OXFORD UNIVERSITY PRESS | Molecular Biology and Evolution

TIAMMAt: Leveraging biodiversity to revise protein domain models, evidence from innate immunity

Journal:	*Molecular Biology and Evolution*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Tassia, Michael; Auburn University, Biological Sciences David, Kyle; Auburn University, Townsend, James; Marine Biological Laboratory, Whitman Center; Providence College, Department of Biology Halanych, Kenneth; Auburn University, Biological Sciences
Key Words:	Protein Evolution, Domain Annotation, Animal Evolution, Innate Immunity

Molecular Biology & Evolution: Resource Article

Title:

TIAMMAt: Leveraging biodiversity to revise protein domain models, evidence from innate immunity

Authors:

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

Affiliations:

[1]Department of Biological Sciences, Auburn University, Auburn, Alabama 36849

[2] Whitman Center, Marine Biological Laboratory, Woods Hole, Massachusetts 02543

[3]Department of Biology, Providence College, Providence, Rhode Island 02908

Corresponding Author:

Michael G. Tassia

mgt0007@auburn.edu

ABSTRACT

Sequence annotation is fundamental for studying the evolution of protein families, particularly when working with non-model species. Given the rapid, ever-increasing number of enigmatic species receiving high-quality genome sequencing, accurate domain modeling that is representative of species diversity, instead of a few traditional biomedical model species, is crucial for understanding protein family sequence evolution and their inferred function(s). In this study, we describe a bioinformatic tool, called TIAMMAt (*Taxon-Informed Adjustment of Markov Model Attributes*), which revises domain profile hidden Markov models by incorporating homologous domain sequences from underrepresented, non-model 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 can provide valuable insight into the evolution of protein families. Following domain profile adjustment, domain profiles exhibit changes in their per-site amino acid state emission probabilities and insertion/deletion probabilities while maintaining the overall structure of the consensus sequence. We examine and revise six domains central to several protein families involved in immunity, NLRs, TLRs, RLRs, IRFs, and NF-κBs. Our results show that domain revision can heavily impact evolutionary interpretations for some families (i.e., NLR's NACHT domain), whereas revisional impact on other domains (e.g., rel homology domain and interferon regulatory factor domains) is minimal due to high levels of sequence conservation across our sampled phylogenetic depth (i.e., Metazoa). Future studies on protein evolution incorporating, or focusing, on non-model species will benefit from TIAMMAt improving phylogenetic representation within domain models.

INTRODUCTION

Accurate assignment of protein identity is a fundamental component of molecular studies involving non-model species. Such studies often begin by tethering an uncharacterized protein's identity to a homolog of known function to allow inference of, for example, residue-specific selective pressures (Buckley & Rast 2012), protein-protein interaction networks (Szklarczyk et al. 2014), 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 non-model species can compromise the design of genetic reporter constructs (Cavalieri & Spinelli 2014) or CRISPR/Cas9 targets (Connahs et al. 2019). These errors cost researchers time and financial resources, and/or negatively impact the accuracy of scientific conclusions.

Comparative molecular studies employing non-model species (Buckley & Rast 2015; Brennan & 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). In light of 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 & 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 flexibility across deep evolutionary timescales (Buckley & Rast 2012; Costa-Paiva et al. 2017; Tassia et al. 2017; Gerdol et al. 2017; Costa-Paiva et al. 2018). 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 & 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 catalogue of domain models placed in an evolutionary context for protein studies across the tree of life (Sonnhammer et al. 1997; Finn et al. 2016; El-Gebali et al. 2018). Each Pfam domain entry is created as follows (**Fig. 1**): 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. 2018; Mistry et al. 2020). Encoding 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 (e.g., HMMER, Eddy 2009). 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. 2018), 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, seed alignments are dominated by sequences from a few biomedical model taxa (**Fig. 2**), and the trend in sequencing bias towards these model systems is becoming increasingly exacerbated (David et al. 2019).

Here, we show that revising domain profile seed alignments aids identification of homologs in non-model animal species. The value of phylogenetically representative domain models cannot be overstated. Identifying protein homologs across deep evolutionary timescales (e.g., Buckley & Rast 2012; Costa-Paiva et al. 2017; Tassia et al. 2017; Gerdol et al. 2017; Costa-Paiva et al. 2018) is a challenge that continues to grow as genomes become more accessible, particularly for those of historically underrepresented species. 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 which rely on domain-domain interactions and show considerable variation between taxa.

Innate immunity is an animal's most rapid defense against invading pathogens and 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 & Kasahara 2010), innate immunity pathways and 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 & Rast 2012; Gerdol et al 2017). The capacity for PRRs to bind ligands, transduce a cytoplasmic signal, and initiate transcription of immune factors is reliant upon domain-domain interactions. Among the most well-described PRR signaling pathways are NOD-like receptors (NLRs; Lechtenberg et al. 2014), Toll-like receptors (TLRs; Akira & 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 their unique signal transduction partners (**Fig. 3** & **Supplementary Fig. 1**), 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. Because these protein families rely heavily on domain-domain interactions for both activation and signal transduction, possess defined domain architectures (Akira & Takeda 2004; Kowalinski et al. 2011; Lechtenberg et al. 2014), and have dominantly been studied in biomedical model species (Leulier & Lemaitre 2008), innate immunity proteins present an ideal case study for revising domain models using non-model species to address broader evolutionary patterns across Metazoa. Here, we describe a domain revision protocol called TIAMMAt (*Taxon-Informed Adjustment of Markov Model Attributes*) and apply it to the domains at the core of PRR signaling to reveal the effects of narrow phylogenetic representation within domain seed alignments, and the application of their derived models when searching for domain homologs in non-model species.

NEW APPROACHES

TIAMMAt (*Taxon-Informed Adjustment of Markov Model Attributes*) provides an automated and reproducible method for revising Pfam domain models to capture homologous sequence diversity contingent upon a user-defined taxonomic distribution. TIAMMAt fundamentally relies on HMMER's suite of profile HMM tools and their direct association with Pfam domain database entries. Although TIAMMAt is executed 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 could be applied to investigating the DEATH domain superfamily in all eukaryotes or globin domains solely within arachnids. The extensibility of TIAMMAt makes it a powerful and flexible tool for studies employing non-model or poorly sampled taxa when investigating protein family sequence evolution.

