): 400.
<https://doi.org/10.1186/s12859-019-2988-x>.

Schlitt, Thomas, Kimmo Palin, Johan Rung, Sabine Dietmann, Michael Lappe, Esko Ukkonen, and Alvis Brazma. 2003. 'From Gene Networks to Gene Function'. *Genome Research* 13 (12): 2568–76.

<https://doi.org/10.1101/gr.1111403>.

- Schoch, Conrad L, Stacy Ciuffo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, et al. 2020. 'NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools'. *Database* 2020 (January): baaa062. <https://doi.org/10.1093/database/baaa062>.
- Schwentner, Martin, David J. Combosch, Joey Pakes Nelson, and Gonzalo Giribet. 2017. 'A Phylogenomic Solution to the Origin of Insects by Resolving Crustacean-Hexapod Relationships'. *Current Biology* 27 (12): 1818-1824.e5. <https://doi.org/10.1016/j.cub.2017.05.040>.
- Seppey, Mathieu, Panagiotis Ioannidis, Brent C. Emerson, Camille Pitteloud, Marc Robinson-Rechavi, Julien Roux, Hermes E. Escalona, et al. 2019. 'Genomic Signatures Accompanying the Dietary Shift to Phytophagy in Polyphagan Beetles'. *Genome Biology* 20 (1): 98. <https://doi.org/10.1186/s13059-019-1704-5>.
- Shah, Neethu, Douglas R Dorer, Etsuko N Moriyama, and Alan C Christensen. 2012. 'Evolution of a Large, Conserved, and Syntenic Gene Family in Insects'. *G3 Genes|Genomes|Genetics* 2 (2): 313-19. <https://doi.org/10.1534/g3.111.001412>.
- Simakov, Oleg, Jessen Bredeson, Kodiak Berkoff, Ferdinand Marletaz, Therese Mitros, Darrin T. Schultz, Brendan L. O'Connell, et al. n.d. 'Deeply Conserved Synteny and the Evolution of Metazoan Chromosomes'. *Science Advances* 8 (5): eabi5884. <https://doi.org/10.1126/sciadv.abi5884>.
- Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. 'BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs'. *Bioinformatics* 31 (19): 3210-12. <https://doi.org/10.1093/bioinformatics/btv351>.
- Sinka, M. E., S. Pironon, N. C. Massey, J. Longbottom, J. Hemingway, C. L. Moyes, and K. J. Willis. 2020. 'A New Malaria Vector in Africa: Predicting the Expansion Range of Anopheles Stephensi and Identifying the Urban Populations at Risk'. *Proceedings of the National Academy of Sciences* 117 (40): 24900-908. <https://doi.org/10.1073/pnas.2003976117>.
- Soubrier, Julien, Mike Steel, Michael S.Y. Lee, Clio Der Sarkissian, Stéphane Guindon, Simon Y.W. Ho, and Alan Cooper. 2012. 'The Influence of Rate





- Heterogeneity among Sites on the Time Dependence of Molecular Rates'. *Molecular Biology and Evolution* 29 (11): 3345–58. <https://doi.org/10.1093/molbev/mss140>.
- Stork, Nigel E. 2018. 'How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth?' *Annual Review of Entomology* 63 (1): 31–45. <https://doi.org/10.1146/annurev-ento-020117-043348>.
- Sun, Cheng, Jiaying Huang, Yun Wang, Xiaomeng Zhao, Long Su, Gregg W C Thomas, Mengya Zhao, et al. 2021. 'Genus-Wide Characterization of Bumblebee Genomes Provides Insights into Their Evolution and Variation in Ecological and Behavioral Traits'. Edited by Fuwen Wei. *Molecular Biology and Evolution* 32 (2): 486–501. <https://doi.org/10.1093/molbev/msaa240>.
- Suzuki, R., and H. Shimodaira. 2006. 'Pvclust: An R Package for Assessing the Uncertainty in Hierarchical Clustering'. *Bioinformatics* 22 (12): 1540–42. <https://doi.org/10.1093/bioinformatics/btl117>.
- Tatusov, Roman L., Michael Y. Galperin, Darren A. Natale, and Eugene V. Koonin. 2000. 'The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution'. *Nucleic Acids Research* 28 (1): 33–36. <https://doi.org/10.1093/nar/28.1.33>.
- Tatusov, Roman L., Eugene V. Koonin, and David J. Lipman. 1997. 'A Genomic Perspective on Protein Families'. *Science* 278 (5338): 631–37. <https://doi.org/10.1126/science.278.5338.631>.
- Thomas, Gregg W. C., Elias Dohmen, Daniel S. T. Hughes, Shwetha C. Murali, Monica Poelchau, Karl Glastad, Clare A. Anstead, et al. 2020. 'Gene Content Evolution in the Arthropods'. *Genome Biology* 21 (1): 15. <https://doi.org/10.1186/s13059-019-1925-7>.
- Thurmond, Jim, Joshua L Goodman, Victor B Strelets, Helen Attrill, L Sian Gramates, Steven J Marygold, Beverley B Matthews, et al. 2019. 'FlyBase 2.0: The next Generation'. *Nucleic Acids Research* 47 (D1): D759–65. <https://doi.org/10.1093/nar/gky1003>.
- Tihelka, Erik, Chenyang Cai, Mattia Giacomelli, Jesus Lozano-Fernandez, Omar Rota-Stabelli, Diying Huang, Michael S. Engel, Philip C. J. Donoghue, and Davide Pisani. 2021. 'The Evolution of Insect Biodiversity'. *Current Biology* 31 (19): R1299–1311. <https://doi.org/10.1016/j.cub.2021.08.057>.
- Vakirlis, Nikolaos, Anne-Ruxandra Carvunis, and Aoife McLysaght. 2020.

- 'Synteny-Based Analyses Indicate That Sequence Divergence Is Not the Main Source of Orphan Genes'. *ELife* 9 (February): e53500. <https://doi.org/10.7554/eLife.53500>.
- Walton, Alexander, Michael J. Sheehan, and Amy L. Toth. 2020. 'Going Wild for Functional Genomics: RNA Interference as a Tool to Study Gene-Behavior Associations in Diverse Species and Ecological Contexts'. *Hormones and Behavior* 124 (August): 104774. <https://doi.org/10.1016/j.yhbeh.2020.104774>.
- Waterhouse, R. M., E. V. Kriventseva, S. Meister, Z. Xi, K. S. Alvarez, L. C. Bartholomay, C. Barillas-Mury, et al. 2007. 'Evolutionary Dynamics of Immune-Related Genes and Pathways in Disease-Vector Mosquitoes'. *Science* 316 (5832): 1738–43. <https://doi.org/10.1126/science.1139862>.
- Waterhouse, R. M., E. M. Zdobnov, and E. V. Kriventseva. 2011. 'Correlating Traits of Gene Retention, Sequence Divergence, Duplicability and Essentiality in Vertebrates, Arthropods, and Fungi'. *Genome Biology and Evolution* 3 (0): 75–86. <https://doi.org/10.1093/gbe/evq083>.
- Waterhouse, Robert M. 2015. 'A Maturing Understanding of the Composition of the Insect Gene Repertoire'. *Current Opinion in Insect Science, Insect genomics * Development and regulation*, 7 (February): 15–23. <https://doi.org/10.1016/j.cois.2015.01.004>.
- Waterhouse, Robert M., Sergey Aganezov, Yoann Anselmetti, Jiyoung Lee, Livio Ruzzante, Maarten J. M. F. Reijnders, Romain Feron, et al. 2020. 'Evolutionary Superscaffolding and Chromosome Anchoring to Improve Anopheles Genome Assemblies'. *BMC Biology* 18 (1): 1. <https://doi.org/10.1186/s12915-019-0728-3>.
- Waterhouse, Robert M., Fredrik Tegenfeldt, Jia Li, Evgeny M. Zdobnov, and Evgenia V. Kriventseva. 2013. 'OrthoDB: A Hierarchical Catalog of Animal, Fungal and Bacterial Orthologs'. *Nucleic Acids Research* 41 (D1): D358–65. <https://doi.org/10.1093/nar/gks1116>.
- Wehrens, Ron, and Johannes Kruisselbrink. 2018. 'Flexible Self-Organizing Maps in Kohonen 3.0'. *Journal of Statistical Software* 87 (1): 1–18. <https://doi.org/10.18637/jss.v087.i07>.
- Wilson, Robert J., and Richard Fox. 2021. 'Insect Responses to Global Change Offer Signposts for Biodiversity and Conservation'. *Ecological Entomology* 46 (4): 699–717. <https://doi.org/10.1111/een.12970>.

- Wolf, Yuri I, Liran Carmel, and Eugene V Koonin. 2006. 'Unifying Measures of Gene Function and Evolution'. *Proceedings of the Royal Society B: Biological Sciences* 273 (1593): 1507–15. <https://doi.org/10.1098/rspb.2006.3472>.
- Yang, Z. 1995. 'A Space-Time Process Model for the Evolution of DNA Sequences.' *Genetics* 139 (2): 993–1005. <https://doi.org/10.1093/genetics/139.2.993>.
- Yates, Andrew D, James Allen, Ridwan M Amode, Andrey G Azov, Matthieu Barba, Andrés Becerra, Jyothish Bhai, et al. 2022. 'Ensembl Genomes 2022: An Expanding Genome Resource for Non-Vertebrates'. *Nucleic Acids Research* 50 (D1): D996–1003. <https://doi.org/10.1093/nar/gkab1007>.
- Yon Rhee, Seung, Valerie Wood, Kara Dolinski, and Sorin Draghici. 2008. 'Use and Misuse of the Gene Ontology Annotations'. *Nature Reviews Genetics* 9 (7): 509–15. <https://doi.org/10.1038/nrg2363>.
- Yoshida, Yuki, Nurislam Shaikhutdinov, Olga Kozlova, Masayoshi Itoh, Michihira Tagami, Mitsuyoshi Murata, Hiromi Nishiyori-Sueki, et al. 2022. 'High Quality Genome Assembly of the Anhydrobiotic Midge Provides Insights on a Single Chromosome-Based Emergence of Extreme Desiccation Tolerance'. *NAR Genomics and Bioinformatics* 4 (2): lqac029. <https://doi.org/10.1093/nargab/lqac029>.
- Zdobnov, Evgeny M, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Matthew Berkeley, and Evgenia V Kriventseva. 2021. 'OrthoDB in 2020: Evolutionary and Functional Annotations of Orthologs'. *Nucleic Acids Research* 49 (D1): D389–93. <https://doi.org/10.1093/nar/gkaa1009>.
- Zhang, Jianzhi. 2003. 'Evolution by Gene Duplication: An Update'. *Trends in Ecology & Evolution* 18 (6): 292–98. [https://doi.org/10.1016/S0169-5347\(03\)00033-8](https://doi.org/10.1016/S0169-5347(03)00033-8).

Appendix 1: Immunity Case-Study

Functional Constraints on Insect Immune System Components Govern Their Evolutionary Trajectories

Livio Ruzzante , Romain Feron , Maarten J.M.F. Reijnders , Antonin Thiébaud , and Robert M. Waterhouse *

Department of Ecology and Evolution, Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland

*Corresponding author: E-mail: robert.waterhouse@unil.ch.

Associate editor: Thomas Leitner

Abstract

Roles of constraints in shaping evolutionary outcomes are often considered in the contexts of developmental biology and population genetics, in terms of capacities to generate new variants and how selection limits or promotes consequent phenotypic changes. Comparative genomics also recognizes the role of constraints, in terms of shaping evolution of gene and genome architectures, sequence evolutionary rates, and gene gains or losses, as well as on molecular phenotypes. Characterizing patterns of genomic change where putative functions and interactions of system components are relatively well described offers opportunities to explore whether genes with similar roles exhibit similar evolutionary trajectories. Using insect immunity as our test case system, we hypothesize that characterizing gene evolutionary histories can define distinct dynamics associated with different functional roles. We develop metrics that quantify gene evolutionary histories, employ these to characterize evolutionary features of immune gene repertoires, and explore relationships between gene family evolutionary profiles and their roles in immunity to understand how different constraints may relate to distinct dynamics. We identified three main axes of evolutionary trajectories characterized by gene duplication and synteny, maintenance/stability and sequence conservation, and loss and sequence divergence, highlighting similar and contrasting patterns across these axes amongst subsets of immune genes. Our results suggest that where and how genes participate in immune responses limit the range of possible evolutionary scenarios they exhibit. The test case study system of insect immunity highlights the potential of applying comparative genomics approaches to characterize how functional constraints on different components of biological systems govern their evolutionary trajectories.

Key words: *Anopheles* mosquito, evolutionary profiling, gene expression, gene families, innate immunity.

Introduction

The concept of constraints in evolutionary biology encompasses a diverse array of interpretations and terminologies shaped by the approaches of different research fields (Antonovics and van Tienderen 1991). In general terms, constraints can be described as factors that limit or direct the process of natural selection leading to outcomes representing only a fraction of all theoretically possible scenarios. Constraints may impact the capacity to generate new variants as well as how selection either limits or promotes consequent phenotypic change, often considered in developmental biology (Richardson and Chipman 2003) and population genetics (Hoffmann 2013) contexts. Comparative genomics also recognizes the role of constraints, in shaping the evolution of gene and genome architectures, sequence evolutionary rates, and gene gains and losses, as well as on the molecular phenotypes governed by their functional products (Koonin and Wolf 2010). For example, protein sequence evolution is constrained by requirements for maintaining proper protein structure and function, including during folding and interactions with other macromolecules (Worth et al. 2009). Functional constraints also impact the

evolution of gene families, for example, families of paralogs with or without essential genes exhibit dramatically different evolutionary regimes in terms of sequence divergence and duplication rates (Shakhnovich and Koonin 2006). These likely influence observed trends across the gene duplication spectrum that show a dichotomy of constrained single-copy control versus a multi-copy license for greatly relaxed copy-number restrictions (Waterhouse et al. 2011). Integrative analyses of evolutionary and functional constraints point to emergent properties such as a gene family's "importance" or "status" characterized by low sequence divergence and propensity for gene loss with high expression levels, protein interactions, and essentiality; or a family's "adaptability" manifested by high duplication levels, many genetic interaction partners, and a tendency of genes to be nonessential; or a family's "reactivity" with high gain/loss and expression levels but low sequence divergence, a paucity of essential genes, and few physical or genetic interactions (Wolf et al. 2006). If such constraints limit the realm of possibilities in terms of allowed gene evolutionary trajectories then recurring patterns should be observable for genes evolving under similar constraints. Characterizing these patterns in the context of a relatively

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Open Access

well-studied system, where putative functional roles and interactions of member genes are well described, offers an opportunity to explore whether genes with similar or analogous functions exhibit similar evolutionary trajectories, possibly governed by common constraints.

