Ancient Origins of Vertebrate-Specific Innate Antiviral Immunity

Krishanu Mukherjee,[1] Bryan Korithoski,[1] and Bryan Kolaczkowski*[,1]
[1]Department of Microbiology and Cell Science, University of Florida
***Corresponding author:** E-mail: bryank@ufl.edu.
Associate editor: Naoko Takezaki

Abstract

Animals deploy various molecular sensors to detect pathogen infections. RIG-like receptor (RLR) proteins identify viral RNAs and initiate innate immune responses. The three human RLRs recognize different types of RNA molecules and protect against different viral pathogens. The RLR protein family is widely thought to have originated shortly before the emergence of vertebrates and rapidly diversified through a complex process of domain grafting. Contrary to these findings, here we show that full-length RLRs and their downstream signaling molecules were present in the earliest animals, suggesting that the RLR-based immune system arose with the emergence of multicellularity. Functional differentiation of RLRs occurred early in animal evolution via simple gene duplication followed by modifications of the RNA-binding pocket, many of which may have been adaptively driven. Functional analysis of human and ancestral RLRs revealed that the ancestral RLR displayed RIG-1-like RNA-binding. MDA5-like binding arose through changes in the RNA-binding pocket following the duplication of the ancestral RLR, which may have occurred either early in Bilateria or later, after deuterostomes split from protostomes. The sensitivity and specificity with which RLRs bind different RNA structures has repeatedly adapted throughout mammalian evolution, suggesting a long-term evolutionary arms race with viral RNA or other molecules.

Key words: **RIG-like receptor, antiviral immunity, RNA-binding protein, protein family evolution.**

Introduction

Detecting and responding to foreign RNA in the cytoplasm is a major component of innate antiviral immunity. RIG-like receptors (RLRs) are related proteins that identify viral RNAs and initiate downstream immune responses through direct interactions with the mitochondrial antiviral signaling protein (MAVS) (fig. 1*A*). Humans and other mammals have three RLRs, which collectively recognize a wide variety of viral infections (Yoneyama et al. 2005; Kato et al. 2006). Polymorphisms in RLRs have been linked to autoimmune diseases (Smyth et al. 2006; Colli et al. 2010; Aida et al. 2011), and therapeutic activation of RLRs is being evaluated as a potential antiviral and anticancer approach (Poeck et al. 2008). Given the biological importance of RLRs and their therapeutic potential, RLR function has received considerable recent attention.

RLRs recognize viral RNAs and activate downstream immune responses using a modular domain architecture (fig. 1*B*). A C-terminal RNA recognition domain (RD) binds viral RNA (Li, Ranjith-Kumar, et al. 2009; Takahasi et al. 2009), inducing a protein conformational shift, which allows twin N-terminal caspase activation and recruitment domains (CARDs) to interact with the signal-transducing protein MAVS and ultimately activate cellular immune responses (Kawai et al. 2005; Meylan et al. 2005; Seth et al. 2005; Jiang et al. 2011; Luo et al. 2011).

Although the RNA-binding properties of human RLRs have received considerable attention (Pichlmair et al. 2006; Li, Lu, et al. 2009; Li, Ranjith-Kumar, et al. 2009; Pippig et al. 2009; Shigemoto et al. 2009; Lu et al. 2010; Wang et al. 2010; Jiang et al. 2011), few studies have directly compared RNA binding by different RLRs under the same experimental conditions, making it difficult to synthesize existing results to produce a clear picture of the functional differences among RLRs.

The three RLRs present in humans and other mammals do appear to differ in their RNA-binding sensitivity and specificity. RIG-1 (*DDx58*) is a high-affinity receptor for short, blunt-ended double-stranded RNA (dsRNA) and 5′ triphosphate (5′ppp) RNA (Hornung et al. 2006). MDA5 (*IFIH1*) appears to recognize long blunt-ended dsRNA, although its native ligand remains controversial (Li, Lu, et al. 2009). LGP2 (*DHx58*) seems to bind both blunt-ended and 5′ppp dsRNA with high affinity (Pippig et al. 2009) and regulates RIG-1 and MDA5 through an unknown mechanism (Murali et al. 2008; Satoh et al. 2010).

Structural and mutational studies suggest that differences in RNA binding among human RLRs primarily result from differences in the overall shape and distribution of electrostatic forces across the RNA-binding pocket in the C-terminal RD as well as differences in specific RNA-contacting residues lining the pocket. RIG-1 and LGP2 share an open RNA-binding pocket with a large positively charged binding surface, although the shape of the pocket and the RNA-contacting residues differ between the two molecules (Lu et al. 2011). By contrast, the binding pocket of MDA5 appears to have a more compact RNA-binding surface (Li, Lu, et al. 2009). In

Article

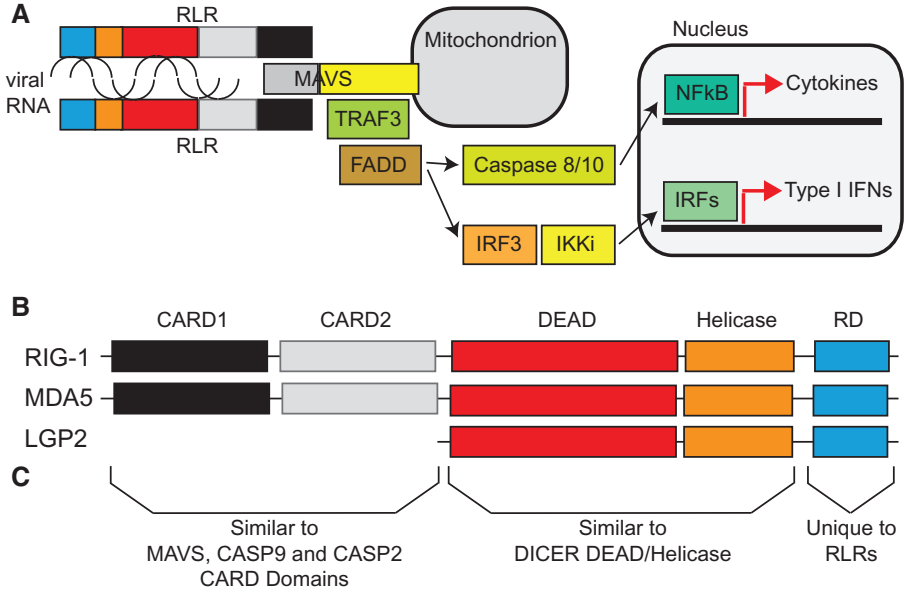

Fig. 1. RLRs recognize viral RNA and activate downstream signaling molecules through a modular domain architecture originating from multiple evolutionary sources. (*A*) Simplified diagram of the molecular interactions involved in RLR-based immune signaling. RLRs bind viral RNAs and interact with MAVS through CARDs (gray). MAVS forms a complex with TRAF3 and FADD, which induces apoptotic cell death through Caspase 8/10 activation as well as type 1 IFN and proinflammatory cytokine production through activation of IRF3/4 and NFkB, respectively (see Loo and Gale 2011 for a review). (*B*) We plot the characteristic domain architecture of human RIG-1, MDA5, and LGP2. The C-terminal RNA RD binds viral RNA (Li, Lu, et al. 2009; Takahasi et al. 2009), while N-terminal CARDs activate MAVS signaling, ultimately generating the immune response (Kawai et al. 2005; Meylan et al. 2005; Seth et al. 2005). LGP2 lacks CARDs and has been recently shown to regulate both RIG-1 and MDA5, although the regulatory mechanism is unclear (Satoh et al. 2010). (*C*) We identified the closest non-RLR homologs of each functional domain via sequence similarity; RD homology was not detectable outside our identified RLR protein family.

all three proteins, binding of different RNA molecules is facilitated by key contact residues, many of which are located within the RNA-binding loop between β5 and β6 (Cui et al. 2008; Li, Ranjith-Kumar, et al. 2009; Pippig et al. 2009; Takahasi et al. 2009). Differences in contact residues among RIG-1, MDA5, and LGP2 are thought to confer differential RNA-binding properties (Li, Lu, et al. 2009).

Evolutionary studies have painted a complex picture of how RLRs arose and functionally diversified. Recent studies suggest that full-length RLRs are a vertebrate-specific evolutionary novelty, although the building blocks of RLRs may have been present in closely related prevertebrate animals (Sarkar et al. 2008; Zou et al. 2009). These studies differ in the ordering of gene-duplication events giving rise to the three RLRs present in mammals, but both studies ultimately conclude that RLR evolution was driven by a complex series of domain-grafting and gene-fusion events. However, these studies relied on phylogenetic methods known to be sensitive to long-branch attraction artifacts, potentially undermining their reliability (Swofford et al. 2001; Susko et al. 2004).

Here, we examine the functional evolution of RLRs using a combination of gene modeling, phylogenetics, structural analysis, population genetics, and molecular-functional characterization of ancestral and extant RLRs. In contrast to previous studies, we find that the RLR-based immune system is not vertebrate specific but originated in the earliest multicellular animals. RLRs functionally diversified through a series of gene duplication events followed by protein-coding changes, which modulated the RNA-binding properties of different RLRs by altering key contact residues within the C-terminal RD. Finally, we find strong evidence that RLRs have been involved in a long-term evolutionary arms race with viral RNA molecules, suggesting that the structure of viral RNA may have shaped the evolution of animal innate immunity.