TIAMMAt executes the following steps for each target domain profile (see **Materials & Methods, Fig. 4A, Supplementary Fig. 2,** and **Supplementary Table 1**). First, each supplied proteome (defined here as the whole collection of protein sequences derived of an organism's genome) is searched for occurrences of the target domain where the target domain meets two conditions: (A) The target domain is the best-hit match within the sequence envelope in which it is identified and, (B) meets HMMER's inclusion thresholds. The stringency of this step is specifically designed to avoid type-I errors which could otherwise be introduced due by incorporating similar, but non-target, motifs that belong to a domain family (e.g., CARD and DED are similar, but exhibit different interactive properties; Jiang et al. 2012). Second, best-hit domains are extracted from their parent sequences in each proteome and aligned 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 domain models, similar to the first step. For the purposes of our study, we revised domains integral to the function of TLRs, RLRs, NLRs, NF-κBs, and IRFs (**Supplementary Table 1, Supplementary Table 2**) using 39 publicly available proteomic datasets representative of taxa across the metazoan phylogeny (**Supplementary Table 3**), focusing particularly on species which have been labeled within scientific literature as emerging or non-model (e.g., Simakov et al. 2013; Simakov et al. 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. 4B**): (1) Overall sequence structure remains consistent between base and revised models, where low information content regions retain low impact on sequence match 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 resides towards the 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 human domain-containing sequences identified before revision were also identified after domain revision. Additionally, the few human sequences found to possess a target domain only after revision (**Supplementary Table 2**) met one of two conditions: A) the sequence phylogenetically clustered with previously described domain-containing proteins suggesting the model revision produced a profile which statistically captured variation within the sequence that could not be accurately described by the original model; or B) the sequence represented a poorly understood protein and the revised domain was assigned to a sequence envelope where no other unrelated domain was previously described. Importantly, the magnitude of change for individual amino acid emission probabilities per-site were not equivalent across all models, suggesting TIAMMAt is sensitive the degree of sequence conservation within the input domain(s). In turn, some domain model revisions in our study had little impact on the broader evolutionary understanding for their associated protein families. For example, domain revision yielded 4 and 8 additional IRF and NF-κB family members, respectively (**Supplemental Table 2**). When all IRFs or NF-κB proteins were placed in a phylogenetic context, resolution of deeper nodes was poor (**Supplemental Fig. 3 & 4, Supplemental File 1 & 2**), and the lack of domain architecture diversity among some of 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 yielded more dramatic changes that could be more confidently placed into an evolutionary framework.

NOD-like Receptors.

NACHT domain revision yielded the greatest increase in the number of domain-containing sequences in our analyses (**Supplementary Tables 4-6**). We defined NLRs as proteins possessing both a NACHT domain and a terminal series of LRRs, consistent with literature on the structural perspectives on NLR signaling kinetics and previous NLR surveys (Laroui et al. 2011; Mo et al. 2012; Meunier & 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**). Aside from novel CARD-containing NLRs (i.e., NLRC subfamily) identified in *P. 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 & Broz, 2017) based solely on domain architecture (**Supplementary Fig. 5; Supplementary Table 6**). 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. 2013). 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, a short adaptor protein containing both a PYD and CARD (Lamkanfi & Dixit, 2012), before signaling for cell-death.

Our evolutionary analysis support studies (Messier-Solek 2010; Hamada et al. 2013; Yuen et al. 2013; Gerdol et al. 2018; **Supplementary Table 6, Supplementary Fig. 6, Supplementary File 3**) that vertebrate-defined NLR subfamilies (i.e., NLRAs, NLRBs, NLRCs, and NLRPs) are insufficient for classifying NLRs outside Vertebrata. Noncanonical NLRs identified in our study include a collection N-terminal Death domain (or juxtaposed Death and CARD) NLRs present in cephalochordates (*Branchiostoma belcheri* and *B. floridae*) and echinoderms (*Acanthaster planci, Patiria miniata, Apostichopus japonicus, Strongylocentrotus purpuratus, Lytechinus variegatus*) (**Supplementary Table 6**). Assuming the overall domain structures of metazoan NLRs retain their functional regionalization (i.e., C-terminal LRRs operate as ligand-binding, NACHT domain promote oligomerization, and the N-terminal domain(s) is/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).

Outside of the death-domain superfamily, NLRs among identified in nine invertebrate taxa possess a higher eukaryotes and prokaryotes nucleotide-binding domain (HEPN) 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 responses. Although we can loosely predict HEPN-NLRs may function as an ancient 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. Nonetheless, the N-terminal domain of NLRs is far more diverse than what has traditionally been represented within vertebrates. The non-canonical 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 incorporates all high-confidence occurrences of the targeted domain during its search, NACHT revision also output several NACHT domain-containing proteins with undocumented affinity for NLR signaling pathways (**Supplementary Fig. 5, Supplementary Table 6**). Given the proclivity for NACHT to facilitate homotypic oligomerization events upon activation (Lamkanfi & Dixit, 2012), this cluster of proteins warrant further study for a potential role in NLR signaling regulation. Importantly, though these proteins were not a direct target of NACHT revision in the context of our study, their return by TIAMMAt nonetheless exemplifies the utility of revising a domain.

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, *Branchiostoma 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 (Halanych & Kocot 2014; Huang et al. 2015; Gerdol et al. 2018), proteins detected in tunicates after revision are proportionally more substantial, doubling the number of reported TLRs in *B. schlosseri* from 1 to 2 (Tassia et al. 2017; Franchi et al. 2019), and in *C. intestinalis* from 3 to 5 (Buckley & 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 statistically outcompeted another high confidence, but unrelated, domain annotation. Thus, TIAMMAt's results yielded a domain architecture fitting the canonical schema for TLRs (Akira & 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 to confidently be placed into the final annotation (per-sequence e-value > 0.01). In contrast, the novel tunicate TLRs lacked any reportable TIR domain prior to revision (**Fig. 5**), suggesting the newly identified tunicate proteins contain a divergent TIR domain relative to sequences present within the original seed alignment. Previous analyses have shown both of *Ciona*'s previously described TLRs as a functional blend of several vertebrate homologs (Sasaki et al. 2009; Satake & Sekiguchi 2012). These previous data describing functional divergence of tunicate TLRs and the newly identified TLRs detected in this study may be causally tied to tunicate's evident rapid rate of molecular evolution relative to their sister phylum, Vertebrata (Berná & 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 and implications for TLR pathway evolution (**Supplementary Table 4; Supplementary Fig. 5**). 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, additionally reporting a 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**). Furthermore, whereas homologs to MYD88 and SARM1 (canonical TIR-DC adaptor proteins responsible for signal transduction and regulation of TLRs, respectively; O'Neill & Bowie 2007) possess ancestry predating the emergence of Vertebrata (Tassia et al. 2017; Toschachev & Neuwald 2020), many evolutionarily conserved TIR-DC proteins (defined in Gerdol et al. 2017) identified here lack any vertebrate homologs (**Supplementary Table 4**). Even when including proteomes from earlier diverging, non-mammalian vertebrate lineages (i.e., hagfish, lamprey, and *Latimeria*) and revising the TIR domain to capture homologous sequence variation from deep within animal phylogeny, vertebrate TIR-DC proteins appear to be confined to TLRs, IL-1Rs, and the five traditional TLR adaptors. Although there may be some causal relationship between the emergence of adaptive immunity and the limited number of TIR-DC protein structures within vertebrates, the non-canonical 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; **Fig. 3**; Esser-Nobis et al. 2020) revealed novel RLR proteins in the cnidarian, *Hydra vulgaris* (n=1 additional), and the sea star, *Patiria miniata* (n=2; **Supplementary Table 5**). Unlike canonical RLRs, novel proteins identified in *H. 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 to the vertebrate LGP2 structure. 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 non-canonical 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 early diverging bilaterian 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. 7, Supplementary File 4**). 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. 7, Supplementary File 4**). Concordant with previous studies (Mukherjee et al. 2013; 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. 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 & Ramos, 2009; Riley et al. 2015; Man & 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 & 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.

