Functional constraints on insect immune system components govern their evolutionary trajectories

Livio Ruzzante¹, Romain Feron¹, Maarten J.M.F. Reijnders¹, Antonin Thiébaut¹, Robert M. Waterhouse¹,∗

¹ Department of Ecology and Evolution, University of Lausanne, and Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
∗ Corresponding author: robert.waterhouse@unil.ch (RMW)

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 recognises 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. Characterising 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 hypothesise that characterising gene evolutionary histories can define distinct dynamics associated with different functional roles. We develop metrics that quantify gene evolutionary histories, employ these to characterise 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 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. 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 characterise how functional constraints on different components of biological systems govern their evolutionary trajectories.

Keywords

Anopheles mosquito, evolutionary profiling, gene expression, gene families, innate immunity

© 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 (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
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 recognises 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, e.g. 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 licence 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” characterised 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 non-essential; 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. Characterising these patterns in the context of a relatively 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-characterised 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 neutralise and clear infections (Rolff and Reynolds 2009; Ligoxygakis 2017). The immune system comprises both humoral and cellular responses with components dedicated to recognising signs of infection, signalling cascades to activate primary defences 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 recognise 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 signalling through the Toll (Valanne et al. 2011), Imd (Myllymäki et al. 2014), or the JAK/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. Defence responses are mediated by various cells and tissues including haemocytes, the fat body, and the midgut, and pathogen killing can occur via processes such as melanisation, phagocytosis, lysis, autophagy, and apoptosis (Hillyer 2016; King 2020), with RNA interference (RNAi) facilitating major antiviral defences (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 characterise different immune phases: (i) gene and domain gains or losses (turnover) can create diversity in recognition modules; (ii) core signalling pathway members are almost always maintained as single-copy orthologues often with elevated levels of sequence divergence; (iii) modulators appear to form lineage-restricted units with members picked from large families often with high gene turnover rates; and (iv) effectors like AMPs show dynamic gains and losses or are lineage-restricted while oxidative defence 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 hypothesise that comprehensive quantitative multi-species and multi-feature characterisation 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; Arensburger 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 (i) develop a suite of metrics that quantify gene and gene family evolutionary histories, (ii) employ these metrics to characterise the evolutionary features of mosquito and fly immune gene repertoires, and (iii) 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 multi-species 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 categorisations with gene co-expression 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 including taxonomic spread, copy-number changes, protein- and DNA-level sequence divergence, conservation, and constraint, as well as genomic organisation, 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 two 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 Methods). Orthologous groups (OGs) are comprised of 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 are computed. OG compositions are used directly to quantify features such as universality (the proportion of species present in an OG) or duplicability (the proportion of species that have multi-copy orthologues). 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 multi-copy orthologues 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 catalogue of 36 canonical mosquito immune-related gene families and subfamilies (Figure 1). The evolutionary feature profiles for all families are visualised 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 (Figure 1, Supplementary Figure S1). 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 universality (UNI) are significantly lower, consistent with there being no detectable LRIM orthologues
beyond mosquitoes (Waterhouse et al. 2010). They also exhibit fairly typical low duplicability (DUP), average copy-number (ACN), and copy-number variation (CNV), reflecting their mostly single-copy orthologue status across mosquitoes. These metrics describe the family as a whole while allowing for differences amongst members, e.g. the gene duplications that gave rise to three APL1/LRIM2 paralogs in one lineage of African Anopheles (Mitri et al. 2020). Estimates of non-synonymous substitutions per non-synonymous 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). Which together produce 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 neighbourhood, or synteny (SYN) is slightly lower than average for LRIMs, while they notably show the most extreme significantly elevated protein sequence evolutionary divergence (EVR). Single nucleotide polymorphism data also show a significantly elevated proportion of non-synonymous SNPs (NSP) and significantly above average non-synonymous 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 (WGA) and significantly less constrained (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. Gram-negative binding proteins (GNBPs) are characterised 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 GNPB subfamily in the mosquito ancestor (Bartholomay et al. 2010). In contrast, Imd pathway signalling genes (IMDSIGs) are characterised 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 signalling 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 non-catalytic members in modulatory cascades and their hierarchies (El Moussawi et al. 2019).

