Distribution Agreement

Yaw Kumi-Ansu April 10, 2018

Investigating evolution of innate immunity in insects: focus on two signaling pathways

by

Yaw Kumi-Ansu

Nicole M. Gerardo, PhD
Adviser

Department of Biology

Nicole M. Gerardo, PhD

Adviser

Jacobus de Roode, PhD

Committee Member

Timothy D. Read, PhD

Committee Member

2018

Investigating evolution of innate immunity in insects: focus on two signaling pathways

By

Yaw Kumi-Ansu

Nicole M. Gerardo, PhD

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Biology

2018

Abstract

Investigating Evolution of Innate Immunity in Insects: Focus on Two Signaling Pathways

By Yaw Kumi-Ansu

Innate immunity is the primary intrinsic immune response in invertebrates, whereas vertebrates are equipped with both innate and acquired immunity. For this project, I sought to investigate the evolution of innate immunity in two signaling pathways: Toll and Immune deficiency (IMD). In the IMD pathway, we investigated the pattern of loss of peptidoglycan recognition proteins (PGRP) in Insecta, and whether this may be linked to homology (loss restricted to a monophyletic cluster) or convergent evolution (independently arising trait). This was achieved by building a presence-absence profile of 173 species, of which 12 showed putative loss of PGRPs. We observed a combination of both types of loss, with Hemipterans (true bugs) displaying a distinct paraphyletic loss of PGRPs in sub-order Sternorrhyncha. Independently arising loss was found in some species flies and beetles. Endosymbiosis was a common trait among most of the species that showed putative loss, but a few species also hint at the possibility of adaptations to an extreme environment contributing to the loss of PGRPs. In the Toll pathway, we focused on signatures of selection of Toll-like receptor (TLR) genes in different populations of monarch butterflies. We hypothesized that trends in signatures of selection would be influenced by variations across populations based on prevalence of their natural protozoan parasite (*Ophryocyctis elektroschirrha*) and coexistence with different host plant communities of milkweed. Specifically, some milkweed species contain varying concentrations of toxic cardenolides which reduce disease burden from *O. elektroschirrha* in adults who fed on these as larvae. We hypothesized that there would be a difference in signatures of selection in immunity-related TLRs between the North American population which encounters lower parasite prevalence and toxic cardenolide concentration relative to other global populations. Alignment and construction of a phylogeny from TLRs sequences identified in monarchs and a few other species revealed two main types of clustering: intra- and interspecific clustering of TLRs for the species studied. Our investigation of signatures of selection via Tajima's D, revealed that our hypothesis was refuted since there wasn't a distinct trend between the North American population and the others regarding the signatures of selection of the TLRs.

Investigating evolution of innate immunity in insects: focus on two signaling pathways

By

Yaw Kumi-Ansu

Nicole M. Gerardo, PhD

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Department of Biology

2018

Acknowledgements

I am truly grateful to my lab mentor, Wen-Hao Tan, for providing me with data for analysis of signatures of selection and giving me great guidance throughout my period of doing research in lab. My sincere gratitude also goes to my PIs, Dr. Jacobus de Roode and Dr. Nicole M. Gerardo, who dedicated a lot of time and energy to my work and helping me to grasp and properly convey my research goals and findings. I would also like to thank Dr. Timothy Read for serving on my thesis committee, the Gerardo and de Roode labs for accepting me into the lab family and Dr. Berry Brossi for helping me to find a research field that fit my interests. I would also like to thank my biology professors from Oxford College, Dr. Eloise Carter, Dr. Nitya Jacob and Dr. Steve Baker, for helping me to hone my curiosity in biology. Finally, I would like to thank the Emory SIRE program for providing me with a grant to carry out my research.

Table of Contents

Introduction

Interactions between living organisms are shaped by evolution of each party in response to the other. One such relationship, parasitism, has been extensively studied across many taxa in order to tease apart the mechanisms that are involved in hosts' resistance and tolerance toward infection and in the ability of parasites to overcome host defenses. The coevolution of hosts and parasites, especially in instances of high virulence, has been suggested to be a driving force for the development of adaptions to keep hosts competitive against the harmful effects of parasites[1]. Hosts that bear the harmful effects of parasites must develop mechanisms to help them cope with or eliminate the parasite in order to improve their own chances of survival and reproduction. From this need arises immunity, which is a mechanism of recognition, signaling and response via many complex pathways that helps the host to defend itself against parasitism. Whereas vertebrates are equipped with both innate and acquired immunity, the former is the primary immune defense in invertebrates. Innate immunity is well conserved due to its crucial role in primary defense against microbial infection in invertebrates[2]. Innate immune pathways include the Toll, Immune deficiency (IMD), JAK/STAT and JNK pathways. Each of these is associated with different pathogen associated molecular patterns (PAMPs), which are sensed by their pathogen recognition receptors (PRRs), leading to a signaling cascade that ends in the transcription of genes that encode for antimicrobial peptides (AMPs), melanization by phenoloxidases, encapsulation, and degradation of RNA (in the case of viruses). In this study, we focus on the IMD and Toll pathways, both of which are NF-κb signaling pathways. We investigate the evolution of recognition genes in the IMD pathway across insects, followed by a more in-depth look at Toll-like receptors (TLRs) in monarch butterflies.

Chapter 1 | The presence and absence of peptidoglycan recognition proteins in Insecta.

BACKGROUND

The Immune deficiency (IMD) pathway is an important component of the innate immune system in insects and is chiefly responsible for responding to infections by Gram-negative bacteria[3]. Peptidoglycan recognition proteins (PGRPs) serve as the pathogen recognition receptors (PRRs) in this pathway. A downstream signaling cascade is initiated upon interaction with DAP-type peptidoglycan fragments, which are mostly unique to Gram-negative bacteria. Signaling leads to the activation and translocation of Relish, an NF-κb transcription factor, into the nucleus and the subsequent expression of anti-microbial peptides (AMPs), such as Diptericin, to fight invading bacteria[4]. Since PGRPs have been suggested to be crucial to activation of this

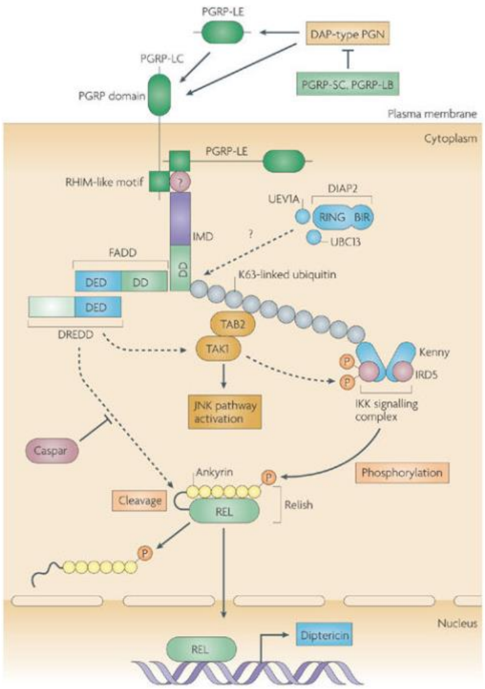


Figure 1. IMD pathway in *Drosophila melanogaster* (Ferrandon et al. 2007, modified)[3]

pathway, it is of importance to study their evolutionary history in relation to host-parasite interactions.

Previous research has identified certain species within phylum Arthropoda (e.g. *Daphnia spp., Ixodes scapularis, Pediculus humanus*) that are missing genes within innate immune pathways[5]. A case in point is the pea aphid (*Acythrosiphon pisum*), in which several genes for factors in the IMD pathway could not be identified either by experimental testing for activation of the IMD pathway, or by searching of the annotated genome[6]. Factors that were absent are known to function in recognition (receptors - PGRP), signaling (signal transduction and transcription factors - IMD, DREDD, Relish) and response (anti-microbial peptides – defensin, cecropin). The loss of such key components of the IMD pathway suggests an elimination of its activity. Analysis of the antimicrobial peptide (AMP) expression upon inoculation of pea aphids with Gram-negative bacteria showed little to no activation of AMPs, which is one of the mechanisms of innate immune response. We sought to expand this search for possible loss of factors in the IMD pathway to more species in class Insecta to determine the pattern of loss of these purportedly essential genes.

We propose two possible predictions for the observation of loss of these canonical immune genes: convergent evolution and homology. With convergent evolution, the loss of canonical innate immune genes may be associated with certain derived traits in different species that influence their interactions with pathogens and parasites. Examples of these are mutualistic interactions with microbial endosymbionts (which is seen across an array of insect species) and sociality, in which some species of ants, bees, termites, as well as some species of aphids, live in close proximity with their relatives. If these traits influence the maintenance of immune defenses, loss is likely to occur in distantly related species. In contrast, the loss of canonical immune genes through homology would suggest that loss occurred in an ancestor relatively long ago and the extent of loss would be restricted to a group of closely related species. Testing these alternatives requires a comprehensive assessment of the presence and absence of genes across a broad set of taxa. Here, we use available, high quality Insecta genomes to assess for patterns of loss of PGRPs.

METHODOLOGY

Selection of Species

Insect genomes were compiled from NCBI by obtaining a list of all insects with sequenced genomes[7]. To control for genome quality, we refined the results to restrict the list to only species that had been sequenced to the level of chromosomes or scaffolds. To further refine this list to perform BLAST, we obtained a list of all insect species in the National Agricultural Library of the USDA using the i5k Workspace[8]. We excluded all species that had a contig N50[9] less than 0.1Mbp as our arbitrary criterion of quality of genome assembly[10] . This was then compared to the list from NCBI to further exclude species that did not meet this requirement.

Obtaining sequences for queries

We first searched for PGRP sequences in NCBI under the protein and nucleotide databases and then restricted the results by taxonomy to include only insects. This list of sequences and species was cross-referenced with the list of insect species with acceptable genome quality from NCBI, and the combined list was used to create an ontology in phyloT and visualized in the Interactive Tree of Life (iTOL) using taxonomic names from NCBI as queries[11]. This produced an ontology that is based on the consensus between taxonomies provided in NCBI for the individual species included in the query. The species that were identified to have PGRP sequences based on the NCBI search, were highlighted and their sequences were used as queries for their close relatives in the ontology. This list was also divided into different orders and a representative species (most well studied/annotated genome) was selected for each of the main orders represented in the set. (*Drosophila melanogaster, Anopheles gambiae* - Diptera, *Tribolium castaneum* - Coleoptera, *Bombyx mori*- Lepidopetra, *Apis mellifera* – Hymenoptera). For orders that had too few members, sequences from the representative species in the closest order were used as queries.

