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, WA 98103

Running title: Adaptive evolution in ubiquitin-ligase adapters

Key words: "adaptive evolution", "positive selection", "ubiquitin ligase", "proteolysis", "innate immunity", "C. elegans", "Arabidopsis".

Abstract

From a complete survey for evidence of positive selection among paralog groups in *C. elegans*, I repeatedly identified two classes of genes that encode ubiquitin-dependent proteasome adapters. In this proteolysis system, adapter proteins recruit diverse substrate proteins for poly-ubiquitination and proteolysis by Cullin-E3 ubiquitin ligase complexes. The adapter proteins consist of a conserved Cullin-binding domain and a variable substrate-binding domain. The adapter genes found in my survey encode proteins in the F-box superfamily and the MATH-BTB family, which are adapters for Cullin1 and Cullin3 complexes, respectively. 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 subject to strong positive selection at multiple sites in their substrate-binding domains but not in their Cullin-binding domains. Structural modeling of the 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.

Supplementary Materials: There are extensive supplementary materials, including 11 figures, 5 tables, and 1 text data set. In addition, protein sequence data may be appropriate to include as additional supplemental material, since many gene predictions were modified as part of this work. Updated gene models have been communicated to WormBase, so this may not be needed. There are no new accession numbers or sequences from this work.

Introduction

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 positive selection and a high degree of population polymorphism (Hughes and Nei 1988; Hughes and Nei 1989; Hughes et al. 1990; Swanson et al. 2001). Positive selection is usually detected by a rate of nonsynonymous codon change higher than synonymous codon change, a pattern the reverse of that produced by the more common negative (purifying) 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; Hughes and Nei 1989; Hughes et al. 1990).

To identify genes that are candidates for pathogen interaction in *C. 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 common 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 2000). The most prominent 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 (Figure 1). Each specific Cullin protein uses a distinct class of adapter protein.

Most or all proteins in the F-box superfamily are adapters for Cullin1 complexes (Bai et al. 1996; Winston et al. 1999; Zheng et al. 2002). 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 (Brunson et al. 2005; Hsiung et al. 2001; Nayak et al. 2005; Winston et al. 1999). 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, WD40 repeat, LRR, FTH, FBA, FBA1, and FBA2 domain-containing families (Andrade et al. 2001; Clifford et al. 2000; Gagne et al. 2002; http://www.sanger.ac.uk/Software/Pfam/ 2005; Ilyin et al. 1999; Jiang and Struhl 1998; Winston et al. 1999). Studies in this 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 protein domain.

Many BTB proteins are adapters for Cullin3 complexes (Figueroa et al. 2005; Furukawa et al. 2003; Pintard et al. 2003; Xu et al. 2003). 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, WD40 repeat, Zn-finger repeat, and kelch repeat domains (Brunson et al. 2005; Figueroa et al. 2005; Pintard et al. 2003; Prag and Adams 2003; Stogios et al. 2005; van den Heuvel 2004; Winston et al. 1999; Xu et al. 2003). 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 about 2.5% of total coding potential. Given their number, remarkably little is known about these genes; 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 these genes function to target foreign proteins for degradation as part of the innate immune system.

Results

Systematic paralog test for positive selection: All *C. elegans* gene families with three or more recent duplicates were tested for evidence of positive selection (see Materials and 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 2,878 genes. 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. Results from this analysis are summarized in Table S1. Among the 544 groups tested, this method identified 86 groups of paralogs that showed potentially significant evidence of positive selection (false discovery rate < 5%). Three gene families or superfamilies were identified repeatedly among these 86 paralog groups: nine paralog groups in the F-box superfamily, four groups in the MATH-BTB family, and five groups in the C-type lectin superfamily. The C-type lectin family is strongly implicated in innate immunity to bacteria and fungi in a variety of organisms (Kanost et al. 2004; Kogelberg and Feizi 2001; Lu et al. 2002; McGreal et al. 2004), though analysis in *C. elegans* is just beginning (Nicholas and Hodgkin 2004). In contrast, the F-box families and 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 about 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 *psi-blast* and *rps-blast* searches (Altschul et al. 1997; Marchler-Bauer and Bryant 2004), I found that approximately 520 loci in *C. elegans* potentially encode a protein with a clear F-box domain (an additional 50 genes, not analyzed here, may contain a highly divergent

F-box like domain). About 40 of these 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: about 220 contain an FTH domain and about 210 contain an FBA2 domain. In both families an N-terminal F-box domain is followed by a more divergent region of about 300 amino acids, which contains the FTH or FBA2 domain (Figure 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 (Figure 2 and data not shown).

Much of my analysis has focused on the 140 genes in the F-box-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. 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 *C. briggsae fbxa* genes, added a few genes that were previously unpredicted, and generated gene models for 228 putative *fbxa* genes in the newly sequenced *C. remanei* (see Materials and Methods).

Birth-death evolution in the F-box superfamily: As shown below, 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 obvious orthologs that is undergoing rapid birth-death evolution. For simplicity I will refer to these as "stable" genes and "unstable" genes throughout this paper, in reference to their apparent rates of gene duplication and deletion. An FBXA protein tree is shown in Figure 3A, 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 these 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*.

Stable genes in the F-box superfamily: Though 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* (Figure 3B). I will refer to orthologs from the three species as ortholog trios. Though the proteins encoded by these ortholog trios all share an F-box domain near their N-terminus, they differ widely in the remainder of the protein, and include members with a WD40 repeat domain, an LRR domain, and other domains. 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 Figure 3B). For the few cases in which the non-F-box region could be aligned across ortholog trios, alignment and bootstrap analysis strongly supported the orthology of specific genes, one from each of the three nematode species. The evolutionary stability of 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. These patterns suggest that these orthologous genes function in an evolutionarily stable role, presumably 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 (*lin-23* (Dreier et al. 2005), *sel-10* (Jager et al. 2004; Li et al. 2002c), and *fsn-1* (Liao et al. 2004)).

F-box-FTH and F-box-FBA2 genes are under positive selection: Tests for positive selection were carried out with most of the 85 unstable *fbxa* genes from *C. elegans* that encode full-length FBXA proteins. 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 and Nielsen 2002; Yang et al. 2000; Yang and Swanson 2002). Based on their protein tree, five sets of closely-related genes were selected for d_N/d_S analysis (see Materials and Methods and Figure S2). This analysis indicated that all five sets had similar patterns of codon conservation; data from the largest set are shown in Figure 4, three additional sets are shown in Figure S3, and summaries of the maximum-likelihood results are given in 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 (Figure 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 (Figures 4 and S1). 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 will 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 are apparent for the data set shown in Figure 4). In addition to positive selection within the extended FTH domain, there is a hypervariable region between the F-box and FTH domains with striking sequence diversity and many sites under probable positive selection. Parts of this hypervariable region align well and contain multiple sites under 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 the possibility that it forms a β-sheet 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 is 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 also showed strong evidence of positive selection with patterns similar to those in Figure 4 (data not shown). Similar analysis of *C. elegans* F-box-FTH genes containing the N-terminal mariner-related domain showed that most or all of these genes also have similar patterns of positive selection restricted to the extended FTH domain and flanking sequences (data not shown).

As expected if they target endogenous ligands, stable F-box gene ortholog trios showed no evidence of positive selection (Figure 4 lower panel and 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 Figure 3B). This combined set also showed no evidence of positive selection (data not shown). In striking contrast to the unstable genes, the substrate-binding domain of stable genes was usually 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 (Figure 4). 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 Figure S4, 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 Table S3 and one alignment with probable selected sites is shown in Figure S5. As expected from the fact that the FBA2 domain is unrelated to the FTH domain, the details of conservation were different, but the general pattern of conserved blocks interspersed with regions subject to positive selection was similar. In addition to probable positive selection between conserved blocks within an extended FBA2 domain, there was a hypervariable region between the F-box and the FBA2 domains with strong evidence of positive selection at multiple sites. 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 suggest 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 3 nematode and 6 mammalian genomes). This expanded set of Skr proteins may mediate binding of the huge number of F-box proteins. Maximum-likelihood tests of codon evolution in the 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 Figure 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 most 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 pseudogenes or encode variant proteins 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 Materials and Methods).

Birth-death evolution in the MATH-BTB family: A MATH-BTB protein tree is shown in Figure 5, which includes all proteins from *C. elegans*, *C. briggsae*, and *C. remanei* that encode at least 80% of an alignable protein. The tree has strong parallels to those of the F-box-FTH and F-box-FBA2 families, including large species-specific clades and a low frequency of orthologs. The MATH-BTB family includes nine sets of stable orthologs, each of which has a single member in each species (marked by black dots in Figure 5). As with the stable F-box genes, this pattern suggests that these 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 one of the ortholog genes (Figure 5). MEL-26 is a Cullin3 adapter that targets a microtubule-severing protein for degradation during early embryonic development (Pintard et al. 2003; Xu et al. 2003). The orthologous proteins are also relatively well conserved across longer phylogenetic distances; for example, they include all eight best scoring *blastp* queries to mouse (marked M in Figure 5). 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-FTH and F-box-FBA2 families. 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 undergone 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 three sets of unstable *C. elegans* MATH-BTB genes. General patterns of codon conservation were similar among the three sets and there was strong evidence for positive selection in all three sets. Results from two of the sets are shown in Figure 6 (upper panel) and in Figure S7, and summaries of all maximum-likelihood results are in Table S4. Sites under likely 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 2005) and data not shown). Evidence in *C. elegans* and *Arabidopsis thaliana* indicates that MATH-BTB proteins dimerize (Weber et al. 2005; Xu et al. 2003), 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.

