

The axon degeneration gene *SARM1* is evolutionarily distinct from other TIR domain-containing proteins

Harsha Malapati¹ · Spencer M. Millen² · William J. Buchser^{1,2}

Received: 16 November 2016 / Accepted: 17 April 2017 / Published online: 26 April 2017
© Springer-Verlag Berlin Heidelberg 2017

Abstract Many forms of neurodegenerative disease are characterized by Wallerian degeneration, an active program of axonal destruction. Recently, the important player which enacts Wallerian degeneration was discovered, the multidomain protein *SARM1*. Since the *SARM1* protein has classically been thought of as an innate immune molecule, its role in Wallerian degeneration has raised questions on the evolutionary forces acting on it. Here, we synthesize a picture of *SARM1*'s evolution through various organisms by examining the molecular and genetic changes of *SARM1* and the genes around it. Using proteins that possess domains homologous to *SARM1*, we established distances and K_a/K_s values through 5671 pairwise species–species comparisons. We demonstrate that *SARM1* diverged across species in a pattern similar to other SAM domain-containing proteins. This is surprising, because it was expected that *SARM1* would behave more like its TIR domain relatives. Going along with this divorce from TIR, we also noted that

SARM1's TIR is under stronger purifying selection than the rest of the TIR domain-containing proteins (remaining highly conserved). In addition, *SARM1*'s synteny analysis reveals that the surrounding gene cluster is highly conserved, functioning as a potential nexus of gene functionality across species. Taken together, *SARM1* demonstrates a unique evolutionary pattern, separate from the TIR domain protein family.

Keywords Wallerian axon degeneration · Neuroscience · Innate immune · Avian · Synteny

Introduction

Wallerian degeneration is the active disassembly of an axon following damage or stress. It is prevalent in many neurodegenerative disorders including neuropathies (Geisler et al. 2016), hearing injury (Furman et al. 2013), glaucoma (Beirowski et al. 2008; Bramley et al. 2016), concussion (Henninger et al. 2016), and likely even ALS (Vérière et al. 2015). The etiology of these diseases can be elucidated through study of the proteins involved in this axon degeneration pathway. Of the various key proteins discovered so far, one is regarded as the central executioner in the axon degeneration pathway and is the focus of this paper: *SARM1*. Sterile alpha and TIR motif containing 1 (*SARM1*) is known to be the upstream initiator of axon degeneration upon axonal stress (Osterloh et al. 2012; Gerdts et al. 2013). Without *SARM1*, axon degeneration does not occur or is drastically delayed.

In *Drosophila*, mutations in *ect4* (*SARM1*'s ortholog) result in a near-lifetime delay in the degeneration of olfactory axons after axotomy (Osterloh et al. 2012). In the murine model, *SARM1* KO adults show a 14 day delay in

Communicated by S. Hohmann.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-017-1320-6) contains supplementary material, which is available to authorized users.

✉ William J. Buchser
wjbuchser@wm.edu

Harsha Malapati
smalapati01@email.wm.edu

Spencer M. Millen
spencer.millen@usuhs.edu

¹ Department of Biology, College of William & Mary, 540 Landrum Dr., Williamsburg, VA, USA

² Neuroscience Program, College of William & Mary, Williamsburg, VA, USA

axon degeneration upon nerve transection (Gerdtts et al. 2013). Recent evidence reveals a likely interaction between SARM1 and a classic regulator of axonal integrity, Nicotinamide nucleotide adenylyltransferase 1 (NMNAT), whose identity was discovered to be the cause of the Wallerian degeneration-slow (Wld^S) neuroprotection and whose product Nicotinamide adenine dinucleotide (NAD⁺) is broken down by *SARM1* (Lunn et al. 1989; Araki et al. 2004; Gerdtts et al. 2015; Summers et al. 2016; Essuman et al. 2017). Thus, various independent sources of evidence suggest that SARM1 is an important player in the axon degeneration pathway.

Prior to the discovery of SARM1 as a key component in the axon degeneration pathway, SARM1 was largely known for its role as an adaptor protein in innate immunity. The presence of a Toll interleukin receptor (TIR) domain in SARM1's structure propelled this hypothesis. In *C. elegans*, *tir-1* (another SARM1 ortholog), controls antimicrobial peptide expression and is important in innate immunity (Couillault et al. 2004). There is evidence that SARM1 KO can actually protect against neuronal damage during a neurotrophic infection like VSV through an inflammatory mechanism (Hou et al. 2013). In mice, SARM1-KO infection-activated T-cells show increased proliferation and decreased apoptosis compared to T-cells with SARM1 intact (Panneerselvam et al. 2013). In **addition**, when SARM1 is overexpressed or the TIR domains are activated (Gerdtts et al. 2013), the molecule can induce a nonapoptotic form of cell death in a variety of cells.

Despite *SARM1*'s importance, there are few published studies which discuss *SARM1*'s evolutionary history. One study on horseshoe crab's *SARM1* determined that *SARM1*'s role as a suppressor of Toll-like receptor (TLR) signaling is highly conserved from horseshoe crabs to humans (Belinda et al. 2008). Another study which focuses on the selection of TIR domain-containing adapter proteins in humans reveals that there was a sweep of purifying selection in *SARM1* and Myeloid differentiation primary response 88 (*MYD88*) in all human populations studied, while the three other adaptors TIR domain-containing adapter protein (*MAL*), TIR domain-containing Adapter-inducing interferon β (*TRIF*), and TRIF related adapter molecule (*TRAM*) displayed purifying selection only in certain subgroups of the human population (Fornarino et al. 2011). To make these claims, the paper relies partly on *Ka/Ks*, the ratio of nonsynonymous to synonymous mutations, which if low suggests purifying selection preserving the resulting amino-acid sequence, and if high suggests positive selection favoring a change in amino-acid sequence over time. Interestingly, a version of the TIR domain also goes back to bacteria, and animal SARM1 TIR is more closely related to bacterial rather than animal TIR domain-containing proteins (Zhang et al. 2011). The

studies mentioned above, despite providing useful information about *SARM1*'s evolution, focus primarily on *SARM1*'s role in the immune system and do not investigate the more recently discovered role of *SARM1* in the nervous system (and a potentially even wider function). With the recent discovery of SARM1 TIR's NADase activity, the evolutionary importance of SARM1 as an immune system protein has become less clear. In an effort to provide a broader perspective on *SARM1*'s evolution, we decided to compare the evolutionary history of *SARM1* across 107 species.

SARM1 demonstrates (1) functionality within the innate immune system, (2) an unusual capacity to induce a nonapoptotic form of cell death, and (3) a regulatory role within the process of Wallerian degeneration. This plethora of potential functions intimates a diverse network of regulatory requirements and evolutionary pressures, which can be explored through phylogenetic analysis of SARM1. In our inquiry, although investigating other aspects of the *SARM1* gene, we focus on SARM1's domains: Sterile alpha motif (SAM), TIR, and Armadillo Repeat (ARM). There are other proteins with these domains, and our study of SARM1 analyzes the protein sequences and domain-coding DNA sequences of various TIR, ARM, and SAM domain-containing proteins to determine how each group of proteins relates to SARM1.

Materials and methods

Protein and organism selection

Most work was done with the genomics suite Geneious 8.1.8 (Kearse et al. 2012) (<http://www.geneious.com>). Proteins were selected through Geneious's NCBI protein database "all fields" search of the domain name (SAM, TIR, and ARM) and restricted to a specific organism list that included mammals, avians, and other chordates. The search results were curated to find distinct SAM, ARM, and TIR domain-containing proteins (full set in Supplemental Table 1). Only SARM1 contained all three of these domains, the other proteins found only contained one of the three domains. A random set of proteins was also selected for comparison by searching the NCBI protein database plugin on Geneious for all human proteins available and then using a random number generator to select 100 of these proteins.

The NCBI database was also searched to acquire specific protein sequences from the organisms of interest. Two search fields called "protein name" and "organism" were used. Then, the search results were parsed by a script that filtered the proteins belonging to the organisms in our data set (all scripts are available in the Supplement). When multiple protein sequences were available for a

single organism, then the longer sequence was chosen. If this method failed to retrieve the protein, then an existing protein sequence was BLASTed through the nr database on Geneious using a BLOSUM62 matrix with a gap cost (open extended) of 11. The BLAST was additionally filtered by an Entrez query with the names of the organisms that were missing in our data set. The max e-value was set to be $1e-1$, and the word size was three. The average percentage of length recovery of the query sequence for the BLAST was ~97%. A set of avian, mammal, amphibian, reptile, and other chordates were chosen based on the organisms highly represented in the pool of SAM, ARM, and TIR domain-containing proteins.