BLAST searches

Protein blast (blastp) was performed in NCBI and/or i5k for the selected species, using the sequences of the closest relative (with annotated/characterized PGRP sequences) as queries. For species that also had separate/specialized databases (e.g. Lepbase[12], Vectorbase[13], Aphidbase[14]), local blasting was performed within these to ensure that the most updated genome was being searched. The E-value was used to evaluate quality of match to the query sequence, for which the cut-off point was set at 10^{-15}. For species in which PGRP sequences were identified, a reverse blast (searching the resulting matches from a blast back within the initial query's database) was used as a test for the blasting methodology, and to provide more confidence as to the homology of the sequences. For species that yielded no results via blastp (protein to protein blast), we searched the nucleotide genome using nucleotide sequences from the representative species as queries. If significant matches were obtained from this method (i.e. E-value > 10^{-15}), they were translated into amino acid sequences in EMBOSS Transeq[15] and searched in the Simple Modular Architecture Research Tool (SMART)[16] to receive an e-value for the PGRP domain of the sequence against representative domain sequences and domain databases (eg, Pfam[17]).

Verification of absence and analysis of PGRP amidase sequences

Further verification of genome quality for species that had shown loss of PGRPs was performed by searching for housekeeping genes (HKGs - e.g. alpha and beta tubulins, and GAPDH)[18], using sequences from *Drosophila melanogaster* as queries. The species that were excluded from further analysis did not show any matches to queries of the HKGs from fruit flies in NCBI, which was interpreted as an indication of poor genome quality (insufficient genome coverage). Therefore, the set of species that was maintained for analysis was reduced to only those in suborder Sternorrhyncha (henceforth referred to as the 'absence group'), which were the only species to provide matches to the HKG queries (Table 1). A nucleotide blast (blastn) was performed for members of the absence group, using the PGRP sequences from *Bemisia tabaci* as query and the identified nucleotide sequences that were translated into proteins in EMBOSS Transeq. These were searched in the Pfam database to determine their level of similarity to the HMM consensus sequence for the PGRP amidase domain. The sequence identified in the pea aphid was searched against reference proteomes in HMMER[19] to observe the distribution of matches across taxa. Also, the PGRP domain sequence of the fruit fly was obtained from Flybase and searched in the same manner. A phylogeny was reconstructed based on the PGRP amidase genes identified in the absence group along with the query sequence from *Bemisia tabaci* and *Occidentia macillensis*. Another phylogeny for of the absence group was reconstructed using Bayesian inference based on sequences of Gram-negative binding proteins (GNBPs) in the absence group, which are known to be functional in pea aphids. For pseudogenes, the e-value is usually quite high (greater than 1.0e-5) indicating a generally conserved sequence but with significant changes that renders them significantly different when compared to the consensus of the Hidden Markov Model (HMM) for the domain in Pfam. In cases where no sequences were obtained at this point, a general database search was done using the species name and PGRP as keywords for queries in NCBI. This was to identify publications or projects that may address PGRP sequences or may have studied general innate immunity in these species. In the case where no results were obtained after this search, it suggested the following: a) the gene could be missing from the genome or significantly altered (identity and/or function) or, b) the gene might be present, but the genome quality was not good enough to come to a reliable conclusion as to its absence.

RESULTS

BLAST search

A total of 173 species of insects were selected for BLAST searches to determine the absence of PGRP genes in their genome. Most of these species belonged to Diptera (flies), Hemiptera (true bugs), Lepidotera (moths and butterflies) and Hymenoptera (bees, ants, etc.) with few representatives for Trichoptera (caddisflies, n=1), Odonata (dragonflies and damselflies, n=3) and Blattodea (termites and cockroaches, n=2) (Figure 2). We identified 12 species that did not yield any matches after blasting, with a few of these occurring in isolated cases where no other species in an immediate cluster/clade also showed a loss (Fig. 2a). Furthermore, a distinct cluster of species that showed a putative loss of PGRPs was found in the sub-order Sternorrhyncha. Specifically, absences were found in the families of Psylloidea (*Diaphorina citri, Pachypsylla venusta*) and Aphididae (*Acythrosiphon pisum, Diuraphis noxia, Myzus persicae*) (Fig.2b).

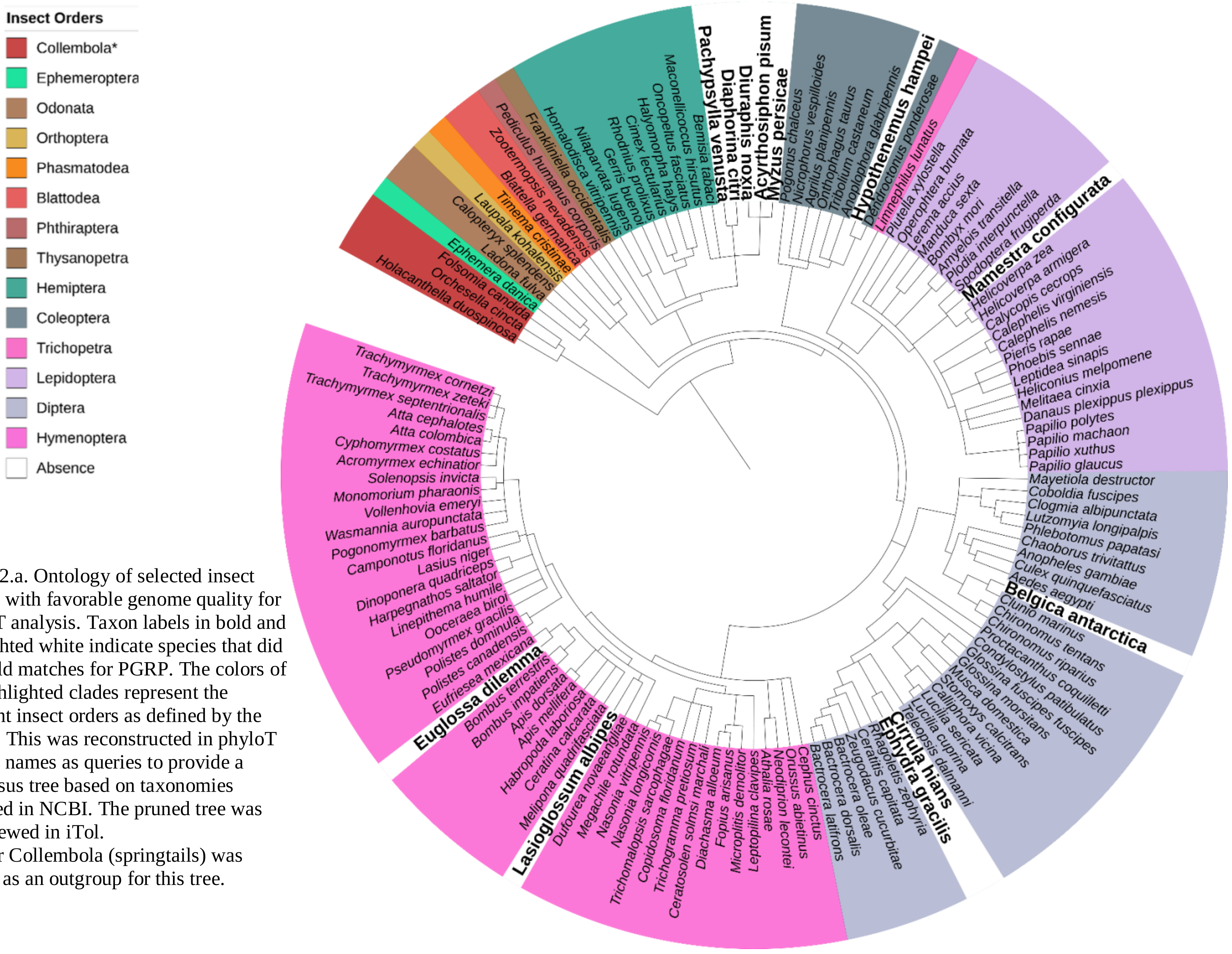


Figure 2.a. Ontology of selected insect species with favorable genome quality for BLAST analysis. Taxon labels in bold and highlighted white indicate species that did not yield matches for PGRP. The colors of the highlighted clades represent the different insect orders as defined by the legend. This was reconstructed in phyloT species names as queries to provide a consensus tree based on taxonomies provided in NCBI. The pruned tree was then viewed in iTol.
* Order Collembola (springtails) was chosen as an outgroup for this tree.

