The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events

Jonathan M. Urbach^a and Frederick M. Ausubel^{a,b,1}

^aDepartment of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114; and ^bDepartment of Genetics, Harvard Medical School, Boston, MA 02115

Contributed by Frederick M. Ausubel, December 18, 2016 (sent for review May 18, 2016; reviewed by William J. Palmer and Jenny Ting)

There are intriguing parallels between plants and animals, with respect to the structures of their innate immune receptors, that suggest universal principles of innate immunity. The cytosolic nucleotide binding site-leucine rich repeat (NBS-LRR) resistance proteins of plants (R-proteins) and the so-called NOD-like receptors of animals (NLRs) share a domain architecture that includes a STAND (signal transduction ATPases with numerous domains) family NTPase followed by a series of LRRs, suggesting inheritance from a common ancestor with that architecture. Focusing on the STAND NTPases of plant R-proteins, animal NLRs, and their homologs that represent the NB-ARC (nucleotide-binding adaptor shared by APAF-1, certain R gene products and CED-4) and NACHT (named for NAIP, CIA, HET-E, and TEP1) subfamilies of the STAND NTPases, we analyzed the phylogenetic distribution of the NBS-LRR domain architecture, used maximum-likelihood methods to infer a phylogeny of the NTPase domains of R-proteins, and reconstructed the domain structure of the protein containing the common ancestor of the STAND NTPase domain of R-proteins and NLRs. Our analyses reject monophyly of plant R-proteins and NLRs and suggest that the protein containing the last common ancestor of the STAND NTPases of plant R-proteins and animal NLRs (and, by extension, all NB-ARC and NACHT domains) possessed a domain structure that included a STAND NTPase paired with a series of tetratricopeptide repeats. These analyses reject the hypothesis that the domain architecture of R-proteins and NLRs was inherited from a common ancestor and instead suggest the domain architecture evolved at least twice. It remains unclear whether the NBS-LRR architectures were innovations of plants and animals themselves or were acquired by one or both lineages through horizontal gene transfer.

NLR | R-protein | innate immunity | evolution | NOD-like receptors

Plants and animals diverged ~1.6–1.8 billion years ago (1) from a common unicellular ancestor (2). Given their long separate evolutionary histories, it is intriguing that the innate immune systems of plants and animals share many common features, such as the use of pattern recognition receptors to recognize the molecular hallmarks of infection, programmed cell death to arrest a localized infection (3), and MAP kinase cascades and hormones to mediate immune signaling (4). It is unclear to what extent similarities between the innate immune systems of plants and animals reflect convergent or parallel evolution and to what extent they reflect inheritance from a common ancestor (5), but instances of convergent evolution in innate immune systems are of interest because they can reveal fundamental principles of immunity.

The similarity of the so-called NOD-like receptors (NLRs) of metazoans and plant NBS-LRR resistance proteins (R-proteins) was noted soon after their discovery, leading to the hypothesis that they evolved from a common ancestor (6, 7). Plant R-proteins and metazoan NLRs share a common nucleotide binding site-leucine rich repeat (NBS-LRR) architecture. R-proteins function directly or (more frequently) indirectly as receptors of pathogen-encoded effector proteins, which are often secreted by pathogens directly into host cells. Some metazoan NLRs, in contrast, function as receptors for pathogen-derived microbe-associated molecular pattern molecules, including bacterial peptidoglycans, flagellin, DNA, viral RNA, and toxins (8), although it has not been demonstrated that all NLRs are actually

receptors. So striking, however, are the structural and functional similarities of these two classes of innate immune proteins that the term NOD-like receptor has been used in numerous publications to refer to all immune-related proteins with an NBS-LRR domain structure (9). In this article, we use the term R-protein to refer to plant NBS-LRR receptors and the term NLR to refer to metazoan proteins with a NACHT-LRR domain structure [also known as NBD-LRRs (10)].

The similar architectures of plant R-proteins and metazoan NLRs are characterized by an N-terminal effector domain, followed by a STAND (signal transduction ATPases with numerous domains) family NTPase domain (11) [either NACHT (named for NAIP, CIA, HET-E, and TEP1) (PF05729) (12) in the case of metazoan NLRs or NB-ARC (nucleotide-binding adaptor shared by APAF-1, certain R gene products and CED-4) (PF00931) (13) in the case of plant R-proteins], and a series of LRRs constituting the sensor domain (*SI Appendix, Fig. S1*). The effector domain may include any of a diverse group of protein-interacting domains, such as a Toll-interleukin receptor (TIR)-domain or a coiled coil (CC) domain in plant R-proteins, or in metazoan NLRs, a Death-fold domain, such as a CARD (caspase activation and recruitment domain), or alternatively, an inhibitor of apoptosis (IAP)/baculoviral inhibitor of apoptosis protein repeat (BIR) domain. The LRR repeats of the sensor domains are principally responsible for pathogen sensing, whereas the STAND NTPase is believed to provide the functionality of a biochemical