Obtaining sequences

The gene sequences of proteins were obtained by retrieval through NCBI's Gene database (<http://www.ncbi.nlm.nih.gov/gene>) through the protein accession numbers. Scripts written with Javascript inside of Geneious were used to automate most of the tasks mentioned here and particularly to ensure that the proteins, sequences, and annotations were properly vetted. The first script gathers the accession numbers of all proteins in our data set and searches NCBI's protein database to recover the annotated protein sequences. Another script takes these results and finds proteins with the word "partial" in their description, so it can replace them with full-length entries. Then, the last script takes the resulting proteins, renames them, and builds a tree. It also validates that the sequence is correctly associated with its accessions (these are listed in Supplemental Table 2).

Ka/Ks

Ka/Ks is the ratio of nonsynonymous to synonymous mutations calculated through the alignment of two sequences with MEGA 7.0 (Kumar et al. 2016) using the Jukes Cantor algorithm (Paglia and Valentine 1967). The DNA sequences of the various domains were exported in MEGA format and imported into MEGA. These sequences were pre-aligned in Geneious using CLUSTALW (Thompson et al. 1994) alignment with an IUB cost matrix and a gap open cost value of 15.0. The Gap extend cost value was 6.66, and the free end gaps value was set to False. MEGA then found the synonymous or nonsynonymous distance matrices using the Jukes–Cantor model and the 'Gaps/missing' data option set to pairwise deletion.

Sequences of domains

To calculate Ka/Ks of domain-containing regions, a Geneious script was created which would perform EMBOSS (<http://www.ebi.ac.uk/Tools/emboss/>)

back-translation on a protein sequence and extract the coding region of the corresponding DNA sequences. Then, the domain annotations were transferred from the ambiguously back-translated protein sequence onto the same organism's DNA coding region, and these annotations were extracted for use in the Ka/Ks analysis. If a given organism did not have a domain annotation in its protein sequence, its domain-containing DNA sequence was extracted using the "Annotate from Database" function of Geneious. Only the domain annotation with the highest transfer similarity was extracted.

Sequence regions

SARM1 sequences from a mammalian (*Equus przewalskii*), an avian (*Aquila chrysaetos canadensis*), and a fish (*Esox lucius*) species were acquired based on the placement of those species in the SARM1 phylogenetic tree and the sequence length of the coding region sequence. Then, a Geneious script extracted the sequence for a 99 bp 'window' and subsequently shifted nine base pairs to extract the next 99 bp window and so on until it reached the end of the sequence. Analysis was conducted with different window sizes and shift steps, but after inspection, a 99 bp window and 9 bp shift were the optimal and corresponds to a data point every three amino acids. The window size corresponds to ~30 amino acids and was the minimum to find local changes in Ka/Ks. During this window-creating process, each extraction was given an annotation marking the base pair region that it was recovered from. These annotations were then transferred to all of the SARM1 coding region sequences and extracted from each organism. The sequences in each folder were aligned using CLUSTALW and then exported to MEGA for Ka/Ks calculations.

Evolutionary distances

Protein sequences were aligned in Geneious using CLUSTALW (Thompson et al. 1994) alignment with a BLOSUM cost matrix and a gap open cost value of 10.0. The Gap extend cost value was 0.1, and the free end gaps value was set to False. These alignments were exported in MEGA format and imported into MEGA 7.0 (Kumar et al. 2016) to attain distance matrices from six different models: number of differences, p-distance, Poisson, equal input, Dayhoff, and Jones–Taylor–Thornton (JTT). This was implemented using an AutoHotKey (<http://AutoHotkey.com>) script (available in the Supplement). To construct distance matrices, the following parameters were used: uniform rates among sites, homogenous pattern among lineages, and pairwise deletion.

Finding protein distances for TIR domain-containing proteins without their TIR domains involved creating

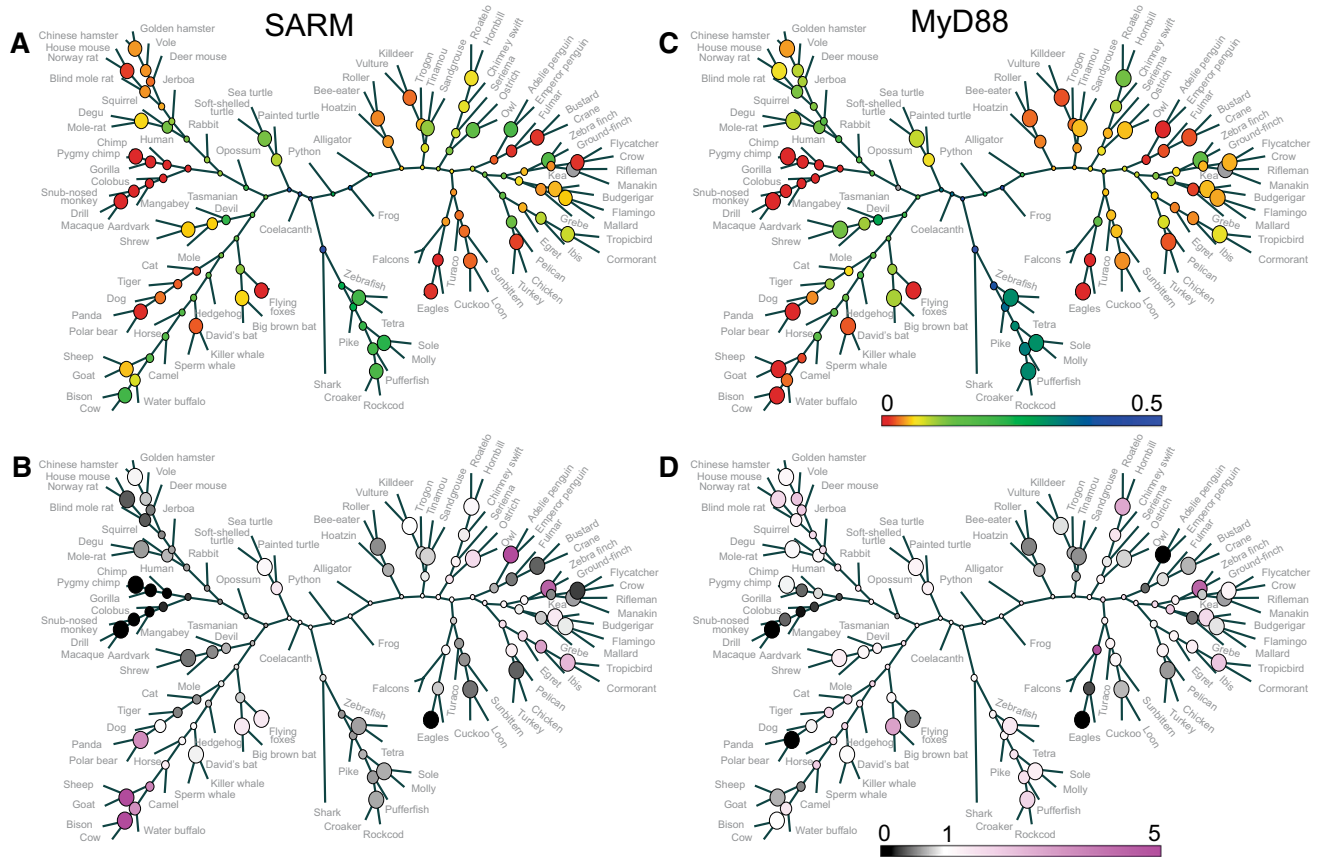


Fig. 1 Distances of *SARM1* across organisms with *MyD88* for reference. The phylogenetic tree of 107 organisms with filled circles representing the distance between nodes of the protein in question. *SARM1* is in **a** and **b**, while *MyD88* is in **c**, **d**. **a**, **c** The red–yellow–green–blue color scale on the right represents lowest (red) and highest (blue) “equal input method” distances between individual

species. As nodes get smaller and move towards the center of the tree, the color value represents the average distance between the next two nodes. **b**, **d** Normalized distances of *SARM1* and *MYD88*. This normalization limits the species–species differences as determined by the median distances of over 100 random proteins, highlighting the protein-specific divergence patterns (color figure online)

a script which extracted the protein sequence surrounding the TIR domain and concatenating the two sequences together. The protein distances were then found using MEGA as above.

Distances were “normalized” to reduce the species–species contribution and highlight the genetic contribution to the distance. This was done by dividing each distance by the median distance of the random set of proteins for that same species–species comparison. $\text{NormDistance}(\text{Species}_i, \text{Species}_j, \text{Gene}_k) = \text{Distance}(\text{Species}_i, \text{Species}_j, \text{Gene}_k) / \text{Median}[\text{Distance}(\text{Species}_i, \text{Species}_j, \text{Gene}_{[\text{RandomSet}]})]$. This is also explained graphically in Supplemental Figure 3.