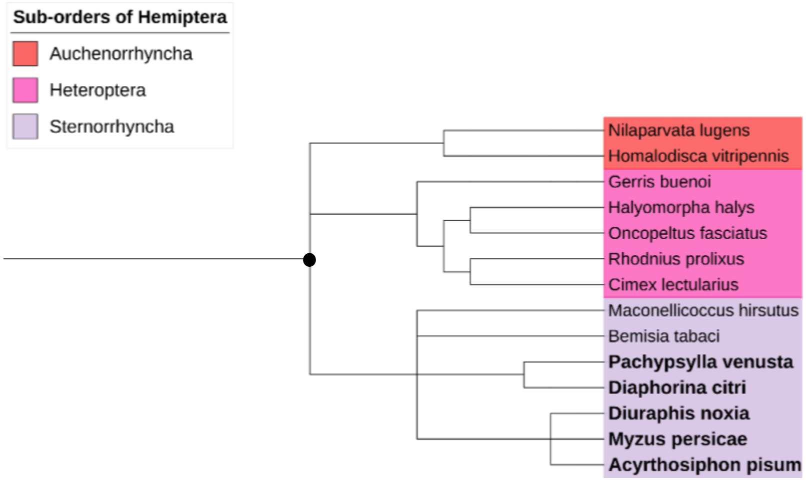
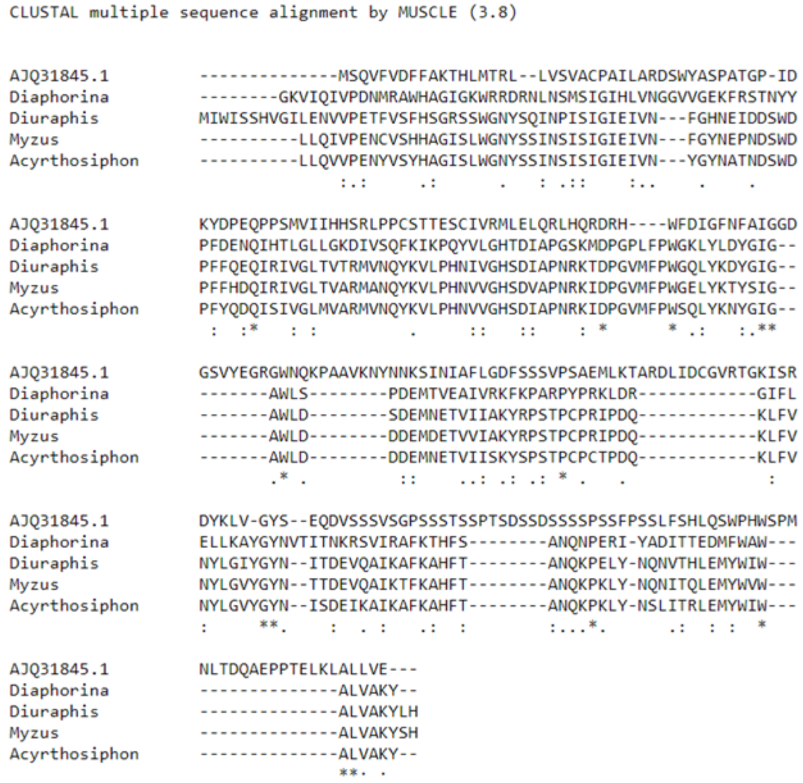


Figure 2.b. Subset of ontology from 2.a focusing on Order Hemiptera (true bugs). Absence group is in bold. Highlighted clades represent the different sub-orders in Hemiptera defined in the legend and black dot represents the internal node for Hemiptera.

Nucleotide blast of the genome of species in the absence group yielded significant matches in the i5k site using the PGRP sequences from the whitefly as the query. Only one PGRP amidase sequence was found in each of the species in the absence group, except *Pachypsylla venusta*, which showed none. Searching these proteins in the Pfam database revealed that they matched the HMM consensus profile for the Type-2 amidase domain, with e-values between 10^{-10} and 10^{-16}(Table 1). The protein alignment between these sequences showed strong similarity between the sequences of the absence group (Fig. 3), which varied significantly from the Type-2 amidase sequence from whiteflies. Searching the Type-2 amidase domain sequence of the pea aphid in HMMER yielded few high matches in Eukaryota (e-value < 10^{-100}), while most matches were in bacteria with E-values below 10^{-30} (Fig. 4a). Performing a similar search using PGRP sequences from fruit flies and whiteflies showed that these had many high-quality matches which were mainly found in Insecta (Fig. 4 c and b).

Table 1. E-values of N-acetylmuramoyl-L-alanine (Type-2) amidase domain from putative PGRP sequences in Pfam

Species	E-value
Diuraphis noxia (Russian wheat aphid)	8.5x10^{-10}
Myzus persicae (green peach aphid)	5.6x10^{-11}
Acythrosiphon pisum (pea aphid)	7.3x10^{-13}
Diaphorina citri (Asian citrus psyllid)	2.9x10^{-15}
Bemisia tabaci (silverleaf whitefly)	7.3x10^{-12}



```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

AJQ31845.1         --------------MSQVFVDFFAKTHLMTRL--LVSVACPAILARDSWYASPATGP-ID
Diaphorina         --------GKVIQIVPDNMRAWHAGIGKWRRDRNLNSMSIGIHLVNGGVVGEKFRSTNYY
Diuraphis          MIWISSHVGILENVVPETFVSFHSGRSSWGNYSQINPISIGIEIVN---FGHNEIDDSWD
Myzus              ----------LLQIVPENCVSHHAGISLWGNYSSINSISIGIEIVN---FGYNEPNDSWD
Acyrthosiphon      ----------LLQVVPENYVSYHAGISLWGNYSSINSISIGIEIVN---YGYNATNDSWD
                                 :.:     .:      .   : .::    :..    .    .

AJQ31845.1         KYDPEQPPSMVIIHHSRLPPCSTTESCIVRMLELQRLHQRDRH----WFDIGFNFAIGGD
Diaphorina         PFDENQIHTLGLLGKDIVSQFKIKPQYVLGHTDIAPGSKMDPGPLFPWGKLYLDYGIG--
Diuraphis          PFFQEQIRIVGLTVTRMVNQYKVLPHNIVGHSDIAPNRKTDPGVMFPWGQLYKDYGIG--
Myzus              PFFHDQIRIVGLTVARMANQYKVLPHNVVGHSDVAPNRKIDPGVMFPWGELYKTYSIG--
Acyrthosiphon      PFYQDQISIVGLMVARMVNQYKVLPHNVVGHSDIAPNRKIDPGVMFPWSQLYKNYGIG--
                    :  :*   : :         .     ::   ::    : *      * .:   :.**

AJQ31845.1         GSVYEGRGWNQKPAAVKNYNNKSINIAFLGDFSSSVPSAEMLKTARDLIDCGVRTGKISR
Diaphorina         -------AWLS--------PDEMTVEAIVRKFKPARPYPRKLDR------------GIFL
Diuraphis          -------AWLD--------SDEMNETVIIAKYRPSTPCPRIPDQ------------KLFV
Myzus              -------AWLD--------DDEMDETVVIAKYRPSTPCPRIPDQ------------KLFV
Acyrthosiphon      -------AWLD--------DDEMNETVIISKYSPSTPCPCTPDQ------------KLFV
                          .* .         ::    ..: .: .: * .   .              :

AJQ31845.1         DYKLV-GYS--EQDVSSSVSGPSSSTSSPTSDSSDSSSSPSSFPSSLFSHLQSWPHWSPM
Diaphorina         ELLKAYGYNVTITNKRSVIRAFKTHFS--------ANQNPERI-YADITTEDMFWAW---
Diuraphis          NYLGIYGYN--ITDEVQAIKAFKAHFT--------ANQKPELY-NQNVTHLEMYWIW---
Myzus              NYLGVYGYN--TTDEVQAIKTFKAHFT--------ANQKPKLY-NQNITQLEMYWVW---
Acyrthosiphon      NYLGVYGYN--ISDEIKAIKAFKAHFT--------ANQKPKLY-NSLITRLEMYWIW---
                   :     **.    :  . :   .:  :        :...*.      .:  : :  *

AJQ31845.1         NLTDQAEPPTELKLALLVE---
Diaphorina         --------------ALVAKY--
Diuraphis          --------------ALVAKYLH
Myzus              --------------ALVAKYSH
Acyrthosiphon      --------------ALVAKY--
                                 **. .
```

Figure 3. Multiple sequence alignment of PGRP amidase sequences from absence group. AJQ31845 is the accession number for the sequence derived from *Bemisia tabaci*. Alignment was done in MUSCLE[14].

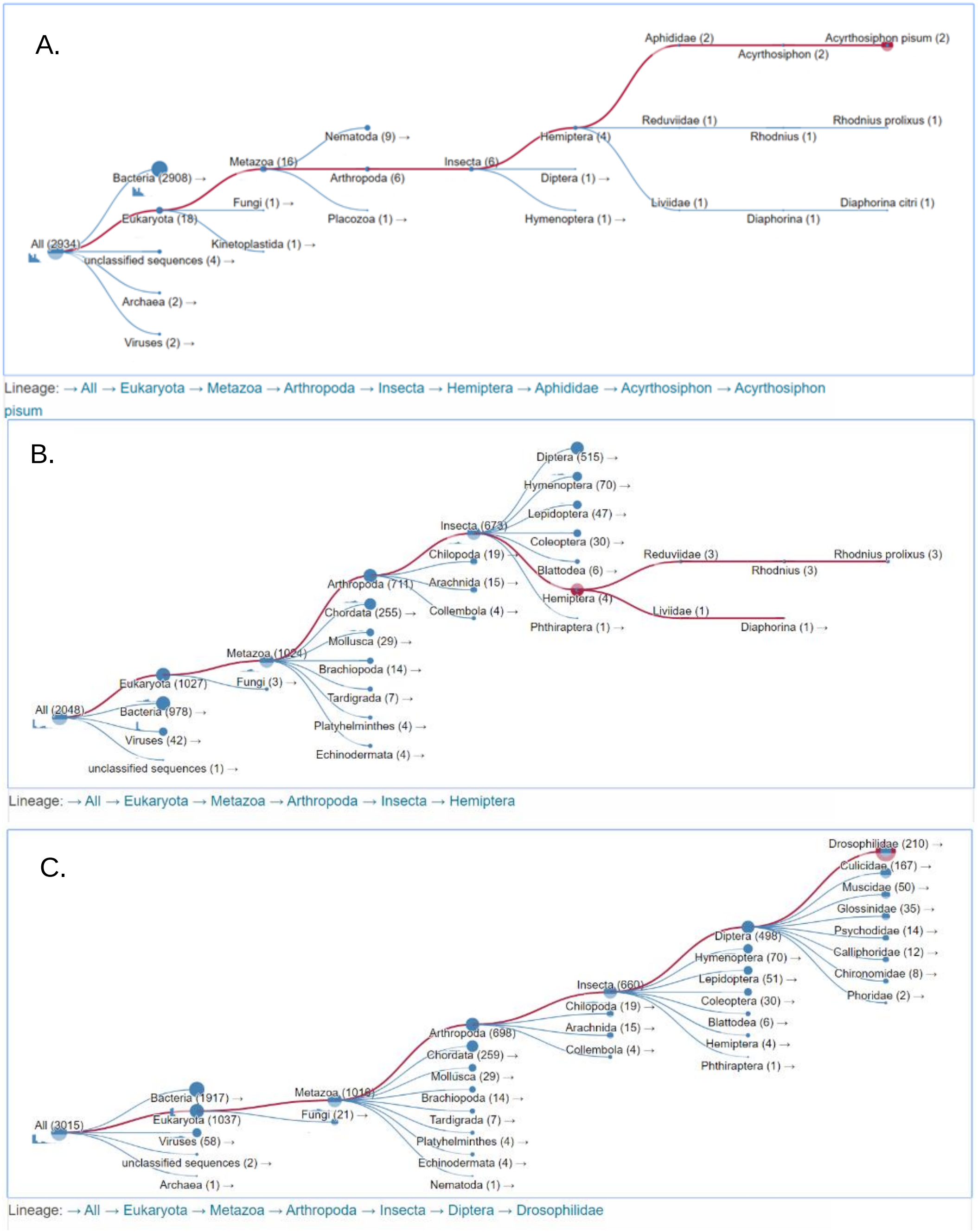


Figure 4. Taxonomic distribution of search matches to PGRP Type-22 amidase domain using sequences from a) *Acythrosiphon pisum*, B) *Bemisia tabaci* and C) *Drosophila melanogaster* as queries in HMMER[19]. This shows that the PGRP amidase domain found in *A. pisum* is more similar to those found in bacteria, rather than in insects based and the very low number of matches in Insecta, relative to matches observed for *B. tabaci* and *D. melanogaster*.