Results and Discussion

Origins of RLR-Based Innate Immunity

The availability of complete genome sequences from many eukaryotes provides an opportunity to systematically examine the origins of protein families currently thought to have limited taxonomic distributions. Using homology-based gene prediction based on confirmed RLRs from humans, we identified full-length RLRs in early branching animal genomes, including Porifera and Cnidaria (supplementary fig. S1, Supplementary Material online). RLRs are absent from nonmetazoan eukaryotes, including fungi and choanoflagellates, arguing that RLRs are an animal-specific evolutionary novelty arising with the origin of multicellularity. Although we identified RLRs from multiple lophotrocozoans and ecdysozoans, we were unable to recover any RLR genes from arthropods, despite the availability of multiple genome sequences, suggesting that RLRs were lost early in arthropod evolution.

Multiple independent lines of evidence suggest that newly identified RLRs are functional. First, gene models were predicted with high confidence, arguing against gene modeling error as an explanation for our results (supplementary table S1, Supplementary Material online). Second, inferred RLR

protein sequences had the same global domain architecture as known RLRs, and multiple sequence alignment confirmed the presence of conserved functional motifs, arguing that newly identified genes are in fact RLRs and not distantly related homologs (supplementary table S2 and alignments, Supplementary Material online). Finally, we found many near-exact matches to new RLRs in expressed sequence tag (EST) databases, suggesting that full-length RLRs are expressed in early animal lineages, including Porifera and Cnidaria (supplementary table S3, Supplementary Material online). Together, these results strongly support the conclusion that functional RLRs are present throughout nonvertebrate animals.

We used sensitive sequence-similarity searches to identify the closest non-RLR homologs of each RLR functional domain (Boratyn et al. 2012). We found that the RLR DEAD/Helicase domain was closely related to DICER, whereas the RLR CARD domains were distantly related to CARDs from MAVS and caspases 9 and 2 (fig. 1*C*). The RLR RD did not exhibit sequence similarity to any proteins outside the RLR family, suggesting that it is either a novel functional domain or so greatly diverged from its homologs as to be undetectable by sequence similarity. These results suggest that the RLR domain architecture probably arose through a fusion of a DICER-like DEAD/Helicase with CARD domains from a MAVS- or caspase-like progenitor.

RLRs do not function in isolation; they initiate cellular immune responses through interactions with downstream signaling molecules (Yoneyama et al. 2004). We found that many of the proteins involved in RLR-based immunity—such as the signaling proteins TRAF3 and FADD—are also present in nonvertebrate animals, including Porifera and Cnidaria (supplementary table S4, Supplementary Material online). Although the immediate target of RLR signaling, MAVS, is too diverged to detect using sequence-homology-based methods, this gene has been identified in the sea urchin (*Strongylocentrotus purpuratus*), suggesting that it predates vertebrates (Hibino et al. 2006). A recent study has also shown that RLRs can activate cellular immune responses independent of MAVS signaling (Poeck et al. 2010), suggesting that functional RLR-based immune pathways may exist in organisms that lack the MAVS gene. We did not find any evidence for canonical RLR-pathway effector molecules (type 1 interferons or proinflammatory cytokines) outside of vertebrates, suggesting that this part of the RLR pathway was a recent innovation, possibly facilitating interactions with the emerging adaptive immune system. However, our findings do suggest that many of the necessary components for a simplified form of RLR-based immunity were present in the earliest multicellular animals and not a later, vertebrate-specific novelty.

Evolutionary History of RLRs

Figure 2 shows the RLR consensus phylogeny obtained using a variety of alignment strategies and phylogenetic inference methods. This tree suggests that two primary gene duplication events gave rise to the three classes of RLRs found in mammals. First, the ancestral RLR (RIG-1/MDA5/LGP2anc) duplicated in Bilateria to give rise to RIG-1 and MDA5/LGP2 lineages, followed by a more recent duplication of the MDA5/LGP2 ancestor, giving rise to MDA5 and LGP2 lineages in jawed vertebrates after their split from jawless vertebrates. This later MDA5/LGP2 duplication event could be related to the proposed whole-genome duplications early in the vertebrate lineage (Dehal and Boore 2005).

According to the consensus phylogeny—in which the ancestral RLR duplicated early in Bilateria—both RIG-1 and MDA5/LGP2 were lost from arthropods, and the MDA5/LGP2 lineage was also lost from other protostomes, including nematodes. However, we remain cautious in our interpretation of this finding, as long-branch attraction between fast-evolving protostome sequences and deuterostome RIG-1 is a potential explanation for this result. An alternative interpretation suggested by the taxonomic distribution of RLR sequences places the first duplication event early in the deuterostome lineage after its divergence from protostomes. This interpretation is more parsimonious in terms of gene loss events, requiring only a single loss of the ancestral RLR in arthropods to explain the observed sequence distribution.

We observed various lineage-specific losses of different RLR genes, suggesting that RLR loss may be common. For example, although MDA5 and LGP2 homologs were found in many teleost fishes, clear RIG-1 homologs have only been identified from salmon and carp, suggesting that RIG-1 was lost from some fish species (Biacchesi et al. 2009). RIG-1 is also missing from the chicken genome, although MDA5 and LGP2 are both present. The sea squirt (*Ciona intestinalis*) exhibited the opposite pattern, encoding two RIG-1 homologs but lacking MDA5 and LGP2, even though closely related species have multiple RIG-1 and MDA5/LGP2 genes (fig. 2). Previous studies have noted the small size of the sea squirt genome, with many protein families appearing extremely simplified compared with other species (Dehal et al. 2002). Together, these findings suggest that the specific content of RLR genes may be fairly labile across evolutionary distances, perhaps driven in part by complex interactions with adaptive immune systems arising very early in vertebrate evolution (Flajnik and Kasahara 2010).

Our consensus phylogeny was robust to alignment ambiguity and use of alternative evolutionary models (supplementary table S5, Supplementary Material online), with the same phylogeny being recovered using techniques robust to common evolutionary model violations (Lartillot and Philippe 2004; Kolaczkowski and Thornton 2008). Removal of fast-evolving protostome sequences did not alter the remaining tree topology (supplementary fig. S2*A*, Supplementary Material online). Although removal of these sequences eliminates any information that could confidently date the first RLR duplication event, it does indicate that the first duplication into RIG-1 and MDA5/LGP2 lineages is not affected by removal of fast-evolving lineages, suggesting that this grouping is probably not caused by long-branch attraction. Together, these results argue against strong phylogenetic biases as a major explanation for our results, as such errors are expected to affect different evolutionary models to different degrees (Sullivan and Swofford 2001; Huelsenbeck and

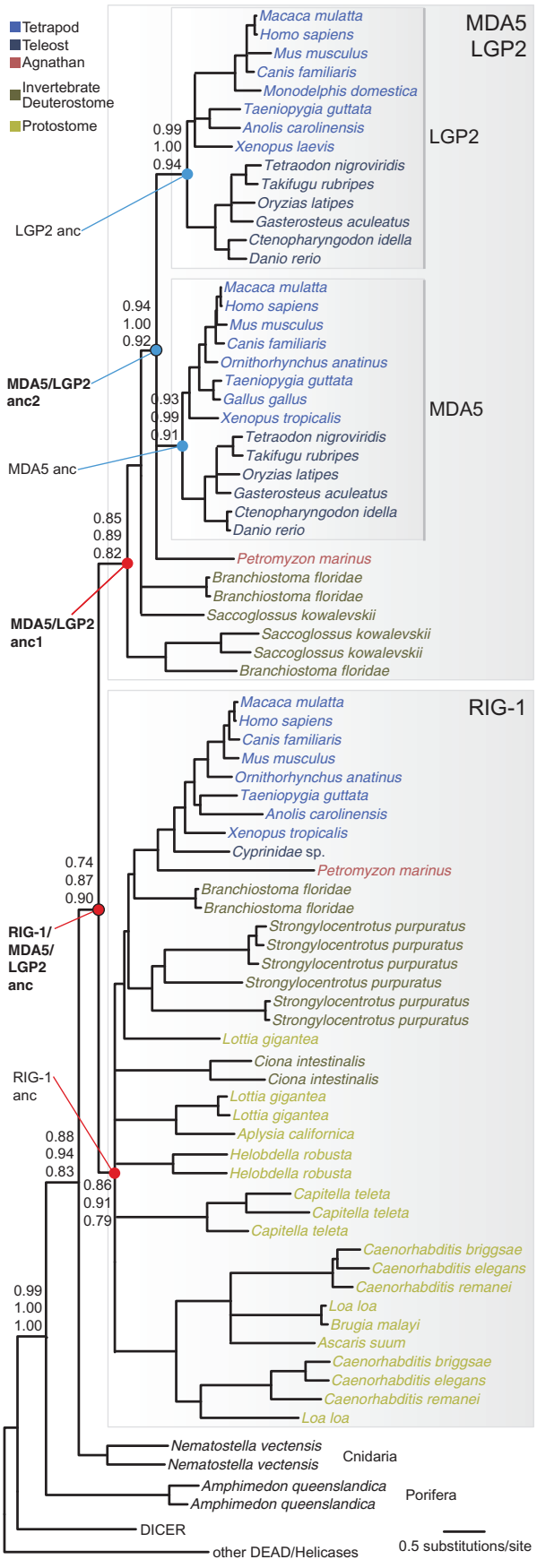

Fig. 2. RLRs evolved through gene duplications in Bilateria and vertebrates. We show the strongly supported RLR consensus tree calculated across a variety of alignment strategies and phylogenetic inference methods. Statistical support is shown for key nodes on the phylogeny, calculated using ML (SH-like aLRT [Anisimova et al. 2011]) (top) and Bayesian posterior probabilities using standard evolutionary models (middle) and the CAT model (Lartillot and Philippe 2004) (bottom). The tree is rooted using closely related DEAD/Helicase domains (Sarkar et al. 2008). Ancestral protein sequences were reconstructed for the nodes indicated on the tree. Species names are colored by taxonomic group.

