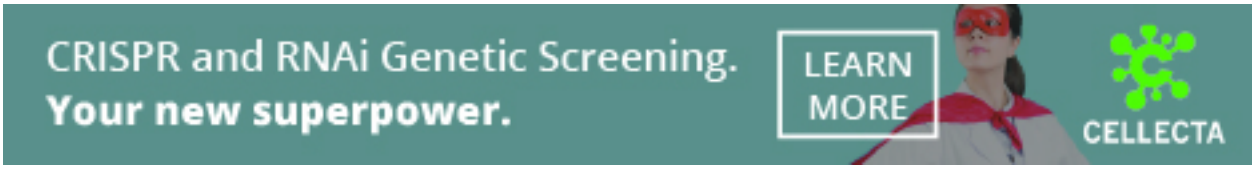
Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants

James H. Thomas

Genome Res. 2006 16: 1017-1030
Access the most recent version at doi:10.1101/gr.5089806



Letter

Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants

James H. Thomas

Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

Host–pathogen arms races can result in adaptive evolution (positive selection) of host genes that mediate pathogen recognition and defense. To identify such genes in nematodes, I used maximum-likelihood analysis of codon evolution to survey all paralogous gene groups in *Caenorhabditis elegans*. This survey found robust evidence of positive selection in two classes of genes not previously implicated in pathogen defense. Both classes of genes encode ubiquitin-dependent proteasome adapters, which recruit diverse substrate proteins for poly-ubiquitination and proteolysis by Cullin-E3 ubiquitin-ligase complexes. The adapter proteins are members of the F-box superfamily and the MATH-BTB family, which consist of a conserved Cullin-binding domain and a variable substrate-binding domain. Further analysis showed that most of the ~520 members of the F-box superfamily and ~50 members of the MATH-BTB family in *C. elegans* are under strong positive selection at sites in their substrate-binding domains but not in their Cullin-binding domains. Structural modeling of positively selected sites in MATH-BTB proteins suggests that they are concentrated in the MATH peptide-binding cleft. Comparisons among three *Caenorhabditis* species also indicate an extremely high rate of gene duplication and deletion (birth–death evolution) in F-box and MATH-BTB families. Finally, I found strikingly similar patterns of positive selection and birth–death evolution in the large F-box superfamily in plants. Based on these patterns of molecular evolution, I propose that most members of the MATH-BTB family and the F-box superfamily are adapters that target foreign proteins for proteolysis. I speculate that this system functions to combat viral pathogens or bacterial protein toxins.

[Supplemental material is available online at www.genome.org.]

Host genes encoding proteins directly involved in recognizing pathogens are expected to be subject to unusual patterns of molecular evolution, driven by an arms race with the pathogens. One expected pattern, typified by mammalian MHC genes, includes site-specific adaptive evolution (positive selection) and a high degree of population polymorphism (Hughes and Nei 1988, 1989; Hughes et al. 1990; Swanson et al. 2001). Positive selection is often detected by a rate of nonsynonymous codon change higher than synonymous codon change, a pattern the reverse of that produced by the more common purifying (negative) selection. Such positive selection in MHC proteins results in regions of rapidly evolving amino acid sequence that interact with foreign proteins, interspersed with regions of highly conserved amino acid sequence that form the structural core of the protein (Hughes and Nei 1988, 1989; Hughes et al. 1990).

To identify genes that are candidates for pathogen interaction in *Caenorhabditis elegans*, I conducted a systematic test for positive selection. Lack of sufficient population sequence data and the absence of close sibling species eliminate two of the methods used to detect positive selection. However, recent paralogous gene duplicates can be analyzed for evidence of positive selection acting on the paralogs relative to each other (Thomas et al. 2005). To apply this method systematically, I clustered the entire gene complement of *C. elegans* to define 544 paralog groups and analyzed each paralog group for positive selection by the maximum-likelihood method of Yang and Nielsen (Yang 1997; Yang and Nielsen 2000). The most prominent novel gene classes identified in this search were the MATH-BTB family and the F-box superfamily (PFAM domains PF00917, PF00651, and PF00646, respectively). F-box and MATH-BTB proteins function as adapters that target substrate proteins for poly-ubiquitination and proteolysis. Ubiquitin-dependent protein degradation is initiated by the transfer of ubiquitin to substrate proteins by E3 ubiquitin ligases. Ubiquitinated substrate proteins are then targeted to the 26S proteasome for degradation (Moon et al. 2004; van den Heuvel 2004; Varshavsky 2005). Substrates for ubiquitination are recruited by large Cullin complexes (also called SCF complexes), which include the E3 ligase, regulatory subunits, a Cullin scaffold protein, and an adapter protein that binds specific substrate proteins. There are several distinct Cullin complexes, which differ primarily in the Cullin scaffold protein and adapter proteins (Fig. 1). Each specific Cullin protein uses a distinct class of adapter protein.

Several members of the F-box superfamily are known adapters for Cullin1 complexes (Bai et al. 1996; Winston et al. 1999; Zheng et al. 2002; Jin et al. 2004). The F-box domain binds to Cullin1 via Skp1-related (Skr) proteins (Bai et al. 1996; Zheng et al. 2002); diverse regions outside the F-box domain bind to specific substrate proteins (Winston et al. 1999; Hsiung et al. 2001; Brunson et al. 2005; Nayak et al. 2005). In these adapter proteins, the F-box is near the N terminus, and the remainder of the protein falls into several families, including kelch repeat, WD-40 repeat, LRR, FTH, FBA, FBA1, and FBA2 domain-containing families (http://www.sanger.ac.uk/Software/Pfam/2005) (Jiang and Struhl 1998; Ilyin et al. 1999; Winston et al. 1999; Clifford et al. 2000; Andrade et al. 2001; Gagne et al. 2002). Studies in this

E-mail jht@u.washington.edu; fax (206) 685-4467.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5089806.

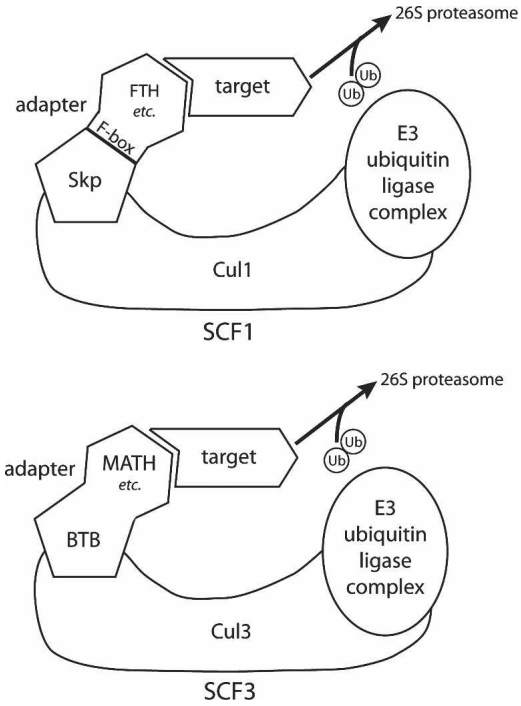


Figure 1. Schematic of ubiquitin-targeting system. The *top* panel shows the SCF1 (Cullin1) complex, which uses Skp-related and F-box proteins as substrate adapters. The domain marked "FTH *etc.*" varies depending on the specific adapter. The *bottom* panel shows the SCF3 (Cullin3) complex, which uses BTB proteins as substrate adapters. The domain marked "MATH *etc.*" varies depending on the specific adapter. (Ub) Ubiquitin.

paper focus mostly on the two largest F-box families in *C. elegans*, the F-box-FTH and the F-box-FBA2 families. Very little is known about the FTH and the FBA2 domains; both are classified as sequence domains of unknown function that are present in large numbers of nematode proteins (http://www.sanger.ac.uk/Software/Pfam/2005). They have no known sequence relationship to each other or to any other known protein domain.

Many BTB proteins are adapters for Cullin3 complexes (Furukawa et al. 2003; Pintard et al. 2003; Xu et al. 2003; Figueroa et al. 2005). The BTB domain binds directly to Cullin3, and other domains in the BTB protein confer substrate specificity. Like F-box proteins, BTB-containing adapters have a variety of substrate-binding domains, including MATH, WD-40 repeat, Zn-finger repeat, and kelch repeat domains (Winston et al. 1999; Pintard et al. 2003; Prag and Adams 2003; Xu et al. 2003; van den Heuvel 2004; Brunson et al. 2005; Figueroa et al. 2005; Stogios et al. 2005). In *C. elegans*, most of these adapters have a BTB domain near their C terminus and an N-terminal MATH domain, which is responsible for substrate binding (Pintard et al. 2003; Xu et al. 2003).

Taken together, the ~520 F-box and 50 MATH-BTB genes in *C. elegans* account for ~2.5% of total coding potential. Given their number, remarkably little is known about these genes; for example, only a few have been identified in forward genetic screens, and the vast majority of the genes tested by RNAi have no observed phenotype (Kamath et al. 2003). Based on results presented in this paper, I propose that most of the genes function to target foreign proteins for degradation as part of the innate immune system.

Results

Global test for positive selection among paralogs

All *C. elegans* gene families with three or more recent duplicates were tested for evidence of positive selection by analyzing rates of nonsynonymous (d_N) and synonymous (d_S) codon evolution (see Methods). Briefly, a complete set of predicted coding sequences was translated and clustered by protein similarity, which generated 544 groups with three or more closely related paralogs, including a total of 2878 genes (see Methods). Coding sequences for each group of genes were subjected to a standardized process of codon alignment and maximum-likelihood analysis of d_N/d_S ratios. The results from this analysis are summarized in Supplemental Table S1. Among the 544 groups tested, this method identified 80 groups of paralogs that showed potentially significant evidence of positive selection (false discovery rate <5%). Failure to identify a specific paralog group in this test does not indicate lack of positive selection; for example, among paralogous genes subject to positive selection, no Srz genes (Thomas et al. 2005) and only a minority of F-box genes were identified (this study). This high false-negative rate is probably due to a combination of suboptimal automated clustering, uncurated alignments, and gene prediction errors. Despite these limitations, three gene families or superfamilies were identified repeatedly among the 80 paralog groups with evidence for positive selection: nine paralog groups in the F-box superfamily, four groups in the MATH-BTB family, and four groups in the C-type lectin superfamily. The C-type lectin superfamily is strongly implicated in innate immunity to bacteria and fungi in a variety of organisms (Kogelberg and Feizi 2001; Lu et al. 2002; Kanost et al. 2004; McGreal et al. 2004), although analysis in *C. elegans* is just beginning (Nicholas and Hodgkin 2004). In contrast, the F-box families and the MATH-BTB family are not known to be involved in innate immunity. Because their repeated identification in this global analysis suggests that positive selection is widespread in F-box and MATH-BTB families, these families were investigated in detail.

