TIAMMAt: Leveraging Biodiversity to Revise Protein Domain Models, Evidence from Innate Immunity

Michael G. Tassia,*[,1] Kyle T. David,[1] James P. Townsend,[2,3] and Kenneth M. Halanych[1]

[1]Department of Biological Sciences, Auburn University, Auburn, AL, USA
[2]Whitman Center, Marine Biological Laboratory, Woods Hole, MA, USA
[3]Department of Biology, Providence College, Providence, RI, USA

*Corresponding author: E-mail: mgt0007@auburn.edu.

Associate editor: Michael Rosenberg

Abstract

Sequence annotation is fundamental for studying the evolution of protein families, particularly when working with nonmodel species. Given the rapid, ever-increasing number of species receiving high-quality genome sequencing, accurate domain modeling that is representative of species diversity is crucial for understanding protein family sequence evolution and their inferred function(s). Here, we describe a bioinformatic tool called Taxon-Informed Adjustment of Markov Model Attributes (TIAMMAt) which revises domain profile hidden Markov models (HMMs) by incorporating homologous domain sequences from underrepresented and nonmodel species. Using innate immunity pathways as a case study, we show that revising profile HMM parameters to directly account for variation in homologs among underrepresented species provides valuable insight into the evolution of protein families. Following adjustment by TIAMMAt, domain profile HMMs exhibit changes in their per-site amino acid state emission probabilities and insertion/deletion probabilities while maintaining the overall structure of the consensus sequence. Our results show that domain revision can heavily impact evolutionary interpretations for some families (i.e., NLR's NACHT domain), whereas impact on other domains (e.g., rel homology domain and interferon regulatory factor domains) is minimal due to high levels of sequence conservation across the sampled phylogenetic depth (i.e., Metazoa). Importantly, TIAMMAt revises target domain models to reflect homologous sequence variation using the taxonomic distribution under consideration by the user. TIAMMAt's flexibility to revise any subset of the Pfam database using a user-defined taxonomic pool will make it a valuable tool for future protein evolution studies, particularly when incorporating (or focusing) on nonmodel species.

Key words: **protein evolution, domain annotation, animal evolution, innate immunity.**

Article

Introduction

Accurate assignment of protein identity is a fundamental component of molecular studies involving nonmodel species. Such studies often begin by tethering an uncharacterized protein's identity to a homolog of known function to infer, for example, residue-specific selective pressures (Buckley and Rast 2012), protein–protein interaction networks (Szklarczyk et al. 2015), or evolutionary divergence (Tassia et al. 2017). Errors in these assessments can be costly. In the field of evolutionary and developmental biology, for example, over- or underestimating the full complement of protein family members in a nonmodel species can compromise the design of genetic reporter constructs (Cavalieri and Spinelli 2014) or CRISPR/Cas9 targets (Connahs et al. 2019). These errors cost researchers time, financial resources, and can negatively impact the accuracy of scientific conclusions.

Comparative molecular studies employing nonmodel species (Buckley and Rast 2015; Brennan and Gilmore 2018) often utilize a common bioinformatic approach when assigning evolutionary affinity and putative function to uncharacterized proteins (Loewenstein et al. 2009). Initially, protein identity is typically labeled using primary sequence similarity, which measures the number of pairwise matches between two sequences. Although similarity metrics aid protein identification (prematurely extrapolated to indicate orthology in some cases; Chen et al. 2007), similarity alone is insufficient to infer function in an evolutionary context (Liu et al. 2018). Given the pitfalls when relying on similarity alone, uncharacterized protein sequences are also placed in a phylogenetic context to verify homology (Tassia et al. 2017) and further annotated with domains—amino acid sequence patterns which can be used to assign function to discrete territories within a full amino acid sequence (Wojcik and Schächter 2001; Zhao et al. 2008). When used in concert, phylogenetic methods and domain annotation can reinforce hypotheses on protein family evolution and their functional variation across deep evolutionary timescales (Buckley and Rast 2012; Costa-Paiva et al. 2017, 2018; Gerdol et al. 2017; Tassia et al. 2017). For example, mammalian inflammatory and apoptotic caspases invariably possess a carboxy-terminal protease effector domain, and paralogs within the family can be categorized by their

amino-terminus CARD or DED domain(s) (Man and Kanneganti 2016). These same rules remain consistent when applied to categorizing caspases in *Hydra*, a freshwater cnidarian (Lasi et al. 2010). Importantly, annotation of an uncharacterized protein with domain structure requires a database of known protein domains.

The Pfam database contains a well-curated and frequently updated catalog of domain models placed in an evolutionary context for protein studies across the tree of life (Sonnhammer et al. 1997; El-Gebali et al. 2019). Each Pfam domain entry is created as follows: 1) a seed alignment is generated from representative sequences containing a conserved pattern that has been characterized in at least one of the sampled species; 2) the seed is then used to build a domain profile hidden Markov model (HMM) using the open source HMMER software package (Eddy 2009); lastly, 3) the new profile HMM is searched against Pfam's proteomic sequence database as quality control and to provide evolutionary context (Sonnhammer et al. 1997; Eddy 2009; El-Gebali et al. 2019; Mistry et al. 2020). Encoding Pfam domains as profile HMMs, in turn, allows protein domain searches to adopt the robust statistical framework underlying HMMs and information entropy (Hernando et al. 2005), along with the benefit that domain profile HMMs are rapidly searchable (Eddy 2009, 2011). Although variation encoded within the model is designed to capture homologs from species outside those represented directly within the seed alignment (El-Gebali et al. 2019), many domain profiles are derived of only a few species, reducing the model's capacity to identify homologous domain sequences in phylogenetically distant taxa. Currently, domain seed alignments are dominated by sequences from a few biomedical model taxa (fig. 1), or closely related taxa, and the trend in sequencing bias toward these model systems is becoming increasingly exacerbated (David et al. 2019).

Using innate immunity proteins as a case study, we show that revising domain profile seed alignments to directly account for underrepresented protein diversity aids homolog identification in nonmodel animal species. Innate immunity signaling relies on pattern recognition receptors (PRRs) which recognize broad categories of microbes (such as RNA viruses or Gram-positive bacteria) by binding specific pathogen-associated moieties (Beutler 2004). Unlike adaptive immunities which evolved independently in both jawed- and jawless vertebrates (Flajnik and Kasahara 2010), PRRs were likely present in the last common ancestor to all animal lineages (Bosch 2013) and some innate immunity protein families have undergone several notable lineage-specific diversifications (Buckley and Rast 2012; Gerdol et al. 2017). Important for the context of our study, PRRs rely on domain–domain interactions for activation and signal transduction (O'Neill and Bowie 2007), possess defined domain architectures (Akira and Takeda 2004; Kowalinski et al. 2011; Lechtenberg et al. 2014), and have dominantly been studied in biomedical model species (Leulier and Lemaitre 2008). Among the most well-described PRRs are NOD-like receptors (NLRs; Lechtenberg et al. 2014), Toll-like receptors (TLRs; Akira and Takeda 2004), and RIG-I-like receptors (RLRs; Kowalinski et al. 2011). Although these three PRR families differ from one another in their domain architectures and signal transduction partners (supplementary fig. 1, Supplementary Material online), all three converge on the activation of nuclear factor κB (NF-κB) and/or interferon regulatory factors (IRFs). These transcription factors promote expression of pro-inflammatory cytokines (e.g., interleukins and tumor-necrosis factors), antimicrobial-, and/or antiviral peptides (Hiscott 2007; Zhang et al. 2017). The current perspective on PRR signaling is intimately tied to domain architecture, emphasizing the importance of protein annotation as a fundamental prerequisite when placing PRRs in a comparative and evolutionary framework.

Here, we show that revising profile seed alignments aids identification of domain homologs in nonmodel species and can provide insight into protein family evolution. The value of phylogenetically representative domain models cannot be overstated as identifying protein homologs across deep evolutionary timescales is a challenge that continues to grow as genomes become more accessible, particularly for those of historically underrepresented species (Buckley and Rast 2012; Costa-Paiva et al. 2017, 2018; Gerdol et al. 2017; Tassia et al. 2017). To this end, we explore the effects of revising domains which are essential for animal innate immunity signaling pathways, a group of evolutionarily ancient protein families within Metazoa that rely on domain–domain interactions and show considerable variation between taxa. Below, we describe a domain revision protocol called Taxon-Informed Adjustment of Markov Model Attributes (TIAMMAt; pronounced "TEE-a-mat" or "TEE-a-maht") and apply it to the domains at the core of PRR signaling to reveal the effects of narrow phylogenetic representation within domain seed alignments on domain homolog detection in nonmodel species.

New Approaches

TIAMMAt provides an automated and reproducible method for revising Pfam domain profile HMMs to capture homologous sequence diversity contingent upon a user-defined taxonomic distribution. TIAMMAt fundamentally relies on HMMER's suite of profile-to-sequence comparison tools and their direct association with Pfam domain database entries. Although TIAMMAt is utilized in the context of metazoan innate immunity for our study, the program can revise any domain profile(s) within Pfam based on a user-defined taxonomic pool. For example, TIAMMAt can be applied to investigate the evolution of the death domain superfamily in all eukaryotes just as it can be used to revise and identify globin domains within arachnids. For each domain revised by TIAMMAt, the program will produce 1) a domain profile HMM which directly accounts for homologous sequence variation within the queried taxon/taxa, and 2) the subset of proteins from each taxon which possess the domain of interest (before and after revision) (fig. 2). Importantly, TIAMMAt is a versatile tool for protein evolution studies that can be catered to the investigator's subject of research.