Significance

Instances of convergent evolution in innate immune systems can reveal fundamental principles of immunity. We use phylogenetic reconstruction and tests of alternative evolutionary hypotheses combined with ancestral state reconstruction and an analysis of the phylogenetic distribution of the nucleotide binding site-leucine rich repeat (NBS-LRR) architecture to demonstrate that most likely, two distinct and separate evolutionary events gave rise to the NBS-LRR architecture in plant resistance proteins (R-proteins) and metazoan NOD-like receptors (NLRs), and furthermore, the STAND NTPase of both protein families was likely inherited from ancestral proteins with an NBS-tetratricopeptide (TPR) architecture. The implication is that the NBS-LRR architecture has special functionality for innate immune receptors.

Author contributions: J.M.U. and F.M.A. designed research; J.M.U. performed research; J.M.U. contributed new reagents/analytic tools; J.M.U. analyzed data; and J.M.U. and F.M.A. wrote the paper.

Reviewers: W.J.P., University of California, Davis; and J.T., University of North Carolina. The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The scripts and datasets used in this paper have been deposited in the Dryad Data Repository (www.datadryad.org), [dx.doi.org/10.5061/dryad.2nd84](https://doi.org/10.5061/dryad.2nd84).

¹To whom correspondence should be addressed. Email: ausubel@molbio.mgh.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1619730114/-DCSupplemental.

switch, with discrete “on” and “off” states that can be maintained for a period determined by its catalytic rate (14).

Previous authors have addressed in part the issue of the evolution of NLR and R-proteins based on differences in the NB-ARC and NACHT subfamilies. Analyzing the phylogenetic distribution of the NB-ARC and NACHT domains in combination with LRRs, Yue et al. (15) concluded that the NB-ARC-LRR architecture was an innovation of plants, and hence that R-proteins and NLRs evolved independently. Yuen et al. (16) reached similar conclusions about NLRs. However, neither set of authors presented a rooted tree that included both R-proteins and NLRs, assessed the plausibility of the NBS-LRR architecture as an ancestral form, or accounted for the possibility that plants or metazoans might have acquired the NBS-LRR architecture through horizontal gene transfer, an important issue because eukaryotic lineages are believed to have acquired STAND NTPases repeatedly through horizontal gene transfer events (11). Therefore, in our opinion, the case for convergent evolution of metazoan NLRs and plant NBS-LRR resistance proteins has not been rigorously demonstrated.

In this work, we use phylogenetic analyses to test whether the NBS-LRR architectures of R-proteins and NLRs were inherited from a common ancestor. We focused on the central STAND NTPase domains of R-proteins and NLRs because, as noted elsewhere (11), they have a broad phylogenetic distribution across the tree of life and would likely not be subject to the expansions and contractions that complicate phylogenetic analysis of repeat sequences. STAND NTPases are a class of P-loop-containing NTPases that include not only the closely related NACHT and NB-ARC clades but also the presumed earlier diverging SWACOS (STAND with adenylyl cyclase or serine/threonine protein kinase), MalT (named for the MalT family of bacterial transcription factors in which the domain is commonly found), and MNS (named for three subfamilies comprising this clade, the MJ/PH, the Npun2340/2341, and the SSO families of NTPases) subclades (11). STAND NTPases all share a five-stranded core, a Walker A motif (P-loop) for phosphate binding, and a Walker B motif for coordinating a catalytic magnesium. In addition, STAND NTPases are often associated with C-terminal WD40 (tryptophan- and aspartic acid-containing repeat, approximately 40 amino acids in length), tetratricopeptide (TPR), LRR, ankyrin (ANK), or armadillo (ARM) repeats (*SI Appendix, Fig. S1*). Except for the MNS subclade, they also often include a C-terminal helical third domain of STAND (HETHS domain). On the basis of their shared sequence features and phylogenetic distributions, Leipe et al. surmised that diversification of NACHT and NB-ARC domains from a common ancestral form occurred in the common ancestor of Actinobacteria, Cyanobacteria, and Chloroflexi (11). The hypothesis of monophyly of the NBS-LRR architectures of R-proteins and NLRs presupposes that the common ancestor of NB-ARC and NACHT domains was associated with C-terminal LRR repeats, and that this architecture was conserved in both lineages. Evidence of ancestral architectures other than NBS-LRR in the common ancestor, or in more recent ancestors, would therefore be sufficient to reject the hypothesis of monophyly of NLRs and R-proteins in favor of convergent evolution.

