Vol 451|14 February 2008|doi:10.1038/nature06617

The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans

Nicole King[1,2], M. Jody Westbrook[1]*, Susan L. Young[1]*, Alan Kuo[3], Monika Abedin[1], Jarrod Chapman[1], Stephen Fairclough[1], Uffe Hellsten[3], Yoh Isogai[1], Ivica Letunic[4], Michael Marr[5], David Pincus[6], Nicholas Putnam[1], Antonis Rokas[7], Kevin J. Wright[1], Richard Zuzow[1], William Dirks[1], Matthew Good[6], David Goodstein[1], Derek Lemons[8], Wanqing Li[9], Jessica B. Lyons[1], Andrea Morris[10], Scott Nichols[1], Daniel J. Richter[1], Asaf Salamov[3], JGI Sequencing[3], Peer Bork[4], Wendell A. Lim[6], Gerard Manning[11], W. Todd Miller[9], William McGinnis[8], Harris Shapiro[3], Robert Tjian[1], Igor V. Grigoriev[3] & Daniel Rokhsar[1,3]

Choanoflagellates are the closest known relatives of metazoans. To discover potential molecular mechanisms underlying the evolution of metazoan multicellularity, we sequenced and analysed the genome of the unicellular choanoflagellate *Monosiga brevicollis*. The genome contains approximately 9,200 intron-rich genes, including a number that encode cell adhesion and signalling protein domains that are otherwise restricted to metazoans. Here we show that the physical linkages among protein domains often differ between *M. brevicollis* and metazoans, suggesting that abundant domain shuffling followed the separation of the choanoflagellate and metazoan lineages. The completion of the *M. brevicollis* genome allows us to reconstruct with increasing resolution the genomic changes that accompanied the origin of metazoans.

Choanoflagellates have long fascinated evolutionary biologists for their marked similarity to the 'feeding cells' (choanocytes) of sponges and the possibility that they might represent the closest living relatives of metazoans[1,2]. Over the past decade or so, evidence supporting this relationship has accumulated from phylogenetic analyses of nuclear and mitochondrial genes[3–6], comparative genomics between the mitochondrial genomes of choanoflagellates, sponges and other metazoans[7,8], and the finding that choanoflagellates express homologues of metazoan signalling and adhesion genes[9–12]. Furthermore, species-rich phylogenetic analyses demonstrate that choanoflagellates are not derived from metazoans, but instead represent a distinct lineage that evolved before the origin and diversification of metazoans (Fig. 1a, Supplementary Fig. 1 and Supplementary Note 3.1)[8,13]. By virtue of their position on the tree of life, studies of choanoflagellates provide an unparallelled window into the nature of the unicellular and colonial progenitors of metazoans[14].

Choanoflagellates are abundant and globally distributed microbial eukaryotes found in marine and freshwater environments[15,16]. Like sponge choanocytes, each cell bears an apical flagellum surrounded by a distinctive collar of actin-filled microvilli, with which choanoflagellates trap bacteria and detritus (Fig. 1b). Using this highly effective means of prey capture, choanoflagellates link bacteria to higher trophic levels and thus have critical roles in oceanic carbon cycling and in the microbial food web[17,18].

More than 125 choanoflagellate species have been identified, and all species have a unicellular life-history stage. Some can also form simple colonies of equipotent cells, although these differ substantially from the obligate associations of differentiated cells in metazoans[19]. Studies of basal metazoans indicate that the ancestral metazoan was multicellular and had differentiated cell types, an epithelium, a body plan and regulated development including gastrulation. In contrast, the last common ancestor of choanoflagellates and metazoans was unicellular or possibly capable of forming simple colonies, underscoring the abundant biological innovation that accompanied metazoan origins.

Despite their evolutionary and ecological importance, little is known about the genetics and cell biology of choanoflagellates. To gain insight into the biology of choanoflagellates and to reconstruct the genomic changes attendant on the early evolution of metazoans, we sequenced the genome of the choanoflagellate *M. brevicollis* and compared it with genomes from metazoans and other eukaryotes.

Gene structure and intron evolution

The ~41.6 megabase (Mb) *M. brevicollis* genome contains approximately 9,200 genes (Supplementary Notes 1 and 2) and is comparable in size to the genomes of filamentous fungi (~30–40 Mb) and other free-living unicellular eukaryotes (for example, small diatoms at ~20–35 Mb[20] and ichthyosporeans at ~20–25 Mb[21]). Metazoan genomes are typically larger, with few exceptions[22].

M. brevicollis genes have several distinguishing structural features (Table 1). Whereas the *M. brevicollis* genome is compact, its genes are almost as intron-rich as human genes (6.6 introns per *M. brevicollis* gene versus 7.7 introns per human gene). *M. brevicollis* introns are short (averaging 174 bp) relative to metazoan introns, and with few exceptions do not include the extremely long introns found in some metazoan genes (Supplementary Fig. 2 and Supplementary Note 3.3).

Comparisons of intron positions in a set of conserved genes from *M. brevicollis*, diverse metazoans and representative intron-rich fungi, plants and a ciliate reveal that the last common ancestor of

[1]Department of Molecular and Cell Biology and the Center for Integrative Genomics, [2]Department of Integrative Biology, University of California, Berkeley, California 94720, USA. [3]Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA. [4]EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany. [5]Department of Biology, Brandeis University, Waltham, Massachusetts 02454, USA. [6]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, California 94158, USA. [7]Vanderbilt University, Department of Biological Sciences, Nashville, Tennessee 37235, USA. [8]Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093, USA. [9]Department of Physiology and Biophysics, Stony Brook University, Stony Brook, New York 11794, USA. [10]University of Michigan, Department of Cellular and Molecular Biology, Ann Arbor, Michigan 48109, USA. [11]Razavi Newman Bioinformatics Center, Salk Institute for Biological Studies, La Jolla, California 92037, USA.
*These authors contributed equally to this work.

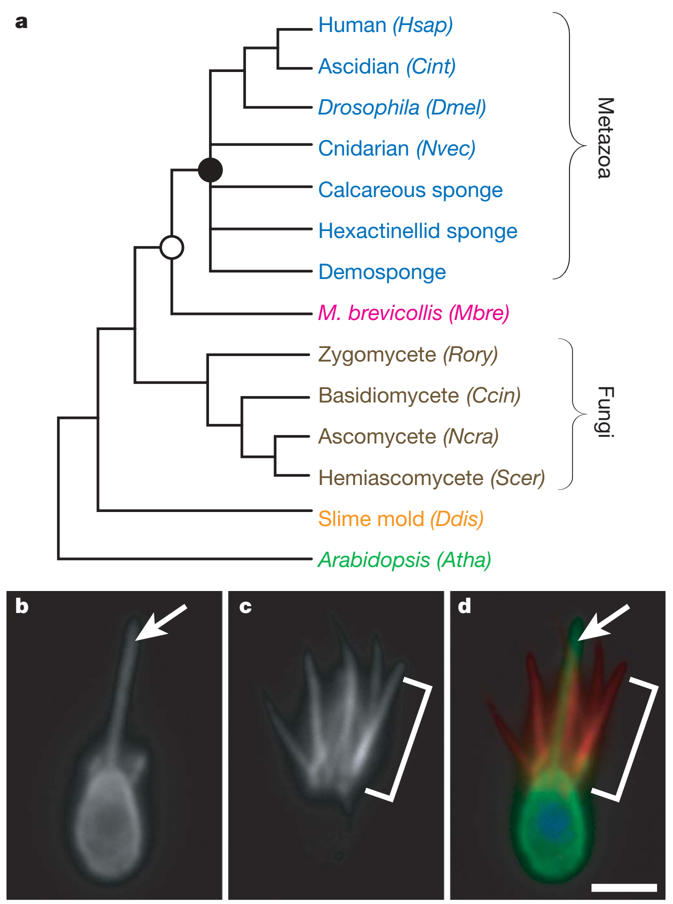


Figure 1 | Introduction to the choanoflagellate *Monosiga brevicollis*. a, The close phylogenetic affinity between choanoflagellates and metazoans highlights the value of the *M. brevicollis* genome for investigations into metazoan origins, the biology of the last common ancestor of metazoans (filled circle) and the biology of the last common ancestor of choanoflagellates and metazoans (open circle). Genomes from species shown with their abbreviation were used for protein domain comparisons in this study: human (*Homo sapiens*, *Hsap*), ascidian (*Ciona intestinalis*, *Cint*), *Drosophila* (*Drosophila melanogaster*, *Dmel*), cnidarian (*N. vectensis*, *Nvec*), *M. brevicollis* (*Mbre*), zygomycete (*Rhizopus oryzae*, *Rory*), basidiomycete (*Coprinus cinereus*, *Ccin*), ascomycete (*Neurospora crassa*, *Ncra*), hemiascomycete (*Saccharomyces cerevisiae*, *Scer*), slime mould (*Dictyostelium discoideum*, *Ddis*) and *Arabidopsis* (*Arabidopsis thaliana*, *Atha*). **b**–**d**, Choanoflagellate cells bear a single apical flagellum (arrow, **b**) and an apical collar of actin-filled microvilli (bracket, **c**). **d**, An overlay of β-tubulin (green), polymerized actin (red) and DNA localization (blue) reveals the position of the flagellum within the collar of microvilli. Scale bar, 2 μm.

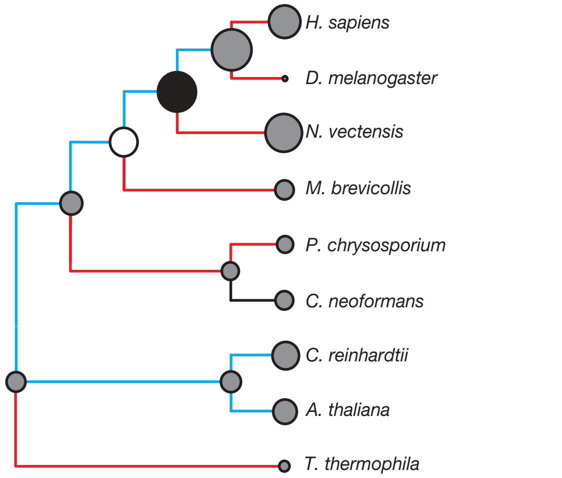


Figure 2 | Intron gain preceded the origin and diversification of metazoans. Ancestral intron content, intron gains and intron losses were inferred by the Csuros maximum likelihood method[45] from a sample of 1,054 intron positions in 473 highly conserved genes in representative metazoans (humans, *D. melanogaster* and *N. vectensis*), *M. brevicollis*, intron-rich fungi (*Cryptococcus neoformans A* and *Phanerochaete chrysosporium*), plants and green algae (*A. thaliana* and *Chlamydomonas reinhardtii*), and a ciliate (*Tetrahymena thermophila*). Branches with more gain than loss are blue, those with more loss than gain are red, and those with comparable amounts of each are black. The inferred or observed number of introns present in ancestral and extant species is indicated by proportionally sized circles. As in Fig. 1, the last common ancestor of metazoans and the last common ancestor of choanoflagellates and metazoans are represented by a filled circle and an open circle, respectively. Other ancestral nodes are indicated by grey circles.

choanoflagellates and metazoans had genes at least as intron-rich as those of modern choanoflagellates (Fig. 2, Supplementary Figs 3 and 4, and Supplementary Note 3.3). Notably, these analyses reveal that the eumetazoan ancestor contained a substantially higher density of introns than the last common ancestor of choanoflagellates and metazoans. This is consistent with a proliferation of introns during the early evolution of the Metazoa[23].

Premetazoan history of protein domains required for multicellularity

The *M. brevicollis* genome provides unprecedented insight into the early evolution of metazoan genes. Annotation of the *M. brevicollis* genome using Pfam and SMART (two protein domain prediction databases) reveals 78 protein domains that are exclusive to choanoflagellates and metazoans, only two of which have been reported previously in choanoflagellates (Supplementary Table 4)[10]. Because genomic features shared by *M. brevicollis* and metazoans were probably present in their last common ancestor, this study extends the evolutionary history of a cohort of important protein domains to the premetazoan era. Many of these domains are central to cell signalling and adhesion processes in metazoans, suggesting a role in the origin of multicellularity. In contrast, metazoan genomic features that are missing from the *M. brevicollis* genome may have evolved within the metazoan lineage, or may have existed in the last common ancestor with choanoflagellates and were subsequently lost on the stem leading to *M. brevicollis*. Presumably, there are many genomic features that evolved in the metazoan lineage, and the *M. brevicollis* genome provides our first glimpse at the complement of genes and protein domains that predate metazoan origins.