Rannala 2004) and should be sensitive to removal of fast-evolving taxa (Aberer et al. 2013).

We rooted our RLR phylogeny using the closest-related protein family (supplementary fig. S3, Supplementary Material online). Although the outgroup branch was long, support for the consensus rooting increased with increasing alignment quality and evolutionary model complexity, arguing against phylogenetic bias, which is expected to affect simpler models more strongly than complex ones (supplementary table S6, Supplementary Material online). The same rooting was also recovered using a gene-species tree reconciliation method that does not rely on outgroup sequences, again arguing against long-branch attraction (supplementary fig. S4, Supplementary Material online).

Although many reasonable alignment methods are available, the choice of how to align sequences and how to postprocess alignment data to remove potentially unreliable positions can have a large impact on resulting clade support, as can the choice of which evolutionary model to use. For this reason, we assessed clade support using a "consensus" approach, which incorporated uncertainty about the sequence alignment, alignment processing and evolutionary model by calculating maximum likelihood (ML) and Bayesian support measures across a variety of different alignments and evolutionary models and then summarizing support on the consensus tree in figure 2. Support values calculated from each alignment are provided in supplementary table S7 (Supplementary Material online).

Consensus support for many of the key nodes on the phylogeny was high (fig. 2). The jawed-vertebrate-specific MDA5 and LGP2 clades were recovered with 0.93 and 0.99 Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-like aLRT) support under an ML framework (0.99 and 1.0 Bayesian posterior probability, respectively). The clade uniting MDA5 and LGP2, along with an RLR from the sea lamprey (*Petromynus marinus*) was also recovered with strong support (0.94 SH-like aLRT, 1.0 Bayesian posterior probability), as was the much older RIG-1 clade (0.86 SH-like aLRT, 0.91 Bayesian posterior probability).

However, support for some of the oldest nodes on the phylogeny was relatively low. The clade uniting vertebrate MDA5 and LGP2 with cephalochordate and hemichordate RLRs had 0.85 SH-like aLRT (0.89 Bayesian posterior probability), and the node uniting all MDA5, LGP2, and RIG-1 sequences was recovered with 0.74 SH-like aLRT (0.87 posterior probability). Lower support at these nodes is not unexpected, as reconstructing deep branching patterns

using limited data from fast-evolving immune receptors is a challenging phylogenetic problem. Nonetheless, we wanted to examine support at these key nodes in greater detail, to assess possible phylogenetic error.

Although consensus support for some of the oldest nodes on the phylogeny was low, support for all nodes increased with increasing alignment and evolutionary-model quality (supplementary table S7, Supplementary Material online). If the phylogeny was erroneous, we would expect support to decrease when potentially unreliable alignment positions were removed and/or parameters were added to the evolutionary model to better fit biological reality. That we observed the opposite effect—increasing support with increasingly reliable data and increasingly realistic evolutionary models—argues against phylogenetic bias or stochastic error as a major explanation for our results.

Furthermore, nearly all the phylogenetic ambiguity in our tree ultimately resulted from difficulty resolving short internodes associated with fast-evolving nonvertebrate sequences. Removal of these sequences completely resolved the tree with 100% support under any method, suggesting that ample phylogenetic signal exists to reconstruct the underlying core branching pattern, even though the precise placement of some fast-evolving sequences is ambiguous (supplementary fig. S2*B*, Supplementary Material online).

We additionally reconstructed the full RLR phylogeny using a concatenation of DNA and protein sequence alignments (Agosti et al. 1996). Both ML and Bayesian support for all key nodes on the phylogeny increased dramatically using this approach (supplementary fig. S5, Supplementary Material online). All but two of the nodes on the phylogeny had maximal support across all methods. The node uniting protostome and deuterostome RIG-1 sequences had maximal Bayesian posterior probability and between 0.97 and 1.0 ML bootstrap support, depending on whether the alignment was processed to remove potentially unreliable positions. The remaining node separating RIG-1, MDA5, and LGP2 sequences from Cnidarian and Poriferan RLRs also had maximal posterior probability and 0.87–0.92 bootstrap.

Although our consensus tree is supported by the bulk of phylogenetic analyses and appears robust to common sources of error, high levels of sequence divergence raise the real possibility that unaccounted-for factors may generate spurious phylogenetic groupings, particularly for deep nodes on the tree. Specifically, the timing of the first RLR duplication event—giving rise to RIG-1 and MDA5/LGP2 lineages—remains weakly supported in many analyses. Although Bayesian posterior probabilities supporting this node are typically high, concerns have been raised that Bayesian approaches may overestimate clade support (Suzuki et al. 2002; Misawa and Nei 2003; Simmons et al. 2004). Bootstrap support for this node was >0.8 in our concatenated analysis, which is typically considered acceptable, but concerns have been raised that this "rule of thumb" may not correspond to traditional expectations of statistical confidence (Efron et al. 1996), and other estimates of ML support are low in some analyses. Finally, reconstruction of the RLR protein family within a larger context of DEAD/Helicase-containing proteins suggests that the first RLR gene duplication may have occurred after the divergence of protostomes from deuterostomes (supplementary fig. S3, Supplementary Material online). We therefore remain cautious in our interpretation of the precise timing of the gene duplication giving rise to RIG-1 and MDA5/LGP2, which may have occurred early in Bilateria or after the protostome/deuterostome divergence.

Previous studies have suggested that the evolutionary history of RLRs is complex, with multiple gene-fusion and domain-grafting events (Sarkar et al. 2008; Zou et al. 2009). Contrary to these findings, we observed no evidence for phylogenetic incongruence among individual functional domains (supplementary table S8, Supplementary Material online). We suspect that previous results may have been compromised by incomplete taxon sampling and less-reliable phylogenetic methods. Our results suggest that, after the full RLR domain-architecture was assembled very early in the animal lineage, a simple model of gene duplication and functional divergence through protein-coding changes explains RLR evolution.

Functional Evolution of RLRs

Extant RLRs recognize different viruses and differ in their sensitivity and specificity for various RNA ligands (Hornung et al. 2006; Li, Lu, et al. 2009; Pippig et al. 2009; Loo and Gale 2011). These differences in RNA binding are primarily determined by differences in the C-terminal RNA-RD (Pippig et al. 2009). To better understand how the RNA-binding properties of RLRs functionally diversified, we reconstructed ancestral RLR proteins, inferred the 3D structures of ancestral RDs, and compared ancestral residues with known functional residues in extant proteins to predict the likely binding properties of ancient RLRs. These predictions were then examined experimentally by measuring RNA binding by ancestral and human RDs in vitro. Because RDs from different human RLRs are known to bind blunt ended and 5′ppp dsRNA with different affinities, and crystal structures are available for some of these interactions, we chose to focus our analyses on a comparison of these two RNA types.

We found that the ancestral RLR exhibited RIG-1-like RNA binding. Not only is the shape and electrostatic distribution of the ancestral RNA-binding pocket very similar to that of human RIG-1 (fig. 3), but also nearly all the key contact residues in human RIG-1 are conserved in both the ancestral RIG-1 and ancestral RLR, but not in other RLRs, including human MDA5, human LGP2, and their ancestors (fig. 4; supplementary fig. S6 and table S9, Supplementary Material online). The poriferan RLR has an RNA-binding pocket that is very similar in both shape and electrostatic distribution to that of human RIG-1, the ancestral RIG-1 and the ancestral RLR, further suggesting that the ancestral RLR bound blunt ended and 5′ppp RNA with high affinity, similar to human RIG-1 (Lu et al. 2011).

Consistent with our structural predictions, we observed that both human RIG-1 and the ancestral RLR bound blunt ended and 5′ppp dsRNA with equal affinities (fig. 3;