TIAMMAt's strength lies in its flexibility to be executed using a user-defined taxonomic distribution and any number of Pfam domain models. 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, given the computational resources. One could also apply TIAMMAt 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. The design of TIAMMAt was stimulated through a combination of uncovering a lack of even phylogenetic representation within domain profile seed alignments (**Figure 2**), and the many studies which rely on domain annotation as an inferential tool to estimate a non-model species' capacity to perform any given molecular pathway (e.g, Hibino et al. 2006; Costa-Paiva et al. 2017; Gerdol et al. 2017; Tassia et al. 2017). As such, future studies focusing on domain/protein evolution within non-model taxa (particularly for those with particularly poor representation within molecular evolution literature) will be the major benefactors of TIAMMAt.

Conclusions.

TIAMMAt leverages biodiversity to revise domain profile HMMs to alleviate taxonomic bias within the Pfam database. Because the assignment of protein identity and functional inference is contingent upon the premise that the associated database is evenly representative of phylogeny, revising domain models to capture homologous sequence variation using a strict, yet unsupervised, method is a critical improvement for studies focusing on species outside the scope of the few biomedical models. In our application of TIAMMAt, we revised domains central to the signaling pathways which participate in animal innate immunity. As shown in our case study, the evolutionary implications for individual domain revisions varies on a case-by-case basis, such that some model revisions (e.g., NACHT) may yield comparably dramatic changes in the number of domain-containing sequences detected, whereas other domain revisions (e.g., IRF) may have little impact on the evolutionary interpretations surrounding the surrounding protein family. Interestingly, among our results we found several cases where one death-domain superfamily member was being substituted for another, implicating some evolutionary plasticity within the death-domain superfamily to perhaps interact with other downstream effectors (e.g., caspases). As genomic and proteomic sequences become increasingly available, TIAMMAt will be a valuable asset for revising domains to garner confidence in protein evolution studies.

MATERIALS & METHODS

Input Dataset Acquisition.

Protein sequence accessions for the 39 metazoan taxa used in this study are available in **Supplementary Table 3**. Species were chosen to represent a deep phylogenetic timescale and their selection was contingent upon genomic/proteomic data availability. Regarding the two species where protein sequence datasets were not directly downloadable (i.e., *Hofstenia miamia* and *Schmidtea mediterranea*), the scaffolded genome 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.** Particularly, we chose domains traditionally associated with TLRs (i.e., TIR & TIR_2 domains; Tassia et al. 2017), RLRs (i.e., RIG-I_C-RD & CARD domains; Liu et al. 2017), NLRs (i.e., NACHT & CARD domains; Elinav et al. 2011), IRFs (i.e., IRF & IRF3 domains; Nehyba et al. 2009), and NFkB (i.e., RHD domain; Hayden & Ghosh 2011). All domain models and their seeds were obtained from Pfam version 32.0 (El-Gebali et al. 2018). 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).

As with any bioinformatic software, the quality of the model revision by TIAMMAt is impacted by the quality of input proteomes. As such, we recommend using TIAMMAt only after careful consideration of input 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 (Simão et al. 2015; Waterhouse et al. 2018) before using the tool.

Database Bias.

Pfam domain profile seed alignments were downloaded from the Pfam 32.0 FTP server on April 7th, 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. 2**).

Domain Profile HMM Revision.

TIAMMAt (*Taxon-Informed Adjustment of Markov Model Attributes*) automates revision of Pfam domain models to capture homologous sequence diversity based upon taxonomic distribution provided by the user (**Supplementary Fig. 2**). The program is written using open-source software packages and is publicly available via GitHub (*author's note to editor/reviewer: URL provide upon manuscript publication*). 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 default parameters. 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 parameters. 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 conventional per-sequence and per-domain (both conditional and independent)

E-value inclusion thresholds of 0.01. 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 non-overlapping list of highest-confidence domain hits per sequence. 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 to be noise from the perspective of the program and are omitted from the following steps due to lack of statistical substantiation.

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. Because the revised model necessarily relies on Pfam's base domain profile HMM and seed sequences, the revision's consensus residue per node is often unchanged when compared with the original model; however, the state-transition probabilities are adjusted to capture the larger domain alignment matrix. After the domain model has been revised, TIAMMAt loops once more through *hmmsearch* and *hmmscan* as it did before, this time isolating sequences which possess either the base or revised domain model hits.

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 considered 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 (both sequences and markov models) are also available via the TIAMMAt github repository (*author's note to editor/reviewer: URL provided upon manuscript publication*).

Phylogenetic Methods.

Each protein family was aligned using MAFFT version 7's L-INS-I protocol (Katoh & 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 & 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 x 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 & Bork, 2019). Unedited tree files for both likelihood and Bayesian phylogenetic inferences from this study are available via the TIAMMAt github repository (*author's note to editor/reviewer: URL provided upon manuscript publication*).

ACKNOWLEDGEMENTS

This work was supported by The National Science Foundation (grant number IOS – 1755377 to KMH), and KTD 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. Computational resources were made available by the Alabama Supercomputer Authority and the Auburn University Hopper Cluster. This publication is Molette Biology Laboratory contribution ## and Auburn University Marine Biology Program contribution ## (*author's note to editor/reviewer: number placeholders will be filled at the time of publication*). Code and raw data obtained in our study are publicly available via the TIAMMAt Github page (*URL available at the time of publication*).

REFERENCES

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

Akira S, Takeda K. 2004. Toll-like receptor signaling. *Nat Rev Immunol* 4: 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.

Beutler B. 2004. Innate immunity: an overview. *Molecular Immunol.* 40: 845-859.

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

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) doi: 10.3389/fimmu.2012.00136

Buckley KM, Rast JP. 2015. Diversity of animal immune receptors and the origins of recognition complexity in the deuterostomes. *Dev. Comp. Immunol.* 49, 179-189.

Callahan N, Tullman J, Kelman Z, Marino M. 2020. Strategies for development of a next-generation protein sequencing platform. *Trends in biochemical sciences* 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: doi:10.7554/eLife.04664

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: doi:10.1242/dev.169367

Costa-Paiva EM, Shrago CG, Halanych KM. 2017. Broad phylogenetic occurrence of oxygen-binding hemerythrins in bilaterians. *Genome Biol. Evol.* 9: 2580-2591.

Costa-Paiva EM, Shrago CG, Coates CJ, Halanych KM. 2018. Discovery of novel hemocyanin genes in metazoans. *Biol. Bull.* 235: 134-151

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 Informatics* 23: 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. 2018. The Pfam protein families database in 2019. *Nucleic Acids Res.* doi: 10.1093/nar/gky995.

