J Mol Evol (2009) 69:54–64
DOI 10.1007/s00239-009-9252-3

RBR Ubiquitin Ligases: Diversification and Streamlining in Animal Lineages

Ignacio Marín

Received: 19 November 2008 / Accepted: 26 May 2009 / Published online: 13 June 2009

Abstract The patterns of emergence and disappearance in animal species of genes encoding RBR ubiquitin ligases are described. RBR genes can be classified into subfamilies (Parkin, Ariadne, Dorfin, ARA54, etc.) according to sequence and structural data. Here, I show that most animal-specific RBR subfamilies emerged early in animal evolution, and that ancient animals, before the cnidarian/bilaterian split, had a set of RBR genes, which was as complex as the one currently found in mammals. However, some lineages (nematodes, dipteran insects) have recently suffered multiple losses, leading to a highly simplified set of RBR genes. Genes of a particular RBR subfamily, characterized by containing a helicase domain and so far found only in plants, are present also in some animal species. The meaning of these patterns of diversification and streamlining are discussed at the light of functional data. Extreme evolutionary conservation may be related to gene products having housekeeping functions.

Keywords Ubiquitination · Gene loss · Gene expression · Placozoa · Cnidaria

Electronic supplementary material The online version of this article (doi:10.1007/s00239-009-9252-3) contains supplementary material, which is available to authorized users.

I. Marín (✉)
Instituto de Biomedicina de Valencia, Consejo Superior de Investigaciones Científicas (IBV-CSIC), Calle Jaime Roig, 11, 46010 Valencia, Spain
e-mail: imarin@ibv.csic.es

Introduction

In eukaryotic cells, ubiquitination serves multiple purposes. The best known, generally associated to the addition of a polyubiquitin chain, is to tag a protein for destruction via the proteasome (Glickman and Ciechanover 2002; Schwartz and Ciechanover 2009). In addition, the correct function of molecular pathways as different as DNA repair, cellular signaling, or endocytosis, and also the activation of gene expression require particular proteins being ubiquitinated but not degraded. These regulatory roles often involve monoubiquitination at specific residues or, some times, assembly of atypical polyubiquitin chains (Welchman et al. 2005; Kerscher et al. 2006; Mukhopadhyay and Riezman 2007; Ikeda and Dikic 2008; Weake and Workman 2008). This relevance of ubiquitination in multiple processes has led to a great interest in unveiling the complex biochemical machinery involved in this finely regulated protein modification system.

Ubiquitination has three steps, each one of them involving a different type of protein: E1s or ubiquitin-activating enzymes, E2s or ubiquitin-conjugating enzymes, and E3s or ubiquitin ligases. Ubiquitin ligases are the most interesting, given that they provide most of the specificity to the ubiquitination process. This is easily deduced from the fact that E3s are much more numerous and diverse than E1s and E2s. In the human genome, current estimations suggest that there is a single E1 and less than 30 E2s, while there are many hundreds of ubiquitin ligases (Schwartz and Ciechanover 2009). It has been found that ubiquitin ligases contain characteristic protein domains, being the two most common the RING finger and the HECT domain. RING finger containing ubiquitin ligases are the most common type of E3s, with hundreds of members encoded in mammalian genomes. They can be divided into different types according to their sequence and structural similarity. Among the most characteristic RING finger containing E3s are the members of the RBR family of ubiquitin ligases (reviewed in Marín et al. 2004). Some of the members of the RBR family are being extensively studied due to their implication in human diseases. The best known is Parkin. Mutations in the *parkin* gene have been found in cases of juvenile, autosomal recessive familial Parkinson disease (Kitada et al. 1998). Evidence for a potential involvement of other RBR proteins in human diseases has been also obtained (see summary in Marín et al. 2004). RBR proteins are characterized by having a complex supradomain, composed of three consecutive domains, rich in cysteins and histidines. The first, N-terminal domain (RING1) is a canonical RING finger with a C_3HC_4 signature of conserved cysteine and histidine residues. The second domain, known as IBR, typically has a C_6HC signature and is present only in RBR proteins. Recent structural data have shown that the IBR domain in fact consists of two consecutive zinc-binding domains, respectively, based on C_4 and C_2HC ligands (Beasley et al. 2007). Finally, the third, C-terminal domain, often called RING2, resembles a RING finger (e.g., it also contains a C_3HC_4 signature). However, both the spacing of residues and the sequence of this domain are quite different from those characteristic of RING fingers (Marín and Ferrús 2002). Structural data indicated that the RING2 domain is not a canonical RING finger, but a totally different domain (Capili et al. 2004), which so far as been also found only in RBR proteins.

Given that the RING1–IBR–RING2 signature (or RBR signature), which characterizes the family is exclusively found in this type of proteins, it is easy to detect them in database searches. In previous studies, my group described the origin and evolution of the RBR protein family (Marín and Ferrús 2002; Marín et al. 2004; Lucas et al. 2006). We found that they are ancient, being present in all eukaryotes for which sequence data are available. In our most recent analyses, which included 347 protein sequences, we determined that, according to sequence similarity and structural features, the RBR family can be subdivided into at least 14 subfamilies (Lucas et al. 2006; see the Supplementary information of that work for complete phylogenetic analyses). Two of them, called Ariadne and ARA54, were found in both unikonts and bikonts, implying a very ancient origin. Most families, however, are restricted to some lineages. Particularly, we found that no less than six subfamilies (called RNF144, Dorfin, XAP3, PAUL, IBRDC1, and Parkin) were animal-specific. A latter study confirmed all these findings (Eisenhaber et al. 2007).

The recent sequencing of the genomes of two key organisms, the placozoan *Trichoplax adhaerens* and the cnidarian *Nematostella vectensis* (Putnam et al. 2007; Srivastava et al. 2008), which belong to basal branches of the animal tree, allow us to address three of the questions that remained to be answered regarding the evolution of the RBR family in animals. A first significant question is to determine when the animal-specific subfamilies emerged. The second question is to establish how the RBR family has expanded or contracted in different animal lineages. The third, final question is to propose a model of why the expansions and contractions of this gene family have occurred, based on our knowledge of the functions of the RBR proteins. Here, after considering the sequences and structures of 530 animal-specific RBR sequences, I trace the pattern of diversification of this family in animal lineages. The evolutionary history of the RBR family turns to be very complex. While most subfamilies emerged very early in animal evolution and some lineages have conserved virtually intact their RBR gene complement for hundreds of millions of years, other lineages have suffered loss of many genes, some times in short periods of time. Lineage-specific amplifications of some subfamilies are also found. The functional implications of these results are discussed.

