A Phylogenomic Investigation into the Origin of Metazoa

Iñaki Ruiz-Trillo,† Andrew J. Roger,* Gertraud Burger,‡ Michael W. Gray,* and B. Franz Lang‡*

*Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada; †ICREA Researcher at Departament de Genètica, Universitat de Barcelona, Barcelona, Spain; and ‡Département de Biochimie, Robert Cedergren Center for Bioinformatics and Genomics, Université de Montréal, Program in Evolutionary Biology, Canadian Institute for Advanced Research, Boulevard Edouard-Montpetit, Montréal, Québec, Canada

The evolution of multicellular animals (Metazoa) from their unicellular ancestors was a key transition that was accompanied by the emergence and diversification of gene families associated with multicellularity. To clarify the timing and order of specific events in this transition, we conducted expressed sequence tag surveys on 4 putative protistan relatives of Metazoa including the choanoflagellate *Monosiga ovata*, the ichthyosporeans *Sphaeroforma arctica* and *Amoebidium parasiticum*, and the amoeba *Capsaspora owczarzaki*, and 2 members of Amoebozoa, *Acanthamoeba castellanii* and *Mastigamoeba balamuthi*. We find that homologs of genes involved in metazoan multicellularity exist in several of these unicellular organisms, including 1 encoding a membrane-associated guanylate kinase with an inverted arrangement of protein-protein interaction domains (MAGI) in *Capsaspora*. In Metazoa, MAGI regulates tight junctions involved in cell-cell communication. By phylogenomic analyses of genes encoded in nuclear and mitochondrial genomes, we show that the choanoflagellates are the closest relatives of the Metazoa, followed by the *Capsaspora* and Ichthyosporea lineages, although the branching order between the latter 2 groups remains unclear. Understanding the function of "metazoan-specific" proteins we have identified in these protists will clarify the evolutionary steps that led to the emergence of the Metazoa.

Introduction

"When animals appeared, the world changed, in some ways, forever," according to Simon Conway Morris in his book *The Crucible of Creation* (1998). Multicellularity has been independently acquired several times within eukaryotes in groups such as animals, fungi, plants, slime molds, and various algal lineages. The emergence of the multicellular Metazoa from protist ancestors is unquestionably the most spectacular case, with the abrupt appearance of an astounding variety of metazoans in the fossil record about 530 MYA, an event known as the Cambrian explosion (Conway Morris 2003, 2006). Attempts to explain this evolutionary transition range from environmental (e.g., the increase in free oxygen levels) to ecological (e.g., changes in ecosystems) and evolutionary (such as horizontal gene transfer; for a review, see Valentine 2004). To be complete, any answer will not only involve environmental or ecological factors but must also take into account the suite of genes present in the genomes of the protistan ancestors of Metazoa.

Molecular phylogenetic analyses have definitively shown that Metazoa and Fungi share a common ancestor and form a eukaryotic supergroup known as the opisthokonts (Baldauf and Palmer 1993; Steenkamp and Baldauf 2004). Recent analyses have changed our view of opisthokont phylogeny by adding new unicellular lineages to this clade, such as the nucleariids (Nucleariidae), the choanoflagellates (Choanoflagellata), the ichthyosporeans (Ichthyosporea), and the genera *Capsaspora*, *Corallochytrium*, and *Ministeria* (Ragan et al. 1996, 2003; Lang et al. 2002; Mendoza et al. 2002; Medina et al. 2003; Philippe et al. 2004; Ruiz-Trillo et al. 2006; Steenkamp et al. 2006; Moreira et al. 2007). In these studies, nucleariids consistently appear to be the sister group to Fungi, whereas choanoflagellates, ichthyosporeans, *Capsaspora*, *Corallochytrium*, and *Ministeria* are close to Metazoa. However, previous attempts to resolve opisthokont phylogeny have suffered from several drawbacks. The phylogenomic analyses published thus far have a very limited taxonomic sampling, including just 1 or 2 unicellular lineages (Lang et al. 2002; Philippe et al. 2004). On the other hand, all taxon-rich analyses have been based on a small number of sequence characters derived from at most 4 genes (Ragan et al. 1996, 2003; Mendoza et al. 2002; Medina et al. 2003; Ruiz-Trillo et al. 2004, 2006; Steenkamp et al. 2006; Moreira et al. 2007). Therefore, several aspects of the opisthokont phylogeny remain contentious. For instance, the position of *Capsaspora* relative to Metazoa, Choanoflagellata, and Ichthyosporea varies from study to study depending on the taxa included and the genes analyzed.

To clarify the unicellular-to-multicellular transition that occurred at the origin of Metazoa, comparative genomic analyses must include both Metazoa and their closest unicellular relatives. However, the choanoflagellates are the only unicellular relative of Metazoa to have been studied on the genomic level (King and Carroll 2001; Snell et al. 2001, 2006; King et al. 2003; King 2004). These studies demonstrated that the choanoflagellates express a wide variety of genes associated with multicellularity, such as cadherins, C-type lectins, tyrosine kinases, and, most recently, a Hedgehog homolog (Snell et al. 2006). In *Monosiga*, these genes are proposed to be involved in 2 processes: sex and predation, both of which require cell–cell recognition, adhesion and endocytosis, or fusion (King et al. 2003; King 2004). Most cell signaling, cell adhesion, and transcription factor genes are widely conserved across Metazoa including morphologically simple taxa such as cnidarians and sponges (Finnerty and Martindale 1999; Kusserow et al. 2005; Miller et al. 2005; Technau et al. 2005; Nichols et al. 2006; Ryan et al. 2006, 2007; Adamska et al. 2007; Putnam et al. 2007; Sullivan et al. 2007). However, several genes associated with multicellularity have, so far, not been detected outside Metazoa. These genes include the Antennapedia (ANTP) and paired classes within the homeobox superfamily and the T-box and Pax families of transcription factors. Actually, a recent examination of the complete genome sequence of the sponge *Amphimedon queenslandica*, which branches prior to the common ancestor of Cnidaria

Key words: multicellularity, metazoa, phylogenomics, opisthokonts, MAGI.