Similar codon analysis indicates that the nine sets of MATH-BTB ortholog trios have patterns of conservation 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 appeared to be 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 amino acid changes in the BTB domain (Figure 6 lower panel). Since the degree of divergence within ortholog trios is not be 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 Figure 5). 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 organismal 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 highly 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. (Li et al. 2002a; McWhirter et al. 1999; Park et al. 1999; Ye et al. 2002)). A structural alignment for one of the *C. elegans* MATH domains was generated by the 3D-PSSM method (see Materials and 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 8-stranded β-sandwich; the protein binding cleft is formed by one 4-stranded sandwich face plus one adjacent β-strand (Figure 7). 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 almost exclusively to the three other β-strands (Figure 7). The region of alternating sites of conservation and positive selection (Figure 6) aligned to a β-sheet in TRAF6, with the positive selection sites facing the protein binding cleft and the conserved sites facing the protein core. These results are consistent with positive selection in the MATH domains being driven by their protein binding partners.

Genome arrangements in the F-box and MATH-BTB families: A simple hypothesis that explains the trees in Figures 3 and 5 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 rapid birth-death evolution. Since most gene duplications in nematodes occur locally (Katju and Lynch 2003; Thomas 2005), this hypothesis predicts that the unstable (birth-death) 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 (Figure 8 and Figure S8). Most unstable genes are strongly clustered, are distributed unevenly among the chromosomes, and are biased toward chromosome arms (a hallmark of gene clusters in *C. elegans* (Thomas 2005)). In contrast, stable genes are scattered widely in the genome, with no apparent clustering or bias toward specific chromosomes. I speculate that the physical clusters of unstable genes reside in regions of active gene duplication. 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 rearrangements during their long period of stability. The modest number of unstable genes that are not in physical clusters may have arisen recently in a similar manner; their low frequency would be expected if gene loss were frequent and stochastic.

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. Like 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; Dharmasiri et al. 2005a; Kepinski and Leyser 2005; Wang et al. 2004)). 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 (Figure S10; see also (Gagne et al. 2002)). These expanded families include 428 (61%) 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 (Table S5 and Figure S11). 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. 89 of the *Arabidopsis* proteins had *blastp* hits to *Oryza* with an E-value less than 10^{-80}, and none of these 89 were in the expanded families that are subject to positive selection (Figure S10). Two sets of genes with close *Oryza* 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 highly conserved and a larger class that is unstable and rapidly diverging. A preliminary protein tree comparison with rice supports rapid 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 approximately 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 modest in size and preliminary analysis suggested that most of these 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, though the surveys were insufficiently complete to warrant a strong conclusion.

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) positive selection acting specifically in substrate-binding regions, and 3) a high rate of gene duplication and loss. The large number of diverse genes implies a demand for high diversity in their specific functional roles. Most of this diversity resides in the substrate-binding domains of the proteins. The fact that most or all of the unstable genes are subject to positive selection in their substrate-binding domain suggests that evolution of their amino acid sequence is driven by changing substrates. The high rate of gene loss implies that natural selection to retain any specific gene is low (or varies with time), resulting in frequent fixation of defective genes and their eventual deletion from the genome. Despite the dispensability of individual genes, the fact that gene loss is balanced by a high rate of gene duplication and diversification implies that breadth of specificity in the family is functionally important.

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 that are potential anti-viral targets for host defense systems. Since these proteins are expressed in host cells, they are 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 anti-viral or anti-bacterial 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 domain. 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 readily 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 strictly indirect. Nevertheless, I found it very difficult to find any other plausible explanation of these results. All MATH-BTB and F-box proteins that target identified substrates for proteolysis act on endogenous proteins, regulating protein turnover as part of normal development or physiology.

The evolutionarily stable members of the MATH-BTB and F-box gene families may be additional genes with this type of function. However the 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. A barely plausible possibility is a function in clearance of endogenous garbage proteins that are weakly deleterious. For example, aberrant mRNA forms that escape mRNA surveillance (Vasudevan and Peltz 2003) or errors in translation might produce deleterious proteins that encourage targeted proteolysis. This explanation is hard pressed to explain the observed positive selection. Another alternative explanation suggested by Richard Palmiter (personal communication), is that these genes function to eliminate toxic proteins released during intestinal digestion of the nematode's bacterial food. If these toxic proteins are part of a bacterial defense against their nematode predators, then this explanation is merely a variant of the innate immune system hypothesis. However it is possible that toxic proteins might arise purely as an accident of digestion. Since such proteins might arise and disappear as a result of bacterial evolution, this situation 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 abscicic acid (Devoto et al. 2002; Dharmasiri et al. 2005a; Dharmasiri et al. 2005b; Dieterle et al. 2001; Gagne et al. 2004; Kepinski and Leyser 2005; McGinnis et al. 2003; Somers et al. 2000; Xu et al. 2002). None of the F-box genes implicated in these responses are members of the expanded groups for which I found evidence of positive selection. I speculate that, as in nematodes, plant F-box genes will divide into evolutionarily stable genes that mediate conserved physiological functions and unstable genes that are involved in environmental interactions, probably 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 (Figure S10 and data not shown).

In *C. elegans*, the only member of the unstable F-box-FTH and F-box-FBA2 gene families with a defined function is *fog-2*. Though 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-1during germ line development, resulting in transient production of sperm in the otherwise female *C. elegans* hermaphrodite (Clifford et al. 2000; Nayak et al. 2005; Schedl and Kimble 1988). 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). Though 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 pathogen interactions were one that was recently co-opted for development.

Studies of pathogenesis in *C. elegans* are in their infancy. Only two clear cases of co-evolved specific pathogens have been described, the bacterium *Microbacterium nematophilum* (Gravato-Nobre et al. 2005; Hodgkin et al. 2000) and the fungus *Drechmaria coniospora* (Couillault et al. 2004; Jansson 1994). If my hypothesis concerning the unstable F-box and MATH-BTB families is correct, it is likely that their main target is nematode viruses, since viral proliferation requires host proteasome accessible proteins. There are no known *Caenorhabditis* viruses, but two recent studies have shown that some animal viruses can infect and 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. The finding that most members of the huge F-box superfamily in plants are also subject to positive selection suggests the possibility that foreign protein degradation is an ancient pathogen defense system and that it may be widespread in animals and plants.

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 regulated degradation. 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, though little is known 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, one 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* (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 in the optimal range for *codeml* analysis (Anisimova et al. 2002). These parameters resulted in clustering 2,878 genes into 544 groups with 3 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 appeared to include dubious alignment regions and these paralog clusters were discarded (these assessments were performed blind to molecular identification, in order to avoid investigator bias). For the remaining groups, a p-value was determined by a χ-square test on twice the difference in log-likelihood between models 7 and 8, with two degrees of freedom (see next section). These p-values were 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 estimated d_N/d_S value was significantly greater than 1.0, by comparison to the log-likelihood value from model 8 with an 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. Table S1 summarizes the results of these analyses, including *codeml* results, χ-square 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 families stood out by appearing multiple times: 5 groups (23 genes) from the F-box-FTH family, 4 groups (15 genes) from the F-box-FBA2 family, 4 groups (36 genes) from the MATH-BTB family, and 6 groups (37 genes) containing one or two C-type lectin domains (note - two of the C-type lectin alignments (10 genes) were noted as potentially problematic, 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 sequences 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 5 to 15 closely-related proteins were selected and aligned using *clustalw* or *clustalx* with default settings (Thompson et al. 1997; Thompson et al. 1994). This protein alignment was used 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 were 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, whereas the selection model 8 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-square 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 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 11^{th} d_N/d_S class fixed at 1.0; the result was compared to model 8 with a free 11^{th} d_N/d_S class, using a similar χ-square test (1 degree of freedom) to determine whether the free 11^{th} 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; 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; all remained highly significant ($P < 0.00001$). 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 Figure 4); both cases remained highly significant ($P < 0.00001$).

Gene Prediction: A combination of hand curation and homology-based gene prediction were 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 by hand curation with existing predictions. 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 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 (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 target DNA-query protein 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 by request.

Protein data sets, alignments, and trees: Nematode proteins used in this paper are found as fasta sequences in Datasets 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 dataset WS140 are marked by a terminal "m" in the fasta name, or by "-A" and "-B" in cases where 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 041227 (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 alignments were made with *clustalw* or *clustalx* with BLOSUM matrices. For protein trees, sequences were aligned and ends were trimmed to shared sequence. In the nematode MATH-BTB family 14 proteins that included less than 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 to 1,000 samples. 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 2005; Kelley et al. 2000) 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 Figure S9 and possible sites of positive selection were determined from the codon analysis shown in Figure 6.

Plant Genes: Complete sets of coding sequences for *Arabidopsis thaliana* were from TAIR release 6 (http://www.arabidopsis.org/) and for *Oryza sativa* from TIGR 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 1,327 from *Oryza*). A few proteins were removed prior to tree construction because the proteins were less than 150 amino acids or more than 800 amino acids, suggesting a prediction abnormality. A tree of 701 *Arabidopsis* proteins 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 Figure S10. Strong evidence for positive selection was obtained for all six sets (summarized in Table S5) and data from one set are shown in Figure S11. 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, PFAM, and SMART (http://smart.embl-heidelberg.de/; http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi; http://www.sanger.ac.uk/Software/Pfam/ 2005).

Acknowledgements

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, and Richard Palmiter for the idea that these proteins might be directed toward toxins produced by digestion.

Figure Legends

Figure 1. Schematic of ubiquitin-targeting system. 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. 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 is ubiquitin.

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 protein and MATH-BTB domain proteins. Domains that bind Cullins are shades of blue and domains that bind substrate are shades of orange. F-box family A2 has an additional domain (M, mariner) of unknown function that is related to the DNA-binding domain of mariner transposases (green). The region of highest sequence diversity in each F-box family is labeled "hypervariable".