Methods

A total of 34 different RBR signature sequences were used as queries in TBLASTN searches using the NR, EST, WGS, GSS, and HTGS databases at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/). These 34 sequences were selected because either they were representatives of the known RBR subfamilies or they were found in preliminary analyses to be significant dissimilar to the already characterized sequences. After compiling all the information obtained with the TBLASTN 170 searches, I detected that they had become saturated, i.e., all sequences appeared in multiple searches. After eliminating incomplete fragments (<180 amino acids long) and duplicates, the sequences were aligned using CLUSTALX 2.07 (Larkin et al. 2007) and the alignments were manually refined with the GeneDoc 2.7 sequence editor (Nicholas and Nicholas 1997). Finally, a database containing 1174 RBR sequences was obtained. I selected for this work the 530 animal sequences that were present in that database.

From the final alignment of the animal protein sequences, phylogenetic analyses were performed according to three different methods. First, neighbor-joining (NJ) trees were obtained using MEGA 4 (Tamura et al. 2007). As a second procedure, maximum-parsimony (MP) trees were obtained using PAUP*, beta 10 version (Swofford 2003). Finally, maximum-likelihood (ML) trees were obtained using PHYML (Guindon and Gascuel 2003). Parameters used were the same described in Lucas and Marín (2007).

For NJ, sites with gaps were included and Kimura's correction was used. For MP, the parameters were as follows: (1) all sites were included, (2) randomly generated trees were used as seeds, (3) the maximum number of tied trees saved was equal to 20, and (4) a heuristic search using the subtree pruning-regrafting algorithm was performed. This is not a very exhaustive procedure, and may lead to an underestimate of bootstrap support (see below), but had to be used in this case, given the large number of sequences to be analyzed. For ML trees, the BIONJ tree was taken as a starting point for the iterative ML searches, and calculations were performed using the Blosum62 matrix of amino acidic similarity. A total of 1,000 bootstrap replicates were performed to establish the reliability of the NJ and MP trees obtained. For ML, which is more computer intensive, I obtained 100 bootstrap replicates.

Structures were analyzed using a combination of analyses using both TBLASTN and the integrated tool InterProScan (http://www.ebi.ac.uk/Tools/InterProScan/), which detects known protein domains in multiple databases (Zdobnov and Apweiler 2001). When a human representative of a given subfamily existed, its full-length sequence (instead of just the RBR signature, as in the previously described analyses) was used as query in TBLASTN searches to determine whether the proteins that appear closely located in the phylogenetic trees based on the RBR signature also contained significant similarities in the regions that corresponded to the protein domains already characterized in the human proteins (described in Marín and Ferrús 2002; Marín et al. 2004; Lucas et al. 2006). Searches with canonical members of subfamilies for which no human proteins were present were similarly performed. The ORF finder routine, available at NCBI, (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) was used to expand the ORFs detected, if required. InterProScan was used to establish the domains present in all the proteins of the basal animals *T. adhaerens* and *N. vectensis*.

Functional data were obtained from SymAtlas (http://symatlas.gnf.org/SymAtlas/; Su et al. 2002, 2004) for human genes, FlyAtlas (http://130.209.132.177/atlas/atlas.cgi; Chintapalli et al. 2007), FlyBase (http://www.flybase.org/; Wilson et al. 2008), and the Vienna *Drosophila* RNAi Center (http://stockcenter.vdrc.at/control/main; Dietzl et al. 2007) for *Drosophila melanogaster* genes and Wormbase (http://www.wormbase.org/; Bieri et al. 2007) for *Caenorhabditis elegans* genes. For human RBR genes, the microarray data in SymAtlas (datasets GNF1H.gcRMA and GNF1H.MAS5, each one with 79 independent results for different tissues or cell types) were downloaded. For each RBR gene present in those datasets, the probe with the highest average level of expression was chosen and the number of tissues with a significantly high level of expression (i.e., expression value $\geq$ 200; see Su et al. 2002) was determined. For *D. melanogaster*, both microarray and genetic data for the five RBR genes present in that species were also downloaded in order to estimate the presence/absence of expression in different tissues and the effect of mutations. A gene was considered to have positive expression in a tissue when at least three of the four "calls" obtained in the four independent microarrays were positive (see http://130.209.132.177/atlas/about_atlas.html for a detailed explanation). Finally, genetic data for all *C. elegans* RBR genes were obtained in order to determine the phenotypic effects of either null mutations or down-expression using RNA interference.

Results

Nearly All Animal RBR Sequences Can Be Classified into just 12 Subfamilies

Figure 1 summarizes the results of all the phylogenetic analyses. At the subfamily and orthology group levels, NJ, MP, and ML results were congruent enough as to be fitted together into a single tree. Any reader interested in the database of animal RBR proteins may access Supplementary Files 1 and 2, which, respectively, contain the alignment and the NJ phylogenetic tree (in MEGA 4 format), with all the detailed information for the 530 animal sequences. In our latest analyses (Lucas et al. 2006), we were able to classify all animal sequences then available into nine subfamilies (Ariadne, RNF144, Dorfin, XAP3, Paul, IBRD1, Parkin, ARA54, Triad3; names according to the corresponding human genes), leaving out just a few sequences of *Caenorhabditis* nematodes, which were very divergent. These analyses, with a much larger number of animal sequences, largely confirmed those results: all but 9 of the 530 sequences can be ascribed to 1 of those 10 ensembles (Fig. 1). A single difference is apparent when these new results are compared with those in our previous studies. The Ariadne subfamily appeared as a single group, albeit with no significant bootstrap support, when sequences for all type of organisms were mixed together (see Supplementary Fig. 1 in Lucas et al. 2006). However, in the animal-specific analyses performed here, the Ariadne family is divided into two groups. On one hand, sequences closely related to those of the human genes *Ariadne1*, *Ariadne2*, and *ANKIB1* appear together in the tree, although again the branch that includes them all does not have significant bootstrap support (Fig. 1). On the other hand, the sequences related to a fourth Ariadne gene, called *PARC*, appeared in a separated position in one phylogenetic analysis (NJ, shown in Fig. 1), but together with other Ariadnes in the other two (MP, ML; not shown). In my view, these results do not challenge the hypothesis, put

