Evolution of RNA-Binding Proteins in Animals: Insights from Genome-Wide Analysis in the Sponge Amphimedon queenslandica

Pierre Kerner,1 Sandie M. Degnan,2 Lauriane Marchand,1 Bernard M. Degnan,2 and Michel Vervoort*,1,3
1Development and Neurobiology Programme, Institut Jacques Monod, Centre national de la recherche scientifique/Université Paris Diderot-Paris 7, Paris cedex, France
2Centre for Marine Science, School of Biological Sciences, University of Queensland, Brisbane, Queensland, Australia
3Institut Universitaire de France, Paris, France
†Present address: Department of Cell and Developmental Biology, Weill Medical College of Cornell University, New York, New York.
*Corresponding author: E-mail: vervoort.michel@ijm.univ-paris-diderot.fr.
Associate editor: Billie Swalla

Abstract
RNA-binding proteins (RBPs) are key players in various biological processes, most notably regulation of gene expression at the posttranscriptional level. Although many RBPs have been carefully studied in model organisms, very few studies have addressed the evolution of these proteins at the scale of the animal kingdom. We identified a large set of putative RBPs encoded by the genome of the demosponge Amphimedon queenslandica, a species representing a basal animal lineage. We compared the Amphimedon RBPs with those encoded by the genomes of two bilaterians (human and Drosophila), representatives of two other basal metazoan lineages (a placozoan and a cnidarian), a choanoflagellate (probable sister group of animals), and two fungi. We established the evolutionary history of 32 families of RBPs and found that most of the diversity of RBPs present in contemporary metazoans, including humans, was already established in the last common ancestor (LCA) of animals. This includes RBPs known to be involved in key processes in bilaterians, such as development, stem and/or germ cells properties, and noncoding RNA pathways. From this analysis, we infer that a complex toolkit of RBPs was present in the LCA of animals and that it has been recruited to perform new functions during early animal evolution, in particular in relation to the acquisition of multicellularity.

Key words: RNA-binding proteins, evolution, ncRNA, miRNA, development, stem cells.

Introduction
RNA-binding proteins (RBPs) are key components in RNA metabolism. They regulate all aspects of posttranscriptional RNA biogenesis, including RNA maturation, nucleocytoplasmic transport, splicing, subcellular localization, translation, and degradation (Dreyfuss et al. 2002; Orphanides and Reinberg 2002; Lasko 2003; Glisovic et al. 2008). This posttranscriptional processing of pre-mRNAs and mRNAs lies at the heart of animal gene regulation, making RBPs important players in many aspects of animal development, such as axis determination, maternal to zygotic transition, control of cell proliferation and differentiation, and nervous system development (Dreyfuss et al. 2002; Orphanides and Reinberg 2002; Lasko 2003; Glisovic et al. 2008). Not surprisingly then, it has recently become apparent that RBPs are also involved in many human diseases, ranging from neurologic disorders to cancer (He and Hannon 2004; Amaral et al. 2008; Ghildiyal and Zamore 2009).

Eukaryotic RBPs are characterized by the presence of RNA-binding domains containing 60–100 residues that are found in single or multiple copies within a protein (Burd and Dreyfuss 1994; Dreyfuss et al. 1994; Dreyfuss et al. 2002; Chen and Varani 2005; Lunde et al. 2007). Many different RNA-binding domains have been identified (Anantharaman et al. 2002); the most common classes are the RNA recognition motif (RRM), found in >50% of RBPs, the heterogeneous nuclear RNP K-homology (KH) domain, the DEAD/DExH-box helicase domain ( DEAD-box), and the double-stranded RNA-binding domain (DsRM) (Lunde et al. 2007). Other less common types of domains used to bind RNA include CCCH and CCHH zinc fingers, PAZ and Piwi domains found in proteins involved in the miRNA pathways, the Sm domain found in the small nuclear ribonucleoprotein-associated (Sm and LSm) proteins and the PUF domain characteristic of the FBF/Pumilio proteins (Anantharaman et al. 2002). The presence of any of these domains in a protein is suggestive of an RNA-binding function and has been used to predict RBPs in
several genomes, such as in *Drosophila* and mouse (Lasko 2000; Keene 2001; McKee et al. 2005; Gambeti et al. 2006; Lee and Schedl 2006). These bioinformatic studies have revealed that eukaryotic genomes encode a large number of putative RBPs: 5–8% of genes encode proteins predicted to function as RBPs in yeast and approximately 2% in *Caenorhabditis elegans, Drosophila melanogaster,* and mouse (Lasko 2000; Keene 2001; McKee et al. 2005; Gambeti et al. 2006; Lee and Schedl 2006).

The number of genome-wide studies of RBPs conducted to date has been too small to permit analysis of the evolution of these proteins, in particular in animals. In this article, we report a large set of RBPs encoded by the completely sequenced genome of the sponge *Amphimedon queenslandica*. Sponges are an ancient group of animals that diverged from other animals over 600 Ma and therefore play a critical role in the understanding of metazoan evolution, in particular for the evolutionary origin of key animal properties, such as multicellularity, developmental processes, immunity, and the acquisition of specific cell types such as neurons (Srivastava et al. 2010). We compared the repertoire of *Amphimedon* RBPs with those encoded by the genomes of two bilaterians (human and *Drosophila*), two other metazoans of basal emergence (a placozoan and a cnidarian), a choanoflagellate (choanoflagellates are unicellular organisms which most likely constitute the sister group of metazoans), and two fungi (fig. 1). We also studied the evolution of several RBPs subfamilies known to be involved in bilaterian development, stem and/or germ cell development and behavior, and ncRNA pathways, by retrieving and phylogenetically analyzing members of these subfamilies from 28 species whose genome is completely sequenced. Our study provides new insights into the evolution of RBPs in animals.