Figure 3. Protein trees for nematode F-box-FTH proteins. A) Unrooted neighbor-joining distance tree for 383 F-box proteins from *C. elegans* (green), *C. briggsae* (blue), and *C. remanei* (red). Scale bar indicates distance in amino acid changes per site. Selected bootstrap values are marked in black (% bootstrap support from 200 replicates); all cases of large species-specific clades with bootstrap support ≥ 30% and all cases of possible one-to-one orthologs are shown. Possible one-to-one orthologs with bootstrap values ≥ 90% are marked with a black dot. B) Protein trees for 117 F-box proteins, colored as in panel A, including 69 proteins that constitute 23 ortholog trios from the three species. Additional randomly selected F-box-FTH and F-box-FBA2 proteins from *C. elegans* were included in the tree for comparison; the full tree has the same clear separation from all stable proteins. The 23 sets of probable orthologs are marked with a filled black circle. Where branch joins are absent proteins were unalignable outside their F-box domain (see Materials and Methods). Relevant subtrees were tested by 1,000 bootstrap replicates and the percent bootstrap support is marked in black. Cases in which the C-terminal region contained a domain identified by rps-blast or Pfam searches are marked by brackets and the domain name ("?" indicates hits with poor E-values). Genes with described phenotypes in *C. elegans* are marked "Phen" and those with *blastp* scores > 50 (E-value < 10^{-6} for the search) to a mouse protein are marked "M". One *C. remanei* protein that appears somewhat divergent is marked by an asterisk because the prediction is truncated by the end of a contig in the current *C. remanei* assembly.

Figure 4. 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 through G) of the extended FTH domain are marked above the top alignment. The jagged line indicates the position of a possible β-sheet. 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 possible investigator bias the alignments shown were not hand modified - a few places with possible artifactual alignment are apparent (for example, misaligned R residues near the N-terminal end of the PFAM-designated FTH domain). 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.

Figure 5. Protein tree for nematode MATH-BTB family. Unrooted maximum-likelihood tree for 164 MATH-BTB proteins from *C. elegans* (green), *C. briggsae* (blue), and *C. remanei* (red). Scale bar indicates distance in amino acid changes per site. Selected bootstrap values are marked in black (% bootstrap support from 1000 replicates). Probable ortholog trios are marked with a black dot. The eight *C. elegans* proteins with the best *blastp* matches to mouse proteins are marked M.

Figure 6. 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 from *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 an amino acid change in any of the three sequences in the MATH or BTB domains are marked with a black dot.

Figure 7. Structural model of MATH domain. The structure is TRAF6 with bound RANK peptide (PDB 1LB5), colored according to degree of amino acid conservation among nematode MATH domains (Figure S9). Mapping from nematode MATH domains to TRAF6 is based on a 3D-PSSM structural alignment to the MATH domain of C08C3.2 (see Text S1). In the space-filled models, long regions of high conservation are colored dark green, long regions of diversity are colored yellow, sites of putative positive selection are colored red, and other regions are colored khaki. Residues in TRAF6 that were not aligned with C08C3.2 are colored grey. The bound peptide is shown in grey-blue wireframe. The two large views are rotated nearly 180^o from each other and are rotation-centered on the bound peptide (front) and the most conserved regions (back). The smaller space-filled side view below shows the binding cleft in TRAF6 more clearly. The ribbon-view shows the 8-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 grey.

Figure 8. 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 ortholog trios plus four genes that did not have specific one-to-one orthologs because of a single recent duplication or loss in one species. Gene bins are 100 Kb in length.

Figure S1. Multiple alignment of a large set of F-box-FTH proteins. Multiple alignment of a diverse set of full-length F-box-FTH (FBXA) proteins. Blue alignment shading is proportional to the sum-of-pairs score for each amino acid residue relative to its aligned column. The F-box domain and the eight conserved blocks of the extended FTH domain are marked in blue above the alignment. Each of these eight blocks is separated by regions with high amino acid diversity and changes in length. The region between the F-box domain and FTH block A is notably long and diverse and is marked in red as the hypervariable domain.

Figure S2. F-box-FTH family d_N/d_S tree sets. Protein tree of a subset of F-box-FTH proteins, including those used for d_N/d_S analysis. Proteins used for each d_N/d_S test are color coded and lettered to correspond to other figures. Two proteins (faded blue) in the set A clade were eliminated from the d_N/d_S analysis because a region aligned poorly.

Figure S3. d_N/d_S results for three additional sets of F-box-FTH genes. See Figure 4 legend.

Figure S4. Multiple alignment of a large set of F-box-FBA2 proteins. See Figure S1 legend. The FBA2 domain has no detectable sequence similarity to the FTH domain.

Figure S5. d_N/d_S results for one set F-box-FBA2 genes. See Figure 4 legend. The domains marked in blue are the F-box domain, the PFAM-recognized FBA2 domain and two conserved segments of an extended FBA2 domain (labeled A and B).

Figure S6. MATH-BTB family d_N/d_S tree sets. Protein distance tree of MATH-BTB proteins considered for d_N/d_S analysis. Proteins used for two d_N/d_S tests are color coded and lettered to correspond to other figures. The genes analyzed for Figure 6 do not form a clade; they were selected for high quality alignment to ensure accurate d_N/d_S analysis and are not marked on this tree. Genes excluded from Figure 6 are labeled in grey because they contained a region of questionable alignment.

Figure S7. d_N/d_S for an additional set of MATH-BTB genes. See Figure 6 legend. Alignment and maximum-likelihood d_N/d_S values for ten MATH-BTB proteins from *C. elegans* (set A from Figure S6).

Figure S8. MATH-BTB family genome positions. The genome positions of all identified MATH-BTB genes in *C. elegans*. The nine genes with bootstrap supported orthologs in *C. briggsae* and *C. remanei* (see Figure 5) are shown in red. All other genes are shown in blue and are strongly clustered, consistent with evolution by local gene duplication. Gene bins were 100 Kb in length.

Figure S9. Alignment of 84 proteins in the unstable MATH-BTB family. All well-aligned full-length members of the unstable MATH-BTB families from *C. elegans*, *C. briggsae*, and *C. remanei* are shown. MATH and BTB domains are marked in blue above the alignment. Blue alignment shading is proportional to the sum-of-pairs score for each amino acid residue relative to its aligned column. For the MATH domain only, regions of four or more amino acids with high diversity are marked in yellow, regions of four or more amino acids with high average conservation are marked in green, and sites of probable positive selection are marked in red. Additional variation is present in family members not shown, largely in the regions marked yellow in this alignment. The figure was end-trimmed to the first sites that are well-conserved among all family members.

Figure S10. Protein classification tree for the *Arabidopsis thaliana* F-box superfamily. Neighbor-joining distance tree of 701 *Arabidopsis* F-box proteins, based on pairwise protein distances. Note that this method does not necessarily produce a tree that accurately reflects evolutionary ancestry, as it includes pair alignment distances between proteins that are unrelated outside of their F-box domain. Six of the largest expanded groups are shown in color and are labeled A through F; the letters correspond to the groups tested for positive selection (Table S5). 89 proteins with a *blastp* match to any *Oryza sativa* protein with E-value $\leq 10^{-80}$ (blast score density ~0.75) are marked with filled black circles. 52 proteins with an E-value between 10^{-30} (blast score density ~0.4) and 10^{-80} are marked with open black circles. Proteins with known functions are labeled in blue, based on the following references (Devoto et al. 2002; Dharmasiri et al. 2005a; Dieterle et al. 2001; Dill et al. 2004; Gagne et al. 2004; Kepinski and Leyser 2005; Kim and Delaney 2002; McGinnis et al. 2003; Nelson et al. 2000; Qiao et al. 2004; Samach et al. 1999; Somers et al. 2000; Strader et al. 2004; Wang et al. 2004; Xu et al. 2002).

Figure S11. d_N/d_S example from *Arabidopsis thaliana* F-box family A. Sample alignment and maximum-likelihood d_N/d_S values for a set of 9 F-box proteins from *Arabidopsis* family A (see Figure S10). The F-box and FBD (Smart00579) domains are marked above the alignment. Blue alignment shading is proportional to the sum-of-pairs score for each amino acid residue relative to its aligned column. The histogram section 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 positive selection are marked with a red asterisk ($P \geq 0.9$) or red square ($P \geq 0.8$).

References (web site references not separated)

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

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

Andrade, M.A., M. Gonzalez-Guzman, R. Serrano, and P.L. Rodriguez. 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., J.P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18: 1585-1592.

Anisimova, M., J.P. Bielawski, and Z. Yang. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* 19: 950-958.

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

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

Brunson, L.E., C. Dixon, A. LeFebvre, L. Sun, and N. Mathias. 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., C.C. Norbury, Y. Cho, J.W. Yewdell, and J.R. Bennink. 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., S.W. Jin, A. Cohen, and R.E. Ellis. 2004. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. *Genome Res* 14: 1207-1220.

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

Clifford, R., M.H. Lee, S. Nayak, M. Ohmachi, F. Giorgini, and T. Schedl. 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., N. Pujol, J. Reboul, L. Sabatier, J.F. Guichou, Y. Kohara, and J.J. Ewbank. 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 J.R. Krebs. 1979. Arms races between and within species. *Proc R Soc Lond B Biol Sci* 205: 489-511.

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

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

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

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

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

Dill, A., S.G. Thomas, J. Hu, C.M. Steber, and T.P. Sun. 2004. The Arabidopsis F-box protein SLEEPY1 targets gibberellin signaling repressors for gibberellin-induced degradation. *Plant Cell* 16: 1392-1405.