Synteny

To find the various chromosomes on which *SARM1* resides in different organisms, a script utilizes the

accession numbers for *SARM1* to return the chromosome sequence from which it originates. Then, another script was used to get the gene annotations in a 1 million base pair region before and after *SARM1*. If there was a gene annotation that did not have a valid symbol (for example whose name started with “loc”), the annotation name was used to find an annotation that was homologous to the other species. For each annotation (gene), its position relative to *SARM1* or another “center” gene was found.

Analysis and statistics

Statistics were run by binning the pairwise species comparisons into ten random bins and running ANOVA with post-tests. In other analyses, linear models were used to test linear regression between variables. R (CRAN project) was used to conduct linear regression analysis (with bootstrapping) of *SARM1* by extracting species–species comparisons. Spotfire DecisionSite (Tibco, Ca) was used

to create the plots including cumulative probability histograms, synteny, and iceberg graphs. The graphics software CorelDraw (Corel Corp, Ottawa, Ca) was used to assemble the figures.

Results

SARM1's evolutionary variation from phylogenetic distance

We first sought to qualify the divergence of the SARM1 protein across 107 chordate species. As expected, most of SARM1's divergence follows the predictable evolutionary divergence relative to the source animal. In other words, related species have similar SARM1 and more distantly related species have more divergent SARM1. To see cases where SARM1 evolves in different ways, we overlaid the SARM1 distance (shown explicitly in Supplemental Figure 1) among different organisms to a generalized organism tree (Fig. 1a). The same was done for MYD88 (Fig. 1c), since it is commonly perceived as a close relative. All four of the trees in Fig. 1 have the same relationships represented by the branches, but the nodes plotted are colored by distances derived from SARM1 or MYD88. For both SARM1 and MYD88 the distances are smallest (red colors) towards the leaves of the tree but get greater as more dissimilar animals are compared. A few clades where there is high conservation of either protein (like the primates) can be observed.

Although the direct juxtaposing of the distances between SARM1 or MYD88 sequences gives some information, normalizing these distances to account for the expected phylogenetic distance reveals more about the gene-specific variations while limiting the potential influence of species–species divergence, resulting in Fig. 1b, d. The normalized value in some clades is close to 5, meaning that some species have five times higher distance for MYD88 and SARM1 than a random protein, suggesting rapid change of SARM1 and MYD88 after these animals diverged. Other clades such as the Human/Macaque clade have a relatively high conservation of SARM1 and MYD88 with a low normalized value. More generally, there is a high conservation of SARM1 in mammals compared to a random set of proteins with the exception of panda/polar bear and sheep/cow clade. These two clades show a trend entirely different from the rest of the mammals as their normalized values are over 1 and demonstrate a lowered conservation of SARM1.

Moving further away from mammals, fish show an interesting rift between SARM1 and MYD88. SARM1 is far more conserved than MYD88 relative to the median of a random set of proteins in fish, thus giving special

importance to SARM1's conserved role not only in mammals but also in fish. In general, the species–species distances of SARM1 are, with notable exceptions, lower than the median of a random set of proteins, even more so than MYD88.

We next sought to determine how SARM1's pattern-of-distances among organisms was similar to other proteins. To do this, we identified a large list of proteins that contain at least one of the three major domains of SARM1 (ARM, SAM, and TIR). For each protein, we found the evolutionary distances between pairs of organisms using six different methods: Jones–Taylor–Thornton, equal input method, p-distance (proportion of sites), Dayhoff, and Poisson model as well as a simple count of the # of differences. This constituted $(n/2) \times (n - 1) = 5671$ different sets of species–species comparisons per protein. These species comparisons were put into 4 bins, bordered by distances of 0.08, 0.188, and 0.36, based on the distributions of the distances (Supplemental Figure 2). We then aggregated all species comparisons within each distance bin to produce a median distance value (separately for each of the 250 proteins). The results of this analysis are plotted as cumulative probability histograms in Fig. 2a, b. The plots are split up based on whether the proteins were ARM domain-containing, SAM domain-containing, TIR domain-containing (with and without the TIR domain itself included in the sequence), or a set of random proteins that had none of these domains. A lower percentile demonstrates that a protein is changing rapidly and higher percentile indicates high sequence conservation. SARM1's rank relative to all of these different sets of proteins is plotted in Fig. 2c. All six methods demonstrate that SARM1 is in the 40th percentile of the ARM domain-containing proteins, suggesting that SARM1 is diverging somewhat more than other ARM domain-containing proteins. The same analysis demonstrates that SARM1 is around the 50th percentile of all SAM domain-containing proteins (thus, it is similar to SAM domain-containing proteins). This finding is tempered by the observation that SARM1 also finds itself near the middle compared with a random set of proteins (except for # of differences, where SARM1's change is somewhat greater).

Compared with other TIR domain-containing proteins, SARM1 demonstrates a significant difference, rising above the 90th percentile when compared to other TIR domain-containing proteins (Fig. 2c). On average, these TIR domain-containing proteins have larger distances, since they evolve quickly under selective pressure from outside forces (presumably pathogens). In stark contrast, SARM1 is one of the slowest to evolve, implying that its function is either protected from or independent of the traditional TIR domain evolutionary pressures. To establish the contribution of the TIR domain sequence itself in the divergence

of SARM1 among the TIR domain-containing proteins, we compared SARM1 with the TIR domain-containing proteins without the TIR domains (–TIR). The slow divergence of SARM1 indicated in Fig. 2c was preserved in the –TIR analysis, suggesting that SARM1 as a whole, not just the highly conserved TIR domain, is responsible for the slow divergence of SARM1 among TIR domain-containing proteins.

Although aggregating all the pairwise comparisons for each individual protein gave this interesting result, we thought that species–species distances of other proteins might be used to predict SARM1’s influences. We asked

whether SARM1’s species–species distances could be predicted based on simple linear combinations of distances (a linear model) from the other domain classes (ARM, SAM, and TIR). Since phylogenetic evolution of the organisms represents the most important influence on the species–species distances of proteins, we used normalized distances (in the same way as in Fig. 1b, d). Supplemental Figure 3 outlines the logic of this line of thinking. We constructed separate models for each distance calculation method (Fig. 2d). In most models, the SAM and the TIR domain significantly influenced SARM1’s distances. The highest R^2 value found among all six fits was 8%. If outlier values were removed