**Materials and Methods**

**Sequences Retrieval**

We used the previously identified RBPs from *Drosophila* and mouse (Lasko 2000; McKee et al. 2005; Gambeti et al. 2006) as queries to undertake similarity searches using the BlastP algorithm (Altschul et al. 1990). Blast searches were made using the KoriBlast platform (Korilog company, Muzillac, France) against local databases for *Amphimedon* and databases hosted by the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) for the other species. Some Blast searches were also done on the Department of Energy (DOE) Joint Genome Institute web site using its Genome Portal Blast tools (http://genome.jgi-psf.org/). All retrieved sequences can be found in supplementary tables S5 and S7, Supplementary Material online.

**Phylogenetic Analyses**

Multiple alignments were performed with MUSCLE 3.7 (Edgar 2004) using the Phylogeny.fr web server (Dereeper et al. 2008) and were subsequently manually improved. Handling of the multiple alignments was done using SEAVIEW (Gouy et al. 2010). Maximum likelihood analyses were performed with PHYML (Guindon and Gascuel 2003) using the PHYML web server (Guindon et al. 2005) hosted at the Montpellier bioinformatics platform (http://www.atgc-montpellier.fr/phyml/). PHYML analyses were performed using the LCA of holozoans, metazoans, and eumetazoans (i.e., bilaterians plus cnidarians) are shown, and their putative complements of RBPs are also shown. Due to the occurrence of many secondary gene losses in *Trichoplax* and the uncertainties about its phylogenetic position, we think that it is meaningless to indicate the RBP complement of the LCA of eumetazoans plus placozoans. The five types of RBPs (RRM, KH, DEAD, DS, and others) are described in the main text. RBP families involved in developmental processes and/or ncRNA pathways that were present in the LCA of opisthokonts are indicated as well as those that were subsequently added and present in the LCA of holozoans, metazoans, and eumetazoans.

![Fig. 1. Summary of the evolution of RBPs in opisthokonts.](image-url)
were launched, each containing from 1,500,000 to 4,000,000 Monte Carlo steps (depending on the number of steps required to get chain convergence). We estimated that convergence was obtained if the “average standard deviation of split frequencies” reached a threshold value of 0.05. One of every 200 trees was saved during the runs. The trees obtained in the two runs were mixed, and the first 25% of the trees were discarded as “burnin.” Marginal probabilities at each internal branch were taken as a measure of statistical support. Phylogenetic trees were visualized and rooted using TreeDyn (Chevenet et al. 2006) hosted by the Phylogeny.fr web site. Multiple alignments are available upon request.

**Results and Discussion**

**Evolution of the Repertoire of RBPs in Animals**

We made Blast searches (Altschul et al. 1990) to identify *Amphimedon* proteins that harbor one or more domains known to bind either single or double-stranded RNA, that is, the RRM domain, the KH domain, the DEAD-box domain, and the DsRM domain (Lunde et al. 2007). We also searched for homologs of other proteins known to bind RNA through other types of domains, such as PUF, zinc fingers, Sm, Piwi, and PAZ domains. We identified 266 putative RBPs in *Amphimedon*, including 121 RRM, 18 KH, 72 DEAD-box, and 10 DsRM proteins (table 1). This is similar to the number of proteins with these domains found in *Drosophila*, but smaller than the number of mouse proteins harboring these domains (Lasko 2000; McKee et al. 2005; Gamberi et al. 2006) (table 1). We used the reciprocal best Blast hit method (Tatusov et al. 1997; Bork et al. 1998) to define the orthologs of all the putative *Amphimedon* RBPs encoded by the *Drosophila* and human genomes. We found that about 80% of the *Amphimedon* RBPs have orthologs of human and/or *Drosophila* proteins (see table 2 for details). A list of all the identified *Amphimedon* RBPs, with their orthologs in *Drosophila* and/or human, can be found in supplementary table S1, Supplementary Material online.

To analyze the early evolution of RBPs in animals, we extended our study to three additional species whose genomes have been completely sequenced, namely the choanoflagellate *Monosiga brevicollis* and two nonbilaterian animals, the placozoan *Trichoplax adhaerens* and the cnidian *Nematostella vectensis* (Putnam et al. 2007; King et al. 2008; Srivastava et al. 2008) (fig. 1). We found 151 RBPs in *Monosiga*, 203 in *Trichoplax*, and 257 in *Nematostella* (table 1). Using the reciprocal best Blast hit approach, we defined the orthologs of all these proteins in both human and *Amphimedon*—all the identified *Monosiga*, *Trichoplax*, and *Nematostella* proteins with their putative orthologs in human and *Amphimedon* can be found in supplementary tables S2–S4, Supplementary Material online. Conversely, we defined the *Monosiga*, *Trichoplax*, and *Nematostella* orthologs of all the *Amphimedon* RBPs—these orthologs are listed in supplementary table S1, Supplementary Material online. We also defined the orthologs of all the *Amphimedon* RBPs in two distantly related species of fungi, the sordariomycetes *Neurospora crassa*, and the taphrinomycotina *Schizosaccharomyces pombe* (supplementary table S1, Supplementary Material online). A summary of these analyses is shown in table 2. Taken together, these data allow several conclusions to be drawn about the evolution of RBPs in animals.

Given the probable phylogenetic position of sponges as a sister group to all other animals (eumetazoans, i.e., the last common ancestor [LCA] of placozoans, cnidarians, and bilaterians as well as all its descendants) (fig. 1) (Philippe et al. 2009; Pick et al. 2010; Srivastava et al. 2010), eumetazoan or bilaterian genes that have orthologs in sponges (and vice versa) derive from ancestral genes already present in the LCA of all extant metazoans. Our data therefore allow us to infer the set of RBPs likely to have been present in this remote ancestor. We must first address the possibility that some of the genes encoding RBPs may have been lost or duplicated in the sponge lineage (or specifically in *Amphimedon*) since the time that it diverged from the eumetazoan lineage. To enhance our detection of lineage-specific putative gene losses, we used the *Monosiga* data set to search for eumetazoan RBPs that have orthologs in *Monosiga* but not in *Amphimedon*—these RBPs would be encoded by ancestral genes secondarily lost in the *Amphimedon* genome. We found seven cases (four RRM, one KH, one DEAD, zero DsRM, and three for RBPs with other RNA-binding domains) of such probable secondary gene loss in *Amphimedon*. To detect lineage-specific gene duplications, we searched for closely related paralogs in *Amphimedon*, that is, *Amphimedon* proteins for which the most similar protein(s) (using Blast statistics, e-value, and score, as criteria) is another (others) *Amphimedon* protein(s). We detected ten cases of duplications for RRM proteins (two with two paralogs, six with three paralogs, one with four paralogs, and one with five paralogs), two cases for KH proteins (with two paralogs), five cases for DEAD-box proteins (one with two paralogs and four with three paralogs), one case for DsRM proteins (with two paralogs), and three cases for RBPs with other RNA-binding domains (with two paralogs).