Dreier, L., M. Burbea, and J.M. Kaplan. 2005. LIN-23-mediated degradation of beta-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 K. Sandvig. 2000. Penetration of protein toxins into cells. *Curr Opin Cell Biol* 12: 407-413.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.6a2., pp. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Figueroa, P., G. Gusmaroli, G. Serino, J. Habashi, L. Ma, Y. Shen, S. Feng, M. Bostick, J. Callis, H. Hellmann, and X.W. Deng. 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.

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.9/. 2004. NCBI Blast download 2.2.9.

ftp://genome.wustl.edu/pub/seqmgr/remanei/pcap/remanei_041227/. Caenorhabditis remanei Pcap genome assembly download site.

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

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

Gagne, J.M., J. Smalle, D.J. Gingerich, J.M. Walker, S.D. Yoo, S. Yanagisawa, and R.D. Vierstra. 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 U S A* 101: 6803-6808.

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

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

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

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

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

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

http://faculty.washington.edu/~jstorey/qvalue/. Q-value software download site.

http://smart.embl-heidelberg.de/. SMART: Simple Modular Architecture Research Tool, protein domain search page.

http://wormbase.org/. WormBase home page, pp. WormBase home page.

http://www.arabidopsis.org/. TAIR: The Arabidopsis Information Resource home page.

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi. CD Search: NCBI conserved domain search page.

http://www.sanger.ac.uk/Software/Pfam/. 2005. Pfam: Protein families database of alignments and HMMs.

http://www.tigr.org/tdb/e2k1/osa1/. The Institute for Genomic Research (TIGR): rice genome annotation.

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

Hughes, A.L. and M. Nei. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci U S A* 86: 958-962.

Hughes, A.L., T. Ota, and M. Nei. 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., M. Rialland, D. Glaise, and C. Guguen-Guillouzo. 1999. Identification of a novel Skp2-like mammalian protein containing F-box and leucine-rich repeats. *FEBS Lett* 459: 75-79.

Jager, S., H.T. Schwartz, H.R. Horvitz, and B. Conradt. 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 U S A* 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 G. Struhl. 1998. Regulation of the Hedgehog and Wingless signalling pathways by the F-box/WD40-repeat protein Slimb. *Nature* 391: 493-496.

Kamath, R.S., A.G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, D.P. Welchman, P. Zipperlen, and J. Ahringer. 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* 421: 231-237.

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

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

Kelley, L.A. 2005. 3D-PSSM Web Server, pp. 3D-PSSM Web Server 2.6.0 - protein fold recognition.

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

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

Kim, H.S. and T.P. Delaney. 2002. Arabidopsis SON1 is an F-box protein that regulates a novel induced defense response independent of both salicylic acid and systemic acquired resistance. *Plant Cell* 14: 1469-1482.

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 J.L. King. 1979. Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. *Proc Natl Acad Sci U S A* 76: 2858-2861.

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

Kloetzel, P.M. and F. Ossendorp. 2004. Proteasome and peptidase function in MHC-class-I-mediated antigen presentation. *Curr Opin Immunol* 16: 76-81.

Kogelberg, H. and T. Feizi. 2001. New structural insights into lectin-type proteins of the immune system. *Curr Opin Struct Biol* 11: 635-643.

Li, C., C.Z. Ni, M.L. Havert, E. Cabezas, J. He, D. Kaiser, J.C. Reed, A.C. Satterthwait, G. Cheng, and K.R. Ely. 2002a. Downstream regulator TANK binds to the CD40 recognition site on TRAF3. *Structure (Camb)* 10: 403-411.

Li, H., W.X. Li, and S.W. Ding. 2002b. Induction and suppression of RNA silencing by an animal virus. *Science* 296: 1319-1321.

Li, J., A.M. Pauley, R.L. Myers, R. Shuang, J.R. Brashler, R. Yan, A.E. Buhl, C. Ruble, and M.E. Gurney. 2002c. SEL-10 interacts with presenilin 1, facilitates its ubiquitination, and alters A-beta peptide production. *J Neurochem* 82: 1540-1548.

Liao, E.H., W. Hung, B. Abrams, and M. Zhen. 2004. An SCF-like ubiquitin ligase complex that controls presynaptic differentiation. *Nature* 430: 345-350.

Lu, J., C. Teh, U. Kishore, and K.B. Reid. 2002. Collectins and ficolins: sugar pattern recognition molecules of the mammalian innate immune system. *Biochim Biophys Acta* 1572: 387-400.

Lu, R., M. Maduro, F. Li, H.W. Li, G. Broitman-Maduro, W.X. Li, and S.W. Ding. 2005. Animal virus replication and RNAi-mediated antiviral silencing in Caenorhabditis elegans. *Nature* 436: 1040-1043.

Marchler-Bauer, A. and S.H. Bryant. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 32: W327-331.

McGinnis, K.M., S.G. Thomas, J.D. Soule, L.C. Strader, J.M. Zale, T.P. Sun, and C.M. Steber. 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., L. Martinez-Pomares, and S. Gordon. 2004. Divergent roles for C-type lectins expressed by cells of the innate immune system. *Mol Immunol* 41: 1109-1121.

McWhirter, S.M., S.S. Pullen, J.M. Holton, J.J. Crute, M.R. Kehry, and T. Alber. 1999. Crystallographic analysis of CD40 recognition and signaling by human TRAF2. *Proc Natl Acad Sci U S A* 96: 8408-8413.

Moon, J., G. Parry, and M. Estelle. 2004. The ubiquitin-proteasome pathway and plant development. *Plant Cell* 16: 3181-3195.

Mota, L.J. and G.R. Cornelis. 2005. The bacterial injection kit: type III secretion systems. *Ann Med* 37: 234-249.

Nandi, D., H. Jiang, and J.J. Monaco. 1996. Identification of MECL-1 (LMP-10) as the third IFN-gamma-inducible proteasome subunit. *J Immunol* 156: 2361-2364.

Nayak, S., J. Goree, and T. Schedl. 2005. fog-2 and the evolution of self-fertile hermaphroditism in Caenorhabditis. *PLoS Biol* 3: e6.

Nelson, D.C., J. Lasswell, L.E. Rogg, M.A. Cohen, and B. Bartel. 2000. FKF1, a clock-controlled gene that regulates the transition to flowering in Arabidopsis. *Cell* 101: 331-340.

Nicholas, H.R. and J. Hodgkin. 2004. Responses to infection and possible recognition strategies in the innate immune system of Caenorhabditis elegans. *Mol Immunol* 41: 479-493.

Nielsen, R. and Z. Yang. 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., V. Burkitt, A.R. Villa, L. Tong, and H. Wu. 1999. Structural basis for self-association and receptor recognition of human TRAF2. *Nature* 398: 533-538.

Pintard, L., J.H. Willis, A. Willems, J.L. Johnson, M. Srayko, T. Kurz, S. Glaser, P.E. Mains, M. Tyers, B. Bowerman, and M. Peter. 2003. The BTB protein MEL-26 is a substrate-specific adaptor of the CUL-3 ubiquitin-ligase. *Nature* 425: 311-316.

Prag, S. and J.C. Adams. 2003. Molecular phylogeny of the kelch-repeat superfamily reveals an expansion of BTB/kelch proteins in animals. *BMC Bioinformatics* 4: 42.

Qiao, H., H. Wang, L. Zhao, J. Zhou, J. Huang, Y. Zhang, and Y. Xue. 2004. The F-box protein AhSLF-S2 physically interacts with S-RNases that may be inhibited by the ubiquitin/26S proteasome pathway of protein degradation during compatible pollination in Antirrhinum. *Plant Cell* 16: 582-595.

Samach, A., J.E. Klenz, S.E. Kohalmi, E. Risseeuw, G.W. Haughn, and W.L. Crosby. 1999. The UNUSUAL FLORAL ORGANS gene of Arabidopsis thaliana is an F-box protein required for normal patterning and growth in the floral meristem. *Plant J* 20: 433-445.

Sayle, R.A. and E.J. Milner-White. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem Sci* 20: 374.

Schedl, T. and J. Kimble. 1988. fog-2, a germ-line-specific sex determination gene required for hermaphrodite spermatogenesis in Caenorhabditis elegans. *Genetics* 119: 43-61.

Schulman, B.A., A.C. Carrano, P.D. Jeffrey, Z. Bowen, E.R. Kinnucan, M.S. Finnin, S.J. Elledge, J.W. Harper, M. Pagano, and N.P. Pavletich. 2000. Insights into SCF ubiquitin ligases from the structure of the Skp1-Skp2 complex. *Nature* 408: 381-386.

Somers, D.E., T.F. Schultz, M. Milnamow, and S.A. Kay. 2000. ZEITLUPE encodes a novel clock-associated PAS protein from Arabidopsis. *Cell* 101: 319-329.

Stogios, P.J., G.S. Downs, J.J. Jauhal, S.K. Nandra, and G.G. Prive. 2005. Sequence and structural analysis of BTB domain proteins. *Genome Biol* 6: R82.

Storey, J.D. and R. Tibshirani. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100: 9440-9445.

Strader, L.C., S. Ritchie, J.D. Soule, K.M. McGinnis, and C.M. Steber. 2004. Recessive-interfering mutations in the gibberellin signaling gene SLEEPY1 are rescued by overexpression of its homologue, SNEEZY. *Proc Natl Acad Sci U S A* 101: 12771-12776.

Swanson, W.J., Z. Yang, M.F. Wolfner, and C.F. Aquadro. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc Natl Acad Sci U S A* 98: 2509-2514.

Thomas, J.H. 2005. Analysis of homologous gene clusters in C. elegans reveals striking regional cluster domains. *Genetics* In press.

Thomas, J.H., J.L. Kelley, H.M. Robertson, K. Ly, and W.J. Swanson. 2005. Adaptive evolution in the SRZ chemoreceptor families of Caenorhabditis elegans and Caenorhabditis briggsae. *Proc Natl Acad Sci U S A* 102: 4476-4481.

Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25: 4876-4882.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 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.

Toes, R.E., A.K. Nussbaum, S. Degermann, M. Schirle, N.P. Emmerich, M. Kraft, C. Laplace, A. Zwinderman, T.P. Dick, J. Muller, B. Schonfisch, C. Schmid, H.J. Fehling, S. Stevanovic, H.G. Rammensee, and H. Schild. 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: R59-61.

Varshavsky, A. 2005. Regulated protein degradation. *Trends Biochem Sci* 30: 283-286.

Vasudevan, S. and S.W. Peltz. 2003. Nuclear mRNA surveillance. *Curr Opin Cell Biol* 15: 332-337.

Wang, L., L. Dong, Y. Zhang, Y. Zhang, W. Wu, X. Deng, and Y. Xue. 2004. Genome-wide analysis of S-Locus F-box-like genes in Arabidopsis thaliana. *Plant Mol Biol* 56: 929-945.

Weber, H., A. Bernhardt, M. Dieterle, P. Hano, A. Mutlu, M. Estelle, P. Genschik, and H. Hellmann. 2005. Arabidopsis AtCUL3a and AtCUL3b form complexes with members of the BTB/POZ-MATH protein family. *Plant Physiol* 137: 83-93.

Wilkins, C., R. Dishongh, S.C. Moore, M.A. Whitt, M. Chow, and K. Machaca. 2005. RNA interference is an antiviral defence mechanism in Caenorhabditis elegans. *Nature* 436: 1044-1047.

Winston, J.T., D.M. Koepp, C. Zhu, S.J. Elledge, and J.W. Harper. 1999. A family of mammalian F-box proteins. *Curr Biol* 9: 1180-1182.

Xu, L., F. Liu, E. Lechner, P. Genschik, W.L. Crosby, H. Ma, W. Peng, D. Huang, and D. Xie. 2002. The SCF(COI1) ubiquitin-ligase complexes are required for jasmonate response in Arabidopsis. *Plant Cell* 14: 1919-1935.

Xu, L., Y. Wei, J. Reboul, P. Vaglio, T.H. Shin, M. Vidal, S.J. Elledge, and J.W. Harper. 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., Bielawski, B. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15: 496-503.

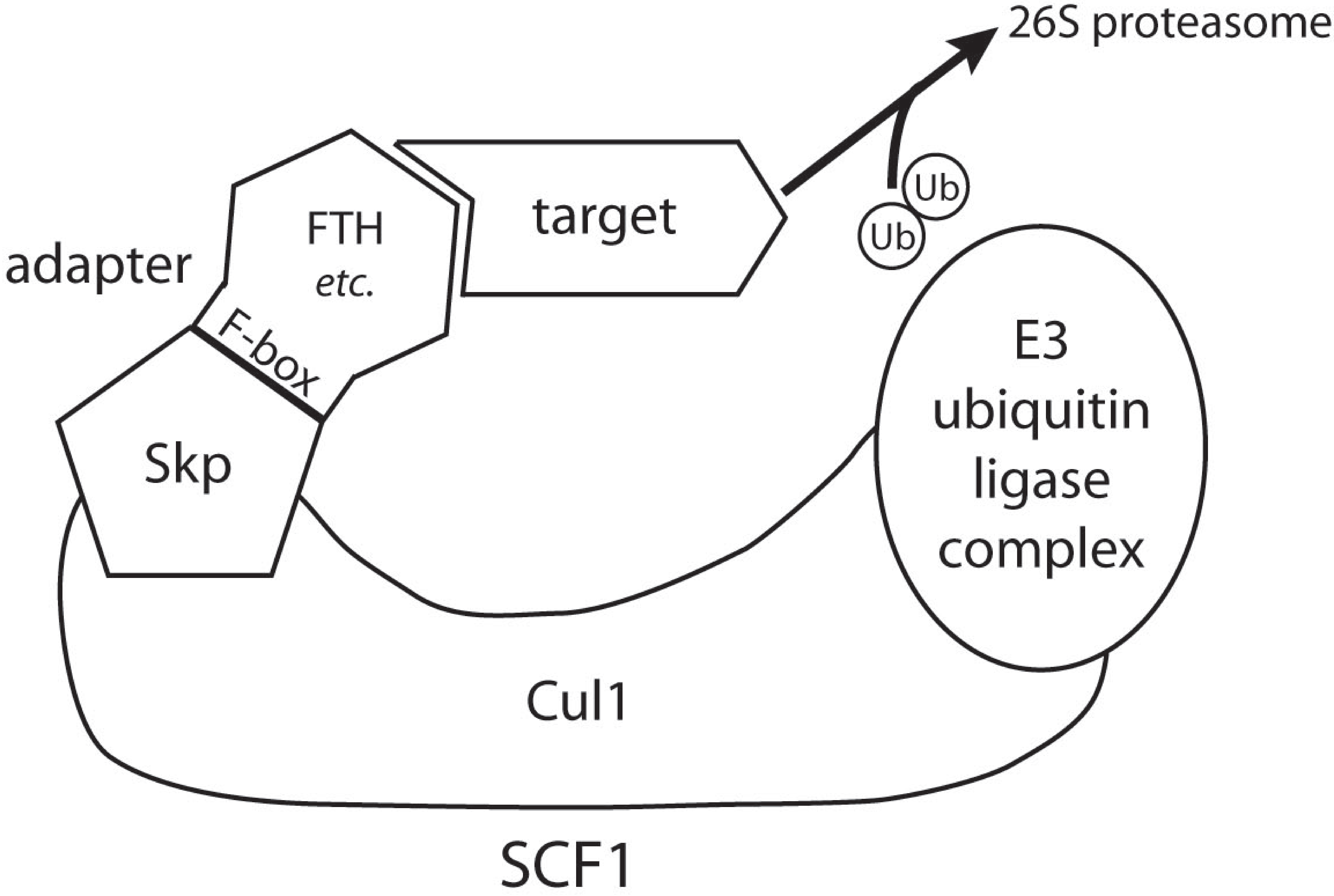
Yang, Z. and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19: 908-917.

Yang, Z., R. Nielsen, N. Goldman, and A.M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155: 431-449.

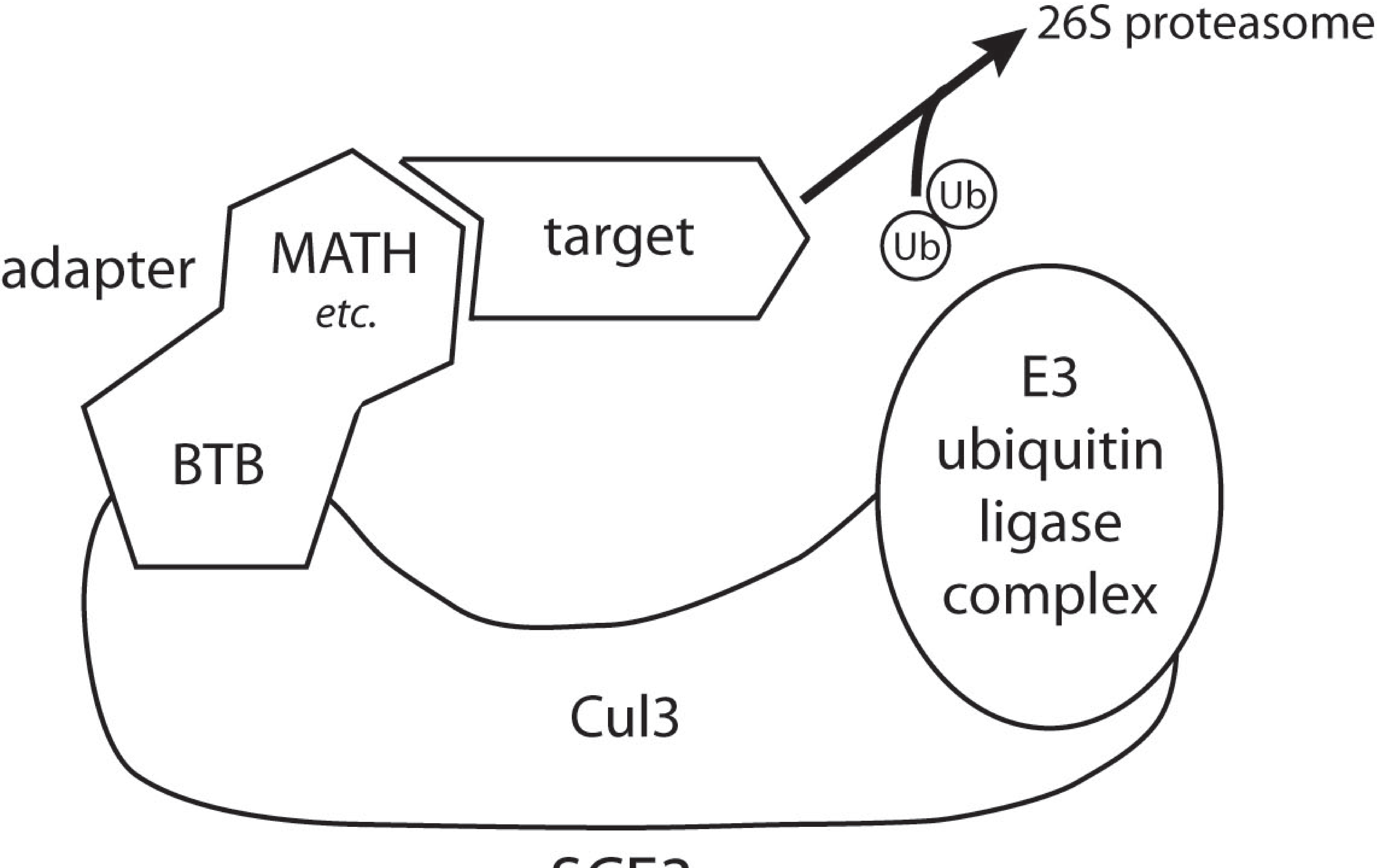
Yang, Z. and W.J. Swanson. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* 19: 49-57.