PGRP amidase sequences vs. GNBP

The PGRP amidase domain sequences from the absence group clustered together with a probability of 0.7056. *Diaphorina citri* had the earliest and highest divergence within this clade with a raw branch length of 3 (Fig. 5a). The genes from *B. tabaci* and *Occidentia massiliensis* showed the highest divergence from this clade (branch length = 4). The phylogenetic tree of GNBPs (Fig. 5b) showed a high posterior probability for the clustering of all the species throughout the phylogeny. All the GNBPs for *B. tabaci* formed a monophyletic cluster with a posterior probability of 1.000, while those belonging to members of the absence group formed mixed clusters throughout the phylogeny with high posterior probabilities.

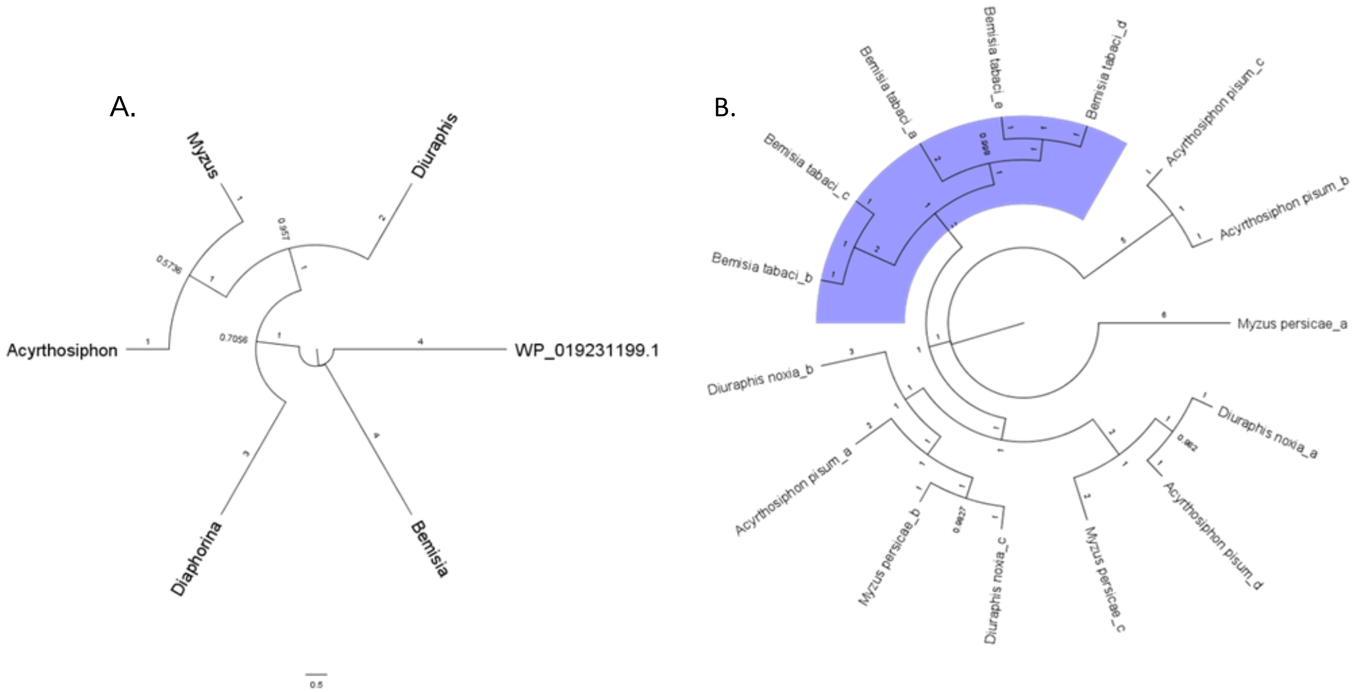


Figure 5. Phylogenetic trees reconstructed using Bayesian inference in Mr.Bayes, based on A) PGRP and B) GNBP sequences. Trees were visualized in FigTree. Monophyletic cluster of *Bemisia tabaci* GNBPs is highlighted. Branches are transformed and raw branch lengths are displayed based on analysis from Mr.Bayes. Note: Taxon label WP_019231199.1 in 5.A represents the PGRP amidase sequence obtained from *Occidentia massiliensis* (bacterium).

DISCUSSION

In this chapter, we sought to identify species of insects that exhibit a putative loss of genes in the IMD pathway, using absence of PGRPs as our main indication of loss. In our investigation, we employed BLAST searches, multi-sequence alignment and phylogenetics. We considered that loss may be due to homology, which would imply a loss of genes in a common ancestor that is then expressed in descendants clustered together monophyletically. Loss could also be potentially due to convergent evolution mediated by independently emerging traits such as sociality or symbiotic relationships with microbes.

From our investigation, there were 12 species (out of 173) with a putative absence of PGRP genes. We observed both independent origins of PGRP absence in single species (n=5), and absence that followed a pattern of clustering within an unresolved clade of Hemipterans (n=5). PGRPs were also detected in the genomes of another monophyletic group within Dipterans (n=2) (Fig 1). This clade consisted of two species of shoreflies (*Cirrula hians* and *Ephydra gracilis*), both in sub-order Brachycera. These species are closely related within the family Ephydridae (shoreflies), which are known to live in hypersaline environments such as Mono Lake in California and the Great Salt Lakes in Utah [20]. It is possible that living in such an environment may have led to the development of adaptations that may affect different factors within pathways of the innate immune system. This may be influenced by the type and diversity of bacteria in these regions owing to the adaptations required to thrive in this unique environment. Similarly, *Belgica antartica*, a dipteran that is endemic to Antarctica, did not produce matches to PGRP queries. This insect species has the smallest known genome which has been hypothesized to be an adaptation to living in an extremely cold environment[21]. The absence of PGRP genes in these two groups of organisms that live in ‘extreme’ environments could be an indication of how changes in the innate immune response may be associated with the development of coping mechanisms to surviving in these environments.

We observed particularly pronounced loss of PGRP genes in Hemipterans, especially within the sub-order Sternorrhynca (Fig 2), which includes aphids. We observed two distinct clades within this suborder corresponding to two super-families that showed absence namely, Psylloidea (*Pachypsylla venusta* and *Diaphorina citri*) and Aphididae (*Acythrosiphon pisum*, *Myzus persicae* and *Diuraphis noxia*) (Fig 2). A key trait of most species within this sub-order is endosymbiosis, which assists them in nutrient extraction from their food sources (i.e. plants)[22]. The putative absence of PGRP sequences may relate to adaptations to reduce the effect of the IMD pathway on Gram-negative endosymbionts. Outside Hemipterans, we also observed a putative loss of PGRP in *Hypothenemus hampei*, commonly known as the coffee-borer beetle. This species has been shown to be heavily reliant on microbial symbiosis to degrade caffeine. Groups that were experimentally treated with antibiotics were found to have significantly reduced reproductive fitness, which may be linked to the essential role symbiosis plays in their nutrition[23]. This points to the possibility of the linkage between a trait derived via convergent evolution and the development of different mechanisms to sustain such relationships, either by upregulation of inhibitory factors in the IMD pathway (e.g. PGRP-LB), or the loss of activation and signaling genes via negative selection[24]. Some studies outside arthropods have shown that some host are able to maintain gram-negative bacterial endosymbionts by resorting to the degradation of their peptidoglycan fragments via amidase domains in EsPGRP2, thereby inhibiting activation of the IMD pathway[25]. This allows the host to create a favorable environment for their endosymbiont while preventing the cost associated with constantly mounting an immune response. Therefore, there could also be other complex adaptations for protecting the gut microbiota while simultaneously preventing hyperactivity of the host's immune system in insects that rely on symbiosis. The presence and retention of the activity of GNBP in some species of the absence group indicates a maintenance of interaction with Lys-type peptidoglycan fragments from Gram-positive bacteria, which is mainly catered to by the Toll pathway. However, the absence of PGRP genes could hint at the loss of function of these genes via negative selection, in favor of endosymbiosis. It is therefore promising to investigate how the diversity of endosymbionts in such species, with regards to Gram-negative and Gram-positive bacteria, relates to the presence and absence of recognition factors within different innate immune pathways. Further investigation is particularly important given that symbiosis is not restricted to the absence group, and therefore symbiosis itself could not be the only determinant of absence of PGRPs.

Secondly, our blast results identified a single Type-2 amidase domain sequence that was very well conserved between *Diuraphis noxia, Diaphorina citri, Acythrosiphon pisum* and *Myzus persicae* as observed from the alignment of their protein sequences (Fig. 2). Blasting these sequences against all protein databases in NCBI and HMMER revealed a high number of matches with Type-2 amidase sequences found in several bacteria. The type-2 amidase domain is an important functional unit of PGRPs that mediates interaction with DAP-type peptidoglycan fragments from Gram-negative bacteria. This domain is also present in bacteria and bacteriophages and helps in degrading peptidoglycan via hydrolase activity[26]. This points to the possibility of the horizontal gene transfer between bacteria and their host. Due to the high level of conservation of this sequence between the insect species, it is likely that this amidase gene was either incorporated into the genome of an ancestor to these species, or recently transferred by the same species of bacteria to the different species in the absence. A well-supported match for this sequence was also identified in *Rhodnius prolixus*, which falls outside the sub-order Sternorrhynca. Studies that have investigated the activation of the IMD pathway in pea aphids have shown little to no activation or production of AMPs after infection with Gram-negative bacteria[6]. This is in contrast to the scenario observed in *Rhodnius prolixus*, which shows a loss of factors within the IMD pathway, but still has functional PGRP genes and the amidase gene in question. Infecting *R. prolixus* with Gram-negative bacteria results in activation of the IMD pathway, ending in transcription of AMPs via homologs of NF-κB transcription factors[27]. *Diaphorina citri* presents a peculiar case, compared to the other species described. Besides showing putative loss of PGRPs, a previous study also failed to locate sequences for both GNBPs and β-1,3-glucan recognition proteins, which are the main recognition proteins in the Toll pathway for interacting with Lys-type peptidoglycan from Gram-positive bacteria[28]. This also presents an interesting system to study the effects of loss of recognition components in the Toll pathway and how this may relate to endosymbiont diversity in such insects.