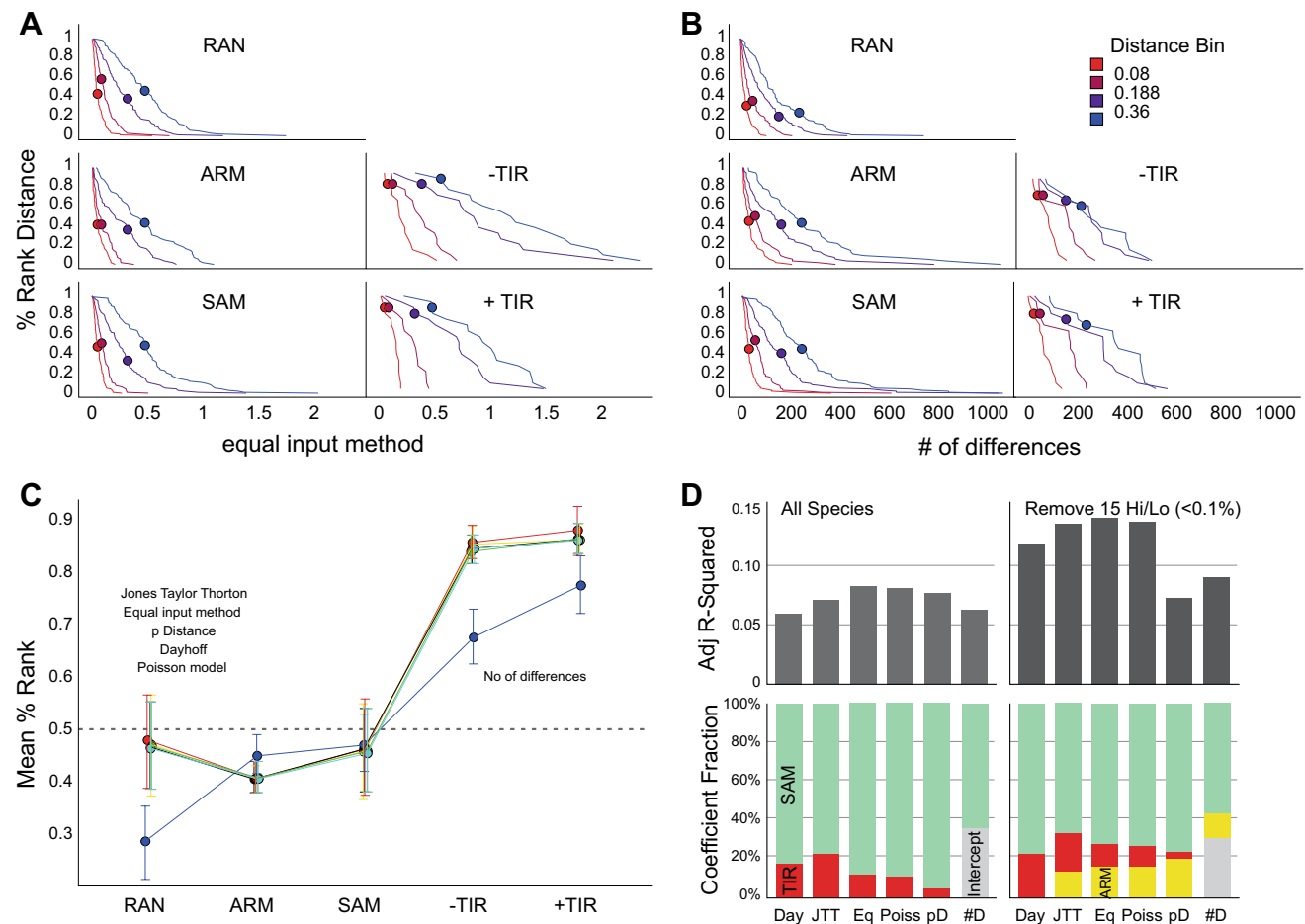


Fig. 2 Aggregate distances from pairwise comparisons of 107 species. Average amino-acid distance as calculated by 5671 species–species comparisons. **a** Cumulative probability histogram representing the average distance of proteins containing SAM, ARM, or TIR domains as calculated by the equal input method. The rank position where SARM1 falls is indicated by a filled circle. A random set of proteins is present for reference. +TIR indicates that the whole TIR domain-containing protein sequence was used to find average distance. –TIR indicates that the TIR domain was removed, and the remaining protein sequence was used to find the average distance. **b** Cumulative histograms (as in **a**) plotting the No. of differences method. **c** Mean percent rank of SARM1 in various protein groups

using different distance calculation methods. **d** Through linear regression, the species–species distances for SARM1 are predicted using species–species differences from aggregates of TIR, SAM, and ARM domain-containing proteins. Linear regression was performed for all six models. For each model, the adjusted r^2 value and the relative weight of each domain in the linear regression formula are shown (upper vs lower panels). The analysis was repeated in the right panels after removing the top and bottom 15 species–species distances of SARM1. Only significant coefficients are plotted in the coefficient fraction graphs (green SAM, red TIR, yellow ARM, gray intercept) (color figure online)

(by eliminating the lowest and highest 15 *SARM1* species–species comparisons), then the new fit predicted ~15% of the total variance of *SARM1*. The SAM domain always had the most effect on *SARM1*, with coefficients representing 70–90% explained variance. The TIR domains were second best at explaining the variance in the all-inclusive model, but the ARM domains also became predictive in the model with outliers removed. We ran multiple iterations with bootstrapping and found that the coefficients were robust to different sampling methods (not shown). Therefore, we were surprised to see that *SARM1*'s evolutionary distances between species was primarily predicted by other SAM domain-containing proteins, and only weakly predicted by other TIR domain-containing proteins.

Evolution of SARM's domains is primarily purifying

After conducting analyses on species–species protein distances, we evaluated the Ka/Ks (nonsynonymous/synonymous) ratio for just the domain-coding regions for all the proteins in our data set. These results indicate how selection was acting on *SARM1*'s domains relative to homologous domains. Ka/Ks was calculated for all 5671 pairwise species comparisons within a single protein. As above, the various species–species comparisons were split into the same four bins by distance (not binned by Ka/Ks). The Ka/Ks values for *SARM1*'s domains in these four bins are plotted in Fig. 3a. *SARM1*'s ARM domain appears to be under the least selective pressure, since the values are closest to one. The two SAM domains are each under purifying selection, and the TIR domain, with values less than 0.1, reflects a sequence subject to strong purifying selection.

The log of the Ka/Ks values was used to rank *SARM1*'s domains relative to the same domain in other proteins (Fig. 3b). Using this analysis, *SARM1*'s ARM domain and two SAM domains were found to be changing slightly more than the average ARM domain and SAM domain, respectively. Depending on the bin, the log of the average Ka/Ks value of *SARM1*'s ARM domain hovers around the 40th percentile of all ARM domains tested, and, using the same metrics, the percentile rank of the second SAM domain is, depending on the bin, in the 10–40 percentile range of all SAM domains tested. The first SAM domain is much more conserved, typically ranking in the 50th percentile of the SAM domains tested. On the other hand, *SARM1*'s TIR domain has the second lowest log of Ka/Ks value out of all of the TIR domains studied in all bins, suggesting that *SARM1*'s TIR domain is under stronger purifying selection than a typical TIR domain. Only IL1RAPL1's TIR has a lower Ka/Ks, which interestingly has also been implicated in neuro-inflammation (Yasumura et al. 2014). These findings suggest that purifying selection is acting similarly on *SARM1*'s SAM and ARM domains as it is on the SAM and

ARM domains of other proteins, but it is having a much stronger effect on *SARM1*'s TIR domain than on the TIR domain of a typical TIR domain-containing protein. The full cumulative histograms are shown for the three domains in Fig. 3c, d.

Now that we showed *SARM1*'s TIR domain is subject to almost the strongest purifying selection (compared with other TIR domain-containing proteins), we sought to uncover other regions of selection pressure. This process determined the Ka/Ks values for various regions of *SARM1*, even ones without a classified domain. We also compared the regions of selection pressure within three other TIR domain-containing proteins, MyD88 (Fig. 4a), Interleukin 18 receptor accessory protein (IL18RAP) (since it was at the 50 percentile of the TIR proteins, Fig. 4b), and Interleukin 1 Receptor Accessory Protein-like 1 (IL1RAPL1) (with the lowest TIR Ka/Ks Fig. 4c). We did the same for *SARM1* but aligned all of the 107 species with three different “template” species for comparison. *SARM1*'s coding region “template” sequence was taken from a mammal, avian, and fish species (Fig. 4d–f, respectively), and the Ka/Ks values were calculated for each. As expected, the regions with the lowest Ka/Ks values are the SAM and TIR domain regions (Fig. 4d–f). The TIR domain also has a consistently low Ka/Ks value in many of its regions, whereas both of the SAM domains and the ARM domain vary considerably in the magnitude of their Ka/Ks values. It is also notable that there is an increase in Ka/Ks in the middle of the TIR domain in all three sequence templates. There is also a spike of purifying selection (a ‘lowpoint’) in the early parts of the TIR domain suggesting an area of importance. A predominant theory about the mechanism of *SARM1*'s action involves the multimerization of SARMS through their SAM domains and, upon stress, the release of the inhibitory N-terminal sequence and attachment of the TIR domains to each other to elicit axon degeneration (Gerdt et al. 2016). The Ka/Ks plot here suggests that the spike in purifying selection of the TIR domain may be important in the TIR–TIR binding or in another important aspect of its function. Interestingly, in the TIR region, the loss of function mutations of *SARM1* [Fig. 4c in (Gerdt et al. 2013), 625–632, 661–668, 697–704] may correspond to the Ka/Ks lowpoint. Furthermore, there has been a recent revelation of a BB loop in *SARM1* TIR which contains a conserved glutamic acid residue essential for *SARM1*'s NADase function (Summers et al. 2016; Essuman et al. 2017). This loop and the amino acids surrounding it fall in what we found to be the lowest Ka/Ks region of *SARM1* (Supplemental Figure 4). This suggests the importance of *SARM1* TIR's NADase activity, required for the axon degeneration process, to its function.

While the TIR domain's low Ka/Ks value is certainly interesting, there is another Ka/Ks lowpoint in *SARM1*