E-mail: inaki.ruiz@icrea.es.

Mol. Biol. Evol. 25(4):664–672. 2008
doi:10.1093/molbev/msn006
Advance Access publication January 9, 2008

and Bilateria (Eumetazoa), shows that within the ANTP class, *A. queenslandica* possesses several NK-like genes but no Hox, ParaHox, or EHGbox genes (Larroux et al. 2007). Similarly, the placozoan *Trichoplax adhaerens*, a basal metazoan of unclear phylogenetic position (Dellaporta et al. 2006), seems also to have a low diversity of ANTP class homeobox genes (Monteiro et al. 2006). Genes coding for proteins involved in cell adhesion and the extracellular matrix are also of key importance to the origin of metazoan multicellularity. Although some of these proteins, such as the cadherins, have already been detected in choanoflagellates (King et al. 2003; King 2004), others have been suggested to be specific to Metazoa, including the membrane-associated guanylate kinases with inverted arrangements (MAGIs) that participate in the assembly of multiprotein complexes at regions of cell–cell contact (Dobrosotskaya et al. 1997; Dobrosotskaya and James 2000). All these observations seem to suggest that the last common ancestor of metazoans was not as genetically complex as the last common ancestor of Eumetazoa (Metazoa excluding Porifera). On the other hand, it is likely that unicellular ancestors of Metazoa possessed some part of the "genetic toolkit" needed to construct a multicellular body plan. Yet, it remains unclear how many of the genes responsible for multicellularity were adapted from preexisting ones in unicells and which ones originated in the first metazoans.

In order to clarify the origin of Metazoa, we need a well-resolved phylogeny of the opisthokonts and genomic data from unicellular, close relatives of metazoans. With this aim, we have undertaken expressed sequence tag (EST) projects from 4 unicellular opisthokonts (*Capsaspora owczarzaki*, the ichthyosporeans *Sphaeroforma arctica* and *Amoebidium parasiticum*, and the choanoflagellate *Monosiga ovata*). We have also undertaken EST projects from 2 taxa (*Mastigamoeba balamuthi* and *Acanthamoeba castellanii*) from Amoebozoa, the sister group to opisthokonts. Furthermore, we have obtained the almost complete sequence of the mitochondrial genome of the single-celled opisthokont *C. owczarzaki*. We have built large concatenated nuclear and mitochondrial alignments that increase both the number of genes and the taxonomic sampling of single-celled relatives of Metazoa and the sister group of Opisthokonta. Moreover, we searched our protistan EST data for the occurrence of genes relevant to the origin of multicellularity in Metazoa. Among them, we have identified the first member of the MAGI outside Metazoa. Here, by combining these 2 approaches, we clarify the origin of metazoan multicellularity by further delineating the phylogenetic placement of these unicellular lineages relative to Metazoa and by identifying genes in their transcriptomes that are associated with metazoan multicellularity.

Materials and Methods

EST Data

Total RNA was extracted using TRI Reagent (Molecular Research Center, Cincinnati, OH) following the manufacturer's guidelines. Complementary DNA (cDNA) libraries were constructed by Amplicon Express (Pullman, WA) except for *A. castellanii* and *M. balamuthi* (DNA Technologies Inc., Gaithersburg, MD), *A. parasiticum* (for details, see Rodriguez-Ezpeleta et al. 2007), and *M. ovata* (a lambda ZAPII library, kindly provided by Dr P. Holland). The number of ESTs passing quality control and submitted to further analysis was 13,770 (*A. castellanii*), 3,623 (*A. parasiticum*), 8,870 (*C. owczarzaki*), 19,182 (*M. balamuthi*), 6,433 (*M. ovata*), and 8,006 (*S. arctica*). EST data were automatically clustered by tools implemented in Taxonomically Broad EST Database (TBestDB) (http://amoebidia.bcm.umontreal.ca/pepdb/searches/login.php?bye=true) (O'Brien et al. 2007) and AnaBench (http://anabench.bcm.umontreal.ca/anabench/) (Badidi et al. 2003). *Capsaspora owczarzaki* and *S. arctica* data were also manually clustered using Phred and Phrap and CAP3 (Ewing and Green 1998; Huang and Madan 1999). All EST data generated for this article are publicly available from the GenBank EST data set, and clusters are available at TBestDB.

Purification and Sequencing of Mitochondrial DNA

Cells were lysed in the presence of 1% sodium dodecyl sulfate in TE buffer, and mitochondrial DNA (mtDNA) was purified from this whole-cell lysate as described previously (Lang and Burger 2007). The mtDNA sequence was determined from a random shotgun sequence library (Burger et al. 2007), using Phred and Phrap (http://www.phrap.org/) for genome assembly.

Phylogenetic Analyses

A concatenated data set of 110 nucleus-encoded proteins was constructed with MacGDE2.3 (Smith et al. 1994) by combining the data sets described in Philippe et al. (2004, 2005), our new EST data, and new data from other publicly available EST or genome projects. All individual gene alignments were manually inspected and edited. All potential paralogs were manually inspected, and if paralogy could not be ruled out, the corresponding proteins were removed from the final alignment. Furthermore, only those positions that were unambiguously aligned were manually included in the final analysis, resulting in a total of 20,711 amino acid positions. The 13 mitochondrial protein sequences inferred from mtDNA sequence were automatically aligned with Muscle (Edgar 2004) and concatenated, after using Gblocks (Castresana 2000) with default parameters to remove regions that were not aligned with confidence. The final nuclear and mitochondrial alignments can be downloaded from http://www.multicellgenome.com/Lab/Welcome.html.