Therefore, there is reason to infer that the level of reliance on endosymbionts for nutrition and metabolism might be related to the strength of selection on genes for factors in the innate immune pathways. In this case, there is the likelihood of differential selection on different factors in the IMD pathway of insects based on reliance on endosymbionts, which would lead to adaptations that would mediate interactions with their microbes. In the future we hope to investigate the possibility of a link between variation in the nutritional and developmental importance of endosymbionts for insect hosts and the signature of selection on factors within the IMD pathway of hosts, especially within recognition factors.

Chapter 2 | Toll-like receptor genes in Monarchs

BACKGROUND

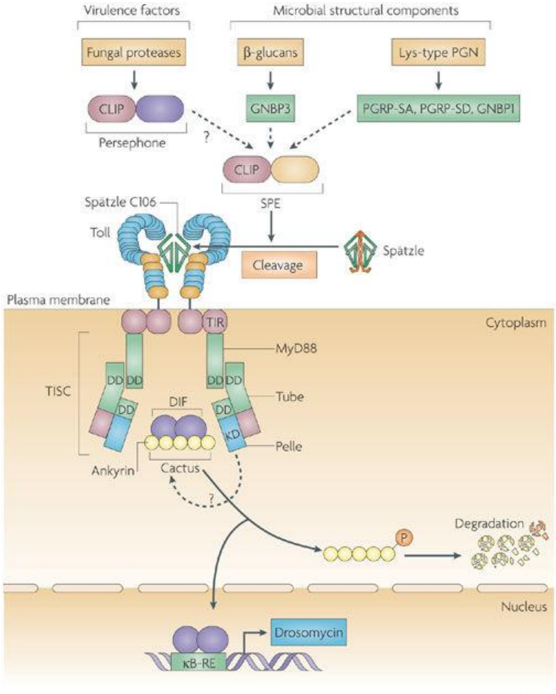
Monarch butterflies are an emerging model organism for population genetics and genomics given their unique migration and global dispersal into distinct populations that are now existing under different ecological conditions. Previous studies have identified variation in signatures of selection on genes associated with flight muscles and wing morphology[29]. This variation has been suggested to be due to differences in migration in different populations of monarchs in two main classes: migratory and non-migratory (resident) populations[30]. Migratory monarchs from North America dispersed to Central America, the Pacific and Atlantic in the 1800s. These new populations stopped migrating and became residents in these regions. Beyond the change in climate and migratory behavior, other changes emerged that are important to their survival and immune defenses. One important difference is disease prevalence. Specifically, there is an inverse relationship between propensity of migration and prevalence of the natural protozoan (apicomplexan) parasite of monarchs, *Ophryocystis elektroscirrha* [31]. The North American population has been shown to have less disease prevalence than the Pacific, Central American and Atlantic populations. These differences set the stage to study the population genetics of immune and defense related genes across monarch populations.

Other ecological conditions within these populations could also shape disease prevalence. During the larval phase, monarch butterflies consume large amounts of milkweed and increase their body mass about 3000 times from first to fifth instar. Besides providing the nutrients to facilitate their metamorphosis from larvae to adult, feeding on milkweed during the larval stage has been shown to reduce spore loads of *Ophryocystis elektroscirrha* in adult monarchs[32]. Specifically, species of milkweed that contain toxic cardenolides (medicinal milkweeds) have been shown to be an alternative, plant-derived form of defense against infection by *O. elektroscirrha*.

The efficacy of medicinal milkweed species in reducing spore load has been suggested to be due to their relative concentrations of non-polar cardenolides, which are more toxic than polar cardenolides due to their ability to cross cell membranes[16]. Studies have shown that species with high concentration of non-polar cardenolides provide significantly higher growth reduction of *O. elektroscirrha* than species with lower concentrations. There is variation in distribution of species of medicinal milkweed, with communities in North America tending to have a lower concentration of non-polar cardenolides than other global communities. These global populations include the Atlantic (Portugal, Morocco, Spain), Pacific (Samoa, Fiji, Australia, New Zealand) and Central American (South Florida, Belize, Costa Rica, Ecuador, Aruba) populations. This therefore posits variation in host-parasite interactions across populations of monarchs when plant-derived defense is factored into this relationship. This assumes that populations coexisting with communities of milkweed with high concentrations of toxic cardenolides would rely more on plant-derived defense in the reduction of parasite growth than would populations coexisting with communities with lower concentrations.

While monarchs may use milkweeds as medicine, they also mount innate immune-based defenses, though these are not well characterized. A few studies have shown that melanization by phenoloxidases and upregulation of hemocytes are some of the innate immune responses to *Ophryocystis elektroscirrha* in monarchs[33] . Since *O. elektroscirrha* is in the phylum Apicomplexa, we sought to draw knowledge from studies into insect interactions with other apicomplexan parasites. Arguably the most widely studied apicomplexans are malaria-associated *Plasmodium spp*. which are carried by female mosquitoes[34]. Studies have shown that different innate immune pathways are activated in response to different species of *Plasmodium*. Specifically, *Plasmodium berghei* activates the Toll pathway, while *Plasmodium falciparum* activates the IMD pathway[35].

In this study, we opted to study the Toll-like receptors (TLRs), which are the signaling factors in the Toll pathway (Fig. 1). Besides being involved in immunity, TLRs play roles in development in many tissues of the insect body and at different stages throughout their lifetime, as characterized in fruit flies and silkworms[36–38]. TLRs are transmembrane proteins with an extracellular domain comprised of leucine rich repeats that function in recognition, and an intracellular Toll-interleukin receptor (TIR) domain responsible for signaling. In fruit flies, activation of the pathway by Gram-positive bacteria, fungus and yeast results in the production of antimicrobial peptides. In development, the *Drosophila melanogaster* Toll gene has been shown to be involved in dorso-ventral axis formation[38].



Nature Reviews | Immunology

Figure 1. Toll pathway in *Drosophila melanogaster* (Ferrandon et al. 2007).

The structure and varied functions of TLRs makes them good candidates to study the possible effects of host-parasite dynamics on the evolution of innate immune recognitions factors. In such a system, we expect that TLRs involved in developmental functions would have a stronger and more uniform signature of selection across populations, while TLRs that are involved in immune functions would be under varying strengths of selection depending on variation in ecological factors and disease exposure. In monarchs, both the variation in distribution of medicinal milkweeds (based on concentration of non-polar cardenolides) and parasite prevalence have the potential to influence the trajectory of evolution of defense across monarch populations.

We sought to answer the following question: is there a difference in the signatures of selection for TLRs in different global populations of monarch butterflies, and is this difference associated with the distribution of medicinal milkweeds and variation in disease prevalence? We hypothesized that populations of monarchs in geographic regions with communities of milkweed that tend to have a high concentration of non-polar cardenolides (outside North America) would have relaxed selection on their TLRs owing to a higher dependence on milkweed for defense against *O. elektroschirra*. In this scenario, there might be increased reliance on medicinal milkweed as an alternative defense mechanism over canonical innate immune pathways. Therefore, we predicted that immune-related Toll genes from the North American population would exhibit signatures of selection that would vary the most from the genes of their background genome compared to other population. To address this question, we identified TLR genes in monarchs based on sequence homology with characterized TLRs from other species of insects. We analyzed signatures of selection of these genes between the different populations to determine if there was a trend between the North American population and the other three populations studied.

METHODOLOGY

Obtaining Toll sequences for BLAST queries

Toll-like receptor sequences (TLR) were selected from a few species that had annotated and/or characterized TLR sequences. These included *Bombyx mori, Anopheles gambiae, Drosophila melanogaster*, *Plutella xylostella* and *Manduca sexta* (Fig.2). Both protein and nucleotide TLR sequences from these species were obtained from Lepbase, Flybase, Vectorbase and NCBI to get the most updated set of TLR sequences for each species. These sequences were then aligned in ClustalX[39] to observe highly conserved regions which could be sub-selected and used in BLAST searches to obtain more refined matches. This was done to prevent false negatives that may occur due to regions of sequences that may be highly variable between and within species.

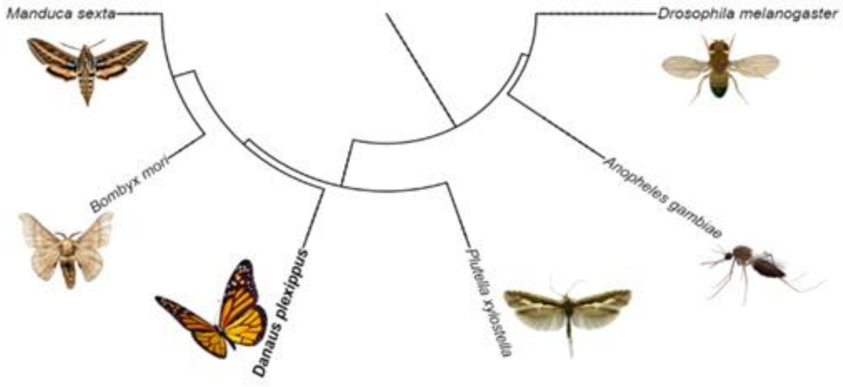


Figure 2. Phylogeny of insect species used as queries for TLRs in Monarchbase. Tree was reconstructed in phyloT online phylogeny creator. Scientific names were used as queries to obtain consensus tree based on taxonomies provided for the various species in NCBI.