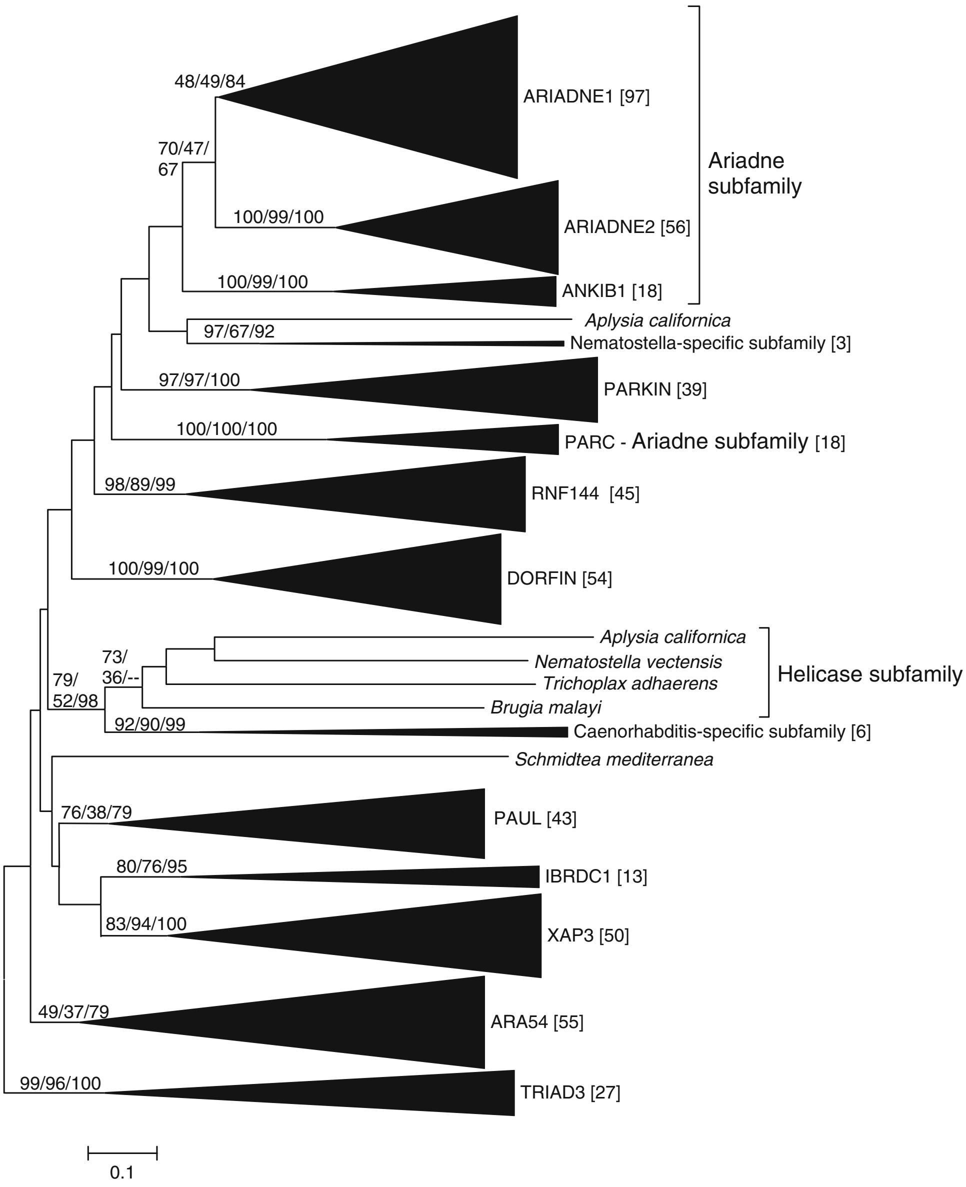


Fig. 1 Phylogenetic tree of RBR sequences. Names refer to both genes (orthology groups) and subfamilies, except for the Ariadne subfamily, which can be divided into four orthology groups (Ariadne1, Ariadne2, ANKIB1, PARC), as indicated. *Numbers in brackets* refer to the number of sequences in each branch. Bootstrap values, in percentages, are shown above the lines, ordered as follows: NJ/MP/ML. MP values tend to be lower than the rest because the MP analyses were not very exhaustive given the large amount of sequences, what makes difficult to detect the optimal trees (see "Methods" section). *Dashes* indicate that a particular branch was not supported by the ML analysis. A full account of this tree, including all species names and accession numbers, can be found in Supplementary Files 1 and 2

forward in Marín and Ferrús (2002), that genes belonging to these four orthology groups must be included into a single subfamily. This is due to the fact, already pointed out in our previous works, that proteins in the four groups contain a large region of similarity beyond the RBR domain. This region, which we called Ariadne domain (Marín and Ferrús, 2002), has not been hitherto found in any protein not belonging to the RBR family. This is a strong indication of a common origin for *Ariadne1*, *Ariadne2*, *ANKIB1*, and *PARC* genes. Ariadne domains have been detected in all new proteins included in these recent analyses, with the exception of some truncated sequences. It is significant to point out that the division of the animal Ariadnes into the four divergent orthology groups shown in Fig. 1 was already described by Lucas et al. (2006), which performed a detailed phylogenetic analysis of this subfamily (See Supplementary Fig. 2 of that paper).

Seven of the other eight subfamilies described in our previous works (Parkin, RNF144, Dorfin, Paul, IBRDC1, XAP3, and Triad3) appear also in these new analyses as significantly supported groups (Fig. 1). The eighth subfamily, ARA54, has a lower degree of support, as occurred in previous works (Marín and Ferrús 2002; Lucas et al. 2006). However, again we can trust that all the genes indeed belong to the same orthology group given that they share a domain called RWD (Marín et al. 2004; Lucas et al. 2006). Finally, the *Caenorhabditis*-specific ensemble detected before also appears in these new analyses as an independent, highly supported group. Given the clear sequence divergence of these nematode genes from the rest of RBR genes, confirmed with this extensive dataset, it seems convenient to classify them as belonging to a new RBR subfamily (*Caenorhabditis*-specific subfamily; Fig. 1).

The nine sequences that cannot be fitted into the 10 subfamilies described above fall into three classes. First, there are two sequences, from the mollusk *Aplysia californica* and the flatworm *Schmidtea mediterranea*, which are highly divergent and cannot be ascribed to any group (Fig. 1). Second, there are three sequences, which appear as a monophyletic, highly supported group, which are all from the cnidarian *N. vectensis*. I will follow the same convention used before for the *Caenorhabditis*-specific sequences and consider that these three sequences constitute a new subfamily (*Nematostella*-specific subfamily, Fig. 1). The simplest interpretation of this finding is that RBR genes have appeared in cnidarians, which are not present in other animals. Finally, and most interestingly, I have detected four genes (from *Trichoplax*, *Nematostella*, *Aplysia*, and the nematode *Brugia malayi*) that must be considered as members of an additional RBR subfamily. Not only these genes have quite closely related RBR signature sequences (Fig. 1), but also they share domains, which are typical of the DEAH/DEAD family of RNA and DNA helicases. This leads to the definition of a Helicase subfamily (Fig. 1). This finding, which has significant implications, is described in detail in the next section.