Ye, H., J.R. Arron, B. Lamothe, M. Cirilli, T. Kobayashi, N.K. Shevde, D. Segal, O.K. Dzivenu, M. Vologodskaia, M. Yim, K. Du, S. Singh, J.W. Pike, B.G. Darnay, Y. Choi, and H. Wu. 2002. Distinct molecular mechanism for initiating TRAF6 signalling. *Nature* 418: 443-447.

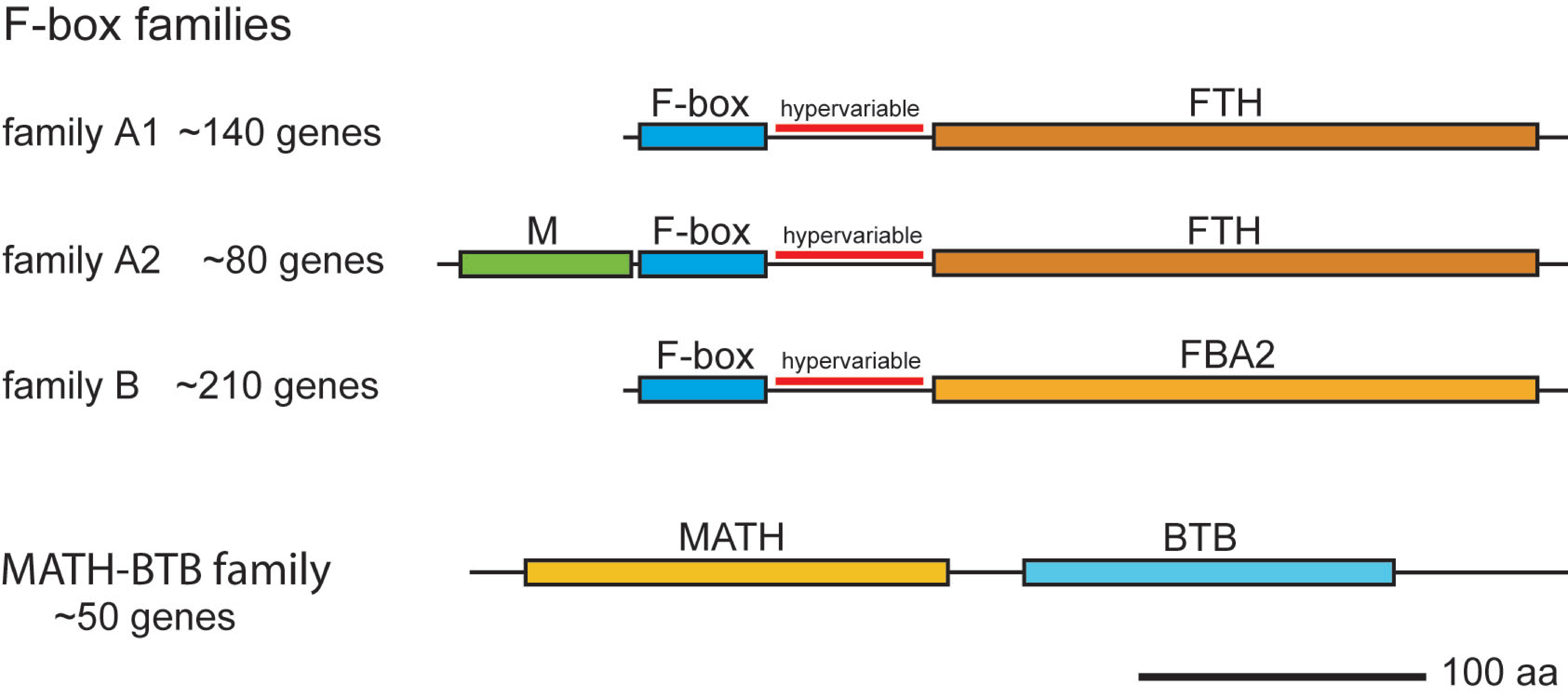
Zheng, N., B.A. Schulman, L. Song, J.J. Miller, P.D. Jeffrey, P. Wang, C. Chu, D.M. Koepp, S.J. Elledge, M. Pagano, R.C. Conaway, J.W. Conaway, J.W. Harper, and N.P. Pavletich. 2002. Structure of the Cul1-Rbx1-Skp1-F boxSkp2 SCF ubiquitin ligase complex. *Nature* 416: 703-709.

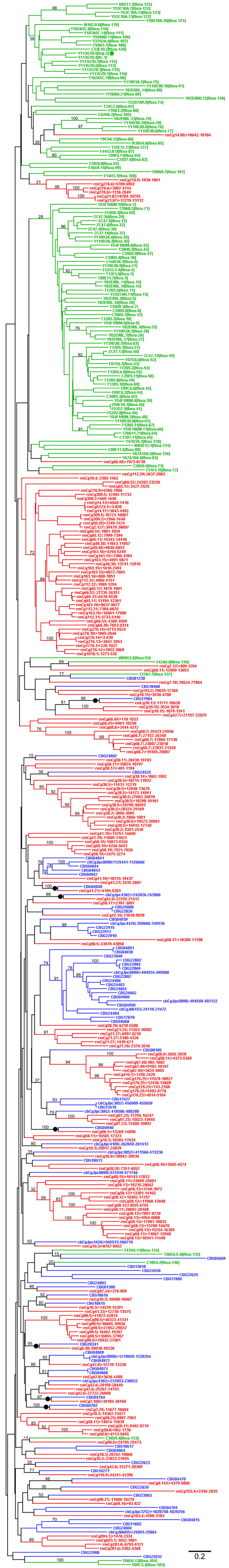

26S proteasome
target
FTH
etc.
adapter
F-box
Ub
Ub
E3
ubiquitin
ligase
complex
Skp
Cul1

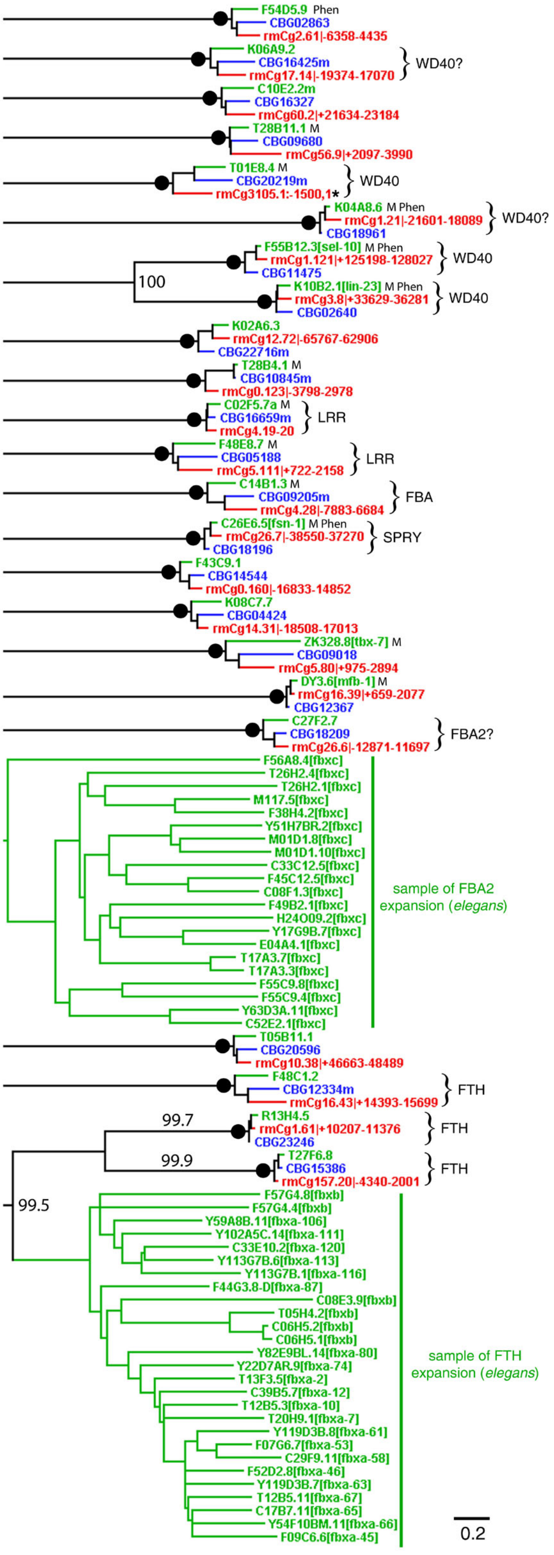
SCF1

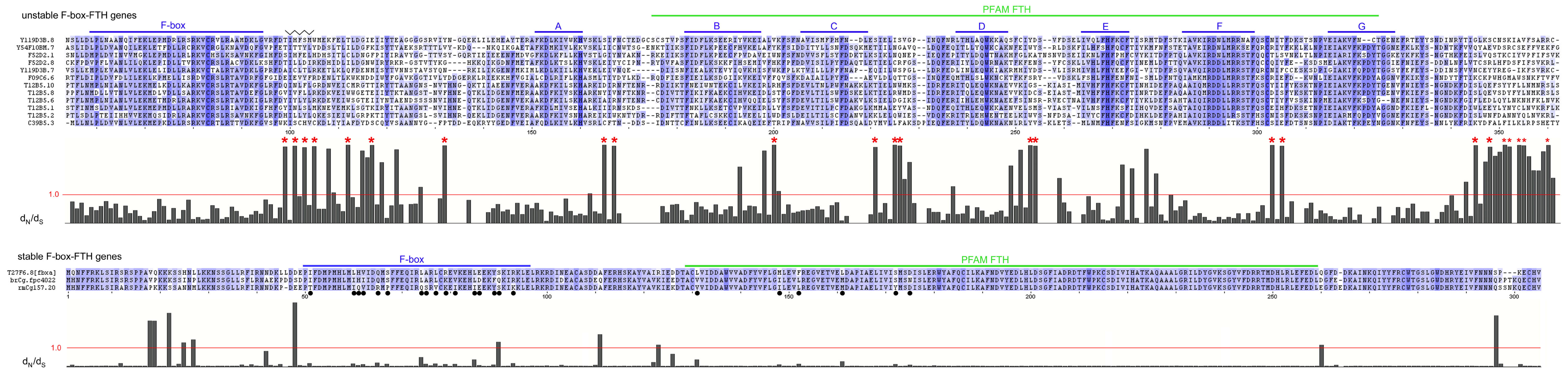

26S proteasome
target
MATH
etc.
adapter
Ub
Ub
E3
ubiquitin
ligase
complex
BTB
Cul3