Phylogenetic trees from the concatenated data set were estimated using a combination of programs and procedures in order to test for any potential systematic errors. We first built the tree using IQPNNI (le and Haeseler 2004) and Raxml (Stamatakis et al. 2005; Stamatakis 2006) programs using a Whelan and Goldman (WAG) model of evolution and with a gamma distribution (8 categories) (WAG + Γ). Statistical support was obtained from 100 bootstrap replicates using a WAG + Γ model (4 rate categories). Bootstrapping in IQPNNI was performed by Iqpnniboot kindly provided by Jessica Leigh, Dalhousie University. Both Raxml and IQPNNI gave nearly identical topologies and bootstrap values. Thus, in the interest of clarity, only IQPNNI results will be shown throughout the manuscript.

Table 1
Deep Phylogenetic Relationships between Metazoa and Their Unicellular Relatives Obtained with Different Data Sets and Using Various Methodologies

	Topologies (% bootstrap support)				
	Nuclear data set			Mitochondrial data set	
Methods	Ecdysozoa monophyletic	Cnidaria monophyletic	*Capsaspora* + Ichthyosporea	*Trichoplax* + Bilateria	*Capsaspora*-independent lineage
ML	− (51)	− (65)	+ (100)	− (85)	+
ML (FSR)	− (<50)	− (83)	+ (100)	NA	NA
ML excluding taxa with >50% missing data	+ (57)	NA	+ (100)	NA	NA
ML (recoded data)	+ (95)	+ (52)	+ (97)	− (39)	+ (57/49)
Bayesian (CAT model)	+ (99)	+ (51)	+ (97)	—	+ (93/72)
Bayesian (CAT model + recoded)	+	+	+	NA	NA

NOTE.—Statistical support is indicated where available. See text and Supplementary Material online for full details and discussion of methods. NA, not available; ML, maximum likelihood (IQPNNI program); and FSR, fast-evolving sites removed.

To minimize potential systematic error, we used several methods. First, we reconstructed the tree with the fastest evolving sites removed as in Ruiz-Trillo et al. (1999), which resulted in a total of 18,341 amino acid positions. The conditional mode site rates of all amino acid positions were estimated with Tree-Puzzle v. 5.2 (Schmidt et al. 2002) with the WAG + Γ model (8 rate categories). Fastest evolving sites (category 8) were excluded from the analysis, and a new tree was estimated using IQPNNI with the WAG + Γ model (8 rate categories). Second, to reduce the effects of compositional heterogeneity and saturation, we recoded the amino acids into 4 functional categories as specified in Rodriguez-Ezpeleta et al. (2007) and reconstructed a tree by IQPNNI using a general time reversible + Γ model (8 rate categories). Third, we used a site-heterogeneous mixture model (CAT + Γ) that accounts for different models of evolution for classes of aligned sites (Lartillot et al. 2007) implemented in a Bayesian approach using the program PhyloBayes (Lartillot and Philippe 2004). Statistical support was obtained by 100 bootstrap replicates in IQPNNI and PhyloBayes. Fourth, we also attempted recoding analyses using the site-heterogeneous CAT + Γ model in PhyloBayes. In this case, we recoded several states most influenced by guanine-cytosine/adenosine-thymine composition biases, R/K and V/I, into one state, yielding an 18-state CAT + Γ model. Finally, to determine whether missing data had a significant effect on the nuclear analyses, taxa with more than 50% missing data (see supplementary table S1, Supplementary Material online) were excluded and phylogenies were constructed using IQPNNI. Statistical tests of alternative topologies (*Capsaspora* as sister group of ichthyosporeans or as sister group of choanoflagellates + Metazoa) were performed on both the nuclear and the mitochondrial data set using the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa 1999) and the expected likelihood weights (ELWs) test (Strimmer and Rambaut 2002), implemented in Tree-Puzzle v. 5.2 (Schmidt et al. 2002).

Searches for Metazoan-Specific Genes

We searched for known metazoan cell signaling, cell adhesion, and transcription factor gene families in all our ESTs. We also searched in all available ESTs and genomic databases from Eukaryota. We used the cnidarian homolog (or if this was not available, we used the human one instead) as a query with TBlastN against our protistan EST data using AnaBench (Badidi et al. 2003). All putative positives (E value $< 1 \times 10^{-05}$) were rechecked by blasting the putative hit against PFAM version 21.0 (Finn et al. 2006). Only those TBlastN hits that also gave in PFAM the same gene family were considered robust positives. For example, we blasted a human Hedgehog homolog against our protistan database, retrieving one putative homolog in *Monosiga* ESTs. The protein sequence inferred from this *Monosiga* EST was blasted against PFAM, identifying the *Monosiga* candidate as a clear Hedgehog protein. Results are shown in table 2.

Characterization of MAGI

Purified polyA$^+$ messenger RNA and cDNA from *Capsaspora* were obtained as in Ruiz-Trillo et al. (2006) and independently from the RNA used to construct the cDNA libraries. We obtained the full sequence of the 5′ and 3′ ends of *Capsaspora* MAGI by rapid amplification of cDNA ends using the GeneRacer kit (Invitrogen, Carlsbad, CA) with primers designed from the original EST sequence. Sequences were obtained and analyzed as in Ruiz-Trillo et al. (2006). Alignment of different MAGI homologs was done using MacGDE2.3 software (Smith et al. 1994). The MAGI alignment is available from http://www.multicellgenome.com/Lab/Welcome.html. Searches for MAGI were carefully conducted using different Blast methodologies at National Center for Biotechnology Information (NCBI), TBestDB, PFAM (Finn et al. 2006), Metazome, Prosite (Hulo et al. 2006), and ongoing genome projects from The Institute for Genomic Research and Joint Genome Institute. A phylogenetic tree, from an alignment comprising all MAGI protein domains, was estimated using IQPNNI with the WAG + Γ model (8 rate categories). Bootstrap values were obtained by 100 replicates in Phyml (Guindon and Gascuel 2003) using a WAG + Γ model (4 rate categories). The domain architectures of MAGI homologs were obtained using Blast against Prosite and NCBI databases.