The autophagy (APHAG) and small regulatory RNA (SRRP) pathway members 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, while 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 orthologues of 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 dataset 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 Additional File 1 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 (Figure 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 Methods). Employing family median (Figure 2) and mean (Supplementary Figure S2) 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 Figures S3-6, Additional File 2). Furthermore, clustering using principal components instead of the metric values themselves also identifies several of the observed family groupings (Supplementary Figure S7). 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 non-synonymous SNPs, contractions (gene losses), and evolutionary rate (protein sequence divergence); (ii) densities of synonymous and non-synonymous SNPs; (iii) average copy-number, expansions, duplications, and copy-number variation, often also including synteny as in Figure 2; and (iv) age, universality, constraint, and alignability, often also 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 Figure S2). 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 (Figure 2, Supplementary Figures S2, S8).

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 non-synonymous 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 orthologues 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 non-synonymous SNPs.

Set (iii) captures all the copy-number features related to gene birth, linked to local genomic organisation (synteny). Gene duplications lead to higher and often more variable copy-numbers that are identified by CAFE as expansions, 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 synteny suggests that duplicated genes often also maintain their local genomic neighbourhoods. However, this phenomenon is driven by only a small subset of families with both elevated duplication and synteny metrics: GNBPs, MD-2-like proteins (MLs), and particularly LYSs (Figure 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 characterised 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 Figures S9, S10). Thus the associations between gene loss and protein sequence divergence, between duplication and synteny, 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 catalogue also recovers the links between gene loss and protein sequence divergence, and between taxonomic spread and DNA-level sequence conservation (Supplementary Figures S11, S12). However, despite MLs and LYSs showing the same trend as for An. gambiae, synteny is no longer associated with copy-number features related to gene birth, indicating that maintaining genomic neighbourhoods 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 neighbourhoods.

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 summarised by three main axes delineated by the major PCA groupings: axis 1, duplication and synteny; 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) antimicrobial peptides and glutathione peroxidases; (b) cysteine-aspartic and CLIPC proteases with serine protease inhibitors; (c) LRIMs, thioester-containing proteins, CLIPA protease homologues, CLIPB&E proteases, C-type lectins, and fibrinogen-related and Nimrod proteins; (d) GNBPs, MD-2-like proteins, lysozymes, and prophenoloxidases; (e) PGRPs and galectins; (f) Toll, Imd, and JAK/STAT signalling proteins; and (g) a large set comprising autophagy and RNA interference (RNAi) related proteins, Toll, Imd, and JAK/STAT pathway modulators, toll receptors, scavenger receptors A and B, CLIPD proteases, superoxide dismutases, as well as haem and thioredoxin peroxidases. Clustering with metric means rather than medians results in different hierarchies but with several broadly similar groupings including: LRIMs, CLIPA protease homologues, CLIPB&E proteases, and fibrinogen-related proteins; cysteine-aspartic and CLIPC proteases; GNBPs, MD-2-like proteins, lysozymes, and prophenoloxidases; 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 haem and thioredoxin peroxidases (Supplementary Figure S2). Similar variations of these groupings are obtained when clustering means or medians using
alternative distance-clustering method combinations (Additional File 2). Combining this variation with results from bootstrapping provides a measure of evolutionary profile similarity between all pairs of families (see Methods). The families that most frequently cluster together using metric means (Supplementary Figure S5) or medians (Supplementary Figure S6) include: PGRPs, galectins, GNBPs, MD-2-like proteins, lysozymes, and phenoxoloxidases; cysteine-aspartic and CLIPC proteases; LRIMs, thioester-containing proteins, CLIPA protease homologues, 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 haem and thioredoxin peroxidases. Thus, while 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 GNBPs and MD-2-like proteins, also indicating that MD-2-like proteins more closely resemble classical than other recognition families, thereby warranting their reclassification (Figure 2). PGRPs can bind bacterial cell wall Dap- or Lys-type peptidoglycans (Wang et al. 2019), while galectins can bind surface β-galactosides (Vasta 2020). Similarly, GNBPs can recognise β-1,3-glucans that make up structural polysaccharides of yeast cell walls (Rao et al. 2018), while MD-2-like proteins 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 Khouy 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 (Figure 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.