### Table 1. Number of RBPs in the Species Under Study.

<table>
<thead>
<tr>
<th>Number of Proteins in Species</th>
<th>Number of Proteins in <em>Amphimedon</em></th>
<th>Number of Proteins in <em>Drosophila</em></th>
<th>Number of Proteins in <em>Mus</em></th>
<th>Number of Proteins in <em>Nematostella</em></th>
<th>Number of Proteins in <em>Trichoplax</em></th>
<th>Number of Proteins in <em>Monosiga</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>RRM</td>
<td>121</td>
<td>117</td>
<td>227</td>
<td>125</td>
<td>85</td>
<td>75</td>
</tr>
<tr>
<td>KH</td>
<td>18</td>
<td>27</td>
<td>42</td>
<td>20</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>DEAD box</td>
<td>72</td>
<td>63</td>
<td>&gt;100</td>
<td>54</td>
<td>60</td>
<td>38</td>
</tr>
<tr>
<td>DsRM</td>
<td>10</td>
<td>8</td>
<td>21</td>
<td>13</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Others</td>
<td>45</td>
<td>41</td>
<td>&gt;100</td>
<td>45</td>
<td>41</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>266</td>
<td>256</td>
<td>&gt;400</td>
<td>257</td>
<td>203</td>
<td>152</td>
</tr>
</tbody>
</table>
Taking into account the number of conserved RBPs between eumetazoans and *Amphimedon* as well as the specific gene losses and gene duplications that occurred in *Amphimedon*, we can estimate that the metazoan LCA possessed at least 88 different RRM proteins, 15 KH proteins, 49 DEAD-box proteins, 9 DsRM, and 39 proteins with other RNA-binding domains (fig. 1). Using the same reasoning, we can also infer the set of RBPs that were present in the LCA of holozoans (choanoflagellates plus metazoans)—in this case, we used our data on fungi to detect putative secondary gene loss in *Monosiga* (11 RRM, 0 KH, 4 DEAD, 2 DsRM, and 12 for RBPs with other RNA-binding domains), and our search for closely related paralogs did not detect any *Monosiga*-specific duplications. We therefore infer that the LCA of holozoans had at least 69 RRM proteins, 9 KH proteins, 35 DEAD-box proteins, 4 DsRM proteins, and 38 proteins with other RNA-binding domains (fig. 1). We also infer that the LCA of opisthokonts (fungi plus holozoans) had at least 51 RRM proteins, 6 KH proteins, 29 DEAD-box proteins, 3 DsRM proteins, and 29 proteins with other RNA-binding domains (fig. 1). These are likely underestimations of the true numbers of ancestral RBPs, as we only searched for specific domains or known RBPs and we probably underestimated gene losses in *Amphimedon* and *Monosiga*. We also searched for eumetazoan-specific RBPs, that is, RBPs present in *Trichoplax* or *Nematostella* and human but with orthologs in neither *Amphimedon* nor *Monosiga*. We only found seven RRM proteins, one KH protein, four DEAD-box protein, zero DsRM protein, and one protein with another RNA-binding domain that appear to be eumetazoan specific, suggesting that very few new RBP types emerged in the eumetazoan lineages. The very high number of RBPs in human and mouse seems to be due to further duplications in the vertebrate lineage of most of the ancestral RBP-encoding genes (not observed)—this is observed for most gene families and is probably a consequence of whole-genome duplications that occurred during early vertebrate evolution (Panopoulou and Poustka 2005).

We can therefore conclude that 1) the LCA of all metazoans had a very complex repertoire of RBPs, 2) the vast majority of the RBP types found in present-day metazoans were already present in the metazoan LCA, 3) a significant proportion of these RBPs were already present in the LCA of holozoans and opisthokonts, and 4) the repertoire of RBPs has been remarkably stable during metazoan evolution. The diversification of RBPs seems therefore to have mainly occurred before the divergence of the main metazoan lineages, followed by a very minor expansion in the eumetazoan lineage. This is in contrast to what has been observed for transcription factors: although most of the main transcription factor types (e.g., different types of homeodomain, basic-Helix-Loop-Helix, Sox, LIM homeodomain) are found in *Amphimedon*, they are significantly less numerous than in bilaterians or eumetazoans, suggesting an important expansion in the eumetazoan lineage after its divergence from sponges (Simionato et al. 2007; Larroux et al. 2008; Degnan et al. 2009; Srivastava et al. 2010). One plausible explanation for this difference is that RBPs are involved in fundamental molecular processes in eukaryotes, such as RNA splicing and translation, and therefore can be expected to be of very ancient origin. However, a lot of RBPs have been shown to have much more specific functions in bilaterians, for example, during development including neurogenesis or in stem cell biology and small RNA pathways. Interestingly, we found orthologs for many of these RBPs in *Amphimedon* (some of them are also in *Monosiga* and/or fungi; see next sections),