To investigate further the extent to which molecular components required for metazoan multicellularity evolved before the origin of

Table 1 | *M. brevicollis* genome properties in a phylogenetic context

	Metazoa				Choanoflagellates	Fungi		*Dictyostelium*	Plants
	Hsap	*Cint*	*Dmel*	*Nvec*	*Mbre*	*Ccin*	*Ncra*	*Ddis*	*Atha*
Genome size (Mb)	2,900	160	180	357	42	38	39	34	125
Total number of genes	23,224	14,182	14,601	18,000	9,196	13,544	9,826	13,607	27,273
Mean gene size (bp)	27,000	4,585	5,247	6,264	3,004	1,679	1,528	1,756	2,287
Mean intron density (introns per gene)	7.7	6.8	4.9	5.8	6.6	4.4	1.8	1.9	4.4
Mean intron length (bp)	3,365	477	1,192	903	174	75	136	146	164
Gene density (kb per gene)	127.9	11.9	13.2	19.8	4.5	2.7	4.0	2.5	4.5

Species names follow the four-letter convention from Fig. 1.

metazoans, we performed targeted searches in the *M. brevicollis* genome and representative metazoan, fungal and plant genomes for homologues of critical metazoan cell adhesion, cell signalling and transcription factor protein families.

An abundance of cell adhesion domains

A critical step in the transition to multicellularity was the evolution of mechanisms for stable cell adhesion. *M. brevicollis* encodes a diverse array of cell adhesion and extracellular matrix (ECM) protein domains previously thought to be restricted to metazoans (Fig. 3). At least 23 *M. brevicollis* genes encode one or more cadherin domains, homologues of which are required for cell sorting and adhesion during metazoan embryogenesis[24], and 12 genes encode C-type lectins, 2 of which are transmembrane proteins. Whereas soluble C-type lectins have functions ranging from pathogen recognition to ECM organization, transmembrane C-type lectins mediate specific adhesive activities such as contact between leukocytes and vascular endothelial cells, cell recognition, and molecular uptake by endocytosis[25–27].

The genome of *M. brevicollis* also contains integrin-α and immunoglobulin domains—cell adhesion domains formerly thought to be restricted to metazoans. In metazoans, integrin-α- and integrin-β-domain-containing proteins heterodimerize before binding to ECM proteins such as collagen[28]. We find that *M. brevicollis* has at least 17 integrin-α-domain-containing proteins, but no integrin-β domains. Metazoan immunoglobulin-domain-containing proteins have both adhesive and immune functions. The *M. brevicollis* genome encodes a total of five immunoglobulin domains that show affinity for the I-set, V-set or C2-set subfamilies, but not the vertebrate-specific C1-set subfamily. In contrast to *M. brevicollis*, metazoan genomes possess from ~150 to ~1,500 immunoglobulin domains (Supplementary Table 7), suggesting that the radiation of the immunoglobulin superfamily occurred after the divergence of choanoflagellates and metazoans.

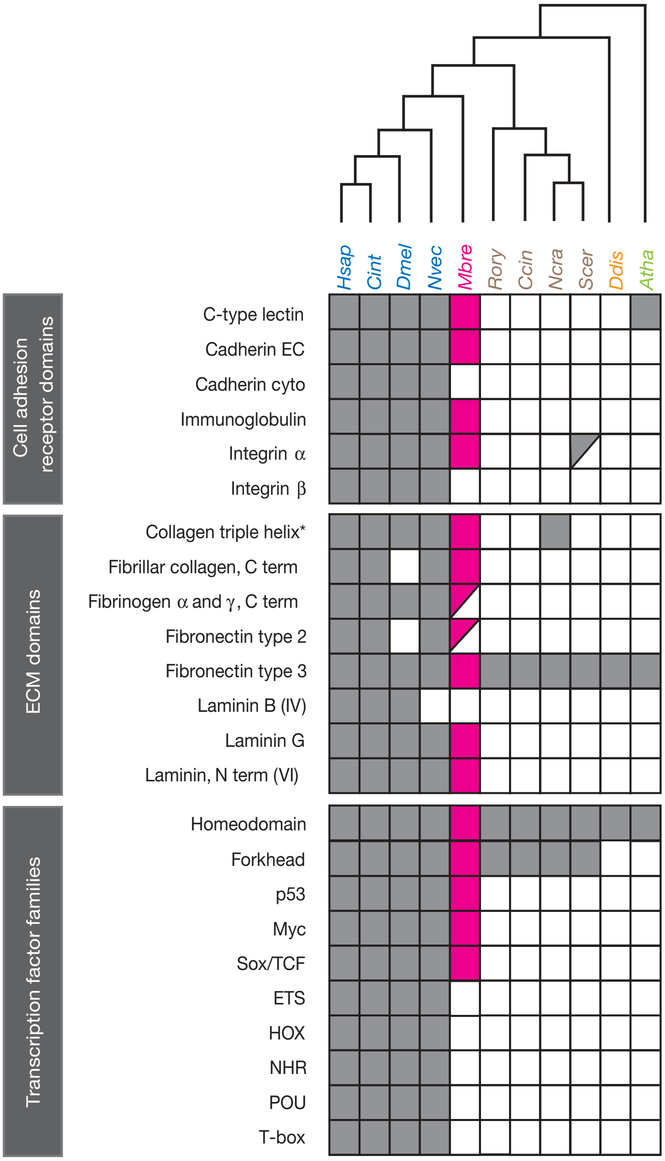
The finding in *M. brevicollis* of cell adhesion domains that were previously known only in metazoans has two important implications. First, the common ancestor of metazoans and choanoflagellates possessed several of the critical structural components used for multicellularity in modern metazoans. Second, given the absence of evidence for stable cell adhesion in *M. brevicollis*, this also suggests that homologues of metazoan cell adhesion domains may act to mediate interactions between *M. brevicollis* and its extracellular environment.

Extracellular-matrix-associated protein domains

As the targets of many adhesion receptors, the question of whether metazoan-type ECM proteins and domains evolved before or after the transition to multicellularity is of great interest. In metazoans, collagens are ECM proteins that polymerize to form a major component of the basement membrane of epithelia and have been invoked as a potential 'key innovation' during the transition to multicellularity[29]. We find five collagen-domain-encoding genes in the *M. brevicollis* genome, two of which encode the diagnostic Gly-X-Y repetitive sequence motif (where X and Y are frequently proline and hydroxyproline, respectively) in an arrangement similar to metazoan collagens[30]. Other ECM-associated domains previously known only from metazoans that occur in *M. brevicollis* include laminin domains (an important class that contributes to the basement membrane), the reeler domain (found in the neuronal ECM protein reelin[31]) and the ependymin domain (an extracellular glycoprotein found in cerebrospinal fluid[32]; Fig. 3 and Supplementary Table 4).

The discovery of putatively secreted ECM proteins in a free-living choanoflagellate suggests that elements of the metazoan ECM evolved in contact with the external environment before being sequestered within an epithelium. Although some choanoflagellates secrete extracellular structures or adhere to form colonial assemblages[19,33,34], *M. brevicollis* is not known to do so. Instead, these ECM protein homologues in *M. brevicollis* may mediate an analogous process such as substrate attachment.

Against the backdrop of abundant conservation of cell adhesion and ECM protein domains among the genomes of *M. brevicollis* and metazoans, it is important to note the differences. Individual cell adhesion and ECM-associated domains in the *M. brevicollis* genome often occur in unique arrangements, and clear orthologues of specific metazoan adhesion proteins are rarely found. Although the domains associated with metazoan adhesion and ECM proteins were present in the ancestor of choanoflagellates and metazoans, the canonical metazoan adhesion protein architectures[35] probably evolved after the divergence of the two lineages.



		Hsap	*Cint*	*Dmel*	*Nvec*	*Mbre*	*Rory*	*Ccin*	*Ncra*	*Scer*	*Ddis*	*Atha*
Cell adhesion receptor domains	C-type lectin	■	■	■	■	■	□	□	□	□	□	■
	Cadherin EC	■	■	■	■	■	□	□	□	□	□	□
	Cadherin cyto	■	■	■	■	□	□	□	□	□	□	□
	Immunoglobulin	■	■	■	■	■	□	□	□	□	□	□
	Integrin α	■	■	■	■	■	□	□	□	◩	□	□
	Integrin β	■	■	■	■	□	□	□	□	□	□	□
ECM domains	Collagen triple helix*	■	■	■	■	■	□	□	■	□	□	□
	Fibrillar collagen, C term	■	■	□	■	■	□	□	□	□	□	□
	Fibrinogen α and γ, C term	■	■	■	■	◩	□	□	□	□	□	□
	Fibronectin type 2	■	■	□	■	◩	□	□	□	□	□	□
	Fibronectin type 3	■	■	■	■	■	■	■	■	■	■	■
	Laminin B (IV)	■	■	■	□	□	□	□	□	□	□	□
	Laminin G	■	■	■	■	■	□	□	□	□	□	□
	Laminin, N term (VI)	■	■	■	■	■	□	□	□	□	□	□
Transcription factor families	Homeodomain	■	■	■	■	■	■	■	■	■	■	■
	Forkhead	■	■	■	■	■	■	■	■	■	□	□
	p53	■	■	■	■	■	□	□	□	□	□	□
	Myc	■	■	■	■	■	□	□	□	□	□	□
	Sox/TCF	■	■	■	■	■	□	□	□	□	□	□
	ETS	■	■	■	■	□	□	□	□	□	□	□
	HOX	■	■	■	■	□	□	□	□	□	□	□
	NHR	■	■	■	■	□	□	□	□	□	□	□
	POU	■	■	■	■	□	□	□	□	□	□	□
	T-box	■	■	■	■	□	□	□	□	□	□	□

Figure 3 | Phylogenetic distribution of metazoan-type cell adhesion domains and sequence-specific transcription factor families. *M. brevicollis* possesses diverse adhesion and ECM domains previously thought to be unique to metazoans (magenta). In contrast, many metazoan sequence-specific transcription factors are absent from the *M. brevicollis* gene catalogue. For adhesion and ECM domains, a filled box indicates a domain identified by both SMART and Pfam[37,48], a half-filled box indicates a domain identified by either SMART or Pfam, and an open box indicates a domain that is not encoded by the current set of gene models. The presence (filled box) or absence (empty box) of transcription factor families was determined by reciprocal BLAST and SMART/Pfam domain annotations (Supplementary Note 3.5). Species names follow the convention from Fig. 1. EC, extracellular domain; cyto, cytoplasmic domain; asterisk, collagen triple-helix-domains occur in the extended tandem arrays diagnostic of collagen proteins found only in metazoans and choanoflagellates.

Domain shuffling in the evolution of intercellular signalling pathways

Our analysis of the *M. brevicollis* genome reveals little evidence that metazoan-specific signalling pathways were present in the last common ancestor of choanoflagellates and metazoans. Many pathways are missing entirely, and *M. brevicollis* genes with some similarity to metazoan signalling machinery are largely found to share conserved domains without aligning across the full span of what are often complex multidomain proteins (for example, epidermal growth factor (EGF) repeats are common to Notch, and also to many other proteins; Supplementary Table 8). Specifically, no receptors or ligands were identified from the NHR (nuclear hormone receptor), WNT and TGF-β signalling pathways. The only evidence of the JAK (Janus kinase)/STAT (signal transducer and activator of transcription) pathway is an apparent *STAT*-like gene that encodes a STAT DNA-binding domain and a partial SH2 domain. Convincing evidence is also lacking for the Toll signalling pathway—a signalling system important both for development and for innate immunity in metazoans.

Nonetheless, the genome of *M. brevicollis* does provide insights into the evolution of Notch and hedgehog (Hh) signalling pathways. Cassettes of protein domains found in metazoan Notch receptors (EGF, NL and ANK (ankyrin repeats)) are encoded on separate *M. brevicollis* genes in arrangements that differ from metazoan Notch proteins, and definitive domains, such as the NOD domain and the MNNL region, are absent (Fig. 4a).