The insect innate immune system is relatively well characterized with respect to the functional roles and evolutionary histories of key implicated pathways and component gene families. It confers remarkable resilience to encountered pathogens through the activation of powerful responses to neutralize and clear infections (Roff and Reynolds 2009; Ligoxygakis 2017). The immune system comprises both humoral and cellular responses with components dedicated to recognizing signs of infection, signaling cascades to activate primary defenses and induce transcriptional responses, modulators that control the intensity and direction of responses, and effector proteins and biomolecules for pathogen killing. Many of the genes and their protein products implicated in these complex processes were first identified in the fruit fly, *Drosophila melanogaster* (Lemaitre and Hoffmann 2007; Imler 2014). Classical receptor proteins that recognize pathogen-associated molecular patterns include peptidoglycan recognition proteins (PGRPs) (Wang et al. 2019) and β -1,3-glucan recognition or gram-negative bacteria-binding proteins (GNBPs) (Rao et al. 2018). Pathogen recognition may then trigger immune signaling through the Toll (Valanne et al. 2011), Imd (Myllymäki et al. 2014), or the JAnus kinase protein (JAK)/signal transducer and activator of transcription (STAT) (Myllymäki and Rämet 2014) pathways. Their activation leads to the translocation of transcription factors to the nucleus where the expression of effector genes such as those encoding antimicrobial peptides (AMPs) (Lazzaro et al. 2020) is upregulated. Defense responses are mediated by various cells and tissues including hemocytes, the fat body, and the midgut, and pathogen killing can occur via processes such as melanization, phagocytosis, lysis, autophagy, and apoptosis (Hillyer 2016; King 2020), with RNA interference (RNAi) facilitating major antiviral defenses (Mussabekova et al. 2017). These complex interactions collectively offer insects protection from a vast array of viruses, bacteria, fungi, protozoa, and nematodes.

Sequencing the *D. melanogaster* and *Anopheles gambiae* genomes provided the first opportunity for comparative genomic analysis of immune-related genes in insects (Christophides et al. 2002). Advances in genome sequencing technologies have facilitated an increasingly dense sampling of species to explore insect gene repertoires and perform cross-species comparisons to trace gene evolutionary histories (Waterhouse 2015). This has allowed comparisons beyond Diptera to include Hymenoptera (Evans et al. 2006; Brucker et al. 2012; Barribeau et al. 2015), Coleoptera (Zou et al. 2007), Lepidoptera (Tanaka et al. 2008), and Hemiptera (Gerardo et al. 2010), as well as expanded sampling of flies and mosquitoes (Sackton et al. 2007; Waterhouse et al. 2007; Bartholomay et al. 2010; Sackton et al. 2017). These comparative studies generally focused on the canonical immune-related gene repertoire, comprising genes that have been directly implicated in immune responses through experimental research, or

indirectly linked to immunity through homology to known immune proteins (Bartholomay and Michel 2018; Waterhouse et al. 2020). Emerging patterns pointed to distinct evolutionary dynamics that characterize different immune phases: (1) gene and domain gains or losses (turnover) can create diversity in recognition modules; (2) core signaling pathway members are almost always maintained as single-copy orthologs often with elevated levels of sequence divergence; (3) modulators appear to form lineage-restricted units with members picked from large families often with high gene turnover rates; and (4) effectors like AMPs show dynamic gains and losses or are lineage-restricted whereas oxidative defense effectors are widespread with low levels of sequence divergence. These observations provide specific examples and strong expectations of types of genes with similar functions that exhibit similar evolutionary trajectories, within the established framework of insect innate immunity that classifies genes and families into broad functional categories of recognition, signal transduction, modulation, and effector components.

These trends are based on observations from cross-species comparisons of insect immune gene repertoires. Here we hypothesize that comprehensive quantitative multispecies and multifeature characterization of gene family evolutionary histories can define distinct dynamics associated with different functional roles in immune responses. Such detailed evolutionary profiling can then be used to address the question of whether gene families involved in common immune functional categories, modules, or processes exhibit similar evolutionary trajectories possibly driven by shared constraints. We take advantage of genomic resources available for 22 mosquito species (Holt et al. 2002; Nene et al. 2007; Arensbarger et al. 2010; Lawniczak et al. 2010; Marinotti et al. 2013; Jiang et al. 2014; Chen et al. 2015; Neafsey et al. 2015; Matthews et al. 2018; Ruzzante et al. 2019) and 46 other insects to (1) develop a suite of metrics that quantify gene and gene family evolutionary histories, (2) employ these metrics to characterize the evolutionary features of mosquito and fly immune gene repertoires, and (3) explore the relationships between gene family evolutionary profiles and their functional roles in immunity to understand how different constraints may relate to distinct dynamics. The resolution afforded by multispecies comparative analyses and our suite of gene sequence and copy-number evolutionary metrics reveals the evolutionary features that most clearly distinguish each family, and highlights similar and contrasting patterns across all immune gene families. Complementing knowledge-based functional categorizations with gene coexpression analyses identifies immune families that function in concert, revealing evolutionary-functional correspondences where most prominently, families involved in mosquito complement system responses show both high evolutionary similarities and high expression similarities.

Results and Discussion

A Suite of Metrics to Quantify Gene and Gene Family Evolutionary Histories

The developed set of evolutionary feature metrics is designed to capture a broad spectrum of gene evolutionary dynamics

Table 1. Evolutionary Feature Metric Descriptions.

Evolutionary feature	Acronym	Description	Data source
Taxonomic age	AGE	Age of the last common ancestor of species in an OG, in terms of millions of years since divergence, computed from the ultrametric species phylogeny	43-insect orthology
Universality	UNI	The proportion of the total species present in an OG (all species, UNI = 1)	43-insect orthology
Duplicability	DUP	The proportion of species present in an OG that have multicopy orthologs	43-insect orthology
Average copy number	ACN	The average (mean) ortholog copy number across all species present in an OG	43-insect orthology
Copy number variation	CNV	The standard deviation of ortholog counts per species present in an OG divided by the ACN	43-insect orthology
Expansions	EXP	CAFE quantified proportions of gene gain nodes for an OG	43-insect orthology
Contractions	CON	CAFE quantified proportions of gene loss nodes for an OG	43-insect orthology
Stability	STA	CAFE quantified proportions of no copy-number change nodes for an OG	43-insect orthology
Synteny	SYN	The proportion of orthologs in an OG that maintains their orthologous neighbors in the genomes of the other species	43-insect orthology
Evolutionary rate	EVR	The average rate of protein sequence divergence normalized by the distance (% identity) between each pair of species as computed by OrthoDB	43-insect orthology
PAML's dS	PDS	The number of synonymous substitutions per synonymous site as computed by PAML	19- <i>Anopheles</i> orthology
PAML's dN	PDN	The number of nonsynonymous substitutions per nonsynonymous site as computed by PAML	19- <i>Anopheles</i> orthology
PAML's dN/dS	SEL	The nonsynonymous to synonymous substitution ratio (dN/dS) as computed by PAML	19- <i>Anopheles</i> orthology
Nonsynonymous SNP proportion	NSP	The proportion of all coding-sequence SNPs that were nonsynonymous (averaged over genes per OG)	<i>An. gambiae</i> variation
Nonsynonymous SNP density	NSD	The density of nonsynonymous SNPs over a gene's coding-sequence length (averaged over genes per OG)	<i>An. gambiae</i> variation
Synonymous SNP density	SSD	The density of synonymous SNPs over a gene's coding-sequence length (averaged over genes per OG)	<i>An. gambiae</i> variation
Whole genome alignability	WGA	The number of species aligned, per nucleotide from the whole-genome alignment, averaged over coding-sequence length (averaged over genes per OG)	22 mosquitoes 36 <i>Drosophila</i>
PhastCons constraint	PHC	PhastCons quantified constraint scores, per nucleotide from the whole-genome alignment, averaged over coding-sequence length (averaged over genes per PG)	22 mosquitoes 36 <i>Drosophila</i>

NOTE.—For each evolutionary feature, the metric name, acronym, description, and source data are presented (see Materials and Methods for details).

including taxonomic spread, copy-number changes, protein- and DNA-level sequence divergence, conservation, and constraint, as well as genomic organization, and population-level sequence variation (table 1). The 18 metrics are computed using gene orthology delineated across 43 insect species (21 mosquitoes, 15 other dipterans, and 2 outgroup representatives each for Lepidoptera, Coleoptera, Hymenoptera, and Exopterygota), sets of whole genome alignments with 22 mosquitoes or with 36 *Drosophila*, or polymorphism data from *An. gambiae* (see Materials and Methods). Orthologous groups (OGs) comprised all genes descended from a single gene in the last common ancestor of the set of the extant species under consideration. As such they form the principal unit for which the suite of metrics is computed. OG compositions are used directly to quantify features such as universality (UNI; the proportion of species present in an OG) or duplicability (DUP; the proportion of species that have multicopy orthologs). They are used as inputs for gene copy-number turnover analysis to quantify gain (expansion) and loss (contraction) events. Their aligned sequences are

used to compute protein- and DNA-level divergence metrics per OG. Nucleotide-level measurements from whole genome alignment or population variation data are computed over each gene's coding-sequence length and averaged over multicopy orthologs in an OG. Compositions of families range from just a single OG for prophenoloxidases (PPO, 9 *An. gambiae* genes, 3 *D. melanogaster* genes), to 23 OGs with 28 *An. gambiae* genes for small regulatory RNA pathway members (SRRP) or 29 OGs with 37 *D. melanogaster* genes for C-type lectins (CTL) (table 2). The suite of metrics represents a comprehensive quantitative framework to enable detailed evolutionary feature profiling analyses, here applied to 298 OGs containing 420 *An. gambiae* immune-related genes and 276 OGs with 354 *D. melanogaster* immunity genes.

The Evolutionary Feature Landscape of Mosquito Immunity

Profiles built from the 18 quantified evolutionary features successfully delineate key similarities and differences amongst the catalog of 36 canonical mosquito immune-related gene

Table 2. The *Anopheles gambiae* and *Drosophila melanogaster* Immunity Gene Catalogs.

Acronym	Summary description	<i>An. gambiae</i>		<i>D. melanogaster</i>	
		Genes	OGs	Genes	OGs
GALE	Galectins bind specifically to β -galactoside sugars and can function as pattern recognition receptors in innate immunity	9	6	6	5
GNBP	Gram-negative binding proteins (or β -1,3-glucan-binding proteins) are a family of carbohydrate-binding pattern recognition receptors	7	3	3	3
PGRP	Peptidoglycan recognition proteins are pattern recognition receptors capable of recognizing the peptidoglycan from bacterial cell walls	7	5	12	6
SCRA	Scavenger receptors are made up of different classes that function as pattern recognition receptors for a broad range of ligands including from pathogens	5	5	5	4
SCRB		13	10	14	9
CTL	C-type lectins are carbohydrate-binding proteins with roles in pathogen opsonization, encapsulation, and melanization, as well as immune signaling cascades	25	20	37	29
FREP	Fibrinogen-related proteins (also known as FBNs) are a family of pattern recognition receptors with homology to the C terminus of the fibrinogen β and γ chains	38	15	13	6
LRIM	Leucine-rich repeat immune proteins are mosquito immune factors that activate complement-like defense responses against pathogens	24	20	0	0
ML	MD-2-like proteins, also known as Niemann-Pick Type C-2 proteins, possess myeloid-differentiation-2-related lipid-recognition domains involved in recognizing lipopolysaccharide	16	7	8	5
NIMROD	Nimrods have been shown to bind bacteria leading to their phagocytosis by hemocytes, they contain epidermal growth factor-like domains	3	3	12	8
TEP	Thioester-containing proteins are related to vertebrate complement factors and α 2-macroglobulin protease inhibitors, their activation through proteolytic cleavage leads to phagocytosis or killing of pathogens	10	5	5	5
IMDSIG	The immune deficiency pathway is characterized by peptidoglycan recognition protein receptors, intracellular signal transducers (IMDSIG) and modulators (IMDMOD), and the NF- κ B transcription factor Relish	9	9	10	10
IMDMOD		6	6	6	6
JASTSIG	The JAK and the STAT are two core components of the JAK/STAT pathway, with signal transducers (JASTSIG) and modulators (JASTMOD) involved in cellular responses to stress or injury	3	3	6	6
JASTMOD		3	3	4	4
TOLLSIG	The intracellular components of the Toll pathway are homologous to the toll-like receptor innate immune pathway in mammals, with signal transducers (TOLLSIG) and modulators (TOLLMOD) culminating in activation of the NF- κ B transcription factors Dorsal	5	5	6	6
TOLLMOD		8	8	8	8
CASP	Caspases are cysteine-aspartic proteases involved in immune signaling cascades and apoptosis	15	6	7	5
CLIPA	Subfamilies of CLIP-domain serine proteases are defined by patterns of cysteine residues, several CLIPs have roles as activators or modulators of immune signaling cascades	20	13	12	10
CLIPB		27	20	15	13
CLIPC		8	6	7	7
CLIPD		9	8	10	10
CLIPE		9	7	3	3
IAP	Inhibitors of apoptosis are important in antiviral responses and are involved in regulating immune signaling and suppressing apoptotic cell death	8	5	4	4
SRPN	Serine protease inhibitors, or serpins, modulate many signaling cascades; they act as suicide substrates to inhibit their target proteases	18	16	30	20
AMP	Antimicrobial peptides are the classical effector molecules of innate immunity; they include defensins, cecropins, and attacins that are involved in bacterial killing by disrupting their membranes	9	8	10	5
LYS	Lysozymes are key effector enzymes that hydrolyze peptidoglycans present in the cell walls of many bacteria, causing cell lysis	7	1	17	3
PPO	Prophenoloxidases are key enzymes in the melanization cascade that helps to kill invading pathogens and is important for wound healing	9	1	3	1
GPX	Glutathione, heme, and thioredoxin peroxidases are enzymes involved in the metabolism of reactive oxygen species that are toxic to pathogens	2	2	2	2
HPX		15	10	10	9
TPX		5	5	6	6
SOD	Superoxide dismutases are antioxidant enzymes involved in the metabolism of toxic superoxide into oxygen or hydrogen peroxide	4	4	4	4
APHAG	Autophagy-related genes participate in a form of cell death characterized by the formation of an internal autophagosome where pathogens are degraded	19	19	22	22
SRRP		28	23	22	20

(continued)

Table 2. Continued

Acronym	Summary description	<i>An. gambiae</i>		<i>D. melanogaster</i>	
		Genes	OGs	Genes	OGs
SPZ	Small regulatory RNA pathway members are involved in RNA interference and include argonautes, dicers, piwis, and helicases Spaetzle-like proteins contain a cysteine knot domain, the cleavage of Spaetzle results in binding of the product to the Toll receptor and subsequent activation of the Toll pathway	5	5	6	6
TOLL	Toll receptors connect extracellular pathogen recognition to intracellular Toll pathway signaling and activation of immune defense responses	12	6	9	6
Totals:		420	298	354	276

NOTE.—Brief descriptions of immune gene families or pathway components are presented along with counts of the numbers of genes and OGs for the mosquito and fly catalogs.

families and subfamilies (fig. 1). The evolutionary feature profiles for all families are visualized by averaging the metrics over all OGs with genes belonging to each family. Contrasting the profile of a given family against the profiles of all other immune-related families reveals the evolutionary features that most clearly distinguish each family (fig. 1; supplementary fig. S1, Supplementary Material online). This is clearly illustrated by the leucine-rich repeat immune genes (LRIMs) comprising 24 *An. gambiae* genes from 20 OGs, members of which interact with thioester-containing proteins (TEPs) to activate complement-like responses against pathogens (Povelones et al. 2009; Levashina and Baxter 2018). Their taxonomic age (AGE) and UNI are significantly lower, consistent with there being no detectable LRIM orthologs beyond mosquitoes (Waterhouse et al. 2010). They also exhibit fairly typical low DUP, average copy-number (ACN), and copy-number variation (CNV), reflecting their mostly single-copy ortholog status across mosquitoes. These metrics describe the family as a whole although allowing for differences amongst members, for example, the gene duplications that gave rise to three *APL1/LRIM2* paralogs in one lineage of African *Anopheles* (Mitri et al. 2020). Estimates of nonsynonymous substitutions per nonsynonymous site (PDN) are higher than for other families, and significantly so. They are not as extreme, but still significantly higher than other families, for synonymous substitutions per synonymous site (PDS). Together this produces PDN:PDS ratios (SEL, i.e. dN/dS ratios) that are significantly higher than other families, consistent with positive selection or relaxed constraint as observed in previous genus-wide analyses (Neafsey et al. 2015).