Elinav E, Strowig T, Henao-Mejia J, Flavell RA. 2011. Regulation of the antimicrobial responses by NLR proteins. *Immunity* 34: 665-679.

Esser-Nobis K, Hatfield LD, Gale Jr M. 2020. Spatiotemporal dynamics of innate immune signaling via RIG-I-like receptors. *PNAS.* doi: 10.1073/pnas.1921861117.

Flajnik MF, Kasahara M. 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* 11: 47-59.

Forterre P. 2006. The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res.* 117, 5-16

Franchi N, Ballarin L, Peronato A, Cima F, Grimabaldi 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, Srivastava M. 2019. Acoel genome reveals the regulatory landscape of whole-body regeneration. *Science* 363, eauu6173.

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. *Devel. Comp. Immunol.* 70: 145-164.

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.

Halanych KM, Kocot KM. 2014. Repurposed transcriptomic data facilitate discover of innate immunity Toll-like receptor (TLR) genes across Lophotrochozoa. *Biol. Bull.* 227: 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, 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 Research* 21: 223-244.

Hernando D, Crespi V, Cybenko G. 2005. Efficient computation of the hidden Markov model entropy for a given observation sequence. *IEEE Transactions on Information 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: 349-365.

Hiscott J. 2007. Convergence of the NF-κB and IRF pathways in the regulation of innate antiviral response. *Cytokine & Growth Factor Reviews* 18: 483-490.

Huang S, Yan S, Guo L, Yanhong Y, Li J, Wu T, Liu T, Yang M, Wu K, Liu H, et al. 2015. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Research* 18: 1112-1126.

Jiang F, Ramanathan A, Miller MT, Tang G, Gale Jr. M, Patel SS, Marcotrigiano J. 2011. Structural basis of RNA recognition and activation by innate immune receptor RIG-I. *Nature* 479: 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, 959-973.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587-589.

Kanneganti T, Lamkanfi M, Núñez G. 2007. Intracellular NOD-like receptors in host defense and disease. *Immunity* 27: 549-559.

Katoh K, Standly 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: 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 Reas.* 40: D331-336

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.

Lamkanfi M, Dixit VM. 2012. Inflammasomes and their role in health and disease. *Annu. Rev. Cell Dev. Biol.* 28:137-161

Laroui H, Yan Y, Narui Y, Ingersoll SA, Ayyadurai S, Charania MA, Zhou F, Wang B, Salaita K, Sitaraman SV, et al. 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, 269-278.

Laumer CE, Fernández R, Lemer S, Combosch D, Kocot KM, Resgo A, Andrade SCS, STerrer W, Sørensen MV, Giribet G. 2019. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. R. Soc. B* 286: doi:10.1098/rspb.2019.0831

Lechtenberg BC, Mace PD, Riedl SJ. 2014. Structural mechanisms in NLR inflammasome signaling. *Current Opinion in Structural Biology* 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.* D1, D240-D245.

Letunic I, Bork P. 2019. Interactive Tree of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47(W1): W256-259

Leulier F, Lemaitre B. 2008. Toll-like receptors – taking an evolutionary approach. *Nat. Rev. Genet.* 9: 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, doi:10/1038/s41467-018-04273-x

Liu Y, Olagnier D, Lin R. 2017. Host and viral modulation of RIG-I-mediated antiviral immunity. *Front. Immuol.* 7(662), doi: 10.3389/Immu.2016.00662.

Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Oregno C, Thornston J, Tramantano A. 2009. Protein function annotation by homology-based inference. *Genome Biology* 10(207), doi:10.1186/gb-2009-10-2-207.

Luo D, Ding SSC, Vela A, Kohlway A, Lindenbach BD, Pyle AM. 2011. Structural insights into RNA recognition by RIG-I. *Cell* 147: *409-422.*

Man SM, Kanneganti T. 2016. Converging roles of caspases in inflammasome activation, cell death, and innate immunity. *Nat Rev Immunol* 16, 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, et al. 2020. Pfam: The proteins families database in 2021. *Nucleic Acids Res.* 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. 2013. 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 AD. 2007. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nat. Rev. Immunol.* 7: 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): doi:10.3389/fbioe.2015.00143

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. *Nuc. Acids. Res.* 37(6): 2014-2025.

Pugh C, Kolaczkowski O, Manny A, Korinthoski 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(241). doi:10.1186/s12862-016-0818-6.

Reikine S, Nguyen JB, Modis Y. 2014. Pattern recognition and signaling mechanisms of RIG-I and MDA5. *Front. Immunol.* 5(342), doi:10.3389/fimmu.2014.00342

Riley JS, Malik A, Holohan C, Longley DB. 2015. DED or alive: assembly and regulation of death effect domain complexes. *Cell Death and Disease* 6. doi: 10.1038/cddis.2015.213

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: 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: 27336-27343.

Satake H, Sakeiguchi T. 2012. Toll-like receptors of deuterostome invertebrates. *Front. Immunol.* 3(34). doi:10.3389/fimmu.2012.00034

Simakov O, Marletaz F, Cho S, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo D, Larsson T, Lv J, Arendt D, et al. 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature.* 493, 526-531.

Simakov O, Kwashima 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, 459-465.

Simão FA, Waterhouse RM, Ioannida P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19), 3210-3212

Sonnhammer ELL, Eddy SR, Durbin R. 1997. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics* 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. 2014. STRING v10: protein-protein interaction networks integrated over the tree of life. *Nucleic Acids Res.* 43: D447-452

Tassia MG, Whelan NV, Halanych KM. 2017. Toll-like receptor pathway evolution in deuterostomes. *PNAS* 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:* 1-23

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): 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. *MBE* 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: 8243-8257.

Yuen B, Bayes JM, Degnan SM. 2013. 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. 30 years of NF-κB: a blossoming of relevance to human pathobiology. *Cell* 168: 37-57.

Zhao X, Wang Y, Chen L, Aihara K. 2008. Protein domain annotation with integration of heterogeneous information sources. *Proteins* 72: 461-473.

FIGURE CAPTIONS

Figure 1. Conceptual schematic for the generation of a novel domain profile HMM. (A) Target protein homologs are isolated from their parent proteome and subsequence patterns are identified. These patterns are extracted and aligned to generate a domain seed alignment. (B) Network visualization of a profile HMM generated from a hypothetical seed alignment (Eddy 2009). Domain profile HMM encodes transition probabilities per-site for consensus state match → match, insertion, or deletion; insertion → match, insertion; and deletion → match, deletion for all positions in the model. (C) Visual representation of a profile HMM generated by Skylign (Wheeler et al. 2014). Relative height of amino acid symbols reflects their emission probability relative to all other amino acid states at that site.