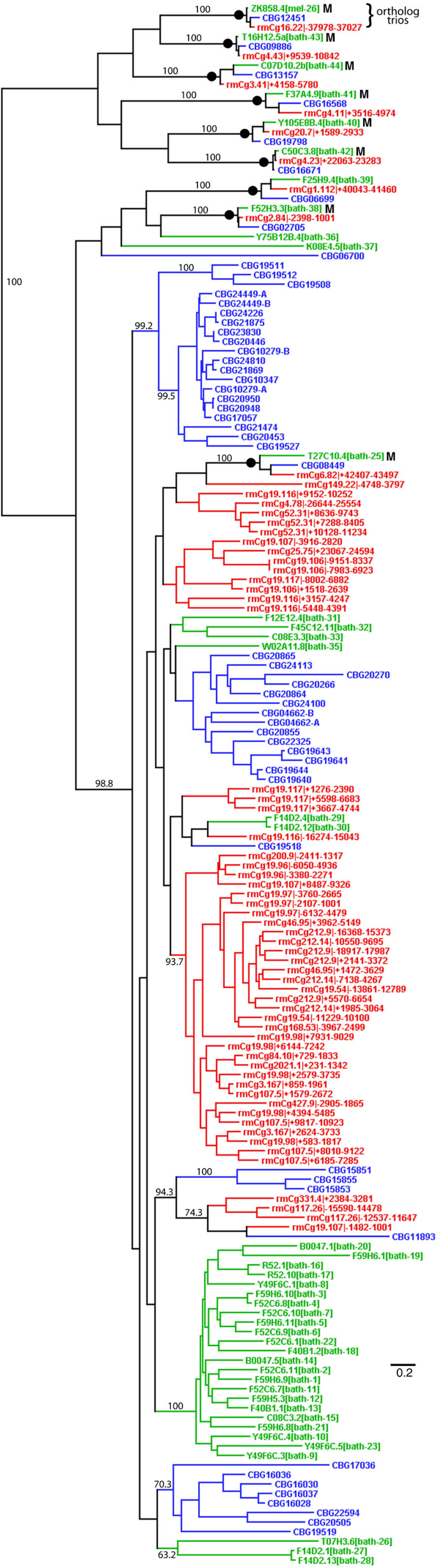
SCF3

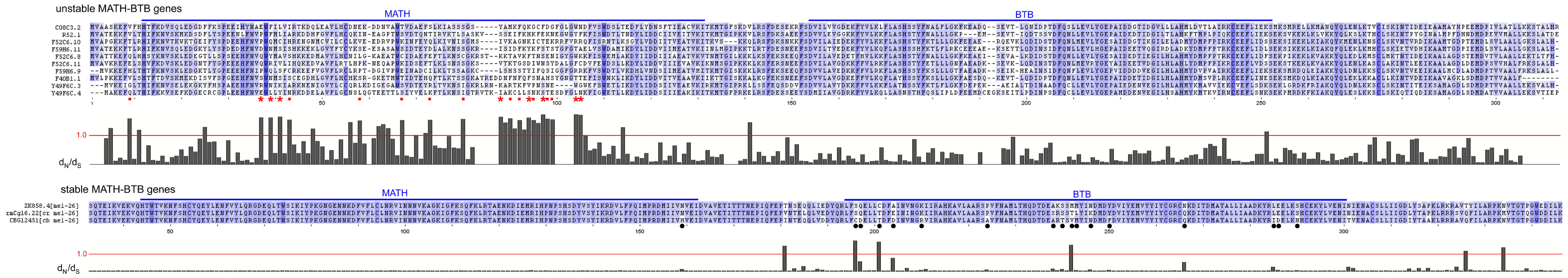

F-box families
family A1 ~140 genes
F-box
hypervariable
FTH
family A2 ~80 genes
M
F-box
hypervariable
FTH
family B ~210 genes
F-box
hypervariable
FBA2
MATH-BTB family
~50 genes
MATH
BTB
100 aa