Table 2
Phylogenetic Distribution of Genes Associated with Metazoan Multicellularity

Function	Gene Product	Organisms	GenBank Accession Number	Other Nonmetazoa Taxa Where Gene Is Present
Cell adhesion and adhesion related	Tetraspanin	*Capsaspora*	EF693744	Fungi and Amoebozoa
	Laminin A	*Capsaspora* and *Monosiga ovata*	EC736556, EC165586, EC164818	Amoebozoa and *Trypanosoma*
	Beta-catenin–interacting protein (ICAT)	*Capsaspora*, *Acanthamoeba*	EC740811, EF693748	*Dictyostelium*
	MAGI	*Capsaspora*	EF611870	—
	Ankyrin	*Capsaspora*, *Mastigamoeba*	EC737721, EC705671	Plants and Fungi
	Fascin	*Capsaspora*, *Monosiga ovata*	EF693745, EF693746, EF693747	*Dictyostelium*
Transcription factor	Forkhead	*Amoebidium*	EC627545, EC629343	Fungi
Cell signaling	Hedgehog	*Monosiga ovata*	ABA55664	—

NOTE.—See main text and Supplementary Material online for details of the methods used.

Results and Discussion

Molecular Phylogeny

Here, we address the phylogeny of Opisthokonta by increasing both the number of genes and the taxonomic sampling of single-celled relatives of Metazoa and members of Amoebozoa, the sister group to opisthokonts. Two concatenated alignments were constructed using data from both our EST and mitochondrial genome projects plus data from publicly available EST and genome projects. The nuclear data set includes a total of 30 taxa, 6 of them from our ESTs projects and 110 nucleus-encoded proteins (20,711 amino acid positions). The mitochondrial data set includes a total of 38 taxa and 13 mitochondrion-encoded proteins (2,619 amino acid positions), including homologs from the complete mitochondrial genome of *Capsaspora* that we have determined. Phylogenetic trees were inferred using a variety of methods (see table 1 and Supplementary Material online). Because phylogenetic analyses regularly suffer from systematic error such as long-branch attraction (Felsenstein 1978), we used methods and evolutionary models known to minimize these artifacts (see table 1 and Supplementary Material online). These measures include the following: 1) removing fast-evolving positions as in Ruiz-Trillo et al. (1999), 2) recoding the amino acids into functional categories as in Rodriguez-Ezpeleta et al. (2007), and 3) using a site-heterogeneous mixture (CAT) model that accounts for heterogeneity in the evolutionary

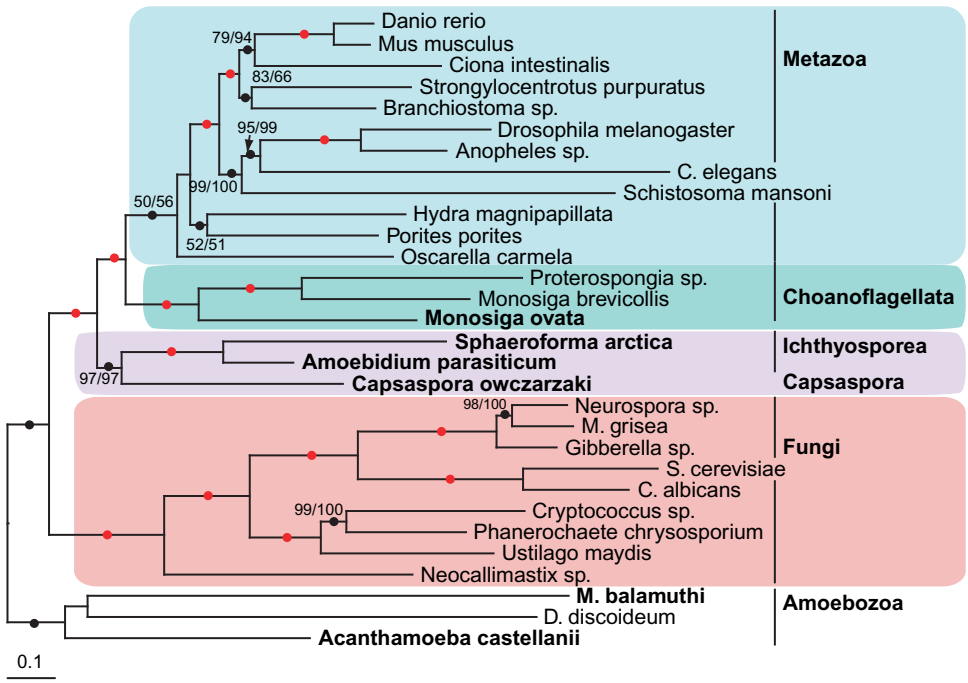

FIG. 1.—Phylogeny of the opisthokonts based on concatenation of 110 nucleus-encoded proteins. The topology and branch lengths were obtained using Bayesian analyses (PhyloBayes) with the amino acids recoded into functional categories. All branches with posterior probability values of 1.0 are marked with a filled dot (black). The dot is colored red if maximum likelihood (ML) bootstrap analysis (IQPNNI) and Bayesian (PhyloBayes) bootstrap also yields 100% support. For other relevant nodes, ML (with amino acids recoded into functional categories) bootstrap and Bayesian bootstrap support values are indicated (in %). Taxa from which new data have been obtained from an EST project are depicted in bold. See Materials and Methods for further details and supplementary table S1 (Supplementary Material online) for complete names of taxa.