TIAMMAt executes the following steps for each target domain profile (see Materials and Methods, fig. 2, and supplementary fig. 2 and table 1, Supplementary Material online). First, each supplied proteome (defined here as the whole

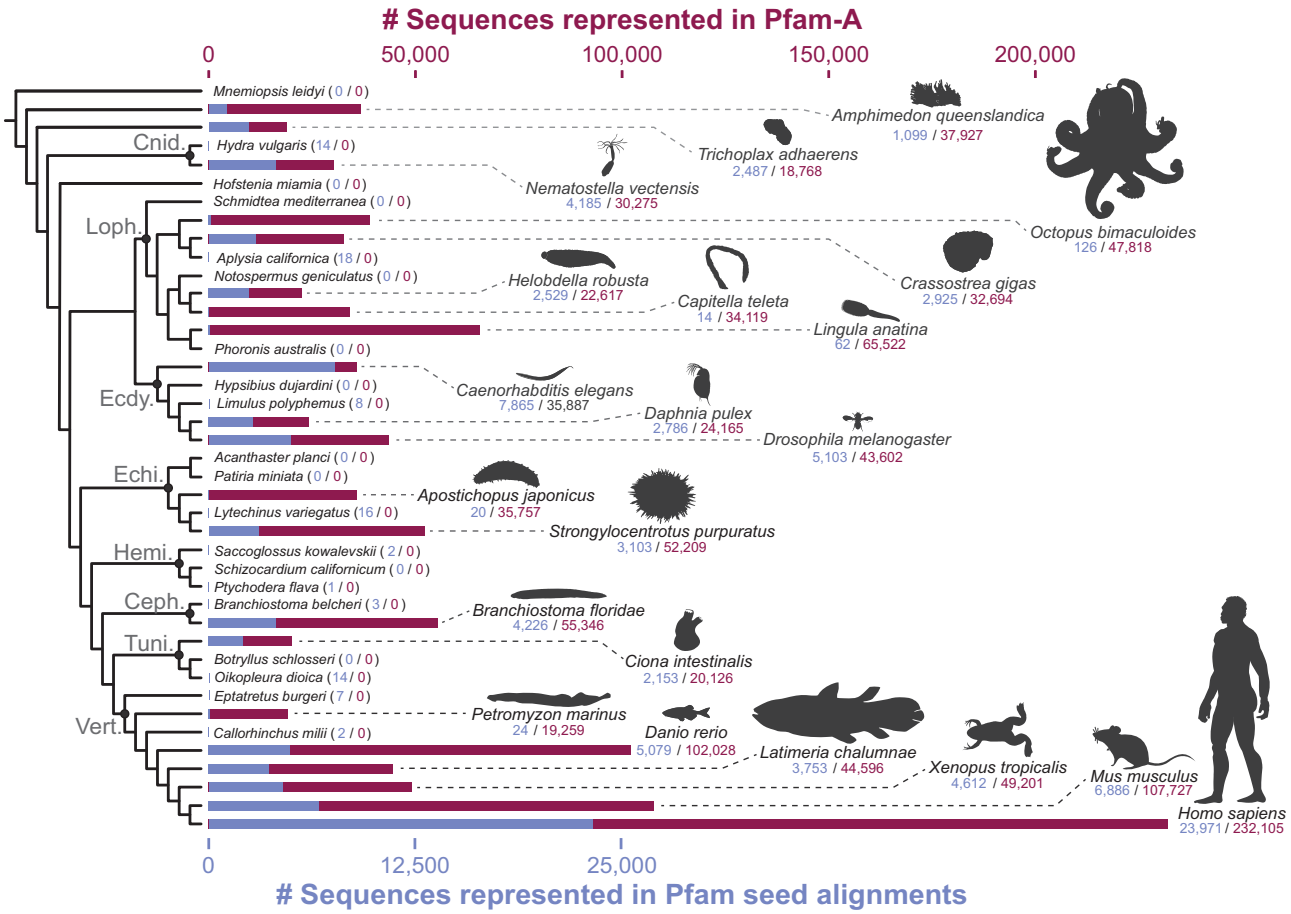

Fig. 1. Taxon representation within Pfam database. In blue (left values following species names) is the total number of occurrences a species appears across all Pfam seed alignments. In red (right values following species names) is the total number of sequences within a species' reference proteome captured by all Pfam domain profiles. Blue and red bars duplicate the numeric values next to species names and the blue bar is superimposed on the red bar, each with its own independent scale displayed at the top (Red) and bottom (Blue) of the plot. Cladogram depicts consensus phylogenetic relationships derived from Laumer et al. (2019). Cnid., Cnidaria; Loph., Lophotrochozoa; Ecdy., Ecdysozoa; Echi., Echinodermata; Hemi., Hemichordata; Ceph., Cephalochordata; Tuni., Tunicata; Vert., Vertebrata.

collection of protein sequences derived of an organism's genome) is searched for occurrences of the target domain where the target domain must meet two conditions: 1) The per-target/per-domain expectation values (e-values) do not exceed HMMER's reporting or inclusion thresholds (default per-domain/-target reporting threshold $\leq$10.0 and per-domain/-target inclusion threshold $\leq$0.01, respectively), and 2) the target domain has the lowest per-domain e-value within the sequence envelope in which it is identified (relative to every other domain in the Pfam database). Conceptually, coordination of these two filtering conditions constrains the domain revision process to not only incorporate amino acid sequences labeled as "true homologs" to the target domain (as defined by HMMER's profile-to-sequence comparison pipeline; Eddy 2009), but also that the target domain has the lowest probability of being a false positive within the sequence envelope it was identified (relative to all other domains). These constraints are specifically designed to avoid false positives which could otherwise be introduced by incorporating similar, but nontarget, motifs into the domain seed alignment (e.g., CARD and DED are similar and related, but exhibit distinct interactive properties; Jiang et al. 2012). Importantly, TIAMMAt can also be run with user-specified per-target/per-domain e-value thresholds. Such flexibility may be useful if, for example, investigators intend to experimentally test a domain's function in a pharmacological context using a protein derived from nonmodel species (Agrawal et al. 2016). In this scenario, increasing the stringency of these thresholds permits investigators to identify sequence structures strictly similar to those captured in the original domain model, while also accounting for evolutionary distance between their subject species and those used to build the original seed alignment.

Following annotation filtering, TIAMMAt extracts all best-hit domain sequences and aligns them to the profile HMM along with the original Pfam seed sequences. This revised seed alignment, which is a direct derivation of the original alignment (and constrained a priori to align to original domain profile HMM), is then reconstructed into a single new revised domain profile HMM. Finally, all proteomes are searched once again for the target domain(s) using all the original and revised models generated during the run (fig. 2). For the purposes of our study, we revised domains integral to the functions of TLRs, RLRs, NLRs, NF-κBs, and

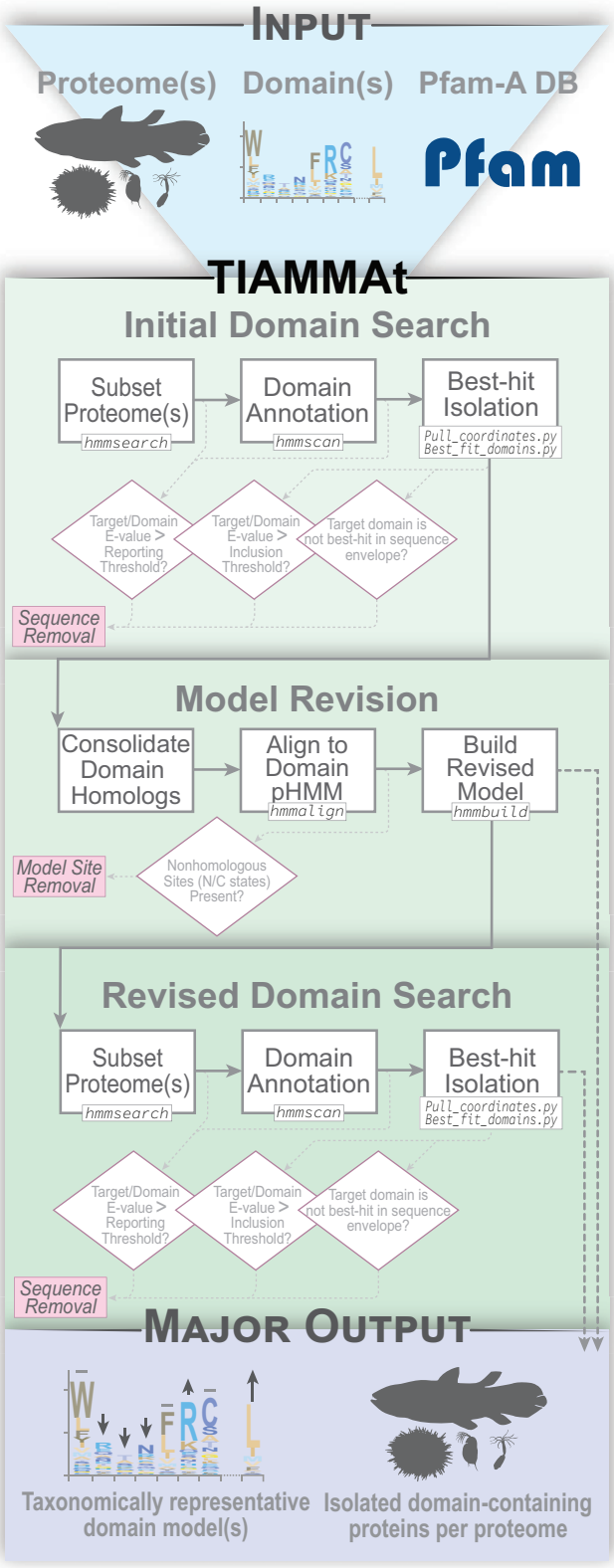

FIG. 2. Bioinformatic operations performed by TIAMMAt. Boxes and solid arrows symbolize major workflow of TIAMMAt. Dashed lines and diamonds show filtering criteria used by TIAMMAt. For further detail, see supplementary figure 2, Supplementary Material online.