Categorization of TLR sequences for BLAST

Based on the knowledge that there is a high level of conservation of Toll genes across Insecta[40], we reconstructed a phylogeny using TLR genes from each of the selected species. From observing the clustering pattern of these genes, we sought to find matches in the monarch genome by using sequences from the different species that form monophyletic groups for particular TLR genes as queries to blast against the monarch genome. For each set of queries used for a blast in Monarchbase[41], the top three matches (based on e-value) were recorded to determine the best matching sequence to the query sequences from a monophyletic cluster in the initial phylogeny. The best matches were reverse blasted in NCBI against the genomes of the reference species to make sure the match quality was reciprocated. The top matches for each set of TLR genes in monarchs was named and the sequences were aligned with the sequences that were used in the primary phylogeny. The TIR domain, which is known to be highly conserved in arthropods and mammals relative to other portions of the TLR sequences[36], was sub-selected and realigned to be used in reconstructing the final phylogeny by Bayesian inference in MrBayes v3.2.6. The analysis was run by setting priors to mixed-fixed rate amino acid models. A million generations were run, and the sample frequency was set at 1000. Convergence was reached when the standard deviation of split frequencies for the two simultaneous runs approached 0.01. Pattern of clustering of *B. mori* was used in identifying orthologs in monarchs since they are the closest relative to monarchs with fairly well characterized TLRs.

Obtaining sequences from geographic populations and setting up analysis

A pipeline was constructed by Wen-Hao Tan (graduate student and lab mentor) in Linux and Python to analyze signatures of selection across innate immune genes against the background genome for different geographic populations of monarchs. Sequences were extracted from the monarch genome databases compiled by Zhang et. al (2011)[42]. Populations included in this investigation were selected based on quality of genome coverage and number of replicates. Those selected are the Atlantic (n=6, from Spain and Morocco, excluding Portugal), Pacific (n=12, excluding Hawaii), North American (n=12) and South Florida (n=7) populations[43]. Immune genes were identified based on annotations from the Heliconius Genome Consortium[44] and those located on sex chromosomes were excluded from the set.

Calculation of Tajima's D and data analysis

Tajima's D was calculated for all genes across the selected populations in ANGSD[45] (which analyzes next generation sequencing data). This is based on analysis of observed polymorphisms in sequences at different sites in a population, compared against expected variation at these sites. For comparative analysis of immune genes against the background genome, sex-linked immune genes were excluded, and a subset of signaling genes (which includes TLRs) was sub-selected to compare against the background. Outliers of genes in the subset, with respect to the background, were classified as those that had values falling above 90% or below 10% of the background's range of coverage. TLR genes that satisfied this condition were identified based on the results of BLAST searches and phylogeny of TLRs of selected species in comparison to the putative TLR genes.

RESULTS

A total of 10 putative TLR sequences were identified in the monarch genome using characterized sequences from the selected species of insects (Fig.3). The phylogenetic tree reconstructed in MrBayes using TIR domain sequences yielded two main types of clustering: intraspecific and interspecific clustering of the TIR domain of the different species (Figure 3, Table 1). A major cluster of TLRs 3, 4, 5 (Fig.4, clade C) contained the intraspecifically clustering TLRs while other TLRs mainly clustered interspecifically (Table 1). There are two clusters of TLR 9 which both diverged earliest from the main tree (Figure 4, clades G and H), each with posterior probabilities of 1.000. Both clades contained a unique TLR 9 gene from monarchs, as was observed for most of the species in the tree.

At the next major node, we observed a divergence of clade C, which contains the intraspecifically clustering TLRs, from the remainder of the TLRs. This formed an unresolved node with a posterior probability of 1.000. Within clade C, TLRs for the dipterans (*Anopheles gambiae* and *Drosophila melanogaster*) diverged and clustered independent of one another. Within the lepidopteran cluster, two unique monarch TLRs were identified while three to four unique sequences were observed in the other species within this cluster. They clustered at an unresolved node with three branches which were split into separate groups for moths, butterflies and a combined monophyletic group of both. Genes in the combined group displayed relatively shorter branch lengths than the exclusively moth and butterfly groups. There was a monophyletic group of TLRs 1, 3, 4 5 from *Manduca sexta* within the combined group that shared a node with monarch TLR-3/4 with a probability of 0.9187 (Fig. 4, Table 1).

The remainder of the tree contained the interspecifically clustering TLRs. Clade F contained only TLRs from Lepidoptera while the TLRs from *A. gambiae* and *D. melanogaster* were observed to form independent and indistinct associations to the main node. No genes from monarchs were found to cluster in clade B, which contained TLR-2 for only *H. melpomene, B. mori.* and *M. sexta*. Clade E, which represents TLRs-7, contained two distinct monophyletic groups for lepidopterans and dipterans. This was more representative of a paraphyletic clade since only the genes that had been identified as TLR-7 were included. The dipteran sequences in this paraphyletic clade were in a monophyletic cluster with TLRs 2 and 7 with a relatively low posterior probability (0.5341).

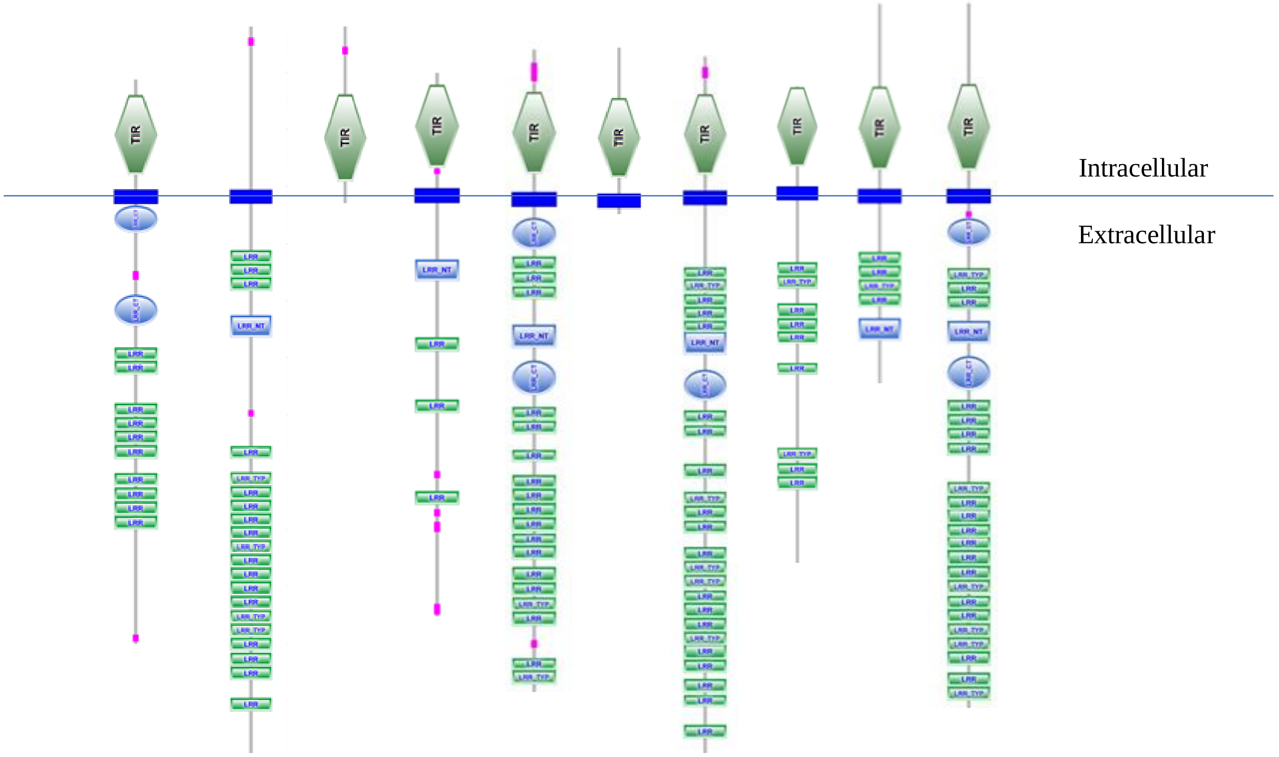


Figure 3. Structure of predicted TLRs from Monarch genome based on BLAST results in Monarchbase. Protein structure and putative domains were obtained from SMART. From Left to Right: DPOGS200002, DPOGS205295-PA, DPOGS205296-PA, DPOGS211472-PA, DPOGS203198-PA, DPOGS205293, DPOGS203200-PA, DPOGS215274-PA, DPOGS205283-PA, DPOGS205279-PA. Abbreviations: Toll-Interleukin receptor (TIR), Leucine Rich Repeat (LRR).