Gene gain/loss estimates for the LRIMs show significantly fewer expansions (EXP) and significantly more contractions (CON), but overall stability (STA) close to the mean, in agreement with the copy-number metrics. Conservation of genomic neighborhood, or synteny (SYN), is slightly lower than average for LRIMs, although they notably show the most extreme significantly elevated protein sequence evolutionary divergence (EVR). Single nucleotide polymorphism (SNP) data also show a significantly elevated proportion of nonsynonymous SNPs (NSP) and significantly above average nonsynonymous SNP density (NSD), with synonymous SNP density (SSD) slightly below the mean. The family as a whole

thus appears to reflect the natural diversity and polymorphism observed for some family members (Rottschaefer et al. 2011; Holm et al. 2012). Finally, whole genome alignment data show that LRIMs are significantly less alignable (whole genome alignability [WGA]) and significantly less constrained (per-nucleotide levels of constraint [PHC]) than other immune gene families, reflecting the patterns observed with protein- and DNA-based measures of sequence divergence.

Family profiles highlight the extent to which each family deviates from or matches the typical metric values for each evolutionary feature. GNBP are characterized by high values for metrics capturing gene duplications (DUP and ACN) with high alignability across mosquito genomes (WGA), consistent with the birth of the B-type GNBP subfamily in the mosquito ancestor (Bartholomay et al. 2010). In contrast, Imd pathway signaling genes (IMDSIGs) are characterized as being relatively ancient (high AGE and UNI) and copy-number stable (low CON and high STA) with nevertheless a high protein sequence evolutionary rate (EVR), in agreement with previously observed evolutionary dynamics of immune signaling pathway members (Waterhouse et al. 2007). The subfamilies of CLIP-domain serine proteases are characteristically young (low AGE and UNI), except for CLIPDs which are older and significantly more taxonomically widespread (UNI), a contrast also reflected by several other evolutionary features. Differences amongst CLIP subfamilies could relate to the roles of catalytic and noncatalytic members in modulatory cascades and their hierarchies (El Moussawi et al. 2019).

The autophagy (APHAG) and SRRPs share many features that are significantly different from the mean: They are ancient (high AGE and UNI), stable (low CON and high STA), and constrained (low SEL, EVR, NSP, NSD with high WGA and PHC). However they differ markedly with respect to estimates of dN and dS with both PDN and PDS being significantly lower for APHAGs and significantly higher for SRRPs. Their overall conservation and stability is consistent with both autophagy and RNAi being ancient cellular processes with roles beyond immunity, although their contrasting levels of substitutions could reflect different structural constraints on protein–protein versus protein–RNA interactions. The SRRPs do show above average DUP and ACN values, but not significantly so, consistent with reported single-copy orthologs of

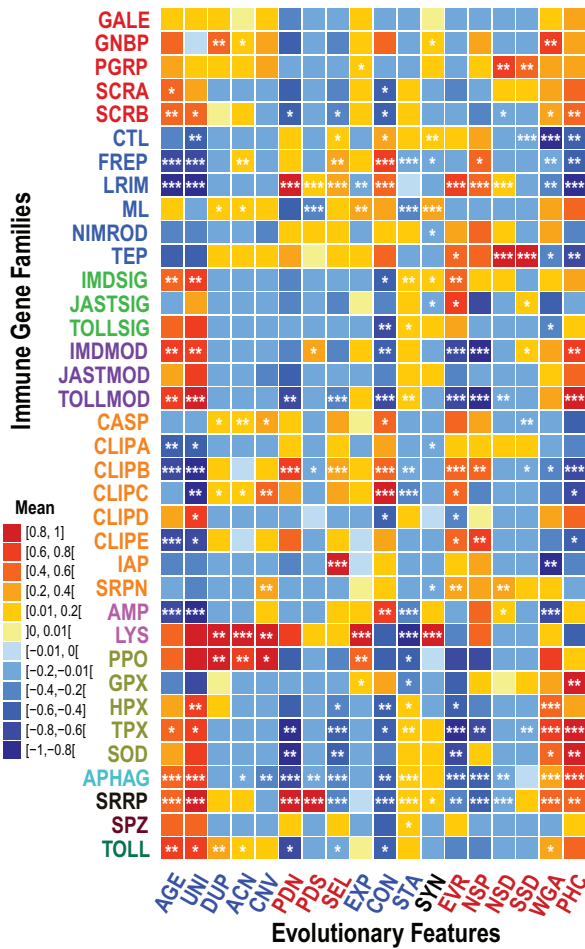


Fig. 1. Evolutionary feature profiles of mosquito immune gene families. Evolutionary profiling highlights similar and contrasting patterns across all 36 immune gene families or subfamilies (rows). Deviations from the typical metric values for the suite of 18 evolutionary feature metrics (columns) are computed as the difference between the family mean and the average over all OGs from other immune-related gene families ($\Delta\bar{x}$). For visualization, values of $\Delta\bar{x}$ are scaled by the absolute maximum $\Delta\bar{x}$ per metric, that is, for each metric the distribution is transformed by dividing all values by the absolute maximum $\Delta\bar{x}$. Values therefore range from a minimum of -1 for metrics where the largest deviation is below the mean, that is lower than other families, and the maximum of 1 for metrics where the largest deviation is above the mean, that is higher than other families. The significance of the difference of the distribution of metric values (no scaling) for each family compared with all other families was assessed using the Wilcoxon rank-sum (Mann–Whitney U) test and a permutation test (asterisks correspond to the lower P value from these two tests; $***P \leq 0.01$, $**P \leq 0.05$, $*P \leq 0.1$). Feature acronyms are defined in table 1. Family acronyms are defined in table 2 and are colored according to categories defined based on their putative roles in the principal immune phases: classical recognition (red), other recognition (blue), pathway signaling (bright green), pathway modulation (purple), cascade modulation (orange), antimicrobial effectors (pink), effector enzymes (olive green), autophagy (dark cyan), RNAi (black), cytokines (brown), and toll receptors (dark green). See text for definitions of evolutionary feature acronyms: taxonomic spread and copy-number features in blue; sequence-based features in red. Evolutionary feature profiles of mosquito immune gene families with median differences ($\Delta\bar{x}$) are presented in supplementary figure S1, Supplementary Material online.

Argonautes 1 and 2 and duplications of *Piwi/Aubergine* in mosquitoes (Lewis et al. 2016). Indeed previous analyses of SRRPs suggested faster evolution in *Aedes* and *Culex* rather than *Anopheles* mosquitoes (Campbell et al. 2008), so conservative patterns observed here could be driven by the data set consisting mainly of anophelines.

The distributions of computed OG metrics for all of the mosquito immune gene families for each evolutionary feature are presented in supplementary additional file 1, Supplementary Material online together with statistical assessments of the significance of deviations from the typical metric values. The trends and significant differences observed across the suite of quantified features facilitate evolutionary profiling that recovers previous mostly qualitative observations and highlights similar and contrasting patterns across all immune gene families (table 3).

Families with Similar Functional Roles Exhibit Similar Evolutionary Profiles

Several bootstrap-supported groupings of families and subsets of features are revealed when hierarchical clustering is applied to the matrix of evolutionary feature profiles of all mosquito immune gene families (fig. 2). Clustering aims to objectively delineate the hierarchical similarities amongst families and features to identify subsets of features that vary in concert, and groups of evolutionarily similar families (see Materials and Methods). Employing family median (fig. 2) and mean (supplementary fig. S2, Supplementary Material online) metric values to build a dissimilarity matrix with Pearson’s correlation distances and performing bootstrapped clustering with the average linkage method results in several well-supported subsets and groupings. Using Pearson’s correlation distances for clustering aims to give weight to the metric directionalities rather than their magnitudes or ranks (Kassambara 2017), to identify families with similar evolutionary feature profiles. Nevertheless, clustering with alternative distance functions (Spearman’s and Kendall’s correlation, and Euclidean distances) and additional agglomerative clustering methods (Single, Complete, and Median linkage) confirms support for many of the observed hierarchical similarities (supplementary figs. S3–S6, additional file 2, Supplementary Material online). Furthermore, clustering using principal components instead of the metric values themselves also identifies several of the observed family groupings (supplementary fig. S7, Supplementary Material online). Overall, there are four main subsets of evolutionary features that consistently cluster together and somewhat more variable groupings of gene families depending on the combinations of metrics and methods applied.

First, with respect to evolutionary features (see table 1 for feature summary descriptions), four subsets of features are repeatedly and robustly recovered: (i) PAML’s dN, PAML’s dN/dS (SEL), the proportion of nonsynonymous SNPs, CON (gene losses), and evolutionary rate (protein sequence divergence); (ii) densities of synonymous and nonsynonymous SNPs; (iii) ACN, EXP, duplications, and CNV, often also including SYN as in figure 2; and (iv) age, universality, constraint, and alignability, often also

Table 3. Characteristic Evolutionary Features of Immune Gene Families and Subfamilies.

Family	Significantly higher	Significantly lower	Interpretation summary
GALE	–	–	No extreme features
GGBP	DUP, ACN, SYN, WGA	–	Duplications, maintained neighborhood, widely alignable
PGRP	EXP, NSD, SSD	–	Duplications, population variation
SCRA	AGE	CON	Ancient, stable copy-number
SCRB	AGE, UNI, WGA, PHC	PDN, SEL, CON, NSD	Ancient, widespread, widely alignable, constrained sequence, constrained substitutions, stable copy-number, population conservation
CTL	SEL, CON, SYN	UNI, SSD, WGA, PHC	Relaxed substitutions, losses, maintained neighborhood, widespread, population conservation, sparsely alignable, relaxed sequence
FREP	ACN, SEL, CON, NSP	AGE, UNI, STA, SYN, WGA, PHC	Duplications, relaxed substitutions, losses, amino acid divergence, young, sparse, unstable copy-number, shuffled neighborhood, sparsely alignable, relaxed sequence
LRIM	PDN, PDS, SEL, CON, EVR, NSP, NSD	AGE, UNI, EXP, WGA, PHC	Relaxed substitutions, losses, amino acid divergence, population variation, young, sparse, stable copy-number, sparsely alignable, relaxed sequence
ML	DUP, ACN, EXP, SYN	PDS, STA	Duplications, maintained neighborhood, constrained substitutions, unstable copy-number
NIMROD	–	SYN	Shuffled neighborhood
TEP	EVR, NSD, SSD	WGA, PHC	Amino acid divergence, population variation, sparsely alignable, relaxed sequence
IMDSIG	AGE, UNI, STA, SYN, EVR	CON	Ancient, widespread, stable copy-number, maintained neighborhood, amino acid divergence
JASTSIG	EVR, SSD	SYN	Amino acid divergence, population variation, shuffled neighborhood
TOLLSIG	STA	CON, WGA	Stable copy-number, sparsely alignable
IMDMOD	AGE, UNI, PDS, SSD, PHC	CON, EVR, NSP	Ancient, widespread, relaxed synonymous substitutions, population variation, constrained sequence, stable copy-number, amino acid conservation
JASTMOD	–	–	No extreme features
TOLLMOD	AGE, UNI, STA, PHC	PDN, SEL, CON, EVR, NSP, NSD	Ancient, widespread, stable copy-number, constrained sequence, relaxed substitutions, amino acid divergence, population variation
CASP	DUP, ACN, CNV, CON	SSD	Duplications, losses, population conservation
CLIPA	–	AGE, UNI, SYN	Young, sparse, shuffled neighborhood
CLIPB	PDN, SEL, CON, EVR, NSP	AGE, UNI, PDS, STA, SSD, WGA, PHC	Relaxed substitutions, losses, amino acid divergence, young, sparse, constrained synonymous substitutions, unstable copy-number, population conservation, sparsely alignable, relaxed sequence
CLIPC	DUP, ACN, CNV, CON, EVR	UNI, STA, PHC	Duplications, losses, amino acid divergence, sparse, unstable copy-number, relaxed sequence
CLIPD	UNI	CON, EVR	Widespread, stable copy-number, amino acid conservation
CLIFE	EVR, NSP	AGE, UNI, PHC	Amino acid divergence, young, sparse, relaxed sequence
IAP	SEL	WGA	Relaxed substitutions, sparsely alignable
SRPN	CNV, EVR, NSD	SYN	Duplications, amino acid divergence, shuffled neighborhood
AMP	CON, NSD	AGE, UNI, STA, WGA	Losses, amino acid divergence, young, sparse, unstable copy-number, sparsely alignable
LYS	DUP, ACN, CNV, EXP, SYN	STA	Duplications, maintained neighborhood, unstable copy-number
PPO	DUP, ACN, CNV, EXP	STA	Duplications, unstable copy-number
GPX	EXP, PHC	STA	Duplications, constrained sequence, unstable copy-number
HPX	UNI, STA, WGA	SEL, CON, EVR	Widespread, stable copy-number, widely alignable, relaxed substitutions, amino acid conservation
TPX	AGE, UNI, STA, WGA, PHC	PDN, SEL, CON, EVR, NSP, SSD	Ancient, widespread, stable copy-number, widely alignable, constrained sequence, constrained substitutions, amino acid conservation, population conservation
SOD	WGA, PHC	PDN, SEL, EVR	Widely alignable, constrained sequence, constrained substitutions, amino acid conservation
APHAG	AGE, UNI, STA, WGA, PHC	ACN, CNV, PDN, PDS, SEL, CON, EVR, NSP, NSD	Ancient, widespread, stable copy-number, widely alignable, constrained sequence, constrained substitutions, amino acid conservation, population conservation
SRRP	AGE, UNI, PDN, PDS, STA, SYN, WGA, PHC	SEL, CON, EVR, NSP, NSD	Ancient, widespread, relaxed substitutions, stable copy-number, maintained neighborhood, widely alignable, constrained sequence, amino acid conservation, population conservation
SPZ	STA	–	Stable copy-number
TOLL	AGE, UNI, DUP, ACN, WGA	PDN, SEL, CON	Ancient, widespread, duplications, widely alignable, constrained substitutions

NOTE.—For each immune-related immune family, evolutionary features with significantly higher or significantly lower metrics compared with other immune families are listed with summarized interpretations.