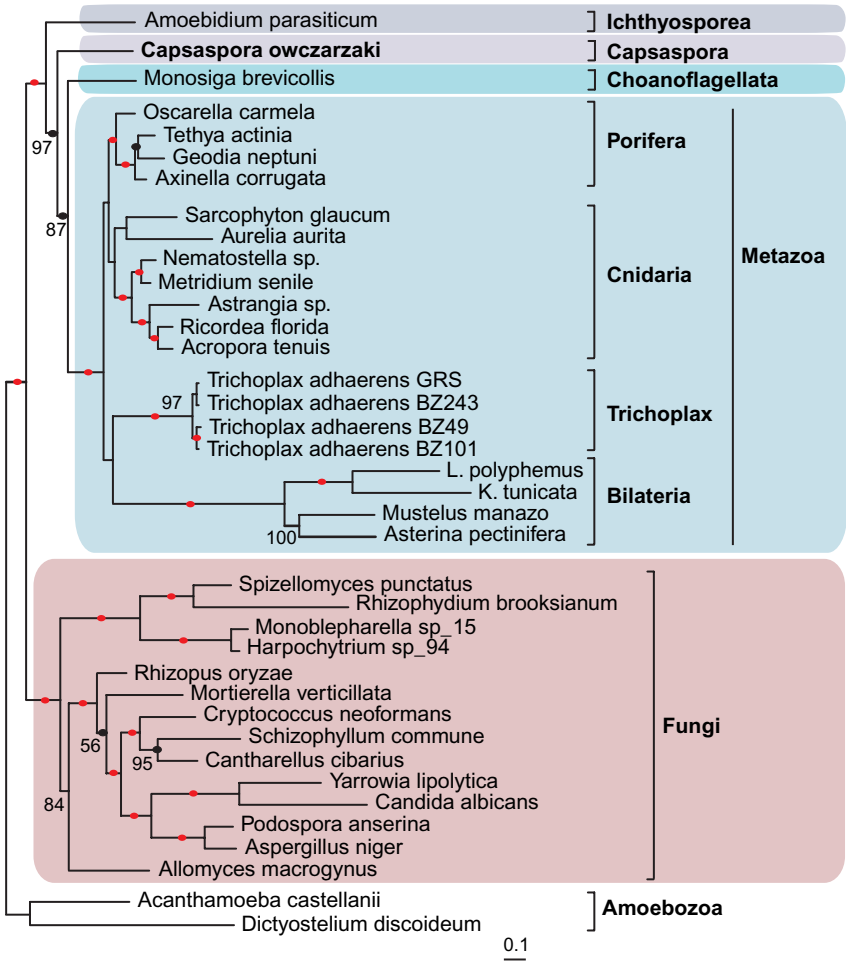

FIG. 2.—Phylogenomic analysis based on mitochondrial proteins. The alignment (2,619 amino acid positions, after trimming with Gblocks) was built from 13 protein sequences that are encoded in most mtDNAs. The topology and branch lengths were obtained in a PhyloBayes analysis. All branches with posterior support values of 1.0 are marked with a filled dot (black). The dot is colored red when, in addition, maximum likelihood (ML) bootstrap analysis with IQPNNI yields 100% support. Other ML (IQPNNI) bootstrap support values of interest are indicated. Genus abbreviations are: *L. polyphemus*, *Limulus polyphemus* and *K. tunicata*, *Katharina tunicata*.

process across sites (Lartillot and Philippe 2004; Lartillot et al. 2007; Rodriguez-Ezpeleta et al. 2007). As expected, the use of recoding or of complex models such as CAT improved the results. For example, some widely accepted metazoan clades such as Ecdysozoa or Cnidaria were recovered only when using more complex models with the nuclear data set (see table 1 and fig. 1) (Ruiz-Trillo et al. 2002; Lavrov and Lang 2005; Philippe et al. 2005; Philippe and Telford 2006; but see Wolf et al. 2004; Philip et al. 2005; Rogozin et al. 2007). Curiously, the monophyly of Metazoa has a very low bootstrap support (fig. 1). This is probably due to the effect of the missing data for both *Oscarella carmela* (70.12% missing data) and *Porites porites* (56.60% missing data; see supplementary table S1, Supplementary Material online). Consistent with this hypothesis, a tree excluding those taxa with more than 50% missing data shows a maximum likelihood bootstrap support of 100% for Metazoa (supplementary fig. S1, Supplementary Material online). An important point is that the position of *Capsaspora*, ichthyosporeans, and choanoflagellates remained identical regardless of the method (table 1). Curiously, the nuclear tree shows *Capsaspora* as the sister group to ichthyosporeans (fig. 1), whereas the mitochondrial tree shows *Capsaspora* in an intermediate position between ichthyosporeans and choanoflagellates (fig. 2). Using the SH and the ELWs tests, we found that we could statistically reject the mitochondrial topology using the nuclear data set (P values = 0.04 and 0, respectively). The reciprocal test, the nuclear tree imposed upon the mitochondrial data set was also rejected (P value = 0.04 and 0.02). The incongruity between these data sets most likely results from a phylogenetic artifact, and it is difficult to assess which topology is correct. One possible contributing factor could be the different taxonomic sampling in the mitochondrial versus the nuclear data set (the mitochondrial data set includes just one representative of each of the 3 unicellular opisthokont lineages, but a wider sampling of metazoans). However, the nuclear tree excluding taxa with more than 50% missing data not only has a similar sampling of unicellular taxa as the mitochondrial tree but also recovers ichthyosporeans and *Capsaspora* as sister groups (supplementary fig. S1, Supplementary Material online). The source of the apparent strong incongruity between these data sets remains unclear, but because the mitochondrial analysis is based on fewer aligned positions

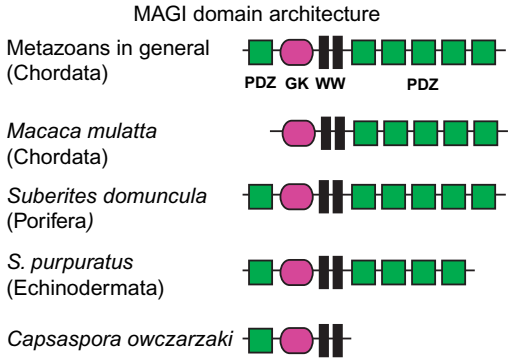

FIG. 3.—Protein domain architecture of MAGI from *Capsaspora* and Metazoa. Protein domains were identified by Blast searches against the amino acid sequences found in Prosite and NCBI databases. PDZ domains are found in many signaling proteins and seem to be important domains of scaffolding proteins. They bind either the C-terminal sequences of proteins or internal peptide sequences. Canonical GK domains catalyze the ATP-dependent phosphorylation of GMP into GDP. The WW domain, a short conserved region present in a number of unrelated proteins, binds proline-rich peptide motifs.