Here we present a phylogenetic reconstruction of the NB-ARC and NACHT domains, using a large set of sequences, analyses to test the likelihood of alternative evolutionary hypotheses, as well as reconstruction of the ancestral domain structure of the STAND NTPase that was the common ancestor of the STAND NTPases of R-proteins and NLRs. We also present surveys of the phylogenetic distribution of the NBS-LRR and related domain combinations. We find that the NBS-LRR domain architecture of R-proteins and NLR proteins likely evolved at least twice, independently in plants and metazoans, and that the common ancestor of the NTPases of these two protein families most likely had an NBS-TPR architecture, not an NBS-LRR architecture. Furthermore, these findings show that the similarity of R-proteins and NLRs likely results from convergence.

Results

Mining of GenBank for NTPase Domains and Subsequent Iterative Alignments and Phylogeny. Using representative NB-ARC and NACHT sequences, we generated a hidden Markov model (HMM) sensitive to both NB-ARC and NACHT NTPases that we refer to as STAND-HMM (*SI Appendix, Results*). It was also sensitive to the SWACOS and MalT clades, but only weakly to the MNS clade. We used this HMM to probe 10,565,004 sequences in the NCBI non-redundant (NR) protein database and obtained 15,500 STAND NTPase domain hits. To add additional data to our alignment, we appended the corresponding C-terminal HETHS domain to each NTPase domain (*SI Appendix, Results*), and a subset of the resulting STAND-HETHS domain combinations was used to generate an alignment of 964 NB-ARC and NACHT NTPases, plus some representative SWACOS, and MalT NTPases to serve as outgroups (*SI Appendix, Fig. S2 and Table S1*).

Maximum-Likelihood Phylogeny of R-Proteins. A maximum-likelihood (ML) tree of the alignment of 964 STAND NTPases (with appended HETHS domains) was generated using RAxML with 1,000 bootstraps (Fig. 1 and *SI Appendix, Table S2*). The tree was rooted using, as outgroups, the SWACOS (XXI) and MalT (XXII) subclades of the STAND family, which were inferred by Leipe et al. (11) to have diverged before the NACHT and NB-ARC subclades. Some important relationships are immediately apparent from the ML tree (Fig. 1). Clades containing known NACHT domains (XVI–XIX) form a group with a small clade of firmicute and cyanobacterial NTPases (XX) with 89% bootstrap confidence that we classified as putative early-diverging NACHT domains. The known NACHT domains form four groups, including a metazoan and bacterial clade containing the NOD-like receptors with NBS-LRR architectures (XVIII, 91% bootstrap confidence), a set of bacterial NACHT domains associated with NBS-WD40 and other architectures (XIX, 94% bootstrap confidence), a clade including metazoan TEP1 (telomerase associated protein 1) and related bacterial proteins associated with NBS-WD40 and NBS-TPR architectures (XVII, 72% bootstrap support), and a weakly supported clade including fungal, bacterial, and metazoan proteins with NBS-WD40, NBS-ANK, NBS-TPR, and other domain architectures (XVI, 49% bootstrap support). The NACHT clades (XVI–XX) form a group that is sister to all of the STAND NTPases except the SWACOS (XXI) and MalT (XXII) clades. Despite having a long branch and moderately good bootstrap support (89%), the NACHTs (XVI–XX) do not appear to be firmly excluded from the NB-ARC clade, as the bootstrap value for grouping of the NB-ARC clades (I–XV) is 71.

Consistent with previously noted similarities among R-proteins, plant R-proteins group together with 99% bootstrap support in four distinct clades (I, II, III, and IV with 100%, 100%, 84%, and 100% bootstrap support, respectively). Of these four clades, clade I includes NBS-LRR proteins with N-terminal TIR domains (TIR-NBS-LRR structure) but excludes proteins with N-terminal CC domains. Conversely, clades II, III, and IV contain NB-ARC proteins with N-terminal CC domains (CC-NBS-LRR structure), but none with N-terminal TIR domains. Individual groupings among clades I–IV have poor bootstrap support and are therefore uncertain.

It is notable that the ML tree places the STAND NTPases of R-proteins and the STAND NTPases of NLRs nested in separate well-supported (or moderately well supported) clades of proteins with non-NBS-LRR domain structures, specifically the I–X clade [96% bootstrap support, itself nested in the NB-ARC clade (I–XV) with 71% support] and the NACHT clade (XVI–XX, 89% support). This topology suggests the NLRs and R-proteins do not form a monophyletic clade, and most likely arose in separate evolutionary events. Although there is uncertainty in the tree regarding the early branching relationships among the NB-ARC clades, clades I–X form a well-supported group with 96%