Homologues of hedgehog, dispatched and patched genes are also present; however, there is no evidence for smoothened nor for its defining frizzled domain. In metazoans, hedgehog consists of an amino-terminal signalling domain and carboxy-terminal hedgehog/intein (Hint) domain responsible for autocatalytically cleaving the protein. In one *M. brevicollis* hedgehog-like protein, a hedgehog N-terminal signalling domain is found at the N terminus of a large transmembrane protein that, instead of a Hint domain, includes von Willebrand A, cadherin, TNFR (tumor necrosis factor receptor), furin and EGF domains. Similar proteins are found in the sponge *Amphimedon queenslandica* and in the cnidarian *Nematostella vectensis*[36], revealing that the *M. brevicollis* genome captures an ancestral arrangement of protein domains rather than representing a lineage-specific domain-shuffling event. Another *M. brevicollis* hedgehog-like protein contains a Hint domain—a key region involved in the autocatalytic processing of hedgehog (Fig. 4b). The identification of a hedgehog-like gene in a choanoflagellate is not without precedent. A distinct Hint-domain-containing protein, named Hoglet, was identified in the distantly related *Monosiga ovata*[12], supporting the idea that isolated signalling components were present in the last common ancestor of choanoflagellates and metazoans.

Divergent use of phosphotyrosine signalling machinery

Phosphotyrosine (pTyr)-based signalling was considered unique to metazoans until its recent discovery in choanoflagellates[9,11]. The key domains involved in pTyr signalling are found in abundance in the *M. brevicollis* genome: tyrosine kinase domains that phosphorylate tyrosine (~120 occurrences), pTyr-specific phosphatases (PTP) that remove the phosphate modification (~30) and SH2 domains that bind pTyr-containing peptides (~80) (Supplementary Fig. 7). In contrast, these domains are rare in non-metazoans; for example, *S. cerevisiae* has no tyrosine kinase domains, only three PTP domains and a single SH2 domain. These findings support a model in which the full set of pTyr signalling machinery evolved before the separation of the choanoflagellate and metazoan lineages.

Although pTyr signalling machinery is present in metazoans and choanoflagellates, the mode of usage in *M. brevicollis* may be distinct from that in metazoans. A simple metric for the use of a particular domain is the range of domain types with which it is found in combination[37]. In the *M. brevicollis* genome, more than half of the observed pairwise domain combinations involving tyrosine kinase, PTP and SH2 domains are distinct from those seen in any metazoan genome (Fig. 5 and Supplementary Note 3.7). In contrast, for other sets of common signalling domains (those involved in phosphoserine/threonine (pSer/Thr), Ras–GTP and Rho–GTP signalling), most observed combinations are shared between *M. brevicollis* and metazoans. These observations are consistent with a simple model in which pSer/Thr, Ras–GTP and Rho–GTP signalling were fully elaborated before the branching of the choanoflagellate and metazoan lineages (consistent with the presence of these systems in other eukaryotes, including fungi, *Dictyostelium* and plants). In contrast, simple pTyr signalling may have emerged in the common ancestor and diverged radically between choanoflagellates and metazoans.

Streamlined transcriptional regulation

The core transcriptional apparatus of *M. brevicollis* is, in many ways, typical of most eukaryotes examined to date (Supplementary Table 10) including, for example, all 12 RNA polymerase II subunits and most of the transcription elongation factors (TFIIS, NELF, PAF, DSIF and P-TEFb, but not elongin). However, homologues of the largest subunit of TFIIF and several subunits of TFIIH are apparently lacking from the genome and the expressed-sequence-tag collection (Supplementary Fig. 8), reminiscent of the absence of several basal factors from the *Giardia lamblia* genome, suggesting alternative strategies for interacting with core promoter elements[38]. Similarly, only a limited number of general co-activators are identifiable in *M. brevicollis*, including the components of several chromatin-remodelling complexes (Supplementary Fig. 9 and Supplementary Note 3.8).

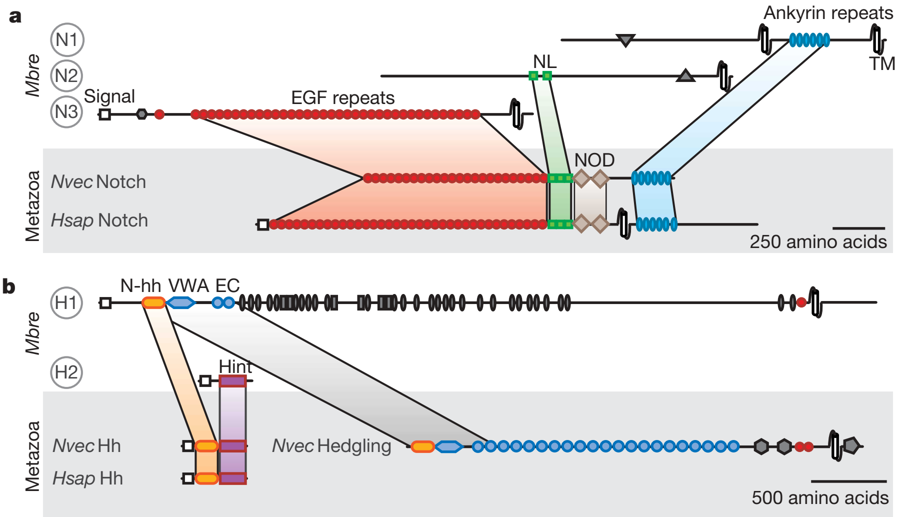


Figure 4 | Domain shuffling and the evolution of Notch and hedgehog. Analysis of the draft gene set reveals that *M. brevicollis* possesses proteins containing domains characteristic of metazoan Notch (**a**, N1–N3) and hedgehog (**b**, H1 and H2). Some of these protein domains were previously thought to be unique to metazoans. The presence of these domains in separate *M. brevicollis* proteins implicates domain shuffling in the evolution of Notch and Hedgehog. Hh, hedgehog; N-hh, hedgehog N-terminal signalling domain; Hint, hedgehog/intein domain; TM, transmembrane domain; VWA, von Willebrand A domain. See Supplementary Note 3.6 for protein accession numbers and Supplementary Fig. 6 for identification of all displayed protein domains. Species names follow the convention from Fig. 1.

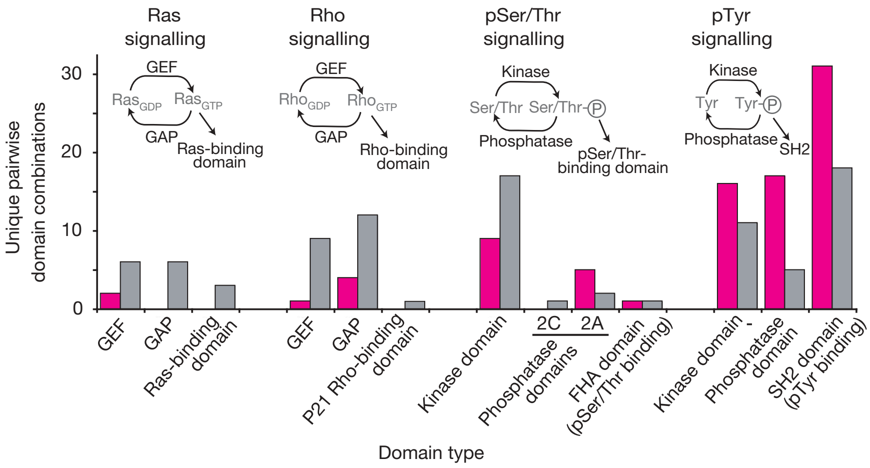


Figure 5 | Divergent usage of protein domains involved in pTyr-based signalling between *M. brevicollis* and metazoans. A metric for functional usage of a domain within a genome is the number of other domains with which it co-occurs in a single protein. Numbers of pairwise domain combinations are indicated for classes of signalling domains involved in Ras, Rho, pSer/Thr and pTyr signalling. In cases in which a domain combination occurs multiple times within an individual protein or genome, it is only counted once. All combinations observed in *M. brevicollis* are indicated either as those that are only observed in the *M. brevicollis* genome (magenta) or as those that are observed both in *M. brevicollis* and metazoan genomes (grey). pTyr signalling domains in *M. brevicollis* are unique in that most of their observed pairwise domain combinations are distinct from those observed in metazoans. GEF, guanine-nucleotide exchange factor; GAP, GTPase-activating protein.

Perhaps not surprisingly, *M. brevicollis* possesses members from most of the ubiquitous families of eukaryotic transcription factors (Supplementary Fig. 10). Most of the predicted transcription factors are zinc-coordinating; approximately 44% are C2H2-type zinc fingers. Eight proteins (5% of a total of 155 predicted transcription factors) are forkhead transcription factors, otherwise known only from metazoans and fungi.

The homeodomain transcription factors are an ancient protein family found in all known eukaryotes. At least two major superclasses of homeodomain proteins evolved before the origin of metazoans: those containing homeodomains of 60 amino acids (the 'typical', or non-TALE superclass), and those containing homeodomains of more than 63 amino acids (the TALE superclass)[39]. The *M. brevicollis* genome encodes only two homeodomain proteins, both of which group with the MEIS sub-class of TALE homeodomains (Supplementary Fig. 12). Apparently, genes encoding non-TALE homeodomain proteins have been lost in the lineage leading to *M. brevicollis*. *Bona fide* HOX class homeobox genes—a subclass of the non-TALE superclass—are absent from both *M. brevicollis* and the *Amphimedon queenslandica* (demosponge) genome sequence reads, indicating that this characteristic metazoan gene family probably emerged along the stem leading to eumetazoans[40].

M. brevicollis contains a subset of the transcription factor families previously thought to be specific to metazoans. Members of the metazoan p53, Myc and Sox/TCF families were identified, whereas many transcription factor families associated with metazoan patterning and development (ETS, HOX, NHR, POU and T-box) seem to be absent (Fig. 3).

Discussion

Choanoflagellates, sponges and other metazoans last shared a unicellular common ancestor in the late Precambrian period, more than 600 million years ago[41,42]. Although the origin of metazoans was a pivotal event in life's history, little is known about the genetic underpinnings of the requisite transition to multicellularity. Comparisons of modern genomes provide our most direct insights into the ancient genomic conditions from which metazoans emerged. By comparing choanoflagellate and metazoan genomes, we infer that their common ancestor had intron-rich genes, some of which encoded protein domains characteristically associated with cell adhesion and the ECM in animals.

In addition to containing protein domains associated with metazoan cell adhesion, *M. brevicollis* possesses a surprising abundance of tyrosine kinases and their downstream signalling targets. In contrast, components of most other intercellular signalling pathways, as well as many of the diverse transcription factors that comprise the developmental toolkit of modern animals, are absent. These presumably reached their modern form on the metazoan stem, although it is formally possible that they were in place much earlier and degenerated in the *M. brevicollis* lineage. Likewise, it is possible that the last common ancestor of choanoflagellates and metazoans had an early form of multicellularity that became more robust in metazoans and was lost in the choanoflagellate lineage. In any event, the evolutionary distance between choanoflagellates and metazoans is substantial, and evidently few, if any, intermediate lineages survived. There are, for example, no other known microbial eukaryotes that possess any of the eight developmental signalling pathways characteristic of metazoans.

The mechanism of invention of new genes on the metazoan stem, and their integration to create the emergent network of cell signalling and transcriptional regulation fundamental to metazoan biology, remains mysterious. Domain shuffling, which has frequently been proposed as an important mechanism for the evolution of metazoan multidomain proteins[43,44], is implicated by the presence of essential metazoan signalling domains in *M. brevicollis* that appear in unique combinations relative to animals. For pTyr-based signalling in particular, the marked divergence of domain combinations suggests that this mode of cellular interaction existed in a nascent form in the common choanoflagellate–metazoan ancestor, and was subsequently specialized and elaborated on in each lineage.

Given the limited transcription factor diversity in *M. brevicollis*, it is notable that the genome encodes representatives of the otherwise metazoan-specific p53, Myc and Sox/TCF transcription factor families. These transcription factors may have had early and critical roles in the evolution of metazoan ancestors by regulating the differential expression of genes to allow multiple cell types to exist in a single organism, and their study in choanoflagellates is a promising future direction.