(2,619 vs. 20,711 amino acid positions), we favor the topology obtained with nuclear data. However, the position of *Capsaspora* should be reexamined once full genomic data become available from a wider variety of unicellular and multicellular opisthokonts.

In summary, our results show that choanoflagellates, ichthyosporeans, and *Capsaspora* are more closely related to metazoans than to fungi, confirming previous results (Ruiz-Trillo et al. 2004, 2006). Our results also demonstrate with high confidence that choanoflagellates are the closest sister group of Metazoa, to the exclusion of the other 2 groups (figs. 1 and 2). The positions of *Ministeria* and *Corallochytrium* relative to these taxa should be examined once significant genomic information from them becomes available.

Genes Involved in Multicellularity

To investigate the origin of gene families involved in multicellularity in Metazoa, we searched our protistan EST data as well as publicly available data for the occurrence of developmental and other genes relevant to multicellularity (see Supplementary Material online for details). We found that both amoebozoans and unicellular opisthokonts share with metazoans a number of genes involved in cell signaling or cell adhesion (see table 2 and Supplementary Material online). *Capsaspora* and the choanoflagellate *Monosiga* express a significantly wider range of these genes (table 2). Of note is the Hedgehog homolog of *Monosiga* (Snell et al. 2006), which so far has not been found in other nonmetazoan taxa, and the *Capsaspora* gene encoding a MAGI-like protein, that functions in the regulation of metazoan "tight junctions" (Adell et al. 2004).

Tight junctions are intracellular structures that mediate adhesion between epithelial cells. They control paracellular permeability and act as barriers to intramembrane diffusion of components. As noted above, one of the proteins known to regulate tight junctions is MAGI, a member of the MAGUK (membrane-associated guanylate kinase) family of proteins that participate in the assembly of multiprotein complexes at regions of cell–cell contact (Dobrosotskaya et al. 1997; Dobrosotskaya and James 2000). MAGUK proteins are specific to Metazoa, and MAGI has been described only in vertebrates, echinoderms, and, most recently, in sponges (Adell et al. 2004). The *Capsaspora* MAGI therefore represents the first reported occurrence of such a protein (or even of any member of the larger MAGUK family) in a nonmetazoan organism. Curiously, despite the closer relationship of choanoflagellates to Metazoa, we could not identify any MAGI homologs in choanoflagellates (or in any other eukaryotes).

Phylogenetic analyses show that the *Capsaspora* homolog is a basal member of the MAGI subgroup of the MAGUK family (see supplementary fig. S2, Supplementary Material online). MAGI (with an inverted arrangement of protein–protein interaction domains) can be distinguished from other MAGUK proteins by several specific features (Dobrosotskaya et al. 1997) (fig. 3): 1) the presence of a PDZ domain at the N-terminal end, a feature shared by other members of the MAGUK family; 2) a GK (guanylate kinase) domain near the N-terminus, in contrast to other MAGUKs; 3) 2 WW domains instead of the typical SH3 domain; and 4) 5 PDZ domains at the C-terminal end (fig. 3). Analysis of the protein architecture shows that *Capsaspora* MAGI has the first 3 of these features but lacks the C-terminal PDZ domains. Moreover, *Capsaspora* MAGI shares with all other MAGUK proteins a GK domain that conserves guanylic acid (GMP)-binding residues but lacks an ATP-binding motif (supplementary fig. S3, Supplementary Material online). Thus, the domain architecture of the *Capsaspora* MAGI homolog is unique, potentially representing a "transitional structure" between an ancestral protein and the canonical metazoan MAGI. It seems most probable that the common ancestor of *Capsaspora* and the metazoan lineage had a MAGI protein such as the one described here for *Capsaspora* (PDZ-GK-WW), which is a domain architecture not shared with any other MAGUK protein or any other gene family. We infer that, in a more recent common ancestor of extant Metazoa, the extra C-terminal PDZ signal domains were introduced, modifying the function of the protein. To better understand the selective forces at work in this scenario, more detailed functional investigations of the *Capsaspora* MAGI-like protein will be required.

We have detected the presence of other genes involved in cell adhesion in *Capsaspora*, and, in some cases, in *M. ovata*, *M. balamuthi*, and *A. castellanii* (table 2). Based on their wide distribution among eukaryotes, some of these, such as ankyrin, laminin A, or tetraspanin, appear to be genes or domains ancestrally present in unicellular eukaryotes. Their importance in shaping metazoan multicellularity likely derives from their new domain arrangements and/or domain compositions in metazoan genes, a conjecture that will require more extensive analyses when full-length gene and genomic sequences become available. Particular mention may be made of the presence of fascin in both *M. ovata* and *C. owczarzaki*. Fascin has a clear role in cell adhesion and migration (Kureishy et al. 2002) and to date has only been identified, within eukaryotes, in Metazoa and *Dictyostelium*. Interestingly, *Dictyostelium* fascin proteins

exhibit the simplest architecture (a single fascin domain), whereas in Metazoa, most fascin homologs are organized into 2, 3, 4,or 6 contiguous fascin domains (supplementary fig. S4, Supplementary Material online). Fascin proteins in both *Capsaspora* and *M. ovata* possess 4 contiguous fascin domains, probably representing an intermediate form between fascin protein architecture in *Dictyostelium* and that found in Metazoa. Elucidation of the function of the fascin domain proteins in the unicellular opisthokonts will likely be key to understanding the evolution of cell adhesion and migration in Metazoa.