IRFs (supplementary tables 1 and 2 and figs. 3 and 4, Supplementary Material online) using 39 publicly available proteomic datasets representative of taxa across the metazoan phylogeny (supplementary table 3, Supplementary Material online), focusing particularly on species which have been labeled within scientific literature as emerging or nonmodel (e.g., Simakov et al. 2013, 2015; Hall et al. 2017; Gehrke et al. 2019).

Results and Discussion

Trends in Model Revision

Given that the model revision process employed by TIAMMAt is dependent upon the original model, the revised model's properties are constrained in a few key ways (fig. 3): 1) Overall sequence structure remains consistent between base and revised models, where low information content regions retain low impact on sequence hit probabilities and high information content regions retain their overall structure and comparatively high statistical weight; 2) the revised model's length is partially constrained to the base model's length by trimming nonhomologous residues near ends of the alignment (via *hmmalign*'s *–trim* flag), avoiding overparameterization which could be produced as the new seed alignment incorporates new sequences; 3) most changes are adjustments in the emission probabilities per residue per site of the domain profile and changes in insertion probabilities, but not changes in overall consensus sequence structure.

For all analyses, we included the human proteome as a positive control for domain revision. All target domain-containing sequences identified before revision in human were also identified after revision. Additionally, the human sequences found to possess a target domain only after revision (supplementary table 2, Supplementary Material online) met one of two conditions: 1) following phylogenetic analysis, the newly identified protein clustered with sequences known to possess the original (prerevision) domain, suggesting model revision produced a profile which describes sequence variation absent in the original model; or 2) the sequence represented a poorly understood protein and the revised domain was assigned to a sequence envelope where no other domain met HMMER's inclusion threshold (i.e., the revised domain fell within an unannotated sequence envelope). Importantly, the magnitude of change for individual amino acid emission probabilities per-site was not equivalent across all model revisions, suggesting TIAMMAt is sensitive to the degree of sequence variation within the input domain(s). In turn, revision of some domains showed limited change. For example, domain revision yielded four and eight additional IRF and NF-κB family members, respectively (supplementary table 2, Supplementary Material online). When all IRF and NF-κB proteins were placed in a phylogenetic context, resolution of deeper nodes was poor (supplementary figs. 5 and 6 and files 1 and 2, Supplementary Material online), and lack of domain architecture diversity among these transcription factor family members further exacerbated the challenge of interpreting novel IRF and NF-κB members in an evolutionary context. In contrast, revision of domains central to NLR, TLR, and RLR signaling pathways produced more dramatic changes that could be interpreted in an evolutionary framework.

TIAMMAt revises domain profile HMMs to capture homologous sequence variation represented within the proteomes provided; as such, taxon selection (and dataset quality) directly influences the evolutionary context for revised domains and their potential use in subsequent searches. Revision of immunity-associated domains using, for instance, a single genus of crustaceans may not produce revised domains appropriate for studies at the scale of Metazoa. However, revising a domain using a single clade of organisms would yield interesting and valuable results if that clade of organisms is already known to possess divergent proteins, particularly for domains directly implicated in protein–protein interactions (as is the case for the TIR and NACHT domains). Under these circumstances, model revision using narrow taxon sampling would facilitate identification of lineage-specific domain structures. Importantly, because each domain profile HMM describes the variation observed for a single Pfam domain, the original and revised models are not mutually exclusive.

Below, we report and discuss the effects of domain model revision on three key innate immunity PRR families: NLRs, TLRs, and RLRs. Direct comparisons between studies investigating PRR diversity across Metazoa can be difficult due to differences in bioinformatics and the definitions used to define each family (e.g., Nehyba et al. 2009; Buckley and Rast 2012; Yuen et al. 2014; Tassia et al. 2017). Nevertheless, the number of domain-containing sequences identified by TIAMMAt prior to domain revision are reported in supplementary tables 4–7, Supplementary Material online, and have been categorized to reflect conservative estimates of the number of PRR family members recognized in the literature (Buckley and Rast 2015; Pugh et al. 2016; Gerdol et al. 2018). Although TIAMMAt's domain filtering conditions are explicitly designed to avoid false positives, phylogenetic methods serve as a valuable framework to support evolutionary relationships between protein sequences identified before and after model revision. In each tree generated during our analyses, all domain-containing proteins exclusively identified after revision fell within orthology groups comprised proteins identified prior to revision (supplementary files 1–4, Supplementary Material online).

NOD-Like Receptors

NACHT domain revision yielded the greatest increase in the number of domain-containing sequences in our analyses (fig. 4 and supplementary tables 4–6, Supplementary Material online). We defined NLRs as proteins possessing both a NACHT domain and a terminal series of LRRs, consistent with literature on the structural perspectives of NLR signaling kinetics and previous NLR surveys (Laroui et al. 2011; Mo et al. 2012; Meunier and Broz 2017). Following NACHT revision, we identified novel NLRs in the sea snail, *Aplysia californica* ($n = 1$ additional), seastars *Acanthaster planci* ($n = 1$) and *Patiria miniata* ($n = 5$), the sea cucumber, *Apostichopus japonicus* ($n = 1$), acorn worms *Ptychodera flava* ($n = 2$) and *Schizocardium californicum* ($n = 1$), and the purple urchin, *Strongylocentrotus purpuratus* ($n = 1$; supplementary table 6, Supplementary Material online). Aside from novel CARD-containing NLRs (i.e., NLRC subfamily) identified in *Ptychodera flava* and *S. californicum*, all other NLRs identified after revision by TIAMMAt could not be classified into the four canonical NLR subfamilies (Kanneganti et al. 2007; Meunier and Broz 2017) based solely on domain architecture

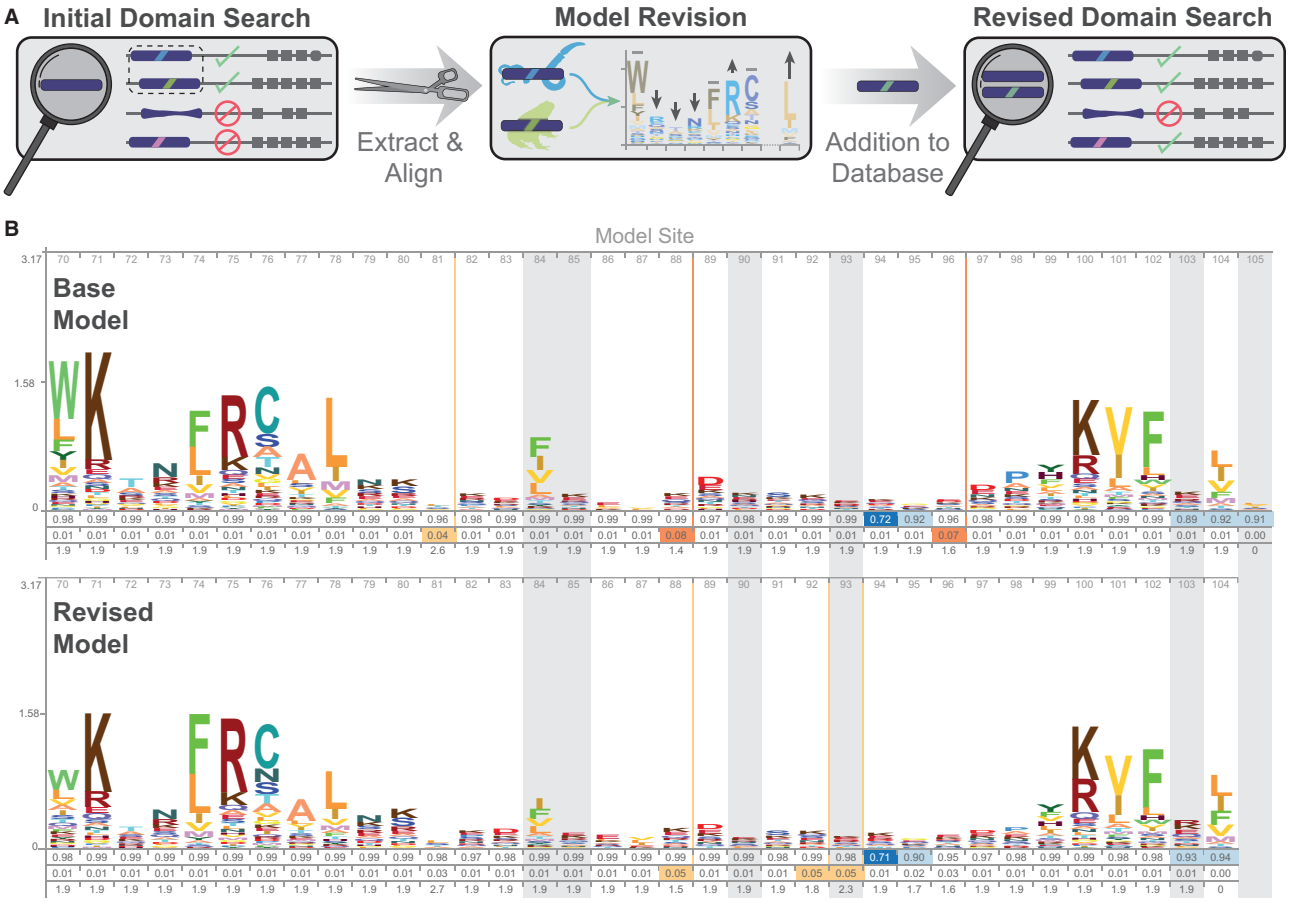

FIG. 3. Domain revision by TIAMMAt. (*A*) Schematic overview of the three major operations performed by TIAMMAt (see Materials and Methods for details). First, target domains are searched for among input proteomes. These domains are extracted and aligned to the associated domain profile HMM. Second, the alignment is recompiled into a revised domain profile HMM. Lastly, revised domains are appended to a local installation of Pfam and used to reannotate all sequences which possess either the base or revised model. (*B*) Visual alignment of IRF domain (PF00605) C-terminus Skylign graphs (Wheeler et al. 2014) showing common parameter adjustments after domain revision, including changes in most probable amino acid state emission per site (gray columns), nonconsensus state trimming (last column), and overall adjustments in information content (bit score) per site (*Y* axis value per site). *X* axes below each diagram are as follows (from top to bottom): occupancy probability, probability of insertion following site, and length of insertion following site. Vertical bars mark sites where the insertion probability >0.01. Relative height of amino acid symbols reflects their emission probability relative to all other amino acid states at that site.