In total, I found six RBR sequences in the *Trichoplax* genome, which corresponded to orthologs of *Ariadne1*, *parkin*, *dorfin*, *PAUL*, and *ARA54* plus the helicase just mentioned. In *Nematostella*, 15 RBR sequences were detected, corresponding to orthologs of *Ariadne1*, *Ariadne2*, *ANKIB1*, *RNF144*, *dorfin*, *PAUL*, *IBRDC1*, *XAP3*, *ARA54* (two genes), and *Triad3*, plus the helicase-encoding and the three *Nematostella*-specific subfamily genes already mentioned. Structures of all the RBR proteins of *Trichoplax* and *Nematostella* were examined by a combination of TBLASTN and InterProScan searches (see “Methods” section). I confirmed that most of the proteins of those two organisms were structurally identical to those in other animals, a result that reinforces the phylogenetic analyses. This includes finding the following typical domains: (1) Ariadne domains, C-terminally located in the three proteins in *Nematostella* and the single protein in *Trichoplax*, which belong to the Ariadne subfamily; (2) N-terminal Ankyrin repeats in the ortholog of *ANKIB1* in *Nematostella*; (3) Short dorfin domains, C-terminally located in both the *Nematostella* and *Trichoplax dorfin* orthologs. These domains may be truncated, given that it is likely that the reconstructions I obtained for those two proteins were incomplete; (4) Two RanBP2 (“little fing”) fingers in the *Nematostella* ortholog of *PAUL* and a single one in the ortholog of *XAP3* in that same species, all of them located N-terminally with respect to the RBR supradomain; (5) A C-terminal ubiquitin-like domain in the *XAP3* ortholog in *Nematostella*; (6) A C-terminal RWD domain in the *ARA54* orthologs in both *Trichoplax* and *Nematostella*. No additional domains were found in the RNF144, TRIAD3 or IBRDC1 proteins of those two species, as also occurs in their mammalian counterparts. The *Nematostella*-specific subfamily sequences do not contain any additional domain, in spite of that they are 897–924 amino acids long. In summary, all but one of the structures of *Trichoplax* and *Nematostella* proteins were very similar or identical to those found in the orthologous proteins in mammals (see Marín et al. 2004). The only discrepancy with the mammalian sequences was found for the *Trichoplax* ortholog of *parkin*. I could not confirm that the *Trichoplax* parkin protein contains a ubiquitin domain, as occurs in all parkin proteins in other species. However, this may be due to the sequence being incomplete.

Helicase-RBR Proteins Are Ancient and Have Been Lost in Many Lineages

The finding of a few helicase-encoding RBR genes in animals was quite unexpected. In our previous analyses, we already detected RBR genes with helicase domains, but only in plants, leading to the definition of a subfamily called “Plant I” (Marín and Ferrús 2002; Lucas et al. 2006). The fact that only a few animal genes from species recently sequenced turn to encode helicase domains explains why we missed them before. Given the significance of this finding, I have performed specific analyses to detect additional helicase-containing RBR sequences. TBLASTN and BLASTP analyses at NCBI detected the already known plant sequences, the four sequences shown in Fig. 1 and a single additional gene, which, given that the available nucleotide sequences can be translated into protein sequences with multiple stop codons, may correspond to a pseudogene. This gene derives from the ciliate *Plasmodium tetraurelia*. Figure 2 summarizes the current information about the structures of the proteins encoded by all these genes. All of them are very similar. They first have an N-terminal region of variable length without any obvious domains. Then, three domains typical of DEAD/DEAH helicases (known as DEAD, C-terminal helicase domain and HA2) appear. They are separated by about 600–700 amino acids from the RBR supradomain, which is always located in the C-terminus of the protein. Often, other RNA-binding domains (RRM, KH) or the DUF1605 domain, of unknown function, are present between the helicase and RBR domains. It is very unlikely that this peculiar structure has appeared multiple times, so the alternative hypothesis, namely that the patchy phylogenetic range is due to multiple independent losses, must be preferred. If this is correct, we must conclude that the Helicase subfamily originated before animals arose. It would, therefore, be the fourth subfamily, the others being Ariadne, ARA54, and Triad3 (Marín and Ferrús 2002; Lucas et al. 2006), which emerged before the advent of animals.

Patterns of Expansion and Contraction of RBR Subfamilies

We have seen that essentially all animal RBR sequences can be classified into 12 subfamilies and 15 orthology groups (Fig. 1). The large amount of sequences available, together with the large number of species sampled, allows for the first time to establish with a great precision how these subfamilies and orthology groups have appeared and disappeared along the evolution of different animal lineages. Figure 3 summarizes my current hypothesis regarding how RBR genes have evolved in animals. This figure describes the most parsimonious view that accounts for gene emergence and loss, using all the available information. Figure 3 was derived considering whether each of the organisms or groups depicted contained or not each of the orthology groups and then minimizing the number of gains or losses required to explain the pattern observed. The reader may infer this figure, or test the likelihood of alternative hypotheses, by studying Supplementary File 2.

A fundamental conclusion of this work is that most of the diversification of the RBR family occurred very early in animal evolution. Already before the split that separated the placozoan lineage from the rest of animals, we must postulate the presence of at least seven subfamilies. This is due to the facts that the placozoan *T. adhaerens* has members of five subfamilies (Ariadne, ARA54, Parkin, Dorfin, and Paul), that Triad3 genes were already present before the fungi/animals split (Lucas et al. 2006) and that the Helicase subfamily was also present when animals originated (as deduced from the data shown in the previous section). At that time, it is possible that a single Ariadne gene was present: an ortholog of the human gene *Ariadne1* is present in *Trichoplax*. A second RBR expansion most likely occurred after the separation of the placozoans but before the cnidarian/bilaterian split. To explain the diversity of RBR genes currently found in *Nematostella*, we must hypothesize the emergence of other three subfamilies (RNF144, IBRDC1, and XAP) plus the diversification of the Ariadne subfamily, to give rise to the *Ariadne2* and *ANKIB1* genes (Fig. 3). In summary, we find that, quite notably, *Nematostella* has a RBR family, which is comparable both in number of genes and in diversity to that found in mammals: 15 genes from 10 subfamilies are present in the cnidarian and about the same number of genes but from just 9 subfamilies are found in mammalian species. In summary, both lineages, mammals and cnidarians, have conserved almost all genes present in ancient animals and later incorporated a few additional duplicates.