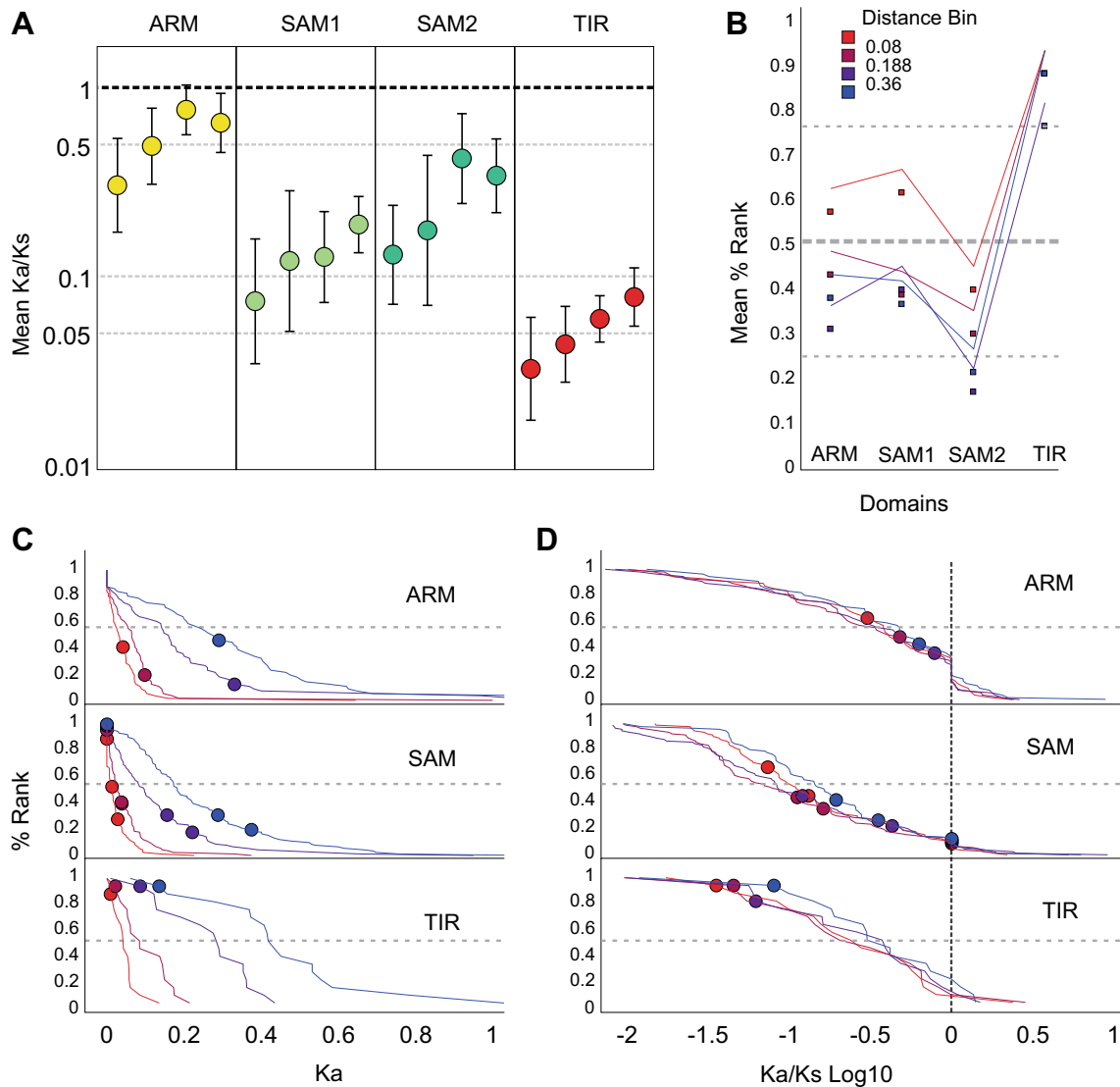


Fig. 3 Purifying selection acting on *SARM1*. **a** Mean Ka/Ks of *SARM1*'s TIR, SAM, and ARM domain-containing regions. The four points represent the four bins of pairwise species comparisons, with the first bin representing comparisons that are low distance and the last the highest distance. **b** Percent rank of the log of the Ka/Ks of *SARM1*'s ARM, SAM, and TIR domain compared to that of other ARM, SAM, and TIR domain-containing proteins. The Ka/Ks is only calculated from the domain listed. The values are derived from

the plot in **d**. **c** Cumulative probability histogram of the Ka of ARM, SAM, and TIR domains; *SARM1* is indicated by the filled circle. The SAM shows twice as many circles, since *SARM1* has two SAM domains. Each line is for a particular distances bin. **d** Cumulative probability histogram of the log of the Ka/Ks of ARM, SAM, and TIR domains with *SARM1*'s domains as filled circles. The data are separated into the same bins

which is equally fascinating. The region at the C-terminal end of the first SAM domain has a lower Ka/Ks value than any of the surrounding areas suggesting that it may be important in *SARM1* multimerization required for the eventual TIR–TIR binding and axon degeneration. Furthermore, this region, or the Ka/Ks lowpoint at the N-terminal end of the ARM domain, may be important in the release of the inhibitory N-terminal sequence upon activation. These regions of low Ka/Ks may be contributing to the high percentile rank of *SARM1* in Fig. 2c.

For comparison with *SARM1*'s 'iceberg' graph, we looked at three other genes in Fig. 4a–c. When we analyzed the StdDev of Ka/Ks across each gene itself (i.e. within a gene), the three *SARM1* templates were most similar to MyD88 (*SARMs* 0.55, Myd88 0.53, ILRAPL1 0.6, IL18RAP 0.32). In addition, the percent of the overall sequence that represented a change from purifying to positive selection (or vice versa) was highest in Myd88 and *SARM1* (30.1 and 34.6%), and lower in the other two genes (19.9% IL18RAP, 16.2% IL1RAP1). *SARM1* is

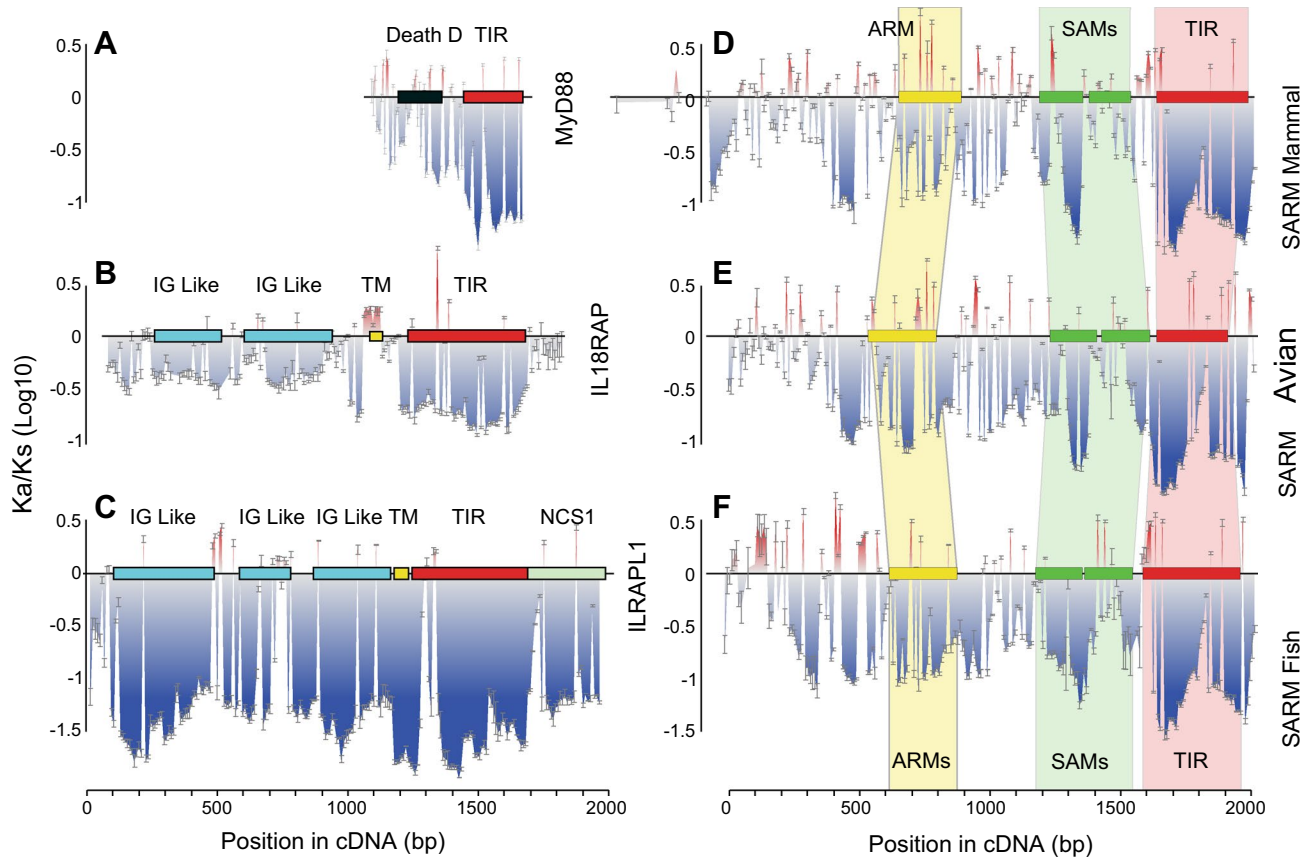


Fig. 4 Clusters of purifying selection plotted with Ka/Ks throughout *SARM1* sequence with other genes. The Ka/Ks analysis of 99 bp ‘windows’ nine bp apart along the coding sequence of various genes. The Ka/Ks values are plotted as log10 to distinguish between values over one (positive selection, red) and below one (purifying selection, blue). The height plotted represents the log of the average Ka/Ks of

up to 5761 species–species comparisons (with error bars representing StDev). The horizontal rectangles represent the location of various domains. **a** ‘Iceberg’ graph of *MYD88*. **b** Graph of *IL18RAP*. **c** Graph of *ILRAPL1*. **d–f** Graphs of *SARM1* with template generated from a mammal (*Equus przewalskii*), a bird (*Aquila chrysaetos canadensis*), and a fish (*Esox lucius*) (color figure online)

under strong purifying selection pressure, but its pattern of Ka/Ks change is still different than the other purified gene *IL1RAPL1*, making it an evolutionarily distinct member of the TIR domain-containing protein family.