Conclusions

In summary, our analyses show definitively that Ichthyosporea and *Capsaspora* diverged prior to choanoflagellates and that the latter organisms are the closest unicellular relatives of Metazoa. More importantly, our comparisons of EST and genomic data indicate that unicellular opisthokont and amoebozoan lineages possess a number of genes involved in cell signaling and cell adhesion. Some of these genes have already been described in *Dictyostelium* or fungi (table 2), some (e.g., fascin) have a unique domain organization in unicellular opisthokonts, whereas others, such as MAGI, were previously thought to be metazoan specific. Thus, several genes involved in multicellularity and development in metazoans were already present in their single-celled ancestors. More speculatively, some of these genes could have ancestrally been involved in sex, cell contact, and environmental sensing. Our EST data yield just a glimpse of the genomic composition of these organisms. We expect that additional Metazoa-specific genes will be uncovered in whole-genome sequencing initiatives such as National Human Genome Research Institute UNICellular Opisthokont Research iNitiative (Ruiz-Trillo et al. 2007), an NHGRI-endorsed multitaxon genome sequencing project that will generate genome sequences from 11 taxa at the base of Metazoa and Fungi.

Supplementary Material

Supplementary figures S1–S4 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Lise Forget and Shona Teijeiro for construction of cDNA and genomic libraries for sequencing and Emmet O'Brien for organizing the EST database, TBestDB. We thank Jacquie de Mestral, Jessica Leigh, and Maria J. Barberà for technical help; and Bjarne Landfald for sharing the *Sphaeroforma arctica* culture. This work was supported by Genome Canada (through Genome Atlantic and Génome Québec), the Atlantic Innovation Fund, and by Canadian Institutes of Health Research (CIHR) Grant MOP-62809 awarded to A.J.R. Interaction support from the Canadian Institute for Advanced Research, Program in Evolutionary Biology (G.B., B.F.L., A.J.R., and I.R.-T.), is gratefully acknowledged. I.R.-T. has been supported by European Molecular Biology Organization and CIHR postdoctoral fellowships and by an Institució Catalana de Recerca i Estudis Avançats contract.

Literature Cited

Adamska M, Degnan SM, Green KM, Adamski M, Craigie A, Larroux C, Degnan BM. 2007. Wnt and TGF-beta expression in the sponge Amphimedon queenslandica and the origin of metazoan embryonic patterning. PLoS ONE. 2:e1031.

Adell T, Gamulin V, Perovic-Ottstadt S, Wiens M, Korzhev M, Muller IM, Muller WE. 2004. Evolution of metazoan cell junction proteins: the scaffold protein MAGI and the transmembrane receptor tetraspanin in the demosponge *Suberites domuncula*. J Mol Evol. 59:41–50.

Badidi E, De Sousa C, Lang BF, Burger G. 2003. AnaBench: a Web/CORBA-based workbench for biomolecular sequence analysis. BMC Bioinformatics. 4:63.

Baldauf SL, Palmer JD. 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci USA. 90:11558–11562.

Burger G, Lavrov DV, Forget L, Lang BF. 2007. Sequencing complete mitochondrial and plastid genomes. Nat Protoc. 2:603–614.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Conway Morris S. 1998. The crucible of creation. Oxford: Oxford University Press.

Conway Morris S. 2003. The Cambrian "explosion" of metazoans and molecular biology: would Darwin be satisfied? Int J Dev Biol. 47:505–515.

Conway Morris S. 2006. Darwin's dilemma: the realities of the Cambrian "explosion". Philos Trans R Soc Lond B Biol Sci. 361:1069–1083.

Dellaporta S, Xu AL, Sagasser S, Jakob W, Moreno M, Buss LA, Schierwater BW. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports Placozoa as the basal lower metazoan phylum. Proc Natl Acad Sci USA. 103:8751–8756.

Dobrosotskaya I, Guy RK, James GL. 1997. MAGI-1, a membrane-associated guanylate kinase with a unique arrangement of protein-protein interaction domains. J Biol Chem. 272:31589–31597.

Dobrosotskaya IY, James GL. 2000. MAGI-1 interacts with beta-catenin and is associated with cell-cell adhesion structures. Biochem Biophys Res Commun. 270:903–909.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8:186–194.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27:401–410.

Finn RD, Mistry J, Schuster-Bockler B, et al. 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34:D247–D251.

Finnerty JR, Martindale MQ. 1999. Ancient origins of axial patterning genes: Hox genes and ParaHox genes in the Cnidaria. Evol Dev. 1:16–23.

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

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9:868–877.

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. 2006. The PROSITE database. Nucleic Acids Res. 34:D227–D30.