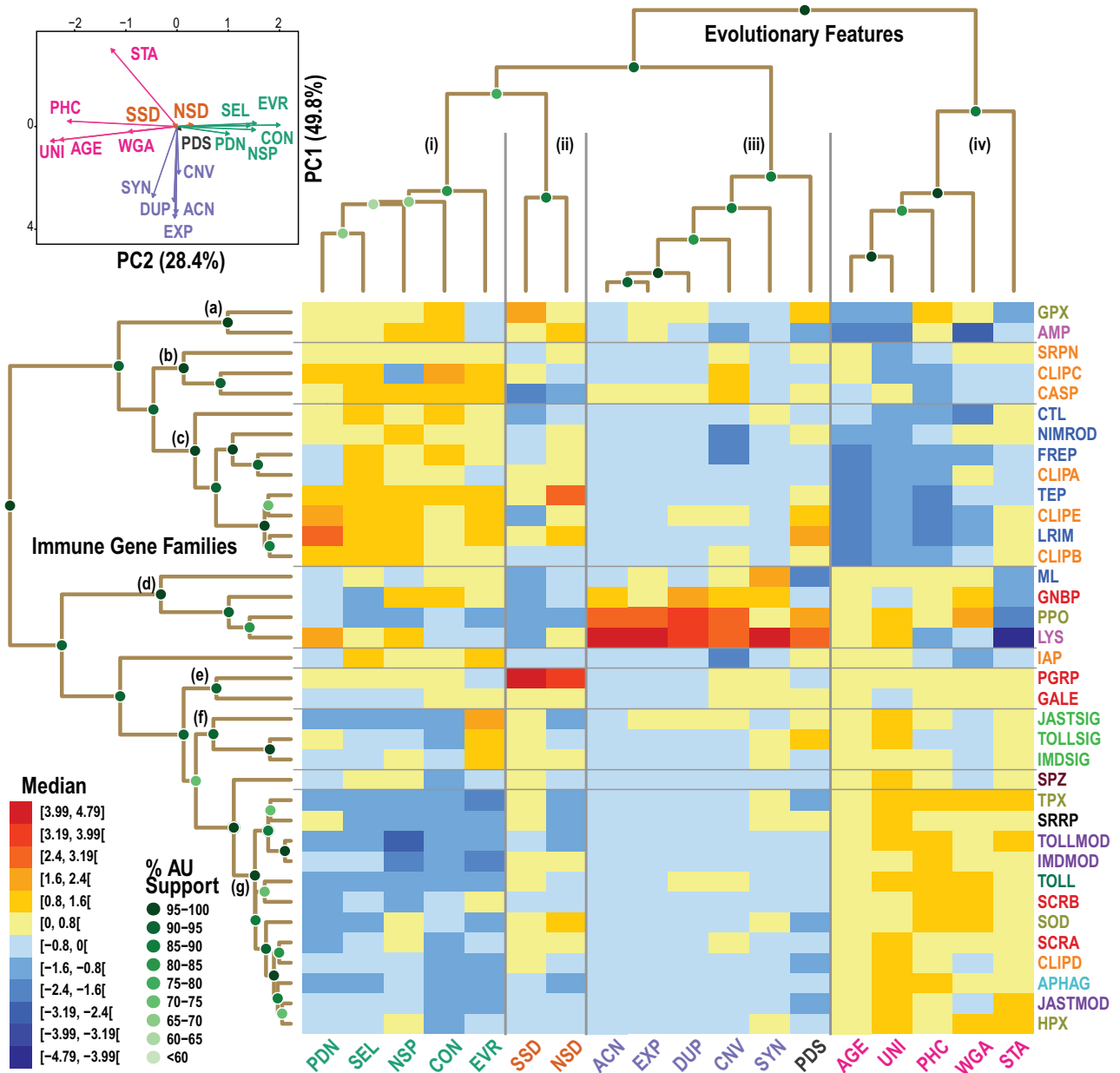


Fig. 2. Clustering heatmap and dendrograms of immune families and their evolutionary features. Groupings of families and subsets of features delineated by hierarchical clustering using the matrix of evolutionary feature profiles of all immune gene families. Hierarchical clustering results are visualized for the immune families ($n = 36$) and evolutionary features ($n = 18$) using scaled median metrics with a Pearson’s correlation-based distance matrix and average linkage agglomerative clustering. The heatmap displays the relative values of the scaled metrics from low in blue to high in red. The dendrograms show the quantified distances (similarities) between each of the families, and between each of the features, and their groupings, determined by the clustering algorithm and distance method. Support for each node of the two dendrograms is shown with green-filled circles, using multiscale bootstrap resampling to estimate AU support values. PCA supports three major groupings of the four subsets of evolutionary features with PC1 and PC2 capturing 78.2% of the variance. Feature acronyms are defined in table 1. Family acronyms are defined in table 2 and are colored according to categories defined based on their putative roles in the principal immune phases: classical recognition (red), other recognition (blue), pathway signaling (bright green), pathway modulation (purple), cascade modulation (orange), antimicrobial effectors (pink), effector enzymes (olive green), autophagy (dark cyan), RNAi (black), cytokines (brown), and toll receptors (dark green). See text for definitions of evolutionary feature acronyms, colored according to groupings in the dendrogram and PCA. Clustering heatmap and dendrograms of immune families and their evolutionary features using mean metrics are presented in supplementary figure S2, Supplementary Material online.

including stability as in figure 2. These subsets are also recovered when clustering using metric means rather than medians, with the exception of PAML’s dS (supplementary fig. S2, Supplementary Material online). Principal component analysis (PCA) of both the median and mean

metrics supports three major groupings of the four subsets, with PC1 dominated by set (iii) features contrasted by stability, and with PC2 clearly separating set (i) from set (iv) features (fig. 2; supplementary figs. S2 and S8, Supplementary Material online).

Set (i) captures both gene losses and several features related to protein sequence divergence. PAML's dN and dN/dS are computed from codon analysis of multiple protein sequence alignments, and the evolutionary rates are computed from amino acid similarities from all-against-all protein alignments, thus they are expected to vary in concert. The observed grouping of the proportion of nonsynonymous SNPs with these protein-alignment-based metrics suggests that long-term divergence over millions of years of mosquito evolution is reflected in population-level polymorphism today. The grouping of gene losses with these sequence divergence and diversity features may appear less intuitive; however, correlations between the propensity for gene loss and sequence evolutionary rates have been observed previously from analyses of orthologs from seven distantly related eukaryotes (Krylov 2003). Here with a larger set of more closely related species (43 insects but mostly mosquitoes and other dipterans) this pattern is recovered while focusing exclusively on immune-related genes. Set (ii) groups together the expected correlated densities of genome-wide synonymous and nonsynonymous SNPs.

Set (iii) captures all the copy-number features related to gene birth, linked to local genomic organization (SYN). Gene duplications lead to higher and often more variable copy-numbers that are identified by computational analysis of gene family evolution (CAFE) as EXP, so these metrics should define different aspects of these features arising from the same underlying process and hence are expected to vary in concert. The link with maintained SYN suggests that duplicated genes often also maintain their local genomic neighborhoods. However, this phenomenon is driven by only a small subset of families with both elevated DUP and SYN metrics: GNBPs, MD-2-like proteins (MLs), and particularly lysozymes (LYSs; fig. 1). For these immune genes it appears that retaining their relative genomic locations played an important role in maintaining their functionalities after duplicating in the mosquito or anopheline ancestor. Set (iv) captures the taxonomic spread features together with DNA-level sequence conservation and constraint, linked to gene family copy-number stability. This grouping clearly connects conservation at whole-gene and nucleotide levels, with older widespread immunity genes generally showing signs of greater constraints. In general, older genes do appear to evolve more slowly than younger ones (Albà and Castresana 2005); they are also longer, more highly expressed, and subject to stronger purifying selection (Wolf et al. 2009). In addition to constrained sequence evolution, genes functionally characterized as essential are more likely to be ancient and widespread (Waterhouse et al. 2011). This highlights the ancient origins and essential roles of several core components of the insect immune system that have been maintained over millions of years of evolution.

Clustering with a subset of 12 evolutionary features after excluding PAML-based (dN, dS) and variation-based (SNPs) metrics recovers sets (i), (iii), and (iv) observed with the full suite of metrics (supplementary figs. S9 and S10, Supplementary Material online). Thus the associations between gene loss and protein sequence divergence, between

DUP and SYN, and between taxonomic spread and DNA-level sequence conservation, are identifiable using this subset of features. Performing the same clustering analyses with the *D. melanogaster* immune gene catalog also recovers the links between gene loss and protein sequence divergence, and between taxonomic spread and DNA-level sequence conservation (supplementary figs. S11 and S12, Supplementary Material online). However, despite MLs and LYSs showing the same trend as for *An. gambiae*, SYN is no longer associated with copy-number features related to gene birth, indicating that maintaining genomic neighborhoods after gene duplication events is a family-dependent phenomenon rather than a global trend. The GNBPs offer a specific example, where the birth of the B-type GNBPs in the mosquito ancestor produced a new subfamily with members showing elevated conservation of their genomic neighborhoods.

The evolutionary profiles describe contrasting features of gene families and pathway members implicated in immune responses. The suite of features covers a wide spectrum of gene family evolutionary dynamics that can be broadly summarized by three main axes delineated by the major PCA groupings: axis 1, DUP and SYN; axis 2, maintenance/stability and sequence conservation; and axis 3, loss and sequence divergence. Axis 1 might be driven by only a subset of families, but the pattern is intuitive when considering the advantage of maintaining expression regulatory coordination across sets of duplicated genes. Axes 2 and 3 appear to reflect global trends in gene evolutionary dynamics observed in different taxa and over different timescales, suggesting that a broadly common set of rules also applies to the evolution of components of the immune system.

With respect to gene families (see table 2 for family summary descriptions), several groupings of different sizes are recovered: from top to bottom in figure 2 (a) AMPs and glutathione peroxidases; (b) cysteine-aspartic and CLIPC proteases with serine protease inhibitors; (c) LRIMs, TEPs, CLIPA protease homologs, CLIPB&E proteases, CTL, and fibrinogen-related and Nimrod proteins; (d) GNBPs, MLs, lysozymes, and PPOs; (e) PGRPs and galectins; (f) Toll, Imd, and JAK/STAT signaling proteins; and (g) a large set comprising autophagy and RNAi-related proteins, Toll, Imd, and JAK/STAT pathway modulators, toll receptors, scavenger receptors A and B, CLIPD proteases, superoxide dismutases, as well as heme and thioredoxin peroxidases. Clustering with metric means rather than medians results in different hierarchies but with several broadly similar groupings including: LRIMs, CLIPA protease homologs, CLIPB&E proteases, and fibrinogen-related proteins; cysteine-aspartic and CLIPC proteases; GNBPs, MLs, lysozymes, and PPOs; and a large set comprising autophagy and RNAi-related proteins, Toll, Imd, and JAK/STAT pathway modulators, toll receptors and galectins, scavenger receptors A and B, CLIPD proteases, superoxide dismutases, as well as heme and thioredoxin peroxidases (supplementary fig. S2, Supplementary Material online). Similar variations of these groupings are obtained when clustering means or medians using alternative distance-clustering method combinations (supplementary additional file 2, Supplementary Material online). Combining this variation with results from

bootstrapping provides a measure of evolutionary profile similarity between all pairs of families (see Materials and Methods). The families that most frequently cluster together using metric means (supplementary fig. S5, Supplementary Material online) or medians (supplementary fig. S6, Supplementary Material online) include: PGRPs, galectins, GNBP, MLs, lysozymes, and PPOs; cysteine-aspartic and CLIP proteases; LRIMs, TEPs, CLIPA protease homologs, CLIPB&E proteases, and fibrinogen-related proteins; and a large set comprising autophagy and RNAi-related proteins, Toll, Imd, and JAK/STAT pathway modulators, toll receptors, scavenger receptors A and B, CLIPD proteases, superoxide dismutases, as well as heme and thioredoxin peroxidases. Thus, although the gene family groupings are more variable across different distance-clustering method combinations than those of the evolutionary features, the results identify families with consistently similar evolutionary profiles.

Evolutionary profile clustering identifies features that are shared by genes and families within each of the major immune phases. Pairs of recognition protein families with similar profiles include PGRPs and galectins, A- and B-type scavenger receptors, and GNBP and MLs, also indicating that MLs more closely resemble classical than other recognition families, thereby warranting their reclassification (fig. 2). PGRPs can bind bacterial cell wall Dap- or Lys-type peptidoglycans (Wang et al. 2019), whereas galectins can bind surface β -galactosides (Vasta 2020). Similarly, GNBP can recognize β -1,3-glucans that make up structural polysaccharides of yeast cell walls (Rao et al. 2018), whereas MLs can bind lipopolysaccharides from the outer membrane of Gram-negative bacteria (Shi et al. 2012). A- and B-type scavenger receptors may have broader ligand specificities including lipoproteins and surface molecules of Gram-negative and Gram-positive bacteria (Alquraini and El Khoury 2020). As important pattern recognition receptors in animal immunity, these are all expectedly old families; however, despite interacting with pathogens they remain relatively constrained (DNA-level) and do not show extreme protein sequence divergence (fig. 2). This apparent lack of evidence for an arms race scenario may in fact reflect the relatively limited structural diversity of the main microbial ligands—peptidoglycan, β -1,3-glucan, lipopolysaccharide—they must bind to or cleave.

Signaling genes of the Toll, Imd, and JAK/STAT pathways group together, being generally ancient and stable but with remarkably elevated rates of protein sequence divergence. Their copy-number stability is possibly a reflection of constraints imposed by the large disruptive potential of duplicates on core signal transduction functionality. Their protein products work together as interacting partners, including the death-domain-mediated MyD88-Tube-Pelle complex of the Toll pathway (Valanne et al. 2011), the Imd pathway's Imd-Fadd-Dredd, Tab 2-Tak1, and I κ B kinase complexes (Myllymäki et al. 2014), and the Domeless-Hopscotch complex of the JAK/STAT pathway (Myllymäki and Rämet 2014). Their greater sequence divergence could therefore be explained by the accumulation of compensatory amino acid changes that maintain key interactions amongst these partners, and overall pathway functionality. The signaling

pathway modulators are also old and stable, but instead show constrained sequence evolution. These include several enzymes, such as ubiquitinases like Effete and Bendless, or E3 ligases like Pellino and Pias, which are under strong constraints to maintain their enzymatic activities. They are involved in proteasomal degradation and are therefore also critical for many other processes beyond immune signaling (Glickman and Ciechanover 2002). Other enzymes including the superoxide dismutases as well as the heme and thioredoxin peroxidases involved in reactive oxygen species metabolism (Hillyer 2016), show similarly conservative evolutionary profiles (fig. 2). Proteolytically activated PPOs oxidize phenols in the melanin production process (Nakhleh, El Moussawi, et al. 2017) and also show similar sequence constraints; however, multiple gene duplications result in an evolutionary profile that is radically different. Thus although there is some variation, in general the functional constraints on these types of enzymes appear to restrict their patterns of molecular divergence.