(supplementary fig. 7 and table 6, Supplementary Material online). This result is consistent with previous findings showing NLRs exhibit more variety in their N-terminal domains among invertebrate taxa than within Vertebrata (Lange et al. 2011; Hamada et al. 2013; Yuen et al. 2014). Moreover, PYD-containing NLRs (i.e., NLRP subfamily) appear to be exclusive to euteleosts in our dataset (i.e., *Latimeria*, zebrafish, and human), even after domain revision. Coincidentally, PYD, independent of NACHT, could only be identified in euteleost taxa (data not shown). Unlike the NLRCs which can directly elicit cell-death behaviors through homotypic CARD interactions, NLRPs (which possess an N-terminal PYD in place of a CARD) require ASC as a signaling intermediate (supplementary fig. 1, Supplementary Material online), a short adaptor protein containing both a PYD and CARD (Lamkanfi and Dixit 2012), before signaling for cell-death.

Our evolutionary analysis supports previous studies (Messier-Solek 2010; Hamada et al. 2013; Yuen et al. 2014; Gerdol et al. 2018) which suggest vertebrate-defined NLR subfamilies (i.e., NLRAs, NLRBs, NLRCs, and NLRPs) are insufficient for classifying NLRs outside Vertebrata (supplementary table 6, fig. 8, and file 3, Supplementary Material online). Noncanonical NLRs identified in our study include a collection N-terminal death domain (or juxtaposed death and CARD domains) NLRs present in cephalochordates (*Branchiostoma belcheri* and *B. floridae*) and echinoderms (*Acanthaster planci*, *P. miniata*, *Apostichopus japonicus*, *Strongylocentrotus purpuratus*, *Lytechinus variegatus*) (supplementary table 6, Supplementary Material online). Assuming the overall domain structures of metazoan NLRs retain their functional regionalization (i.e., C-terminal LRRs operate as ligand-binding, NACHT domains promote oligomerization, and the N-terminal domains are responsible for protein–protein interaction and signal transduction), the presence of noncanonical death domain superfamily members among NLRs may indicate a degree of evolutionary flexibility connecting pathogen recognition to the various death domain superfamily-associated signaling effects such as inflammation, apoptosis, cytokine/chemokine expression, and transcriptional regulation (Park et al. 2007; Kwon et al. 2012).

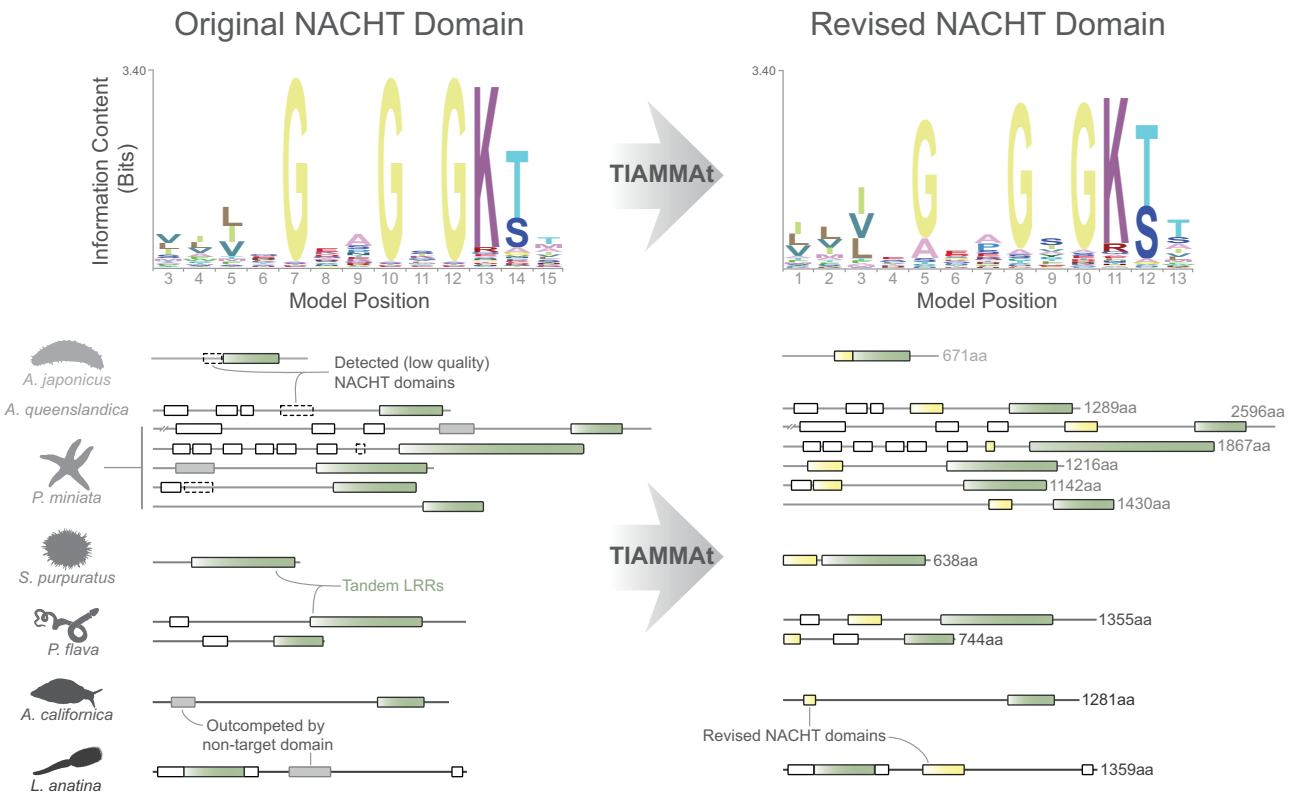

FIG. 4. NACHT domain model revision identifies previously unrecognized NLRs. Top: Skylign (Wheeler et al. 2014) graph for positions 3–15 of the original NACHT domain model (PF05729; left) and positions 1–13 of the revised domain (right). Relative height of amino acid symbols reflects their emission probability relative to all other amino acid states at that site. Bottom: Domain diagrams of NLR structures before (left) and after (right) domain revision, highlighting the utility of incorporating taxonomic diversity into the NACHT domain seed alignment when working with underrepresented taxa. Regarding domain diagrams, green bars represent tandem leucine-rich repeats, yellow bars represent NACHT domains, dotted-outline empty bars represent NACHT domains below inclusion threshold, gray bars represent domains which outcompete NACHT for best-fit domain within the sequence envelope, and white bars represent other NLR accessory domains.

Outside of the death domain superfamily, NLRs identified in nine invertebrate taxa possess a higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domain at their N-terminus. In a previous survey of HEPN domain sequence evolution across the tree of life (Anantharaman et al. 2013), HEPN proteins were predicted to act as either RNA sensors or catabolic RNases associated with RNA-dependent host-defense and stress response. Although we can loosely predict HEPN-NLRs may function as a cytoplasmic sensor for some category of RNAs, broader taxon sampling among underrepresented animal phyla and targeted molecular studies will be required to validate these proteins' hypothetical role in immunity.

The N-terminal domain of NLRs is far more diverse than what has traditionally been represented within vertebrates. The noncanonical NLRs identified in this study represent an underappreciated subset of the NLR protein family, perhaps indicative of more diverse functional roles for the family over the course of animal evolution. Moreover, because the search protocol employed by TIAMMAt isolates all proteins containing a target domain (which meet TIAMMAt's statistical prerequisites), several NACHT domain-containing proteins with undocumented affinity for NLR signaling pathways were identified before and after revision (supplementary fig. 7 and table 6, Supplementary Material online). Given their role in facilitating protein–protein interactions between two or more NACHT-containing proteins (Lamkanfi and Dixit 2012), these unclassified NACHT domain-containing proteins warrant further investigation for their potential role in NLR signaling regulation across Metazoa.