Fig. 1. Phylogram representing 1,000-bootstrap RAXML tree. The ML tree was generated using 16 cores of a Linux cluster running 1,000 bootstraps of maximum-likelihood analysis on the 964-protein alignment, using RAXML's fast bootstrap method (27), the WAG substitution matrix, and the Γ evolutionary model with four rate classes assumed and empirically determined amino acid frequencies. The resulting tree was rerooted as described in *Methods*, using the clade defined by the common parental node of MalT_VIBORI (ZP_05944565.1, representative MalT domain) and ThcG#1_RHOERY (AAD28307.1, representative SWACOS domain) as the out-group. Tree nodes representing the common ancestors of the various clades were collapsed. Represented taxa and domain combinations of interest, that is, NBS in combination with LRR, WD40, TPR, ANK, and ARM, are represented symbolically, as indicated in the key. Where they occur, TIR and CC domains are also indicated. Branch lengths reflect lengths from the ML tree excluding collapsed branches, except for extremely short branches indicated with an * that were lengthened so as to be distinguishable to the reader.

bootstrap support in the ML tree (Fig. 1), indicating that the NB-ARC domains of plant R-proteins (I–IV) have a close relationship with those of several groups of NBS-WD40 and NBS-ARM proteins (V–X) including metazoan APAF1 and CED4 proteins (VIII) and others from metazoans, prokaryotes, and plants.

ML tree topology also suggests a close relationship between the NB-ARC domains of metazoan apoptotic ATPases (VIII) and those of clade VII associated with bacterial NBS-WD40 proteins. That the NTPase and WD40 repeat domains, together comprising 84–94% of the length of human APAF1 (ABQ59028.1) sequence, align by BLAST with numerous presumed bacterial transcriptional activators (*SI Appendix, Table S3*), raises the possibility that the NBS-WD40 domain architecture of APAF1 and related metazoan proteins was conserved from a common possibly prokaryotic ancestral protein and acquired by metazoans via horizontal gene transfer.

Testing Alternate Evolutionary Hypotheses with the AU and SOWH Tests. We used the AU (approximately unbiased) and SOWH (Swofford, Olsen, Waddell, and Hillis) tests to further assess support for the clade groupings observed in the ML tree and to test alternate evolutionary hypotheses about the relative positions of the STAND NTPases of NLRs (XVIII) and of R-proteins (I–IV). These tests offer statistically robust, yet complimentary, ways to evaluate alternate evolutionary hypotheses and, based on a given sequence alignment, generate *P* values indicating the relative likelihoods of different evolutionary scenarios represented as constraints on tree topology. Low *P* values (≤ 0.05) indicate that the evolutionary hypothesis represented by a given constraint is significantly less likely than the one represented by the unconstrained ML tree. Whereas the AU test is versatile and relatively easy to perform (17), the SOWH test is intuitive, statistically powerful, and mostly independent of the best ML tree itself, although computationally intensive and known to be sensitive to model misspecification (18). For both tests, we defined constraints that

imposed tree topologies consistent with the various evolutionary hypotheses by grouping certain clades together and excluding others. For the AU test, ML trees were generated on the basis of these constraints and then compared with the unconstrained ML tree, using the CONSEL package to generate the AU and other test statistics. For the SOWH test, in contrast, pairs of constrained and unconstrained parametric bootstraps were conducted and used to generate the test statistic. Because of the computationally intensive nature of the SOWH test, we created a modified version of the SOWHAT program (19) to allow us to run 250 pairs of bootstraps as an array of parallel jobs on a Linux computing cluster and then assemble the data to calculate *P* values. Because each set of SOWH bootstraps required ~ 3 y of CPU time, we tested only two constraints we believed to be the most important: CON4 and CON6 (Table 1).

The first three constraints excluded the NTPases of R-proteins from their nested position in the NB-ARC clades (Table 1). Of these, the first one (CON1), which excluded R-protein NTPases (I–IV) from clades V–X, yielded an AU test *P* value of 0.319 ± 0.101 , and is therefore very plausible. The other constraints, which excluded R-protein NTPases (I–IV) from clades V–XIII (CON2; AU *P* = 0.076 ± 0.027) or clades V–XV (CON4; AU *P* = 0.056 ± 0.044), gave AU test *P* values very slightly above α (0.05), rejecting at slightly less than 95% confidence tree topologies that move R-proteins out of their deeply nested position among NB-ARC sequences. However, the SOWH test rejects CON4 with high statistical significance (SOWH *P* < 0.004), indicating that topologies that move the STAND-NTPases of R-proteins out of the NB-ARC clade are very unlikely.