SARM1 has high GC content among TIR domain-containing proteins

To determine any potential effects of GC content on *SARM1* evolutionary history, we recorded the %GC within the coding regions (CDs), the whole gene (Gene), and the intronic regions (Gene–CDs). Teleologically, one might expect regions of coding sequences, specifically protein domains, to contain low levels of GC sequences, minimizing the rate of nonsynonymous mutation while maintaining both protein function and regulation of *SARM1* expression. If higher GC content was demonstrated within the *SARM1* CDS than in the surrounding gene, we would predict the GC content codes for essential functional or regulatory

requirements of *SARM1*. We found that *SARM1*, while highly conserved across TIR domain-containing proteins, contains high GC content when compared with other TIR domains (from the TIR domain-containing family) (Fig. 5c) falling in the highest percentile for GC content. When compared between species, reptiles (including birds) have high GC content in the noncoding and intronic regions of *SARM1*, while mammalian *SARM1* has high GC content within CDs (Fig. 5a). When the whole gene GC content is ranked and all the animals are put into a waterfall plot, a distinct pattern for mammalian *SARM1* is revealed. While most *SARM1* genes correlate GC content between both coding and noncoding regions, mammals have a strong divergence of the coding and noncoding GC content, in the range of 15% (Fig. 5b).

When we compare regions within *SARM1* for Ka and Ka/Ks and GC content, we see that regions of high Ka often correspond with high Ka/Ks (with heavy positive selection Supplemental Figure 5). Several of these

high K_a regions also correspond with high GC content. Prominently, the SAM domain region functions as an area of high positive correlation (>0.5) between the K_a and GC contents. This region of high correlation is followed by a low negative correlation, marking the SAM and TIR domain transition. These variations between K_a and GC content correlation potentially highlight different epigenetic regulation or functional requirements placed between regions. Finally, while we found evidence of a positive correlation between K_a and GC content, no strong correlation was found between K_a/K_s and GC content (data not shown), perhaps, indicating a combination of both purifying and selective pressures on GC content within the CDs.

SARM1 genomic context and synteny

We next evaluated the synteny of the genome around *SARM1*. For 47 organisms, the chromosome sequence containing *SARM1* was acquired, and genes within a 1 megabp window before and after *SARM1* were evaluated. Then, these genes and their positions relative to *SARM1* were graphed (Fig. 6a). Positions of the genes DNA polymerase delta interacting protein 2 (*POLDIP2*), transmembrane protein 199 (*TMEM199*), Vitronectin (*VTN*), *SARM1*, solute carrier family 46 member 1 (*SLC46A1*), Forkhead box N1 (*FOXN1*), and UNC-119 lipid binding chaperone (*UNC119*) are extremely well preserved in both mammals and avian species (with the exception of turkey) suggesting that their arrangement may be functionally important. A smaller set of genes, from *VTN* to *UNC119*, are present in almost all of the 47 species tested (Fig. 6b, c). In contrast, certain gene positions relative to *SARM1* are preserved only in certain clades. Kinesin family member 12's (*KIF12*) position next to *VTN* in avians is replaced by SEBOX homeobox (*SEBOX*) in mammals for example. Furthermore, there are genes farther away from *SARM1* whose positions are stable over a short position span in *Aves*, but these genes are located elsewhere in mammals.

To look at *SARM1*'s synteny in perspective, we compared *SARM1*'s synteny with that of 24 other genes (Supplemental Figure 6). As examples, we have included the synteny graphs of three genes: *NMNAT1* (important to axon degeneration), *STEAP3* Metalloreductase (*STEAP3*, chosen at random from the larger random group of proteins), and *ILIRAPL1* in Fig. 6d. *SARM1* and *NMNAT1* have similar syntenic profiles, while *ILIRAPL1* displays a poor synteny when compared to *SARM1*. *STEAP3* is somewhere between *NMNAT1* and *ILIRAPL1* in its syntenic similarity to *SARM1*. After analyzing the gene regions around each of 24 genes (Soderlund et al. 2006; Ghiurcuta and Moret 2014), we found that *SARM1* is contained in a chromosomal region with high regional density of genes. However,

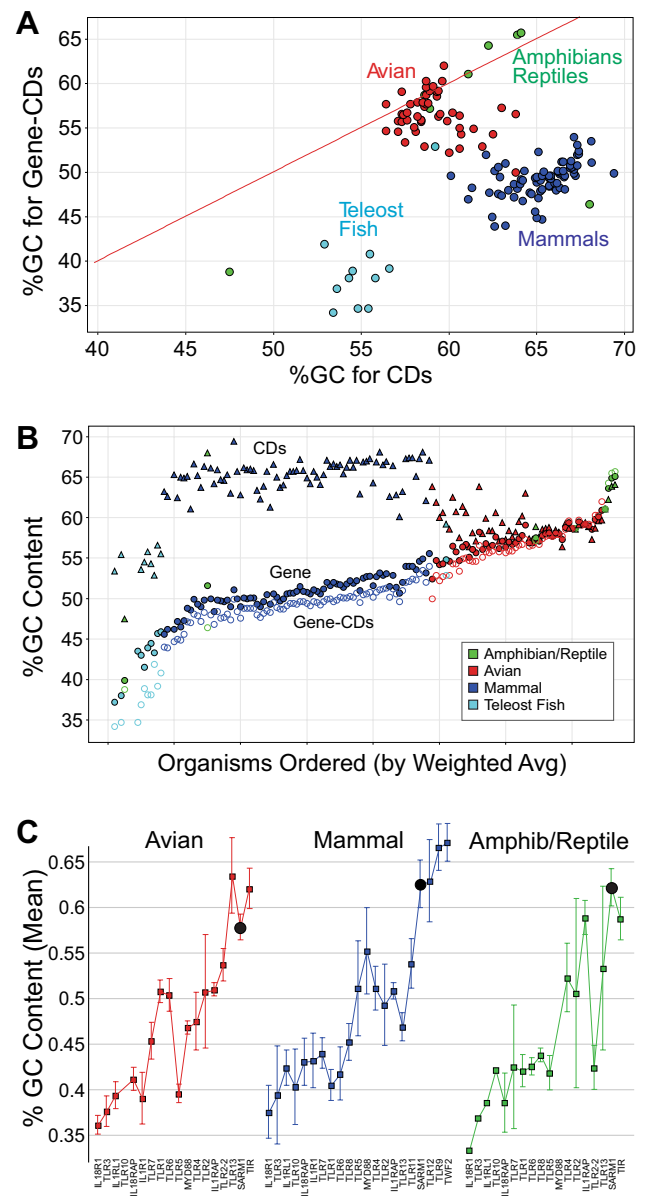


Fig. 5 GC content of *SARM1* in different groups of animals. **a** Each marker represents a single organism, and it is plotted for its *SARM1* GC content within the coding region (CDs, *x*-axis) and its nonexonic GC content (Gene-CDs, *y*-axis). Clustering of animal classes can be seen clearly. **b** Same analysis now plotted, so that the *x*-axis ranks the animals by (primarily) their GC content over the whole gene (*closed circles* gene; *open circles* gene-CDs; *triangles* CDs). **c** GC content within the TIR domain for several TIR domain-containing proteins, including *SARM1*. Each marker plots the average and standard deviation of all the organisms in the stated clade. *SARM1* is marked by a black filled circle

to further test whether this high regional density of genes corresponded to a high level of organization (high synteny), we examined three values: # co-occurrence, % co-occurrence, and correlation. The number of genes co-occurring in multiple organisms was highest in *SARM1* by almost

two folds, but the percent co-occurrence, which accounts for high local density of genes by dividing # co-occurrence by the total number of genes present in the 2 MBP region, had a below average value for *SARM1* compared to the other genes. In addition, the number of genes that occur in the same organization (order) among the organisms tested (correlation of gene order) was also lower for *SARM1* than the average gene analyzed. Although our methodology for examining the synteny of *SARM1* has benefits not associated with other synteny analyses (such as ease of comparability among species and clear quantification of synteny), its reliance on annotations, length of chromosome analyzed, and the physical location of the gene on the chromosome will affect the results.