The *M. brevicollis* sequence opens the door to genome-enabled studies of choanoflagellates, a diverse group of microbial eukaryotes that are important in their own right as bacterial predators in both marine and freshwater ecosystems. Although *M. brevicollis* is strictly unicellular, other choanoflagellates facultatively form colonies, and the modulation of these associations by cell signalling, adhesion, transcriptional regulation and environmental influences is poorly understood. An integrative approach that unites studies of choanoflagellate genomes, cell biology and ecology with the biogeochemistry of the Precambrian promises to reveal the intrinsic and extrinsic factors underlying metazoan origins.

METHODS SUMMARY

All analyses described were performed on Version 1.0 of the genome sequence. Details can be found in the Supplementary Information.

Separation of choanoflagellate and bacterial DNA. Using physical separation techniques combined with antibiotic treatments, a culture line with only a single bacterial food source, *Flavobacterium* sp., was developed. The GC content of *Flavobacterium* DNA (33%) is sufficiently different from that of *M. brevicollis* (55%) to allow separation over a CsCl gradient. *M. brevicollis* genomic DNA was used to construct replicate libraries containing inserts of 2–3 kilobases (kb), 6–8 kb and 35–40 kb.

Genome sequencing, assembly and validation. The draft sequence of the *M. brevicollis* genome was generated from ~8.5-fold redundant paired-end whole-genome shotgun sequence coverage (Supplementary Information). Sequence data derived from six whole-genome shotgun libraries were assembled using release 2.9.2 of the whole-genome shotgun assembler Jazz. Completeness of the draft genome was assessed by capturing ~98.5% of sequenced expressed sequence tags.

Gene prediction and annotation. Gene models (9,196) were predicted and annotated using the Joint Genome Institute (JGI) Annotation Pipeline (Supplementary Information).

Intron analysis. Homologues of 473 highly conserved genes from *M. brevicollis* and representative eukaryotes were aligned to reveal the position and phylogenetic distribution of 1,989 highly reliable intron splice sites at 1,054 conserved positions. The evolutionary history of introns in orthologous genes was inferred using Dollo parsimony, Roy-Gilbert maximum likelihood and Csuros maximum likelihood[45–47].

Analysis of signalling, adhesion and transcription factor protein domains. Gene models containing metazoan signalling, adhesion and transcription factor domains were identified by text and protein domain ID searches of the JGI *M. brevicollis* genome portal, local BLAST searches within the *M.brevicollis* genome scaffolds, the online Pfam and SMART tools, and reciprocal BLAST searches in the NCBI non-redundant protein database (Supplementary Information).

Received 21 August; accepted 17 December 2007.

1. James-Clark, H. On the spongiae ciliatae as infusoria flagellata; or observations on the structure, animality, and relationship of *Leucosolenia botryoides*. *Ann. Mag. Nat. His.* **1**, 133–142, 188–215 250–264 (1868).
2. Saville Kent, W. *A Manual of the Infusoria* (Bogue, London, 1880–1882).
3. Steenkamp, E. T., Wright, J. & Baldauf, S. L. The protistan origins of animals and fungi. *Mol. Biol. Evol.* **23**, 93–106 (2006).
4. Medina, M. *et al.* Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. *Int. J. Astrobiology* **2**, 203–211 (2003).
5. Philippe, H. *et al.* Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol. Biol. Evol.* **21**, 1740–1752 (2004).
6. Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W. & Burger, G. The closest unicellular relatives of animals. *Curr. Biol.* **12**, 1773–1778 (2002).
7. Burger, G., Forget, L., Zhu, Y., Gray, M. W. & Lang, B. F. Unique mitochondrial genome architecture in unicellular relatives of animals. *Proc. Natl Acad. Sci. USA* **100**, 892–897 (2003).
8. Lavrov, D. V., Forget, L., Kelly, M. & Lang, B. F. Mitochondrial genomes of two demosponges provide insights into an early stage of animal evolution. *Mol. Biol. Evol.* **22**, 1231–1239 (2005).
9. King, N. & Carroll, S. B. A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. *Proc. Natl Acad. Sci. USA* **98**, 15032–15037 (2001).
10. King, N., Hittinger, C. T. & Carroll, S. B. Evolution of key cell signaling and adhesion protein families predates animal origins. *Science* **301**, 361–363 (2003).
11. Segawa, Y. *et al.* Functional development of Src tyrosine kinases during evolution from a unicellular ancestor to multicellular animals. *Proc. Natl Acad. Sci. USA* **103**, 12021–12026 (2006).
12. Snell, E. A. *et al.* An unusual choanoflagellate protein released by Hedgehog autocatalytic processing. *Proc. Biol. Sci.* **273**, 401–407 (2006).
13. Rokas, A., Kruger, D. & Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. *Science* **310**, 1933–1938 (2005).
14. King, N. The unicellular ancestry of animal development. *Dev. Cell* **7**, 313–325 (2004).
15. Buck, K. R. & Garrison, D. L. Distribution and abundance of choanoflagellates (Acanthoecidae) across the ice-edge zone in the Weddell Sea, Antarctica. *Mar. Biol.* **98**, 263–269 (1988).
16. Thomsen, H. A. & Larsen, J. Loricate choanoflagellates of the Southern Ocean with new observations on cell-division in *Bicosta spinifera* (Throndsen, 1970) from Antarctica and *Saroeca attenuata* Thomsen, 1979, from the Baltic Sea. *Polar Biol.* **12**, 53–63 (1992).
17. Arndt, H. *et al.* in *The Flagellates* (ed. Leadbeater, B. S. C. & Green, J. C.) 240–268 (Taylor & Francis, London, 2000).
18. Boenigk, J. & Arndt, H. Bacterivory by heterotrophic flagellates: community structure and feeding strategies. *Antonie Van Leeuwenhoek* **81**, 465–480 (2002).
19. Leadbeater, B. S. C. Life-history and ultrastructure of a new marine species of *Proterospongia* (Choanoflagellida). *J. Mar. Biol. Assoc.UK* **63**, 135–160 (1983).
20. Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79–86 (2004).
21. Ruiz-Trillo, I., Lane, C. E., Archibald, J. M. & Roger, A. J. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. *J. Eukaryot. Microbiol.* **53**, 379–384 (2006).
22. Seo, H. C. *et al.* Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**, 2506 (2001).
23. Sullivan, J. C., Reitzel, A. M. & Finnerty, J. R. A high percentage of introns in human genes were present early in animal evolution: evidence from the basal metazoan *Nematostella vectensis*. *Genome Inform.* **17**, 219–229 (2006).
24. Halbleib, J. M. & Nelson, W. J. Cadherins in development: cell adhesion, sorting, and tissue morphogenesis. *Genes Dev.* **20**, 3199–3214 (2006).
25. Gupta, G. & Surolia, A. Collectins: sentinels of innate immunity. *Bioessays* **29**, 452–464 (2007).
26. Yamaguchi, Y. Lecticans: organizers of the brain extracellular matrix. *Cell. Mol. Life Sci.* **57**, 276–289 (2000).
27. Zelensky, A. N. & Gready, J. E. The C-type lectin-like domain superfamily. *FEBS J.* **272**, 6179–6217 (2005).
28. Akiyama, S. K. Integrins in cell adhesion and signaling. *Hum. Cell* **9**, 181–186 (1996).
29. Erwin, D. H. The origin of metazoan development — a paleobiological perspective. *Biol. J. Linn. Soc.* **50**, 255–274 (1993).
30. van der Rest, M. & Garrone, R. Collagen family of proteins. *FASEB J.* **5**, 2814–2823 (1991).
31. Tissir, F. & Goffinet, A. M. Reelin and brain development. *Nature Rev. Neurosci.* **4**, 496–505 (2003).
32. Suarez-Castillo, E. C. & Garcia-Arraras, J. E. Molecular evolution of the ependymin protein family: a necessary update. *BMC Evol. Biol.* **7**, 23 (2007).
33. Leadbeater, B. S. Developmental and ultrastructural observations on two stalked marine choanoflagellates, *Acanthoecopsis spiculifera* Norris and *Acanthoeca spectabilis* Ellis. *Proc. R. Soc. Lond. B* **204**, 57–66 (1979).
34. Leadbeater, B. S. C. Developmental studies on the loricate choanoflagellate *Stephanoeca diplocostata* Ellis. 7. Dynamics of costal strip accumulation and lorica assembly. *Eur. J. Protistol.* **30**, 111–124 (1994).
35. Hutter, H. *et al.* Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science* **287**, 989–994 (2000).
36. Adamska, M. *et al.* The evolutionary origin of hedgehog proteins. *Curr. Biol.* **17**, R836–R837 (2007).
37. Letunic, I. *et al.* SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, D257–D260 (2006).
38. Best, A. A., Morrison, H. G., McArthur, A. G., Sogin, M. L. & Olsen, G. J. Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. *Genome Res.* **14**, 1537–1547 (2004).
39. Derelle, R., Lopez, P., Le Guyader, H. & Manuel, M. Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. *Evol. Dev.* **9**, 212–219 (2007).
40. Larroux, C. *et al.* The NK homeobox gene cluster predates the origin of Hox genes. *Curr. Biol.* **17**, 706–710 (2007).
41. Knoll, A. H. *Life on a Young Planet* (Princeton Univ. Press, Princeton, 2003).
42. Peterson, K. J. & Butterfield, N. J. Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record. *Proc. Natl Acad. Sci. USA* **102**, 9547–9552 (2005).
43. Ekman, D., Bjorklund, A. K. & Elofsson, A. Quantification of the elevated rate of domain rearrangements in Metazoa. *J. Mol. Biol.* **327**, 1337–1348 (2007).
44. Tordai, H., Nagy, A., Farkas, K., Banyai, L. & Patthy, L. Modules, multidomain proteins and organismic complexity. *FEBS J.* **272**, 5064–5078 (2005).
45. Csuros, M. in *Proceedings of the Comparative Genomics: RECOMB 2005 International Workshop; Dublin, Ireland* (eds McLysaght, A. & Huson, D. H.) 47–60 (Springer, Berlin, 2005).
46. Rogozin, I. B., Wolf, Y. I., Sorokin, A. V., Mirkin, B. G. & Koonin, E. V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **13**, 1512–1517 (2003).
47. Roy, S. W. & Gilbert, W. Complex early genes. *Proc. Natl Acad. Sci. USA* **102**, 1986–1991 (2005).
48. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory, Lawrence Berkeley National Laboratory, and Los Alamos National Laboratory. Work in the King laboratory is supported by funding from the Gordon and Betty Moore Foundation, the Pew Scholars program, and R. Melmon. The Rokhsar group is supported by the Gordon and Betty Moore Foundation and R. Melmon. We thank J. Stajich, P. Johnson and R. Lusk for discussions, E. Hare, E. Meltzer and K. Osoegawa for technical advice, E. Begovic for assistance with Fig. 1, M. Dayel and N. Patel for critical reading of the manuscript, and S. Carroll for early support of this project. N.K. is a Scholar in the Integrated Microbial Biodiversity Program of the Canadian Institute for Advanced Research.

Author Contributions N.K. and D.R. are co-senior authors.

Author Information The sequenced strain of *M. brevicollis* has been deposited at ATCC.org under accession number PRA-258. The genome assembly and annotation data are deposited at DBJ, EMBL and GenBank under the project accession ABFJ00000000. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.K. (nking@berkeley.edu) or D.R. (dsrokhsar@lbl.gov).