### Table 2. Summary of the Comparative Analysis of the RBP Repertoire in Different Opisthokont Species.

<table>
<thead>
<tr>
<th></th>
<th>Number of Proteins in <em>Amphimedon</em></th>
<th>% of the Amphimedon Proteins With Orthologs in Human or <em>Drosophila</em></th>
<th>% of the Amphimedon Proteins With Orthologs in <em>Nematostella</em></th>
<th>% of the Amphimedon Proteins With Orthologs in <em>Trichoplax</em></th>
<th>% of the Amphimedon Proteins With Orthologs in <em>Schizosaccharomyces pombe</em> or <em>Neurospora crassa</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>RRM</td>
<td>121</td>
<td>78</td>
<td>64</td>
<td>56</td>
<td>41</td>
</tr>
<tr>
<td>KH</td>
<td>18</td>
<td>89</td>
<td>83</td>
<td>72</td>
<td>39</td>
</tr>
<tr>
<td>DEAD box</td>
<td>72</td>
<td>74</td>
<td>57</td>
<td>44</td>
<td>46</td>
</tr>
<tr>
<td>DsRM</td>
<td>10</td>
<td>100</td>
<td>60</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Others</td>
<td>45</td>
<td>85</td>
<td>74</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>266</td>
<td>79</td>
<td>65</td>
<td>54</td>
<td>41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>% of the Nematostella Proteins With Orthologs in Human</th>
<th>% of the Nematostella Proteins With Orthologs in <em>Amphimedon</em></th>
<th>% of the Trichoplax Proteins With Orthologs in Human</th>
<th>% of the Monosiga Proteins With Orthologs in Human</th>
<th>% of the Monosiga Proteins With Orthologs in <em>Amphimedon</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>RRM</td>
<td>69</td>
<td>64</td>
<td>96</td>
<td>87</td>
<td>77</td>
</tr>
<tr>
<td>KH</td>
<td>85</td>
<td>80</td>
<td>93</td>
<td>53</td>
<td>100</td>
</tr>
<tr>
<td>DEAD box</td>
<td>81</td>
<td>76</td>
<td>82</td>
<td>68</td>
<td>83</td>
</tr>
<tr>
<td>DsRM</td>
<td>85</td>
<td>70</td>
<td>100</td>
<td>100</td>
<td>67</td>
</tr>
<tr>
<td>Others</td>
<td>80</td>
<td>75</td>
<td>98</td>
<td>95</td>
<td>96</td>
</tr>
<tr>
<td>Total</td>
<td>75</td>
<td>70</td>
<td>92</td>
<td>81</td>
<td>83</td>
</tr>
</tbody>
</table>
### Table 3. Summary of the Expressions and Functions of the 32 Families of RBPs That Have Been Studied at the Phylogenetic Level.

<table>
<thead>
<tr>
<th>Gene Family</th>
<th>Domain(s)</th>
<th>Functions/Expressions</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>elav/Hu</td>
<td>RRM</td>
<td>Differentiation and plasticity of neurons in bilaterians; expression in neuronal cells in the cnidarian <em>Nematostella</em></td>
<td>Deschénes-Furry et al. 2006; Marlow et al. 2009; Pascale et al. 2008</td>
</tr>
<tr>
<td>msi</td>
<td>RRM</td>
<td>Expression in the nervous system in neural progenitors and/or neural stem cells in bilaterians; expression in neuronal cells in the cnidarian <em>Nematostella</em></td>
<td>Okano et al. 2005; Marlow et al. 2009</td>
</tr>
<tr>
<td>Dazap1</td>
<td>RRM</td>
<td>Expression in the gonads in mammals</td>
<td>Tsui et al. 2000</td>
</tr>
<tr>
<td>tra-2</td>
<td>RRM</td>
<td>Diversification of vascular smooth muscles during mammalian development; in <em>Drosophila</em>, acquisition of sex-specific properties during development</td>
<td>Heinrichs et al. 1998; Shukla and Fisher 2008</td>
</tr>
<tr>
<td>spen</td>
<td>RRM</td>
<td>Various developmental processes; participate in WNT signaling both in <em>Drosophila</em> and vertebrates</td>
<td>Wiellette et al. 1999; Feng et al. 2007; Chang et al. 2008</td>
</tr>
<tr>
<td>spenito/RBM15</td>
<td>RRM</td>
<td>Wide expression in embryonic and adult tissues in vertebrates; in <em>Drosophila</em>, gametogenesis, anterior–posterior axis and germ cell formation; stem cell maintenance in the flatworm <em>Schmidtea</em></td>
<td>Barreau et al. 2006; Chekulaeva et al. 2006; Guo et al. 2006</td>
</tr>
<tr>
<td>CELF</td>
<td>RRM</td>
<td>Germ line development and control of anterior-posterior axis determination in the early embryo of the nematode <em>Caenorhabditis</em> and the insect <em>Tribolium</em></td>
<td>Draper et al. 1996; Ciosk et al. 2006; Schoppmeier et al. 2009</td>
</tr>
<tr>
<td>FMR1</td>
<td>KH</td>
<td>Various aspects of development; interact with components of the miRNA pathway</td>
<td>Caudy et al. 2002; Li et al. 2008; Lukong et al. 2008</td>
</tr>
<tr>
<td>BicC</td>
<td>KH</td>
<td>Kidney development and left–right asymmetry determination in vertebrates; patterning of the <em>Drosophila</em> embryo</td>
<td>Chicoine et al. 2007; Tran et al. 2007; Maisonneuve et al. 2009</td>
</tr>
<tr>
<td>Mex-3</td>
<td>KH</td>
<td>Germ line development and control of anterior-posterior axis determination in the early embryo of the nematode <em>Caenorhabditis</em> and the insect <em>Tribolium</em></td>
<td>Draper et al. 1996; Ciosk et al. 2006; Schoppmeier et al. 2009</td>
</tr>
<tr>
<td>vasa</td>
<td>DEAD box</td>
<td>Expression in somatic and germinal stem cells in many animals including nonbilaterians; role in germ cell development in vertebrates and <em>Drosophila</em></td>
<td>Extavour et al. 2005; Rebscher et al. 2007; Gustafson and Wessel 2010; Kranz et al. 2010; Alié et al. 2010</td>
</tr>
<tr>
<td>Pl10/DDX3</td>
<td>DEAD box</td>
<td>Expression in somatic and germinal stem cells in many animals including nonbilaterians; role in germ cell development in vertebrates and <em>Drosophila</em></td>
<td>Extavour et al. 2005; Rebscher et al. 2007; Gustafson and Wessel 2010; Kranz et al. 2010; Alié et al. 2010</td>
</tr>
<tr>
<td>armitage/Mov10L</td>
<td>DEAD box</td>
<td>Association to RISC, role for effective miRNA-mediated silencing</td>
<td>Tomari et al. 2004; Chendrimada et al. 2007</td>
</tr>
<tr>
<td>Mov10</td>
<td>DEAD box</td>
<td>miRNA-mediated silencing</td>
<td>Tomari et al. 2004; Chendrimada et al. 2007</td>
</tr>
<tr>
<td>drosha/RNASEN</td>
<td>DsRM</td>
<td>Constitute the «Microprocessor complex» that control the first step in the biogenesis of functional miRNA, the cleavage in the nucleus of “pri-miRNA” (around 1,000 nt-long) into 60–70 nt-long “pre-miRNA”</td>
<td>He and Hannon 2004; Denli et al. 2004; Hammond 2005</td>
</tr>
<tr>
<td>Pasha/DiGeorge syndrome critical region gene 8 (DGCR8)</td>
<td>DsRM</td>
<td>Formation of mature miRNA though the cleavage in the cytoplasm of “pre-miRNA” into imperfect double-stranded RNA duplex</td>
<td>He and Hannon 2004; Hammond 2005</td>
</tr>
<tr>
<td>dicer</td>
<td>DsRM + DEAD box</td>
<td>Formation of mature miRNA though the cleavage in the cytoplasm of “pre-miRNA” into imperfect double-stranded RNA duplex</td>
<td>He and Hannon 2004; Hammond 2005</td>
</tr>
<tr>
<td>Staufen</td>
<td>DsRM</td>
<td>RNA transport along microtubules</td>
<td>Miki et al. 2005</td>
</tr>
<tr>
<td>TARBP2/PKRA</td>
<td>DsRM</td>
<td>Recruitment of Dicer-associated miRNAs or siRNA to RISC; regulation of the activity of PKR, a dsRNA-regulated interferon-inducible protein kinase</td>
<td>Rossi 2005; Kim et al. 2009</td>
</tr>
<tr>
<td>smaug</td>
<td>Others</td>
<td>SAM domain; anterior-posterior axis formation and maternal to zygotic transition in <em>Drosophila</em>; translational repression in mouse neurons</td>
<td>Aviv et al. 2003; Baez and Boccaccio 2005; Tadros et al. 2007</td>
</tr>
<tr>
<td>mbl</td>
<td>Others</td>
<td>Zinc fingers; control of alternative splicing of RNA during eye and muscle development</td>
<td>Begemann et al. 1997; Kanadia et al. 2003; Pascual et al. 2006</td>
</tr>
</tbody>
</table>
indicating that the corresponding genes were also present in the metazoan LCA.