Signalling 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, Tab2-Tak1, and IkB 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 signalling 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 signalling (Glickman and Ciechanover 2002). Other enzymes including the superoxide dismutases as well as the haem and thioredoxin peroxidases involved in reactive oxygen species metabolism (Hillyer 2016), show similarly conservative evolutionary profiles (Figure 2). Proteolytically activated PPOs oxidise 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 while 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 characterised by low gene turnover and low sequence evolutionary rates (Figure 2). In contrast, much more dynamic evolutionary profiles characterise the grouping of families of immune cascade modulators like C-type lectins, CLIPA protease homologues and CLIPB&E proteases, regulators of melanisation responses like serine protease inhibitors, and key players in mosquito complement-like responses, like thioester-containing proteins and LRIMs. While melanisation 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 centred 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; Povelones et al. 2013; Povelones et al. 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 characterised with evolutionary profiles showing a pattern of younger and less widespread orthologues, 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.

**Co-expression analyses identify immune families that function in concert**

Analysis of multi-sample 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* (Figure 3, Supplementary Figure S13). Thus, without presupposing any functional categorisations, 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 categorisations 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 normalised 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 Methods). Visualising pairwise family similarities as a spring model layout network optimised 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 (Figure 3). These prominently include a quintet of families with highly and significantly overlapping expression patterns: LRIMs, TEPs, FREPs, CLIPAs, and CLIPBs (Figure 3), with members implicated in coordinating and executing mosquito complement system responses (Povelones et al. 2016; Reyes Ruiz et al. 2019).

An. gambiae TEP1 forms a stable protein complex with a heterodimer of LRIM1 and APL1A/B/C (LRIM2 paralogues) in the haemolymph until the complement response is activated (Fraiture et al. 2009; Povelones et al. 2009; Williams et al. 2015), so coordinated co-expression of these genes is important for their functions. Like the LRIMs and TEPs, the FREPs are also found in the haemolymph, and several members are infection-responsive and important for defence, e.g. FREP57/FBN8, FREP13/FBN9, and FREP40/FBN39 (Dong et al. 2006; Dong and Dimopoulos 2009; Simões et al. 2017). FREPs themselves might dimerise or oligomerise, 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 homologues are positive and negative regulators of immune responses mediated by TEP1, e.g. 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 e.g. 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 melanisation responses (Volz et al. 2006). The available evidence therefore supports the family-level expression analyses that demonstrate highly and significantly overlapping expression patterns (Figure 3) of members of this quintet of families that function in concert.

Of this quintet, expression of CLIPA protease homologues 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 melanisation 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 (Figure 2), remain largely unknown. Serpins themselves are most similar in expression to prophenoloxidases, both of which would need to be replenished after being depleted during melanisation responses (Gulley et al. 2013; Nakhleh,
The PPOs in turn appear significantly similar to the C-type lectins, which are generally considered glycan-binding recognition proteins, but at least two members - CTL4 and CTLMA2- are key regulators of melanisation downstream of immune recognition (Schnitger et al. 2009; Bishnoi et al. 2019). The family-level expression similarities (Figure 3) therefore derive from the functional links amongst the CLIP, CTL, and SRPN family members that modulate the activation of melanisation, and the PPO enzyme effectors of melanisation activity.