Supplementary materials for: The genome of the choanoflagellate *Monosiga brevicollis* and the origins of metazoan multicellularity

Nicole King[1,2], M. Jody Westbrook[1*], Susan L. Young[1*], Alan Kuo[3], Monika Abedin[1], Jarrod Chapman[1], Stephen Fairclough[1], Uffe Hellsten[3], Yoh Isogai[1], Ivica Letunic[4], Michael Marr[5], David Pincus[6], Nicholas Putnam[1], Antonis Rokas[7], Kevin J. Wright[1], Richard Zuzow[1], William Dirks[1], Matthew Good[6], David Goodstein[1], Derek Lemons[8], Wanqing Li[9], Jessica Lyons[1], Andrea Morris[10], Scott Nichols[1], Daniel J. Richter[1], Asaf Salamov[3], JGI Sequencing[3], Peer Bork[4], Wendell A. Lim[6], Gerard Manning[11], W. Todd Miller[9], William McGinnis[8], Harris Shapiro[3], Robert Tjian[1], Igor V. Grigoriev[3], Daniel Rokhsar[1,3]

[1]Department of Molecular and Cell Biology and the Center for Integrative Genomics,
University of California, Berkeley, CA 94720, USA

[2]Department of Integrative Biology, University of California, Berkeley, CA 94720, USA

[3]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

[4]EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany

[5]Department of Biology, Brandeis University, Waltham, MA 02454

[6]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA

[7]Vanderbilt University, Department of Biological Sciences, Nashville, TN 37235, USA

[8]Division of Biological Sciences, University of California, San Diego
La Jolla, CA 92093

[9]Department of Physiology and Biophysics, Stony Brook University, Stony Brook, NY 11794

[10]University of Michigan, Department of Cellular and Molecular Biology, Ann Arbor MI 48109

[11]Razavi Newman Bioinformatics Center, Salk Institute for Biological Studies, La Jolla, CA 92037

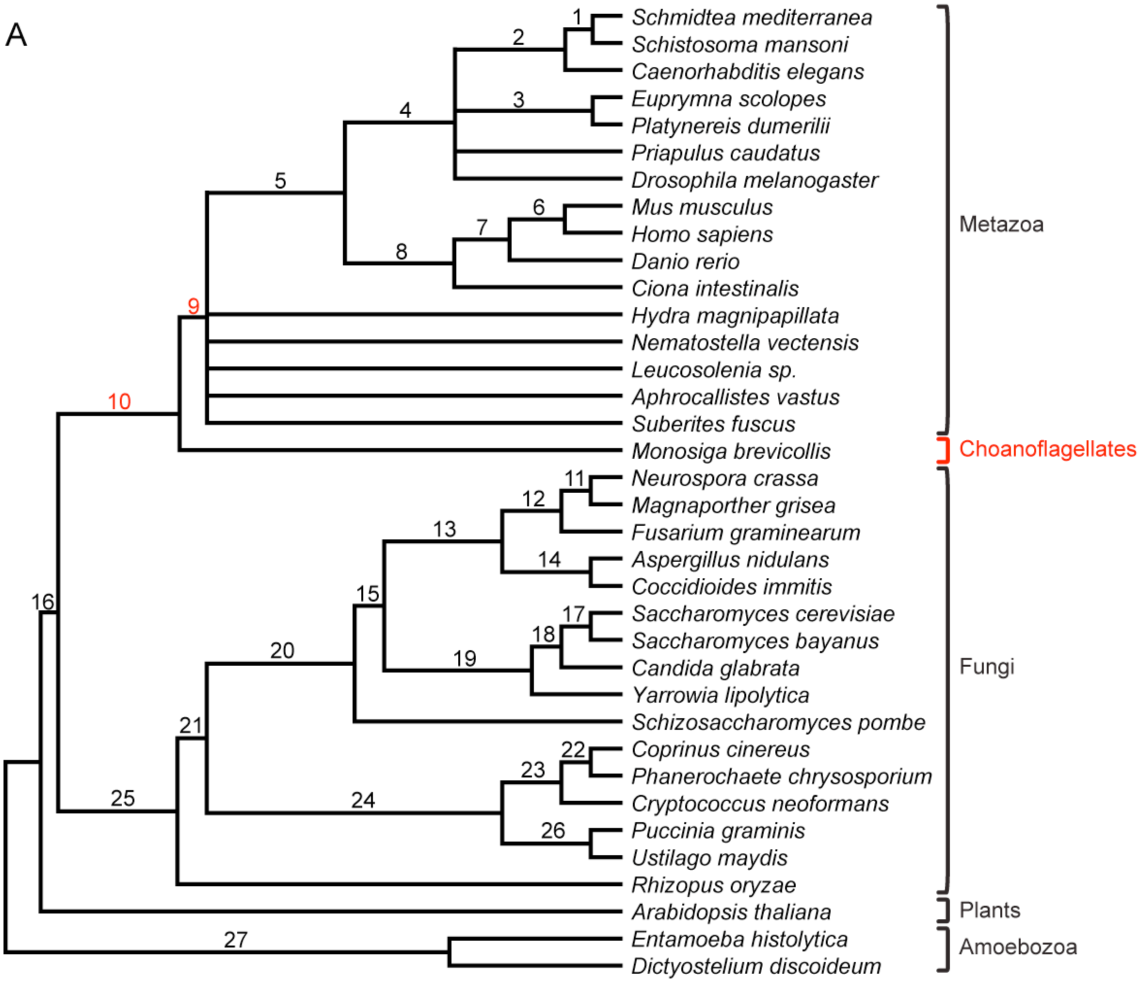
*These authors contributed equally to this work.

Contents:

Supplementary Figures

Supplementary Tables

Supplementary Notes



B

Branch	Support	Branch	Support	Branch	Support
1	100 / 100	10	100 / 100	19	100 / 100
2	98 / 83	11	62 / 69	20	100 / 100
3	88 / 49	12	100 / 100	21	100 / 100
4	83 / 27	13	100 / 100	22	100 / 100
5	96 / 76	14	100 / 100	23	100 / 100
6	100 / 100	15	100 / 99	24	100 / 100
7	100 / 100	16	100 / 100	25	100 / 100
8	95 / 42	17	100 / 100	26	100 / 100
9	100 / 74	18	100 / 100	27	100 / 100

Figure S1. Choanoflagellates are a close outgroup of Metazoa. A phylogenetic analysis of 50 genes shows that *M. brevicollis* is placed outside metazoans (including poriferans and cnidarians), and justifies its choice for comparative genomic investigations into the transition from a unicellular to the multicellular metazoan lifestyle. (A) The tree with the highest likelihood in the maximum likelihood analyses is shown. (B) Boostrap support values for all branches shown in A are shown. For each branch, the bootstrap support values from the maximum likelihood and maximum parsimony are shown, respectively.

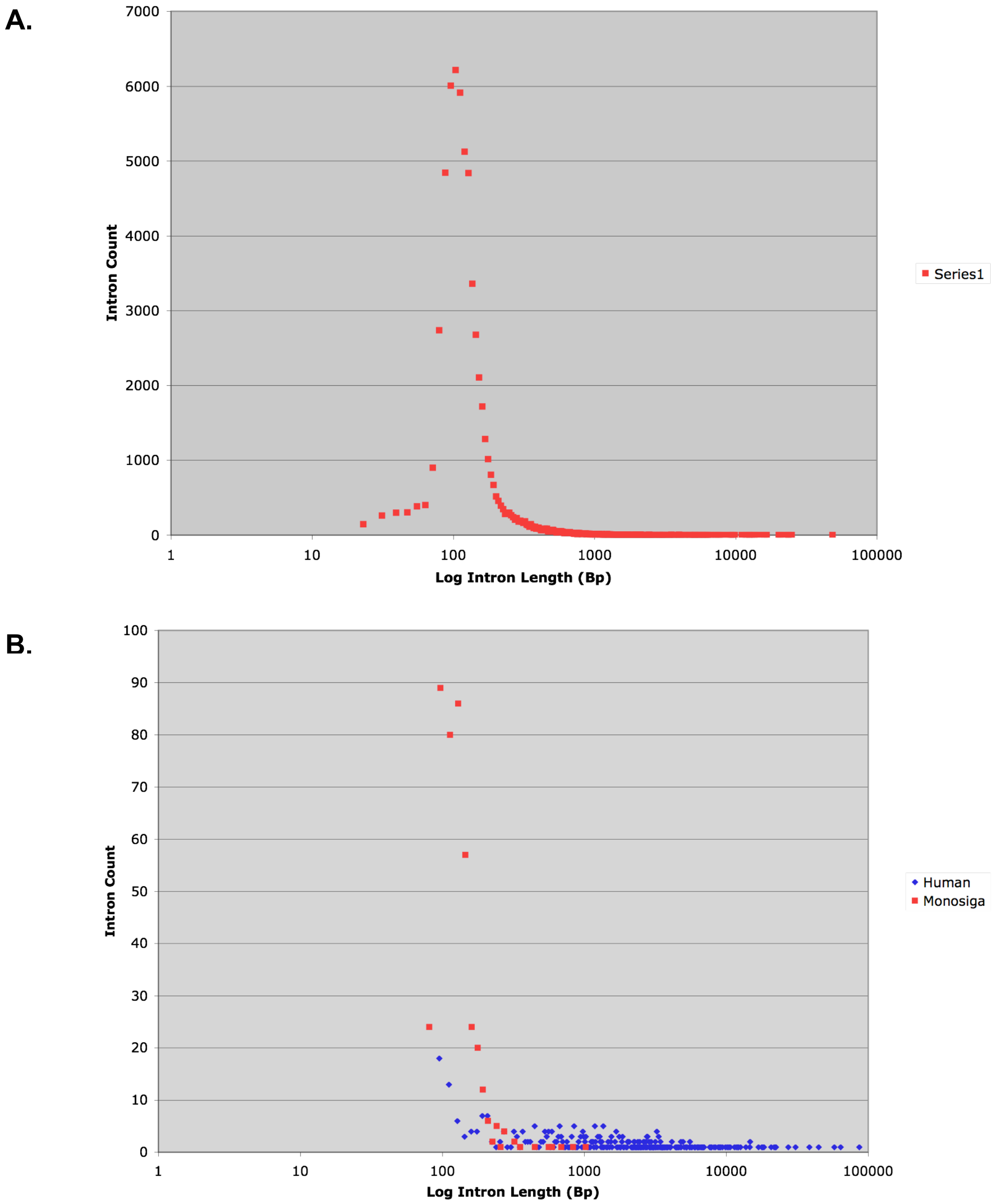


Figure S5. Distribution of *M. brevicollis* intron lengths. A. Distribution of the lengths of the 60,636 introns from the *M. brevicollis* filtered gene models. B. Distribution of the lengths of 419 introns that occur at the same positions in orthologous genes in *M. brevicollis* and humans.