F-box domain families

The F-box domain is ~40 amino acids long and in all well-studied cases acts as a Cullin1 adapter for ubiquitin-mediated proteolysis (Bai et al. 1996; Schulman et al. 2000). Based on Ψ-BLAST and rps-BLAST searches (Altschul et al. 1997; Marchler-Bauer and Bryant 2004), I found that ~520 genes in *C. elegans* potentially encode a protein with a clear F-box domain (an additional 50 genes, not analyzed here, probably contain a highly divergent F-box like domain). About 40 of the 520 genes are predictions that appear to include two copies of the F-box and associated sequences; it was unclear whether or not these are gene prediction errors, and they were not further analyzed. Most of the remaining 480 genes fall into two broad families: ~220 contain an FTH domain and ~210 contain an FBA2 domain. In both families an N-terminal F-box domain is followed by a more divergent region of ~300 amino acids, which contains the FTH or FBA2 domain (Fig. 2). The FTH and FBA2 domains have no detectable sequence similarity to each other, and both appear thus far to be nematode-specific. About 80 members of the F-box-FTH family share an additional ~50-amino-acid domain N-terminal to their F-box domain. This unnamed domain is distantly related to the DNA-binding domain of mariner transposes (Fig. 2; data not shown).

My initial analysis focused on the 140 genes in the F-box-

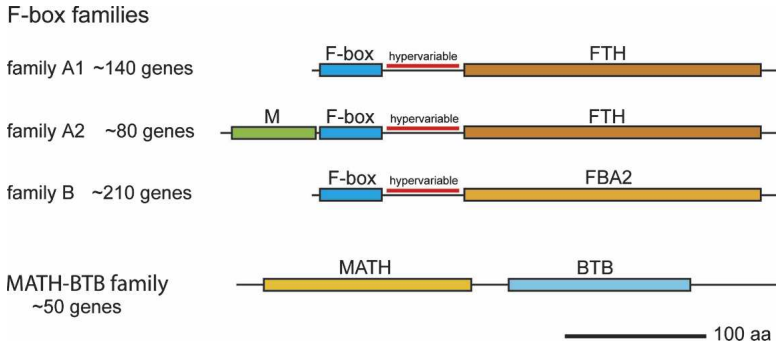


Figure 2. Schematics of F-box and MATH-BTB protein domains. Domain schematics of the main types of proteins analyzed in this paper, including three types of F-box domain proteins and MATH-BTB domain proteins. Domains that bind Cullins are shades of blue, and domains that bind substrate are shades of yellow to brown. F-box family A2 has an additional domain (M for mariner, shown in green) of unknown function that is related to the DNA-binding domain of mariner transposases. The region of highest sequence diversity in each F-box family is labeled hypervariable.

FTH family that lack the mariner-related domain; these have been assigned to the gene class *fbxa* (F-box family A). A sizeable fraction of putative *fbxa* loci are either defective genes or encode variant proteins lacking substantial parts of the typical FBXA protein (see Methods). Attempts to disentangle these possibilities using sequence alignment, improved gene predictions, and existing transcript data were not fully satisfactory; additional experimental evidence will be required to clarify which *fbxa* genes are likely to be functional. Nevertheless, focused gene annotation efforts generated improved gene predictions for many *C. elegans* and *Caenorhabditis briggsae fbxa* genes, added a few genes that were previously unpredicted, and generated gene models for 228 putative *fbxa* genes in the newly sequenced *Caenorhabditis remanei* (see Methods).

Birth–death evolution in the F-box superfamily

As shown below, the F-box and MATH-BTB families include two classes of genes based on evolutionary stability: one class with clear well-conserved orthologs in *C. elegans*, *C. briggsae*, and *C. remanei*; and a second class without clear orthologs that is undergoing rapid birth–death evolution. For simplicity, I refer to these as "stable" genes and "unstable" genes, in reference to their apparent rates of gene duplication and deletion. An FBXA protein tree is shown in Supplemental Figure S1 and summarized in Table 1, including all identified unstable genes that encode at least 80% of an alignable F-box-FTH protein from the three *Caenorhabditis* species. The tree is remarkable for containing several large clades that are completely species-specific. Several of these proteins (those most closely related to FOG-2) from *C. elegans* and *C. briggsae* were previously analyzed with similar findings (Nayak et al. 2005). Bootstrap support varies for the species-specific clades; nevertheless, it is clear that extensive gene duplication and gene loss have occurred in all three species since their divergence. Among the unstable genes, there are very few cases of one-to-one ortholog pairs among any of the species, and *C. elegans* (the first of the three species to diverge) has no bootstrap-supported orthologs in either *C. briggsae* or *C. remanei*. In addition, the number of genes in the three species is variable (~140 loci in *C. elegans*, 112 loci in *C. briggsae*, and 196 loci in *C. remanei*, including probable defective genes). A strikingly similar protein tree was obtained for the F-box-FBA2 family from the three species, with many large species-specific clades, very few ortholog pairs, and variable gene numbers (data not shown). The rates of gene duplication and deletion implied by these results are unparalleled among known gene families in *Caenorhabditis* (data not shown).

Stable genes in the F-box superfamily

Although most genes in the F-box superfamily are subject to rapid birth–death evolution, 23 genes from *C. elegans* have clear orthologs in both *C. briggsae* and *C. remanei* (Supplemental Fig. S2; Methods). I refer to one-to-one orthologs from the three species as "ortholog trios." Ortholog trios differ widely from each other outside of a shared N-terminal F-box domain; they include members with a WD-40 repeat domain, an LRR domain, and other domains (Supplemental Fig. S2). One ortholog trio contains a possible FBA2 domain, and three trios contain FTH domains that are divergent from each other and from the FTH domains encoded by unstable genes. In all cases, proteins within an ortholog trio are highly conserved across their entire length, but most are unalignable to proteins from other ortholog trios outside of their F-box domain (unjoined tree segments in Supplemental Fig. S2). For the few cases in which the non-F-box region could be aligned across ortholog trios, bootstrap analysis strongly supported the orthology of specific genes, one from each of the three nematode species. The relative evolutionary stability of

Table 1. Summary of unstable F-box-FTH and MATH-BTB trees

Species	Gene family	Unstable genes	Mean expansion size	Expansion size normalized	Positive selection detected
C. elegans	F-box-FTH	116	43.8	27.5	66/91
C. briggsae	F-box-FTH	82	4.6	4.1	ND
C. remanei	F-box-FTH	185	13.7	5.4	ND
C. elegans	MATH-BTB	34	12.1	25.9	18/24
C. briggsae	MATH-BTB	48	8.1	12.3	ND
C. remanei	MATH-BTB	55	21.7	28.8	ND

Unstable genes include all near-full-length genes that are not members of ortholog trios. No evidence for positive selection was detected in the 23 ortholog trios in the F-box-FTH family or in the nine ortholog trios in the MATH-BTB family. "Mean expansion size" is based on a maximum-likelihood protein tree for each family and is the size of each apparent species-specific clade of genes divided by the number of such clades (without bootstrap testing). "Expansion size normalized" is the mean expansion size divided by the number of genes in the species, adjusted to give the same sum as mean expansion size. The criteria for analyzing positive selection are described in Methods. Less systematic analysis in *C. briggsae* and *C. remanei* suggested similarly high frequencies of positive selection (data not shown).

many of these ortholog trios extends beyond nematodes; for example, proteins encoded by 12 of the 23 stable *C. elegans* genes had better BLASTP matches to mouse proteins than any of the ~450 unstable genes tested (labeled M in Supplemental Fig. S2). These patterns suggest that these orthologous genes function in an evolutionarily stable role, probably to target endogenous proteins for ubiquitin-mediated degradation as part of normal development or physiology. Supporting this interpretation, there are three cases in *C. elegans* in which known endogenous substrates are targeted for ubiquitin-dependent proteolysis by F-box proteins, and all three are among the stable orthologs (genes *lin-23* [Dreier et al. 2005], *sel-10* [Li et al. 2002c; Jager et al. 2004], and *fsn-1* [Liao et al. 2004]).

Unstable F-box-FTH and F-box-FBA2 genes are under positive selection

Tests for positive selection were carried out with all suitable full-length *fbxa* genes from *C. elegans* (see Methods). The analysis used a maximum-likelihood test for codon evolution, which can detect specific sites under positive selection in sequences that are otherwise subject to purifying selection (Nielsen and Yang 1998; Yang et al. 2000; Yang and Nielsen 2002; Yang and Swanson 2002). Based on their protein tree, five sets of closely related genes were selected for rigorous d_N/d_S analysis (see Methods; Supplemental Fig. S3). This analysis indicated that all five sets had similar patterns of codon conservation; data from the largest set are shown in Figure 3, three additional sets are shown in Supplemental Figure S5, and summaries of the maximum-likelihood results are given in Supplemental Table S2. All sites in the F-box domain are under purifying selection, consistent with its expected role in binding endogenous Skr and Cullin1 proteins (Fig. 1). C-terminal to their F-box domain, FBXA proteins contain seven blocks of high conservation (labeled A through G) separated by regions of highly divergent sequence (Fig. 3; Supplemental Fig. S4). Six of these seven conserved blocks are identified as part of the FTH domain in Pfam (http://www.sanger.ac.uk/Software/Pfam/2005); I refer to all seven blocks as the extended FTH domain. Between these seven conserved blocks are short divergent regions, all of which show significant evidence of positive selection in at least two of the five gene sets analyzed (five regions are apparent in Fig. 3). In addition to positive selection within the extended FTH domain, there is a hypervariable region between the F-box and extended FTH domains with striking sequence diversity and many sites under probable positive selection. Parts of this hypervariable region align well and contain multiple sites of positive selection for all five gene sets analyzed. The first segment of the hypervariable region alternates between one diverse site and one conserved hydrophobic site, suggesting that it forms a β-strand with one face involved in substrate binding and the other face embedded in the protein core. The region C-terminal to the FTH domain is probably also subject to positive selection, but alignment quality in this region was more problematic, and some sites of apparent positive selection might result from misaligned codons. I used the same method to analyze members of the F-box-FTH family from *C. briggsae* and *C. remanei*. Sets of sequences in both species showed strong evidence of positive selection with patterns similar to those in Figure 3 (data not shown). Similar analysis of *C. elegans* F-box-FTH genes containing the N-terminal mariner-related domain also showed similar patterns of positive selection (data not shown).

As expected if they target endogenous substrates, stable F-box gene ortholog trios showed no evidence of positive selection (Fig. 3, lower panel; data not shown). Since the degree of divergence within ortholog trios is not optimal for detecting positive selection (Anisimova et al. 2002), I also tested a combined alignment of the six genes most closely related to R13H4.5, which includes two ortholog trios with FTH domains (see Supplemental Fig. S2). This combined set also showed no evidence of positive selection (data not shown). Furthermore, in striking contrast to the unstable genes, the substrate-binding domains of stable genes were usually much more conserved than the F-box domain. For example, in the T27F6.8 ortholog trio, the F-box domain had 16 sites of amino acid change, whereas the much longer FTH domain had only six sites of change (Fig. 3). I conclude that stable F-box genes are under strong purifying selection in their substrate-binding domain, presumably because they must bind an endogenous substrate with high specificity.