Amongst classical recognition proteins, PGRPs and GNBPs are most similar, and their expression patterns both closely match those of antimicrobial peptides 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, e.g. 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 melanisation 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 signalling pathways means that these families may be thought of as performing analogous roles rather than functioning in concert per se. However, learning more about signalling 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 signalling and modulation components remain distinct from the recognition and response families: Imd and JAK/STAT pathway modulators are significantly similar, while Toll pathway modulators group together with Toll and JAK/STAT pathway signalling 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 Figure S14). At broad-scale resolution the distinction between pathway signalling/modulation and recognition/response families is accentuated, while the melanisation and complement responses become more closely interlinked. Many of the most similar families also show substantially overlapping expression patterns when quantifying similarities across co-expression modules built from a subset of immune-related experimental conditions (see Methods, Supplementary Figures S15-S16, Table S2, Additional File 3). For example, families implicated in complement system responses again show similar expression patterns (Supplementary Figure S17), and at a broader-scale resolution become more closely associated with melanisation responses (Supplementary Figure S18). At broad-scale resolution pairs of similar recognition families include GNBPs and PGRPs, GNBPs and MD-2-like proteins, as well as galectins and B-type scavenger receptors, while at both resolutions Imd and JAK/STAT pathway signalling members are highly and significantly similar. Multi-condition co-expression 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 (Figure 4, Additional File 4). 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 CLIPEs, for which specific roles in complement responses remain largely unknown, but with which their expression similarity increases at broad-scale resolution, albeit remaining non-significant. The CLIPA protease homologues 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 CLIPE 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 defence (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 (Povedones 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 (CLIPE-FREP). While CLIPAs and SRPNs are known to function together in cascades regulating melanisation (El Moussawi et al. 2019), potential functional interactions between CLIPEs and FREPs remain to be explored. The melanisation 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, MD-2-like 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 to 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 signalling 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 signalling (JASTSIG-TOLLSIG), Imd and JAK/STAT modulation (IMDMOD-JASTMOD), and JAK/STAT signalling 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 (Figure 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 categorisations, our multi-species 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-characterised. 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 within an established framework. Feature analysis within the limits of our study system identified three main axes of evolutionary trajectories characterised by gene duplication and synteny, 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 MD-2-like proteins, showed patterns that can be explained by the limited structural diversity of the principal microbial ligands with which they interact. Pathway signalling 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 co-expression 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 melanisation responses, both of which occur mainly in the haemolymph. 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 defence 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
Orthologous Groups (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). 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. Single nucleotide polymorphisms (SNPs) for \textit{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 datasets 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). Multi-species whole genome alignments were generated from the assemblies of 22 mosquitoes available from VectorBase and 36 \textit{Drosophila} available from The National Center for Biotechnology Information (Supplementary Table S1). 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 \textit{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 co-expression analysis employed these expression statistics using a subset of the conditions to build co-expression modules. Co-expression 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).

\textit{Anopheles gambiae} and \textit{Drosophila melanogaster} immunity gene catalogues
The catalogues of \textit{An. gambiae} and \textit{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). \textit{An. gambiae} and \textit{D. melanogaster} gene and orthologous group membership for 36 immune-related gene families and subfamilies are summarised 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 multi-copy orthologues; the average orthologue copy number (ACN); the copy number variation (CNV) computed as the standard deviation of orthologue counts per species present divided by the ACN. Phylogenetic analysis by maximum likelihood (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 non-synonymous substitutions per non-synonymous site (PDN); and the non-synonymous to synonymous ratio (SEL). Gene turnover was estimated using the computational analysis of gene family evolution (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 synteny conservation (SYN) as the proportion of orthologues that maintain their orthologous neighbours in the genomes of the other species. Finally, the evolutionary rate (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 non-synonymous (NSP), as well as the non-synonymous (NSD) and synonymous (SSD) SNP densities as the number of SNPs divided by the total coding-sequence length. Multi-species whole genome alignments were used to compute per-nucleotide metrics of conservation and constraint. Whole genome alignability (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 per-nucleotide levels of constraint (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 catalogues define immunity gene membership of subfamilies (e.g. cecropins, defensins, attacins), families (e.g. antimicrobial peptides), and broader categories (e.g. antimicrobial effectors), and the orthology datasets define gene membership of OGs. Thus, the gene family evolutionary metrics were computed by averaging values over all OGs containing genes belonging to each catalogued immune gene family. These family-level means for each metric were compared to 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, i.e. delta-mean ($\Delta \bar{x}$). For graphical visualisation, $\Delta \bar{x}$ values were scaled by dividing by the absolute maximum $\Delta \bar{x}$ per evolutionary feature and plotted with the colour-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 to 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 to $\Delta \bar{x}$ 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 Δx̄ greater than the observed Δx̄ 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 normalise all metric values by subtracting the means and then dividing the (centred) values by their standard deviations. Dissimilarity matrices were computed with the normalised 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 normalised as follows: 

\[
\frac{2 \times \text{co-occurrence of Family 1 & Family 2}}{\text{co-occurrence of Family 1 with any Family} + \text{co-occurrence of Family 2 with any Family}}
\]