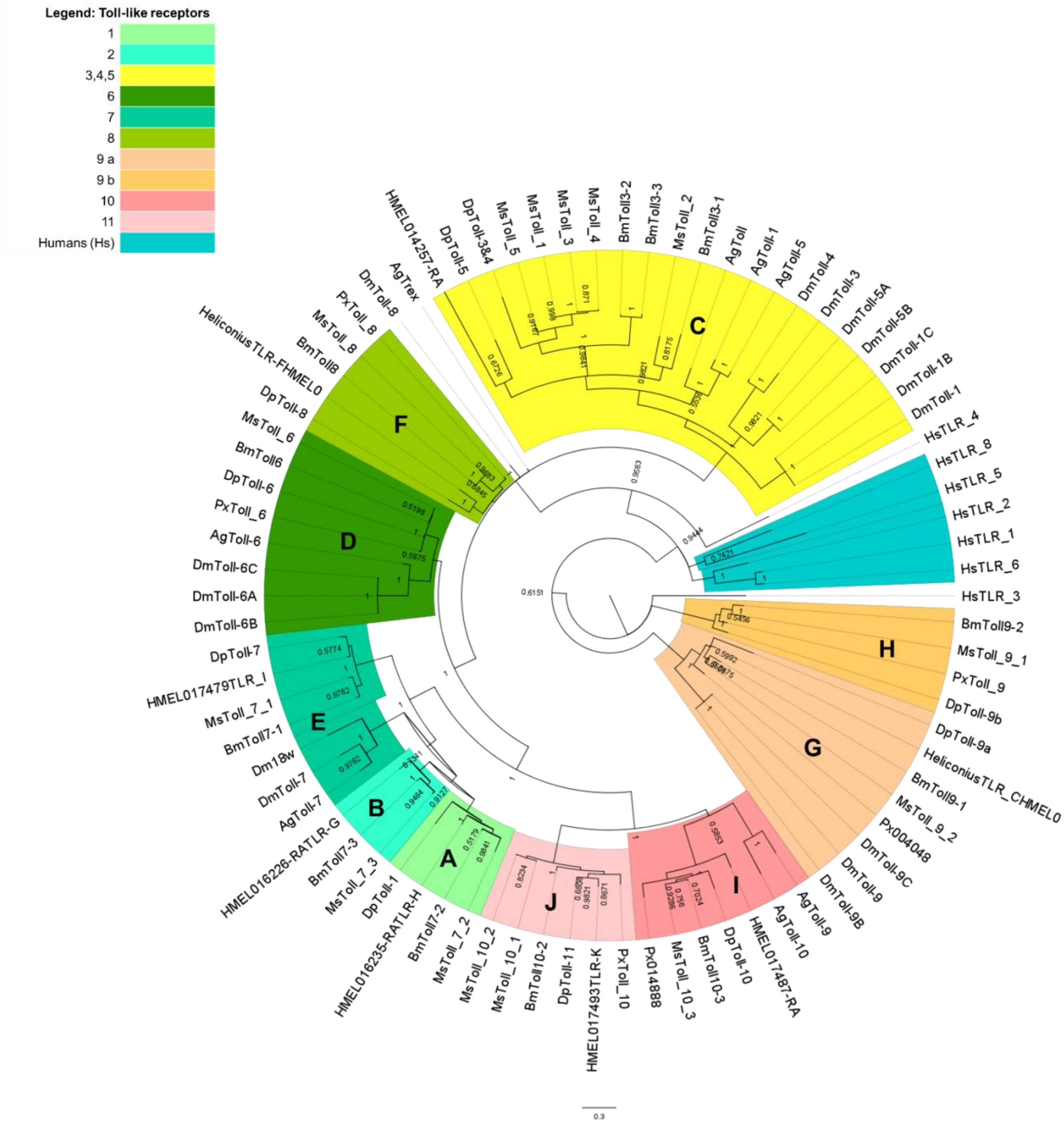


Figure 4. Phylogeny across all species and for TIR domain of all TLRs. TLR genes were obtained from *Bombyx mori* (Bm), *Drosophila melanogaster* (Dm), *Plutella xylostella* (Px), *Anopheles gambiae* (Ag), *Manduca sexta* (Ms), and *Danaus plexippus* (Dp). Phylogenies were created using Bayesian inference in Mr. Bayes with priors set at mixed fixed rate for amino acid models and a million generations were run with a sample frequency of 1000. Clade posterior probabilities are displayed at each node and labels of clusters are indicated in bold within the highlighted portions of the clades as follows: A-TLR1, B-TLR2, C-TLRs 3,4 and 5, D-TLR6, E-TLR7, F-TLR-8, G-TLR9a, H-TLR9b, I-TLR10, J-TLR-11. The tree was visualized in FigTree^{TM} v1.4.3[46].

Table 1. Types of clustering of TLR genes of all species expressed in phylogenetic tree, with posterior probability of clustering out of a million generations

TLR	*Represented species*	*Posterior probability of cluster*
	Interspecific clustering	
1	Lepidopterans	1.000
2	Lepidopterans (except *Danaus plexippus*)	1.000
6	all species	0.5675
7	Lepidopterans	1.000
	Dipterans	1.000
8	all species	0.6845
9	all species	1.000
10	all species (except *Drosophila melanogaster*)	0.5853
11	Lepidopterans	1.000
	Intraspecific clustering	
3,4,5	*Manduca sexta*	0.9980
3,4,5	*Drosophila melanogaster*	0.9821
1,3,5	*Anopheles gambiae*	1.000

The Tajima's D test was employed in investigating signatures of selection to identify deviations of TLR genes (diversifying or balancing selection) from the background genome. A positive Tajima's D value means that more diversity (higher frequency of polymorphisms) was observed than was expected, which could be indicative of either balancing selection or population contraction. A negative Tajima's D reflects the observation of lower diversity than expected which may be indicative of balancing selection or population expansion. In the four geographic populations analyzed, positive mean Tajima's D values were obtained for the Pacific (1.001) and Atlantic (0.348) populations, while the North American (-0.992) and South Florida (-0.908) populations produced negative values. All these populations had normally distributed Tajima's D values of the background genome which differed significantly from a standard mean distribution of the same sample size (n=14209) (mean=0, standard deviation=1) as determined by a Welch's t-test (p < 2e-16). In each population, the mean Tajima's D value for Toll genes was more extreme and in the same direction as the background. A statistical analysis for the significance of these trends could not be performed owing to the large magnitude of difference between the sample sizes (n of background = 14209, n of Toll genes = 23). TLR 11 was identified as an outlier in all four populations, while the majority of TLRs fell within the 10% – 90% coverage of the background genome across populations.

Table 2. Status of predicted TLR sequences in four geographic populations of *Danaus plexippus* relative to background genome based on Tajima's D values within a 10%-90% range of background genome coverage.

Predicted TLR	*Monarchbase*		*Geographic Populations*		
sequences	*ACCESSION ID*	*NA*	*PAC*	*FL*	*ATL*
1	DPOGS205295-PA	●	●	●	●
3/4*	DPOGS211472-PA	●	↓(-1.069018)	↓ (-1.692181)	●
5	DPOGS200002-PA	●	●	●	●
6	DPOGS203198-PA	●	●	●	●
7	DPOGS205293-PA	●	●	●	↑ (1.216113)
8	DPOGS203200-PA	↓ (-1.253691)	●	↓ (-1.473077)	●
9a	DPOGS205123-PA	●	●	●	●
9b	DPOGS215274-PA	●	●	●	●
10	DPOGS205279-PA	↓ (-0.430824)	●	●	↑(1.763864)
11	DPOGS205283-PA	↓ (-2.216873)	↑ (2.551167)	↓ (-2.324272)	↑ (1.00649)
		Mean Tajima's D (background, Toll pathway)			
		(-0.992, -1.406)	(1.001, 1.229)	(-0.908, -1.342)	(0.348, 0.5)

The 'status' of the different TLR sequences in each population refers to whether it falls within (●) the range 10% – 90% coverage of the background (genome). Outliers fall above (↑) or below (↓) the arbitrarily determined range. Values in parentheses indicate the Tajima's D value for that sequence in the corresponding population. Populations represented are North America (NA), Pacific-excluding Hawaii (PAC), South Florida (FL) and Atlantic-excluding Portugal (ATL).

* BLAST search provided the highest match to this sequence for all queries of annotated genes (TLRs 3, 4) obtained from the selected species.

TLRs 1, 3/4, 6, 9a and 9b had Tajima's D values that were within the arbitrarily chosen range of the background genome in all four populations. Genes from the Pacific showed two outliers and the broadest range of the Tajima's D value. TLR 11 was an upper outlier in this population (2.551) while TLR 3/4 had a TJD value of -1.069, placing it in the range of a lower outlier.

DISCUSSION

We sought to identify and categorize TLR sequences in monarchs based on sequence homology with annotated TLRs from selected species of insects which had varying levels of relatedness to monarchs. In our phylogeny of TLRs across the selected species, we observed that some TLRs clustered across species while other groups tended to maintain strong within-specie clustering. The relatively low posterior probability of the clades of interspecifically clustering TLRs (Table 1) may be attributed to within species differences arising from divergence. The high posterior probability and short branch lengths observed in the order-specific clusters within these clades may be due to stronger signatures of selection against mutations, which would otherwise increase divergence. Therefore, there is the possibility that the genes in order-specific clusters are more likely to be orthologs, maintaining form and function from a common ancestor. TLR-6 (Fig 3, clade D) has been shown to be involved in development and nervous system function in *Drosophila melanogaster*[47] and expressed in diapause eggs and during the pupal stage in *Bombyx mori*[36]. Such developmental expressions and functions are often well-conserved owing to their importance in the development of the organism[18]. Therefore, there is likely to be strong selection against deleterious mutations in such genes. Genes involved in immunity and defense are often under more varied signatures of selection since their activity and interactions may be dictated more by environmental cues. TLR 9 (Fig 3, clades G and H) has been shown to be involved in immunity in both *B. mori*[48] and *D. melanogaster* against fungal and Gram-positive bacterial infection. This clade is also the first to diverge from the main tree out of all the TLRs studied. The conservation of immune function in this interspecific cluster could suggest that the immune function of TLR 9 is very important to the Toll pathway and may also share strong similarities to ancestral TLRs that gave rise to the other classes of TLRs. It would be interesting to investigate how developmental and immune functions in the different classes of TLRs evolved, and the evolutionary processes that may be involved.

Next, we analyzed the observed Tajima's D values for the different populations (Table 2) in relation to the functions observed in orthologs of *B. mori* and *D. melanogaster* from previous studies. We hypothesized that there would be a difference in signature of selection between the North American population and the other populations of monarch (Pacific, South Florida and Atlantic), which could be due to the differential distribution of milkweed species between these two groups. We predicted that TLRs that had immune functions would diverge less (neutral selection), either by diversifying (negative Tajima's D values) or balancing selection (positive Tajima's D values), from the background genome in the North American population owing to the relatively low parasite prevalence and toxic cardenolide content of milkweed communities in this region. In a study by Tanaka et al., they identified TLRs 1, 5 and 9 to have orthologs across *D. melanogaster*, *A. gambiae* and *B. mori* that are involved in immunity[49]. Recent studies have also shown that TLR-7 is also involved in immune response against viral infection in *D. melanogaster*. These four TLRs were observed to fall within the 10%-90% range of the distribution of Tajima's D values of the background genome in all four populations (Table 2). It is therefore likely that these TLRs might not differ significantly from the mean of the background genome which is an indication of these genes not undergoing any additional effects besides that which may be observed in population contraction (positive mean TJD value of background) or expansion (negative TJD value of background). Only TLR 7 was identified as an upper outlier in the Atlantic population. In the other TLRs, which are purported to be associated with developmental functions (TLRs 3/4 6, 8, 10 and 11), a distinct trend was not observed in the signatures of selection between the migratory North American population and the other populations. Therefore, our hypothesis of differential signatures of selection based on geographic location was not supported. Sequencing of TLRs from more individuals from each of these populations could help to draw a stronger conclusion. Furthermore, we should consider that there has been shown to be substantial variation in the prevalence of O. *elektroscirrha* across populations. For example, prevalence is suggested to increase from north to south in populations of monarchs within North America. Also, migration distance was found to be inversely correlated with prevalence of O. *elektroscirrha* also in North America[50]. Similarly, there is high heterogeneity in prevalence of O. *elektroscirrha* in Hawaiian populations[51], which are known to be non-migratory. These differences within different population in host-parasite interactions is likely to affect how we analyze signatures of selection and which populations should be categorized together for our future analyses.