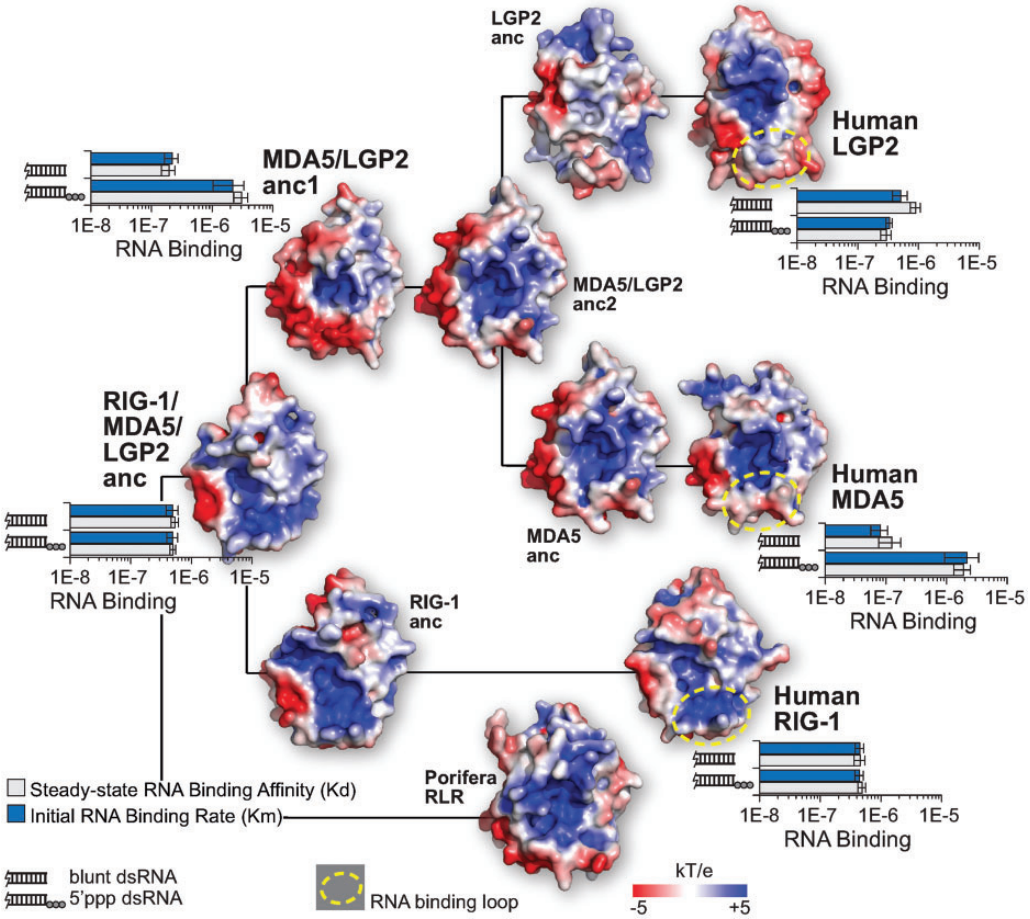

Fig. 3. Primary functional differences in RNA binding were established early in RLR evolution. We plot the RNA-binding pocket shape and electrostatic potential (kT/e) for human RIG-1, MDA5, and LGP2 as well as for reconstructed ancestral proteins (see fig. 2 for ancestral nodes on the phylogeny). Dotted yellow line indicates RNA-binding loop. Bar graphs display RLR concentrations at which half-maximal blunt-ended and 5′ppp dsRNA binding occurs (steady-state binding affinity [Kd] and initial RNA-binding rate [Km], measured using an in vitro kinetics assay). Note that lower Kd and Km indicate tighter binding. Bars indicate standard errors. See supplementary fig. S7 (Supplementary Material online) for complete binding curves.

supplementary fig. S7, Supplementary Material online). Human RIG-1 exhibited equivalent steady-state binding of blunt ended and 5′ppp RNA (Kd = 0.46 and 0.49 μM, respectively; two-sample *t*, $P = 0.38$), as did the ancestral RLR (Kd = 0.54 and 0.50 μM, respectively; $P = 0.34$).

These steady-state binding results were corroborated by our analysis of initial RNA-binding rates, in which blunt ended and 5′ppp dsRNA were bound with equal rates by both human RIG-1 (Km = 0.45 and 0.44 μM, respectively; $P = 0.48$) and the ancestral RLR (Km = 0.49 μM for both RNA types; $P = 0.49$). Remarkably, both the steady-state affinities and initial RNA-binding rates of the ancestral RLR and human RIG-1 were the same across these two RNA types (two-sample *t*, $P > 0.25$ across all comparisons), suggesting an amazing degree of molecular-functional conservation over millions of years of animal evolution.

After duplication of the ancestral RLR, a number of substitutions occurred in the MDA5/LGP2 ancestor to create an MDA5-like protein. Specifically, the MDA5/LGP2 ancestor had a much more acidic RNA-binding loop, compared with the ancestral RLR and RIG-1, and the distribution of positive charge across the RNA-binding pocket has shifted (fig. 3). These structural features are retained in human MDA5, but not in human RIG-1 or LGP2, suggesting that the MDA5/LGP2 ancestor likely exhibited MDA5-like RNA binding.

Indeed, functional analysis of the MDA5/LGP2 ancestral RD revealed a strong preference for blunt-ended over 5′ppp dsRNA compared with the ancestral RLR, characteristics that are retained in human MDA5 but not human RIG-1 or LGP2 (fig. 3; supplementary fig. S7, Supplementary Material online). The MDA5/LGP2 ancestor had slightly stronger steady-state affinity and faster initial binding rates for blunt-ended dsRNA, compared with the ancestral RLR (two-sample *t*, $P = 0.006$ and 0.038, respectively). At the same time, the MDA5/LGP2 ancestor weakened its 5′ppp RNA-binding affinity compared with the ancestral RLR by nearly an order of magnitude ($P < 0.000002$). These binding properties of the MDA5/LGP2 ancestor are retained in human MDA5, which binds blunt-ended dsRNA with about tenfold stronger affinity and higher initial rate, compared with 5′ppp dsRNA ($P < 0.01$). These results are consistent with at least one previous study of MDA5 RNA binding (Li, Lu, et al. 2009).

The physiological ligand of human MDA5 has remained mysterious, with recent studies suggesting that MDA5 may

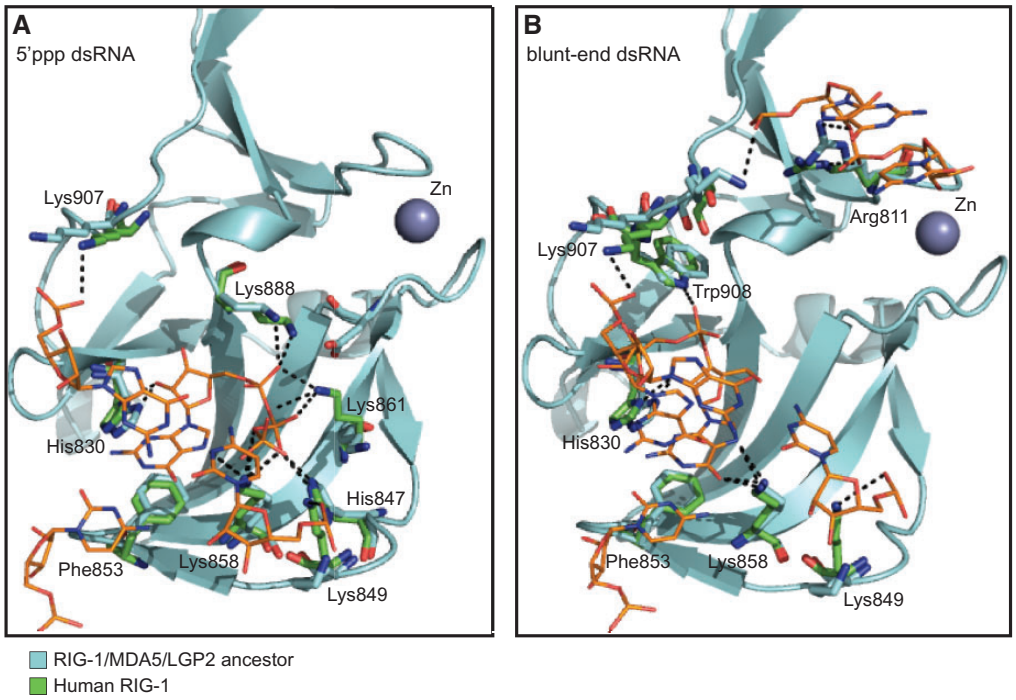

Fig. 4. The ancestral RLR binds 5′ppp and blunt-ended dsRNA. We used homology-based molecular docking to fit 5′ppp dsRNA (*A*) and blunt-ended dsRNA (*B*) into the binding pocket of the RIG-1/MDA5/LGP2 ancestral RLR (fig. 2). Contact residues are shown as sticks, with polar contacts indicated by black dotted lines. RNA contact residues found in human RIG-1 are shown in green and labeled (Lu et al. 2010, 2011).

cooperatively bind long, blunt-ended dsRNAs to produce immune-signaling "filaments" (Berke et al. 2012; Wu et al. 2013). Although our experiments using short dsRNAs do not directly address cooperative MDA5 binding, our finding that the human MDA5 RD—as well as the MDA5/LGP2 ancestral RD—prefer blunt-ended over 5′ppp dsRNA is consistent with previous results.

Although we do not know whether differences in blunt ended versus 5′ppp dsRNA binding are the only—or even the primary—functional differences separating human MDA5 and the MDA5/LGP2 ancestor from human RIG-1 and the RLR ancestor, our results do suggest that the RNA-binding sensitivity and specificity of the MDA5/LGP2 ancestor shifted immediately following its duplication from the ancestral RLR, and this functional shift was retained in human MDA5.

Although it may be tempting to dismiss relatively small changes in RNA-binding affinities as potentially "biologically irrelevant," that these changes are associated with large-scale structural differences conserved over millions of years of evolution argues against such an interpretation. Furthermore, small differences in an organism's ability to quickly recognize and respond to viral infections may translate into relatively large differences in pathogen sensitivity and, ultimately, fitness.

LGP2 arose from a jawed-vertebrate-specific gene duplication of the MDA5/LGP2 ancestor and immediately lost its twin N-terminal CARD signaling domains (supplementary fig. S8 and table S2, Supplementary Material online). Human LGP2 also appears to have evolved a slightly stronger steady-state affinity for 5′ppp over blunt-ended dsRNA (fig. 3; supplementary fig. S7, Supplementary Material online; two-sample *t*, $P = 0.01$), opposite of what we observed for both human MDA5 and the MDA5/LGP2 ancestor. However, this difference in binding affinity is small, and no significant difference in initial RNA-binding rate was observed ($P = 0.13$). Nonetheless, our results suggest that human LGP2's RNA-binding properties—particularly its tight binding of 5′ppp dsRNA—are much more similar to those of human RIG-1 than MDA5, which is consistent with previous findings (Li, Lu, et al. 2009).