Toll-Like Receptors

Following TIR domain revision (PF01582 and PF13676), additional Toll-like receptor (TLR) proteins were identified in the tunicates *Ciona intestinalis* ($n = 2$ additional) and *Botryllus schlosseri* ($n = 1$), the stalked brachiopod, *Lingula anatina* ($n = 1$), and the lancelet chordate, *B. belcheri* ($n = 1$) (fig. 5). Whereas novel TLRs identified in *L. anatina* and *B. belcheri* occur in a background of >20 and >40 TLRs, respectively (Huang et al. 2008; Halanych and Kocot 2014; Gerdol et al. 2018), proteins detected in tunicates after revision are proportionally more substantial, doubling the number of reported TLRs in *Botryllus schlosseri* from 1 to 2 (Tassia et al. 2017; Franchi et al. 2019) and in *C. intestinalis* from 3 to 5 (Buckley and Rast 2015; Tassia et al. 2017). For all novel TLRs identified, the revised TIR domain was exclusively predicted in previously unannotated space downstream of tandem LRR cassettes, not within a territory where it

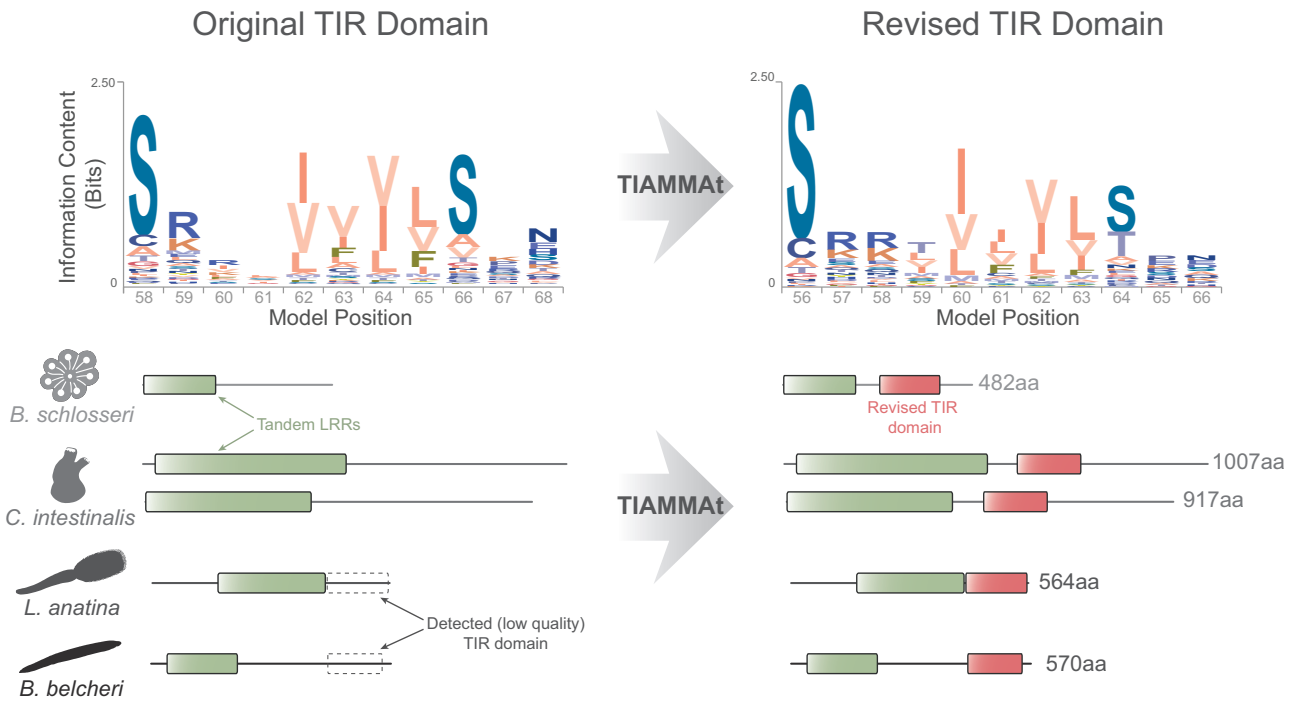

FIG. 5. TIR domain model revision identifies previously unrecognized TLRs. Top: Skylign (Wheeler et al. 2014) graph for positions 58–68 of the original TIR domain model (PF01582; left) and after revision (right). Relative height of amino acid symbols reflects their emission probability relative to all other amino acid states at that site. Bottom: Domain diagrams of TLR structures before (left) and after (right) domain revision, highlighting the utility of incorporating taxonomic diversity into the TIR domain seed alignment when working with underrepresented taxa. Regarding domain diagrams, green bars represent tandem leucine-rich repeats, red bars represent TIR domains, dotted-outline empty bars represent TIR domains below inclusion threshold.

statistically outcompeted another high confidence, but unrelated, domain annotation. Thus, TIAMMAt's results yielded a domain architecture fitting the canonical schema for TLRs (Akira and Takeda 2004). A TIR domain was omitted in the original annotations of these TLRs for two different reasons. For *Lingula*'s and *Branchiostoma*'s novel TLRs, a TIR domain met HMMER's default reporting threshold prior to revision (per-domain and per-sequence e-values <10.0). However, the domain did not meet the inclusion threshold requirement (per-sequence e-value <0.01) to confidently be labeled as a statistically significant homolog. In contrast, the novel tunicate TLRs lacked any reportable TIR domain prior to revision (fig. 5), suggesting the newly identified tunicate proteins contain divergent TIR domains relative to sequences in the original seed alignment. Prior analyses have shown both of *Ciona*'s previously described TLRs act as a functional blend of several vertebrate homologs (Sasaki et al. 2009; Satake and Sekiguchi 2012). Notably, the divergent tunicate TLR may be causally tied to tunicate's rapid rate of molecular evolution relative to their sister phylum, Vertebrata (Berná and Alvarez-Valin 2014).

TIR domain revision also supported previous data (Gerdol et al. 2017) suggesting TIR-domain-containing (TIR-DC) proteins have experienced a high degree of evolutionary change across Metazoa. Several TIR-DC families possess notable taxonomic distributions with implications for TLR pathway evolution (supplementary table 4 and fig. 7, Supplementary Material online). Stimulator of interferon genes (STING), an evolutionarily ancient facilitator of innate immunity responses against exogenous RNA and dsDNA (Wu et al. 2014), was reported to uniquely possess a TIR domain in several lophotrochozoan lineages (Gerdol et al. 2017), implicating an intersection between TLR- and STING-facilitated immunity. Our results corroborate these findings, reporting an additional TIR-DC STING protein in the nemertean, *Notospermus geniculatus*, and two more copies in the oyster, *Crassostrea virginica*, following TIR domain revision (supplementary table 4, Supplementary Material online). Furthermore, whereas homologs to MYD88 and SARM1 (canonical TIR-DC adaptor proteins responsible for signal transduction and regulation of TLRs, respectively; O'Neill and Bowie 2007) possess ancestry predating the emergence of Vertebrata (Tassia et al. 2017; Toshchakov and Neuwald 2020), many evolutionarily conserved TIR-DC proteins (defined in Gerdol et al. [2017]) identified here lack any homologs within Vertebrata (supplementary table 4, Supplementary Material online). Even when including proteomes from non-mammalian vertebrate lineages (i.e., hagfish, lamprey, and *Latimeria*) when running TIAMMAt, vertebrate TIR-DC proteins appear to be restricted to TLRs, IL-1Rs, and the five traditional TLR adaptors (O'Neill and Bowie 2007). Although there may be some relationship between the emergence of adaptive immunity and the limited number of TIR-DC protein structures within vertebrates, the noncanonical TIR-DC proteins identified across metazoan taxa may also represent a more flexible role for TIR domains outside the confines of the TLR pathway.

RIG-I-Like Receptors

Revision of the RLR C-terminal domain (CTD), which is unique to three canonical RLR family members (retinoic acid-inducible gene, RIG-I; melanoma differentiation antigen 5, MDA5; and laboratory of genetic and physiology 2, LGP2; supplementary fig. 1, Supplementary Material online; Esser-Nobis et al. 2020), revealed novel RLR proteins in the cnidarian, *Hydra vulgaris* ($n = 1$ additional), and the sea star, *P. miniata* ($n = 2$; supplementary table 5, Supplementary Material online). Unlike canonical RLRs, novel proteins identified in *Hydra vulgaris* and *P. miniata* have atypical and individually distinct domain organizations. The novel protein identified in *Hydra* has a reversed architecture (with an N-terminal RLR "C-terminal domain"), an incomplete central helicase, and lacks CARD domains, similar in structure to the vertebrate LGP2 protein. In contrast, *Patiria*'s novel proteins both possess appropriately positioned C-terminal CTDs. However, one of the two newly identified *Patiria* RLRs lacks a central helicase, the second possesses a duplicated CTD, and both possess a single N-terminal death effector domain (DED). Moreover, the novel domain architectures described above are not unique to the post-domain-revision dataset as several noncanonical RLR-related domain architectures (defined by the presence of the RLR-specific C-terminal domain) were detected across Metazoa even before domain revision. For example, *Hydra* possesses a second reversed RLR protein and *Hofstenia miamia* (a member of the clade, Xenacoelamorpha) possesses two reverse RLR proteins which, together with *Hydra*'s proteins, comprise a well-supported monophyletic orthology group (>90% posterior probability; supplementary fig. 9 and file 4, Supplementary Material online). Given that all canonical RLRs (i.e., RIG-I, MDA5, and LGP2) share a central DExD/H-box helicase and a CTD, which together give RLR's their RNA recognition capacities (Pippig et al. 2009; Jiang et al. 2011; Luo et al. 2011; Reikine et al. 2014), the proteins with incomplete helicases described in this paragraph provide an interesting opportunity to investigate the function of the RLR CTD independent of a proximal helicase.