We may thus conclude that innovation in the RBR family has been very rare after the cnidaria/bilateria split. At most three novelties can be detected (Fig. 3): (1) the few *Nematostella*-specific genes that may correspond to a new subfamily; (2) the appearance of the *Caenorhabditis*-specific subfamily; and (3) the emergence of the complex *PARC* genes (described in detail in Marín and Ferrús 2002; Marín et al. 2004). In fact, even whether the *Caenorhabditis* genes are truly new is unclear. They consistently appear in the phylogenetic trees together with the Helicase subfamily genes (Fig. 1), although they do not encode helicase-containing proteins. In my opinion, the finding of a helicase-containing gene in the closely related nematode *Brugia malayi* might be interpreted as the *Caenorhabditis*

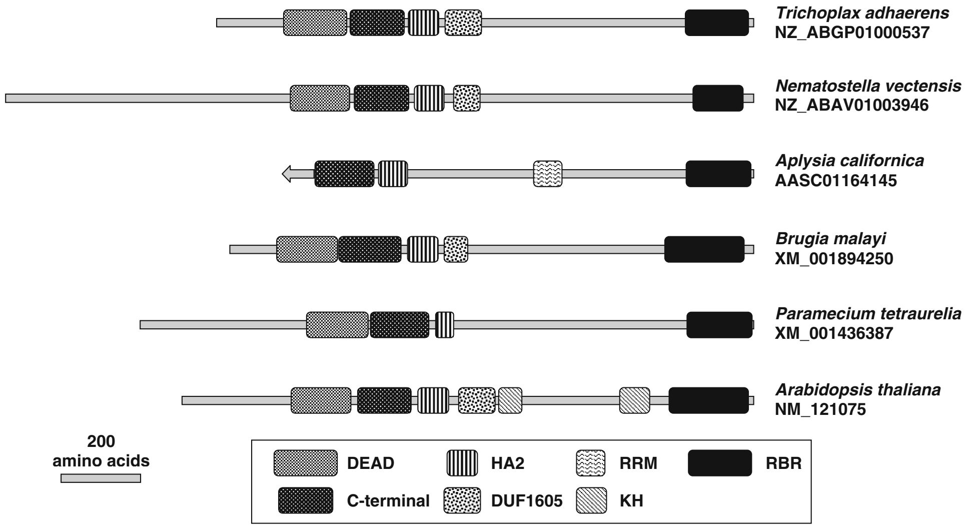
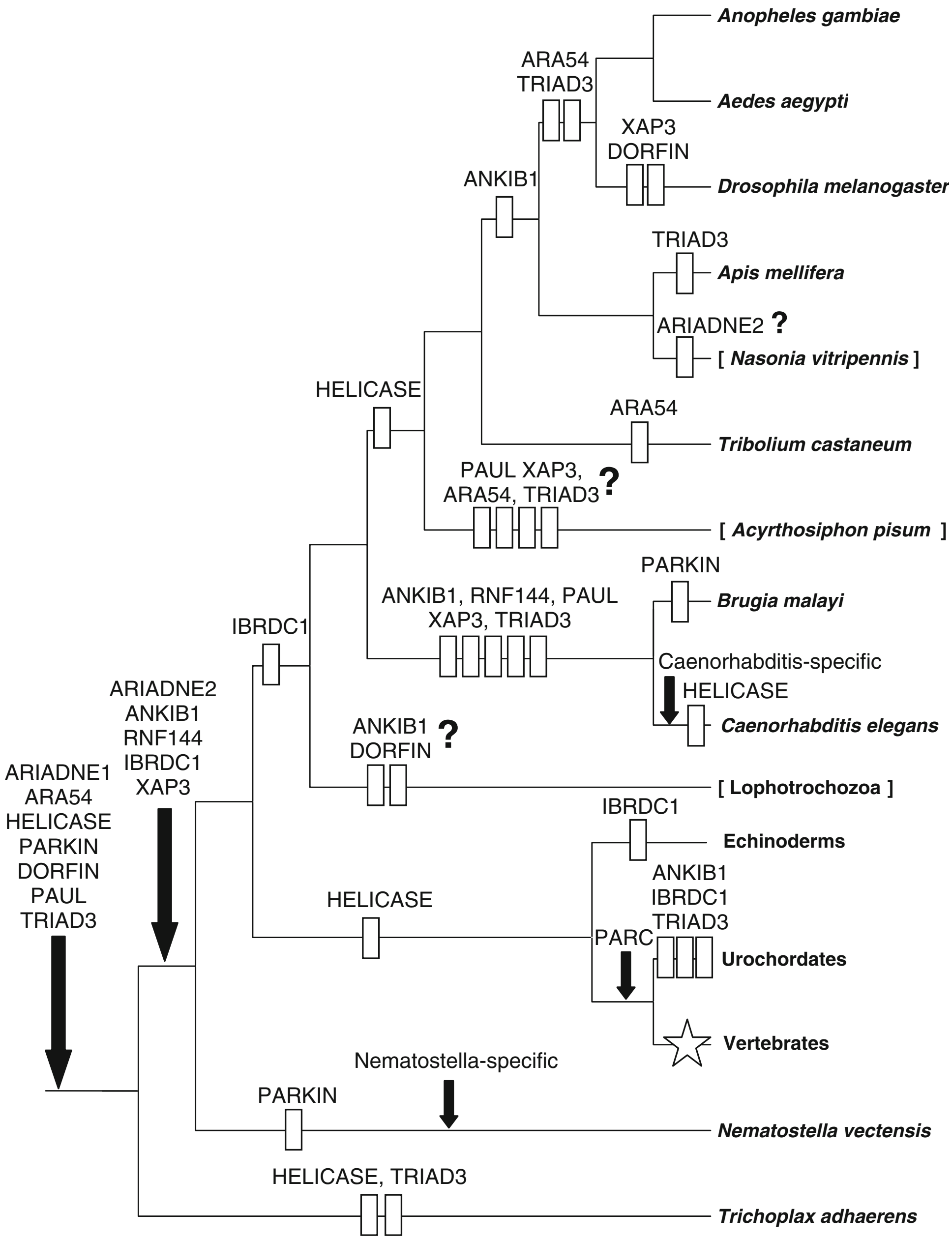


Fig. 2 Structures of helicase-containing RBR proteins. The *arrow* in the *Aplysia* sequence indicates that it has not been possible to establish its N-terminal end

Fig. 3 Diversification and simplification of the RBR family. The simplest hypothesis that explains all available data is summarized. I have assumed a basal phylogenetic position for placozoans, as described by Srivastava et al. (2008), and, for protostomes, the ecdysozoa/lophotrochozoa split. *Arrows* indicate appearances and *rectangles* indicate losses. The star indicates multiple appearances and losses associated to the genome duplications in vertebrates. *Question marks* have been added to indicate that results for branches for which there is no complete genomic data (in which the names are indicated *in brackets*) are still provisional



genes being truncated derivates of Helicase subfamily genes. If this is true, the number of innovations after the cnidaria/bilateria split would be even further limited.