We performed detailed phylogenetic analyses for 32 families of such RBPs. For these phylogenetic analyses, which are reported in the following sections, we identified and analyzed members of each of these RBP families from 28 different completely sequenced opisthokont genomes. The consensus view of the phylogenetic relationships between the studied species is shown in supplementary figure S1, Supplementary Material online. We summarized the phylogenetic analyses in supplementary table 6, Supplementary Material online and our inference about the timing of appearance of the 32 families in figure 1 as well as the occurrence of putative gene loss and duplication events in nonbilaterians (supplementary fig. S2, Supplementary Material online) and bilaterians (supplementary fig. S3, Supplementary Material online).

### Phylogenetic Analysis of RBPs With RRM Domains

We analyzed eight different families of RRM proteins—Elav/Hu, Musashi (Msi), DAZ associated protein 1 (Dazap1), Epithelial splicing Regulatory Protein (ESRP), Transformer-2 (Tra-2), Split ends (Spen), Spenito/RBM15, and Bruno/CUG-BP1 and ETR-3 Like Factors (CELF)—known to be involved in various aspects of development in bilaterians (CUG-BP1 and ETR-3 Like Factors (CELF)—known to be involved in various aspects of development in bilaterians (see supplementary fig. S2, Supplementary Material online). We found one or more elav/Hu genes in all of the 26 studied holozoan species, including Amphimedon and Monosiga, but not in fungi (supplementary table 6, Supplementary Material online). This gene was therefore already present in the LCA of holozoans but probably not in that of opisthokonts. The phylogenetic analysis suggests that a single elav/Hu gene was present in the LCAs of holozoans, metazoans, and bilaterians and that multiple independent duplications have since occurred in various metazoan lineages (supplementary fig. S4, Supplementary Material online). msi and Dazap1 are closely related genes, and we therefore made phylogenetic analyses of these two families together (supplementary fig. S5, Supplementary Material online). We found two well-supported monophyletic groups, one including the bilaterian msi genes, the other the bilaterian Dazap1 genes (supplementary table 6, Supplementary Material online; supplementary fig. S5, Supplementary Material online). Both groups also include Amphimedon sequences, indicating that both msi and Dazap1 genes were already present in the LCA of metazoans. The Monosiga genome includes a Dazap1 gene but not a msi gene, suggesting that either the two gene families were already present in the LCA of holozoans but that the msi gene has been lost (or became much divergent) in Monosiga or that msi genes are specific to metazoans (supplementary fig. S5, Supplementary Material online). Neither msi nor Dazap1 genes were found in fungi (supplementary table 6, Supplementary Material online), indicating that these genes likely represent holozoan innovations.