Members of ancient pathways controlling RNAi (SRRP) and autophagy (APHAG) responses group with other conservative evolutionary profiles characterized by low gene turnover and low sequence evolutionary rates (fig. 2). In contrast, much more dynamic evolutionary profiles characterize the grouping of families of immune cascade modulators like CTL, CLIPA protease homologs and CLIPB&E proteases, regulators of melanization responses like serine protease inhibitors, and key players in mosquito complement-like responses, like TEPs and LRIMs. Although melanization is conserved across arthropods (Hillyer 2016), the proteolytic cascades that trigger or dampen melanin production often involve lineage-specific members of large gene families including these dynamically evolving modulators (Gulley et al. 2013; Meekins et al. 2017; Bishnoi et al. 2019; El Moussawi et al. 2019). The complement-like responses centered on TEPs and LRIMs are specific to mosquitoes, and are also triggered and regulated by members of these large and dynamic families (Blandin et al. 2008; Fraiture et al. 2009; Povelones et al. 2009, 2013, 2016). Based on understanding molecular functions of only a limited number of genes from these families, it appears that immune responses requiring such finely tuned activation, amplification, and deactivation processes source components from dynamically evolving families from which to build functional modules. The families involved are characterized with evolutionary profiles showing a pattern of younger and less widespread orthologs, with lower-sequence constraints, and often elevated signatures of selection and population-level variation. This dynamism is more consistent with an arms race scenario, where the effectiveness of such functional modules is continuously being tested by evolving pathogen attacks and evasion strategies.

Coexpression Analyses Identify Immune Families That Function in Concert

Analysis of multisample gene expression data shows that families with the strongest fine-scale or broad-scale expression similarities include many pairs whose members are known to function together in vivo (fig. 3; supplementary

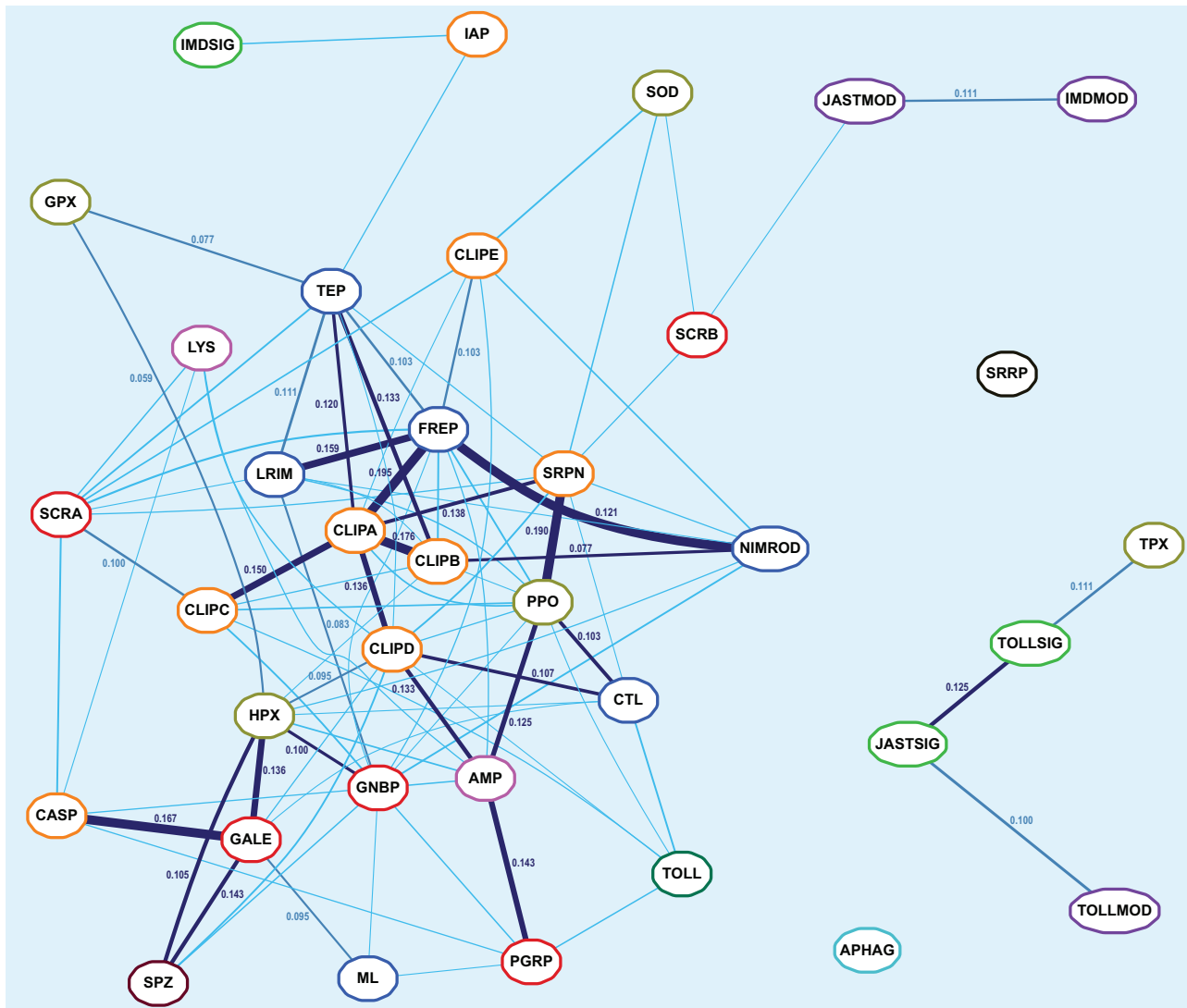


Fig. 3. Network of immune family expression similarities based on the VectorBase Expression Map. The network layout optimized with a spring model provides a 2D visualization of expression similarities for pairwise comparisons of all 36 immune-related gene families computed as gene co-occurrence scores across the VectorBase Expression Map (AgamP4.11 VB-2019-02). Families with more similar gene expression profiles are placed closer together in the graph. Significant co-occurrences are indicated with connecting lines: light blue <0.05 , royal blue <0.01 , and dark blue <0.005 with line thickness scaled to the P value and co-occurrence scores indicated for all pairs with $P < 0.01$. Family acronyms are defined in table 2 and are colored according to categories defined based on their putative roles in the principal immune phases: classical recognition (red), other recognition (blue), pathway signaling (bright green), pathway modulation (purple), cascade modulation (orange), antimicrobial effectors (pink), effector enzymes (olive green), autophagy (dark cyan), RNAi (black), cytokines (brown), and toll receptors (dark green).

fig. S13, Supplementary Material online). Thus, without presupposing any functional categorizations, the similar expression profiles highlight families whose members are likely working together across different conditions. Gene expression-based quantification of functional similarities amongst immune gene families provides an alternative objective classification that complements the classical categorizations based on their putative roles in key immune responses. The VectorBase Expression Map (MacCallum et al. 2011) defines clusters of genes with similar expression profiles for 12,672 genes using normalized data across 202 conditions, enabling the quantification of fine-scale or broad-scale gene expression similarities amongst all pairs of immune-related families. Pairwise family similarities are

computed as the frequency of co-occurrences of gene family members in the same region of the map, with significance assessed taking into account family sizes and expression cluster sizes (see Materials and Methods). Visualizing pairwise family similarities as a spring model layout network optimized with the *neato* tool from the Graphviz package (Gansner and North 2000; Gansner et al. 2005) identifies subsets of families with putative roles in common immune processes (fig. 3). These prominently include a quintet of families with highly and significantly overlapping expression patterns: LRIMs, TEPs, FREPs, CLIPAs, and CLIPBs (fig. 3), with members implicated in coordinating and executing mosquito complement system responses (Povelones et al. 2016; Reyes Ruiz et al. 2019).

Anopheles gambiae TEP1 forms a stable protein complex with a heterodimer of LRIM1 and APL1A/B/C (LRIM2 paralogs) in the hemolymph until the complement response is activated (Fraiture et al. 2009; Povelones et al. 2009; Williams et al. 2015), so coordinated coexpression of these genes is important for their functions. Like the LRIMs and TEPs, the FREPs are also found in the hemolymph, and several members are infection-responsive and important for defense, for example, FREP57/FBN8, FREP13/FBN9, and FREP40/FBN39 (Dong et al. 2006; Dong and Dimopoulos 2009; Simões et al. 2017). FREPs themselves might dimerize or oligomerize, but whether they interact directly with TEPs and/or LRIMs in mosquitoes remains unknown, although evidence from snails indicates that FREPs and TEPs do interact (Li et al. 2020), and the observed expression similarities support at least some functional, if not physical, interaction. CLIPA serine protease homologs are positive and negative regulators of immune responses mediated by TEP1, for example, CLIPA8 (Schnitger et al. 2007), SPCLIP1/CLIPA30 (Povelones et al. 2013), CLIPA2 (Yassine et al. 2014), CLIPA14 (Nakhleh, Christophides, et al. 2017), and CLIPA28 (El Moussawi et al. 2019). These regulatory modules also involve the catalytically active CLIPBs, for example, CLIPB14 and CLIPB15 (Volz et al. 2005), CLIPB8 (Zhang et al. 2016), and CLIPB10 (Zhang et al. 2021), and together CLIPAs and CLIPBs are also key modulators of melanization responses (Volz et al. 2006). The available evidence therefore supports the family-level expression analyses that demonstrate highly and significantly overlapping expression patterns (fig. 3) of members of this quintet of families that function in concert.

Of this quintet, expression of CLIPA protease homologs is additionally strongly and significantly similar to that of CLIPCs, CLIPDs, and SRPNs (serpins, or serine protease inhibitors). The CLIPC9 protease has recently been shown to regulate melanization downstream of SPCLIP1/CLIPA30, CLIPA8, and CLIPA28, and to be inhibited by SRPN2 (Sousa et al. 2020). CLIPC2 may function together with SRPN7 controlling the activation of effector mechanisms (Blumberg et al. 2013). Specific roles for CLIPDs, which show an evolutionary profile distinct from the other CLIPs (fig. 2), remain largely unknown. Serpins themselves are most similar in expression to PPOs, both of which would need to be replenished after being depleted during melanization responses (Gulley et al. 2013; Nakhleh, El Moussawi, et al. 2017). The PPOs in turn appear significantly similar to the CTL, which are generally considered glycan-binding recognition proteins, but at least two members—CTL4 and CTLMA2—are key regulators of melanization downstream of immune recognition (Schnitger et al. 2009; Bishnoi et al. 2019). The family-level expression similarities (fig. 3) therefore derive from the functional links amongst the CLIP, CTL, and SRPN family members that modulate the activation of melanization, and the PPO enzyme effectors of melanization activity.

Amongst classical recognition proteins, PGRPs and GNBPs are most similar, and their expression patterns both closely match those of AMPs and MD-2-like lipid recognition proteins. These similarities are driven by the upregulation of members of these gene families upon infection or following

a blood meal, which promotes growth of the gut microbiota, for example, in response to blood-feeding (Dana et al. 2005), microbes (Aguilar et al. 2005), *Plasmodium* (Dong et al. 2006), or fungi (Ramirez et al. 2020). They are nevertheless not as tightly interconnected as components of the complement and melanization responses, possibly reflecting the contrast between broad-scope protection of these systems versus the generally much more pathogen-specific activities of different families of recognition proteins and antimicrobial effectors. Indeed feeding into and/or being transcriptionally activated by different immune signaling pathways means that these families may be thought of as performing analogous roles rather than functioning in concert per se. However, learning more about signaling crosstalk and response overlap has shifted thinking from traditional functional distinctions amongst immune pathways (Kounatidis and Ligoxygakis 2012). Thus, these similarities might reflect somewhat overlapping responses, but also a common readiness or priming to face newly perceived threats.

Notably, expression patterns of pathway signaling and modulation components remain distinct from the recognition and response families: Imd and JAK/STAT pathway modulators are significantly similar, whereas Toll pathway modulators group together with Toll and JAK/STAT pathway signaling members. Genes involved in RNAi (SRRP) and autophagy (APHAG) responses do not show significant similarities in expression patterns to other families; however, SRRP and APHAG genes have highly and significantly overlapping expression patterns at broad-scale resolution, and are most similar to modulators of all three pathways (supplementary fig. S14, Supplementary Material online). At broad-scale resolution, the distinction between pathway signaling/modulation and recognition/response families is accentuated, whereas the melanization and complement responses become more closely interlinked. Many of the most similar families also show substantially overlapping expression patterns when quantifying similarities across coexpression modules built from a subset of immune-related experimental conditions (see Materials and Methods, supplementary figs. S15 and S16, table S2, additional file 3, Supplementary Material online). For example, families implicated in complement system responses again show similar expression patterns (supplementary fig. S17, Supplementary Material online), and at a broader-scale resolution become more closely associated with melanization responses (supplementary fig. S18, Supplementary Material online). At broad-scale resolution pairs of similar recognition families include GNBPs and PGRPs, GNBPs and MLs, as well as galectins and B-type scavenger receptors, whereas at both resolutions Imd and JAK/STAT pathway signaling members are highly and significantly similar. Multicondition coexpression analysis therefore identifies gene expression similarities amongst sets of immune-related families with members that are known or inferred to function in concert.

Complement-Related Families Exhibit Elevated Evolutionary-Functional Similarities

Immune gene family evolutionary-functional correspondences are revealed by employing quantifications of

evolutionary similarities based on gene family feature profiling and of functional similarities based on gene family expression patterns (fig. 4; supplementary additional file 4, Supplementary Material online). Most prominently, families involved in mosquito complement system responses show both high evolutionary similarities and high fine-scale and broad-scale expression similarities: recognition family pairs of LRIMs–TEPs, FREPs–TEPs, and FREPs–LRIMs, as well as modulator-recognition family pairs of CLIPAs with FREPs and TEPs, and CLIPBs with FREPs, TEPs, and LRIMs. Members of these principal complement–response gene families exhibit common expression and evolutionary profiles suggestive of common constraints. Both TEPs and LRIMs are also highly evolutionarily similar to CLIPs, for which specific roles in complement responses remain largely unknown, but with which their expression similarity increases at broad-scale resolution, albeit remaining nonsignificant. The CLIPA protease homologs and CLIPB proteases form a highly similar pair, but their strong and significant expression similarity is not maintained at broad-scale resolution, suggesting tight functional coupling of these key modulators. Conversely, CLIPB and CLIPB modulators also form a highly similar pair, but with strong and significant expression similarity only at broad-scale resolution. In contrast, FREP-NIMROD expression similarity is maintained at both resolutions and it is amongst the most significant of all family pairs that also show high evolutionary similarities. Although a much smaller gene family than the FREPs, NIMRODs including *draper*, *nimrod*, and *eater*, are also infection-responsive and important for defense (Midega et al. 2013; Estévez-Lao and Hillyer 2014). Combining results from evolutionary profiling and knowledge-blind functional clustering therefore identifies families that appear both evolutionarily and functionally similar. These similarities are notably pronounced for families with members known to function in concert to coordinate and execute mosquito complement system responses (Povelones et al. 2016; Reyes Ruiz et al. 2019).