We placed all RLRs identified in our study into a Bayesian phylogenetic framework to compare with previous phylogenetic hypotheses on RLR evolution and to expand RLR sampling to include the less conventional RLR family members described above (supplementary fig. 9 and file 4, Supplementary Material online). Concordant with previous studies (Mukherjee et al. 2014; Pugh et al. 2016), we resolve RIG-I and MDA5/LGP2 orthology groups with deep representation of deuterostome taxa, except tunicates which possess their own RLR orthogroup. Interestingly, an orthology group comprised of RLRs with N-terminal DED domains (including the two novel *Patiria* sequences described above) was recovered with maximal support. DED, like CARD, is a member of the death domain superfamily (Park et al. 2007). Independent of RLR signaling, DED operates through homotypic domain–domain interactions and is vital for the regulation of cell death, including interactions mediated by caspase-8 and -10 (Valmiki and Ramos 2009; Riley et al. 2015; Man and Kanneganti 2016). Although they belong to the same superfamily, functional evidence has shown the CARDs of RLRs and the DED of caspase-8 are not functionally equivalent (Jiang et al. 2012), suggesting DED-containing RLRs present among invertebrates may function independently of the canonical RLR signaling pathway. Given the ancient origins of cell death regulation through DED–DED interactions among animals (Sakamaki et al. 2015; Man and Kanneganti 2016), the ubiquitous threat of viral infection (Forterre 2006), and the potential coupling of DED-dependent signaling to the dsRNA recognition via RLRs containing an N-terminal DED, we hypothesize that RLRs possess additional family members among invertebrates which act through rapid DED-dependent apoptotic pathways.

Future Prospects of TIAMMAt

In our application of TIAMMAt on innate immunity protein families, we demonstrated the value of improving representation of nonmodel species in Pfam domain seed alignments. Strikingly, for each of the PRR signaling families we considered, protein domain architecture diversity appears to be underestimated across Metazoa even independent of domain revision (supplementary fig. 7, Supplementary Material online), and the effect becomes more severe when directly accounting for homologous sequence variation in domains among nonmodel species. These findings are consistent with previous studies that highlight the value of leveraging underrepresented species to investigate protein family evolution (Zhang et al. 2012; Yuen et al. 2014; Gerdol et al. 2017).

The design of TIAMMAt was stimulated through a combination of uncovering a lack of even phylogenetic representation within domain profile seed alignments (fig. 1) and the inferential value of annotating uncharacterized proteins with Pfam domains (e.g., Hibino et al. 2006; Costa-Paiva et al. 2017; Gerdol et al. 2017; Tassia et al. 2017). As such, TIAMMAt is designed to be compatible with any taxonomic distribution and collection of Pfam domains of interest to the user. Given enough computational and proteomic resources, TIAMMAt could be applied, for example, to domain/protein evolution studies among all opisthokonts, eukaryotes, or even all organisms. In contrast, TIAMMAt can also be used to revise a domain based on a single genus, yielding a revised domain profile where the per-site amino acid emission probabilities are narrowed to be a strict representation of that genus' domain sequence variance. For example, where one study could revise all globin domains using a broadly sampled metazoan dataset to help identify and investigate oxygen transport protein evolution, another study could revise the same domains using only hydrothermal vent and/or cold seep species where oxygen-transport has become highly specialized (Hourdez and Lallier 2007). Though these two scenarios may focus on the same Pfam domains, the revised domains produced by TIAMMAt reflect different taxonomic assemblages and address two discrete evolutionary questions.

TIAMMAt can be flexibly inserted into bioinformatic pipelines where domain annotation plays a role in inferring function of uncharacterized proteins. As nonmodel and underrepresented species continue to be sequenced at an accelerated rate (David et al. 2019), TIAMMAt provides bioinformaticians the option to account for phylogenetic

distance during protein domain annotation—being particularly valuable when data are generated with the direct intent of addressing protein family/pathway evolution. In studies which concentrate on protein families instead of newly sequenced species, TIAMMAt can be used in a fashion that is similar to our case study on innate immunity, shifting focus from the traditional biomedical model species to a broader comparative scope by leveraging biodiversity already available in public data repositories.

Materials and Methods

Input Dataset Acquisition

Protein sequence accessions for the 39 metazoan taxa used in this study are available in supplementary table 3, Supplementary Material online. Species were chosen to represent a broad phylogenetic distribution across Metazoa with compensation for representation bias within the Pfam database (fig. 1). Regarding the two species where protein sequence datasets were not directly downloadable at the time of acquisition (i.e., *Hofstenia miamia* and *Schmidtea mediterranea*), scaffolded genomes and accompanying protein models were used to generate a protein sequence dataset using *gffread* (https://github.com/gpertea/gffread). In the context of our study, we do not discriminate between protein sequences derived of direct protein sequencing (reviewed in Callahan et al. 2020) and those inferred through bioinformatic translation of nucleotide datasets. Similarly, we recognize each species' proteome is not reflective of the same degree of sequencing revision or protein annotation (David et al. 2019). As a result, proteomes belonging to deeply sequenced species, such as humans, encode a high number of isoforms per protein when compared with more enigmatic taxa (Uhlén et al. 2015). To compensate for uneven annotations across taxa, we make the assumption that all protein isoform predictions possess equal probability to be expressed and are functional. Importantly, because we employ proteomic datasets derived primarily of genome sequencing projects, assessments made in our study are at the level of unique protein species encoded within the genome (accounting for all modeled isoforms of a single gene), not the measure of genes present.

The domain profile HMMs and seeds associated with key innate immunity proteins are summarized in supplementary table 1, Supplementary Material online. Particularly, we chose domains traditionally associated with TLRs (i.e., TIR and TIR_2 domains; Tassia et al. 2017), RLRs (i.e., RIG-I_C-RD and CARD domains; Liu et al. 2016), NLRs (i.e., NACHT and CARD domains; Elinav et al. 2011), IRFs (i.e., IRF and IRF3 domains; Nehyba et al. 2009), and NFkB (i.e., RHD domain; Hayden and Ghosh 2011). All domain models and their seeds were obtained from Pfam version 32.0 (El-Gebali et al. 2019). Phylogenetic breadth represented within seed alignments before and after revision is shown in supplementary figures 3 and 4, Supplementary Material online. Additional LRR annotation was supplemented with Interproscan's (version 5.26-65.0) Gene3D annotation (version 4.1.0; Lees et al. 2014) due to HMMER's difficulty for positively annotating boundaries between individual repeat cassettes (Pellegrini 2015; Mistry et al. 2020).

Database Bias

Pfam domain profile seed alignments were downloaded from the Pfam 32.0 FTP server on April 7, 2020. The Pfam-A database, which is generated from HMMs constructed from the seed alignments, was also downloaded. Species codes were then extracted and aggregated from both Pfam-A and the seed alignments to get a count estimate of species representation in the seeds themselves, as well as how those seeds may contribute to representation (or lack thereof) in the full database (fig. 1).

Domain Profile HMM Revision

TIAMMAt automates revision of Pfam domain models to capture homologous sequence diversity based upon taxonomic distribution provided by the user (fig. 2 and supplementary fig. 2, Supplementary Material online). The program is written using open-source software packages and is publicly available via GitHub (https://github.com/mtassia/TIAMMAt.git). Looping through the individual domain profile HMMs compiled above, TIAMMAt begins by searching proteomes for a single domain signature using HMMER's *hmmsearch* (version 3.1b2; Eddy 2009) under either default reporting/inclusion thresholds (used in this study) or user-defined thresholds (supplementary table 7, Supplementary Material online). For each target sequence reporting a hit to the target domain, the target sequence is isolated from its parent proteome and scanned for all Pfam domains using *hmmscan*, again with default thresholds (used here) or user-defined values. TIAMMAt then parses *hmmscan*'s domain table output to identify the best-fit domain architecture per sequence. Specifically, TIAMMAt first omits any hits which do not meet the per-sequence and per-domain (both conditional and independent) E-value inclusion thresholds of 0.01 (or the value specified by the user). The remaining hits are then ranked in ascending order of per-domain conditional E-values (with a lower bound of zero) and filtered of overlapping annotations, always maintaining the better-scoring domain hit over an overlapping weaker-scoring hit. This annotation parsing schema produces a nonoverlapping list of highest-confidence domains per sequence which must at least meet the per-domain E-value inclusion threshold. Notably, some sequences which reported a potential hit to the target domain during the *hmmsearch* step may not report the same target domain after filtering due to conditional statistics after including all other domains in the Pfam database. Such annotations are considered noise from the perspective of the program and are omitted from the following steps due to lack of statistical substantiation, avoiding incorporation of false positives during revision.

Following domain annotation and identification of sequences with a best-fitting target domain, TIAMMAt extracts all best-fitting domain targets from their parent sequences (e.g., all TIR domains found within the *Saccoglossus kowalevskii* proteome). All isolated domains and the domain's seed sequences are aligned to the relevant

domain profile HMM using *hmmalign* with the optional –*trim* argument to trim nonhomologous residues—particularly those which may accumulate at the termini of the model. Next, TIAMMAt runs *hmmbuild* to generate a revised domain profile HMM from *hmmalign*'s output Stockholm alignment. After the domain model has been revised, TIAMMAt loops once more through *hmmsearch* and *hmmscan*, this time isolating sequences which possess either the base or revised domain model hits (which meet the same threshold requirements specified above).

Once all domains have been revised, TIAMMAt executes a final *hmmscan* using a Pfam database appended with all revised domain models from the current TIAMMAt run—permitting each sequence to be annotated with base or revised domains of all those considered which, until this point, had all been revised in isolation of one another. This step is particularly important if the domains being revised are, in combination, descriptive of a single protein family (e.g., NACHT and CARD domain revisions as they relate to NOD-like receptors). Post-revision datasets from our immunity study (both sequences and Markov models) are also available via the TIAMMAt github repository (https://github.com/mtassia/TIAMMAt.git).

We recommend using TIAMMAt only after careful consideration of input proteome dataset quality and completeness, such as using protein datasets derived of published genomes where such effects have been considered and explicitly controlled or performing genome quality assessments like BUSCO (Waterhouse et al. 2018) or BlobTools (Laetsch and Blaxter 2017).