Figure 2. Taxon representation within Pfam database. In blue (left values following species names) are the total number of occurrences a species appears across all Pfam seed alignments. In red (right values following species names) are the total number of sequences within a species' reference proteome captured by all Pfam domain profiles. Cladogram depicts consensus phylogenetic relationships derived from Laumer et al. 2019. Abbreviations: Cnid. – Cnidaria, Loph. – Lophotrochozoa, Ecdy. – Ecdysozoa, Echi. – Echinodermata, Hemi. – Hemichordata, Ceph. – Cephalochordata, Tuni. – Tunicata, Vert. – Vertebrata.

Figure 3. Diagram of NLR, TLR, and RLR signaling pathways. (Left) NLRs are cytoplasmically localized and possess a C-terminal series of leucine-rich repeats responsible for ligand binding, and a central NACHT domain involved in oligomerization and activation (Lechtenberg et al. 2014). NLR subfamilies differ in their N-terminal domain(s) which promote transcription factor activation or inflammasome assembly (Meunier & Broz, 2017). (Middle) TLRs are type-I transmembrane proteins localized to cell or endosomal membranes. Their N-terminal leucine-rich repeats bind pathogen-associated moieties, and the C-terminal TIR domain undergoes homotypic TIR domain interactions with one of five TIR-domain-containing adaptor proteins (Akira & Takeda 2004). (Right) RLRs are cytoplasmically localized and are exclusively involved in nucleic acid sensing. The central helicase

and C-terminal regulatory domain are involved in ligand binding and autoregulation, whereas the N-terminal CARD domains are involved in signal transduction (Reikine et al. 2014). All three pathways converge on the activation of NF-κB and IRF activation, transcription factors which promote the expression of host-defense compounds like pro-inflammatory cytokines and antiviral peptides, respectively.

Figure 4. Domain revision by TIAMMAt. A) Schematic overview of the three major operations performed by TIAMMAt (see **Materials & 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 re-annotate 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 (grey columns), non-consensus 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, 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.

Figure 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.

Supplemental Figure 1. Diagram of NLR, TLR, and RLR signaling pathways with all domains labeled. (Left) NLRs are cytoplasmically localized and possess a C-terminal series of leucine-rich repeats responsible for ligand binding, and a central NACHT domain involved in oligomerization and activation (Lechtenberg et al. 2014). NLR subfamilies differ in their N-terminal domain(s) which promote transcription factor activation or inflammasome assembly (Meunier & Broz, 2017). (Middle) TLRs are type-I transmembrane proteins localized to cell or endosomal membranes. Their N-terminal leucine-rich repeats bind pathogen-associated moieties, and the C-terminal TIR domain undergoes homotypic TIR domain interactions with one of five TIR-domain-containing adaptor proteins (Akira & Takeda 2004). (Right) RLRs are cytoplasmically localized and are exclusively involved in nucleic acid sensing. The central helicase and C-terminal regulatory domain are involved in ligand binding and autoregulation, whereas the N-terminal CARD domains are involved in signal transduction (Reikine et al. 2014). All three pathways converge on the activation of NF-κB and IRF activation, transcription factors which promote the expression of host-defense compounds like pro-inflammatory cytokines and antiviral peptides, respectively.

Supplemental Figure 2. Schematic of commands and data analysis performed by TIAMMAt (see **Materials & Methods** for further detail).

Supplemental Figure 3. Maximum-likelihood phylogenetic reconstruction of NF-κB family proteins using IQ-TREE. All nodes possess ultrafast-bootstrap support ≥95%. Best-fit model by BIC: VT+F+R6. Scale bar in number of substitutions per site.

Supplemental Figure 4. Maximum-likelihood phylogenetic reconstruction of IRFs using IQ-TREE. All nodes possess ultrafast-bootstrap support ≥95%. Best-fit model by BIC: VT+R6. Scale bar in number of substitutions per site.

Supplemental Figure 5. Domain architectures associated TIR, RLR C-terminal domain, and NACHT domain revision. Dotted outlines denote domains which are not necessary for protein classification, though may be present.

Supplemental Figure 6. Summarized topology of maximum-likelihood phylogenetic reconstruction of all proteins containing both NACHT domains and LRRs using IQ-TREE. All nodes possess ultrafast-bootstrap support ≥95%. Best-fit substitution model by BIC: VT+R10. Scale bar in number of substitutions per site.

Supplemental Figure 7. Summarized topology of Bayesian RLR phylogenetic reconstruction. Bold branches mark RLRs identified only after domain revision. Nodes with posterior probabilities (PP) <90% are collapsed. Nodes 90≤PP<100% are marked with white circles. Nodes with 100% PP are marked with black circles. Best-fit substitution model by BIC: VT+F+G. Scale bar in number of substitutions per site.