Additional families with above average evolutionary and expression similarities at both resolutions include another pair of modulators (CLIPA–SRPN), and another modulator-recognition pair (CLIPB–FREP). Although CLIPAs and SRPNs are known to function together in cascades regulating melanization (El Moussawi et al. 2019), potential functional interactions between CLIPs and FREPs remain to be explored. The melanization modulator-effector pair of SRPNs and PPOs shows the highest expression similarity at both resolutions, but with negligible evolutionary similarity, suggesting that regulating these responses and executing them are subject to different constraints. Amongst other recognition proteins, MLs show above average evolutionary and expression similarities to the classical recognition families of galectins (GALEs) at fine-scale resolution, and PGRPs at broad-scale resolution with lower but still significant expression similarity at fine-scale resolution. Compared with galectins or PGRPs, the MLs are evolutionarily more similar to GNBPs, with which they show lower, but still significant, expression similarity. These patterns suggest analogous functionalities—recognition of foreign—with different specificities for

lipopolysaccharides, β -galactosides, peptidoglycans, or β -1,3-glucans, that arise depending on the pathogen/microbe community composition. Common constraints faced by classical recognition phase families appear to produce similarities amongst their evolutionary trajectories, with functional similarities quantified through gene expression patterns possibly arising through immune pathway signaling crosstalk and priming (Kounatidis and Ligoxygakis 2012).

Evolutionarily similar families that only show high expression similarities at broad-scale resolution include modulators of the Imd and Toll pathways (IMDMOD–TOLLMOD) and genes involved in autophagy and RNAi responses (APHAG–SRRP). At fine-scale resolution, pathway components from JAK/STAT and Toll signaling (JASTSIG–TOLLSIG), Imd and JAK/STAT modulation (IMDMOD–JASTMOD), and JAK/STAT signaling and Toll modulation (JASTSIG–TOLLMOD) also show above average evolutionary and expression similarities. These pathways and responses play key roles in processes other than immunity, including in development and morphogenesis, so their gene expression-based functional similarities will vary depending on the conditions examined. This also means that the functional constraints they experience are not solely derived from their roles in immune processes. Their functional similarities are more stably evident when the modules are abstracted to analogous phases of signal input, signal processing, and signal output. Whether functionally similar or analogous, these immune-related pathways and responses exhibit common conservative evolutionary profiles that distinguish them from other more dynamically evolving components of the immune system (fig. 2). These constrained evolutionary features could result from the effects of pleiotropy, and possibly the modular architectures, on the trade-offs during adaptive evolution producing a limited range of available trajectories (Mauro and Ghalambor 2020).

Conclusions

Through quantitative evolutionary feature profiling of genes and gene families, integrated with knowledge- and expression-based functional categorizations, our multispecies comparative immunogenomic analyses identified evolutionary-functional correspondences suggesting that constraints on genes with similar or analogous functions govern their evolutionary trajectories. The profiles delineate whether and how each family deviates from the feature value distributions of other families, and provide the substrate for clustering to define similarities amongst families and features. We employed insect innate immunity as our test case study system because the key implicated pathways and component gene families have been well characterized. While acknowledging that responses to infections involve diverse processes beyond the canonical immune system (Sackton 2019) and that immune-related genes may also function in other biological processes, this prior knowledge provided specific examples and strong expectations of types of genes with similar functions and distinguishing patterns of evolution, enabling the interpretation of observed correspondences

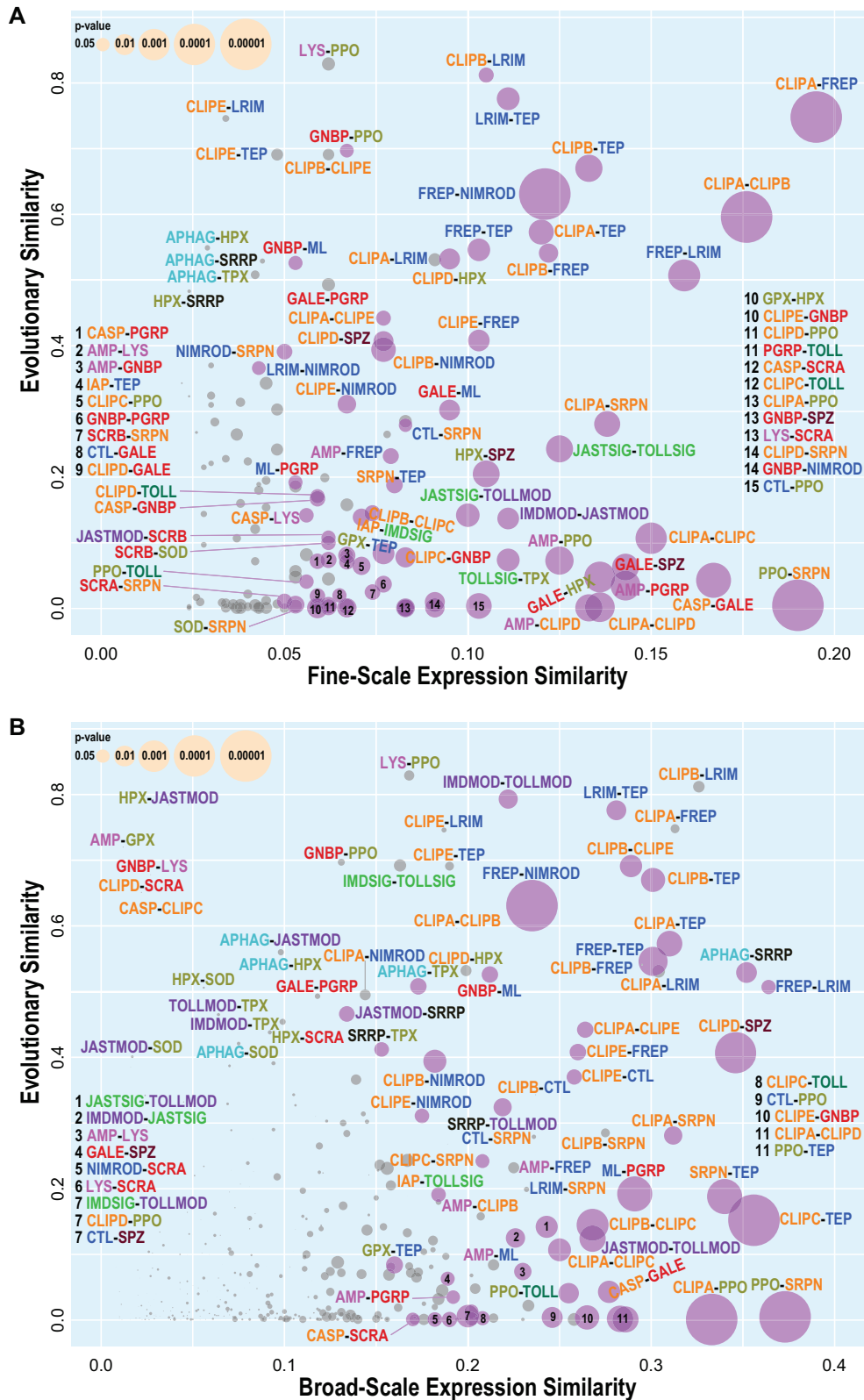


Fig. 4. Pairwise comparisons of immune family expression similarity and evolutionary similarity. Evolutionary similarities (based on feature metric medians) of pairs of gene families are compared with their expression similarities at (A) fine-scale resolution and (B) broad-scale resolution. Pairs of families with significant ($P < 0.05$) gene expression co-occurrence scores are shown with purple circles, with nonsignificant pairs shown in gray, and with circle sizes scaled by the P value. Family acronyms are defined in table 2 and are colored according to categories defined based on their putative roles in the principal immune phases: classical recognition (red), other recognition (blue), pathway signaling (bright green), pathway modulation (purple), cascade modulation (orange), antimicrobial effectors (pink), effector enzymes (olive green), autophagy (dark cyan), RNAi (black), cytokines (brown), and toll receptors (dark green).

within an established framework. Feature analysis within the limits of our study system identified three main axes of evolutionary trajectories characterized by gene duplication and SYN, gene maintenance/stability and sequence conservation, and gene loss and sequence divergence. Clustering highlighted similar and contrasting patterns across these axes amongst subsets of immune gene families. For example, classical recognition families, including the herein reclassified MLs, showed patterns that can be explained by the limited structural diversity of the principal microbial ligands with which they interact. Pathway signaling genes on the other hand exhibited trajectories that could relate to physical interactions of protein complexes and constraints from the effects of pleiotropy and disruptive effects of gene duplicates on signal transduction. Functional similarities defined by coexpression analyses recovered sets of immune-related families with members that are known or inferred to function in concert. Most prominently, these included families involved in the complement system and melanization responses, both of which occur mainly in the hemolymph. Comparing these with feature-based clustering results identified evolutionary-functional correspondences that were particularly striking amongst families with members known to function together in the coordination and execution of complement system responses. Our results suggest that where and how different genes participate in immune defense responses limit the range of possible evolutionary scenarios that are tolerated by natural selection. Our test case analyses of insect immunity that explored approaches to quantify gene evolutionary histories and relate these to gene functions highlight the potential for future applications to advancing understanding of functional constraints on evolution. Further developing and applying such comparative genomics approaches to explore constraints in evolutionary biology could offer opportunities to advance the understanding of how functional constraints on different components of biological systems govern their evolutionary trajectories.

Materials and Methods

Orthology, Variation, Alignment, and Expression Data
OGs of genes were defined using the OrthoDB (Kriventseva et al. 2015) orthology delineation procedure across 21 mosquitoes and 22 other insects (see [supplementary materials](#), orthology data, [Supplementary Material online](#)). OrthoDB employs all-against-all protein sequence alignments to first identify all best reciprocal hits (BRHs) between all genes from each pair of species (Zdobnov et al. 2017). It then uses a graph-based clustering procedure that starts with BRH triangulation to progressively build OGs that include all genes descended from a single gene in the last common ancestor. SNPs for *An. gambiae* PEST, including all synonymous and nonsynonymous SNPs in annotated coding regions, were retrieved using the BioMart data mining tool from VectorBase (Giraldo-Calderón et al. 2015). The SNPs derive from eight variation data sets hosted at VectorBase (Neafsey et al. 2010; White et al. 2011; Weetman et al. 2012; Markianos et al. 2016; Hammond et al. 2017; Miles et al. 2017; Wiltshire et al. 2018).

Multispecies whole genome alignments were generated from the assemblies of 22 mosquitoes available from VectorBase and 36 *Drosophila* available from The National Center for Biotechnology Information ([supplementary table S1](#), [Supplementary Material online](#)). The alignment process starts with pairwise sequence comparisons that are then progressively combined following the species phylogeny using the MultiZ approach of the Threaded Blockset Aligner (Blanchette et al. 2004). Expression data for *An. gambiae* genes were retrieved from VectorBase (Expression Stats VB-2019-06) as log₂ transformed expression values for 13,201 genes across 291 conditions (mean, variance, and number of replicates). Immune gene family coexpression analysis employed these expression statistics using a subset of the conditions to build coexpression modules. Coexpression analysis also employed clusters of genes defined by the VectorBase Expression Map (MacCallum et al. 2011), with gene membership of all clusters/cells retrieved from the AgamP4.11 VB-2019-02 map (comprising 12,672 genes and based on 202 conditions).

Anopheles gambiae and *D. melanogaster* Immunity Gene Catalogs

The catalogs of *An. gambiae* and *D. melanogaster* immune-related genes were built by combining and updating the results of previous comparative immunogenomics studies (Christophides et al. 2002; Waterhouse et al. 2007; Bartholomay et al. 2010; Neafsey et al. 2015). *Anopheles gambiae* and *D. melanogaster* gene and OG membership for 36 immune-related gene families and subfamilies are summarized in [table 2](#).

Orthology-Based Evolutionary Features

Features were quantified as a suite of 13 orthology-based evolutionary metrics per OG that included: the evolutionary age (AGE) of the last common ancestor in terms of millions of years since divergence from the ultrametric species phylogeny; the universality (UNI) computed as the proportion of the total species present; the duplicability (DUP) computed as the proportion of species present with multicopy orthologs; the average ortholog copy number (ACN); the copy number variation (CNV) computed as the standard deviation of ortholog counts per species present divided by the ACN. PAML (Yang 2007) was employed using the M0 model on the alignments of OG member sequences to compute the number of synonymous substitutions per synonymous site (PDS); the number of nonsynonymous substitutions per nonsynonymous site (PDN); and the nonsynonymous to synonymous ratio (SEL). Gene turnover was estimated using the CAFE (Han et al. 2013) tool in order to quantify proportions of gene gains (expansions, EXP), gene losses (contractions, CON), or no copy-number changes (stable, STA). Orthology data combined with genomic location data were used to quantify SYN conservation as the proportion of orthologs that maintain their orthologous neighbors in the genomes of the other species. Finally, the EVR of each OG corresponds to the average rate of protein sequence divergence

normalized by the distance between each pair of species as computed by OrthoDB (Waterhouse et al. 2013).

Variation-Based and Alignment-Based Evolutionary Features

Five additional evolutionary feature metrics were computed from polymorphism data and whole genome alignments. The population genomics data for *An. gambiae* retrieved from VectorBase were used to compute per-gene metrics of the proportion of all coding-sequence SNPs that were nonsynonymous (NSP), as well as the nonsynonymous (NSD) and synonymous (SSD) SNP densities as the number of SNPs divided by the total coding-sequence length. Multispecies whole genome alignments were used to compute per-nucleotide metrics of conservation and constraint. WGA measures the proportion of the full set of 22 mosquitoes or 36 *Drosophila* that were aligned to the *An. gambiae* or *D. melanogaster* reference genomes, respectively, for each nucleotide. PhastCons (Siepel et al. 2005) was used to estimate PHC from the whole genome alignments. Per-nucleotide values were averaged over the full coding-sequence lengths of all genes to obtain per-gene metrics. The variation-based and alignment-based per-gene metrics were averaged over all genes in each OG to obtain the per-OG values for each of the metrics.