Discussion

Our analysis of *SARM1* reveals a highly regular and conserved genetic sequence across species. Such a strict structure underlies a potential evolutionary pressure to regulate the protean functionality of *SARM1*. Protein distances found using 5671 species–species comparisons and with six different distance estimation models revealed *SARM1*'s distance to be the second lowest of all TIR domain-containing proteins tested, with or without the TIR domain region taken into account. In addition, when we modeled *SARM1*'s species–species distances only from its domains, most variance was predicted by *SARM1*'s SAM domain (70–90%) and only minimally from its TIR domain. Then, we tried to answer questions about how purifying selection was acting on different regions of *SARM1*. Purifying selection acted on *SARM1*'s SAM and ARM domain similarly to other SAM and ARM domains, but it operated on *SARM1*'s TIR to a much greater extent than other TIR domains. This suggests that, contrary to the popular notion of *SARM1* as a TIR domain molecule, *SARM1*'s evolutionary behavior is unlike other TIR domain-containing proteins.

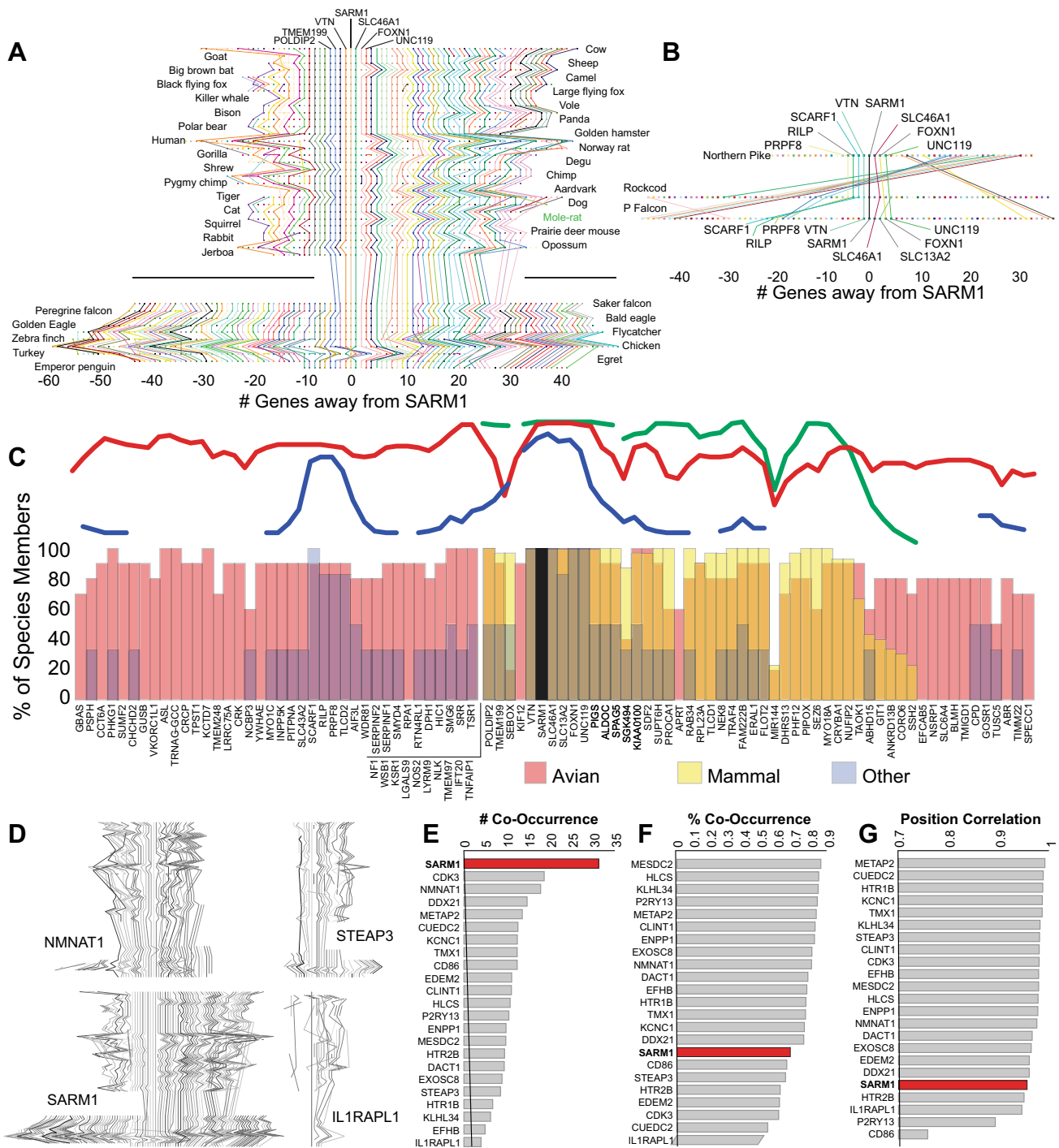
After getting an idea of where *SARM1* falls relative to other proteins, we wanted to uncover regions (not necessarily known domains) of *SARM1* undergoing purifying selection. We created a Ka/Ks map of *SARM1* and found that its TIR domain region was highly purified with regions in the first SAM domain and a region just before the ARM also showing extremely low Ka/Ks value. Despite generally low Ka/Ks values across the *SARM1* gene, there were quite a few spikes of positive selection at various points on the gene, an unusual pattern compared to the other genes tested. With the results of the protein distance and Ka/Ks analyses in mind, *SARM1* further distinguishes itself from the group of TIR domain-containing proteins.

The process of 5-methyl cytosine transition to thymine is believed to have dramatically reduced the number of CG

sequences found within mammalian genomes to less than 21% the expected frequency (when assuming a random succession of guanosine following cytosine) (Lander 2011). Regions with high GC content can benefit from increased regulation and a predicted increased rate of transitions ($C \rightarrow T$, $G \rightarrow A$) accounting for some established pathological conditions (Cooper and Youssoufian 1988). Interestingly, regional maintenance of high GC content within the *SARM1* CDS generally exhibits a positive correlation with an increased rate of nonsynonymous mutations. However, GC content and Ka/Ks juxtaposition revealed no significant correlation, exposing regions caught between the functional requirement of GC content and the steady pull of spontaneous deamination.

Once we were done characterizing the *SARM1* gene itself, we concerned ourselves with looking at genes around *SARM1* and the conservation of their position around the *SARM1* gene. Taking the chromosomal sequence of over 47 organisms, we found a highly conserved cluster of genes around *SARM1* from *VTN* to *UNC119*. In addition, when we juxtaposed the *SARM1* synteny to the synteny of other genes, *SARM1* was found to have a high local density of genes (high # co-occurrence), but these genes did not always maintain a precise order, as *SARM1* fell below the average % co-occurrence and correlation. Although *SARM1* may not have high synteny as determined by our measures, our findings imply that the genes immediately around *SARM1* may be important in *SARM1*'s regulation and function due to their presence in a wide range of species.

Since its initial discovery, *SARM1*'s role in the immune system has widely been documented, but, recently, *SARM1* TIR domain's role as an NADase has been described, and a disruption of its catalytic ability blocks NAD⁺ depletion and subsequent degeneration of the axon. *SARM1* also displays saturation kinetics characteristic of enzymes and shares functional similarities with ADP-ribosyl cyclase family of enzymes. Prior to (Summers et al. 2016; Essuman et al. 2017), no other TIR domain had been associated with enzymatic function, so *SARM1* may be functionally very different from other TIR domain-containing molecules. Additional recent work by (Zhang et al. 2011; Patot et al. 2017) has provided some corroborating evidence by proposing that *SARM1* may not have evolved like other TIR domain-containing proteins, as it clusters with bacterial TIR domain proteins rather than animal TIR domain proteins. The evolutionary similarity between *SARM1* and bacterial TIR domain-containing proteins may be due to a horizontal gene transfer event which occurred between bacteria and the ancestor of ecdysozoa and deuterostomes. Certainly, this bacterial past could explain *SARM1*'s distinguishing role among other TIR domain-containing adaptor proteins as a negative regulator of TLR signaling. This



negative regulation of TLRs has been found in Methicillin-resistant *S. aureus*'s (MRSA) TIR domain protein TirS which diminishes the response of macrophages to infection (Patot et al. 2017). More interestingly, the protein is found in the staphylococcal chromosomal cassette (SCC) which is a mobile genetic element; perhaps, SARM1 may have been part of a mobile genetic element which got integrated into the genome of the ecdysozoan/deuterostome ancestor. The results mentioned above suggest that SARM1 may not

behave like a typical TIR domain-containing protein, but the importance of SARM's immune system role to its evolution has not been adequately described and knocking out SARM1 may have devastating effects on the function of immune cells.