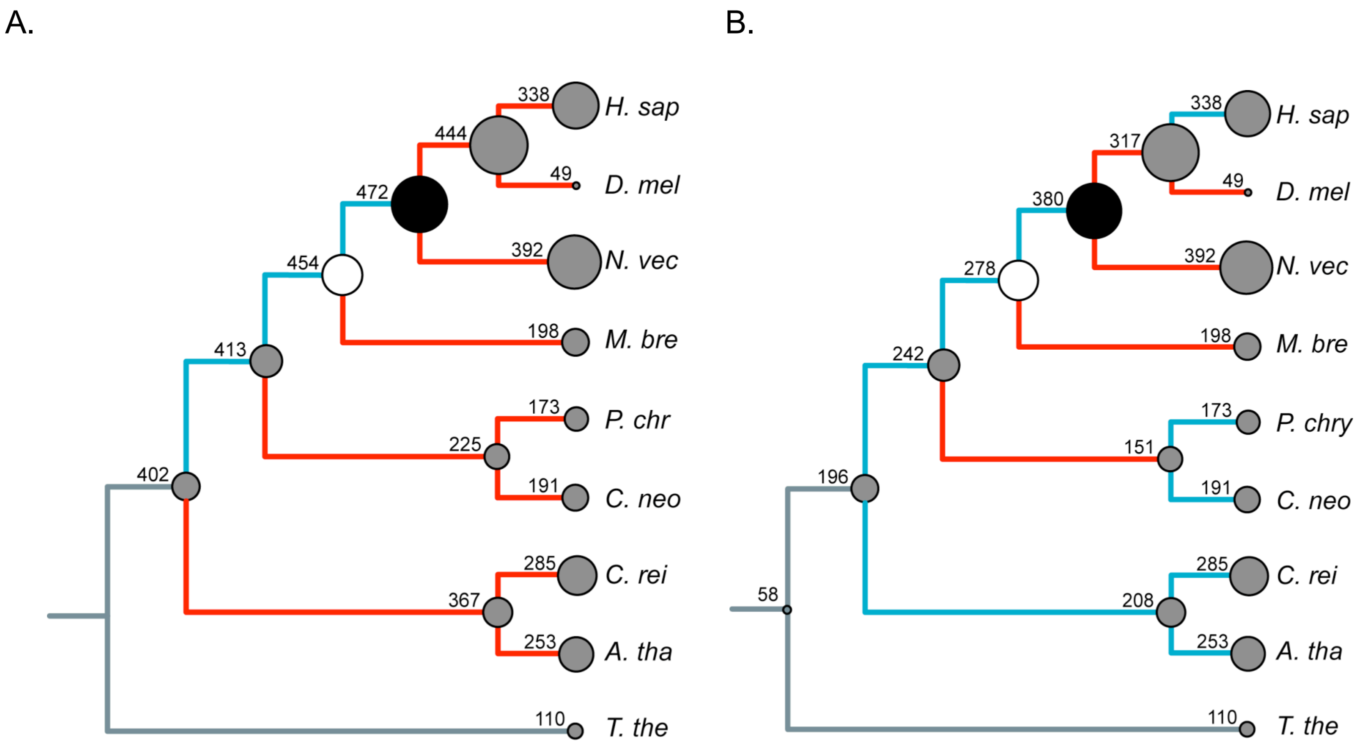


Figure S3. Analysis of intron evolution in nine species. Ancestral intron content and intron gains and losses were inferred using two additional methods: A. Roy-Gilbert maximum likelihood and B. Dollo parsimony methods. A sample of 1,054 intron positions in highly conserved sequences from 473 orthologs were used. Branches with at least 10% more gain than loss are blue, those with more loss than gain are red, and those with comparable amounts are black. Outgroup branches, for which intron loss could not be calculated, are grey. The inferred or observed number of introns present in ancestors and extant taxa are indicated next to proportionally sized circles. Species included are *Tetrahymena thermophila* (*T. the*), *Chlamydomonas reinhardtii* (*C. rei*), *Arabadopsis thaliana* (*A. tha*), *Cryptococcus neoformans* A (*C. neo*), *Phanerochaete chrysosporium* (*P. chr*), *Monosiga brevicollis* (*M. bre*), *Nematostella vectensis* (*N. vec*), *Drosophila melanogaster* (*D. mel*) and humans (*H. sap*).

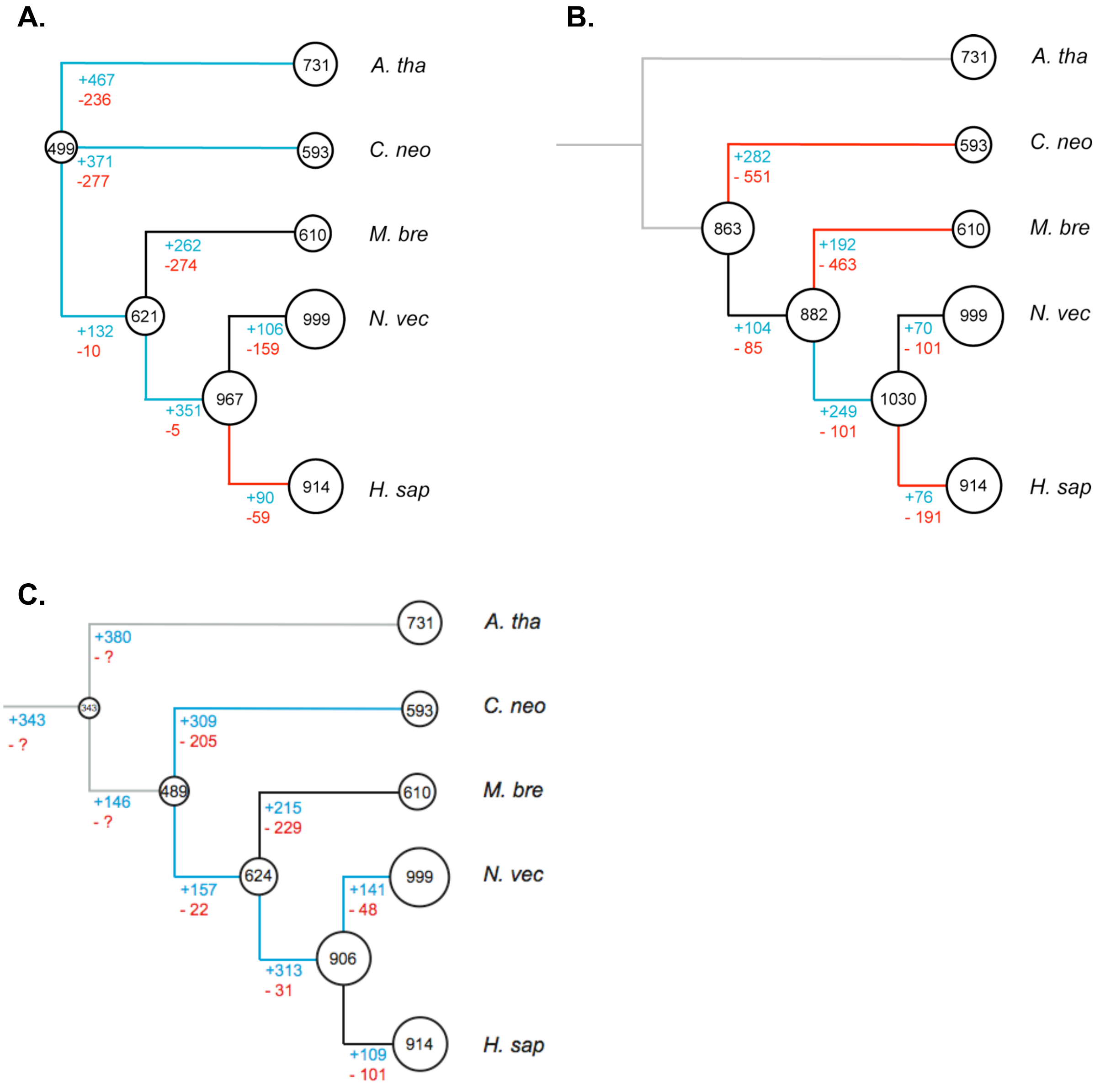


Figure S4. Analysis of intron evolution in five species. Ancestral intron content and intron gains and losses were inferred using three methods: **A.** Csuuros maximum liklihood, **B.** Roy-Gilbert maximum likelihood and **C**. Dollo parsimony methods. A sample of 2121 intron positions in highly conserved sequences from 538 orthologs were used. Branches with 10% more gain than loss are blue, those with more loss than gain are red, and those with comparable amounts are black. Outgroup branches are grey. The numbers of introns gained and lost are shown in blue and red respectively. Using Dollo parsimony, the number of introns lost cannot be inferred without an outgroup, and this is indicated by question marks. The inferred or observed number of introns present in ancestors and extant taxa are in proportionally sized circles. Species included are the plant *Arabadopsis thaliana* (*A. tha*), the fungus *Cryptococcus neoformans* A (*C. neo*), the choanoflagellate *M. brevicollis* (*M. bre*) and the metazoans *Nematostella vectensis* (*N. vec*) and humans (*H. sap*).

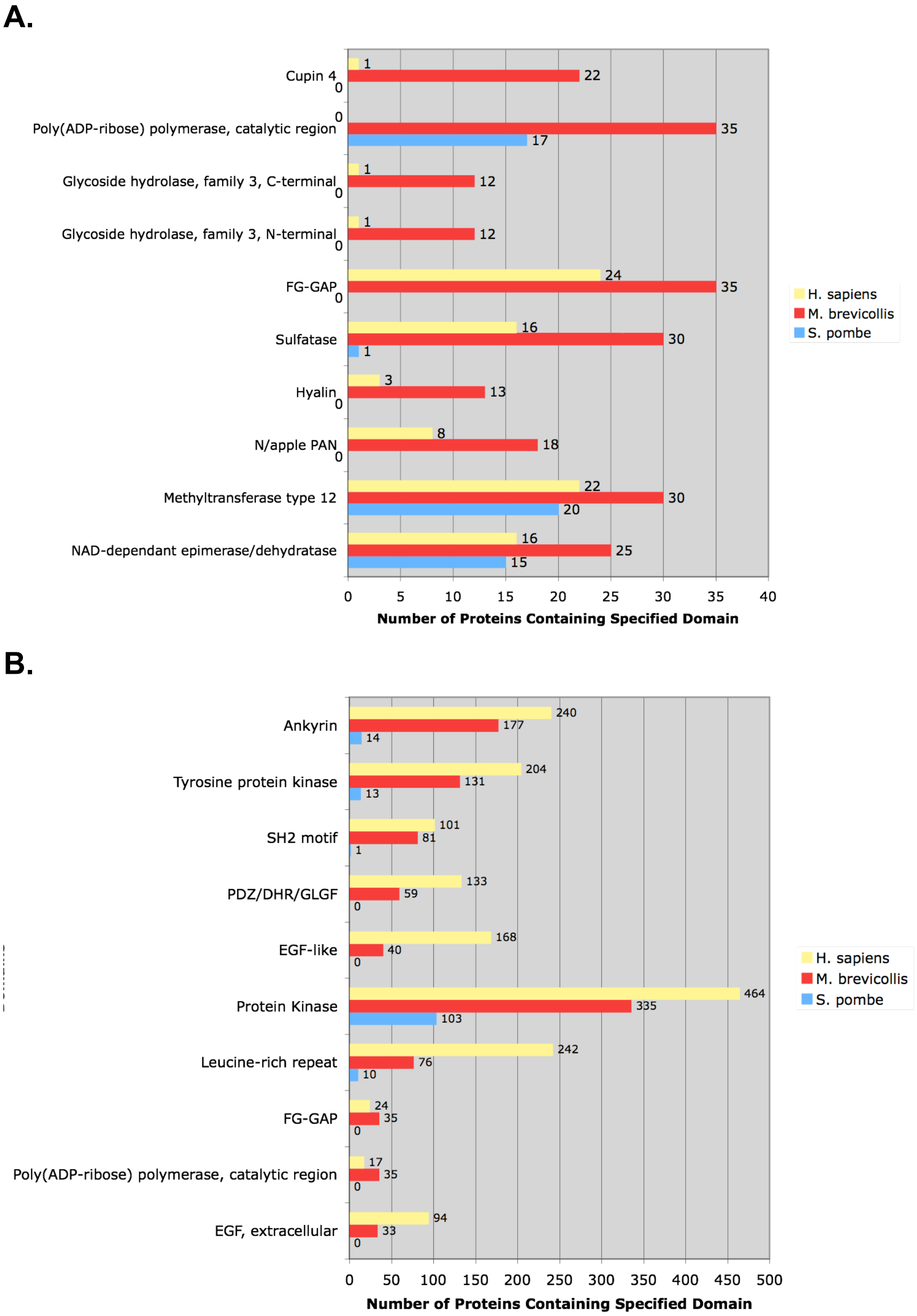


Figure S5. Domains significantly over-represented in choanoflagellates. Significantly over-represented domains in the choanoflagellate genome were identified by comparing the occurrence of PFAM domains excluding repeats (one hit per protein) in *M. brevicollis* to the human (panel A) and *S. pombe* (panel B) genomes. The ten most significantly over represented domains from each comparison as determined by a Chi-squared test are shown, with the most significantly over-represented domain shown at the top of the graphs. The number of proteins containing each domain is indicated.

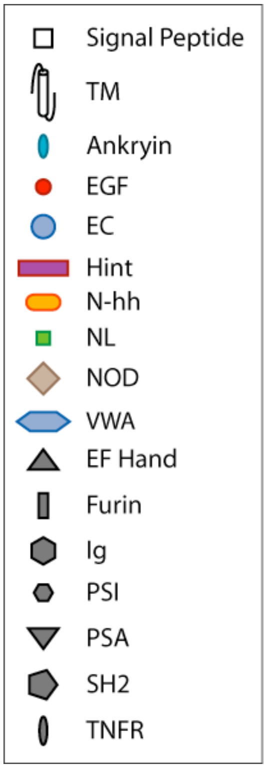


Figure S6. Legend for domains shown in Figure 4 - Domain shuffling and the evolution of Notch and Hedgehog. Analysis of the draft gene set reveals that *M. brevicollis* possesses protein domains characteristic of metazoan Notch and Hedgehog (Hh) proteins, some of which were previously thought to be unique to metazoans. The presence of these domains in disparate peptides in *M. brevicollis* suggests that domain shuffling has occurred in these proteins since the separation of the choanoflagellate and metazoan lineages.