The other three constraints force NACHTs (XVI–XX) into an exclusive grouping with R-proteins (I–IV) (CON5; AU *P* = 0.110 ± 0.072) or force NLRs and their closest bacterial homologs (XVIII) to be grouped with R-proteins (I–IV) (CON6; AU *P* = 0.069 ± 0.035 ; SOWH *P* < 0.004) or with clades I–X (CON7;

Table 1. AU and SOWH test results

Constraint	AU	SOWH	SOWH CI	Description
CON1	0.319 ± 0.101			R-protein STANDs (I–IV) excluded from V–X
CON2	0.076 ± 0.027			R-protein STANDs (I–IV) excluded from V–XIII
CON4	0.056 ± 0.044	<0.004	0–0.01465	R-protein STANDs (I–IV) excluded from V–XV
CON5	0.110 ± 0.072			R-protein STANDs (I–IV) grouped with NACHTs (XVI–XX)
CON6	0.069 ± 0.035	<0.004	0–0.01465	R-protein STANDs (I–IV) grouped with clade XVIII (NACHTs from metazoan NLRs and bacterial homologs)
CON7	0.085 ± 0.058			Metazoan NLR and related bacterial proteins (XVIII) grouped with clades I–X (R-proteins, APAF1-like NBS-WD40, and bacterial homologs)

The AU and SOWH tests were performed as detailed in *Methods* to test various alternative evolutionary hypotheses. The *P* values of the AU and SOWH tests are indicated, along with the corresponding upper and lower SOWH 95% CI (confidence interval) and a description of each constraint.

AU $P = 0.085 \pm 0.058$). Again, these AU tests produced *P* values close to or slightly above α , rejecting at slightly less than 95% confidence tree topologies that group the NTPases of R-proteins, NLRs, and the clades in which they are nested. Once again, however, the SOWH test rejects the evolutionary scenario represented by CON6 with high significance, thereby rejecting monophyly of R-proteins and clade XVIII, which includes NLRs and their bacterial homologs. Additional AU tests using strictly defined topologies also reject monophyly of R-proteins and NLRs and are consistent with the nested position of R-proteins within the NB-ARC clade (*SI Appendix, Results* and Table S4 and Fig. S3).

Reconstruction of the Ancestral Domain Structure Associated with the Common Ancestor of the STAND NTPases of R-Proteins and NLRs. Reasoning that an additional way to distinguish among the various hypotheses about the origins of NLRs and R-proteins was to reconstruct the likely domain structures of the ancient proteins containing the ancestors of their NTPase domains, and in particular their last common ancestor (LCA), we applied ancestral state reconstruction methods to reconstruct putative ancestors represented in the ML tree. Furthermore, since ancestral state reconstruction based on a single ML tree containing numerous nodes with low bootstrap support would be insufficient for confidence in the reconstructed ancestral states, we performed a second analysis in parallel, in which we used 280 bootstraps as a plausible sampling of tree topologies, which we individually rooted and used for ancestral state reconstruction, averaging the bootstrap-derived ancestral state probabilities for ancestors of interest. We used three different standard ancestral state reconstruction methods, including two that calculated marginal ancestral state likelihoods by either an ML (Fig. 2 and *SI Appendix, Tables S5 and S6*) or a continuous-time Markov chain approach (*SI Appendix, Fig. S4A and Tables S5 and S6*), and one ML-based method that calculated conditional scaled ancestral state likelihoods (*SI Appendix, Fig. S4B and Tables S5 and S6*). The three methods gave overall similar results (see *SI Appendix, Results* for detailed analysis), indicating that the protein containing the common ancestor of the STAND NTPases of R-proteins and NLRs most likely did not have a NBS-LRR domain structure (ML marginal bootstrap average likelihood of NBS-LRR = .00013), but more likely, an NBS-TPR domain structure (ML marginal bootstrap average likelihood of NBS-TPR = 0.98155) (*SI Appendix, Table S5*), which further supports the conclusion that the NBS-LRR architectures of R-proteins and of NLRs evolved in separate events. Furthermore, it is likely that the protein containing the common ancestor of clades I–X had an NBS-WD40 domain structure (marginal ML bootstrap average, NBS-WD40 likelihood = 0.68577), and the common ancestor of NACHTs appears to have had an NBS without associated repeats (marginal ML bootstrap average likelihood = 0.83084; *SI Appendix, Table S5*), which suggests that the evolutionary path from NBS-TPR architecture in the common ancestor of the STAND NTPases of R-proteins and NLRs proceeded through NBS-WD40 and nonrepeat associated NBS architectures to

give rise to the NBS-LRR domain structures of R-proteins and NLRs, respectively. Although the outgroup method is generally used to root a tree, and we have applied this method based on previous work (11), we also carried out an alternative midpoint rooting method, which also excludes NBS-LRR as the domain structure of the LCA but raises the possibility of a non-repeat-associated NBS architecture instead of an NBS-TPR architecture (*SI Appendix, Results* and Fig. S4 D and E and Table S5 D and E).