Gene Family Metrics and Comparisons

The canonical immunity gene catalogs define immunity gene membership of subfamilies (e.g. cecropins, defensins, attacins), families (e.g. AMPs), and broader categories (e.g. antimicrobial effectors), and the orthology data sets define gene membership of OGs. Thus, the gene family evolutionary metrics were computed by averaging values over all OGs containing genes belonging to each cataloged immune gene family. These family-level means for each metric were compared with the means of all other OGs that contain at least one *An. gambiae* immune gene to quantify the extent to which the metrics of the OGs of a given immune gene family differ from all other immune gene containing OGs, that is, delta-mean ($\Delta\bar{x}$). For graphical visualization, $\Delta\bar{x}$ values were scaled by dividing by the absolute maximum $\Delta\bar{x}$ per evolutionary feature and plotted with the color-blind safe RdYlBu palette from the *RColorBrewer* package from R (R Core Team 2021). The Wilcoxon rank-sum (Mann–Whitney *U*) test implemented in the *wilcox.test* function in R (default two-sided test) was used to test the significance of the difference of the distribution of each family's OGs metric values (no scaling) compared with all other immune-related OGs for all metrics and each family. As several families contain few OGs, a permutation test implemented in R was also used to test the significance of the difference of the metric distributions. Observed $\Delta\bar{x}$ was compared with $\Delta\bar{x}$ from permutations of all OG metric values randomly assigned to size-matched sets. The number of permutation differences that were greater than the observed difference, divided by the total number of permutations provides an empirical estimate of the probability of obtaining a $\Delta\bar{x}$ greater than the observed $\Delta\bar{x}$ by chance.

Clustering of Gene Family Metrics

To assess and quantify the similarities of the evolutionary feature profiles, hierarchical clustering of the evolutionary features and families was performed with the *hclust* function in R. For the *An. gambiae* analyses these comprised 18 features and 36 families, whereas for the mosquito-fly comparisons these comprised a common subset of 12 features and 35 families. For all evolutionary feature metrics, both the means and the medians of all OGs per family were assessed. Prior to clustering, the *scale* function in R was used to normalize all metric values by subtracting the means and then dividing the (centered) values by their standard deviations. Dissimilarity matrices were computed with the normalized metric values using three correlation-based distance methods and the Euclidean distance method in R. Clustering with *hclust* was performed with all dissimilarity matrices using single, complete, average, and median linkage agglomeration methods. To estimate statistical support for the clustering of families and features, 10,000 bootstrap replicates were performed with the *pvclust* R package. In *pvclust*, the approximately unbiased (AU) *P* values are computed using multiscale bootstrap resampling (Suzuki and Shimodaira 2006), and provide a confidence measure for each node of the cluster dendrograms of families and evolutionary features. The robustness of gene family clustering across all 16 tested distance–method combinations was further assessed by quantifying the co-occurrence of all pairs of families within subtrees of all 160,000 *pvclust* bootstrap replicates. This evolutionary profile similarity score (family subtree co-occurrence score) was computed and normalized as follows: $(2 \times \text{co-occurrence of Family 1 and Family 2}) / (\text{co-occurrence of Family 1 with any Family} + \text{co-occurrence of Family 2 with any Family})$. Normalized scores of zero indicate that these pairs of families never appear in the same subtree and scores of one would indicate that they occur as sister lineages in all bootstrap samples from all distance–method combinations. Based on these assessments of clustering stability, the dissimilarity matrix from Pearson's correlation method with the average linkage agglomeration method was selected. Specifically, the bootstrap replication analysis showed that the Pearson's correlation distances with the average linkage method produced the fewest poorly supported nodes (based on AU *P* values) across immune families and evolutionary features (see [supplementary methods, figs. S3 and S4, Supplementary Material online](#)). The hierarchical clustering results were visualized as heatmaps with corresponding family and evolutionary feature dendrograms showing AU support, plotted with the *gplots* and *dendextend* (Gallili 2015) R packages. PCA of the family by feature matrices of both median and mean metrics were performed with the *prcomp* function from the *stats* package in R. As well as producing well-supported nodes, the Pearson.Average distance–method approach on the scaled metrics produces similar family groupings to using the top ten principal components with the standard Euclidean-Ward.D2 distance–method approach ([supplementary fig. S7, Supplementary Material online](#)), that is, when applying standard clustering techniques after transforming the correlated metrics into principal components.

Gene and Family Coexpression Analyses

Gene expression similarities amongst all pairs of *An. gambiae* immune-related families were quantified using the gene expression data and Expression Map (MacCallum et al. 2011) retrieved from VectorBase (Giraldo-Calderón et al. 2015). The map was analyzed to quantify co-occurrences of gene family members in the same cell on the map (fine-scale resolution of gene coexpression), and in the same supercell, the cell and its immediate eight neighboring cells on the map including toroidal neighbors (broader-scale resolution of gene coexpression). Pairwise family cell/supercell co-occurrence scores (expression similarity scores) were computed as the intersection, Family 1 \cap Family 2, divided by the union, Family 1 \cup Family 2 (i.e. number of cells with at least one gene from both Family 1 and Family 2/number of cells with at least one gene from either Family 1 or Family 2). A score of zero: the pair of families have no member genes that cluster in the same cell/supercell. A score of one: all member genes from both families always cluster in cells/supercells with at least one member of the other family. Statistical significance of the family cell/supercell co-occurrence scores was assessed with a permutation test: scores were recomputed after gene to cell assignments were randomly shuffled (10,000 permutations) preserving the total number of cells and families, and the number of genes in each cell and each family. These were used to calculate an empirical estimate of the probability (P value) of obtaining a co-occurrence score greater than the observed co-occurrence score by chance: the number of permutation scores that were greater than the observed score, divided by the total number of permutations. Complementary assessments of gene expression clustering were performed using the weighted correlation network analysis approach (Langfelder and Horvath 2008) on a subset of 24 conditions selected from the VectorBase gene expression data set including blood feeding experiments and tissues from Marinotti et al. (2006), Neira Oviedo et al. (2008), and Baker et al. (2011). Expression similarities of pairs of immune gene families and the significance of their co-occurrences were computed as for the Expression Map but using module membership rather than cell/supercell membership.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors acknowledge the constructive feedback received from the anonymous reviewers. This research was supported by Novartis Foundation for medical-biological research grant #18B116 and Swiss National Science Foundation grants PP00P3_170664, PP00P3_202669, and CRSII5_186397.

Author Contributions

R.M.W. conceived the study. L.R., R.F., M.J.M.F.R., A.T., and R.M.W. designed the analyses. L.R. analyzed orthology data, quantified evolutionary features, and performed clustering analyses and statistical testing. R.F. built whole genome

alignments, and analyzed alignment and variation data. M.R. analyzed variation data. A.T. curated data and assisted with gene expression data analysis. L.R. and R.M.W. wrote the manuscript with input from all authors. All authors read and approved the manuscript.

Data Availability

No new genomic data were produced as part of this study and all data analyzed herein are available from public databases.

Additional File 1: Distributions of computed OG metrics for all of the immune gene families for each evolutionary feature together with statistical assessments of the significance of deviations from the typical metric values.

Additional File 2: Heatmaps and dendrograms resulting from clustering with family mean metric values with different distance functions and agglomerative clustering methods.

Additional File 3: Visualization of expression patterns per expression cluster defined by weighted correlation network analysis, with results from Gene Ontology term enrichment tests.

Additional File 4: Plots of evolutionary similarities versus expression similarities based on evolutionary feature metric means and medians for different sets of expression clusters.

References

- Aguilar R, Jedlicka AE, Mintz M, Mahairaki V, Scott AL, Dimopoulos G. 2005. Global gene expression analysis of *Anopheles gambiae* responses to microbial challenge. *Insect Biochem Mol Biol*. 35(7):709–719.
- Albà MM, Castresana J. 2005. Inverse relationship between evolutionary rate and age of mammalian genes. *Mol Biol Evol*. 22(3):598–606.
- Alquraini A, El Khoury J. 2020. Scavenger receptors. *Curr Biol*. 30(14):R790–R795.
- Antonovics J, van Tienderen PH. 1991. Ontoecogenophyloconstraints? The chaos of constraint terminology. *Trends Ecol Evol*. 6(5):166–168.
- Arensburger P, Megy K, Waterhouse RM, Abrudan J, Amedeo P, Antelo B, Bartholomay L, Bidwell S, Caler E, Camara F, et al. 2010. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science* 330(6000):86–88.
- Baker DA, Nolan T, Fischer B, Pinder A, Crisanti A, Russell S. 2011. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, *Anopheles gambiae*. *BMC Genomics* 12:296.
- Barribeau SM, Sadd BM, du Plessis L, Brown MJ, Buechel SD, Cappelle K, Carolan JC, Christiaens O, Colgan TJ, Erler S, et al. 2015. A depauperate immune repertoire precedes evolution of sociality in bees. *Genome Biol*. 16:83.
- Bartholomay LC, Michel K. 2018. Mosquito immunobiology: the intersection of vector health and vector competence. *Annu Rev Entomol*. 63:145–167.
- Bartholomay LC, Waterhouse RM, Mayhew GF, Campbell CL, Michel K, Zou Z, Ramirez JL, Das S, Alvarez K, Arensburger P, et al. 2010. Pathogenomics of *Culex quinquefasciatus* and meta-analysis of infection responses to diverse pathogens. *Science* 330(6000):88–90.
- Bishnoi R, Sousa GL, Contet A, Day CJ, Hou C-FD, Profitt LA, Singla D, Jennings MP, Valentine AM, Povelones M, et al. 2019. Solution structure, glycan specificity and of phenol oxidase inhibitory activity of *Anopheles* C-type lectins CTL4 and CTLMA2. *Sci Rep*. 9(1):15191.
- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. 14(4):708–715.

- Blandin SA, Marois E, Levashina EA. 2008. Antimalarial responses in *Anopheles gambiae*: from a complement-like protein to a complement-like pathway. *Cell Host Microbe*. 3(6):364–374.
- Blumberg BJ, Trop S, Das S, Dimopoulos G. 2013. Bacteria- and IMD pathway-independent immune defenses against *Plasmodium falciparum* in *Anopheles gambiae*. *PLoS One*. 8(9):e72130.
- Brucker RM, Funkhouser LJ, Setia S, Pauly R, Bordenstein SR. 2012. Insect innate immunity database (IID): an annotation tool for identifying immune genes in insect genomes. *PLoS One*. 7(9):e45125.
- Campbell CL, Black WC, Hess AM, Foy BD. 2008. Comparative genomics of small RNA regulatory pathway components in vector mosquitoes. *BMC Genomics* 9:425.
- Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W, Vontas J, et al. 2015. Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution. *Proc Natl Acad Sci U S A*. 112(44):E5907–E5915.
- Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, et al. 2002. Immunity-related genes and gene families in *Anopheles gambiae*. *Science* 298(5591):159–165.
- Dana AN, Hong YS, Kern MK, Hillenmeyer ME, Harker BW, Lobo NF, Hogan JR, Romans P, Collins FH. 2005. Gene expression patterns associated with blood-feeding in the malaria mosquito *Anopheles gambiae*. *BMC Genomics* 6(1):5.
- Dong Y, Aguilar R, Xi Z, Warr E, Mongin E, Dimopoulos G. 2006. *Anopheles gambiae* immune responses to human and rodent *Plasmodium* parasite species. *PLoS Pathog*. 2(6):e52.
- Dong Y, Dimopoulos G. 2009. Anopheles fibrinogen-related proteins provide expanded pattern recognition capacity against bacteria and malaria parasites. *J Biol Chem*. 284(15):9835–9844.
- El Moussawi L, Nakhleh J, Kamareddine L, Osta MA. 2019. The mosquito melanization response requires hierarchical activation of non-catalytic clip domain serine protease homologs. *PLoS Pathog*. 15(11):e1008194.
- Estévez-Lao TY, Hillyer JF. 2014. Involvement of the *Anopheles gambiae* Nimrod gene family in mosquito immune responses. *Insect Biochem Mol Biol*. 44:12–22.
- Evans JD, Aronstein K, Chen YP, Hetru C, Imler J-L, Jiang H, Kanost M, Thompson GJ, Zou Z, Hultmark D. 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. *Insect Mol Biol*. 15(5):645–656.
- Fraiture M, Baxter RHG, Steinert S, Chelliah Y, Frolet C, Quispe-Tintaya W, Hoffmann JA, Blandin SA, Levashina EA. 2009. Two mosquito LRR proteins function as complement control factors in the TEP1-mediated killing of *Plasmodium*. *Cell Host Microbe*. 5(3):273–284.
- Galili T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31(22):3718–3720.
- Gansner ER, Koren Y, North S. 2005. Graph drawing by stress majorization. In: Pach J, editor. Lecture notes in Computer Science. Vol. 3383. Berlin, Germany: Springer. p. 239–250.
- Gansner ER, North SC. 2000. An open graph visualization system and its applications to software engineering. *Softw Pract Exp*. 30(11):1203–1233.
- Gerardo NM, Altincicek B, Anselme C, Atamian H, Barribeau SM, de Vos M, Duncan EJ, Evans JD, Gabaldón T, Ghanim M, et al. 2010. Immunity and other defenses in pea aphids, *Acyrtosiphon pisum*. *Genome Biol*. 11(2):R21.
- Girardo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S, VectorBase Consortium, Madey G, et al. 2015. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res*. 43(Database issue):D707–D713.
- Glickman MH, Ciechanover A. 2002. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol Rev*. 82(2):373–428.
- Gulley MM, Zhang X, Michel K. 2013. The roles of serpins in mosquito immunology and physiology. *J Insect Physiol*. 59(2):138–147.
- Hammond AM, Kyrou K, Bruttini M, North A, Galizi R, Karlsson X, Kranjc N, Carpi FM, D'Aurizio R, Crisanti A, et al. 2017. The creation and selection of mutations resistant to a gene drive over multiple generations in the malaria mosquito. *PLoS Genet*. 13(10):e1007039.
- Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol*. 30(8):1987–1997.
- Hillyer JF. 2016. Insect immunology and hematopoiesis. *Dev Comp Immunol*. 58:102–118.
- Hoffmann AA. 2013. III.8. Evolutionary limits and constraints. In: Losos JB, Baum DA, Futuyma DJ, Hoekstra HE, Lenski RE, Moore AJ, Peichel CL, Schluter D, Whitlock MC, editors. The Princeton guide to evolution. Princeton (NJ): Princeton University Press. p. 247–252.
- Holm I, Lavazec C, Garnier T, Mitri C, Riehle MM, Bischoff E, Brito-Fravallo E, Takashima E, Thiery I, Zettor A, et al. 2012. Diverged alleles of the *Anopheles gambiae* leucine-rich repeat gene APL1A display distinct protective profiles against *Plasmodium falciparum*. *PLoS One*. 7(12):e52684.
- Holt RA, Subramanian GM, Halpern A, Sutton GC, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro José MC, Wides R, et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298(5591):129–149.
- Imler J-L. 2014. Overview of *Drosophila* immunity: a historical perspective. *Dev Comp Immunol*. 42(1):3–15.
- Jiang X, Peery A, Hall AB, Sharma A, Chen X-G, Waterhouse RM, Komissarov A, Riehle MM, Shouche Y, Sharakhova MV, et al. 2014. Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi*. *Genome Biol*. 15(9):459.
- Kassambara A. 2017. Practical guide to cluster analysis in R: unsupervised machine learning. 1st ed. France: STHDA.
- King JG. 2020. Developmental and comparative perspectives on mosquito immunity. *Dev Comp Immunol*. 103:103458.
- Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nat Rev Genet*. 11(7):487–498.
- Kounatidis I, Ligoxygakis P. 2012. *Drosophila* as a model system to unravel the layers of innate immunity to infection. *Open Biol*. 2(5):120075.
- Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM. 2015. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res*. 43(Database issue):D250–D256.
- Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res*. 13(10):2229–2235.
- Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559.
- Lawniczak MKN, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, Redmond S, Fulton L, Appelbaum E, Godfrey J, et al. 2010. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science* 330(6003):512–514.
- Lazzaro BP, Zasloff M, Rolff J. 2020. Antimicrobial peptides: application informed by evolution. *Science* 368(6490):eaau5480.
- Lemaitre B, Hoffmann J. 2007. The host defense of *Drosophila melanogaster*. *Annu Rev Immunol*. 25:697–743.
- Levashina EA, Baxter RHG. 2018. Complement-like system in the mosquito responses against malaria parasites. In: Stoute JA, editor. Complement activation in malaria immunity and pathogenesis. Cham, Switzerland: Springer International Publishing. p. 139–146.
- Lewis SH, Salmela H, Obbard DJ. 2016. Duplication and diversification of Dipteran Argonaute Genes, and the evolutionary divergence of Piwi and Aubergine. *Genome Biol Evol*. 8(3):507–518.
- Li H, Hambrook JR, Pila EA, Gharamah AA, Fang J, Wu X, Hanington P. 2020. Coordination of humoral immune factors dictates compatibility between *Schistosoma mansoni* and *Biomphalaria glabrata*. *elife* 9:e51708.