We found single ESRP, tra-2, spen, and spenito/RBM15 genes in most species we studied, including Amphimedon (supplementary table 6, Supplementary Material online). None of these genes were found in fungi and only two of them (ESRP and tra-2) in Monosiga (supplementary table 6, Supplementary Material online). We therefore conclude that none of these genes were present in the LCA of opisthokonts, ESRP and tra-2 were present in the LCA of holozoans, and spen and spenito/RBM15 are metazoan-specific genes. Phylogenetic analyses of these four families indicate specific duplications that occurred in vertebrates as well as in a few other species (supplementary figs. S6–S9, Supplementary Material online). We found two or more Bruno/CELF genes in most animal species we investigated, including...
Amphimedon; the choanoflagellate Monosiga has a single bruno gene, and no bruno genes were found in fungi (supplementary table 6, Supplementary Material online). Phylogenetic analysis showed the existence of two large monophyletic groups, indicating that two bruno genes were already present in the LCA of bilaterians (supplementary fig. S10, Supplementary Material online). In contrast, we were unable to confidently determine whether a single or two bruno genes were present in the LCA of metazoans and holozoans due to uncertainties in the phylogenetic trees (supplementary fig. S10, Supplementary Material online).

We conclude that the eight analyzed families of RRM-containing RBPs evolved in holozoans after their divergence from fungi, and at least five of them (elav/hu, Dazap1, ESRP, tra-2, and bruno/CELF) were already present in the LCA of holozoans, whereas the three remaining ones (spen, spenito/RBM15, and nsi) are metazoan-specific genes (fig. 1). Our analysis also reveals the presence of two bruno/CELF genes is the LCA of bilaterians and their subsequent conservation in most present-day bilaterian species.

Phylogenetic Analysis of RBPs With KH or DEAD-box Domains
We analyzed three families of KH proteins—Fragile X Mental Retardation 1 (FMR1), Bicaudal-C (Bicc), and Muscle in excess-3 (Mex-3)—and four families of DEAD-box proteins—Vasa, PI10/DDX3, MOV10, and Armitage—known to be involved in development, including germ/stem cell formation or maintenance, or in the miRNA pathway (see table 3 for details and references). We found single FMR1 and Bicc genes in most species we studied, including Amphimedon and Monosiga, but not in fungi (supplementary table 6, Supplementary Material online), indicating that these two RBP subtypes were already present in the LCA of holozoans but not in that of opisthokonts. Phylogenetic analyses suggest an almost complete absence of duplication events in the case of Bicc and more frequent duplications in the case of FMR1 (supplementary figs. S11 and S12, Supplementary Material online). Notably, we detected five closely related FMR1 genes in the pea aphid Acrithosiphon pisum (supplementary fig. S11, Supplementary Material online). Interestingly, several other genes involved in the miRNA pathway also have undergone duplications in the pea aphid, suggesting a specific expansion of this pathway in this species (see below; Jaubert-Possamai et al. 2010). We found a single Mex-3 gene in most animal genomes, including Amphimedon, but none in Monosiga and fungi (supplementary table 6, Supplementary Material online), suggesting that Mex-3 is a metazoan-specific gene. Specific duplications occurred in vertebrates (supplementary fig. S13, Supplementary Material online).

We found many vasa- and PI10/DDX3-like genes, and their phylogenetic analysis allowed us to define two well-defined vasa and PI10 monophyletic groups (fig. 2). We found vasa and PI10 genes in most animal genomes including Amphimedon (supplementary table 6, Supplementary Material online), indicating that these two genes were already present in the LCA of metazoans. We identified a clear PI10 gene but no vasa gene in Monosiga and the two fungi we analyzed (fig. 2; supplementary table 6, Supplementary Material online). We also retrieved PI10/vasa orthologs in several other fungi and made phylogenetic analyses of these sequences with those of metazoans. We found that all the fungi sequences strongly cluster with the Monosiga and metazoan PI10 genes (supplementary fig. S14, Supplementary Material online). These data therefore suggest that the vasa gene originated in metazoans from a duplication of an ancestral PI10-like gene with one duplicate (metazoan PI10) remaining more similar to the ancestral gene than the other duplicate (metazoan vasa). Our phylogenetic analysis also suggests the occurrence of several independent duplications of the vasa gene in metazoans as well as a few cases of independent duplications for the PI10 family (fig. 2).

We also retrieved MOV10/Armitage proteins and made phylogenetic analyses. We found two monophyletic groups (supplementary fig. S15, Supplementary Material online). One group (that we named Armitage as it includes the Drosophila Armitage protein) contains one or more sequences from most of the species under study, including Amphimedon, Monosiga, and fungi (supplementary table 6, Supplementary Material online), indicating that the corresponding gene is ancestral to opisthokonts. The second monophyletic group (that we named MOV10 as it contains the vertebrate MOV10 proteins) contains sequences from only a few animal species (supplementary fig. S15, Supplementary Material online); however, these species belong to deuterostomes, lophotrochozoans, and cnidarians, suggesting that the corresponding gene was present in the LCA of eumetazoans and has been lost in edyszoans.

We conclude that of the seven families that we studied, two (PI10 and Armitage) are ancestral to opisthokonts, two (FMR1 and Bicc) were present in the LCA of holozoans, two represent metazoan innovations (vasa and Mex-3), and one (MOV10) is specific to eumetazoans (fig. 1).

Phylogenetic Analysis of RBPs With DsRM Domains
We selected five families of DsRM proteins known to be involved either in the biogenesis of functional miRNAs—Drosha (also known as RNASEN), Partner of Drosha (Pasha; also known as DiGeorge syndrome critical region gene 8, DGCR8), Dicer, and TAR (HIV transactivator RNA) RBP (TARBP2)/Protein kinase R activating protein (PKRA, also known as PACT)—or in development—Staufen—(see table 3 for details and references).