The maximum-likelihood method was also used to test for positive selection among members of the F-box-FBA2 family in *C. elegans*. An alignment of 48 F-box-FBA2 proteins is shown in Supplemental Figure S6, which illustrates the regional conservation patterns for the family. Strong evidence for positive selection was obtained with seven different subsets of these sequences. Maximum-likelihood results for all seven sets are summarized in Supplemental Table S3, and one alignment with probable selected sites is shown in Supplemental Figure S7. As expected from the fact that the FBA2 domain is unrelated to the FTH domain, the details of conservation are different, but the general pattern of conserved blocks interspersed with regions subject to positive selection is similar. In addition to these parallels with the F-box-FTH family, protein trees made from F-box-FBA2 gene predictions in *C. elegans*, *C. briggsae*, and *C. remanei* showed a similar pattern of large species-specific clades, indicating frequent gene duplication and deletion (data not shown). These results indicate that molecular evolution in the F-box-FBA2 family is very similar to that in the F-box-FTH family.

Skp-related family

F-box proteins bind to Cullin1 complexes via small Skp-related (Skr) proteins. The Skr gene family in *Caenorhabditis* is expanded relative to mammals, with 22.3 genes compared to 4.8 (averages of three nematode and six mammalian genomes) (data not shown). This expanded set of Skr proteins may mediate binding of the huge number of nematode-specific F-box proteins. Maximum-likelihood tests of codon evolution in the *C. elegans* Skr family found no evidence of positive selection (data not shown). This result is consistent with the Skr proteins acting as bridges between F-box domains and the single *C. elegans* Cullin1 (see Fig. 1), without any direct involvement in the substrate specificity conferred by the F-box proteins.

MATH-BTB family

Approximately 110 genes in *C. elegans* contain an identifiable MATH domain. Of these, 47 also contain a BTB domain, and nearly all of the remainder consist of varying numbers of MATH domain repeats. Several MATH-BTB proteins are known to bind Cullin3 complexes via their BTB domain and are thought to bind substrates via their MATH domain (Pintard et al. 2003; Prag and Adams 2003; Xu et al. 2003). The MATH-BTB-containing genes in *C. elegans* have been assigned the gene name *bath* (BTB and MATH domain). As with the F-box families, a sizeable fraction of *bath* loci are either defective genes or encode variant proteins

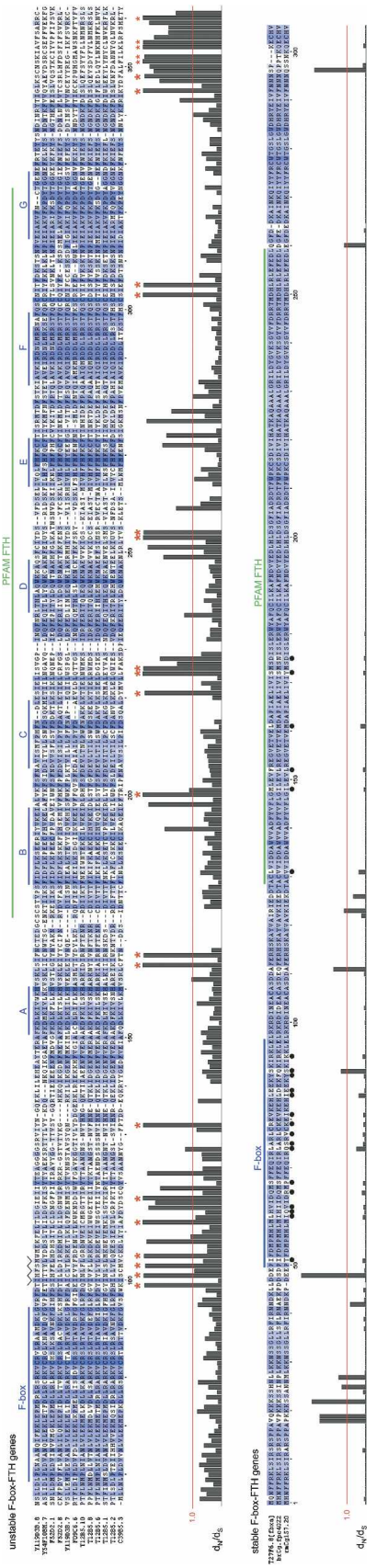


Figure 3. d_N/d_S results for F-box-FTH genes. Alignment and maximum-likelihood d_N/d_S values for a set of 12 unstable F-box-FTH proteins (*top* panel) and one stable ortholog trio of F-box-FTH proteins (*lower* panel). The F-box domain and conserved segments (A–G) of the extended FTH domain are marked *above* the *top* alignment. The jagged line indicates the position of a possible β-strand. Blue alignment shading is proportional to the sum-of-pairs score for each amino acid residue relative to its aligned column. The histogram section of each panel shows estimated d_N/d_S values for each gap-free alignment column, with a red line indicating a value of 1.0. Sites under probable positive selection ($P > 0.9$) are marked with a red asterisk; the five sites near the C terminus have a smaller asterisk to indicate the possibility of misalignment. Evidence for positive selection remained highly significant when this section was removed prior to analysis (data not shown). To avoid investigator bias, the alignments shown were not hand-modified—a few places with possible artifactual alignment are apparent (e.g., misaligned R residues near the N-terminal end of the PFAM-designated FTH domain in the *top* panel). In the *lower* panel, black dots *below* the alignment indicate sites with an amino acid change in any of the three proteins in the F-box or FTH domains.

lacking substantial parts of the typical MATH-BTB protein. Focused gene annotation efforts generated improved gene predictions for many *C. elegans* and *C. briggsae bath* genes, added a few genes that were previously unpredicted, and generated gene predictions for 65 putative *bath* genes in *C. remanei* (see Methods).

Birth–death evolution in the MATH-BTB family

A MATH-BTB protein tree is shown in Supplemental Figure S8 and is summarized in Table 1, including all proteins from *C. elegans*, *C. briggsae*, and *C. remanei* that encode at least 80% of the typical MATH-BTB protein. The tree has strong parallels to those of the F-box-FTH and F-box-FBA2 families, including large species-specific clades and nine sets of stable orthologs, each of which has a single member in each species (marked by black dots in Supplemental Fig. S8). As with the stable F-box genes, this pattern suggests that the ortholog trios target endogenous proteins for degradation as part of normal development or physiology. Supporting this interpretation, the single *C. elegans* MATH-BTB gene with a known function, *mel-26*, is a member of an ortholog trio (Supplemental Fig. S8). The MEL-26 protein is a Cullin3 adapter that targets a microtubule-severing protein for degradation during early embryonic development (Pintard et al. 2003; Xu et al. 2003). Ortholog trio proteins are also relatively well conserved across longer phylogenetic distances; for example, they include all eight best scoring BLASTP queries to mouse (labeled M in Supplemental Fig. S8). Outside of this phylogenetically conserved set of genes, MATH-BTB genes are subject to frequent gene duplication and gene loss in a pattern very similar to genes in the F-box families (Table 1). For example, most of the unstable *C. elegans* MATH-BTB proteins fall into a single clade of 23 proteins, indicating that they all derive from repeated duplication of an ancestral gene that was lost in the *C. briggsae–C. remanei* lineage. *C. remanei* appears to have a modest expansion in the MATH-BTB family relative to the other two species.

Unstable MATH-BTB genes are under positive selection

I used maximum-likelihood tests for positive selection on all suitable unstable *C. elegans* MATH-BTB genes (see Methods), and the results from three sets are presented. General patterns of codon conservation were similar among the three sets, and all three showed strong evidence of positive selection. Results from two of the sets are shown in Figure 4 (upper panel) and Supplemental Figure S10, and summaries of all maximum-likelihood results are in Supplemental Table S4. Sites subject to positive selection are restricted entirely to the MATH domain, consistent with purifying selection in the BTB domain for Cullin3 binding and substrate-driven positive selection in the MATH domain. Sites under positive selection in the MATH domain fall primarily in two regions: a short region with amino acids alternating between strong conservation and positive selection, and a longer region with high variability involving both codon changes and indel mutations. The MATH domain in *C. elegans* is also found in a large family of proteins that consist largely of two or more MATH domains, and many of these genes are also subject to positive selection (Thomas 2006; data not shown). Evidence in *C. elegans* and *Arabidopsis thaliana* indicates that MATH-BTB proteins dimerize (Xu et al. 2003; Weber et al. 2005), suggesting the possibility that the repertoire of MATH-BTB adapters might be extended by formation of dimers or multimers among MATH repeat and MATH-BTB proteins.

As with the F-box families, the nine sets of MATH-BTB ortholog trios have patterns of codon evolution very different from their unstable cousins. Specifically, there was no indication of positive selection in any of the ortholog trios (data not shown), and for most sets the MATH domain was more conserved than the BTB domain. For example, among the three *mel-26* orthologs, there was only one site with an amino acid change in the entire MATH domain, but 16 sites with an amino acid change in the BTB domain (Fig. 4, lower panel). Since the degree of divergence within ortholog trios is not optimal for detecting positive selection (Anisimova et al. 2002), I also tested a combined alignment of the nine genes most closely related to *mel-26* (see Supplemental Fig. S8). This combined set also showed no evidence of positive selection (data not shown). These results indicate that the nine stable MATH-BTB ortholog trios evolve in a manner typical for genes with critical functions that change little with time. Of particular interest here, the MATH domains in the stable MATH-BTB genes are under strong purifying selection, consistent with specific and evolutionarily stable adapter targets.

MATH protein structure

Three-dimensional structures are known for several MATH domain proteins with bound peptide ligands (e.g., McWhirter et al. 1999; Park et al. 1999; Li et al. 2002a; Ye et al. 2002). A structural alignment for one of the *C. elegans* MATH domains was generated by the 3D-PSSM method (see Methods), and variation among nematode MATH domains was mapped to the best-matching protein structure, the TRAF6–RANK complex (Ye et al. 2002). The MATH domain of TRAF6 forms an eight-stranded β-sandwich; the protein-binding cleft is formed by one four-stranded sandwich face plus one adjacent β-strand (Fig. 5). The variable regions and positive-selection sites of the nematode MATH domains map mostly to the peptide-binding face of TRAF6, whereas the most conserved regions map to the other three β-strands (Fig. 5). The region of alternating sites of conservation and positive selection (Fig. 4; Supplemental Fig. S11) aligned to a β-strand in TRAF6, with the positive selection sites facing the binding cleft and the conserved sites facing the protein core. These results suggest that positive selection in the MATH domains is driven by their protein-binding partners.