The second major result deduced from these data is that losses of RBR genes are frequent, and are especially concentrated in some lineages. Particularly striking are the extreme decrease of the diversity of RBR genes in nematodes and some insects. Within insects, there is a progressive disappearance of RBR subfamilies, with a minimum number in dipterans (Fig. 3). We must conclude that model species such as *D. melanogaster* (which has 6 RBR genes, belonging to 4 subfamilies), *C. elegans* (11 genes, but from just 5 subfamilies), or *Brugia malayi* (just 5 genes, also from 5 subfamilies) have an RBR set of genes, which is much less diverse than that found in *Nematostella*, a much simpler organism. It is significant that before the insect/nematode split at least 11 genes from 9 subfamilies were present and that some species, such as the insect *Tribolium castaneum*, still retain much of that diversity (9 genes, from 7 subfamilies) (Fig. 3).

It will be interesting to compare these data from ecdysozoans with results from the lophotrochozoans, for which there is still a great paucity of information. Although most of the genes have been found in at least one lophotrochozoan species, the largest set of RBR genes detected so far in a single species is just four, in the annelid *Helobdella* *robusta* (where I have detected orthologs of *Ariadne1*, *RNF144*, *XAP3*, and *TRIAD3*) and the platyhelminth *Schmidtea mediterranea* (which contains a single ortholog of *Ariadne1*, two *XAP3* genes, and the extra, unassigned gene shown in Fig. 1). This means that lophotrochozoan lineages may have suffered losses of RBR genes comparable to those found so far in ecdysozoans. In fact, substantial losses in particular lineages may be a general feature of the evolution of the RBR family not only in protostomes but also in deuterostomes: the urochordate *Ciona intestinalis* (eight genes, from six subfamilies) has lost at least three genes since its lineage separated from the one that gave rise to vertebrates (Fig. 3). Finally, it is also noteworthy that the two rounds of genome duplication that likely occurred in the vertebrate lineage did not generate a significant increase in RBR genes. Only two pairs of vertebrate-specific duplicates (*RNF144/IBRDC2* and *Dorfin/IBRDC3*) can be detected. Therefore, vertebrates have suffered massive losses of RBR genes, although conserving at least a member of each of the original orthology groups.

A final result that deserves attention is that recent duplications of particular RBR genes have occurred in particular lineages, often in those in which a loss of other RBR genes had previously occurred. Thus, *C. elegans* contains four *Ariadne1* genes, two *Dorfin* genes, and two *ARA54* genes and *D. melanogaster* has also two *Ariadne1* genes. For some other species, duplications are even more extensive (a caveat is however that functional data are not available, and therefore some sequences may correspond to non-active pseudogenes). I have detected that larger multiplications seem to have occurred in the wasp *Nasonia vitripennis* (four *Ariadne1* genes), the mosquitoes *Aedes aegypti* (five *XAP3* genes), and *Culex quinquefasciatus* (seven *XAP3* genes) and even more strikingly in the genome of *Caenorhabditis brenneri*, in which no less than 17 different *Ariadne1* signatures can be found.

Long-Term Conservation of Particular RBR Genes and Functional Correlates

From Fig. 3, it can be also deduced the likelihood of each particular gene to be lost in the examined lineages. A simple inspection makes clear than genes have different degrees of conservation. Especially, genes belonging to five orthology groups (Ariadne1, Ariadne2, dorfin, Paul, and RNF144) are generally conserved. They have been confirmed lost in at most one of the lineages shown in Fig. 3. I found interesting to investigate whether conservation may be related to having a broad pattern of expression. Therefore, I explored microarray data from multiple tissues or cell types in humans and flies to determine how broadly expressed are their genes (see "Methods" section for the details). In human microarrays obtained from SymAtlas and corresponding to data obtained from 79 different tissues and cell types, significant *Ariadne1*, *Ariadne2*, *PAUL*, and *RNF144* expression was found in the 99.3%, 100%, 97.5%, and 68.4% of the cases, respectively. The situation is a bit more complicated for *dorfin* genes, given that, as I already indicated above, humans, as other vertebrates, contain two of these genes, *dorfin* and *IBRDC3*. I found that *dorfin* is very restrictively expressed (14.6% of the samples), but *IBRDC3* expression was again detected in 100% of the cases. Thus, the ancestral *dorfin* gene, before the *dorfin/IBRDC3* duplication, might have been also broadly expressed, as *IBRDC3* is still now. In summary, the RBR genes that tend to be evolutionary conserved are broadly expressed in humans (average = 91.3% of the tissues or cell types excluding *IBRDC3* data; 93.0%, including them). In SymAtlas, there can be found also expression data for other six RBR genes: *TRIAD3* (98.1% of the tissues showed positive expression), *XAP3* (63.7%), *ARA54* (61.4%), *PARC* (56.3%), *Parkin* (44.9%), and *IBRDC1* (41.8%). The average for these six less conserved genes is 61.0%.

In flies, data from FlyAtlas (see "Methods" section) provided information for expression in 14 different tissues. The bona-fide orthologs of *Ariadne1* (called *ariadne-1* in flies), *Ariadne2* (*ariadne-2*), *RNF144* (*CG33144*), and *parkin* (also called *parkin*) were found expressed in all tissues. For *ariadne-1*, this was first described by Aguilera et al. (2000). On the other hand, the *PAUL* gene in the fly (*CG11321*) was detected in head, crop, hindgut, male accessory glands, and in the whole adult carcass. A duplicate of *ariadne-1*, *CG12362*, which is characterized by having a quite divergent sequence (Marín and Ferrús 2002 and this study), was found to be expressed in testes and, at lower levels, in larval fat body. Thus, in adult flies, three of the four evolutionary highly conserved genes are broadly expressed.