Normalised 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 and Supplementary Figures S3, S4). The hierarchical clustering results were visualised as heatmaps with corresponding family and evolutionary feature dendrograms showing AU support, plotted with the `gplots` and `dendextend` (Galili 2015) R packages. Principal Component Analysis (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 Figure S7), i.e. when applying standard clustering techniques after transforming the correlated metrics into principal components.
Gene and family co-expression 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 analysed to quantify co-occurrences of gene family members in the same cell on the map (fine-scale resolution of gene co-expression), and in the same supercell, the cell and its immediate eight neighbouring cells on the map including toroidal neighbours (broader-scale resolution of gene co-expression). 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 (WGCNA) approach (Langfelder and Horvath 2008) on a subset of 24 conditions selected from the VectorBase gene expression dataset including blood feeding experiments and tissues from (Marinotti et al. 2006; Neira Oviedo et al. 2008; 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 Materials and Methods**
Additional details on the analysed species, genome assemblies, annotated gene sets, as well as software versions and analysis parameters are presented in the *Supplementary Materials*. 
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
RMW conceived the study. LR, RF, MJMFR, AT, and RMW designed the analyses. LR analysed orthology data, quantified evolutionary features, and performed clustering analyses and statistical testing. RF built whole genome alignments, and analysed alignment and variation data. MR analysed variation data. AT curated data and assisted with gene expression data analysis. LR and RMW wrote the manuscript with input from all authors. All authors read and approved the manuscript.

Conflict of Interest Statement
All authors declare no competing interests.

Data Availability
No new genomic data were produced as part of this study and all data analysed 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: Visualisation 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.* 35:709–719.

Albà MM, Castresana J. 2005. Inverse Relationship Between Evolutionary Rate and Age of Mammalian Genes. *Mol. Biol. Evol.* 22:598–606.

Alquraini A, El Khoury J. 2020. Scavenger receptors. *Curr. Biol.* 30:R790–R795.

Antonovics J, van Tienderen PH. 1991. Ontoecogenophyloconstraints? The chaos of constraint terminology. *Trends Ecol. Evol.* 6:166–168.

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* 330:86–90.

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.

Barrieau SM, Sadd BM, du Plessis L, Brown MJ, Buechel SD, Cappelle K, Carolan JC, Christiaens O, Colgan TJ, Erler S, et al. 2015. A deapurate 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* 330:88–90.

Bishnoi R, Sousa GL, Contet A, Day CJ, Hou C-FD, Profitt LA, Singla D, Jennings MP, Valentine AM, Povolones M, et al. 2019. Solution structure, glycan specificity and of phenol oxidase inhibitory activity of Anopheles C-type lectins CTL4 and CTLMA2. *Sci. Rep.* 9:15191.

Blanchette M, Kent WJ, Riemer C, Elintski 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.* 14:708–715.

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* 3:364–374.

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:e72130.

Brucker RM, Funkhouser LJ, Setia S, Paul C, Hagen NC. 2012. Insect innate immunity database (IIID): an annotation tool for identifying immune genes in insect genomes. *PLoS ONE* 7: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 Xuanting, Gu J, 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. Natl. Acad. Sci.* 112:E5907–E5915.

Christophides GK, Zdobnov E, Barillas-Mury C, Binney 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* 298:159–165.

Dana AN, Alquraini A, El Khoury J. 2020. Scavenger receptors. *Curr. Biol.* 30:R790–R795.

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:e52.

Dong Y, Dimopoulos G. 2009. Anopheles Fibrinogen-related Proteins Provide Expanded Pattern Recognition Capacity against Bacteria and Malaria Parasites. *J. Biol. Chem.* 284:9835–9844.

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: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.* 15:645–656.

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* 5:273–284.

Gallii T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31:3718–3720.

Gansner ER, Koren Y, North S. 2005. Graph Drawing by Stress Majorization. In: Pach J, editor. Graph Drawing. Vol. 3383. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 239–250.
Wolf YI, Carmel L, Koonin EV. 2006. Unifying measures of gene function and evolution. *Proc. R. Soc. B Biol. Sci.* 273:1507–1515.