FIGURES

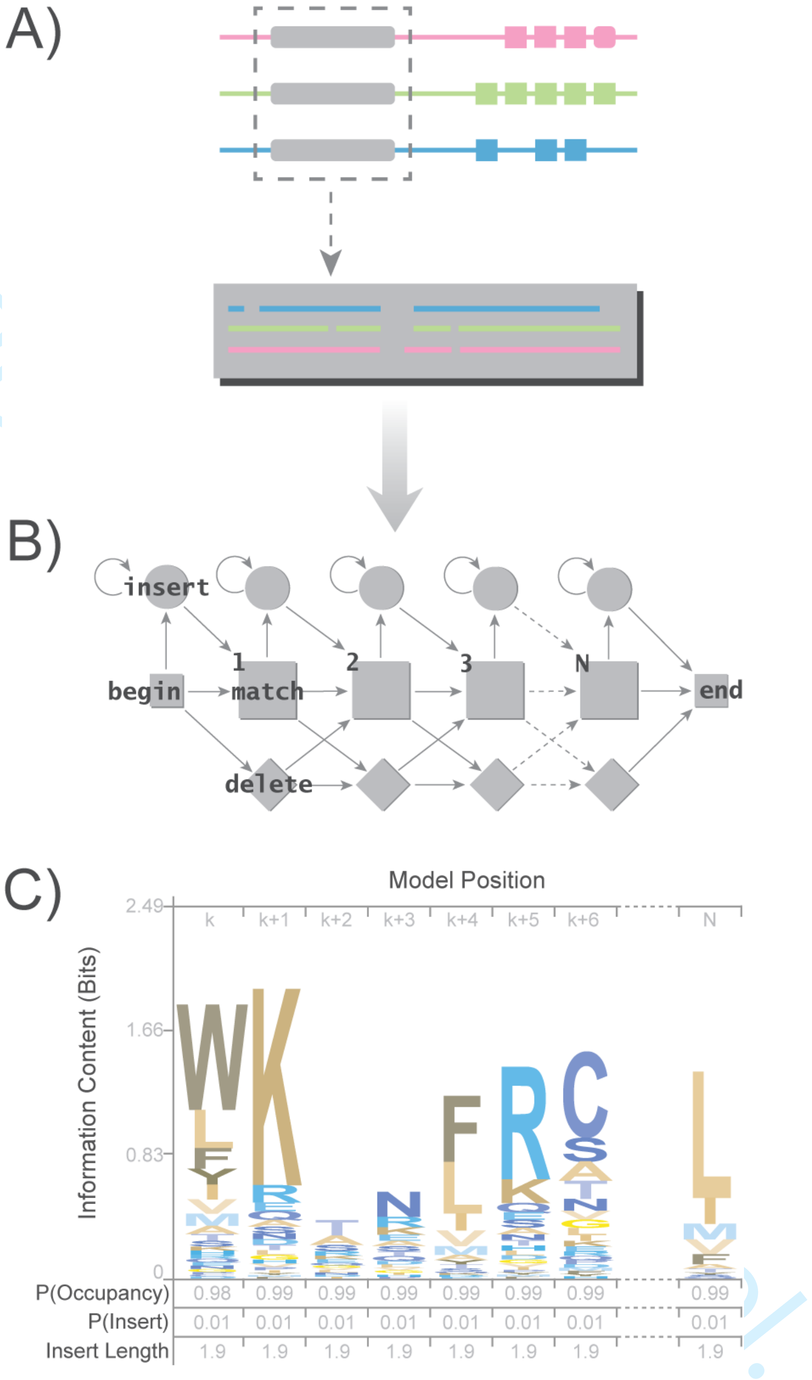
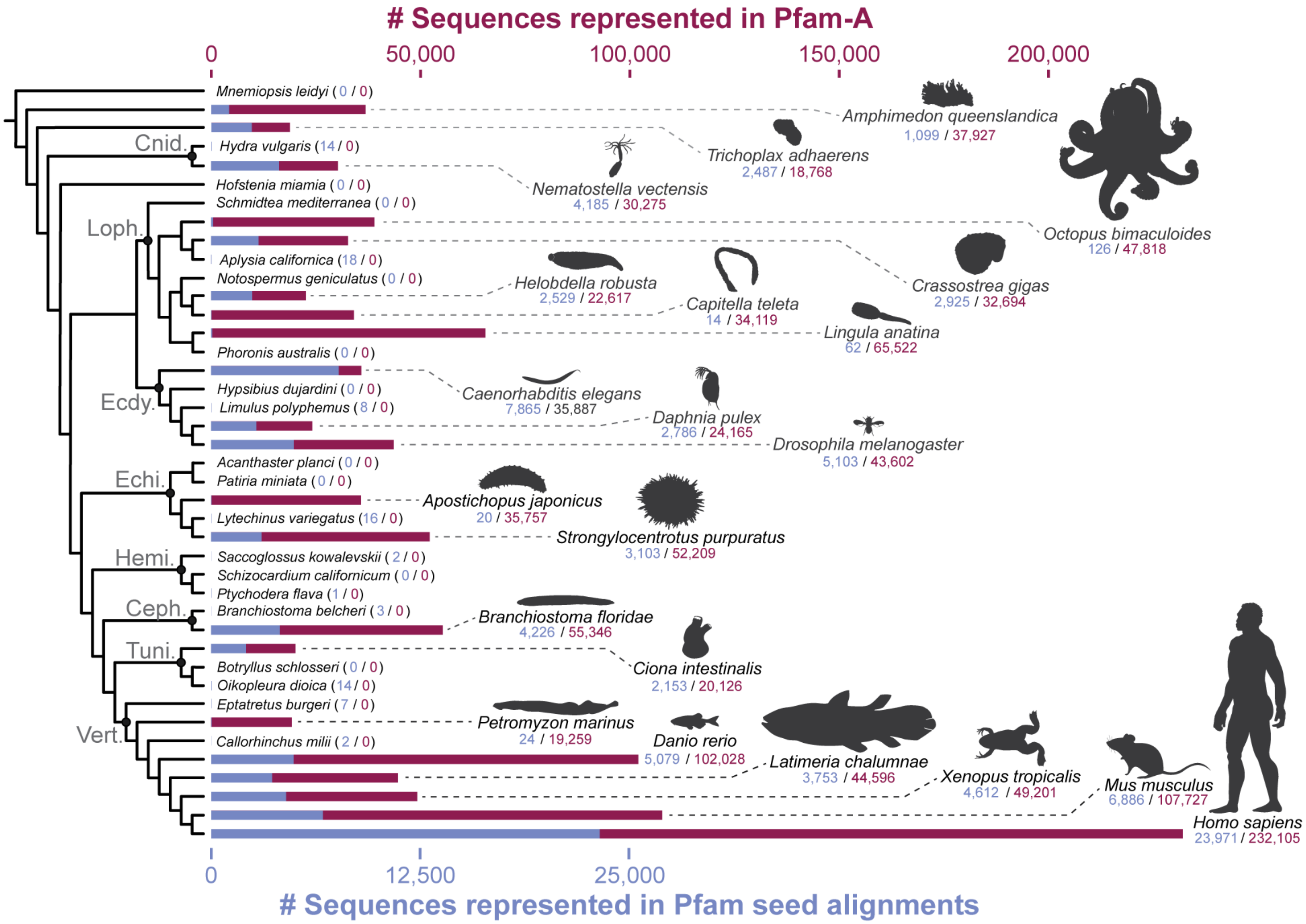


Figure 1



# Sequences represented in Pfam-A
0
50,000
100,000
150,000
200,000
Mnemiopsis leidyi (0 / 0)
Amphimedon queenslandica
1,099 / 37,927
Trichoplax adhaerens
2,487 / 18,768
Cnid.
Hydra vulgaris (14 / 0)
Nematostella vectensis
4,185 / 30,275
Hofstenia miamia (0 / 0)
Schmidtea mediterranea (0 / 0)
Loph.
Octopus bimaculoides
126 / 47,818
Aplysia californica (18 / 0)
Crassostrea gigas
2,925 / 32,694
Notospermus geniculatus (0 / 0)
Helobdella robusta
2,529 / 22,617
Capitella teleta
14 / 34,119
Lingula anatina
62 / 65,522
Phoronis australis (0 / 0)
Ecdy.
Caenorhabditis elegans
7,865 / 35,887
Hypsibius dujardini (0 / 0)
Limulus polyphemus (8 / 0)
Daphnia pulex
2,786 / 24,165
Drosophila melanogaster
5,103 / 43,602
Echi.
Acanthaster planci (0 / 0)
Patiria miniata (0 / 0)
Apostichopus japonicus
20 / 35,757
Lytechinus variegatus (16 / 0)
Strongylocentrotus purpuratus
3,103 / 52,209
Hemi.
Saccoglossus kowalevskii (2 / 0)
Schizocardium californicum (0 / 0)
Ptychodera flava (1 / 0)
Ceph.
Branchiostoma belcheri (3 / 0)
Branchiostoma floridae
4,226 / 55,346
Tuni.
Ciona intestinalis
2,153 / 20,126
Botryllus schlosseri (0 / 0)
Oikopleura dioica (14 / 0)
Vert.
Eptatretus burgeri (7 / 0)
Petromyzon marinus
24 / 19,259
Callorhinchus milii (2 / 0)
Danio rerio
5,079 / 102,028
Latimeria chalumnae
3,753 / 44,596
Xenopus tropicalis
4,612 / 49,201
Mus musculus
6,886 / 107,727
Homo sapiens
23,971 / 232,105
0
12,500
25,000
# Sequences represented in Pfam seed alignments