Genome arrangements in the F-box and MATH-BTB families

A simple hypothesis that explains the trees in Supplemental Figures S1, S2, and S8 is that the stable genes in each family became devoted to specific endogenous substrates a long time ago, whereas the unstable members have continued to evolve by birth–death evolution. Since most gene duplications in nematodes occur locally (Katju and Lynch 2003; Thomas 2006), this hypothesis predicts that the unstable class of genes should be clustered in the genome, a prediction that is strongly supported for both the F-box superfamily and the MATH-BTB family (Fig. 6; Supplemental Fig. S12). Most unstable genes are strongly clustered, with clusters distributed unevenly among the chromosomes and biased toward chromosome arms, hallmarks of gene clusters in *C. elegans* (Thomas 2006). In contrast, stable genes are scattered widely in the genome, with no apparent clustering or bias toward specific chromosomes. Presumably the stable genes in these families also originally arose by local gene duplications, but they became separated from their relatives and from each other by subsequent genome rearrangements during their long period of phylogenetic stability. The modest number of unstable genes that are not in physical clusters may have arisen recently in

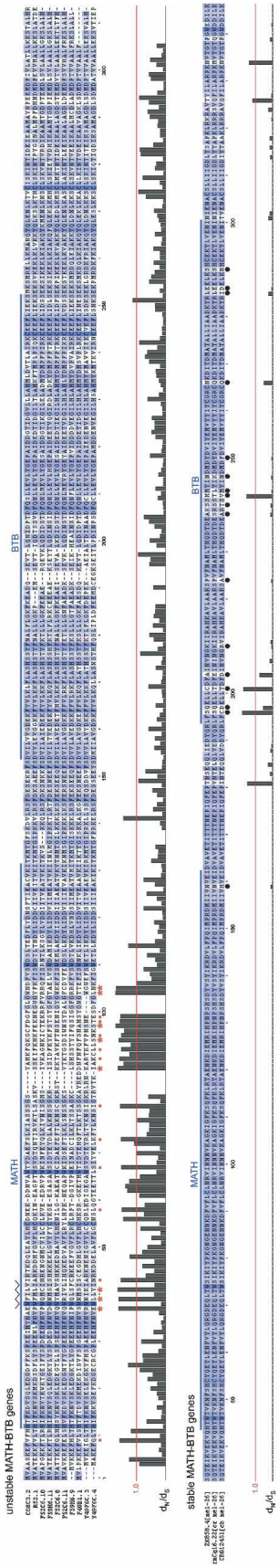


Figure 4. d_N/d_S results for MATH-BTB genes. Alignment and maximum-likelihood d_N/d_S values for a set of 10 unstable MATH-BTB proteins from *C. elegans* (*top* panel) and proteins from one stable ortholog trio, *C. elegans* (*mel-26*), *C. briggsae* (cb *mel-26*), and *C. remanei* (cr *mel-26*) (*lower* panel). The MATH and BTB domains are marked *above* the alignments. Blue alignment shading is proportional to the sum-of-pairs score for each amino acid residue relative to its aligned column. The histogram part of each panel shows estimated d_N/d_S values for each gap-free alignment column, with a red line indicating a value of 1.0. Sites under probable positive selection are marked with a red asterisk ($P \geq 0.9$) or red square ($P \geq 0.8$). In the *lower* panel, sites with an amino acid change in any of the three sequences in the MATH or BTB domains are marked with a black dot.

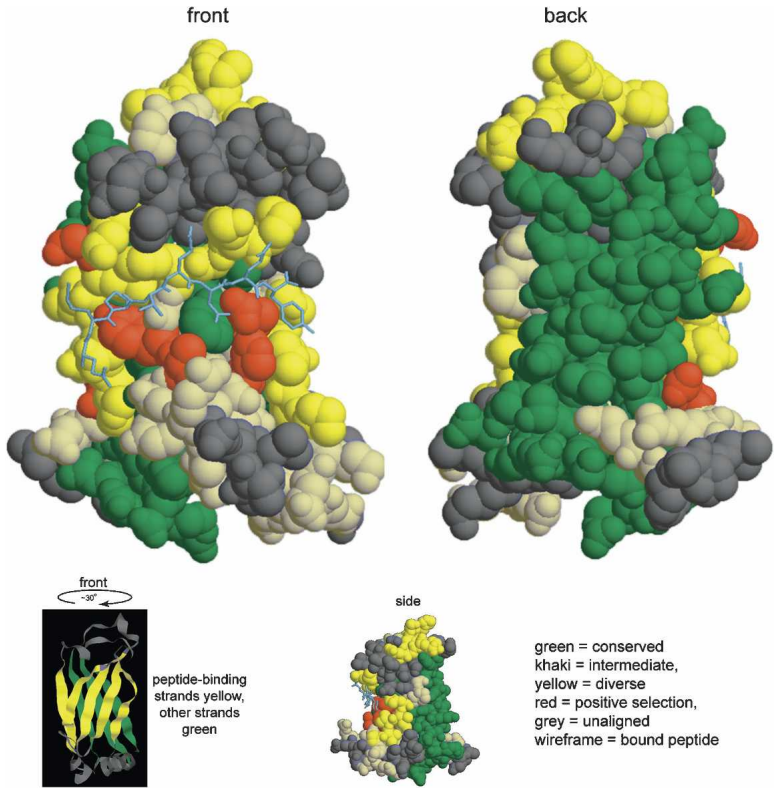


Figure 5. Structural model of MATH domain. The structure is TRAF6 with bound RANK peptide (PDB 1LB5), colored according to the degree of amino acid conservation among nematode MATH domains (Supplemental Fig. S11). Mapping from nematode MATH domains to TRAF6 is based on a 3D-PSSM structural alignment to the MATH domain of C08C3.2 (see Supplemental text S7). In the space-filled models, long regions of high conservation are dark green, long regions of diversity are yellow, sites of probable positive selection are red, and other regions are khaki. Residues in TRAF6 that were not aligned with C08C3.2 are gray. The bound RANK peptide is shown in gray-blue wireframe. The two large views are rotated nearly 180° from each other and are rotation-centered on the bound peptide (*front*) and the most conserved regions (*back*). The small space-filled *side* view shows the binding cleft in TRAF6 more clearly. The ribbon view shows the eight-stranded β-sandwich structure rotated slightly from the *front* view in order to show the β-strands more clearly; peptide binding strands are yellow, other strands are dark green, and non-strand regions are gray.

a similar manner; their low frequency is expected if gene loss is stochastic and frequent relative to dispersing rearrangements.

The F-box superfamily in plants has similar patterns of molecular evolution

Organisms in other phyla also have sizeable F-box and BTB-domain families. Some of these proteins are known Cullin adapters with endogenous substrate targets, but the vast majority are orphan adapters. In a few cases, I tested whether the patterns of positive selection observed in nematodes are also seen in families from these other phyla. As with nematodes, plants have a huge and diverse F-box gene superfamily, with the same basic structural pattern seen in nematodes: an N-terminal F-box domain and a larger C-terminal domain that is very diverse (e.g., Andrade et al. 2001; Wang et al. 2004; Dharmasiri et al. 2005a; Kepinski and Leyser 2005). BLAST searches of predicted *Arabidopsis thaliana* proteins identified 718 genes that encode members of the F-box superfamily; these correspond largely with the set of 693 F-box genes previously analyzed (Gagne et al. 2002). After eliminating a few apparently aberrant gene predictions, I analyzed the remaining 701 genes using the same methods applied to the nematode F-box superfamily. A protein tree for these 701 genes revealed several families with large numbers of genes that encode closely related proteins (Supplemental Fig. S13; see also Gagne et al. 2002). These expanded families include 428 of the 701 genes, with various C-terminal domains (154 LRR-FBD, 88 FBA1, 117 FBA3, and 68 Kelch repeat). I used maximum-likelihood codon analysis to test for positive selection among genes in the six largest expanded groups. Remarkably, strong evidence for positive selection was found in all six cases (Supplemental Table S5; Supplemental Fig. S14). As with the nematode F-box families, sites under positive selection were almost exclusively in the C-terminal substrate-binding domains. I also used BLASTP to test the degree of conservation of these 701 proteins to the nearly complete *Oryza sativa* (rice) gene predictions. Of the 701 *Arabidopsis* proteins, 89 had BLASTP hits to rice with an *E*-value $<10^{-80}$, and none of these 89 were in the expanded families that are subject to positive selection (Supplemental Fig. S13). Two sets of genes with close rice matches were tested by maximum-likelihood codon analysis, and no evidence of positive selection was found (data not shown, both $P > 0.05$). These results suggest that molecular evolution of *Arabidopsis* F-box genes is similar to that in nematodes. Specifically, the genes fall into two evolutionary classes: a smaller class that is relatively stable and conserved and a larger class that is unstable and rapidly diverging. A preliminary protein tree comparison with rice supports birth–death evolution in the F-box families for most genes from these two species (data not shown). As in nematodes, the Skr family in *Arabidopsis* includes ~20 genes, perhaps to bridge the huge F-box superfamily (Gagne et al. 2002). Based on these properties, I speculate that the plant F-box superfamily, like that in nematodes, is involved primarily in recognizing foreign proteins and targeting them for degradation. In contrast to the F-box families, the MATH-BTB family in *Arabidopsis* is small, and preliminary analysis suggests that most of the genes are stable and thus may target endogenous substrates (data not shown).

Mammals and insects also have sizeable F-box and BTB superfamilies, with a variety of putative substrate-binding domains. Samples of genes from several specific mammalian families were analyzed, and none showed evidence of positive selection (data not shown). Paralogs from these families in *Drosophila melanogaster* were too divergent to be used for maximum-likelihood codon analysis (data not shown). It is possible that mammalian and insect members of these families are involved only in endogenous protein degradation, although my surveys were insufficiently complete to warrant a strong conclusion.