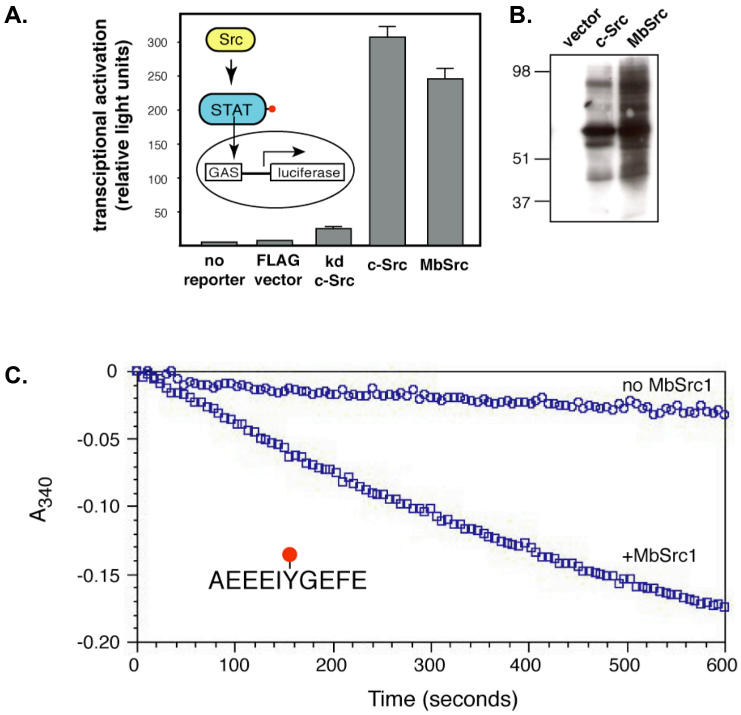


Figure S7. MbSrc functions like human c-Src. A. MbSrc can substitute for c-Src in a reporter assay. Src/Fyn/Yes triple knockout (SYF) cells were transfected with the indicated FLAG-constructs and with a luciferase reporter gene regulated by the interferon-gamma activation sequence. kd = kinase-dead c-Src. B. MbSrc phosphorylates substrates in mammalian cells. SYF cells were transfected with wild-type c-Src, Y527F c-Src, or MbSrc. Tyrosine-phosphorylated proteins in whole cell lysates were visualized by anti-pY Western blotting. C. Kinase activity of purified MbSrc. MbSrc was expressed and purified using the Sf9/baculovirus system. Phosphorylation of a synthetic peptide substrate containing the Src optimal motif was measured by a continuous spectrophotometric assay.

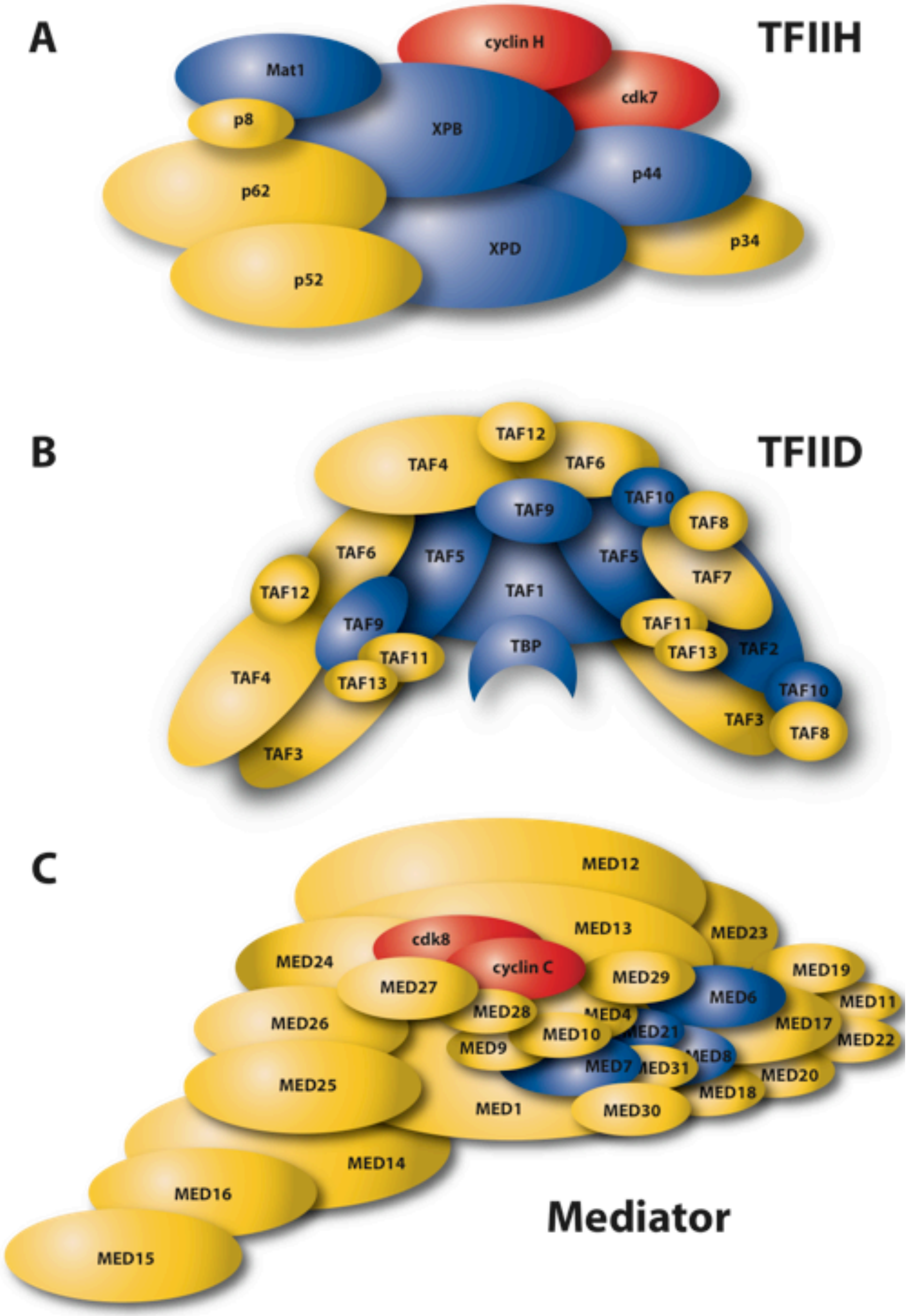


Figure S8. Diagrams of metazoan general transcription factors and coactivators. Blue indicates subunits found in M. brevicollis; yellow indicates a subunit not found in M. brevicollis; and red indicates a possible homolog in *M. brevicollis*. A. Diagram of TFIIH. B. Diagram of TFIID. C. Diagram of Mediator.

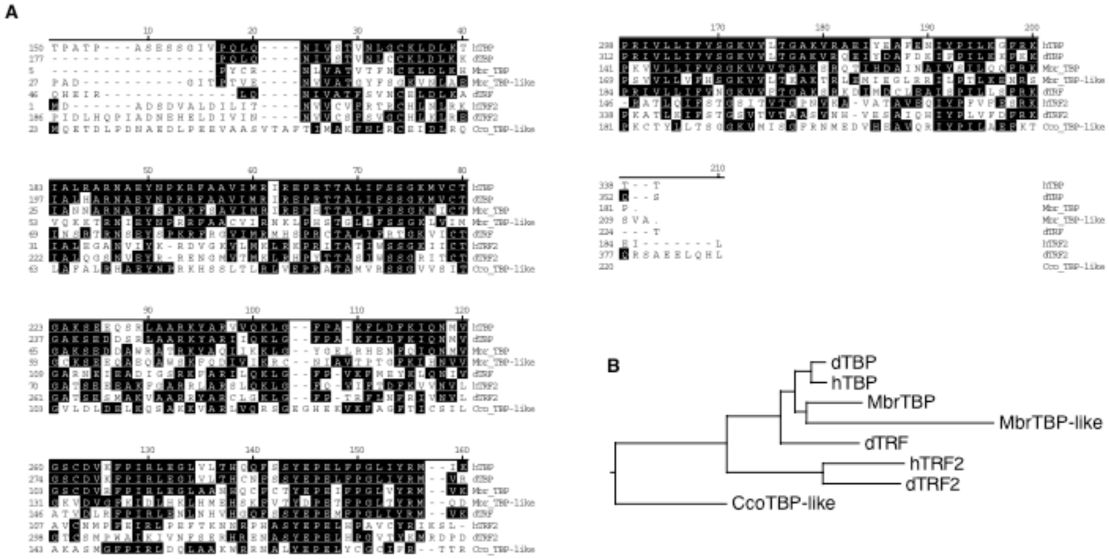


Figure S9. TBP-like factor in *M. brevicollis*. A. ClustalW alignment of Drosophila, human, *M. brevicollis* TBPs and TRFs. Only the highly conserved region corresponding to the saddle domain of TBP is shown. A dinoflagellate (*Crypthecodinium cohnii*) TBP-like factor[1] is used as an outgroup. B. Tree diagram generated from ClustalW alignment. The tree was generated using Megalign program (DNASTAR).