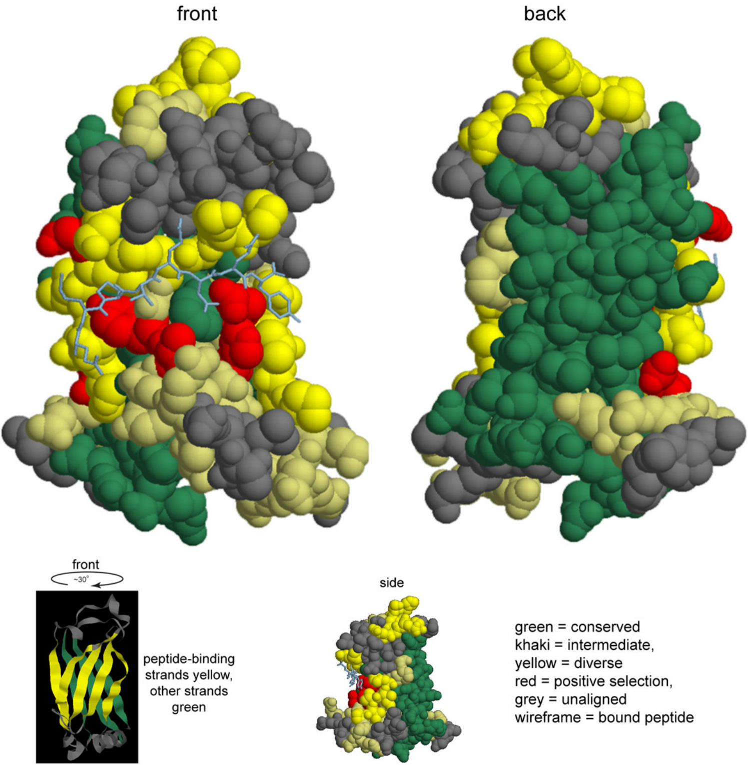

0.2

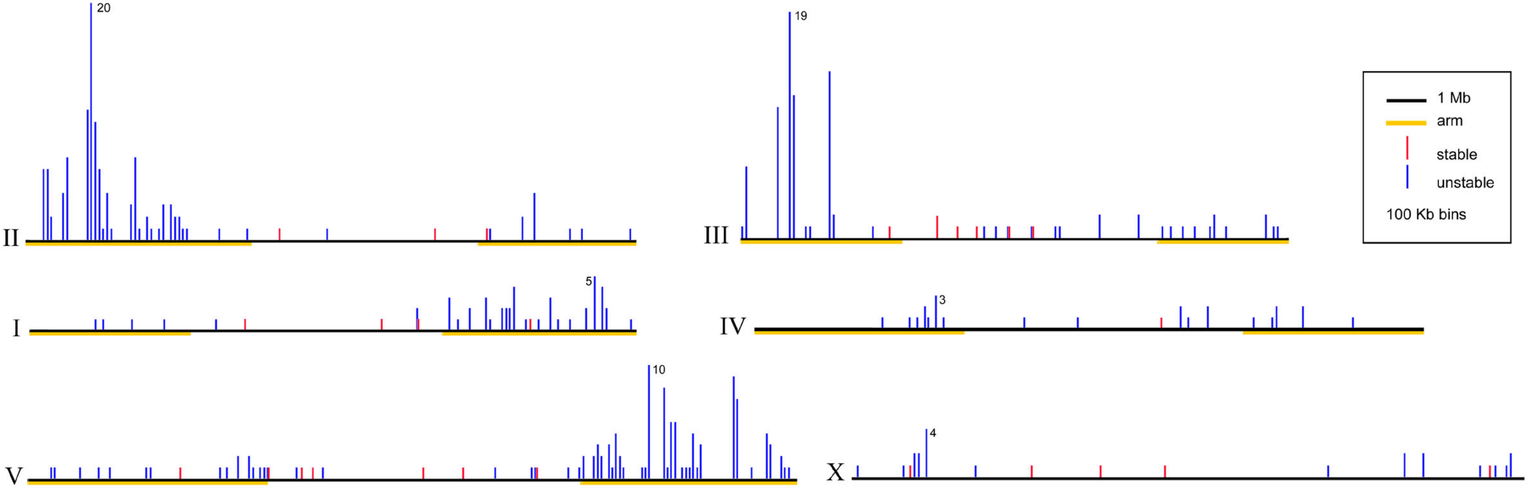

F54D5.9 Phen
CBG02863
rmCg2.61|-6358-4435
K06A9.2
CBG16425m
rmCg17.14|-19374-17070
WD40?
C10E2.2m
CBG16327
rmCg60.2|+21634-23184
T28B11.1 M
CBG09680
rmCg56.9|+2097-3990
T01E8.4 M
CBG20219m
rmCg3105.1:-1500,1*
WD40
K04A8.6 M Phen
rmCg1.21|-21601-18089
CBG18961
WD40?
F55B12.3[sel-10] M Phen
rmCg1.121|+125198-128027
CBG11475
WD40
100
K10B2.1[lin-23] M Phen
rmCg3.8|+33629-36281
CBG02640
WD40
K02A6.3
rmCg12.72|-65767-62906
CBG22716m
T28B4.1 M
CBG10845m
rmCg0.123|-3798-2978
C02F5.7a M
CBG16659m
rmCg4.19-20
LRR
F48E8.7 M
CBG05188
rmCg5.111|+722-2158
LRR
C14B1.3 M
CBG09205m
rmCg4.28|-7883-6684
FBA
C26E6.5[fsn-1] M Phen
rmCg26.7|-38550-37270
CBG18196
SPRY
F43C9.1
CBG14544
rmCg0.160|-16833-14852
K08C7.7
CBG04424
rmCg14.31|-18508-17013
ZK328.8[tbx-7] M
CBG09018
rmCg5.80|+975-2894
DY3.6[mfb-1] M
rmCg16.39|+659-2077
CBG12367
C27F2.7
CBG18209
rmCg26.6|-12871-11697
FBA2?
F56A8.4[fbxc]
T26H2.4[fbxc]
T26H2.1[fbxc]
M117.5[fbxc]
F38H4.2[fbxc]
Y51H7BR.2[fbxc]
M01D1.8[fbxc]
M01D1.10[fbxc]
C33C12.5[fbxc]
F45C12.5[fbxc]
C08F1.3[fbxc]
F49B2.1[fbxc]
H24O09.2[fbxc]
Y17G9B.7[fbxc]
E04A4.1[fbxc]
T17A3.7[fbxc]
T17A3.3[fbxc]
F55C9.8[fbxc]
F55C9.4[fbxc]
Y63D3A.11[fbxc]
C52E2.1[fbxc]
sample of FBA2 expansion (*elegans*)
T05B11.1
CBG20596
rmCg10.38|+46663-48489
F48C1.2
CBG12334m
rmCg16.43|+14393-15699
FTH
99.7
R13H4.5
rmCg1.61|+10207-11376
CBG23246
FTH
99.9
T27F6.8
CBG15386
rmCg157.20|-4340-2001
FTH
99.5
F57G4.8[fbxb]
F57G4.4[fbxb]
Y59A8B.11[fbxa-106]
Y102A5C.14[fbxa-111]
C33E10.2[fbxa-120]
Y113G7B.6[fbxa-113]
Y113G7B.1[fbxa-116]
F44G3.8-D[fbxa-87]
C08E3.9[fbxb]
T05H4.2[fbxb]
C06H5.2[fbxb]
C06H5.1[fbxb]
Y82E9BL.14[fbxa-80]
Y22D7AR.9[fbxa-74]
T13F3.5[fbxa-2]
C39B5.7[fbxa-12]
T12B5.3[fbxa-10]
T20H9.1[fbxa-7]
Y119D3B.8[fbxa-61]
F07G6.7[fbxa-53]
C29F9.11[fbxa-58]
F52D2.8[fbxa-46]
Y119D3B.7[fbxa-63]
T12B5.11[fbxa-67]
C17B7.11[fbxa-65]
Y54F10BM.11[fbxa-66]
F09C6.6[fbxa-45]
sample of FTH expansion (*elegans*)
0.2


unstable F-box-FTH genes
F-box
A
PFAM FTH
B
C
D
E
F
G
Y119D3B.8
Y54F10BM.7
F52D2.1
F52D2.8
Y119D3B.7
F09C6.6
T12B5.10
T12B5.8
T12B5.6
T12B5.1
T12B5.2
C39B5.3
100
150
200
250
300
350
1.0
d_N/d_S
stable F-box-FTH genes
F-box
PFAM FTH
T27F6.8[fbxa]
brCg.fpc4022
rmCg157.20
1
50
100
150
200
250
300
1.0
d_N/d_S


ortholog trios
M
100
ZK858.4[mel-26]
CBG12451
rmCg16.22|-37978-37027
100
T16H12.5a[bath-43]
CBG09886
rmCg4.43|+9539-10842
100
C07D10.2b[bath-44]
CBG13157
rmCg3.41|+4158-5780
100
F37A4.9[bath-41]
CBG16568
rmCg4.11|+3516-4974
100
Y105E8B.4[bath-40]
rmCg20.7|+1589-2933
CBG19798
100
C50C3.8[bath-42]
rmCg4.23|+22063-23283
CBG16671
100
F25H9.4[bath-39]
rmCg1.112|+40043-41460
CBG06699
100
F52H3.3[bath-38]
rmCg2.84|-2398-1001
CBG02705
Y75B12B.4[bath-36]
K08E4.5[bath-37]
CBG06700
100
100
CBG19511
CBG19512
CBG19508
99.2
CBG24449-A
CBG24449-B
CBG24226
CBG21875
CBG23830
CBG20446
CBG10279-B
CBG24810
CBG21869
CBG10347
CBG10279-A
99.5
CBG20950
CBG20948
CBG17057
CBG21474
CBG20453
CBG19527
100
T27C10.4[bath-25]
CBG08449
rmCg6.82|+42407-43497
rmCg149.22|-4748-3797
rmCg19.116|+9152-10252
rmCg4.78|-26644-25554
rmCg52.31|+8636-9743
rmCg52.31|+7288-8405
rmCg52.31|+10128-11234
rmCg19.107|-3916-2820
rmCg25.75|+23067-24594
rmCg19.106|-9151-8337
rmCg19.106|-7983-6923
rmCg19.117|-8002-6882
rmCg19.106|+1518-2639
rmCg19.116|+3157-4247
rmCg19.116|-5448-4391
F12E12.4[bath-31]
F45C12.11[bath-32]
C08E3.3[bath-33]
W02A11.8[bath-35]
CBG20865
CBG24113
CBG20270
CBG20266
CBG20864
CBG24100
CBG04662-B
CBG04662-A
CBG20855
CBG22325
CBG19643
CBG19641
CBG19644
CBG19640
98.8
rmCg19.117|+1276-2390
rmCg19.117|+5598-6683
rmCg19.117|+3667-4744
F14D2.4[bath-29]
F14D2.12[bath-30]
rmCg19.116|-16274-15043
CBG19518
rmCg200.9|-2411-1317
rmCg19.96|-6050-4936
rmCg19.96|-3380-2271
rmCg19.107|+8487-9326
rmCg19.97|-3760-2665
rmCg19.97|-2107-1001
rmCg19.97|-6132-4479
rmCg46.95|+3962-5149
rmCg212.9|-16368-15373
rmCg212.14|-10550-9695
rmCg212.9|-18917-17987
rmCg212.9|+2141-3372
rmCg46.95|+1472-3629
rmCg212.14|-7138-4267
93.7
rmCg19.54|-13861-12789
rmCg212.9|+5570-6654
rmCg212.14|+1985-3064
rmCg19.54|-11229-10100
rmCg168.53|-3967-2499
rmCg19.98|+7931-9029
rmCg19.98|+6144-7242
rmCg84.10|+729-1833
rmCg2021.1|+231-1342
rmCg19.98|+2579-3735
rmCg3.167|+859-1961
rmCg107.5|+1579-2672
rmCg427.9|-2905-1865
rmCg19.98|+4394-5485
rmCg107.5|+9817-10923
rmCg3.167|+2624-3733
rmCg19.98|+583-1817
rmCg107.5|+8010-9122
rmCg107.5|+6185-7285
100
CBG15851
CBG15855
CBG15853
94.3
rmCg331.4|+2384-3281
74.3
rmCg117.26|-15590-14478
rmCg117.26|-12537-11647
rmCg19.107|-1482-1001
CBG11893
B0047.1[bath-20]
F59H6.1[bath-19]
R52.1[bath-16]
R52.10[bath-17]
Y49F6C.1[bath-8]
F59H6.10[bath-3]
F52C6.8[bath-4]
F52C6.10[bath-7]
F59H6.11[bath-5]
F52C6.9[bath-6]
F52C6.1[bath-22]
F40B1.2[bath-18]
B0047.5[bath-14]
F52C6.11[bath-2]
F59H6.9[bath-1]
F52C6.7[bath-11]
0.2
100
F59H5.3[bath-12]
F40B1.1[bath-13]
C08C3.2[bath-15]
F59H6.8[bath-21]
Y49F6C.4[bath-10]
Y49F6C.5[bath-23]
Y49F6C.3[bath-9]
CBG17036
70.3
CBG16036
CBG16030
CBG16037
CBG16028
CBG22594
CBG20505
CBG19519
63.2
T07H3.6[bath-26]
F14D2.1[bath-27]
F14D2.13[bath-28]


unstable MATH-BTB genes
MATH
BTB
C08C3.2
R52.1
F52C6.10
F59H6.11
F52C6.8
F52C6.11
F59H6.9
F40B1.1
Y49F6C.3
Y49F6C.4
50
100
150
200
250
300
1.0
dN/dS
stable MATH-BTB genes
MATH
BTB
ZK858.4[mel-26]
rmCg16.22[cr mel-26]
CBG12451[cb mel-26]
50
100
150
200
250
300
1.0
dN/dS


front
back
front
~30°
peptide-binding strands yellow, other strands green
side
green = conserved
khaki = intermediate,
yellow = diverse
red = positive selection,
grey = unaligned
wireframe = bound peptide


20
II
19
III
1 Mb
arm
stable
unstable
100 Kb bins
5
I
3
IV
10
V
4
X