Like human MDA5 and the MDA5/LGP2 ancestor, human LGP2 has a negatively charged RNA-binding loop, which can also be observed in the LGP2 ancestor and the MDA5 ancestor, but not in any of the RIG-1s or the ancestral RLR (fig. 3). The positively charged RNA-binding pocket of human LGP2 also appears strongly shifted compared with the other human and ancestral RLRs (fig. 3). These observations suggest that LGP2 binds its RNA ligands in a different conformation compared with RIG-1 and MDA5, which is consistent with structural comparisons of human LGP2 and RIG-1 (Lu et al. 2011). These results suggest that the similarities in RNA-binding sensitivity and specificity between human LGP2 and RIG-1 are the result of convergent evolution, in which LGP2 reevolved RIG-1-like RNA binding from an MDA5-like ancestor, but using different structural mechanisms.

Together, our results suggest that the major functional differences in RNA binding among extant RLRs were established early, following gene-duplication events. Structural modeling and functional characterization of ancestral and human RDs confirmed that protein-coding changes and

structural differentiation were associated with functional shifts in the RDs' sensitivity and specificity for blunt-ended versus 5′ppp dsRNA, suggesting that functional differentiation of RNA binding occurred through changes in the shape and electrostatic distribution of RLR RDs.

Because some of the oldest nodes on our phylogeny were not resolved with very high clade support under all analyses, we wanted to make sure our ancestral reconstructions did not depend on the tree topology. We therefore used an empirical Bayesian ancestral-reconstruction approach that integrated over topological uncertainty. Even when accounting for ambiguity in the underlying phylogeny, all ancestral protein sequences were reconstructed with very high confidence, and alternative reconstructions were not supported (supplementary fig. S9, Supplementary Material online). These findings suggest that our results are not strongly affected by ambiguity in either the phylogenetic tree or the reconstructed ancestral sequences. Our structural inferences also do not depend on the particular template used to build structural models of ancestral proteins, as we obtained equivalent results using a range of available structural templates (supplementary fig. S10, Supplementary Material online).

Recent Adaptation of RLRs in Mammals

Although major functional differentiation of RLRs occurred early, we found evidence that RLRs have continued to adapt to viral pathogens by altering their RNA-binding properties throughout mammalian evolution (supplementary fig. S11 and table S10, Supplementary Material online). For example, phylogenetic analysis of nonsynonymous/synonymous substitution rates identified a cluster of adaptive substitutions within the RNA-binding loop of mammalian RIG-1 (fig. 5). Few adaptive substitutions were identified outside the RNA-binding loop, and none occurred in functional domains other than the RD, suggesting that adaptation has specifically targeted the RNA-binding properties of RIG-1 during mammalian evolution.

The functional consequences of many of these substitutions are unknown, but mutational studies suggest that the Tyr853Phe substitution observed in primate RIG-1 is likely to affect RNA binding (Takahasi et al. 2009). A second adaptive Gly811Arg substitution in the human lineage is likely to increase affinity for blunt-ended dsRNA but not 5′ppp dsRNA, because Arg811 forms a hydrogen bond with the backbone of blunt-ended dsRNA but does not contribute to stabilizing bound 5′ppp dsRNA (Lu et al. 2011). The Gly811 residue found in mouse and macaque does not form this stabilizing hydrogen bond (fig. 5), suggesting either a strong evolutionary preference for increasing specific affinity for blunt-ended dsRNA in the human lineage or a compensating change due to substitutions at other positions in the molecule that may destabilize blunt-ended dsRNA binding.

Population-genomic analysis of resequencing data (Durbin et al. 2010) also suggests that RIG-1 may have experienced a recent selective sweep in humans (coalescent $P = 0.04$, fig. 6*A*). A reduction in polymorphism and negative skew in Tajima's *D* around the RIG-1 gene further supports this conclusion (Tajima 1989) (fig. 6*B*), as does analysis of single-nucleotide polymorphisms (SNPs) identified by the HapMap project (Altshuler et al. 2010) (supplementary fig. S12, Supplementary Material online).

Although the support in favor of a selective sweep is weak, and we are cautious regarding the ability of these methods to pinpoint the precise locations of selective sweeps (Nielsen et al. 2005), it is intriguing to note that the strongest signal for a selective sweep occurs in the region of the gene coding

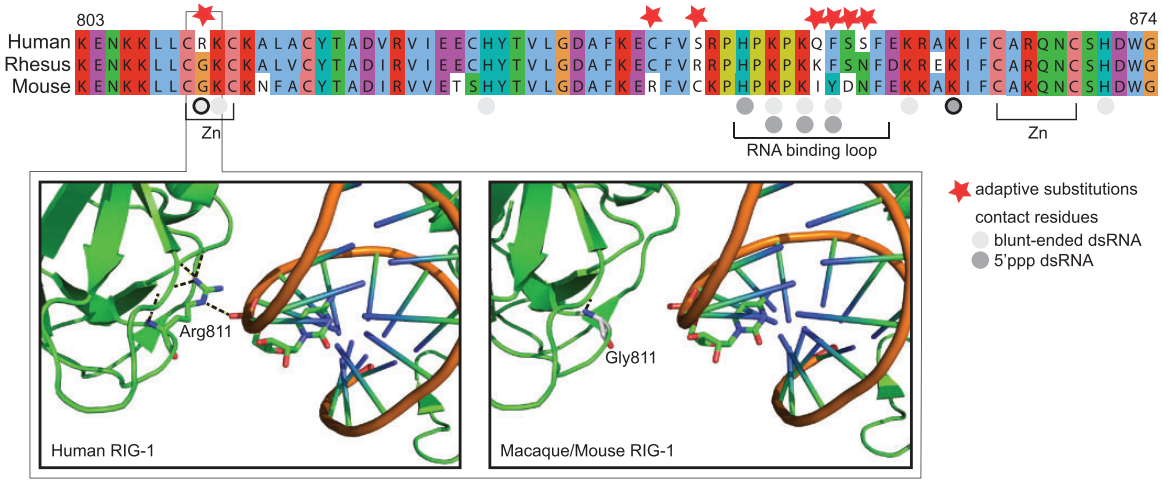

Fig. 5. Adaptive substitutions in mammalian RIG-1 affect RNA binding. We aligned the C-terminal RNA-RD of mammalian RIG-1s (human, rhesus, and mouse shown); zinc-finger motifs (Zn) and the RNA-binding loop are indicated along the bottom of the alignment. Red stars indicate strongly supported adaptive amino acid substitutions. Contact residues binding blunt-ended dsRNA (light gray) and 5′ppp dsRNA (dark gray) are also indicated, with bold outlines indicating residues required for RNA binding. Contact residues were identified from structural and mutational analyses of human RIG-1 (Lu et al. 2010, 2011). Inset shows the structural consequences of the adaptive Gly811Arg substitution observed in the human lineage. Panel at left shows the crystallized human RIG-1 bound to blunt-ended dsRNA (Lu et al. 2011). Polar contacts are indicated with dotted black lines. Right panel shows the corresponding region of macaque/mouse RIG-1. Numbering is based on the full-length human RIG-1 protein sequence.

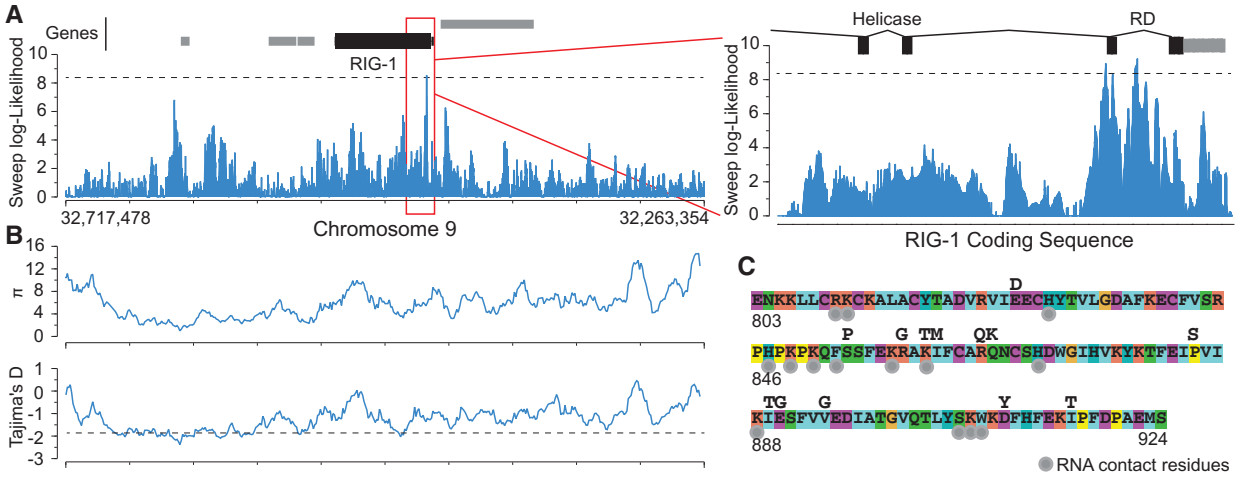

Fig. 6. Population-genomic analysis suggests a recent selective sweep in human RIG-1. (*A*) We used a spatially explicit composite-likelihood model to calculate support for a selective sweep in the region of human chromosome 9 surrounding RIG-1. We plot the log-likelihood support for a selective sweep across this region, with the dotted line indicating the 5% significance cutoff, calculated using coalescent simulations. Note that RIG-1 is coded on the negative DNA strand; we have organized the figure so that the negative strand reads from left to right. The corresponding protein functional domains can then be read from N-terminal (left) to C-terminal (right). Gray bars indicate other genes in the region. We also show a close-up of the RIG-1 coding sequence, with the exons coding for the RNA-RD and the C-terminal end of the Helicase domain indicated. Gray exons indicate the 3′-untranslated region. (*B*) We display sliding-window calculations of polymorphism (π) and Tajima's *D* across the same genomic region as in (*A*). Dotted line indicates lower bound of the 95% confidence interval for mean Tajima's *D* across the region. (*C*) We plot identified protein-coding polymorphisms across the RIG-1 RD (Sherry et al. 2001); RNA contact residues are identified by gray circles (Lu et al. 2010, 2011).

for the C-terminal RD, suggesting that adaptation of RIG-1 RNA binding may have occurred recently in humans.