Drosha, Pasha, and Dicer proteins have been previously shown to be encoded by the Amphimedon genome (Grimson et al. 2008; de Jong et al. 2009). Pasha genes were found only in metazoans, usually as single gene with the exception of the pea aphid Ac. pisum in which specific duplications produced four paralogs (supplementary table 6 and fig. S16, Supplementary Material online), as previously described (Jaubert-Possamai et al. 2010). In the case of the Drosha and Dicer families, we noticed during our Blast searches that some fungi genes could not be confidently assigned to either of these two families by sequence similarities, and we therefore made the phylogenetic analyses of both families together. We found two well-supported monophyletic groups corresponding to dicer and drosha.
genes (supplementary fig. S17, Supplementary Material online). We found drosha genes in animals and Monosiga (supplementary table 6, Supplementary Material online), indicating that a drosha gene was present in the holozoan LCA. We also found two fungi genes that cluster with the holozoan drosha genes, albeit with weak statistical support (supplementary fig. S17, Supplementary Material online)—it is therefore possible that drosha genes are ancestral to opisthokonts (supplementary fig. S17, Supplementary Material online). Our phylogenetic analysis of the dicer family is in agreement with previous studies (Grimson et al. 2008; Murphy et al. 2008; de Jong et al. 2009) and shows the absence of dicer genes in Monosiga (a secondary loss, as dicer genes are found in fungi), the presence of several dicer genes in basal metazoans, such as Amphinemdon and Trichoplax and the occurrence of additional duplications in Daphnia and Acrystisiphon (supplementary table 6 and fig. S17, Supplementary Material online).

We made phylogenetic analyses for TARBP2-, PKRA-, and Staufen-like proteins and found two well-supported monophyletic groups corresponding to the Staufen proteins and the TARBP2 and PKRA proteins, respectively (supplementary fig. S18, Supplementary Material online). Staufen proteins are encoded by the genomes of many bilaterian species as well as by Nematostella but were not found encoded by the Amphinemdon, Trichoplax, Monosiga, or fungi genomes (supplementary fig. S18 and table 6, Supplementary Material online). Staufen may therefore correspond to a eumetazoan innovation or may have been lost in the aforementioned species. Independent duplications of the Staufen gene occurred in some bilaterians, such as vertebrates, Daphnia, Tribolium, and Caenorhabditis (supplementary fig. S18, Supplementary Material online). The second monophyletic group includes TARBP2 and PKRA proteins, Drosophila R2D2 and Loquacious, Caenorhabditis RDE-4 as well as one or more proteins from most of the studied species, including Amphinemdon, with the exception of Hydra, Trichoplax, Monosiga, and fungi (supplementary table 6, Supplementary Material online). The phylogenetic analysis indicates that the TARBP2 and PKRA genes found in vertebrates result from a vertebrate-specific duplication because only a single gene is found in nonvertebrate deuterostomes, lophotrochozoans, and noninsect ecdysozoans (supplementary fig. S18, Supplementary Material online). An independent duplication occurred in insects to give rise to loquacious- and r2d2-like genes (supplementary fig. S18, Supplementary Material online).

We conclude that Dicer and probably Drosha were present in the opisthokont LCA, TARBP2 and Pasha are likely to be metazoan innovations, and Staufen is specific to eumetazoans (fig. 1).

Phylogenetic Analysis of RBPs With Other RNA-Binding Domains

We analyzed twelve families of RBPs which bind RNA with various other types of RNA-binding domains—Smaug: Muscleblind (Mbl); Nanos, Pumilio; PUF; Exportin-1, -5, and -T; Piwi; Argonaute (Ago); and Tudor-SN—and that are known to be involved in development, especially stem or germ cell development, and/or ncRNA pathways (see table 3 for details and references). Single smaug and mbl genes were found in most metazoan species, including Amphinemdon, but not in the choanoflagellate Monosiga (supplementary table 6, Supplementary Material online). A smaug gene was nevertheless found in fungi (supplementary table 6, Supplementary Material online), suggesting that this gene was present in the opisthokont LCA and has been lost in Monosiga. No mbl gene can be found in fungi, suggesting that mbl is specific to metazoa. In only, a few cases duplications can be observed for the smaug and mbl families, in particular in vertebrates (supplementary figs. S19 and S20, Supplementary Material online). We also found nanos genes to be present in most of the animal genomes we studied, including that of Amphinemdon, but neither in Monosiga nor in fungal genomes (supplementary table 6, Supplementary Material online), indicating that nanos is a metazoan-specific gene. Phylogenetic analyses suggest multiple independent gene duplications in various lineages, such as vertebrates, arthropods, nematodes, and cnidarians (supplementary fig. S21, Supplementary Material online).

We found several genes that encode proteins with a PUF domain in all the species we studied—phylogenetic analyses allowed us to define three well-supported monophyletic groups, Pumilio, PUF-A, and PUF-B (fig. 3). We found these three groups to be ancestral to opisthokonts, as present in both Monosiga and fungi (supplementary table 6, Supplementary Material online). Although a Pumilio and a PUF-A gene were found in Amphinemdon, there is no PUF-B gene, indicating a secondary gene loss. Only a single PUF-A and PUF-B gene is found in most species, but many duplications have occurred in the Pumilio family, especially in vertebrates, fungi, insects, and nematodes (fig. 3). In nematodes, a very large expansion of the PUF gene family has occurred, with many of these genes forming a separate monophyletic group in the phylogenetic tree—this may correspond to divergent Pumilio, PUF-A, or PUF-B genes (fig. 3). A recent study has shown an important diversification of the PUF genes in plants with, for example, 26
Phylogenetic analysis of the PUF genes. A Maximum likelihood tree is shown. Midpoint rooting has been used. A very similar tree topology has been obtained by Bayesian inference. Statistical supports are as in figure 2. Full sequence and accession numbers of the proteins shown in this tree can be found in supplementary table 7, Supplementary Material online.
members in *Arabidopsis thaliana* (Tam et al. 2010). As phylogenetic analyses of these plant PUF genes were performed with only a few *Pumilio* sequences (Tam et al. 2010) and in order to get insights into the evolution of the PUF genes at the scale of the eukaryotes, we conducted phylogenetic analyses with our full set of opisthokont sequences and the previously analyzed sequences from *Ar. thaliana*, the moss *Physcomitrella patens*, and the green alga *Chlamydomonas reinhardtii*, as well as additional sequences retrieved from the genomes of *Ostreococcus lucimarinus* (basal Viridiplantae), *Dictyostelium discoideum* (Amoebzoa), *Naegleria gruberi* (Heterolobosea), *Emiliania huxleyi* (Haptophyta), *Phaeodactylum tricornutum* (Heteroconota), and *Thalassiosira pseudonana* (Heteroconta). We found that the *Pumilio*, *PUF*-A, and *PUF*-B groups are ancestral to eukaryotes and that the *Pumilio* group can be further subdivided into two subgroups (we named *Pumilio*-A and *Pumilio*-B), both being ancestral to eukaryotes, but with *Pumilio*-B having been lost in the holozoan lineage (supplementary fig. S22, Supplementary Material online).