Survey of NBS in Combination with Repeats in the NCBI NR Protein Database and in Eukaryotic Genome Sequences. The NCBI NR protein database and a set of 46 draft and finished eukaryotic genomes (*Dataset S1*) were scanned for the STAND-NTPase alone and in combination with LRR, WD40, TPR, and ARM domains (*SI Appendix, Results*). In contrast to NBS-WD40 and NBS-TPR proteins, the NBS-LRR domain combination is rare except in plants and metazoans (Fig. 3 and *SI Appendix, Table S7*), with the exceptions being a group of 12 closely related prokaryotic proteins, primarily from *Streptomyces* (*SI Appendix, Table S8*), and a few from nonplant, nonanimal eukaryotes (*SI Appendix, Table S9*). Of the eukaryotic NBS-LRRs, a protein

Fig. 2. Marginal ML Reconstruction of marginal ancestral state likelihoods in the ML tree, using the ace function of the phyttools R package. Pie charts at nodes represent fractional likelihoods of each species, with green representing NBS-LRR, blue for NBS-WD40, red for NBS-TPR, yellow for NBS-ARM, and gray for NBS (nonrepeat associated).

Fig. 3. Survey of 46 sequenced eukaryotic genomes for NBS, NBS-LRR, NBS-TPR, and NBS-WD40 domain architectures (*Methods*). Total counts of each domain combination for each eukaryotic genome are presented, along with a tree indicating the current best understanding of the phylogenetic relationships of these taxa. Major phylogenetic groups are also labeled and include the following abbreviations: Cf., Choanoflagellida; A., Amoebozoa; Exc., Excavata; Rp., Rhodophyta; Rz., Rhizaria; and Str., Stramenopiles. Table cells are colored as a heat map, with more intense red corresponding to higher counts.

from *Monosiga brevicollis* (jgi|Monbr1|25471|fgenesh2_pg.scaffold_10000121) seems to be the closest analog of R-proteins and NLRs but is apparently not closely related to either.

Discussion

In this work, we used maximum-likelihood phylogeny analysis (Fig. 1) to place the STAND-NTPases of NOD-like receptors (XVIII) and R-proteins (I–IV), both with NBS-LRR architectures, nested within two different well-supported clades of NTPases possessing predominately non-NBS-LRR architectures, the NACHTs (XVI–XX) and I–X, respectively. Being nested within different clades suggests that they are not monophyletic, and being nested in clades associated predominantly with architectures other than NBS-LRR suggests that ancestors of the R-protein and NLR NTPases more recent than their last common ancestor resided in proteins with non-NBS-LRR architectures.

Consistent with the ML tree topology, the SOWH test, which uses 250 pairs of constrained and unconstrained parametric bootstraps independent of the ML tree, rejects monophyly of R-proteins and NLRs with high statistical significance when R-protein STANDS (clades I–IV) are grouped with clade XVIII containing NACHTs from metazoan NLRs and bacterial homologs (CON6; Table 1), as does the AU test, although with only marginal significance. Furthermore, the SOWH test supports

with high significance and the AU test with marginal significance the nested position of R-proteins within the NB-ARC clade (V–XV; CON4; Table 1), suggesting that the proteins containing the ancestors of the NTPases of R-proteins (I–IV) possessed architectures similar to those of clades V–XV, which are predominantly NBS-TPR and NBS-WD40.

Ancestral state reconstruction, based on 280 additional bootstraps independent of the ML tree, suggests that neither the LCA of the STAND NTPases of R-proteins and NLRs nor some ancestors more recent than the LCA likely possessed NBS-LRR domain architectures, including the common ancestor of clades I–X and the common ancestor of NACHTs (XVI–XX). Instead, the protein containing the LCA likely possessed an NBS-TPR domain structure. Furthermore, the evolutionary path from ancestral NBS-TPR to the NBS-LRR structures of R-proteins and NOD-like receptors probably occurred via intermediates, most likely an NBS-WD40 architecture in the case of R-proteins and a non-repeat associated NBS architecture in the case of NLRs. Because neither the ancestral protein containing the LCA nor more recent ancestral proteins likely possessed the NBS-LRR architecture, it follows logically that the NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events.

The conclusion that plant R-proteins and metazoan NLRs evolved in independent events is further supported by the phylogenetic distribution of the NBS-LRR domain architecture (Fig. 3), in agreement with earlier conclusions by ourselves and others (15, 20). Whereas NBS-TPR and NBS-WD40 domain architectures have a broad distribution across the tree of life, NBS-LRR proteins are mainly found in land plants and metazoa (*SI Appendix, Table S7* and Fig. 3), suggesting NBS-LRR proteins are plant and metazoan innovations. In contrast, the presence of structural analogs of R-proteins and NOD-like receptors in other clades, including *M. brevicollis* and some Actinobacteria and planctomycetes, suggests acquisition by horizontal gene transfer is still plausible.