Phylogenetic Methods

Each protein family was aligned using MAFFT version 7's L-INS-I protocol (Katoh and Standley 2013). Phylogenetic reconstruction was performed using IQ-TREE version 1.6.12 (Nguyen et al. 2015). We employ IQ-TREE's ModelFinder subprogram (Kalyaanamoorthy et al. 2017) to infer best-fit substitution models and the ultrafast bootstrap approximation method for node support (10,000 generations; Minh et al. 2013). Phylogenetic trees were initially visualized using the iTOL web server (Letunic and Bork 2019) and all nodes with ultrafast bootstrap support less than 95% are collapsed and considered unsupported per IQ-TREE's statistical guidelines. Anchoring sequences were downloaded from the UniProt SwissProt database (The UniProt Consortium 2019) in Fall 2020.

Bayesian phylogenetic reconstruction of RLR protein relationships was performed using ExaBayes version 1.5 (Aberer et al. 2014). Two independent runs of four Metropolis-coupled chains each were executed in parallel for 1×10^7 generations, sampled every 100 generations, using a γ-distributed rate heterogeneity, empirical amino acid state frequencies, and a fixed substitution model of VT, which was determined to be the best-fit amino acid substitution matrix via BIC by ModelFinder (Kalyaanamoorthy et al. 2017). Chain convergence was confirmed by the presence of average standard deviation of split frequencies <0.01 and effective sample size per parameter ≥100. A majority-rule consensus tree was generated after discarding the first 25% of sampled Markov Chain Monte Carlo (MCMC) generations as burn-in and visualized using the iTOL web server (Letunic and Bork 2019). Unedited tree files for both likelihood and Bayesian phylogenetic inferences from this study are available via the TIAMMAt github repository (https://github.com/mtassia/TIAMMAt.git).

Supplementary Material

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

Acknowledgments

This work was supported by The National Science Foundation (Grant No. IOS—1755377 to K.M.H., Rita Graze, and Elizabeth Hiltbold Schwartz), and K.T.D. was supported by The National Science Foundation's Graduate Research Fellowship Program. Thank you to Elizabeth H. Schwartz, Rita Graze, Ryan Range, Jamie Oaks, and Nathan Whelan for discussing topics on individual protein families and code development. We would also like to thank all testers of TIAMMAt during development and helpful feedback from reviewers and editors. Computational resources were made available by the Alabama Supercomputer Authority and the Auburn University Hopper Cluster. This publication is Molette Biology Laboratory contribution number 108 and Auburn University Marine Biology Program contribution number 206.

Data Availability

TIAMMAt and results of this manuscript are publicly available and maintained through the TIAMMAt GitHub repository (https://github.com/mtassia/TIAMMAt.git). Regarding data analyzed in our study, taxonomic accessions and their source repositories are available via supplementary table 3, Supplementary Material online, and the Pfam models are available via supplementary table 1, Supplementary Material online.

References

Aberer AJ, Kobert K, Stamatakis A. 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. *Mol Biol Evol*. 31(10):2553–2556.

Agrawal S, Adholeya A, Deshmukh SK. 2016. The Pharmacological Potential of Non-ribosomal Peptides from Marine Sponge and Tunicates. *Front Pharmacol*. 7:333

Akira S, Takeda K. 2004. Toll-like receptor signaling. *Nat Rev Immunol*. 4(7):499–511.

Anantharaman V, Makarova KS, Burroughs AM, Koonin EV, Aravind L. 2013. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis, and RNA processing. *Biol Direct*. 8:15.

Berná L, Alvarez-Valin F. 2014. Evolutionary genomics of fast evolving tunicates. *Genome Biol Evol*. 6(7):1724–1738.

Beutler B. 2004. Innate immunity: an overview. *Mol Immunol*. 40(12):845–859.

Bosch TCG. 2013. Cnidarian-microbe interactions and the origin of innate immunity in metazoans. *Annu Rev Microbiol*. 67:499–518.

Brennan JJ, Gilmore TD. 2018. Evolutionary origins of Toll-like receptor signaling. *Mol Biol Evol*. 35(7):1576–1587.
Buckley KM, Rast JP. 2012. Dynamic evolution of toll-like receptor multigene families in echinoderms. *Front Immunol*. 3:136.
Buckley KM, Rast JP. 2015. Diversity of animal immune receptors and the origins of recognition complexity in the deuterostomes. *Dev Comp Immunol*. 49(1):179–189.
Callahan N, Tullman J, Kelman Z, Marino M. 2020. Strategies for development of a next-generation protein sequencing platform. *Trends Biochem Sci*. 45(1):76–89.
Cavalieri V, Spinelli G. 2014. Early asymmetric cues triggering the dorsal/ventral gene regulatory network of the sea urchin embryo. *Elife* 3:e04664.
Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2(4):e383.
Connahs H, Tlili S, van Creij J, Loo TYJ, Banerjee TD, Saunders TE, Monteiro A. 2019. Activation of butterfly eyespots by Distal-less is consistent with a reaction-diffusion process. *Development* 146(9):dev169367. doi:10.1242/dev.169367.
Costa-Paiva EM, Schrago CG, Coates CJ, Halanych KM. 2018. Discovery of novel hemocyanin genes in metazoans. *Biol Bull*. 235(3):134–151.
Costa-Paiva EM, Schrago CG, Halanych KM. 2017. Broad phylogenetic occurrence of oxygen-binding hemerythrins in bilaterians. *Genome Biol Evol*. 9(10):2580–2591.
David KT, Wilson AE, Halanych KM. 2019. Sequencing disparity in the genomic era. *Mol Biol Evol*. 36(8):1624–1627.
Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform*. 23(1):205–211.
Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Comput Biol*. 7(10):e1002195.
El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res*. 47(D1):D427–D432.
Elinav E, Strowig T, Henao-Mejia J, Flavell RA. 2011. Regulation of the antimicrobial responses by NLR proteins. *Immunity* 34(5):665–679.
Esser-Nobis K, Hatfield LD, Gale M. 2020. Spatiotemporal dynamics of innate immune signaling via RIG-I-like receptors. *Proc Natl Acad Sci U S A*. 117(27):15778–15788.
Flajnik MF, Kasahara M. 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat Rev Genet*. 11(1):47–59.
Forterre P. 2006. The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res*. 117(1):5–16.
Franchi N, Ballarin L, Peronato A, Cima F, Grimaldi A, Girardello R, de Eguileor M. 2019. Functional amyloidogenesis in immunocytes from the colonial ascidian *Botryllus schlosseri*: evolutionary perspective. *Dev Comp Immunol*. 90:108–120.
Gehrke AR, Neverett E, Luo Y, Brandt A, Ricci L, Hulett RE, Gompers A, Ruby RG, Rokhsar DS, Reddien PW, et al. 2019. Acoel genome reveals the regulatory landscape of whole-body regeneration. *Science* 363(6432):eauu6173.
Gerdol M, Luo Y, Satoh N, Pallavicini A. 2018. Genetic and molecular basis of the immune system in the brachiopod *Lingula anatina*. *Dev Comp Immunol*. 82:7–30.
Gerdol M, Venier P, Edomi P, Pallavicini A. 2017. Diversity and evolution of TIR-domain-containing proteins in bivalves and Metazoa: new insights from comparative genomics. *Dev Comp Immunol*. 70:145–164.
Halanych KM, Kocot KM. 2014. Repurposed transcriptomic data facilitate discover of innate immunity Toll-like receptor (TLR) genes across Lophotrochozoa. *Biol Bull*. 227(2):201–209.
Hall MR, Kocot KM, Baughman KW, Fernandez-Valverde SL, Gauthier MEA, Hatleberg WL, Krishnan A, McDougall C, Motti CA, Shoguchi E, et al. 2017. The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. *Nature* 544(7649):231–234.
Hamada M, Shoguchi E, Shinzato C, Kawashima T, Miller DJ, Satoh N. 2013. The complex NOD-like receptor repertoire of the coral *Acropora digitifera* includes novel domain combinations. *Mol Biol Evol*. 30(1):167–176.
Hayden MS, Ghosh S. 2011. NF-κB in immunobiology. *Cell Res*. 21(2):223–244.
Hernando D, Crespi V, Cybenko G. 2005. Efficient computation of the hidden Markov model entropy for a given observation sequence. *IEEE Trans Inform Theory*. 51(7):2681–2685.
Hibino T, Loza-Coll M, Messier C, Majeske AJ, Cohen AH, Terwilliger DP, Buckley KM, Brockton V, Nair SV, Berney K, et al. 2006. The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol*. 300(1):349–365.
Hiscott J. 2007. Convergence of the NF-κB and IRF pathways in the regulation of innate antiviral response. *Cytokine Growth Factor Rev*. 18(5–6):483–490.
Hourdez S, Lallier FH. 2007. Adaptations to hypoxia in hydrothermal-vent and cold-seep invertebrates. *Rev Environ Sci Biotechnol*. 6(1–3):143–157.
Huang S, Yuan S, Guo L, Yu Y, Li J, Wu T, Liu T, Yang M, Wu K, Liu H, et al. 2008. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Res*. 18(7):1112–1126.
Jiang F, Ramanathan A, Miller MT, Tang G, Gale M, JrPatel SS, Marcotrigiano J. 2011. Structural basis of RNA recognition and activation by innate immune receptor RIG-I. *Nature* 479(7373):423–429.
Jiang X, Kinch LN, Brautigam CA, Chen X, Du F, Grishin NV, Chen ZJ. 2012. Ubiquitin-induced oligomerization of the RNA sensors RIG-I and MDA5 activates antiviral innate immune response. *Immunity* 36(6):959–973.
Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 14(6):587–589.
Kanneganti T, Lamkanfi M, Núñez G. 2007. Intracellular NOD-like receptors in host defense and disease. *Immunity* 27(4):549–559.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.
Kowalinski E, Lunardi T, McCarthy AA, Louber J, Brunel J, Grigorov B, Gerlier D, Cusack S. 2011. Structural basis for the activation of innate immune pattern-recognition receptor RIG-I by viral RNA. *Cell* 147(2):423–435.
Kwon D, Yoon JH, Shin S, Jang T, Kim H, So I, Jeon J, Park HH. 2012. A comprehensive manually curated protein-protein interaction database for the death domain superfamily. *Nucleic Acids Res*. 40(Database issue):D331–D336.
Laetsch DR, Blaxter ML. 2017. BlobTools: interrogation of genome assemblies. *F1000Res*. 6:1287.
Lamkanfi M, Dixit VM. 2012. Inflammasomes and their role in health and disease. *Annu Rev Cell Dev Biol*. 28:137–161.
Lange C, Hemmrich G, Klostermeier UC, López-Quintero JA, Miller DJ, Rahn T, Weiss Y, Bosch TCG, Rosenstiel P. 2011. Defining the origins of the NOD-like receptor system at the base of animal evolution. *Mol Biol Evol*. 28(5):1687–1702.
Laroui H, Yan Y, Narui Y, Ingersoll SA, Ayyadurai S, Charania MA, Zhou F, Wang B, Salaita K, Sitaraman SV. 2011. L-Ala-γ-D-Glu-meso-diaminopimelic acid (DAP) interacts directly with leucin-rich region domain of nucleotide-binding oligomerization domain 1, increasing phosphorylation activity of receptor-interacting serine/threonine-protein kinase 2 and its interaction with nucleotide-binding oligomerization domain 1. *J Biol Chem*. 286(35):31003–31013.
Lasi M, David CN, Böttger A. 2010. Apoptosis in pre-Bilaterians: *hydra* as a model. *Apoptosis* 15(3):269–278.
Laumer CE, Fernández R, Lemer S, Combosch D, Kocot KM, Riesgo A, Andrade SCS, Sterrer W, Sørensen MV, Giribet G. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc Biol Sci*. 286(1906):20190831.
Lechtenberg BC, Mace PD, Riedl SJ. 2014. Structural mechanisms in NLR inflammasome signaling. *Curr Opin Struct Biol*. 29:17–25.
Lees JG, Lee D, Studer RA, Dawson NL, Sillitoe I, Das S, Yeats C, Dessailly BH, Rentzsch R, Orengo CA. 2014. Gene3D: multi-domain