We retrieved all Exportin proteins that show high sequence similarity to known Exportin-5, Exportin-1, or Exportin-T and made phylogenetic analyses of their sequences (supplementary fig. S23, Supplementary Material online). In agreement with Murphy et al. (2008), we found three well-supported monophyletic groups corresponding to the three aforementioned Exportin subtypes. These three subtypes are found in most of the studied genomes, including those of *Arabidopsis thaliana*, *Physcomitrella patens*, and the green alga *Chlamydomonas reinhardtii*, as well as additional sequences retrieved from the genomes of *Ostreococcus lucimarinus* (basal Viridiplantae), *Dictyostelium discoideum* (Amoebzoa), *Naegleria gruberi* (Heterolobosea), *Emiliania huxleyi* (Haptophyta), *Phaeodactylum tricornutum* (Heteroconota), and *Thalassiosira pseudonana* (Heteroconota). We found that the *Pumilio*, *PUF*-A, and *PUF*-B groups are ancestral to eukaryotes and that the *Pumilio* group can be further subdivided into two subgroups (we named *Pumilio*-A and *Pumilio*-B), both being ancestral to eukaryotes, but with *Pumilio*-B having been lost in the holozoan lineage (supplementary fig. S22, Supplementary Material online).

We found several genes encoding proteins with Tudor domains in *Amphimedon*, including one that contains 11 Tudor motifs, like the *Drosophila* Tudor protein (not shown). One of these *Amphimedon* genes encodes a protein with one Tudor motif and five staphylococcal nuclease homology motifs and shows a strong sequence similarity with bilaterian Tudor-SN, indicating the existence of a single *Tudor*-SN gene in *Amphimedon*. We also found a single *Tudor*-SN gene in most species we studied, including *Monosiga* and fungi (supplementary table 6 and fig. S26, Supplementary Material online). *Tudor*-SN genes are also found in trypanosomes and plants (Alsford et al. 2010; dit Frey et al. 2010; M.V., unpublished observations), indicating that *Tudor*-SN originated early during eukaryote evolution.

We conclude that 10 of the 12 analyzed families were present in the opisthokont LCA and that two of them (*mbl* and *nanos*) are metazoan innovations (fig. 1). We also found that two *piwi* genes were present in the LCA of metazoans and that the PUF family can be subdivided into four subfamilies that are ancestral to eukaryotes, one of these having been lost in the holozoan lineage. Furthermore, our analysis indicates the loss of three genes involved in miRNA or piRNA pathways (*dicer*, *argonaute*, and *piwi*) in *Monosiga*, in agreement with the absence of such small ncRNAs in this species (Grimson et al. 2008).

### Conclusions

By analyzing the RBP repertoire in the genomes of living representatives of the earliest branching phyletic lineages of animals—specifically the cnidarian *Nematostella*, the placozoan *Trichoplax*, and the demosponge *Amphimedon* as well as a unicellular sister group of the Metazoa, the choanoflagellate *Monosiga*—we have been able to reconstruct the early evolution of this important class of metazoan regulatory genes. This study reveals that the origin of most RBPs occurred prior to the divergence of eumetazoan and sponge lineages, with the metazoan LCA possessing at least 200 different RBPs; 155 of these appear to have been inherited from
Fig. 4. Phylogenetic analysis of the piwi genes. A part of the Maximum likelihood tree shown in supplementary figure S24, Supplementary Material online. Midpoint rooting has been used. Statistical supports are as in figure 2. Full sequence and accession numbers of the proteins shown in this tree can be found in supplementary table 7, Supplementary Material online.
the holozoan LCA, and 118 from the opisthokont LCA. In contrast, we estimate that only another 12 RBPs evolved along the eumetazoan stem. As many RBPs are instrumental in developmental and regulatory processes that exist in bilaterians but not in the sponge (e.g., nerve and muscle formation), we can infer that many of these genes had ancestral roles differing from their well-documented functions in bilaterians. In many cases, their role in bilaterian development must have been the result of a co-option event some time after the divergence of eumetazoan and sponge lineages. Other metazoan-specific RBPs appear to have more conserved roles in regulating processes associated with animal multicellularity (e.g., in miRNA biogenesis). Indeed, the remarkable conservation of nearly 200 RBPs in animal lineages separated for more than 600 My strongly suggests these proteins have been instrumental in the genesis and maintenance of the multicellular condition since the origin of the animal kingdom.

**Supplementary Material**

Supplementary tables S1–S7 and figures S1–S26 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We are extremely grateful to the DoE Joint Genome Institute, especially D. Rokhsar and M. Srivastava, for making the *Amphimedon* sequences available to us before publication. We are grateful to the ATGC bioinformatics platform and the Laboratoire d’Informatique, de Robotique et de Microélectronique (LIRMM) (Montpellier) for their help for the phylogenetic analyses. Our work has been supported by the Centre national de la recherche scientifique, the Agence Nationale de la Recherche, and the Université Paris Diderot (to M.V.), and the Australian Research Council (to B.M.D).

**References**