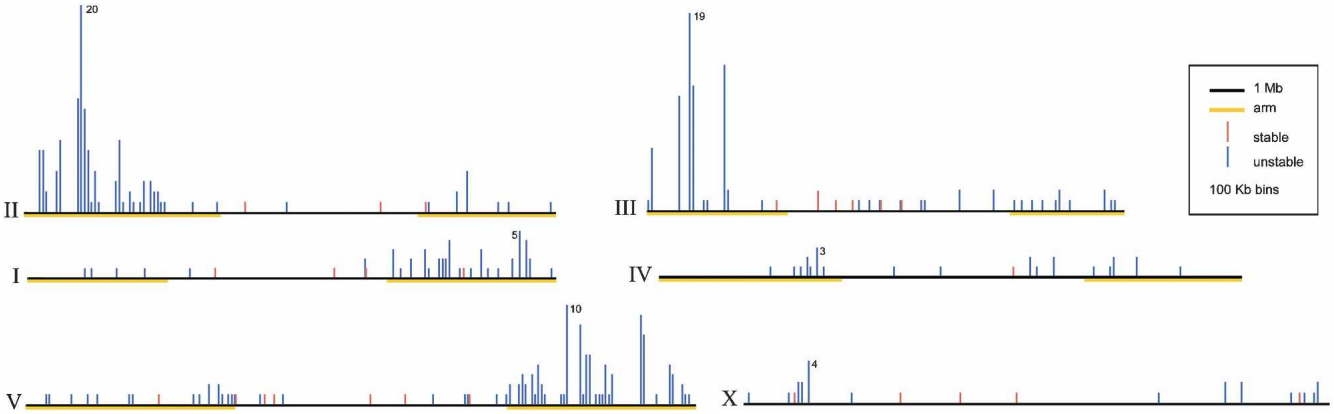


Figure 6. F-box family genome positions. The genome positions of 27 highly conserved (red) and 415 unstable (blue) F-box-containing genes in *C. elegans*. Striking clustering is apparent only for the unstable genes, consistent with evolution by local gene duplication. Highly conserved genes include all 23 ortholog trios plus four genes that did not have specific one-to-one orthologs because of a single duplication or loss in one species. Gene bins are 100 kb in length, and the number of genes in the largest bin on each chromosome is indicated.

Discussion

What is the biological function of the unstable F-box and MATH-BTB family members in nematodes? There is almost no direct experimental evidence addressing the function of these genes, but their patterns of evolution are strongly suggestive. In addition to being consistent with their probable biochemistry as Cullin adapters, an explanation of their function must account for three outstanding features of the families: (1) a large number of diverse genes, (2) a high rate of gene duplication and loss for most genes, and (3) positive selection acting specifically in substrate-binding regions of the unstable genes. I propose that recognition and degradation of foreign proteins explains these patterns of evolution, and that this process is part of the nematode innate immune system. Pathogenic viruses and bacterial protein toxins are plausible specific targets. As part of their life cycle, all viruses express proteins in host cells, making them accessible to ubiquitin-mediated proteolysis. Specifically targeting such viral proteins for degradation should be an effective method for combating viral proliferation. Similarly, many pathogenic bacteria produce protein toxins that translocate into the host cytosol (Falnes and Sandvig 2000) or are secreted into the host cytosol by type III or type IV secretion systems (Christie et al. 2005; Mota and Cornelis 2005). The deleterious effects of such toxins could be combated by targeting them for degradation. Such an antiviral or antibacterial defense system would generate selective pressure for pathogens to evade the defense by evolving target proteins that are not recognized by the host or by evolving functions that antagonize the degradation pathway. This process of pathogen evolution would, in turn, drive evolution of the host defense proteins, resulting in a pattern of positive selection in the adapter substrate-binding domains. In short, the pathogen target proteins and the host adapter proteins would participate in an evolutionary arms race (Dawkins and Krebs 1979). The high frequency of gene loss, gene duplication, and protein diversification in the adapter families is also consistent with this explanation. If any specific host gene conferred only a modest or intermittent selective advantage, it would be prone to stochastic gene loss that could drift to fixation (Kimura 1970; Kimura and King 1979). Once fixed, such gene loss is irreversible, but if extensive adapter diversity were important, then loss events could be balanced by duplication and diversification of the remaining genes, maintaining a genetic complexity whose specific gene components change with time.

The innate immune system hypothesis is speculative, and the evidence for it is indirect. Nevertheless, I found it very difficult to find any other plausible explanation of these results. All MATH-BTB and F-box proteins with known substrates act on endogenous proteins, regulating protein turnover as part of normal development or physiology. The 32 evolutionarily stable members of the MATH-BTB and F-box gene families may function in this manner, but an endogenous substrate explanation for the unstable genes has grave difficulties in accounting for the large number of genes, their pervasive positive selection, and their rapid birth–death evolution. One possibility is a function in clearance of weakly deleterious endogenous garbage proteins resulting from errors in translation or aberrant mRNAs that escape mRNA surveillance (Vasudevan and Peltz 2003). This explanation is hard pressed to explain the observed positive selection because there is no obvious force to drive change. Another possible explanation suggested by Richard Palmiter (pers. comm.), is that these genes function to eliminate toxic proteins released during intestinal digestion of the nematode's bacterial food. Such toxic proteins might be part of a bacterial defense against their nematode predators, but they might also arise purely as an accident of digestion. If such proteins arise and disappear as a result of bacterial evolution, this could conceivably result in positive selection and birth–death evolution of host defense proteins.

In plants, recent findings indicate that several F-box proteins are key regulators of response to light and various small molecules, including auxin, ethylene, jasmonates, gibberellin, and abscisic acid (Somers et al. 2000; Dieterle et al. 2001; Devoto et al. 2002; Xu et al. 2002; McGinnis et al. 2003; Gagne et al. 2004; Dharmasiri et al. 2005a,b; Kepinski and Leyser 2005). None of the F-box genes implicated in these responses are members of the large expanded groups for which I found evidence of positive selection (Supplemental Fig. S13). I speculate that plant F-box genes similarly divide into evolutionarily stable genes that mediate conserved physiological functions and unstable genes that are involved in environmental interactions, possibly including pathogen responses. Insufficient complete genome sequences are currently available in plants to adequately measure the evolutionary stability of the F-box genes, but my preliminary analysis

of *Arabidopsis* and rice F-box genes supports such a possibility (Supplemental Fig. S13; data not shown).

In *C. elegans*, the only member of the unstable F-box and MATH-BTB gene families with a defined function is the F-box-FTH gene *fog-2*. Although the function of *fog-2* is not to recognize foreign proteins, its evolution is remarkable and instructive. The FOG-2 protein binds and probably sequesters the RNA-binding protein GLD-1 during germ-line development, resulting in transient production of sperm in the otherwise female *C. elegans* hermaphrodite (Schedl and Kimble 1988; Clifford et al. 2000; Nayak et al. 2005). The GLD-1-binding region of FOG-2 includes part of the FTH domain and a highly divergent C-terminal segment of the protein, which underwent recent positive selection (Nayak et al. 2005). The ancestral sexual system in *Caenorhabditis* nematodes is male–female, and the hermaphroditic system evolved separately in *C. elegans* and *C. briggsae* (Cho et al. 2004; Kiontke et al. 2004; Nayak et al. 2005). Although the mechanism for sperm generation in *C. briggsae* hermaphrodites is unknown, it is clear that it does not involve a *fog-2* ortholog (Nayak et al. 2005). I speculate that FOG-2 was recently co-opted to bind GLD-1 from an F-box superfamily involved primarily in foreign protein recognition. Given the intensive study of development in *C. elegans* (and very little study of pathogen interactions), it would not be surprising that the only studied member of a large protein family specialized for foreign proteins was one that was recently co-opted for development.

Studies of pathogenesis in *C. elegans* are in their infancy. Only two clear cases of natural (coevolved) pathogens have been described, the gram-positive coryneform bacterium *Microbacterium nematophilum* (Hodgkin et al. 2000; Gravato-Nobre et al. 2005) and the fungus *Drechmaria coniospora* (Jansson 1994; Couillault et al. 2004). If my hypothesis concerning the unstable F-box and MATH-BTB families is correct, it is likely that one target is nematode viruses, since viral proliferation requires expression of proteasome-accessible viral proteins. There are no known *Caenorhabditis* viruses, but two recent studies have shown that some animal viruses can replicate in *C. elegans* (Lu et al. 2005; Wilkins et al. 2005). Both of these studies demonstrate RNAi-mediated gene silencing in nematode antiviral defense, a mechanism previously demonstrated in plants and insects (Hamilton and Baulcombe 1999; Li et al. 2002b). I speculate that this double-stranded RNA targeting mechanism acts in parallel to a foreign protein degradation system in the nematode antiviral defense repertoire. Experimental tests of an innate immunity function for unstable F-box and MATH-BTB genes in *C. elegans* will require identification of suitable natural pathogens, particularly viruses or bacteria known to possess type III or type IV secretion systems. Even if such pathogens were available, the size and diversity of the F-box and MATH-BTB gene families suggests that any specific host gene will target a narrow set of pathogen proteins, a situation that will present a challenge in matching pathogen susceptibilities to specific defense genes. The feasibility of experimental tests in plants is better, since many natural viral and bacterial pathogens are known and disease resistance genetics is much better developed than in nematodes (e.g., Kang et al. 2005).

A limited survey of F-box and MATH-BTB genes in mammals failed to find evidence of positive selection, consistent with the possibility that most or all of them target endogenous substrates for proteolysis. However, a modified ubiquitin-dependent proteasome is thought to be the main source of processed peptides for presentation by MHC class I proteins (Kloetzel and Ossendorp 2004). This immunoproteasome is generated by replacement of three subunits of the constitutive 26S proteasome by interferon-γ-induced subunits (Aki et al. 1994; Groettrup et al. 1996; Nandi et al. 1996), probably to favor production of MHC-presentable peptides (Chen et al. 2001; Toes et al. 2001). Ubiquitinated substrates for the immunoproteasome are thought to be generated by E3 ubiquitin ligases, although I could find no information about the specific Cullins and adapters used. I speculate that an ancestral system of foreign protein degradation via Cullin adapters is the evolutionary antecedent of the MHC class I peptide presentation system. If so, it may be possible to take an evolutionary approach to identifying immune system Cullin adapters responsible for targeting foreign proteins to the immunoproteasome.