annotations for protein sequence and comparative genome analysis. *Nucleic Acids Res.* 42(D1):D240–D245.
Letunic I, Bork P. 2019. Interactive Tree of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47(W1):W256–W259.
Leulier F, Lemaitre B. 2008. Toll-like receptors – taking an evolutionary approach. *Nat Rev Genet.* 9(3):165–178.
Liu M, Liao W, Buckley KM, Yang SY, Rast JP, Fugmann SD. 2018. AID/APOBEC-like cytidine deaminases are ancient immune mediators in invertebrates. *Nat Commun.* 9(1):1948.
Liu Y, Olagnier D, Lin R. 2016. Host and viral modulation of RIG-I-mediated antiviral immunity. *Front Immunol.* 7(662):662.
Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A. 2009. Protein function annotation by homology-based inference. *Genome Biol.* 10(2):207.
Luo D, Ding SSC, Vela A, Kohlway A, Lindenbach BD, Pyle AM. 2011. Structural insights into RNA recognition by RIG-I. *Cell* 147(2):409–422. [PMC][10.1016/j.cell.2011.09.023] [22000018]
Man SM, Kanneganti T. 2016. Converging roles of caspases in inflammasome activation, cell death, and innate immunity. *Nat Rev Immunol.* 16(1):7–21.
Messier-Solek C, Buckley KM, Rast JP. 2010. Highly diversified innate receptor systems and new forms of animal immunity. *Semin Immunol.* 22(1):39–47.
Meunier E, Broz P. 2017. Evolutionary convergence and divergence in NLR function and structure. *Trends Immunol.* 38(10):744–757.
Minh BQ, Nguyen MAT, Haeseler V. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol.* 30(5):1188–1195.
Mistry J, Cihuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ. 2020. Pfam: the proteins families database in 2021. *Nucleic Acids Res.* 41:D412–D419. doi:10.1093/narlgkaa913.
Mo J, Boyle JP, Howard CB, Monie TP, Davis BK, Duncan JA. 2012. Pathogen sensing by nucleotide-binding oligomerization domain-containing protein 2 (NOD2) is mediated by direct binding to muramyl dipeptide and ATP. *J Biol Chem.* 287(27):23057–23067.
Mukherjee K, Korithoski B, Kolaczkowski B. 2014. Ancient origins of vertebrate-specific innate antiviral immunity. *Mol Biol Evol.* 31(1):140–153.
Nehyba J, Hrdličková R, Bose HR. 2009. Dynamic evolution of immune system regulators: the history of interferon regulatory factor family. *Mol Biol Evol.* 26(11):2539–2550.
Nguyen L, Schmidt HA, Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 32(1):268–274.
O'Neill LAJ, Bowie AG. 2007. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nat Rev Immunol.* 7(5):353–364.
Park HH, Lo Y, Lin S, Wang L, Yang JK, Wu H. 2007. The death domain superfamily in intracellular signaling of apoptosis and inflammation. *Annu Rev Immunol.* 25:561–586.
Pellegrini M. 2015. Tandem repeats in proteins: prediction algorithms and biological role. *Front Bioeng Biotechnol.* 3:143.
Pippig DA, Hellmuth JC, Cui S, Kirchhofer A, Lammens K, Lammens A, Schmidt A, Rothenfusser S, Hopfner K. 2009. The regulatory domain of the RIG-I family ATPase LGP2 senses double-stranded RNA. *Nucleic Acids Res.* 37(6):2014–2025.
Pugh C, Kolaczkowski O, Manny A, Korithoski B, Kolaczkowski B. 2016. Resurrecting ancestral structural dynamics of an antiviral immune receptor: adaptive binding pocket reorganization repeatedly shifts RNA preference. *BMC Evol Biol.* 16(1):241.
Reikine S, Nguyen JB, Modis Y. 2014. Pattern recognition and signaling mechanisms of RIG-I and MDA5. *Front Immunol.* 5:342.
Riley JS, Malik A, Holohan C, Longley DB. 2015. DED or alive: assembly and regulation of death effect domain complexes. *Cell Death Dis.* 6:e1866.
Sakamaki K, Imai K, Tomii K, Miller DJ. 2015. Evolutionary analysis of caspase-8 and its paralogs: deep origins of the apoptotic signaling pathways. *Bioessays* 37(7):767–776.
Sasaki N, Ogasawara M, Sekiguchi T, Kusumoto S, Satake H. 2009. Toll-like receptors of the ascidian *Ciona intestinalis* prototypes with hybrid functionalities of vertebrate Toll-like receptors. *J Biol Chem.* 284(40):27336–27343.
Satake H, Sekiguchi T. 2012. Toll-like receptors of deuterostome invertebrates. *Front Immunol.* 3(34):34.
Simakov O, Kawashima T, Marlétaz F, Jenkins J, Koyanagi R, Mitros T, Hisata K, Bredeson J, Shoguchi E, Gyoja F, et al. 2015. Hemichordate genomes and deuterostome origins. *Nature* 527(7579):459–465.
Simakov O, Marletaz F, Cho S-J, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo D-H, Larsson T, Lv J, Arendt D, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature* 493(7433):526–531.
Sonnhammer ELL, Eddy SR, Durbin R. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28(3):405–420.
Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. 2015. STRING v10: protein-protein interaction networks integrated over the tree of life. *Nucleic Acids Res.* 43(Database issue):D447–D452.
Tassia MG, Whelan NV, Halanych KM. 2017. Toll-like receptor pathway evolution in deuterostomes. *Proc Natl Acad Sci U S A.* 114(27):7055–7060.
The UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47:D506–D515.
Toshchakov VY, Neuwald AF. 2020. A survey of TIR domain sequence and structure divergence. *Immunogenetics* 72(3):181–123.
Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. 2015. Tissue-based map of the human proteome. *Science* 347(6220):1260419.
Valmiki MG, Ramos JW. 2009. Death effector domain-containing proteins. *Cell Mol Life Sci.* 66(5):814–830.
Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol.* 35(3):543–548.
Wheeler TJ, Clements J, Finn RD. 2014. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics* 15:7.
Wojcik J, Schächter V. 2001. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 17(1):S296–S305.
Wu X, Wu F, Wang X, Wang L, Siedow JN, Zhang W, Pei Z. 2014. Molecular evolutionary and structural analysis of the cytosolic DNA sensor cGAS and STING. *Nucleic Acids Res.* 42(13):8243–8257.
Yuen B, Bayes JM, Degnan SM. 2014. The characterization of sponge NLRs provides insight into the origin and evolution of this innate immune gene family in animals. *Mol Biol Evol.* 31(1):106–120.
Zhang Q, Lenardo MJ, Baltimore D. 2017. 30 years of NF-κB: a blossoming of relevance to human pathobiology. *Cell* 168(1–2):37–57.
Zhang X, Wang Z, Zhang X, Le MH, Sun J, Xu D, Cheng J, Stacey G. 2012. Evolutionary dynamics of protein domain architecture in plants. *BMC Evol Biol.* 12:6.
Zhao X, Wang Y, Chen L, Aihara K. 2008. Protein domain annotation with integration of heterogeneous information sources. *Proteins* 72(1):461–473.