Although there is no reason to presume that selection must be acting on coding variants, it is interesting to note that a number of protein-coding polymorphisms have been identified in the human RIG-1 RD (fig. 6*C*). Many of these polymorphisms induce radical changes in amino acid biochemical properties either at or nearby critical RNA-contacting residues, suggesting that the RNA-binding properties of RIG-1 may vary markedly across the human population. For example, recent exome resequencing has identified a low-frequency polymorphism (rs142758049, MAF = 0.023) in African Americans (Tennessen et al. 2012) that induces a Lys861Thr change in the RD, disrupting a key interaction stabilizing 5′ppp dsRNA (fig. 4). A second coding polymorphism (rs147393239) on the same African American haplotype changes Arg859 to Gly. Although Arg859 is not an RNA-contact residue, it does stabilize the RNA-binding loop through electrostatic interactions with Glu857 and Glu883 (supplementary fig. S13, Supplementary Material online); disruption of these stabilizing interactions is likely to strongly impact RD function.

In contrast to RIG-1, MDA5 and LGP2 RDs appear to have experienced less protein-coding adaptation across the mammalian phylogeny (supplementary fig. S11, Supplementary Material online), and support for adaptive sweeps within human MDA5 and LGP2 coding sequences was low (supplementary fig. S14, Supplementary Material online). However, we did observe a strong adaptive-sweep signature just upstream of human MDA5. Analyses performed by the ENCODE project have identified a number of transcription-factor binding sites within this region (Dunham et al. 2012), suggesting that MDA5 expression may have adapted recently in humans (supplementary fig. S14*A*, Supplementary Material online).

A recent study of protein-coding polymorphisms in human RLRs also identified a number of adaptive changes in various populations (Vasseur et al. 2011), further supporting the conclusion that RLRs have experienced recent adaptation and possible changes in protein function.

Conclusion

The availability of fully sequenced genomes from many species provides an unprecedented opportunity to systematically evaluate the origins and evolution of protein families and biochemical systems, shedding new light on the old question of how organismal complexity arose. Here, we have shown how an integrative analysis combining techniques from gene prediction, phylogenetics, structural modeling, molecular biochemistry, and population genetics can derive precise information about how the RLR protein family arose and functionally diversified, revealing new information about the origins and evolution of innate antiviral immunity.

Our analysis provides a broad outline of RLR protein evolution while highlighting some of the key events in the evolution of this gene family. Specifically, we found the following:

1) RLR-mediated viral-RNA sensing and response pathways—a major component of innate antiviral immunity—arose at the origin of multicellular animals and are not vertebrate specific, as long believed.
2) The ancestral RLR exhibited RIG-1-like RNA-binding properties, with binding affinities for both blunt ended and 5′ppp dsRNA remarkably similar to those of human RIG-1.

3) When the ancestral RLR duplicated to produce the RIG-1 and MDA5/LGP2 lineages, protein-coding changes in the MDA5/LGP2 ancestor reduced its affinity for 5′ppp dsRNA and increased its affinity for blunt-ended dsRNA, relative to the ancestral RLR. These RNA-binding properties are retained in human MDA5 but not in human RIG-1 and LGP2, suggesting that LGP2 reevolved RIG-1-like RNA binding from an MDA5-like ancestor.
4) RLRs appear to have accumulated adaptive changes altering their RNA-binding properties throughout mammalian evolution, including very recently in human evolution. Human polymorphisms in RIG-1 and other RLRs may contribute to differences in viral susceptibility and/or risk of autoimmune diseases.

Our work provides a general context for understanding RLR functional evolution across the metazoan phylogeny and a framework for elucidating the detailed mechanisms by which RLRs have functionally diversified. Although the biological ramifications of observed RNA-binding differences for RLR-based immune signaling must remain highly speculative at this point, our analysis suggests a model in which a "generalist" antiviral RNA-binding RLR ancestor arose very early in multicellular animals and functionally diversified by gene duplications to produce first a more specialized MDA5-like RLR and then a possibly regulatory LGP2-like molecule lacking direct CARD-signaling capabilities. Given this diversity of RLR functions, we suspect that a thorough understanding of RLR-based antiviral immunity will require not only further refinement of our understanding of specific RLR functions but also elucidation of how interactions among various RLRs may affect immune processes.

Materials and Methods

Gene Prediction

We predicted RLR genes using FGENESH + v2.6, which uses a homology-based hidden Markov model (HMM) framework to predict gene structures from genomic DNA (Salamov and Solovyev 2000). Regions of each genome potentially harboring RLR genes were identified using TBLASTN (Altschul et al. 1997) with human RLR proteins as query sequences. Gene predictions were then made from the identified genomic region ±3 kb. Support for predicted genes was calculated using log-odds ratios, which report the $\log_2$ of the probability of the genomic DNA sequence, given the HMM, divided by the probability of the genomic DNA under a random null model.

Predicted gene sequences were confirmed using four complimentary approaches. First, predicted genes were rejected if a PSI-BLAST search against experimentally verified gene structures in the NCBI database failed to produce a verified RLR gene as the top hit. Second, predicted genes were rejected if the rate of nonsynonymous/synonymous substitutions (d*N*/d*S*) was outside the range of d*N*/d*S* values for confirmed vertebrate RLRs. Third, predicted genes were rejected if their protein sequences did not have the same complete domain architecture as confirmed RLRs, calculated using a sequence search of the Pfam database with an *e*-value cutoff of 10^{-5} (Finn et al. 2010). Finally, transcription of predicted genes was confirmed using BLAST searches of species-specific EST databases, with an *e*-value cutoff of 10^{-10} (Boguski et al. 1993).

Sequence Alignment and Phylogenetic Inference

Sequence alignments were produced using MUSCLE v3.8.31 (Edgar 2004) and MAFFT v6.850b (Katoh and Toh 2008). In addition to full-length sequence alignments, we also used a domain-based alignment strategy. For this approach, each sequence was used to search the Pfam database (Finn et al. 2010) for matching functional domains, which were then aligned individually. Alignments were processed using Gblocks v0.91b at various stringency settings to remove potentially unreliable regions (Castresana 2000; Talavera and Castresana 2007).

For each alignment, the best-fit evolutionary model was identified using the AIC criterion in ProtTest v2.4 (Abascal et al. 2005). ML phylogenies were reconstructed using PhyML v3.0 (Guindon et al. 2010). We additionally reconstructed maximum-likelihood trees using a mixed branch length model, which incorporates evolutionary heterogeneity by calculating the probability of the alignment data, given a specified number of independently optimized branch length classes with corresponding mixing proportions (Kolaczkowski and Thornton 2008). The tree topology and all model parameters were optimized using simulated annealing, and the best-fit number of branch length classes for each alignment was inferred using AIC.

Bayesian analyses were performed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003), integrating over all available evolutionary models. We assumed gamma-distributed among-site rate variation and an estimated proportion of invariant sites. MrBayes analyses were conducted using default parameters, with runs terminating when the average standard deviation in clade posterior probabilities between two independent runs fell below 0.01.

We also conducted Bayesian analyses using the CAT model, implemented in PhyloBayes v3.3b (Lartillot and Philippe 2004; Lartillot et al. 2009). Conceptually, the CAT model incorporates site-specific heterogeneity in qualitative evolutionary constraints using a mixture of amino acid profiles. Recent analyses suggest that this approach may improve tree reconstruction accuracy compared with standard homogenous models (Lartillot et al. 2007).

We constructed a consensus tree using PhySIC_IST v1.0.1 (Scornavacca et al. 2008), which identifies consistently supported clades across a set of input phylogenies, taking into account support values estimated on each input tree. Weakly supported clades (<0.8 SH-like aLRT [Anisimova et al. 2011] or Bayesian posterior probability) were collapsed on each input tree. The same consensus tree was obtained using a modified minimum-cut approach, which has been shown to be among the best-performing supertree methods (Cotton and Wilkinson 2007; Buerki et al. 2011). Consensus support for individual clades was

calculated by averaging over all alignments and evolutionary models.