Methods

Systematic paralog analysis

The complete set of predicted protein-coding sequences was obtained from WormBase release WS150, excluding transposon-related genes (http://wormbase.org/). To avoid clustering alternative transcripts, only the longest coding sequence from each of 20,134 genes was retained. All coding sequences were translated, and sets of closely related paralogs were generated using BLASTCLUST (NCBI BLAST 2.2.9; ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.9/ 2004). Clustering parameters (BLAST score density 1.0 and at least 70% of both sequences aligned) were chosen to produce paralog groups with sequence diversity close to the optimal range for codeml analysis (Anisimova et al. 2002). These parameters resulted in clustering 2878 genes into 544 groups with three or more paralogs. Each of these 544 groups was tested for positive selection by the following pipeline: codon alignment (guided by a protein alignment generated by CLUSTALW with default settings) (Thompson et al. 1994), production of a maximum-likelihood protein tree (PHYML, one rate category) (Guindon and Gascuel 2003), and codon analysis of this alignment and tree by codeml (details are given in the next section). Protein alignments for all potentially significant codeml results were inspected manually; 23 included dubious alignment regions, and these paralog clusters were discarded (these assessments were performed blind to molecular identity, in order to avoid investigator bias). For the remaining groups, a *P*-value was determined by a χ^2 test on twice the difference in log-likelihood between models 7 and 8, with two degrees of freedom (see next section). The list of *P*-values was analyzed for false discovery rate by the Q-value method (http://faculty.washington.edu/~jstorey/qvalue/; Storey and Tibshirani 2003), which identified 109 paralog groups that were below the point of 5% false discovery rate. These 109 groups were further analyzed to determine whether the added d_N/d_S value in model 8 was significantly greater than 1.0, by comparison to the log-likelihood value from model 8 with the additional d_N/d_S class fixed at 1.0. This test left 80 paralog groups that are strong candidates for positive selection, and each of these was assessed in more detail for alignment quality and molecular identity. Supplemental Table S1 summarizes the results of these analyses, including codeml results, χ^2 and false discovery rate statistical tests, and a brief note on alignment quality and molecular identity. Among the 80 groups, many gene families were represented only once and might therefore be false positives. Gene families that appeared twice or more are strong candidates for positive selection; these include galectins, cuticle collagens, nuclear receptors, Str chemoreceptors, and a family containing the DUF672 domain (http://www.sanger.ac.uk/Software/Pfam/ 2005). Four gene fami-

lies stood out by appearing multiple times: five groups (23 genes) from the F-box-FTH family, four groups (15 genes) from the F-box-FBA2 family, four groups (36 genes) from the MATH-BTB family, and six groups (37 genes) containing one or two C-type lectin domains (note: two of the six C-type lectin alignments—10 genes—were noted as potentially problematic) (Supplemental Table S1). The relatively high representation of the MATH-BTB family in this survey (36 of 47 total genes) is likely a result of curated gene model corrections (submitted to WormBase as part of this work); hand curations are much less complete for the larger F-box and C-type lectin superfamilies. A combination of curated gene model corrections and hand-chosen paralog groups showed that a large fraction of F-box-FTH and F-box-FBA2 genes are subject to positive selection (see below and Results). During this analysis, I found that faulty gene predictions result largely in false negatives in this test for positive selection, either because the mispredicted gene fails to cluster (length mismatch or insufficient BLAST score density) or because it clusters but contains a substantial deletion in the paralog alignment. In the latter case, the deletion removes from analysis not only the sites in the mispredicted gene but also the equivalent sites in the entire paralog group (see next section on gap removal). Unlike missing sequence due to misprediction, insertions have little or no effect on the analysis because they align to gaps in other members of the paralog group and are thus discarded for d_N/d_S analysis.

Codon analysis for positive selection

For detailed analysis of positive selection, proteins for codon analysis were derived from hand-curated and vigorously culled gene models, in order to avoid pseudogenes and gene prediction errors. Multiple sets of five to 15 closely related proteins were selected and aligned using CLUSTALW or CLUSTALX with default settings (Thompson et al. 1994, 1997). This protein alignment was used to generate the corresponding codon alignment and to construct a maximum-likelihood protein tree with proml or PHYML (Felsenstein 1993; Guindon and Gascuel 2003). The tree and codon alignment was analyzed with codeml from PAML package 3.14 (Yang 1997), using models 7 and 8, with three starting d_N/d_S (ω) values for model 8 to avoid local optima. The neutral model 7 assumes a β-distribution of 10 d_N/d_S ratio classes constrained to lie between 0 and 1.0; the selection model 8 is similar but permits one additional d_N/d_S ratio class without constraint. In order to minimize the effects of gene prediction and alignment errors, aligned columns with a gap in any sequence were excluded from analysis ("cleandata" option in codeml). For nematode analysis, the transition/transversion ratio (κ) was fixed at 1.7 (Denver et al. 2004), and for other organisms κ was estimated by codeml. Statistical significance was assessed using a χ^2 test on twice the difference in log-likelihood values (ΔML) for models 7 and 8 with two degrees of freedom, a statistic shown to be conservative in simulations (Anisimova et al. 2001). Specific analysis results and statistical tests are summarized in Supplemental Tables S1 through S5. For all systematic paralog tests and for hand-curated F-box-FBA set A and MATH-BTB set A, model 8 was also run with an eleventh d_N/d_S class fixed at 1.0; the result was compared to model 8 with a free eleventh d_N/d_S class, using a similar χ^2 test (one degree of freedom) to determine whether the free eleventh d_N/d_S was significantly greater than 1.0. Assignment of specific sites under likely positive selection was based on the Bayes-Empirical-Bayes test as implemented in PAML 3.14. To rule out alignment artifacts as a source of spuriously high d_N values, the following analyses were added for F-box-FBA set A and MATH-BTB set A. Alignment gap penalties were increased and decreased (CLUSTAL defaults are gap open [go] 10–gap extend [ge] 0.2; others used were go 9–ge 0.15, go 8.0–ge 0.10, go 11.0–ge 0.25, and go 12.0–ge 0.30). Each alignment was subjected to codeml analysis as described above, and ΔML values varied only slightly from default alignments. In addition, for F-box-FBA set A genes, I analyzed a subset of genes with no length variation in the hypervariable domain, and an alignment in which the variable C-terminal region was removed (see Fig. 3); both cases remained highly significant ($P < 0.00001$).

In order to obtain an estimate of the fraction of unstable genes subject to positive selection in the F-box-FTH and MATH-BTB families (Table 1), a separate analysis was performed as follows. All near-full-length proteins for unstable members of each family were used to make maximum-likelihood trees. These trees were inspected by hand to identify suitable gene sets with appropriate tree lengths for sensitive codeml analysis (between two and 15 nucleotide changes per codon for the sum of all tree branches). Of 116 F-box-FTH genes, 91 were members of suitable gene sets; of 34 MATH-BTB genes, 24 were members of suitable gene sets. Full codeml analysis was carried out as described above for systematic paralog analysis, and genes in sets below a false discovery rate of 5% were interpreted as being subject to positive selection. I note that, because of the nature of paralog cluster analysis, these assignments apply to entire paralog clusters and do not strictly show whether each individual gene in a cluster is subject to positive selection.

Gene prediction

A combination of hand curation and homology-based gene prediction was used to define a set of F-box-FTH genes as follows: 56 predicted proteins from WS142 (http://wormbase.org/) were deemed to be correct on the basis of good full-length alignment with other family members. These 56 proteins were used as query in a GeneWise prediction pipeline (see next section) to derive improved F-box-FTH gene predictions in *C. elegans*, and the results were reconciled with existing predictions by hand curation. This analysis resulted in correction of gene models for 30 F-box genes. When combined with WS142 gene predictions, a total of 143 genes were identified that belong to F-box-FTH family and lack the mariner-homology domain. Of these, I was able to obtain 85 probable full-length gene predictions. Four of these 85 contained in-frame stop codons, but were otherwise apparently intact genes. In *C. briggsae* and *C. remanei*, the divergence of unstable F-box family members was so extreme that a bootstrap approach was adopted. An initial collection of well-aligned F-box family members from the cognate genome (from briggpep for *C. briggsae* and from an initial GeneWise run for *C. remanei*) was gathered. This set of proteins was combined with *C. elegans* F-box-FTH proteins to use as query in a final GeneWise prediction (see next section). A similar process of gene model improvement and gene prediction in *C. remanei* was applied to the MATH-BTB family.

GeneWise prediction pipeline

The program GeneWise uses protein homology and a splice junction model to generate gene predictions from genomic DNA segments (Birney et al. 2004). For clustered homologous genes, the main challenge was providing GeneWise with optimal query protein–target DNA pairs that included entire target genes but did not extend through adjacent genes. To identify such DNA segments, a TBLASTN-m8 search was conducted with a set of query proteins against the appropriate genome sequence (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.9/ 2004). The search results were processed heuristically to identify target DNA hit clusters that could encode most or all of a query protein

homolog, and to identify the query protein with the highest combined TBLASTN score for that hit cluster. The hit cluster was extended *N* nucleotides (usually *N* = 1000) upstream and downstream to ensure that the full gene was included. The extended hit cluster was paired with the matched query protein for analysis by GeneWise. GeneWise output was parsed in various ways, including extending predictions to an in-frame Met (when present) and downstream stop codons. All scripts for this pipeline are available on request.

Protein data sets, alignments, and trees

The nematode proteins used in this paper are found as fasta sequences in Supplemental text data sets S1 through S6, which are *C. elegans*, *C. briggsae*, and *C. remanei* F-box proteins and MATH-BTB proteins, respectively. Gene predictions in *C. elegans* and *C. briggsae* that were modified from those available in WormBase data set WS140 are marked by a terminal "m" in the fasta name, or by "-A" and "-B" in cases in which an existing prediction appeared to be a gene fusion and was split into two genes. Predictions in *C. remanei* were de novo based on Pcap assembly 041,227 (ftp://genome.wustl.edu/pub/seqmgr/remanei/pcap/remanei_041227/) and are named by their contig name followed by nucleotide coordinates for the beginning and end of the GeneWise prediction. Protein multiple alignments were made using CLUSTALW with default settings. For protein trees, sequences were aligned, and ends were trimmed to shared sequence. In the nematode MATH-BTB family, 14 proteins that included <80% of the typical family structure were removed, and a few internal insertions unique to one protein were removed as probable gene prediction artifacts. Four predicted proteins from *C. briggsae* appeared to be fusions of two MATH-BTB genes, and these were split into two genes each. In the nematode F-box superfamily, culling was similar to the MATH-BTB case described above. In both F-box families, genes with internal stop codons and small indels were included in the analysis if they encoded a near-full-length protein based on family alignments. In the F-box family, 27 highly divergent proteins were removed from tree alignments to avoid alignment artifacts and long-branch attraction. Protein trees were generated from the final alignments by protdist (JTT matrix, no gamma correction) and Neighbor-Joining from the PHYLIP package (Felsenstein 1993) or by maximum-likelihood with PHYML with one rate class (Guindon and Gascuel 2003). Bootstrap analysis was performed with 200–1000 samples. One-to-one orthologs were defined as single genes from two or more species that clustered on the tree with high bootstrap support. For the *Arabidopsis* and *Oryza* F-box superfamily, analysis was similar, except that improved gene predictions were not attempted, and trees were made from distances determined from pairwise alignments (after trimming large anomalies) rather than from a multiple alignment, with distance correction by the formula $D = -\ln(1 - d)$, where d is the pair alignment score divided by the smaller of the two self-alignment scores using a BLOSUM62 score matrix.

Molecular modeling

The MATH region of protein C08C3.2 (BATH-33) was submitted to 3D-PSSM (Kelley et al. 2000; http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html) for protein structure modeling. The modeled alignment of BATH-33 was compared to the X-ray crystal structure 1LB5 (Ye et al. 2002), which had the best matching protein fold (*E*-value 5×10^{-2}). Rasmol (Sayle and Milner-White 1995) was used to visualize the structure of one of the three identical subunits of the 1LB5 structure, with its bound RANK peptide ligand. Regions of high and low conservation in the MATH domain of *C. elegans* MATH-BTB proteins were determined from sum-of-pairs scores for the alignment shown in Supplemental Figure S11, and probable sites of positive selection were determined from the codon analysis shown in Figure 4.