Figure 2

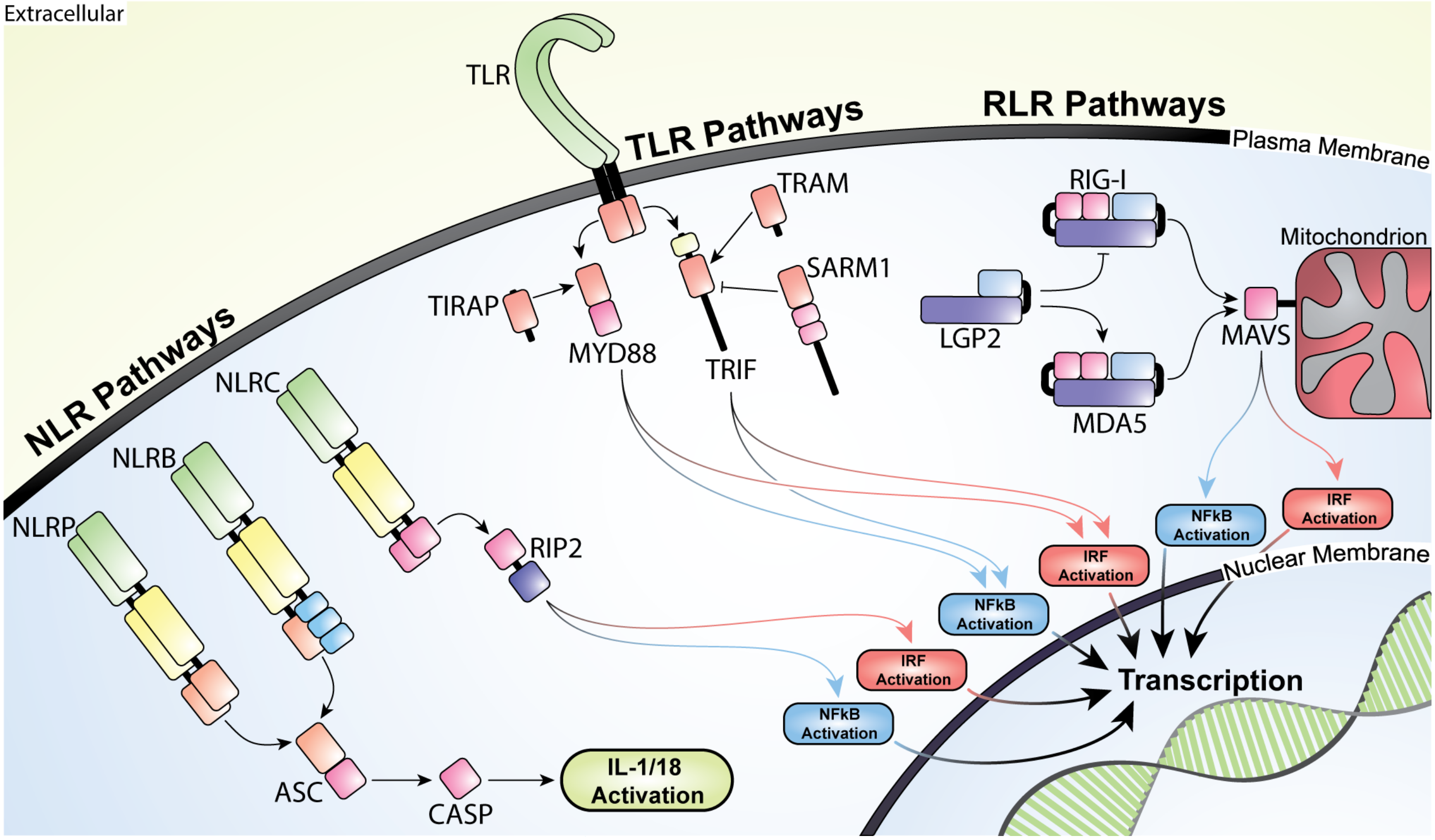

Extracellular
TLR
TLR Pathways
RLR Pathways
Plasma Membrane
TRAM
RIG-I
Mitochondrion
SARM1
TIRAP
MYD88
TRIF
LGP2
MAVS
MDA5
NLR Pathways
NLRC
NLRB
NLRP
RIP2
NFkB Activation
IRF Activation
Nuclear Membrane
Transcription
ASC
CASP
IL-1/18 Activation