We not only provide an extensive genomics analysis of SARM1 but also demonstrate that SARM1 did not evolve as a classic TIR domain-containing protein, suggesting that knocking out SARM1 may not have a dramatic effect on

Fig. 6 *SARM1* is in a highly organized gene cluster and has high local gene density. A synteny analysis was run for 47 animals. **a** Positions of various genes in 47 animals from the perspective of the *SARM1* gene. Each colored line represents a different gene, and each gene's position in a group of animals is represented by the path through which its line travels. The top half of the figure contains mammals, whereas the bottom half contains avian species. Genes that are only present in a small number of organisms or are far away from *SARM1* are omitted. **b** Positions of genes around *SARM1* in a falcon and two fish. **c** Bar graph showing the percentage of avian, mammalian, and other species that have a given gene at a position around *SARM1*. Brown indicates the presence of all three distinguishing groups, whereas orange indicates the presence of mammals and avians. The dark purple region indicates the simultaneous presence of avians (reddish-pink) and other species (light purple). Another representation of a particular gene's co-occurrence is shown just above the bar graph where it is perhaps easier to see the patterns. **d** Synteny graphs of *NMNAT1*, *STEAP3*, *ILIRAP1*, and *SARM1*. After a synteny analysis was conducted for 24 genes, we quantified synteny by describing the **e** #co-occurrence: the number of genes around the gene of interest (on the *x*-axis) that co-occur averaged across every pair of organism tested, **f** % co-occurrence: #co-occurrence divided by the total number of genes present in the 2 MBp segment analyzed around the gene of interest and **g** position correlation. The Pearson's correlation between two organisms in terms of the order (position) in which the genes occur. The name of the gene of interest (that the synteny was centered around) is on the *y*-axis, while the *x*-axis contains the average value taken from comparison of chromosome sequences from pairs of 47 animals (color figure online)

the immune system. However, *SARM1*'s high conservation is noteworthy, since it may result from *SARM1*'s prominence in other pathways or functions, such as its NADase activity. Still, there is more to learn about *SARM1*'s less debilitating functions, and we provide some groundwork through an evolutionary genomics approach which may be useful for future interpretations of *SARM1*'s presence in organisms.

Acknowledgements We would like to thank Benjamin Swack for his help with the R models in this project. Jessica Crowley, James Cole, Micaïlya Mattson, and Lyndah Lovell also provided skilled assistance. Thanks to the Department of Biology and specifically Josh Puzey for software suggestions and Paul Heideman for writing assistance.

Data availability The software scripts written for use with Geneious are available as source code in the supplement. The AutoHotKey scripts are also available in the supplement.

Compliance with ethical standards

Funding William & Mary Charles Center, William & Mary Monroe Scholarship, and William & Mary Department of Biology Ferguson Fund.

Conflict of interest All authors declare they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Araki T, Sasaki Y, Milbrandt J (2004) Increased nuclear NAD biosynthesis and SIRT1 activation prevent axonal degeneration. *Science* 305:1010–1013
- Beirowski B, Babetto E, Coleman MP, Martin KR (2008) The *WldS* gene delays axonal but not somatic degeneration in a rat glaucoma model. *Eur J Neurosci* 28:1166–1179
- Belinda LW-C, Wei WX, Hanh BTH, Lei LX, Bow H, Ling DJ (2008) *SARM*: a novel Toll-like receptor adaptor, is functionally conserved from arthropod to human. *Mol Immunol* 45:1732–1742
- Bramley JC, Collins SVA, Clark KB, Buchser WJ (2016) Avian axons undergo Wallerian degeneration after injury and stress. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol* 202:813–822
- Cooper DN, Youssoufian H (1988) The CpG dinucleotide and human genetic disease. *Hum Genet* 78:151–155
- Couillault C, Pujol N, Reboul J, Sabatier L, Guichou J-F, Kohara Y, Ewbank JJ (2004) TLR-independent control of innate immunity in *Caenorhabditis elegans* by the TIR domain adaptor protein TIR-1, an ortholog of human *SARM*. *Nat Immunol* 5:488–494
- Essuman K, Summers DW, Sasaki Y, Mao X, DiAntonio A, Milbrandt J (2017) The *SARM1* Toll/Interleukin-1 receptor domain possesses intrinsic NAD⁺ cleavage activity that promotes pathological axonal degeneration. *Neuron* 93(1334–1343):e5
- Fornarino S, Laval G, Barreiro LB, Manry J, Vasseur E, Quintana-Murci L (2011) Evolution of the TIR domain-containing adaptors in humans: swinging between constraint and adaptation. *Mol Biol Evol* 28:3087–3097
- Furman AC, Kujawa SG, Liberman MC (2013) Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates. *J Neurophysiol* 110:577–586
- Geisler S, Doan RA, Strickland A, Huang X, Milbrandt J, DiAntonio A (2016) Prevention of vincristine-induced peripheral neuropathy by genetic deletion of *SARM1* in mice. *Brain* 139:3092–3108
- Gerds J, Summers DW, Sasaki Y, DiAntonio A, Milbrandt J (2013) *Sarm1*-mediated axon degeneration requires both SAM and TIR interactions. *J Neurosci* 33:13569–13580
- Gerds J, Brace EJ, Sasaki Y, DiAntonio A, Milbrandt J (2015) *SARM1* activation triggers axon degeneration locally via NAD⁺ destruction. *Science* 348:453–457
- Gerds J, Summers DW, Milbrandt J, DiAntonio A (2016) Axon self-destruction: new links among *SARM1*, MAPKs, and NAD⁺ metabolism. *Neuron* 89:449–460
- Ghiurcuta CG, Moret BME (2014) Evaluating synteny for improved comparative studies. *Bioinformatics* 30:i9–i18
- Henninger N, Bouley J, Sikoglu EM, An J, Moore CM, King JA, Bowser R, Freeman MR, Brown RH (2016) Attenuated traumatic axonal injury and improved functional outcome after traumatic brain injury in mice lacking *Sarm1*. *Brain* 139:1094–1105
- Hou Y-J, Banerjee R, Thomas B, Nathan C, Garcia-Sastre A, Ding A, Uccellini MB (2013) *SARM* is required for neuronal injury and cytokine production in response to central nervous system viral infection. *J Immunol* 191:875–883
- Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28:1647–1649
- Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis Version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870–1874
- Lander ES (2011) Initial impact of the sequencing of the human genome. *Nature* 470:187–197

- Lunn ER, Perry VH, Brown MC, Rosen H, Gordon S (1989) Absence of Wallerian degeneration does not hinder regeneration in peripheral nerve. *Eur J Neurosci* 1:27–33
- Osterloh JM, Yang J, Rooney TM, Fox AN, Adalbert R, Powell EH, Sheehan AE, Avery MA, Hackett R, Logan MA, MacDonald JM, Ziegenfuss JS, Milde S, Hou Y-J, Nathan C, Ding A, Brown RH, Conforti L, Coleman M, Tessier-Lavigne M, Züchner S, Freeman MR (2012) dSarm/Sarm1 is required for activation of an injury-induced axon death pathway. *Science* 337:481–484
- Paglia DE, Valentine WN (1967) Studies on the quantitative and qualitative characterization of erythrocyte glutathione peroxidase. *J Lab Clin Med* 70:158–169
- Panneerselvam P, Singh LP, Selvarajan V, Chng WJ, Ng SB, Tan NS, Ho B, Chen J, Ding JL (2013) T-cell death following immune activation is mediated by mitochondria-localized SARM. *Cell Death Differ* 20:478–489
- Patot S, Imbert P, Baude J, Martins Simões P, Campergue J-B, Louche A, Nijland R, Bès M, Tristan A, Laurent F, Fischer A, Schrenzel J, Vandenesch F, Salcedo S, François P, Lina G (2017) The TIR homologue lies near resistance genes in *Staphylococcus aureus*, coupling modulation of virulence and antimicrobial susceptibility. *PLoS Pathog* 13:e1006092
- Soderlund C, Nelson W, Shoemaker A, Paterson A (2006) SyMAP: a system for discovering and viewing syntenic regions of FPC maps. *Genome Res* 16:1159–1168
- Summers DW, Gibson DA, DiAntonio A, Milbrandt J (2016) SARM1-specific motifs in the TIR domain enable NAD⁺ loss and regulate injury-induced SARM1 activation. *Proc Natl Acad Sci USA* 113:E6271–E6280
- Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W. *Nucleic Acids Res* 22:4673–4680
- Vérièpe J, Fossouo L, Parker JA (2015) Neurodegeneration in *C. elegans* models of ALS requires TIR-1/Sarm1 immune pathway activation in neurons. *Nat Commun* 6:7319
- Yasumura M, Yoshida T, Yamazaki M, Abe M, Natsume R, Kanno K, Uemura T, Takao K, Sakimura K, Kikusui T, Miyakawa T, Mishina M (2014) IL1RAPL1 knockout mice show spine density decrease, learning deficiency, hyperactivity and reduced anxiety-like behaviours. *Sci Rep* 4:6613
- Zhang Q, Zmasek CM, Cai X, Godzik A (2011) TIR domain-containing adaptor SARM is a late addition to the ongoing microbe-host dialog. *Dev Comp Immunol* 35:461–468