Plant genes

Complete sets of coding sequences for *Arabidopsis thaliana* were from TAIR (The *Arabidopsis* Information Resource) release 6 (http://www.arabidopsis.org/) and for *Oryza sativa* from TIGR (The Institute for Genomic Research) release Osa1 (http://www.tigr.org/tdb/e2k1/osa1/). Coding sequences were translated and Ψ-BLAST searches with F-box protein sequences and BLASTP searches with full-length F-box proteins were used to obtain probable F-box-containing proteins (718 from *Arabidopsis* and 1327 from *Oryza*). A few proteins were removed prior to tree construction because the proteins were <150 amino acids or >800 amino acids, suggesting a prediction anomaly. A classification tree of 701 *Arabidopsis* proteins (Supplemental Fig. S13) was used to determine starting sets of closely related paralogs for codon analysis. Multiple alignment of these sets suggested that a substantial fraction of the genes are either mispredicted or are pseudogenes missing large segments of protein. No attempt was made to correct gene models; instead, only paralogs that aligned well across their entire length were used for subsequent analysis. Maximum-likelihood analysis of codon evolution was performed on six sets of genes, each including five to nine full-length paralogous genes. Each set was drawn from a different expanded region of the protein tree, as indicated in Supplemental Figure S13. Strong evidence for positive selection was obtained for all six sets (summarized in Supplemental Table S5), and data from one set are shown in Supplemental Figure S14. Expanded groups A, E, and F are related to each other and are characterized by an N-terminal F-box domain followed by one or two probable matches to a leucine-rich repeat domain (LRR_2, PF07723) and a C-terminal FBD domain (smart00579). Expanded group B is characterized by an N-terminal F-box domain followed by two or more Kelch repeats (Kelch_1, PF01344). Expanded group C is characterized by an N-terminal F-box domain followed by an F-box-associated domain type 3 (FBA_3, PF08268). Expanded group D is characterized by an N-terminal F-box domain followed by an F-box-associated domain type 1 (FBA_1, PF07734). Domain content was determined from conserved domain searches at NCBI (CD Search: NCBI conserved domain search page; http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), Pfam (protein families database of alignments and HMMs; http://www.sanger.ac.uk/Software/Pfam/ 2005), and SMART (Simple Modular Architecture Research Tool, protein domain search page; http://smart.embl-heidelberg.de/).

Acknowledgments

I thank Erica Smith, Nathan Clarke, Willie Swanson, and Stephane Angers for comments on the manuscript; Willie Swanson, Nathan Clarke, Zhirong Bao, and Joe Felsenstein for helpful discussions; Richard Palmiter for the idea that these proteins might be directed toward toxins produced by digestion; and an anonymous reviewer for suggesting a way to summarize complex tree results.

References

Aki, M., Shimbara, N., Takashina, M., Akiyama, K., Kagawa, S., Tamura, T., Tanahashi, N., Yoshimura, T., Tanaka, K., and Ichihara, A. 1994. Interferon-γ induces different subunit organizations and functional diversity of proteasomes. *J. Biochem.* **115:** 257–269.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Andrade, M.A., Gonzalez-Guzman, M., Serrano, R., and Rodriguez, P.L. 2001. A combination of the F-box motif and kelch repeats defines a large *Arabidopsis* family of F-box proteins. *Plant Mol. Biol.* **46:** 603–614.

Anisimova, M., Bielawski, J.P., and Yang, Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* **18:** 1585–1592.

———. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* **19:** 950–958.

Bai, C., Sen, P., Hofmann, K., Ma, L., Goebl, M., Harper, J.W., and Elledge, S.J. 1996. SKP1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. *Cell* **86:** 263–274.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* **14:** 988–995.

Brunson, L.E., Dixon, C., LeFebvre, A., Sun, L., and Mathias, N. 2005. Identification of residues in the WD-40 repeat motif of the F-box protein Met30p required for interaction with its substrate Met4p. *Mol. Genet. Genomics* **273:** 361–370.

Chen, W., Norbury, C.C., Cho, Y., Yewdell, J.W., and Bennink, J.R. 2001. Immunoproteasomes shape immunodominance hierarchies of antiviral $CD8^+$ T cells at the levels of T cell repertoire and presentation of viral antigens. *J. Exp. Med.* **193:** 1319–1326.

Cho, S., Jin, S.W., Cohen, A., and Ellis, R.E. 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res.* **14:** 1207–1220.

Christie, P.J., Atmakuri, K., Krishnamoorthy, V., Jakubowski, S., and Cascales, E. 2005. Biogenesis, architecture, and function of bacterial type IV secretion systems. *Annu. Rev. Microbiol.* **59:** 451–485.

Clifford, R., Lee, M.H., Nayak, S., Ohmachi, M., Giorgini, F., and Schedl, T. 2000. FOG-2, a novel F-box containing protein, associates with the GLD-1 RNA binding protein and directs male sex determination in the *C. elegans* hermaphrodite germline. *Development* **127:** 5265–5276.

Couillault, C., Pujol, N., Reboul, J., Sabatier, L., Guichou, J.F., Kohara, Y., and Ewbank, J.J. 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.

Dawkins, R. and Krebs, J.R. 1979. Arms races between and within species. *Proc. R. Soc. Lond. B. Biol. Sci.* **205:** 489–511.

Denver, D.R., Morris, K., Lynch, M., and Thomas, W.K. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* **430:** 679–682.

Devoto, A., Nieto-Rostro, M., Xie, D., Ellis, C., Harmston, R., Patrick, E., Davis, J., Sherratt, L., Coleman, M., and Turner, J.G. 2002. COI1 links jasmonate signalling and fertility to the SCF ubiquitin-ligase complex in *Arabidopsis*. *Plant J.* **32:** 457–466.

Dharmasiri, N., Dharmasiri, S., and Estelle, M. 2005a. The F-box protein TIR1 is an auxin receptor. *Nature* **435:** 441–445.

Dharmasiri, N., Dharmasiri, S., Weijers, D., Lechner, E., Yamada, M., Hobbie, L., Ehrismann, J.S., Jurgens, G., and Estelle, M. 2005b. Plant development is regulated by a family of auxin receptor F box proteins. *Dev. Cell* **9:** 109–119.

Dieterle, M., Zhou, Y.C., Schafer, E., Funk, M., and Kretsch, T. 2001. EID1, an F-box protein involved in phytochrome A-specific light signaling. *Genes & Dev.* **15:** 939–944.

Dreier, L., Burbea, M., and Kaplan, J.M. 2005. LIN-23-mediated degradation of β-catenin regulates the abundance of GLR-1 glutamate receptors in the ventral nerve cord of *C. elegans*. *Neuron* **46:** 51–64.

Falnes, P.O. and Sandvig, K. 2000. Penetration of protein toxins into cells. *Curr. Opin. Cell Biol.* **12:** 407–413.

Felsenstein, J. 1993. *PHYLIP (Phylogeny Inference Package) version 3.6a2*. Department of Genome Sciences, University of Washington, Seattle, WA.

Figueroa, P., Gusmaroli, G., Serino, G., Habashi, J., Ma, L., Shen, Y., Feng, S., Bostick, M., Callis, J., Hellmann, H., et al. 2005. *Arabidopsis* has two redundant Cullin3 proteins that are essential for embryo development and that interact with RBX1 and BTB proteins to form multisubunit E3 ubiquitin ligase complexes in vivo. *Plant Cell* **17:** 1180–1195.

Furukawa, M., He, Y.J., Borchers, C., and Xiong, Y. 2003. Targeting of protein ubiquitination by BTB-Cullin 3-Roc1 ubiquitin ligases. *Nat. Cell Biol.* **5:** 1001–1007.

Gagne, J.M., Downes, B.P., Shiu, S.H., Durski, A.M., and Vierstra, R.D. 2002. The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in *Arabidopsis*. *Proc. Natl. Acad. Sci.* **99:** 11519–11524.

Gagne, J.M., Smalle, J., Gingerich, D.J., Walker, J.M., Yoo, S.D., Yanagisawa, S., and Vierstra, R.D. 2004. *Arabidopsis* EIN3-binding F-box 1 and 2 form ubiquitin-protein ligases that repress ethylene action and promote growth by directing EIN3 degradation. *Proc. Natl. Acad. Sci.* **101:** 6803–6808.

Gravato-Nobre, M.J., Nicholas, H.R., Nijland, R., O'Rourke, D., Whittington, D.E., Yook, K.J., and Hodgkin, J. 2005. Multiple genes affect sensitivity of *Caenorhabditis elegans* to the bacterial pathogen *Microbacterium nematophilum*. *Genetics* **171:** 1033–1045.

Groettrup, M., Kraft, R., Kostka, S., Standera, S., Stohwasser, R., and Kloetzel, P.M. 1996. A third interferon-γ-induced subunit exchange in the 20S proteasome. *Eur. J. Immunol.* **26:** 863–869.

Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52:** 696–704.

Hamilton, A.J. and Baulcombe, D.C. 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286:** 950–952.

Hodgkin, J., Kuwabara, P.E., and Corneliussen, B. 2000. A novel bacterial pathogen, *Microbacterium nematophilum*, induces morphological change in the nematode *C. elegans*. *Curr. Biol.* **10:** 1615–1618.

Hsiung, Y.G., Chang, H.C., Pellequer, J.L., La Valle, R., Lanker, S., and Wittenberg, C. 2001. F-box protein Grr1 interacts with phosphorylated targets via the cationic surface of its leucine-rich repeat. *Mol. Cell. Biol.* **21:** 2506–2520.

Hughes, A.L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335:** 167–170.

———. 1989. Nucleotide substitution at major histocompatibility complex class II loci: Evidence for overdominant selection. *Proc. Natl. Acad. Sci.* **86:** 958–962.

Hughes, A.L., Ota, T., and Nei, M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol. Biol. Evol.* **7:** 515–524.

Ilyin, G.P., Rialland, M., Glaise, D., and Guguen-Guillouzo, C. 1999. Identification of a novel Skp2-like mammalian protein containing F-box and leucine-rich repeats. *FEBS Lett.* **459:** 75–79.

Jager, S., Schwartz, H.T., Horvitz, H.R., and Conradt, B. 2004. The *Caenorhabditis elegans* F-box protein SEL-10 promotes female development and may target FEM-1 and FEM-3 for degradation by the proteasome. *Proc. Natl. Acad. Sci.* **101:** 12549–12554.

Jansson, H. 1994. Adhesion of conidia of *Drechmaria coniospora* to *Caenorhabditis elegans* wild type and mutants. *J. Nematol.* **26:** 430–435.