Ancestral Sequence Reconstruction and Structural Analysis

Ancestral protein sequences were reconstructed using an empirical Bayesian method to integrate over all tree topologies inferred from ML and Bayesian analyses (Hanson-Smith et al. 2010). We collected all ML trees inferred from any sequence alignment and phylogenetic model as well as all trees in the 95% credible set from each Bayesian analysis and estimated the posterior probability of each topology using Bayes' rule. Nuisance parameters were optimized by ML, and we assumed a flat prior over topologies.

Given a topology, we inferred the marginal posterior-probability distribution over ancestral sequences at each node using PAML v4 (Yang 2007), which implements an empirical Bayesian ancestral reconstruction algorithm (Yang et al. 1995). Next, we integrated over topologies by weighting each ancestral reconstruction by the posterior probability of that tree. Taking the residue with the maximum integrated posterior probability at each alignment column gives the maximum a posteriori (MAP) estimate of the ancestral protein sequence, accounting for uncertainty in tree topology. Support for each sequence reconstruction takes into account uncertainty about the reconstruction, given the tree topology, as well as uncertainty about the topology. Insertions and deletions were inferred using maximum parsimony. We note that the MAP reconstruction was the same as the ML reconstruction, assuming the consensus tree shown in figure 2.

Structural homology models were built using MODELLER v9.9 (Fiser and Sali 2003). Template structures were selected using a sequence search of the RCSB Protein Data Bank (PDB; Rose et al. 2011). In addition, we built homology models of each sequence using crystallized human RIG-1 (3OG8), MDA5 (3GA3), and LGP2 (3EQT) RD domains as template structures. For each template structure and protein sequence, we constructed 10 homology models and selected the best model using the MODELLER objective function as well as DOPE and GA341 assessment scores (Eramian et al. 2008).

Structural models were processed using PROPKA and PDB2PQR to determine residue side-chain pKas, optimize the structure for favorable hydrogen bonding and calculate charge and radius parameters from electrostatic force fields (Li et al. 2005; Dolinsky et al. 2007). Electrostatic surface potentials were estimated from processed structures using APBS (Baker et al. 2001) and projected onto the molecular surface for visualization.

Functional Analysis of RNA Binding

We generated GC-rich 28-base-pair RNA molecules in vitro using T7 RNA transcriptase and synthetic DNA as template, which produces a 5′ppp RNA molecule. DNA template was removed by treatment with DNase, and RNAs were purified by chloroform isoamyl alcohol extraction followed by ethanol precipitation. Complementary purified single-stranded RNAs were annealed to produce dsRNA by combining at 1:1 ratio, heating to 95 °C for 5 min and then cooling to 25 °C. Blunt-ended dsRNA was produced from 5′ppp RNA by two serial reactions with alkaline phosphatase to remove the 5′ppp. The 3′-end of one RNA strand was biotinylated to facilitate kinetics assays using the Pierce 3′ End RNA Biotinylation Kit (Thermo). Biotinylation was confirmed, and RNA concentrations measured, by comparison with a standard dilution. RNAs were visualized on urea gel at each step in the generation process. Short, GC-rich dsRNAs with and without 5′ppp were used to facilitate direct comparisons with previous structural and biochemical studies (Li, Lu, et al. 2009; Li, Ranjith-Kumar, et al. 2009; Lu et al. 2010, 2011); previous work has suggested that base composition has little effect on RNA binding (Lu et al. 2010).

Human and ancestral RLR RDs were expressed in *Escherichia coli* Rosetta 2(DE3)pLysS cells using pET-22b(+) expression constructs, which were verified by Sanger sequencing. Proteins were purified using His-affinity purification and visualized by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) stained with 1% coomassie. Protein concentrations were measured via a linear-transformed Bradford assay (Zor and Selinger 1996).

We measured steady-state RD-RNA–binding affinities using a label-free in vitro kinetics assay at pH 7.0 (Abdiche et al. 2008). Biotinylated RNA molecules were bound to a series of eight streptavidin probes for 15 min, until saturation was observed. Probes were washed and then exposed to 25 μg/ml biocytin to bind any remaining free streptavidin. The probes were then exposed in parallel to RDs at various concentrations in 1× Kinetics Buffer (ForteBio) for 60 min, at which point binding curve saturation was observed. Kinetics buffer included BSA to control for nonspecific binding. Steady-state molecular binding at saturation was measured as the change in laser wavelength when reflected through the probe in solution. Measurements were taken from three experimental replicates, and we estimated the RD concentration at which ½-maximal steady-state RNA binding was achieved (Kd) by fitting a one-site binding curve using nonlinear regression.

Our analysis measured molecular binding at various concentrations over time, sampled every 3 ms (supplementary fig. S15, Supplementary Material online). We fit association curves to these data to estimate the initial rates of RNA binding across RD concentrations and used these rates to calculate the RD concentration at which the ½-maximal RNA-binding rate was achieved (Km).

Protein-Coding Adaptation and Adaptive Sweeps

Protein-coding adaptation was assessed using the branch-sites model implemented in PAML, which uses a mixture distribution to model a combination of negatively selected, neutral, and positively selected positions in the protein sequence (Zhang et al. 2005). For each branch on the phylogeny, we tested the hypothesis that some sites experienced adaptive protein-coding substitutions against the null hypothesis of neutral evolution using a likelihood ratio test. *P* values were calculated using the χ^2 distribution (Zhang et al.

2005), and multiple testing was corrected for using a Bonferroni correction. Because of possible biases introduced by saturation of silent-site substitutions with highly divergent sequences, we restricted our analysis of protein-coding adaptation to within the mammalian lineage.

Individual codons were identified as adaptive using Bayes empirical Bayes (BEB) posterior probabilities calculated by PAML. Briefly, BEB posterior probabilities are calculated for each codon by comparing the probability that the codon evolved adaptively in a given lineage to the sum of the probabilities that the codon evolved under negative, neutral, and positive selection, each weighted by the estimated proportion of sites evolving under that category. Uncertainty in parameter estimates is incorporated by integrating over diffuse prior distributions (Yang et al. 2005). Protein-coding substitutions with BEB posterior probability > 0.95 were considered adaptive. We confirmed adaptive substitutions by reconstructing the ancestral protein sequences at each node on the phylogeny and determining the branch on which each amino acid substitution occurred. Candidate adaptive substitutions were only accepted if the probability of the amino acid substitution was >0.95, given the protein-sequence reconstructions.

We identified adaptive sweeps using SNP frequencies estimated by the 1000 Human Genomes Project (Durbin et al. 2010) and from the combined HapMap 3 data (Altshuler et al. 2010). We excised the region of genomic DNA ± 300 kb surrounding each RLR gene and assessed support for an adaptive sweep using a composite likelihood ratio test (CLRT) (Nielsen et al. 2005). Briefly, the CLRT calculates the likelihood of the local site frequency spectrum (SFS) at a specific location in the genome under two models: 1) the background SFS calculated across the entire region and 2) a one-parameter model that induces a characteristic sweep-like skew in the background SFS. Support for the sweep model is reported as the log-likelihood ratio of the sweep model to the background SFS. We scanned each genomic region for adaptive sweeps sampling every 100 bp.

Significance was assessed using 100,000 coalescent simulations under a standard-neutral model, simulated conditional on the observed number of segregating sites and pattern of sequencing coverage in each region. For each simulated replicate data set, we calculated the log-likelihood ratio in favor of an adaptive sweep using the CLRT, producing a null distribution from which to estimate the *P* value of the observed log-likelihood ratio. This approach has been shown to be robust to population-size changes consistent with our current understanding of human demographic history (Nielsen et al. 2005).

We also calculated polymorphism (π) and Tajima's *D* (Tajima 1989) across each genomic region using a sliding window approach. We used a window size of 8 kb and a step size of 1 kb. We calculated the mean and 95% confidence interval for Tajima's *D* across each region.

Supplementary Material

Supplementary alignments, tables S1–S10, and figures S1–S15 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Takashi Fujita for providing human RLR constructs. This work was supported by the National Institutes of Health (NIAID) grant AI101571-02 and the State of Florida.

References

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104–2105.

Abdiche Y, Malashock D, Pinkerton A, Pons J. 2008. Determining kinetics and affinities of protein interactions using a parallel real-time label-free biosensor, the Octet. *Anal Biochem.* 377: 209–217.

Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst Biol.* 62:162–166.

Agosti D, Jacobs D, DeSalle R. 1996. On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case. *Cladistics* 12:65–82.

Aida K, Nishida Y, Tanaka S, et al. (18 co-authors). 2011. RIG-I- and MDA5-initiated innate immunity linked with adaptive immunity accelerates beta-cell death in fulminant type 1 diabetes. *Diabetes* 60:884–889.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Altshuler DM, Gibbs RA, Peltonen L, et al. (69 co-authors). 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.

Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol.* 60:685–699.

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A.* 98:10037–10041.

Berke IC, Yu X, Modis Y, Egelman EH. 2012. MDA5 assembles into a polar helical filament on dsRNA. *Proc Natl Acad Sci U S A.* 109: 18437–18441.

Biacchesi S, LeBerre M, Lamoureux A, Louise Y, Lauret E, Boudinot P, Bremont M. 2009. Mitochondrial antiviral signaling protein plays a major role in induction of the fish innate immune response against RNA and DNA viruses. *J Virol.* 83:7815–7827.

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST—database for "expressed sequence tags". *Nat Genet.* 4:332–333.

Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. 2012. Domain enhanced lookup time accelerated BLAST. *Biol Direct.* 7:12.

Buerki S, Forest F, Salamin N, Alvarez N. 2011. Comparative performance of supertree algorithms in large data sets using the soapberry family (Sapindaceae) as a case study. *Syst Biol.* 60:32–44.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.