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.* 106:7273–7280.

Worth CL, Gong S, Blundell TL. 2009. Structural and functional constraints in the evolution of protein families. *Nat. Rev. Mol. Cell Biol.* 10:709–720.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591.

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.* 6:806–818.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey 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:D744–D749.

Zhang X, An C, Sprigg K, Michel K. 2016. CLIPB8 is part of the prophenoloxidase 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:R177.
Figure Legends and Tables

Figure 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 orthologous groups from other immune-related gene families (Δx̄). For visualisation, values of Δx̄ are scaled by the absolute maximum Δx̄ per metric, i.e. for each metric the distribution is transformed by dividing all values by the absolute maximum Δx̄. Values therefore range from a minimum of -1 for metrics where the largest deviation is below the mean, i.e. lower than other families, and the maximum of 1 for metrics where the largest deviation is above the mean, i.e. higher than other families. The significance of the difference of the distribution of metric values (no scaling) for each family compared to 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≤0.01, ** p≤0.05, * p≤0.1). Feature acronyms are defined in Table 1. Family acronyms are defined in Table 2 and are coloured according to categories defined based on their putative roles in the principal immune phases: classical recognition (red), other recognition (blue), pathway signalling (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 (Δx̃) are presented in Supplementary Figure S1.

Figure 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 visualised 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 approximately unbiased (AU) support values. Principal Component Analysis (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 coloured according to categories defined based on their putative roles in the principal immune phases: classical recognition (red), other recognition (blue), pathway signalling (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, coloured 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.
Figure 3. Network of immune family expression similarities based on the VectorBase Expression Map.

The network layout optimised with a spring model provides a two-dimensional visualisation 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 coloured according to categories defined based on their putative roles in the principal immune phases: classical recognition (red), other recognition (blue), pathway signalling (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).

Figure 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 non-significant pairs shown in grey, and with circle sizes scaled by the p-value. Family acronyms are defined in Table 2 and are coloured according to categories defined based on their putative roles in the principal immune phases: classical recognition (red), other recognition (blue), pathway signalling (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).
Table 1: Evolutionary feature metric descriptions.
For each evolutionary feature, the metric name, acronym, description, and source data are presented (see methods for details). CAFE, computational analysis of gene family evolution; OG, orthologous group; PAML, Phylogenetic analysis by maximum likelihood.

| 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 multi-copy orthologues  | 43-insect orthology |
| Average Copy Number       | ACN     | The average (mean) orthologue copy number across all species present in an OG | 43-insect orthology |
| Copy Number Variation     | CNV     | The standard deviation of orthologue 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 |
| Constructions             | 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 orthologues in an OG that maintain their orthologous neighbours 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-Anopheles orthology |
| PAML’s dN                 | PDN     | The number of non-synonymous substitutions per non-synonymous site as computed by PAML | 19-Anopheles orthology |
| PAML’s dN/dS              | SEL     | The non-synonymous to synonymous substitution ratio (dN/dS) as computed by PAML | 19-Anopheles orthology |
| Non-synonymous SNP Proportion | NSP   | The proportion of all coding-sequence SNPs that were non-synonymous (averaged over genes per OG) | An. gambiae variation |
| Non-synonymous SNP Density | NSD    | The density of non-synonymous SNPs over a gene’s coding-sequence length (averaged over genes per OG) | An. gambiae variation |
| Synonymous SNP Density     | SSD     | The density of synonymous SNPs over a gene’s coding-sequence length (averaged over genes per OG) | An. gambiae 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 PG) | 22 mosquitoes 36 Drosophila |
| 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 Drosophila |
Table 2. The *Anopheles gambiae* and *Drosophila melanogaster* immunity gene catalogues.