A second aspect that may be significant to understand the relative susceptibility of different genes to loss is the effects of mutations affecting them. There is a limited amount of information regarding the phenotypic effects of mutations in RBR genes. In mammals, such information is restricted to *parkin*. Loss of function mutations in the *parkin* gene lead to just subtle anomalies in mice and human, in this last case leading, a considerable time after birth, to Parkinson disease (Kitada et al. 1998; Goldberg et al. 2003; Itier et al. 2003; Palacino et al. 2004; Perez and Palmiter 2005). Data for model invertebrates is more abundant. In the fly, knockout mutants in *ariadne-1* are semilethal and in *ariadne-2*, fully lethal (Aguilera et al. 2000). This difference may be due to the presence of *CG12362*, a recent duplicate of *ariadne-1* (Marín and Ferrús 2002). On the other hand, *parkin* mutants are viable, although show some phenotype changes, caused by a general mitochondrial dysfunction (Greene et al. 2003; Pesah et al. 2004; Clark et al. 2006; Park et al. 2006; Yang et al. 2006). For the other genes, there are no loss-of-function mutants described. However, down-regulation of the expression of the ariadne-1 duplicate *CG12362* using RNA interference (RNAi) led also to lethality, while RNAi experiments with the other two RBR genes (*CG11321* and *CG33144*) resulted in viable flies. In the nematode, deletions affecting either one of the *Ariadne-1* genes present, *C27A12.8* (also called *ari-1*; Qiu and Fay 2006) or the gene of the *Caenorhabditis*-specific subfamily, *Y57A10A.31*, are lethal. However, loss-of-function mutations in *parkin* (a. k. a. *K08E3.7*, *prd-1*; Springer et al. 2005) and deletions affecting another *Ariadne-1* gene, *Y73F8A.34*, are viable. RNAi experiments have been performed for all *C. elegans* RBR genes, but only one of them, the ARA54 subfamily gene *F56D2.2* showed obvious phenotypes (slow growth, larval arrest, sterility) under the experimental conditions tested. The rest were apparently normal.

Discussion

Comparative genomics results indicate that there were at least three RBR genes, belonging to the Ariadne, ARA54, and Helicase subfamilies, before the split between unikonts and bikonts that may be at the basis of the eukaryotic tree. Later, the Triad3 subfamily, which is present in both fungi and animals, emerged (Marín and Ferrús 2002; Lucas et al. 2006; this study). The results presented in this study demonstrate that many new animal-specific genes emerged very early in animal evolution. If the hypothesis summarized in Fig. 3 is correct, eight novel RBR genes emerged before the cnidarian/bilaterian split. After that, the story reverses, with losses of genes occurring in many lineages and only a few new genes appearing. With the available data, cnidarians and mammals are the only groups that conserve a set of RBR genes that is similar to that of their common ancestors, while other lineages, both protostomates and deuterostomates, have lost several or even most of those genes, a result observed before for other gene families (e.g., Kusserow et al. 2005; Marín 2008) This streamlining process was especially extreme in nematodes and dipteran insects. However, after this occurred, some of the organisms with less RBR genes have generated some or even many new genes, duplicates of the few that were still left, thus secondarily increasing the number of RBR genes available.

Those expansions and contractions are difficult to explain with the available data. In my opinion, it will be possible to formulate a reasonable hypothesis of why nematodes or dipterans have lost so many RBR genes only when enough comparative data about the ubiquitination systems in multiple organisms becomes available. This is a promising line of research. On the other hand, a simple hypothesis to explain the differences in loss propensity for different genes is possible. RBR genes may be preferentially conserved if they have broad patterns of expression and if their mutations strongly affect fitness (as previously suggested for other genes, based on general analyses, by Krylov et al. 2003). Although the functional results obtained so far are very fragmentary, genes like *Ariadne-1* and *Ariadne-2*, prototypes of the most conservative genes, fit that pattern. First, they are present in all species with completely sequenced genomes, and I have found them also in all the genomes for which incomplete but extensive data exist, with the exception of the wasp *Nasonia vitripennis*, in which genome no *Ariadne-2* gene has been so far found. Second, they are expressed in most/all tissues tested in both mammals and flies, suggesting that this broad pattern is ancient and may be present in all animals. Finally, the three *Ariadne-1* and *Ariadne-2* genes that exist in *Drosophila* flies and at least one in the nematode *C. elegans* are lethal or semilethal when mutated. This hypothesis may be easily tested when more functional results of RBR genes in different species are obtained. If supported by additional data, it may allow to functionally dividing RBR genes into two classes, one of them corresponding to vital, perhaps housekeeping genes, and the other including more specialized, lineage-specific genes. This functional division may have significant biomedical relevance. The secondary duplications found in some species, I think may contribute to ameliorate the impact of having few RBR genes: generating some new ones may be easier than recruiting any of the few RBR genes already present to perform new roles. It is relevant here to point out that, although the particular genes present may be different, no species for which we have full-genome sequences has less than five different RBR genes. From placozoans to mammals, this seems the minimal number required to sustain animal life.

Diversity of the RBR genes in animals seems to be restricted to about 11–12 different subfamilies (see "Results" section). Since our last analysis of the family (Lucas et al. 2006), there have been just a few novel findings, the most important being the detection of helicase-containing RBR proteins in a few species, which has been already commented upon in detail above. However, it would not be surprising that some new diversity could be detected in lophotrochozoans, which are still barely explored. In fact, although most of the sequences detected so far can be ascribed to the already known subfamilies, the fact that two sequences from lophotrochozoans are the only ones that do not fit well in any of the subfamilies suggest a few more of them may be found when this large group of protostomates is examined in detail. Extensive results from sponges, for which only a few partial RBR sequences have been detected so far (none of them complete enough as to be included here), will be also very interesting. In any case, although the details of the evolution of the RBR family in animals may be further clarified, the results shown here most likely characterize all the main features of RBR diversification in metazoan species.

Acknowledgments Research supported by grant 200720I021 (Proyectos intramurales especiales, CSIC, Spain) and grant BIO2008-05067 (Programa Nacional de Biotecnología; Ministerio de Ciencia e Innovación, Spain).

References

Aguilera M, Oliveros M, Martínez-Padrón M, Barbas JA, Ferrús A (2000) *Ariadne-1*: a vital *Drosophila* gene is required for development and defines a new conserved family of RING-finger proteins. Genetics 155:1231–1244

Beasley SA, Hristova VA, Shaw GS (2007) Structure of the Parkin in-between-ring domain provides insights for E3-ligase dysfunction in autosomal recessive Parkinson's disease. Proc Natl Acad Sci USA 104:3095–3100

Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, Chan J, Chen N, Chen WJ, Davis P, Fiedler TJ, Girard L, Han M, Harris TW, Kishore R, Lee R, McKay S, Müller HM, Nakamura C, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Spooner W, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Durbin R, Stein LD, Sternberg PW, Spieth J (2007) WormBase: new content and better access. Nucleic Acids Res 35(Database issue):D506–D510

Capili AD, Edghill EL, Wu K, Borden KLB (2004) Structure of the C-terminal RING finger from a RING-IBR-RING/TRIAD motif reveals a novel zinc-binding domain distinct from a RING. J Mol Biol 340:1117–1129