Colli ML, Moore F, Gurzov EN, Ortis F, Eizirik DL. 2010. MDA5 and PTPN2, two candidate genes for type 1 diabetes, modify pancreatic beta-cell responses to the viral by-product double-stranded RNA. *Hum Mol Genet.* 19:135–146.

Cotton JA, Wilkinson M. 2007. Majority-rule supertrees. *Syst Biol.* 56: 445–452.

Cui S, Eisenacher K, Kirchhofer A, Brzozka K, Lammens A, Lammens K, Fujita T, Conzelmann KK, Krug A, Hopfner KP. 2008. The C-terminal regulatory domain is the RNA 5′-triphosphate sensor of RIG-I. *Mol Cell.* 29:169–179.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3:e314.

Dehal P, Satou Y, Campbell RK, et al. (87 co-authors). 2002. The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* 298:2157–2167.

Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA. 2007. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Res.* 35:W522–W525.

Dunham I, Kundaje A, Aldred SF, et al. (602 co-authors). 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A.* 93:13429–13434.

Eramian D, Eswar N, Shen MY, Sali A. 2008. How well can the accuracy of comparative protein structure models be predicted? *Protein Sci.* 17: 1881–1893.

Finn RD, Mistry J, Tate J, et al. (14 co-authors). 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211–D222.

Fiser A, Sali A. 2003. Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol.* 374:461–491.

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

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Hanson-Smith V, Kolaczkowski B, Thornton JW. 2010. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol.* 27:1988–1999.

Hibino T, Loza-Coll M, Messier C, et al. (16 co-authors). 2006. The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol.* 300:349–365.

Hornung V, Ellegast J, Kim S, et al. (12 co-authors). 2006. 5′-Triphosphate RNA is the ligand for RIG-I. *Science* 314:994–997.

Huelsenbeck J, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst Biol.* 53:904–913.

Jiang F, Ramanathan A, Miller MT, Tang GQ, Gale M Jr, Patel SS, Marcotrigiano J. 2011. Structural basis of RNA recognition and activation by innate immune receptor RIG-I. *Nature* 479:423–427.

Kato H, Takeuchi O, Sato S, et al. (18 co-authors). 2006. Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature* 441:101–105.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286–298.

Kawai T, Takahashi K, Sato S, Coban C, Kumar H, Kato H, Ishii KJ, Takeuchi O, Akira S. 2005. IPS-1, an adaptor triggering RIG-I- and Mda5-mediated type I interferon induction. *Nat Immunol.* 6: 981–988.

Kolaczkowski B, Thornton JW. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol Biol Evol.* 25: 1054–1066.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol.* 7:S4.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol.* 21:1095–1109.

Li X, Lu C, Stewart M, Xu H, Strong RK, Igumenova T, Li P. 2009. Structural basis of double-stranded RNA recognition by the RIG-I like receptor MDA5. *Arch Biochem Biophys.* 488:23–33.

Li X, Ranjith-Kumar CT, Brooks MT, Dharmaiah S, Herr AB, Kao C, Li P. 2009. The RIG-I-like receptor LGP2 recognizes the termini of double-stranded RNA. *J Biol Chem.* 284:13881–13891.

Li H, Robertson AD, Jensen JH. 2005. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* 61:704–721.

Loo YM, Gale M Jr. 2011. Immune signaling by RIG-I-like receptors. *Immunity* 34:680–692.

Lu C, Ranjith-Kumar CT, Hao L, Kao CC, Li P. 2011. Crystal structure of RIG-I C-terminal domain bound to blunt-ended double-strand RNA without 5′ triphosphate. *Nucleic Acids Res.* 39:1565–1575.

Lu C, Xu H, Ranjith-Kumar CT, Brooks MT, Hou TY, Hu F, Herr AB, Strong RK, Kao CC, Li P. 2010. The structural basis of 5′ triphosphate double-stranded RNA recognition by RIG-I C-terminal domain. *Structure* 18:1032–1043.

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

Meylan E, Curran J, Hofmann K, Moradpour D, Binder M, Bartenschlager R, Tschopp J. 2005. Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. *Nature* 437: 1167–1172.

Misawa K, Nei M. 2003. Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. *J Mol Evol.* 57(1 suppl), S290–S296.

Murali A, Li X, Ranjith-Kumar CT, Bhardwaj K, Holzenburg A, Li P, Kao CC. 2008. Structure and function of LGP2, a DEX(D/H) helicase that regulates the innate immunity response. *J Biol Chem.* 283: 15825–15833.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.* 15:1566–1575.

Pichlmair A, Schulz O, Tan CP, Naslund TI, Liljestrom P, Weber F, Reis e Sousa C. 2006. RIG-I-mediated antiviral responses to single-stranded RNA bearing 5′-phosphates. *Science* 314:997–1001.

Pippig DA, Hellmuth JC, Cui S, Kirchhofer A, Lammens K, Lammens A, Schmidt A, Rothenfusser S, Hopfner KP. 2009. The regulatory domain of the RIG-I family ATPase LGP2 senses double-stranded RNA. *Nucleic Acids Res.* 37:2014–2025.

Poeck H, Besch R, Maihoefer C, et al. (31 co-authors). 2008. 5′-Triphosphate-siRNA: turning gene silencing and Rig-I activation against melanoma. *Nat Med.* 14:1256–1263.

Poeck H, Bscheider M, Gross O, et al. (18 co-authors). 2010. Recognition of RNA virus by RIG-I results in activation of CARD9 and inflammasome signaling for interleukin 1 beta production. *Nat Immunol.* 11: 63–69.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Rose PW, Beran B, Bi C, et al. (15 co-authors). 2011. The RCSB protein data bank: redesigned web site and web services. *Nucleic Acids Res.* 39:D392–D401.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* 10:516–522.

Sarkar D, Desalle R, Fisher PB. 2008. Evolution of MDA-5/RIG-I-dependent innate immunity: independent evolution by domain grafting. *Proc Natl Acad Sci U S A.* 105:17040–17045.

Satoh T, Kato H, Kumagai Y, Yoneyama M, Sato S, Matsushita K, Tsujimura T, Fujita T, Akira S, Takeuchi O. 2010. LGP2 is a positive regulator of RIG-I- and MDA5-mediated antiviral responses. *Proc Natl Acad Sci U S A.* 107:1512–1517.

Scornavacca C, Berry V, Lefort V, Douzery EJ, Ranwez V. 2008. PhySIC_IST: cleaning source trees to infer more informative supertrees. *BMC Bioinformatics* 9:413.

Seth RB, Sun L, Ea CK, Chen ZJ. 2005. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. *Cell* 122:669–682.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29:308–311.

Shigemoto T, Kageyama M, Hirai R, Zheng J, Yoneyama M, Fujita T. 2009. Identification of loss of function mutations in human genes

encoding RIG-I and MDA5: implications for resistance to type I diabetes. *J Biol Chem.* 284:13348–13354.
Simmons MP, Pickett KM, Miya M. 2004. How meaningful are Bayesian support values? *Mol Biol Evol.* 21:188–199.
Smyth DJ, Cooper JD, Bailey R, et al. (14 co-authors). 2006. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nat Genet.* 38:617–619.
Sullivan J, Swofford DL. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst Biol.* 50:723–729.
Susko E, Inagaki Y, Roger AJ. 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. *Mol Biol Evol.* 21: 1629–1642.
Suzuki Y, Glazko GV, Nei M. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc Natl Acad Sci U. S. A.* 99:16138–16143.
Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol.* 50:525–539.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
Takahasi K, Kumeta H, Tsuduki N, et al. (11 co-authors). 2009. Solution structures of cytosolic RNA sensor MDA5 and LGP2 C-terminal domains: identification of the RNA recognition loop in RIG-I-like receptors. *J Biol Chem.* 284:17465–17474.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* 56:564–577.
Tennessen JA, Bigham AW, O'Connor TD, et al. (23 co-authors). 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69.
Vasseur E, Patin E, Laval G, Pajon S, Fornarino S, Crouau-Roy B, Quintana-Murci L. 2011. The selective footprints of viral pressures at the human RIG-I-like receptor family. *Hum Mol Genet.* 20: 4462–4474.
Wang Y, Ludwig J, Schuberth C, et al. (12 co-authors). 2010. Structural and functional insights into 5′-ppp RNA pattern recognition by the innate immune receptor RIG-I. *Nat Struct Mol Biol.* 17: 781–787.
Wu B, Peisley A, Richards C, Yao H, Zeng X, Lin C, Chu F, Walz T, Hur S. 2013. Structural basis for dsRNA recognition, filament formation, and antiviral signal activation by MDA5. *Cell* 152:276–289.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 22: 1107–1118.
Yoneyama M, Kikuchi M, Matsumoto K, et al. (13 co-authors). 2005. Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity. *J Immunol.* 175: 2851–2858.
Yoneyama M, Kikuchi M, Natsukawa T, Shinobu N, Imaizumi T, Miyagishi M, Taira K, Akira S, Fujita T. 2004. The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nat Immunol.* 5:730–737.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 22:2472–2479.
Zor T, Selinger Z. 1996. Linearization of the Bradford protein assay increases its sensitivity: theoretical and experimental studies. *Anal Biochem.* 236:302–308.
Zou J, Chang M, Nie P, Secombes CJ. 2009. Origin and evolution of the RIG-I like RNA helicase gene family. *BMC Evol Biol.* 9:85.