Brief descriptions of immune gene families or pathway components are presented along with counts of the numbers of genes and orthologous groups (OGs) for the mosquito and fly catalogues.

| Acronym | Summary Description | An. gambiae | D. melanogaster |
|---------|---------------------|-------------|----------------|
|         |                     | 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, BGBP)s are a family of carbohydrate-binding pattern recognition receptors | 7    | 3   | 3     | 3   |
| PGRP    | Peptidoglycan recognition proteins are pattern recognition receptors capable of recognising 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 opsonisation, encapsulation, and melanisation, as well as immune signalling 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 defence 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 recognising lipopolysaccharide | 16   | 7   | 8     | 5   |
| NIMROD  | Nimrods have been shown to bind bacteria leading to their phagocytosis by haemocytes, they contain epidermal growth factor (EGF)-like domains | 3    | 3   | 12    | 8   |
| TEP     | Thioester-containing proteins are related to vertebrate complement factors and α2-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 characterised by peptidoglycan recognition protein receptors, intracellular signal transducers (IMDSIG) and modulators (IMDMOD), and the NF-κB transcription factor Relish | 9    | 9   | 10    | 10  |
| IMDMOD  |                                      | 6    | 6   | 6     | 6   |
| JASTSIG | The JAnus Kinase protein (JAK) and the Signal Transducer and Activator of Transcription (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-κB transcription factors Dorsal | 5    | 5   | 6     | 6   |
| TOLLMOD |                                      | 8    | 8   | 8     | 8   |
| CASP    | Caspases are cysteine-aspartic proteases involved in immune signalling 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 signalling 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 are important in antiviral responses and are involved in regulating immune signalling and suppressing apoptotic cell death | 8    | 5   | 4     | 4   |
| SRPN    | Serine protease inhibitors, or serpins, modulate many signalling 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 hydrolyse peptidoglycans present in the cell walls of many bacteria, causing cell lysis | 7    | 1   | 17    | 3   |
| PPO     | Prophenoloxidases are key enzymes in the melanisation cascade that helps to kill invading pathogens | 9    | 1   | 3     | 1   |
Pathogens and is important for wound healing

| Pathway | Description |
|---------|-------------|
| **GPX** | Glutathione, haem, and thioredoxin peroxidases are enzymes involved in the metabolism of reactive oxygen species (ROS) that are toxic to pathogens |
| **HPX** | Superoxide dismutases are antioxidant enzymes involved in the metabolism of toxic superoxide into oxygen or hydrogen peroxide |
| **TPX** | Autophagy-related genes participate in a form of cell death characterised by the formation of an internal autophagosome where pathogens are degraded |
| **SOD** | Small regulatory RNA pathway members are involved in RNA interference and include argonautes, dicers, pixis, and helicases |
| **SPZ** | Spaeztle-like proteins contain a cysteine knot domain, the cleavage of Spaeztle results in binding of the product to the Toll receptor and subsequent activation of the Toll pathway |
| **TOLL** | Toll receptors connect extracellular pathogen recognition to intracellular Toll pathway signalling and activation of immune defence responses |

| Pathway | Count |
|---------|-------|
| **GPX** | 2 2 2 2 |
| **HPX** | 15 10 10 9 |
| **TPX** | 5 5 6 6 |
| **SOD** | 4 4 4 4 |
| **APHAG** | 19 19 22 22 |
| **SRRP** | 28 23 22 20 |
| **SPZ** | 5 5 6 6 |
| **TOLL** | 12 6 9 6 |