King N. 2004. The unicellular ancestry of animal development. Dev Cell. 7:313–325.
King N, Carroll SB. 2001. A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. Proc Natl Acad Sci USA. 98:15032–15037.
King N, Hittinger CT, Carroll SB. 2003. Evolution of key cell signaling and adhesion protein families predates animal origins. Science. 301:361–363.
Kureishy N, Sapountzi V, Prag S, Anilkumar N, Adams JC. 2002. Fascins, and their roles in cell structure and function. BioEssays. 24:350–361.
Kusserow A, Pang K, Sturm C, et al. 2005. Unexpected complexity of the Wnt gene family in a sea anemone. Nature. 433:156–160.
Lang BF, Burger G. 2007. Purification of mitochondrial and plastid DNA. Nat Protoc. 2:652–660.
Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G. 2002. The closest unicellular relatives of animals. Curr Biol. 12:1773–1778.
Larroux C, Fahey B, Degnan SM, Adamski M, Rokhsar DS, Degnan BM. 2007. The NK homeobox gene cluster predates the origin of Hox genes. Curr Biol. 17:706–710.
Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7(1 Suppl):S4.
Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095–1109.
Lavrov DV, Lang BF. 2005. Poriferan mtDNA and animal phylogeny based on mitochondrial gene arrangements. Syst Biol. 54:651–659.
le SV, Haeseler AV. 2004. IQPNNI: moving fast through tree space and stopping in time. Mol Biol Evol. 21:1565–1571.
Medina M, Collins AG, Taylor JW, Valentine JW, Lipps JH, Amaral-Zettler L, Sogin ML. 2003. Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. Int J Astrobiology. 2:203–211.
Mendoza L, Taylor JW, Ajello L. 2002. The class mesomycetozoea: a heterogeneous group of microorganisms at the animal-fungal boundary. Annu Rev Microbiol. 56:315–344.
Miller DJ, Ball E, Technau U. 2005. Cnidarians and ancestral genetic complexity in the animal kingdom. Trends Genet. 21:536–539.
Monteiro AS, Schierwater B, Dellaporta SL, Holland PW. 2006. A low diversity of ANTP class homeobox genes in Placozoa. Evol Dev. 8:174–182.
Moreira D, von der Heyden S, Bass D, Lopez-Garcia P, Chao E, Cavalier-Smith T. 2007. Global eukaryote phylogeny: combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. Mol Phylogenet Evol. 44:255–266.
Nichols AS, Dirks W, John SP, Nicole K. 2006. Early evolution of animal cell signaling and adhesion genes. Proc Natl Acad Sci USA. 103:1251–1256.
O'Brien EA, Koski LB, Zhang Y, Yang L, Wang E, Gray MW, Burger G, Lang BF. 2007. TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). Nucleic Acids Res. 35:D445–D451.
Philip GK, Creevey CJ, McInerney JO. 2005. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol Biol Evol. 22:1175–1184.
Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol Biol Evol. 22:1246–1253.
Philippe H, Snell EA, Bapteste E, Lopez P, Holl PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol. 21:1740–1752.
Philippe H, Telford MJ. 2006. Large-scale sequencing and the new animal phylogeny. Trends Ecol Evol. 21:614–620.
Putnam NH, Srivastava M, Hellsten U, et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science. 317:86–94.
Ragan MA, Goggin CL, Cawthorn RJ, Cerenius L, Jamieson AV, Plourde SM, Rand TG, Soderhall K, Gutell RR. 1996. A novel clade of protistan parasites near the animal-fungal divergence. Proc Natl Acad Sci USA. 93:11907–11912.
Ragan MA, Murphy CA, Rand TG. 2003. Are Ichthyosporea animals or fungi? Bayesian phylogenetic analysis of elongation factor 1alpha of *Ichthyophonus irregularis*. Mol Phylogenet Evol. 29:550–562.
Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. 56:389–399.
Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. Mol Biol Evol. 24:1080–1090.
Ruiz-Trillo I, Burger G, Holland PW, King N, Lang BF, Roger AJ, Gray MW. 2007. The origins of multicellularity: a multi-taxon genome initiative. Trends Genet. 23:113–118.
Ruiz-Trillo I, Inagaki Y, Davis LA, Sperstad S, Landfald B, Roger AJ. 2004. *Capsaspora owczarzaki* is an independent opisthokont lineage. Curr Biol. 14(22):R946–R947.
Ruiz-Trillo I, Lane CE, Archibald JM, Roger AJ. 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. J Eukaryot Microbiol. 53:1–6.
Ruiz-Trillo I, Paps J, Loukota M, Ribera C, Jondelius U, Baguna J, Riutort M. 2002. A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proc Natl Acad Sci USA. 99:11246–11251.
Ruiz-Trillo I, Riutort M, Littlewood DT, Herniou EA, Baguna J. 1999. Acoel flatworms: earliest extant bilaterian Metazoans, not members of Platyhelminthes. Science. 283:1919–1923.
Ryan JF, Burton PM, Mazza ME, Kwong GK, Mullikin JC, Finnerty JR. 2006. The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes. Evidence from the starlet sea anemone, Nematostella vectensis. Genome Biol. 7:R64.
Ryan JF, Mazza ME, Pang K, Matus DQ, Baxevanis AD, Martindale MQ, Finnerty JR. 2007. Pre-bilaterian origins of the Hox cluster and the Hox code: evidence from the sea anemone, Nematostella vectensis. PLoS ONE. 2:e153.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18:502–504.
Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.
Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM. 1994. The genetic data environment an expandable GUI for multiple sequence analysis. Comput Appl Biosci. 10:671–675.
Snell EA, Brooke NM, Taylor WR, Casane D, Philippe H, Holland PW. 2006. An unusual choanoflagellate protein released by Hedgehog autocatalytic processing. Proc R Soc Lond B Biol Sci. 273:401–407.
Snell EA, Furlong RF, Holland PW. 2001. Hsp70 sequences indicate that choanoflagellates are closely related to animals. Curr Biol. 11:967–970.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688–2690.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 21:456–463.
Steenkamp ET, Baldauf SL. 2004. Origin and evolution of animals, fungi and their unicellular allies (Opisthokonta). In: Hirt RP, Homer DS, editors. Boca Raton (FL): CRC Press. p. 109–129.
Steenkamp ET, Wright J, Baldauf SL. 2006. The protistan origins of animals and fungi. Mol Biol Evol. 23:93–106.
Strimmer K, Rambaut A. 2002. Inferring confidence sets of possibly misspecified gene trees. Proc R Soc Lond B Biol Sci. 269:137–142.
Sullivan JC, Ryan JF, Mullikin JC, Finnerty JR. 2007. Conserved and novel Wnt clusters in the basal eumetazoan Nematostella vectensis. Dev Genes Evol. 217:235–239.
Technau U, Rudd S, Maxwell P, et al. 2005. Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians. Trends Genet. 21:633–639.
Valentine JW. 2004. On the origin of phyla. Chicago: The University of Chicago Press.
Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14:29–36.

Laura Katz, Associate Editor

Accepted December 29, 2007