Conclusion

In this project we sought to investigate evolution in innate immune genes in signaling pathways, mainly looking at PGRPs in the IMD pathway and TLRs in the innate immune pathway. The presence-absence profile of PGRPs for insects showed that there is the possibility of loss of PGRPs for reasons other than those currently studied, such as endosymbiosis. This is owing to the observance of some independently arising putative absences in insect orders besides the distinct absences studied in hemipterans. In the future, we hope to expand this study to other genes within the IMD pathway.

Concerning TLRs, we observed that there was no distinct trend in the signatures of selection of TLRs between the North American population and the other populations. The TLRs that are purported to be involved in immunity (5, 7 and 9) were mostly found to be in the 10%-90% range of the Tajima's D statistics of the background. Most of the other TLRs that are suggested to be involved or expressed in development were found to have Tajima's D values that often fell outside this arbitrary range as upper (above 90% of the background) or lower (below 10% of the background) outliers in some populations, but with no particular trend between populations. Due to our inability to draw solid conclusions as to the significance of the differences in signatures of selection on TLRS that we observed within and between populations, we hope to sequence more TLR genes within each of the populations and compare these across populations, assuming that the background genome is not significantly variable within a population.

References Cited

1. Marques, J. T. & Carthew, R. W. A call to arms: coevolution of animal viruses and host innate immune responses. Trends Genet. 23, 359–364 (2007).

2. Ausubel, F. M. Are innate immune signaling pathways in plants and animals conserved? Nat. Immunol. 6, 973–979 (2005).

3. Lemaitre, B. et al. A recessive mutation, immune deficiency (imd), defines two distinct control pathways in the Drosophila host defense. Proc. Natl. Acad. Sci. U. S. A. 92, 9465–9469 (1995).

4. Takehana, A. et al. Peptidoglycan recognition protein (PGRP)-LE and PGRP-LC act synergistically in Drosophila immunity. EMBO J. 23, 4690–4700 (2004).

5. Alonso Zumaya-Estrada, F., Barnetche, J., Lavore, A., Rivera Pomar, R. & Rodríguez, M. Comparative genomics analysis of triatomines reveals common first line and inducible immunity-related genes and the absence of IMD canonical components among hemimetabolous arthropods. Parasit. Vectors 11, (2018).

6. Gerardo, N. M. et al. Immunity and other defenses in pea aphids, Acyrthosiphon pisum. Genome Biol. 11, R21 (2010).

7. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. 45, D12–D17 (2017).

8. Poelchau, M. et al. The i5k Workspace@NAL—enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Res. 43, D714–D719 (2015).

9. Choudhuri, S. Chapter 7 - Additional Bioinformatic Analyses Involving Nucleic-Acid Sequences*. in Bioinformatics for Beginners 157–181 (Academic Press, 2014). doi:10.1016/B978-0-12-410471-6.00007-4

10. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212 (2015).

11. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242-245 (2016).

12. Challis, R. J., Kumar, S., Dasmahapatra, K. K. K., Jiggins, C. D. & Blaxter, M. Lepbase: the Lepidopteran genome database. bioRxiv 056994 (2016). doi:10.1101/056994

13. Giraldo-Calderón, G. I. et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res. 43, D707–D713 (2015).

14. Legeai, F. et al. AphidBase: A centralized bioinformatic resource for annotation of the pea aphid genome. Insect Mol. Biol. 19, 5–12 (2010).

15. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. TIG 16, 276–277 (2000).

16. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 46, D493–D496 (2018).

17. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–D285 (2016).

18. Sagri, E. et al. Housekeeping in Tephritid insects: the best gene choice for expression analyses in the medfly and the olive fly. Sci. Rep. 7, (2017).

19. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37 (2011).

20. Truong, A., Sondossi, M. & Clark, J. B. Genetic characterization of Wolbachia from Great Salt Lake brine flies. Symbiosis 72, 95–102 (2017).

21. Teets, N. M., Kawarasaki, Y., Lee, R. E. & Denlinger, D. L. Survival and energetic costs of repeated cold exposure in the Antarctic midge, Belgica antarctica: a comparison between frozen and supercooled larvae. J. Exp. Biol. 214, 806–814 (2011).

22. Schwemmler, W. Ecological significance of endosymbiosis: An overall concept. Acta Biotheor. 22, 113–119 (1973).

23. Ceja-Navarro, J. A. et al. Gut microbiota mediate caffeine detoxification in the primary insect pest of coffee. Nat. Commun. 6, 7618 (2015).

24. Royet, J., Gupta, D. & Dziarski, R. Peptidoglycan recognition proteins: modulators of the microbiome and inflammation. Nat. Rev. Immunol. 11, 837–851 (2011).

25. Troll, J. V. et al. Taming the Symbiont for Coexistence: A Host PGRP Neutralizes a Bacterial Symbiont Toxin. Environ. Microbiol. 12, 2190–2203 (2010).

26. Guan, R. et al. Structural basis for peptidoglycan binding by peptidoglycan recognition proteins. Proc. Natl. Acad. Sci. 101, 17168–17173 (2004).

27. Mesquita, R. D. et al. Genome of Rhodnius prolixus, an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proc. Natl. Acad. Sci. 112, 14936–14941 (2015).

28. Arp, A. P., Hunter, W. B. & Pelz-Stelinski, K. S. Annotation of the Asian Citrus Psyllid Genome Reveals a Reduced Innate Immune System. Front. Physiol. 7, (2016).

29. Altizer, S. & Davis, A. K. Populations of Monarch butterflies with different migratory behaviors show divergence in wing morphology. Evol. Int. J. Org. Evol. 64, 1018–1028 (2010).

30. Zhan, S. et al. The genetics of monarch butterfly migration and warning colouration. Nature 514, 317–321 (2014).

31. Altizer Sonia M., Oberhauser Karen S. & Brower Lincoln P. Associations between host migration and the prevalence of a protozoan parasite in natural populations of adult monarch butterflies. Ecol. Entomol. 25, 125–139 (2001).

32. de Roode, J. C., Pedersen, A. B., Hunter, M. D. & Altizer, S. Host plant species affects virulence in monarch butterfly parasites. J. Anim. Ecol. 77, 120–126 (2008).

33. Fritzsche McKay, A., Ezenwa, V. O. & Altizer, S. Unravelling the Costs of Flight for Immune Defenses in the Migratory Monarch Butterfly. Integr. Comp. Biol. 56, 278–289 (2016).

34. Alavi, Y. et al. The dynamics of interactions between Plasmodium and the mosquito: a study of the infectivity of Plasmodium berghei and Plasmodium gallinaceum, and their transmission by Anopheles stephensi, Anopheles gambiae and Aedes aegypti. Int. J. Parasitol. 33, 933–943 (2003).

35. Cirimotich, C. M., Dong, Y., Garver, L. S., Sim, S. & Dimopoulos, G. Mosquito immune defenses against Plasmodium infection. Dev. Comp. Immunol. 34, 387–395 (2010).

36. Cheng, T.-C. et al. Identification and analysis of Toll-related genes in the domesticated silkworm, Bombyx mori. Dev. Comp. Immunol. 32, 464–475 (2008).

37. Tauszig, S., Jouanguy, E., Hoffmann, J. A. & Imler, J.-L. Toll-related receptors and the control of antimicrobial peptide expression in Drosophila. Proc. Natl. Acad. Sci. 97, 10520–10525 (2000).

38. Hashimoto, C., Hudson, K. L. & Anderson, K. V. The Toll gene of Drosophila, required for dorsal-ventral embryonic polarity, appears to encode a transmembrane protein. Cell 52, 269–279 (1988).

39. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinforma. Chapter 2, Unit 2.3 (2002).

40. Leulier, F. & Lemaitre, B. Toll-like receptors — taking an evolutionary approach. Nat. Rev. Genet. 9, 165–178 (2008).

41. Zhan, S. & Reppert, S. M. MonarchBase: the monarch butterfly genome database. Nucleic Acids Res. 41, D758–D763 (2013).

42. Zhan, S., Merlin, C., Boore, J. L. & Reppert, S. M. The Monarch Butterfly Genome Yields Insights into Long-Distance Migration. Cell 147, 1171–1185 (2011).

43. Pierce, A. A. et al. Serial founder effects and genetic differentiation during worldwide range expansion of monarch butterflies. Proc. R. Soc. Lond. B Biol. Sci. 281, 20142230 (2014).

44. The Heliconius Genome Consortium et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487, 94–98 (2012).

45. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15, 356 (2014).

46. FigTree. Available at: http://tree.bio.ed.ac.uk/software/figtree/. (Accessed: 24th March 2018)

47. McIlroy, G. et al. Toll-6 and Toll-7 function as neurotrophin receptors in the Drosophila central nervous system. Nat. Neurosci. 16, 1248–1256 (2013).

48. Wu, S. et al. BmToll9, an Arthropod conservative Toll, is likely involved in the local gut immune response in the silkworm, Bombyx mori. Dev. Comp. Immunol. 34, 93–96 (2010).

49. Tanaka, H. et al. A genome-wide analysis of genes and gene families involved in innate immunity of Bombyx mori. Insect Biochem. Mol. Biol. 38, 1087–1110 (2008).

50. Flockhart D. T. Tyler et al. Patterns of parasitism in monarch butterflies during the breeding season in eastern North America. Ecol. Entomol. 43, 28–36 (2017).

51. Pierce, A. A., de Roode, J. C., Altizer, S. & Bartel, R. A. Extreme Heterogeneity in Parasitism Despite Low Population Genetic Structure among Monarch Butterflies Inhabiting the Hawaiian Islands. PLoS ONE 9, (2014).