**Totals:** 420 298 354 276
| Family | Significantly Higher | Significantly Lower | Interpretation Summary |
|--------|---------------------|--------------------|------------------------|
| GALE   | -                   | -                  | No extreme features    |
| GNBP   | D, A, C, N, S, W     | -                  | Duplications, maintained neighbourhood, widely alignable |
| PGPR   | E, N, S, S, S        | -                  | Duplications, population variation |
| SCRA   | A, G                 | C                 | Ancient, stable copy-number |
| SCRB   | A, U, W, G, P       | P, S, C, K, N, S  | Ancient, widespread, widely alignable, constrained sequence, constrained substitutions, stable copy-number, population conservation |
| CTL    | S, C, S, N          | U, S, W, G, P     | Relaxed substitutions, losses, maintained neighbourhood, widespread, population conservation, sparsely alignable, relaxed sequence |
| FREP   | A, S, C, N, S, N     | A, U, S, T, S, W, G, P | Duplications, relaxed substitutions, losses, amino acid divergence, young, sparse, unstable copy-number, shuffled neighbourhood, sparsely alignable, relaxed sequence |
| LRIM   | P, D, S, C, E, V, N, S, N | A, U, C, P, W, P | Relaxed substitutions, losses, amino acid divergence, population variation, young, sparse, stable copy-number, sparsely alignable, relaxed sequence |
| ML     | D, A, C, E, S, Y     | P, D, S           | Duplications, maintained neighbourhood, constrained substitutions, unstable copy-number |
| NIMROD | -                   | -                  | Shuffled neighbourhood |
| TEP    | E, V, S, S          | G, P              | Amino acid divergence, population variation, sparsely alignable, relaxed sequence |
| IMDSIG | A, U, S, T, S, N, E, V | C         | Ancient, widespread, stable copy-number, maintained neighbourhood, amino acid divergence |
| JASTSIG| E, V, S             | S                 | Amino acid divergence, population variation, shuffled neighbourhood |
| TOLL SIG| S, T, W             | C, W              | Stable copy-number, sparsely alignable |
| IMDMOD | A, U, P, S, S, P, C, F | C, V, E, N, S | Ancient, widespread, relaxed synonymous substitutions, population variation, constrained sequence, stable copy-number, amino acid conservation |
| JAST MOD| -                   | -                  | No extreme features |
| TOLL MOD| A, U, S, T, P       | P, S, C, E, V, N, S | Ancient, widespread, stable copy-number, constrained sequence, relaxed substitutions, amino acid divergence, population variation |
| CASP   | D, A, C, N, Y       | S                 | Duplications, losses, population variation |
| CLIPA  | -                   | A, U, S, Y        | Young, sparse, shuffled neighbourhood |
| CLIPB  | P, S, C, E, N, V    | A, U, S, T, S, D, S, W, G, P | Relaxed substitutions, losses, amino acid divergence, young, sparse, constrained synonymous substitutions, unstable copy-number, population conservation, sparsely alignable, relaxed sequence |
| CLIPC  | D, A, C, N, Y       | U, S, T          | Duplications, losses, amino acid divergence, sparse, unstable copy-number, relaxed sequence |
| CLIPD  | U                   | C                 | Widespread, stable copy-number, amino acid conservation |
| CLIP E | E, V, S             | A, U, G          | Amino acid divergence, young, sparse, relaxed sequence |
| IAP    | S                   | G                 | Relaxed substitutions, sparsely alignable |
| SRPN   | C, V, E, N          | S                 | Duplications, amino acid divergence, shuffled neighbourhood |
| AMP    | C, N, S             | A, U, S, W, G    | Losses, amino acid divergence, young, sparse, unstable copy-number, sparsely alignable |
| LYS    | D, A, C, N, E, Y    | S                 | Duplications, maintained neighbourhood, unstable copy-number |
| PPO    | D, A, C, N, Y       | S                 | Duplications, unstable copy-number |
| GPX    | E, P, H             | S                 | Duplications, constrained sequence, unstable copy-number |
| HPX    | U, S, W             | S, C, E, V        | Widespread, stable copy-number, widely alignable, relaxed substitutions, amino acid conservation |
| TPX    | A, U, S, T, W, G    | P, S, C, E, V, N, S, D | Ancient, widespread, stable copy-number, widely alignable, constrained sequence, constrained substitutions, amino acid conservation, population conservation |
| SOD    | W, G, P             | P, S, C, E        | Widely alignable, constrained sequence, constrained substitutions, amino acid conservation |
| APHAG  | A, U, S, T, W, G    | A, C, N, P, D, S, C, E, V, N, S, N | Ancient, widespread, stable copy-number, widely alignable, constrained sequence, constrained substitutions, amino acid conservation, population conservation |
| SRRP   | A, U, P, S, S, T, W, G, S, C, E, V, N, S | S, C, E, V, N, S, N | Ancient, widespread, relaxed substitutions, stable copy-number, maintained neighbourhood, widely alignable, constrained sequence, amino acid conservation, population conservation |
| SPZ    | S                   | -                 | Stable copy-number |
| TOLL   | A, U, D, A, C, N, W | P, S, C, E | Ancient, widespread, duplications, widely alignable, constrained substitutions |
Figure 1
Figure 2
Figure 4