Figure 3

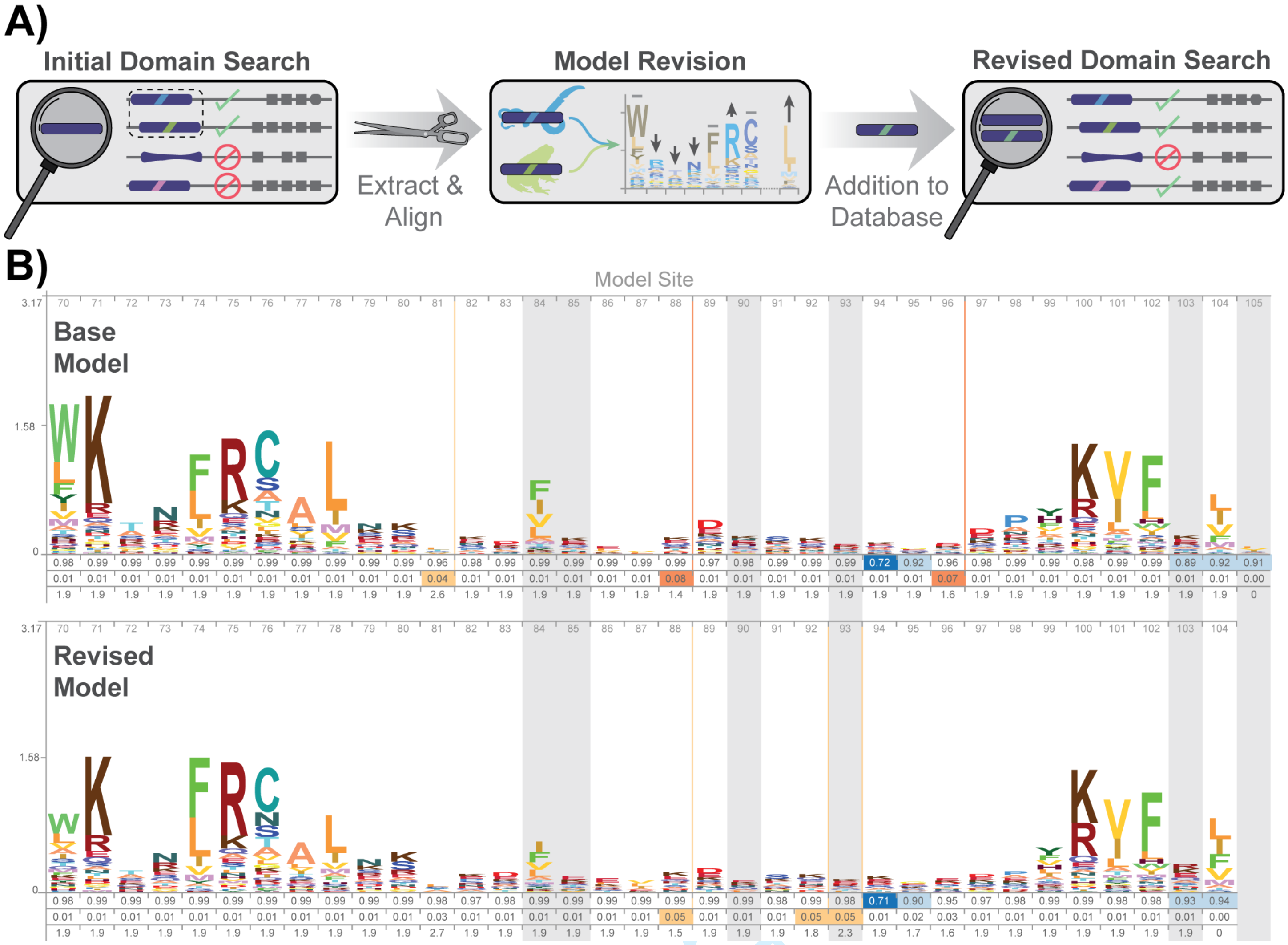

A)
Initial Domain Search
Extract &
Align
Model Revision
Addition to
Database
Revised Domain Search
B)
Model Site
3.17
1.58
0
70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105
Base
Model
0.98 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.96 0.98 0.99 0.99 0.99 0.99 0.99 0.99 0.97 0.98 0.99 0.99 0.99 0.72 0.92 0.96 0.98 0.99 0.99 0.98 0.99 0.99 0.89 0.92 0.91
0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.04 0.01 0.01 0.01 0.01 0.01 0.01 0.08 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.07 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.00
1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 2.6 1.9 1.9 1.9 1.9 1.9 1.9 1.4 1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.6 1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 0
Revised
Model
0.98 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.98 0.97 0.98 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.98 0.99 0.98 0.71 0.90 0.95 0.97 0.98 0.99 0.99 0.98 0.98 0.93 0.94
0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.01 0.01 0.01 0.01 0.01 0.01 0.05 0.01 0.01 0.01 0.05 0.05 0.01 0.02 0.03 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.00
1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 1.9 2.7 1.9 1.9 1.9 1.9 1.9 1.9 1.5 1.9 1.9 1.9 1.8 2.3 1.9 1.7 1.6 1.9 1.9 1.9 1.9 1.9 1.9 1.9 0

Figure 4

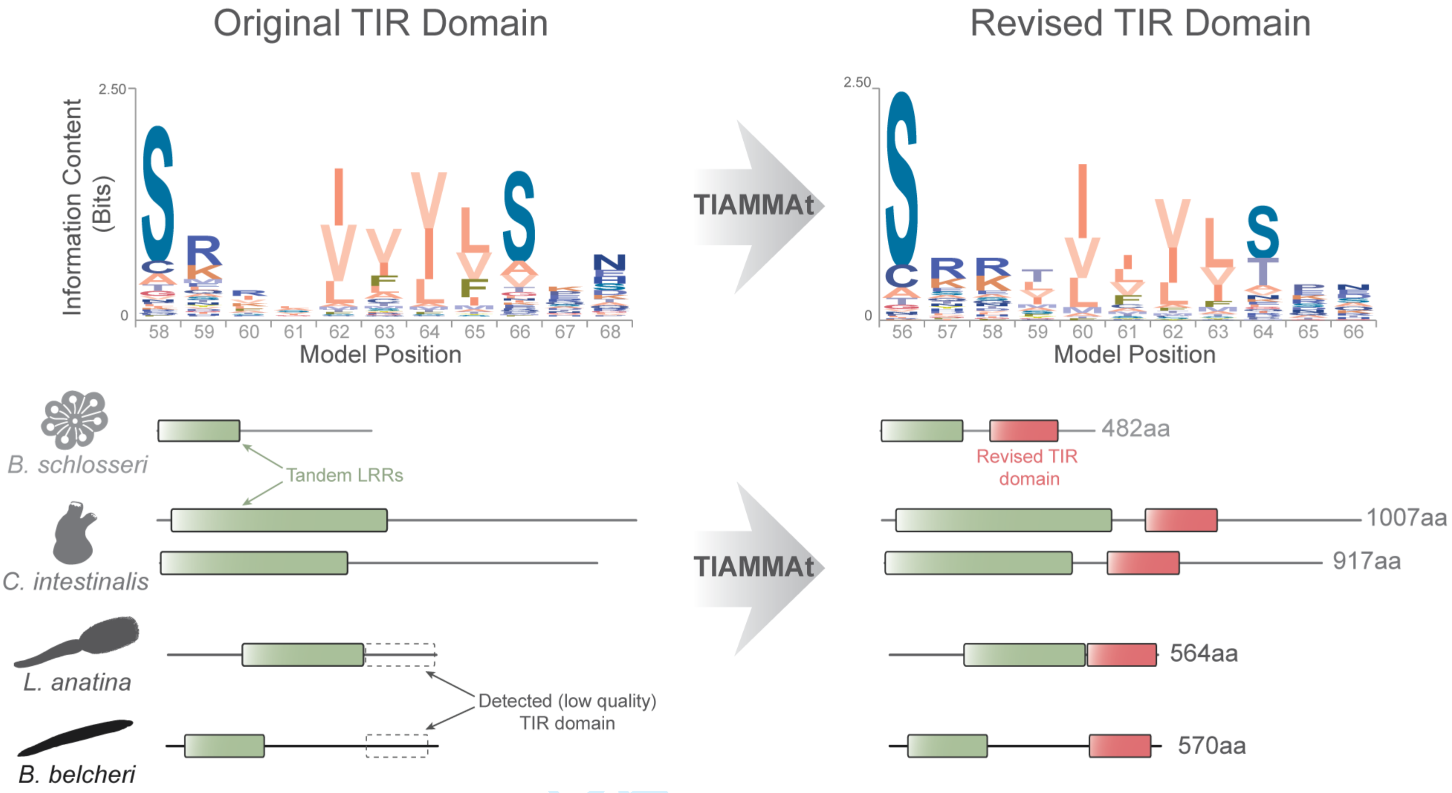

Original TIR Domain
Revised TIR Domain
Information Content (Bits)
2.50
0
58
59
60
61
62
63
64
65
66
67
68
Model Position
TIAMMAt
2.50
0
56
57
58
59
60
61
62
63
64
65
66
Model Position
B. schlosseri
Tandem LRRs
C. intestinalis
L. anatina
Detected (low quality) TIR domain
B. belcheri
TIAMMAt
482aa
Revised TIR domain
1007aa
917aa
564aa
570aa

Figure 5