Chintapalli VR, Wang J, Dow JAT (2007) Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet 39:715–730

Clark IE, Dodson MW, Jiang C, Cao JH, Huh JR, Seol JH, Yoo SJ, Hay BA, Guo M (2006) *Drosophila pink1* is required for mitochondrial function and interacts genetically with *parkin*. Nature 441:1162–1166

Dietzl G, Chen D, Schnorrer F, Su KC, Barinova Y, Fellner M, Gasser B, Kinsey K, Oppel S, Scheiblauer S, Couto A, Marra V, Keleman K, Dickson BJ (2007) A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. Nature 448:151–156

Eisenhaber B, Chumak N, Eisenhaber F, Hauser MT (2007) The ring between ring fingers (RBR) protein family. Genome Biol 8:209

Glickman MH, Ciechanover A (2002) The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. Physiol Rev 82:373–428

Goldberg MS, Fleming SM, Palacino JJ, Cepeda C, Lam HA, Bhatnagar A, Meloni EG, Wu N, Ackerson LC, Klapstein GJ, Gajendiran M, Roth BL, Chesselet MF, Maidment NT, Levine MS, Shen J (2003) *Parkin*-deficient mice exhibit nigrostriatal deficits but not loss of dopaminergic neurons. J Biol Chem 278:43628–43635

Greene JC, Whitworth AJ, Kuo I, Andrews LA, Feany MB, Pallanck LJ (2003) Mitochondrial pathology and apoptotic muscle degeneration in *Drosophila parkin* mutants. Proc Natl Acad Sci USA 100:4078–4083

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

Ikeda F, Dikic I (2008) Atypical ubiquitin chains: new molecular signals. EMBO Rep 9:536–542

Itier JM, Ibanez P, Mena MA, Abbas N, Cohen-Salmon C, Bohme GA, Laville M, Pratt J, Corti O, Pradier L, Ret G, Joubert C, Periquet M, Araujo F, Negroni J, Casarejos MJ, Canals S, Solano R, Serrano A, Gallego E, Sanchez M, Denefle P, Benavides J, Tremp G, Rooney TA, Brice A, Garcia de Yebenes J (2003) *Parkin* gene inactivation alters behaviour and dopamine neurotransmission in the mouse. Hum Mol Genet 12:2277–2291

Kerscher O, Felberbaum R, Hochstrasser M (2006) Modification of proteins by ubiquitin and ubiquitin-like proteins. Annu Rev Cell Dev Biol 22:159–180

Kitada T, Asakawa S, Hattori N, Matsumine H, Yamamura Y, Minoshima S, Yokochi M, Mizuno Y, Shimizu N (1998) Mutations in the *parkin* gene cause autosomal recessive juvenile parkinsonism. Nature 392:605–608

Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res 13:222–2235

Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, Technau U, von Haeseler A, Hobmayer B, Martindale MQ, Holstein TW (2005) Unexpected complexity of the *Wnt* gene family in a sea anemone. Nature 433:156–160

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Lucas JI, Marín I (2007) A new evolutionary paradigm for the Parkinson disease gene *DJ-1*. Mol Biol Evol 24:551–561

Lucas JI, Arnau V, Marín I (2006) Comparative genomics and protein domain graph analyses link ubiquitination and RNA metabolism. J Mol Biol 357:9–17

Marín I (2008) Ancient origin of the Parkinson disease gene *LRRK2*. J Mol Evol 67:41–50

Marín I, Ferrús A (2002) Comparative genomics of the RBR family, including the Parkinson's disease-related gene *parkin* and the genes of the ariadne subfamily. Mol Biol Evol 19:2039–2050

Marín I, Lucas JI, Gradilla AC, Ferrús A (2004) Parkin and relatives: the RBR family of ubiquitin ligases. Physiol Genomics 17:253–263

Mukhopadhyay D, Riezman H (2007) Proteasome-independent functions of ubiquitin in endocytosis and signaling. Science 315:201–205

Nicholas KB, Nicholas HB Jr (1997) GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author

Palacino JJ, Sagi D, Goldberg MS, Krauss S, Motz C, Wacker M, Klose J, Shen J (2004) Mitochondrial dysfunction and oxidative damage in parkin-deficient mice. J Biol Chem 279:18614–18622

Park J, Lee SB, Lee S, Kim Y, Song S, Kim S, Bae E, Kim J, Shong M, Kim JM, Chung J (2006) Mitochondrial dysfunction in *Drosophila PINK1* mutants is complemented by *parkin*. Nature 441:1157–1161

Perez FA, Palmiter RD (2005) *Parkin*-deficient mice are not a robust model of parkinsonism. Proc Natl Acad Sci USA 102:2174–2179

Pesah Y, Pham T, Burgess H, Middlebrooks B, Verstreken P, Zhou Y, Harding M, Bellen H, Mardon G (2004) *Drosophila parkin* mutants have decreased mass and cell size and increased sensitivity to oxygen radical stress. Development 131:2183–2194

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94

Qiu X, Fay DS (2006) ARI-1, an RBR family ubiquitin-ligase, functions with UBC-18 to regulate pharyngeal development in *C. elegans*. Dev Biol 291:239–252

Schwartz AL, Ciechanover A (2009) Targeting protein for destruction by the ubiquitin system: implications for human pathobiology. Annu Rev Pharmacol Toxicol 49:73–96

Springer W, Hoppe T, Schmidt E, Baumeister R (2005) A *Caenorhabditis elegans Parkin* mutant with altered solubility couples alpha-synuclein aggregation to proteotoxic stress. Hum Mol Genet 14:3407–3423

Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss LW, Schierwater B, Dellaporta SL, Rokhsar DS (2008) The *Trichoplax* genome and the nature of placozoans. Nature 454:955–960

Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB (2002) Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA 99:4465–4470

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101:6062–6067

Swofford DL (2003) PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sinauer Associates, Sunderland, MA

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Weake VM, Workman JL (2008) Histone ubiquitination: triggering gene activity. Mol Cell 29:653–663

Welchman RL, Gordon C, Mayer RJ (2005) Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol 6:599–609

Wilson RJ, Goodman JL, Strelets VB, FlyBase Consortium (2008) FlyBase: integration and improvements to query tools. Nucleic Acids Res 36(Database issue):D588–D593

Yang Y, Gehrke S, Imai Y, Huang Z, Ouyang Y, Wang JW, Yang L, Beal MF, Vogel H, Lu B (2006) Mitochondrial pathology and muscle and dopaminergic neuron degeneration caused by inactivation of *Drosophila Pink1* is rescued by Parkin. Proc Natl Acad Sci USA 103:10793–10798

Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848