The conclusion that R-proteins and NLRs represent independent innovations of the NBS-LRR domain architecture raises the question of what is so special about the NBS-LRR combination in the context of immune receptors. Both the STAND NTPase and the LRR repeat motifs offer numerous potential benefits. Takken et al. have proposed that the STAND NTPase domain has at least two potentially useful features for immune receptors (14) (*SI Appendix, Fig. S5*); namely, its ability to act as a timed switch and the fact that it often requires oligomerization to signal in its “on” state. The oligomerization requirement may create a threshold by which multiple signals combine to produce activation, thereby providing a mechanism by which to filter out spurious signals and contravene inappropriate and costly activation of immune responses such as programmed cell death. Furthermore, the lifetime of the activated state of the receptor is finite and is dependent on the catalytic rate of the NTPase active site (*SI Appendix, Fig. S5*). Another beneficial characteristic of the NB-ARC domain may be its modularity and ability to tolerate mutations in the sensor domains that might otherwise affect the protein’s stability.

It has been argued that repeat protein repeat domains such as LRRs offer evolutionary advantages over nonrepeat domains (21). Repeats on the level of amino acid sequence are often repeats on the level of DNA sequence, which may allow facile extension and deletion of repeat units through genetic recombination. Intragenic tandem duplication can provide a rapid mechanism to develop new specificities without sacrificing old ones. Furthermore, repeats, especially those that form open rod-like or helical structures, have the potential to be arbitrarily large, allowing a single protein to amass a large number of binding specificities. The LRR repeat domains of R-proteins are often highly variable and under pressure of diversifying selection (22). What makes LRR repeats any more suited to ligand-binding roles in immune receptors relative to other repeat domains is unclear, but it is possible that they are in some way exceptional in their ability to allow access to a

wide range of binding specificities through common mutation and recombination events while maintaining a stable tertiary structure. The ubiquity of LRRs in the ligand-binding domains of both plant and animal membrane-associated and soluble innate immune receptors, including receptor-like kinases, Toll-like receptors, NLRs and R-proteins, suggests a special versatility for binding. Perhaps the greatest testament to the LRR's ability to easily evolve a wide range of binding specificities is the fact that recombination of LRRs is the basis for generating the diversity of VLR proteins, the immunoglobulin equivalents in the adaptive immune systems of lampreys and hagfish (23).

Ultimately, the advantage of the NBS-LRR architecture over alternative architectures may be that by combining individual advantages of the STAND NTPase and LRR domains, it provides an innate immune receptor with a finite lifetime of its activated state that can safely initiate programmed cell death once a suitable threshold of activation has been crossed, and that can easily evolve new binding specificities under the strains of diversifying selection without sacrificing stability.

In this work, we have demonstrated plausible paths for the evolution of metazoan NOD-like receptors and plant R-proteins with NBS-LRR domain structures, and have shown that the protein containing the common ancestor of the STAND NTPase of R-proteins and NLRs likely possessed an NBS-TPR architecture, that inheritance of the STAND NTPase domain likely occurred through intermediate architectures, and that independent innovations gave rise to the NBS-LRR architectures of plant R-proteins and metazoan NLRs. This intriguing convergence hints at possible universal principles of innate immunity and the importance of factors such as the ease of accessing new binding specificities through common mechanisms of genetic recombination, secondary and tertiary structure stability with respect to mutational disruption, the finite half-life of the activated state, and resistance to spurious activation. How these and other potential factors have contributed to the evolution of R-proteins and NOD-like receptors is worth exploring.

Methods

Sequence Alignment. Sequences were aligned manually using the SeaView sequence editor version 4.4.1 (24). Multiple sequence alignments that were prealigned using hmalign (HMMER package) were improved by using

MUSCLE version 3.8.31 (25), MAFFT version 6.811b (26), and manual adjustment to adjust the poorly conserved portions of the sequence alignment between highly conserved regions, using SeaView's "Align selected sites" feature. SeaView's "Add external method" feature allowed the execution of MAFFT from within the SeaView program.

Phylogeny Reconstruction. Maximum-likelihood phylogeny reconstruction was performed as indicated, either using the MPI version of RAxML (27) on a Linux cluster or, where indicated, using PhyML version 3.0 (28). On the basis of an initial analysis of a subset of sequences using the MIX option of the Bayesian phylogeny program MrBAYES version 3.2.1 (29), and model selection using the ProteinModelSelection.pl script developed by Stamatakis (27), the WAG substitution matrix was selected for the ML phylogeny, using the Γ multiple rate model and assuming four discrete rate classes and empirically determined amino acid frequencies.

Trees were rerooted using the reroot_tree.pl script written in house, which reroots Newick format trees to a node specified by the user as the common parent node of two leaves.

The AU and SOWH Tests. RAxML version 8.2.3 was used to generate ML trees based on predefined constraints (via the -f d, -g and -no-bgfs options). The AU and other tests were conducted using the CONSEL package version 0.20 (30). Because it was observed that optimization of the ML tree that followed multiscale bootstraps analysis did not always give identical *P* values, these two steps were performed in nine parallel replicates (except for CON4, which was performed in 18 parallel replicates; details in *SI Appendix, Methods*).