- Ligoxygakis P. 2017. Insect immunity. In: Ligoxygakis P, editor. *Advances in insect physiology*. Vol. 52. Cambridge (MA): Academic Press. p. 1–248.
- MacCallum RM, Redmond SN, Christophides GK. 2011. An expression map for *Anopheles gambiae*. *BMC Genomics* 12:620.
- Marinotti O, Calvo E, Nguyen QK, Dissanayake S, Ribeiro JMC, James AA. 2006. Genome-wide analysis of gene expression in adult *Anopheles gambiae*. *Insect Mol Biol*. 15(1):1–12.
- Marinotti O, Cerqueira GC, de Almeida LGP, Ferro MIT, Loreto EL da S, Zaha A, Teixeira SMR, Wespiser AR, Almeida e Silva A, Schlindwein AD, et al. 2013. The genome of *Anopheles darlingi*, the main neotropical malaria vector. *Nucleic Acids Res*. 41(15):7387–7400.
- Markianos K, Bischoff E, Mitri C, Guelbeogo WM, Gneme A, Eiglmeier K, Holm I, Sagnon N, Vernick KD, Riehle MM. 2016. Genetic structure of a local population of the *Anopheles gambiae* complex in Burkina Faso. *PLoS One*. 11(1):e0145308.
- Matthews BJ, Dudchenko O, Kingan SB, Koren S, Antoshechkin I, Crawford JE, Glassford WJ, Herre M, Redmond SN, Rose NH, et al. 2018. Improved reference genome of *Aedes aegypti* informs arbovirus vector control. *Nature* 563(7732):501–507.
- Mauro AA, Ghalambor CK. 2020. Trade-offs, pleiotropy, and shared molecular pathways: a unified view of constraints on adaptation. *Integr Comp Biol*. 60(2):332–347.
- Meekins DA, Kanost MR, Michel K. 2017. Serpins in arthropod biology. *Semin Cell Dev Biol*. 62:105–119.
- Midéga J, Blight J, Lombardo F, Povelones M, Kafatos F, Christophides GK. 2013. Discovery and characterization of two Nimrod superfamily members in *Anopheles gambiae*. *Pathog Glob Health*. 107(8):463–474.
- Miles A; The Anopheles gambiae 1000 Genomes Consortium, Kwiatkowski D. 2017. Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature* 552:96–100.
- Mitri C, Bischoff E, Eiglmeier K, Holm I, Dieme C, Brito-Fravallo E, Raz A, Zakeri S, Nejad MIK, Djadid ND, et al. 2020. Gene copy number and function of the APL1 immune factor changed during *Anopheles* evolution. *Parasit Vectors*. 13(1):18.
- Mussabekova A, Daeffler L, Imler J-L. 2017. Innate and intrinsic antiviral immunity in *Drosophila*. *Cell Mol Life Sci*. 74(11):2039–2054.
- Myllymäki H, Rämét M. 2014. JAK/STAT pathway in *Drosophila* immunity. *Scand J Immunol*. 79(6):377–385.
- Myllymäki H, Valanne S, Rämét M. 2014. The *Drosophila* Imd signaling pathway. *J Immunol*. 192(8):3455–3462.
- Nakhléh J, Christophides GK, Osta MA. 2017. The serine protease homolog CLIPA14 modulates the intensity of the immune response in the mosquito *Anopheles gambiae*. *J Biol Chem*. 292(44):18217–18226.
- Nakhléh J, El Moussawi L, Osta MA. 2017. The melanization response in insect immunity. In: Ligoxygakis P, editor. *Advances in insect physiology*. Vol. 52. Amsterdam: Elsevier. p. 83–109.
- Neafsey DE, Lawniczak MKN, Park DJ, Redmond SN, Coulibaly MB, Traore SF, Sagnon N, Costantini C, Johnson C, Wiegand RC, et al. 2010. SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes. *Science* 330(6003):514–517.
- Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburg P, Artemov G, et al. 2015. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* 347(6217):1258522.
- Neira Oviedo M, VanEkeris L, Corena-Mcleod MDP, Linser PJ. 2008. A microarray-based analysis of transcriptional compartmentalization in the alimentary canal of *Anopheles gambiae* (Diptera: Culicidae) larvae: gut transcriptome of larval *An. gambiae*. *Insect Mol Biol*. 17(1):61–72.
- Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu Z, Loftus B, Xi Z, Megy K, Grabherr M, et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316(5832):1718–1723.
- Povelones M, Bhagavatula L, Yassine H, Tan LA, Upton LM, Osta MA, Christophides GK. 2013. The CLIP-domain serine protease homolog SPCLIP1 regulates complement recruitment to microbial surfaces in the malaria mosquito *Anopheles gambiae*. *PLoS Pathog*. 9(9):e1003623.
- Povelones M, Osta MA, Christophides GK. 2016. The complement system of malaria vector mosquitoes. In: Raikhel AS, editor. *Advances in insect physiology*. Vol. 51. Amsterdam: Elsevier. p. 223–242.
- Povelones M, Waterhouse RM, Kafatos FC, Christophides GK. 2009. Leucine-rich repeat protein complex activates mosquito complement resistance to defense against *Plasmodium* parasites. *Science* 324(5924):258–261.
- R Core Team. 2021. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: <https://www.R-project.org/>. Accessed June 7, 2021.
- Ramirez JL, Muturi EJ, Flor-Weiler LB, Vermillion K, Rooney AP. 2020. Peptidoglycan recognition proteins (PGRPs) modulates mosquito resistance to fungal entomopathogens in a fungal-strain specific manner. *Front Cell Infect Microbiol*. 9:465.
- Rao X-J, Zhan M-Y, Pan Y-M, Liu S, Yang P-J, Yang L-L, Yu X-Q. 2018. Immune functions of insect β GRPs and their potential application. *Dev Comp Immunol*. 83:80–88.
- Reyes Ruiz VM, Sousa GL, Sneed SD, Farrant KV, Christophides GK, Povelones M. 2019. Stimulation of a protease targeting the LRIM1/APL1C complex reveals specificity in complement-like pathway activation in *Anopheles gambiae*. *PLOS One*. 14(4):e0214753.
- Richardson MK, Chipman AD. 2003. Developmental constraints in a comparative framework: a test case using variations in phalanx number during amniote evolution. *J Exp Zool B Mol Dev Evol*. 296(1):8–22.
- Rolf J, Reynolds S. 2009. *Insect infection and immunity*. Oxford (United Kingdom): Oxford University Press.
- Rottschaefer SM, Riehle MM, Coulibaly B, Sacko M, Niarié O, Morlais J, Traoré SF, Vernick KD, Lazzaro BP. 2011. Exceptional diversity, maintenance of polymorphism, and recent directional selection on the APL1 malaria resistance genes of *Anopheles gambiae*. *PLoS Biol*. 9(3):e1000600.
- Ruzzante L, Reijnders MJMF, Waterhouse RM. 2019. Of genes and genomes: mosquito evolution and diversity. *Trends Parasitol*. 35(1):32–51.
- Sackton TB. 2019. Comparative genomics and transcriptomics of host-pathogen interactions in insects: evolutionary insights and future directions. *Curr Opin Insect Sci*. 31:106–113.
- Sackton TB, Lazzaro BP, Clark AG. 2017. Rapid expansion of immune-related gene families in the house fly, *Musca domestica*. *Mol Biol Evol*. 34(4):857–872.
- Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in *Drosophila*. *Nat Genet*. 39(12):1461–1468.
- Schnitger AKD, Kafatos FC, Osta MA. 2007. The melanization reaction is not required for survival of *Anopheles gambiae* mosquitoes after bacterial infections. *J Biol Chem*. 282(30):21884–21888.
- Schnitger AKD, Yassine H, Kafatos FC, Osta MA. 2009. Two C-type lectins cooperate to defend *Anopheles gambiae* against gram-negative bacteria. *J Biol Chem*. 284(26):17616–17624.
- Shakhnovich BE, Koonin EV. 2006. Origins and impact of constraints in evolution of gene families. *Genome Res*. 16(12):1529–1536.
- Shi X-Z, Zhong X, Yu X-Q. 2012. *Drosophila melanogaster* NPC2 proteins bind bacterial cell wall components and may function in immune signal pathways. *Insect Biochem Mol Biol*. 42(8):545–556.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15(8):1034–1050.
- Simões ML, Dong Y, Hammond A, Hall A, Crisanti A, Nolan T, Dimopoulos G. 2017. The Anopheles FBN9 immune factor mediates *Plasmodium* species-specific defense through transgenic fat body expression. *Dev Comp Immunol*. 67:257–265.
- Sousa GL, Bishnoi R, Baxter RHG, Povelones M. 2020. The CLIP-domain serine protease CLIPC9 regulates melanization downstream of SPCLIP1, CLIPA8, and CLIPA28 in the malaria vector *Anopheles gambiae*. *PLoS Pathog*. 16(10):e1008985.

- Suzuki R, Shimodaira H. 2006. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22(12):1540–1542.
- Tanaka H, Ishibashi J, Fujita K, Nakajima Y, Sagisaka A, Tomimoto K, Suzuki N, Yoshiyama M, Kaneko Y, Iwasaki T, et al. 2008. A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori*. *Insect Biochem Mol Biol*. 38(12):1087–1110.
- Valanne S, Wang J-H, Rämet M. 2011. The *Drosophila* Toll signaling pathway. *J Immunol*. 186(2):649–656.
- Vasta GR. 2020. Galectins in host–pathogen interactions: structural, functional and evolutionary aspects. In: Hsieh S-L, editor. *Advances in experimental medicine and biology. Lectin in host defense against microbial infections*. Vol. 1204. Singapore: Springer. p. 169–196.
- Volz J, Muller H-M, Zdanowicz A, Kafatos FC, Osta MA. 2006. A genetic module regulates the melanization response of *Anopheles* to *Plasmodium*. *Cell Microbiol*. 8(9):1392–1405.
- Volz J, Osta MA, Kafatos FC, Müller H-M. 2005. The roles of two clip domain serine proteases in innate immune responses of the malaria vector *Anopheles gambiae*. *J Biol Chem*. 280(48):40161–40168.
- Wang Q, Ren M, Liu X, Xia H, Chen K. 2019. Peptidoglycan recognition proteins in insect immunity. *Mol Immunol*. 106:69–76.
- Waterhouse RM. 2015. A maturing understanding of the composition of the insect gene repertoire. *Curr Opin Insect Sci*. 7:15–23.
- Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, et al. 2007. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* 316(5832):1738–1743.
- Waterhouse RM, Lazzaro BP, Sackton TB. 2020. Characterization of insect immune systems from genomic data. In: Sandrelli F, Tettamanti G, editors. *Immunity in Insects*. Springer Protocols Handbooks. New York: Springer. p. 3–34.
- Waterhouse RM, Povelones M, Christophides GK. 2010. Sequence-structure-function relations of the mosquito leucine-rich repeat immune proteins. *BMC Genomics*. 11:531.
- Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res*. 41(Database issue):D358–D365.
- Waterhouse RM, Zdobnov EM, Kriventseva EV. 2011. Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome Biol Evol*. 3:75–86.
- Weetman D, Wilding CS, Steen K, Pinto J, Donnelly MJ. 2012. Gene flow-dependent genomic divergence between *Anopheles gambiae* M and S forms. *Mol Biol Evol*. 29(1):279–291.
- White BJ, Lawniczak MKN, Cheng C, Coulibaly MB, Wilson MD, Sagnon N, Costantini C, Simard F, Christophides GK, Besansky NJ. 2011. Adaptive divergence between incipient species of *Anopheles gambiae* increases resistance to *Plasmodium*. *Proc Natl Acad Sci U S A*. 108(1):244–249.
- Williams M, Summers BJ, Baxter RHG. 2015. Biophysical analysis of *Anopheles gambiae* leucine-rich repeat proteins APL1A1, APL1B and APL1C and their interaction with LRIM1. *PLoS One*. 10(3):e0118911.
- Wiltshire RM, Bergey CM, Kayondo JK, Birungi J, Mukwaya LG, Emrich SJ, Besansky NJ, Collins FH. 2018. Reduced-representation sequencing identifies small effective population sizes of *Anopheles gambiae* in the north-western Lake Victoria basin, Uganda. *Malar J*. 17:285.
- Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. *Proc Biol Sci*. 273(1593):1507–1515.
- Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A*. 106(18):7273–7280.
- Worth CL, Gong S, Blundell TL. 2009. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol*. 10(10):709–720.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.
- Yassine H, Kamareddine L, Chamat S, Christophides GK, Osta MA. 2014. A serine protease homolog negatively regulates TEP1 consumption in systemic infections of the malaria vector *Anopheles gambiae*. *J Innate Immun*. 6(6):806–818.
- Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppely M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res*. 45(D1):D744–D749.
- Zhang X, An C, Sprigg K, Michel K. 2016. CLIPB8 is part of the phenoloxidase activation system in *Anopheles gambiae* mosquitoes. *Insect Biochem Mol Biol*. 71:106–115.
- Zhang X, Li M, El Moussawi L, Saab S, Zhang S, Osta MA, Michel K. 2021. CLIPB10 is a terminal protease in the regulatory network that controls melanization in the African malaria mosquito *Anopheles gambiae*. *Front Cell Infect Microbiol*. 10:585986.
- Zou Z, Evans JD, Lu Z, Zhao P, Williams M, Sumathipala N, Hetru C, Hultmark D, Jiang H. 2007. Comparative genomic analysis of the *Tribolium* immune system. *Genome Biol*. 8(8):R177.

Appendix 2: Species Tree