Jiang, J. and Struhl, G. 1998. Regulation of the Hedgehog and Wingless signalling pathways by the F-box/WD40-repeat protein Slimb. *Nature* **391:** 493–496.

Jin, J., Cardozo, T., Lovering, R.C., Elledge, S.J., Pagano, M., and Harper, J.W. 2004. Systematic analysis and nomenclature of mammalian F-box proteins. *Genes & Dev.* **18:** 2573–2580.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Le Bot, N., Moreno, S., Sohrmann, M., et al. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421:** 231–237.

Kang, B.C., Yeam, I., and Jahn, M.M. 2005. Genetics of plant virus resistance. *Annu. Rev. Phytopathol.* **43:** 581–621.

Kanost, M.R., Jiang, H., and Yu, X.Q. 2004. Innate immune responses of a lepidopteran insect, *Manduca sexta*. *Immunol. Rev.* **198:** 97–105.

Katju, V. and Lynch, M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics* **165:** 1793–1803.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299:** 499–520.

Kepinski, S. and Leyser, O. 2005. The *Arabidopsis* F-box protein TIR1 is an auxin receptor. *Nature* **435:** 446–451.

Kimura, M. 1970. The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. *Genet. Res.* **15:** 131–133.

Kimura, M. and King, J.L. 1979. Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. *Proc. Natl. Acad. Sci.* **76:** 2858–2861.

Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., and Fitch, D.H. 2004. *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl. Acad. Sci.* **101:** 9003–9008.

Kloetzel, P.M. and Ossendorp, F. 2004. Proteasome and peptidase

function in MHC-class-I-mediated antigen presentation. *Curr. Opin. Immunol.* **16:** 76–81.
Kogelberg, H. and Feizi, T. 2001. New structural insights into lectin-type proteins of the immune system. *Curr. Opin. Struct. Biol.* **11:** 635–643.
Li, C., Ni, C.Z., Havert, M.L., Cabezas, E., He, J., Kaiser, D., Reed, J.C., Satterthwait, A.C., Cheng, G., and Ely, K.R. 2002a. Downstream regulator TANK binds to the CD40 recognition site on TRAF3. *Structure* **10:** 403–411.
Li, H., Li, W.X., and Ding, S.W. 2002b. Induction and suppression of RNA silencing by an animal virus. *Science* **296:** 1319–1321.
Li, J., Pauley, A.M., Myers, R.L., Shuang, R., Brashler, J.R., Yan, R., Buhl, A.E., Ruble, C., and Gurney, M.E. 2002c. SEL-10 interacts with presenilin 1, facilitates its ubiquitination, and alters A-β peptide production. *J. Neurochem.* **82:** 1540–1548.
Liao, E.H., Hung, W., Abrams, B., and Zhen, M. 2004. An SCF-like ubiquitin ligase complex that controls presynaptic differentiation. *Nature* **430:** 345–350.
Lu, J., Teh, C., Kishore, U., and Reid, K.B. 2002. Collectins and ficolins: Sugar pattern recognition molecules of the mammalian innate immune system. *Biochim. Biophys. Acta* **1572:** 387–400.
Lu, R., Maduro, M., Li, F., Li, H.W., Broitman-Maduro, G., Li, W.X., and Ding, S.W. 2005. Animal virus replication and RNAi-mediated antiviral silencing in *Caenorhabditis elegans*. *Nature* **436:** 1040–1043.
Marchler-Bauer, A. and Bryant, S.H. 2004. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* **32:** W327–W331.
McGinnis, K.M., Thomas, S.G., Soule, J.D., Strader, L.C., Zale, J.M., Sun, T.P., and Steber, C.M. 2003. The *Arabidopsis* SLEEPY1 gene encodes a putative F-box subunit of an SCF E3 ubiquitin ligase. *Plant Cell* **15:** 1120–1130.
McGreal, E.P., Martinez-Pomares, L., and Gordon, S. 2004. Divergent roles for C-type lectins expressed by cells of the innate immune system. *Mol. Immunol.* **41:** 1109–1121.
McWhirter, S.M., Pullen, S.S., Holton, J.M., Crute, J.J., Kehry, M.R., and Alber, T. 1999. Crystallographic analysis of CD40 recognition and signaling by human TRAF2. *Proc. Natl. Acad. Sci.* **96:** 8408–8413.
Moon, J., Parry, G., and Estelle, M. 2004. The ubiquitin-proteasome pathway and plant development. *Plant Cell* **16:** 3181–3195.
Mota, L.J. and Cornelis, G.R. 2005. The bacterial injection kit: Type III secretion systems. *Ann. Med.* **37:** 234–249.
Nandi, D., Jiang, H., and Monaco, J.J. 1996. Identification of MECL-1 (LMP-10) as the third IFN-γ-inducible proteasome subunit. *J. Immunol.* **156:** 2361–2364.
Nayak, S., Goree, J., and Schedl, T. 2005. fog-2 and the evolution of self-fertile hermaphroditism in *Caenorhabditis*. *PLoS Biol.* **3:** e6.
Nicholas, H.R. and Hodgkin, J. 2004. Responses to infection and possible recognition strategies in the innate immune system of *Caenorhabditis elegans*. *Mol. Immunol.* **41:** 479–493.
Nielsen, R. and Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929–936.
Park, Y.C., Burkitt, V., Villa, A.R., Tong, L., and Wu, H. 1999. Structural basis for self-association and receptor recognition of human TRAF2. *Nature* **398:** 533–538.
Pintard, L., Willis, J.H., Willems, A., Johnson, J.L., Srayko, M., Kurz, T., Glaser, S., Mains, P.E., Tyers, M., Bowerman, B., et al. 2003. The BTB protein MEL-26 is a substrate-specific adaptor of the CUL-3 ubiquitin-ligase. *Nature* **425:** 311–316.
Prag, S. and Adams, J.C. 2003. Molecular phylogeny of the kelch-repeat superfamily reveals an expansion of BTB/kelch proteins in animals. *BMC Bioinformatics* **4:** 42.
Sayle, R.A. and Milner-White, E.J. 1995. RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.* **20:** 374.
Schedl, T. and Kimble, J. 1988. fog-2, a germ-line-specific sex determination gene required for hermaphrodite spermatogenesis in *Caenorhabditis elegans*. *Genetics* **119:** 43–61.
Schulman, B.A., Carrano, A.C., Jeffrey, P.D., Bowen, Z., Kinnucan, E.R., Finnin, M.S., Elledge, S.J., Harper, J.W., Pagano, M., and Pavletich, N.P. 2000. Insights into SCF ubiquitin ligases from the structure of the Skp1–Skp2 complex. *Nature* **408:** 381–386.
Somers, D.E., Schultz, T.F., Milnamow, M., and Kay, S.A. 2000. ZEITLUPE encodes a novel clock-associated PAS protein from *Arabidopsis*. *Cell* **101:** 319–329.
Stogios, P.J., Downs, G.S., Jauhal, J.J., Nandra, S.K., and Prive, G.G. 2005. Sequence and structural analysis of BTB domain proteins. *Genome Biol.* **6:** R82.
Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100:** 9440–9445.
Swanson, W.J., Yang, Z., Wolfner, M.F., and Aquadro, C.F. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl. Acad. Sci.* **98:** 2509–2514.
Thomas, J.H. 2006. Analysis of homologous gene clusters in *C. elegans* reveals striking regional cluster domains. *Genetics* **172:** 127–143.
Thomas, J.H., Kelley, J.L., Robertson, H.M., Ly, K., and Swanson, W.J. 2005. Adaptive evolution in the SRZ chemoreceptor families of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Proc. Natl. Acad. Sci.* **102:** 4476–4481.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.
Toes, R.E., Nussbaum, A.K., Degermann, S., Schirle, M., Emmerich, N.P., Kraft, M., Laplace, C., Zwinderman, A., Dick, T.P., Muller, J., et al. 2001. Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J. Exp. Med.* **194:** 1–12.
van den Heuvel, S. 2004. Protein degradation: CUL-3 and BTB—Partners in proteolysis. *Curr. Biol.* **14:** R59–R61.
Varshavsky, A. 2005. Regulated protein degradation. *Trends Biochem. Sci.* **30:** 283–286.
Vasudevan, S. and Peltz, S.W. 2003. Nuclear mRNA surveillance. *Curr. Opin. Cell Biol.* **15:** 332–337.
Wang, L., Dong, L., Zhang, Y., Zhang, Y., Wu, W., Deng, X., and Xue, Y. 2004. Genome-wide analysis of S-Locus F-box-like genes in *Arabidopsis thaliana*. *Plant Mol. Biol.* **56:** 929–945.
Weber, H., Bernhardt, A., Dieterle, M., Hano, P., Mutlu, A., Estelle, M., Genschik, P., and Hellmann, H. 2005. *Arabidopsis* AtCUL3a and AtCUL3b form complexes with members of the BTB/POZ-MATH protein family. *Plant Physiol.* **137:** 83–93.
Wilkins, C., Dishongh, R., Moore, S.C., Whitt, M.A., Chow, M., and Machaca, K. 2005. RNA interference is an antiviral defence mechanism in *Caenorhabditis elegans*. *Nature* **436:** 1044–1047.
Winston, J.T., Koepp, D.M., Zhu, C., Elledge, S.J., and Harper, J.W. 1999. A family of mammalian F-box proteins. *Curr. Biol.* **9:** 1180–1182.
Xu, L., Liu, F., Lechner, E., Genschik, P., Crosby, W.L., Ma, H., Peng, W., Huang, D., and Xie, D. 2002. The SCF(COI1) ubiquitin–ligase complexes are required for jasmonate response in *Arabidopsis*. *Plant Cell* **14:** 1919–1935.
Xu, L., Wei, Y., Reboul, J., Vaglio, P., Shin, T.H., Vidal, M., Elledge, S.J., and Harper, J.W. 2003. BTB proteins are substrate-specific adaptors in an SCF-like modular ubiquitin ligase containing CUL-3. *Nature* **425:** 316–321.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17:** 32–43.
———. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19:** 908–917.
Yang, Z. and Swanson, W.J. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol. Biol. Evol.* **19:** 49–57.
Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.
Ye, H., Arron, J.R., Lamothe, B., Cirilli, M., Kobayashi, T., Shevde, N.K., Segal, D., Dzivenu, O.K., Vologodskaia, M., Yim, M., et al. 2002. Distinct molecular mechanism for initiating TRAF6 signalling. *Nature* **418:** 443–447.
Zheng, N., Schulman, B.A., Song, L., Miller, J.J., Jeffrey, P.D., Wang, P., Chu, C., Koepp, D.M., Elledge, S.J., Pagano, M., et al. 2002. Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase complex. *Nature* **416:** 703–709.

Received December 21, 2005; accepted in revised form May 23, 2006.