The SOWH test was performed using a modified version of the SOWHAT program (19), which allowed analysis to be suspended after generation of bootstrap alignments, such that RAxML bootstraps could be run as parallel job arrays on many cores of a computing cluster. After bootstraps were run, the SOWHAT program could then be recommenced to give final results. Constraint trees for use as input into the SOWHAT program were made using a perl script (constraint_tree_maker.pl URL) developed in house.

Ancestral State Reconstruction. Ancestral state reconstruction was performed using the ace (Ancestral Character Estimation) function from the APE package version 3.4 (31) in R version 3.2.3, with both marginal and conditional probabilities of ancestral states calculated. See *SI Appendix, Methods* for details.

ACKNOWLEDGMENTS. We thank Dong Wang and Linda Holland for providing DNA samples; Samuel Church for advice relating to the SOWH analysis; and Sarah Mathews, Noah Whiteman, Slim Sassi, Xuecheng Zhang, and Richard Michelmore for helpful conversations. This work was supported by National Science Foundation Grants MCB-0519898 and IOS-0929226 and National Institutes of Health Grants R37-GM48707 and P30 DK040561.

- Parfrey LW, Lahr DJG, Knoll AH, Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci USA* 108(33):13624–13629.
- Meyerowitz EM (2002) Plants compared to animals: The broadest comparative study of development. *Science* 295(5559):1482–1485.
- Ausubel FM (2005) Are innate immune signaling pathways in plants and animals conserved? *Nat Immunol* 6(10):973–979.
- Tena G, Boudsocq M, Sheen J (2011) Protein kinase signaling networks in plant innate immunity. *Curr Opin Plant Biol* 14(5):519–529.
- Hunter P (2005) Common defences. *EMBO Rep* 6(6):504–507.
- Inohara N, Ogura Y, Chen FF, Muto A, Nuñez G (2001) Human Nod1 confers responsiveness to bacterial lipopolysaccharides. *J Biol Chem* 276(4):2551–2554.
- Ting JPY, Davis BK (2005) CATERPILLER: A novel gene family important in immunity, cell death, and diseases. *Annu Rev Immunol* 23(1):387–414.
- Monie TP (2013) NLR activation takes a direct route. *Trends Biochem Sci* 38(3):131–139.
- Jacob F, Vernaldi S, Maekawa T (2013) Evolution and conservation of plant NLR functions. *Front Immunol* 4:1–16.
- Ting JPY, et al. (2008) The NLR gene family: A standard nomenclature. *Immunity* 28(3):285–287.
- Leipe DD, Koonin EV, Aravind L (2004) STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: Multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J Mol Biol* 343(1):1–28.
- Koonin EV, Aravind L (2000) The NACHT family - a new group of predicted NTPases implicated in apoptosis and MHC transcription activation. *Trends Biochem Sci* 25(5):223–224.
- van der Biezen EA, Jones JGD (1998) The NB-ARC domain: A novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr Biol* 8(7):R226–R227.
- Takken FLW, Tameling WIL (2009) To nibble at plant resistance proteins. *Science* 324(5928):744–746.
- Yue J-X, Meyers BC, Chen J-Q, Tian D, Yang S (2012) Tracing the origin and evolutionary history of plant nucleotide-binding site-leucine-rich repeat (NBS-LRR) genes. *New Phytol* 193(4):1049–1063.
- Yuen B, Bayes JM, Degnan SM (2014) The characterization of sponge NLRs provides insight into the origin and evolution of this innate immune gene family in animals. *Mol Biol Evol* 31(1):106–120.
- Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51(3):492–508.
- Swofford D, Olsen G, Waddell P, Hillis D (1996) *Phylogenetic inference. Molecular systematics* (Sinauer, Sunderland, MA), pp 407–514.
- Church SH, Ryan JF, Dunn CW (2015) Automation and evaluation of the SOWH test with SOWHAT. *Syst Biol* 64(6):1048–1058.
- Haney CH, Urbach J, Ausubel FM (2014) Innate immunity in plants and animals. *Biochemist* 36(5):40–44.
- Andrade MA, Perez-Iratxeta C, Ponting CP (2001) Protein repeats: Structures, functions, and evolution. *J Struct Biol* 134(2–3):117–131.
- Parniske M, et al. (1997) Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell* 91(6):821–832.
- Boehm T, et al. (2012) VLR-based adaptive immunity. *Annu Rev Immunol* 30:203–220.
- Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27(2):221–224.
- Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792–1797.
- Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9(4):286–298.
- Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
- Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–321.
- Ronquist F, et al. (2012) MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61(3):539–42.
- Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246–1247.
- Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.