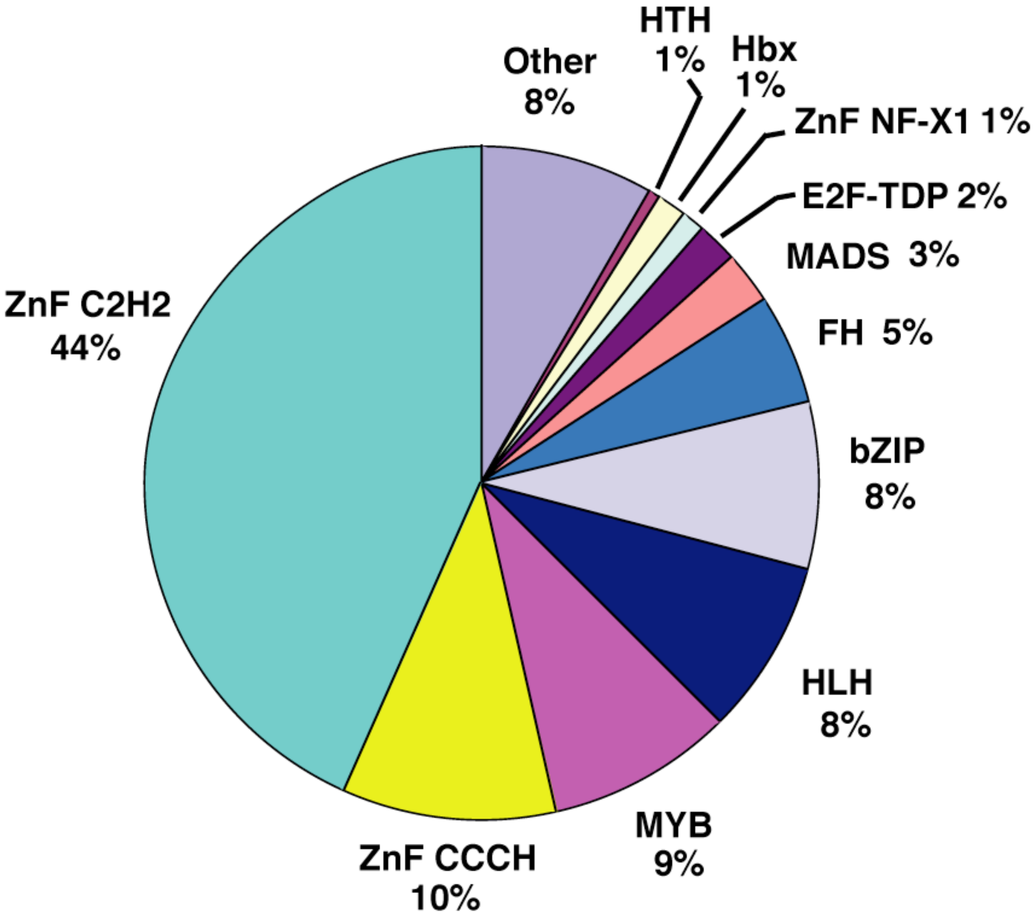


Figure S10. Relative abundance of transcription factor families in *M. brevicollis*. Of 155 protein models containing transcription factor associated domains, the percentage of protein models containing the indicated family specific domain is shown. bZip: basic-leucine zipper; E2f-TDP: E2F/DP (dimerizaton partner) family winged-helix DNA-binding domain; FH: forkhead; Hbx: homeobox; HLH: helix-loop-helix; HTH: helix-turn-helix; ZnF: zinc finger.

```
HESX_HUMAN          GRRPRTAFTQNQIEVLENVF~~~RVNCYPGIDIREDLAQKLNLEEDRIQIWFQNRRAKLKRSH
PMXA_HUMAN          QRRIRTTFTSAQLKELERVF~~~AETHYPDIYTREELALKIDLTEARVQVWFQNRRAKFRKQE
PMX1_HUMAN          QRRNRTTFNSSQLQALERVF~~~ERTHYPDAFVREDLARRVNLTEARVQVWFQNRRAKFRRNE
OTX1_HUMAN          QRRERTTFTRSQLDVLEALF~~~AKTRYPDIFMREEVALKINLPESRVQVWFKNRRAKCRQQQ
CRT1_HUMAN          KRRHRTTFTSLQLEELEKVF~~~QKTHYPDVYVREQLALRTELTEARVQVWFQNRRAKWRKRE
PRH1_HUMAN          RRRHRTTFSPVQLEQLESAF~~~GRNQYPDIWARESLARDTGLSEARIQVWFQNRRAKQRKQE
PIX1_HUMAN          QRRQRTHFTSQQLQELEATF~~~QRNRYPDMSMREEIAVWTNLTEPRVRVWFKNRRAKWRKRE
GSC_HUMAN           KRRHRTIFTDEQLEALENLF~~~QETKYPDVGTREQLARKVHLREEKVEVWFKNRRAKWRRQK
PAX6_HUMAN          LQRNRTSFTQEQIEALEKEF~~~ERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRREE
Renprd1             QRRHRTNFTSHQLEELEKAF~~~EKTRYPDVFMREELAMKISLTEARVQVWFQNRRAKWRKAE
Renprd2             SKRNRTTFTAHQLDELEMIF~~~RQTHYPDVLLREKLAQRIGLPESRVQVWFQNRRAKWRKRE
Renprd3             KRRYRTTFTSFQLRELEKAF~~~ERTHYPDVFTREDLANRVELTEARVQVWFQNRRAKWRKKE
Renprd4             QRRFRTTFTSYQLQELEAAF~~~AKTHYPDVFMREDLALRINLTEARVQVWFQNRRAKWRRAQ
Renprd5             PKRTRTAYSNSQLDQLELIF~~~ATTHYPDVFTREDLSRRLGIREDRIQVWFQNRRARFRKQE
Renprd6             IKKKRMTYTKQQKDALESYF~~~YQDSYPDTQARENMSEALGITPEKVQVWFQNRRAKCRKRE
Renprd7             PKKTRTQFSPKQLVYLEECF~~~LKNRFPSAKERESIAEELDLTTQHIQVWFQNRRAKHRRKS
LHX2_HUMAN          TKRMRTSFKHHQLRTMKSYF~~~AINHNPDAKDLKQLAQKTGLTKRVLQVWFQNARAKFRRNL
LH61_HUMAN          AKRARTSFTAEQLQVMQAQF~~~AQDNNPDAQTLQKLADMTGLSRRVIQVWFQNCRARHKKHT
ISL1_HUMAN          TTRVRTVLNEKQLHTLRTCY~~~AANPRPDALMKEQLVEMTGLSPRVIRVWFQNKRCKDKKRS
LHX3_HUMAN          AKRPRTTITAKQLETLKSAY~~~NTSPKPARHVREQLSSETGLDMRVVQVWFQNRRAKEKRLK
LMXB_HUMAN          PKRPRTILTTQQRRAFKASF~~~EVSSKPCRKVRETLAAETGLSVRVVQVWFQNQRAKMKKLA
RenLIM1             KGKTRTSINPKQLIVLQATY~~~EKEPRPSRSMREELAAQTGLTAKVIQVWFQNRRSKDKKDG
RenLIM2             QPRIRTVLTEQQLQTLRSVY~~~QTNPRPDALLKEQLCELTGLSPRVIRVWFQNRRCKDKKAL
RenLIM3             QKRPRTTISQKQLDLLKTAY~~~CVSPKPSRHVRQELSDKTGLDMRVVQVWFQNKRAKDKRTK
OCT6_HUMAN          KRKKRTSIEVGVKGALESHF~~~LKCPKPSAHEITGLADSLQLEKEVVRVWFCNRRQKEKRMT
PO61_HUMAN          KRKRRTSFTPQAIEALNAYF~~~EKNPLPTGQEITEIAKELNYDREVVRVWFCNRRQTLKNTS
BR3A_HUMAN          KKRKRTSIAAPEKRSLEAYF~~~AVQPRPSSEKIAAIAEKLDLKKNVVRVWFCNQRQKQKRMK
OC3A_HUMAN          RKRKRTSIENRVRGNLENLF~~~LQCPKPTLQQISHIAQQLGLEKDVVRVWFCNRRQKGKRSS
RenPOU1             HRKRRTTIGMSAKERLEQHF~~~QVQPKPSSSDITKVADSLNLDKEVIRVWFCNRRQREKRVR
RenPOU2             RRRRRTAIPVQTKKQLLKEF~~~ENNPKPSVKALKALAEKLGIRFEVVRVWFCNKRAKKKAGK
RenPOU3             KRKGRTAISVQTKKQLLKEF~~~ENDPKPSPKDLKAISEKLGIGFEVVRVWFCNKRAKRKAGK
RenPOU4             KRKKRVVYTPHALSILNKYF~~~LKEPRPNRQIIEMVAEELDLLPEEVRVWFCNKRQKYKTSN
A.nid1              KNNKRQRATQDQLVLLEMEF~~~NKNPTPTAATRERIAQEINMTERSVQIWFQNRRAKIKMLA
N.cra1              KNQKRQRATQDQLTTLEMEF~~~NKNPTPTATVRERIAEEINMTERSVQIWFQNRRAKIKLLA
R.ory1              STRKRTHLSTEQVSLLESSF~~~NENSLPDSAVRSRLAQELSVTERTVQIWFQNRRAKEKKIK
P.bla3              AKPKRKRISPDQFRVLSDLF~~~EKTDTPNYELRERMAGRLNMTNREVQVWFQNRRAKATRAK
R.ory8              IRPKRKRITPNQLEVLTSIF~~~ERTKTPNYQLREHTAKELNMTNREVQVWFQNRRAKLNRKR
R.ory2              RTRKRTRATPEQLAILEKSF~~~NVNPSPNSRVREQLSLQLGMTERSIQIWFQNRRAKVKNQT
P.bla1              QPRKRTRASPEQLGILEKTF~~~NINPSPNNRVREQLSQQLSMSERSIQIWFQNRRAKVKNIA
R.ory3              PVRKRTRATADQLSVLEDTF~~~AMNVSPNSKLRKQLAEQLQMSERSIQIWFQNRRAKVKHMQ
R.ory4              DTKKRTRVTPGQLAILEETF~~~SMTATPDSKLRKQLAERLKMPERSIQIWFQNRRAKVKMLQ
L.bic3              EKRKRSRVTQEQLVHLEQYF~~~KADRCPTATRRREISEQLGMQERQTQIWFQNRRAKAKLQE
P.chr3              EQKKRGRVTPEQLAVLEAIF~~~AANRSPNAVRRKEISEQLGMTERQTQIWFQNRRAKEKHAG
R.ory5              EIKHRRRTSRAQLKVLEESF~~~SENPKPNATVRRILAQQLDMTPRGVQIWFQNRRAKAKLLR
R.ory6              ETKHRRRTSRGQVKILEKAF~~~HDNPKPNGRARERLAESLSMSPRGVQIWFQNRRAKAKNQQ
L.bic1              EVKHRKRTTSAQLKVLETVF~~~KRDTKPNASLRTELAAQLDMTARGVQVWFQNRRAKEKVKA
R.ory7              IKAKRKRASPSQLYILNQVF~~~QQTCFPSTELRIELGKRLGMSPRTVQIWFQNKRQSTRTKE
A.nid3              ARQKRRRTSPEDYAILEAEY~~~QRNPKPDKISRASIVSRVSLGEKEVQIWFQNRRQNDRRKS
N.cra3              PKGKRKRTTAKDKAILEAAY~~~NANPKPDKAARQDIVNRVSLNEKEVQIWFQNRRQNDRRKS
A.nid2              ENLSRPRLTKEQVETLEAQF~~~QAHPKPSSNVKRQLAQQTHLSLPRVANWFQNRRAKAKQQK
N.cra2              QTEPKPRLAKDEVELLEREF~~~AKNPKPNTSLKRELAEQMGVEVPRINNWFQNRRAKEKQMR
P.bla2              FHKKRMMLKPYQYKVLQDHF~~~SANPKPDARVYIDIASRLNVSITKIKNWFQNRRAKARKDK
P.bla4              KIKNRRRFSATEAALLERRY~~~AEEQSPSQHVLQGLADQMSTPRKTITTWFQNRRAKYKRRS
P.bla5              EIKHRHRFSTSELELLEELY~~~RRHPRPSSSEKKAMAAKLDTTPGRVQVWLQNRRAKERKAQ
R.ory9              PIKQRRRFSLEEAQFLEMEY~~~NNNPSPTQDKIQQIASKINSPRKVVTTWFQNRRAKNRRRS
R.ory10             PIRPRKRFTSNQIHLLEMEY~~~MKSDHPSRETKETLANQFKTSIRRIQIWFQNRRAKEKRGE
R.ory11             VARRRMRTSKEEMAVLDEYY~~~RKNPNPNQEEKKEIANLLKMGTKNVHFWFQNRRAKENKKK
A.nig1              KKMKRFRLTHNQTRFLMSEF~~~TRQAHPDAAHRERLSKEIGLTPRQVQVWFQNRRAKLKRLT
N.cra4              RKMKRFRLTHQQTRFLMSEF~~~AKQPHPDAAHRERLSREIGLSPRQVQVWFQNRRAKIKRLT
C.neo3              QVKHRRRTTPEQLKVLEFWY~~~DINPKPDNQLREQLAAQLGMTKRNVQVWFQNRRAKMKGLA
C.neo4              FKSPRKRTNDVQLAMLSEVF~~~RRTQYPSTEERDELAKQLGMTSRSVQIWFQNRRRAVKVDQ
```

Figure continued on next page

```
P.chr2             EKKPRHRMTDKQLERLEALY~~~QQDTHPTREQKQALGEEVGMDTRTVTVWFQNRRQLSKKNT
C.neo1             KMSPRKRFTIPQLQILEVQW~~~SNDISPPKVDRQRLAMWMGTRTKHVNIWFQNRRQYEKKVH
C.neo2             GCKVRRRFTKRELEALEVLW~~~SIAKSPSKYERQRLGAWLGVKTKHITVWFQNRRQEEKRYS
L.bic2             IRKKRKRVDAAQLKVLNETY~~~NRTAFPSTEERHTLAKALDMSARGVQIWFQNKRQSARQTN
C.cin1             SRRTRKRFTNTQLTMLENLF~~~HQTSHPSREEREAVAKAGQMEIKSVTIWFQNKRQTERKSQ
P.chr1             PKKPRHRHSAFQLAALNELY~~~ERDEHPPLEERTSLAERLGMEVKTVNAWFQNKRASTKKRS
P.chr4             VSYGRRRMQPEQLQALQTLY~~~DANTHPTKAQRMQLARELDLDLKSVNVWYQNKRRSMKKKL
P.bla6             IAKRRPRTTPEQSRILNTHF~~~ARNPVPSKNEIKLIAREVKIKPRSTHFWYQNKRASVKREG
CUT1_HUMAN         LKKPRVVLAPEEKEALKRAY~~~QQKPYPSPKTIEDLATQLNLKTSTVINWFHNYRSRIRREL
SIX1_HUMAN         GEETSYCFKEKSRGVLREWY~~~AHNPYPSPREKRELAEATGLTTTQVSNWFKNRRQRDRAAE
SIX3_HUMAN         GEQKTHCFKERTRSLLREWY~~~LQDPYPNPSKKRELAQATGLTPTQVGNWFKNRRQRDRAAA
RenSIX             GEETSYCFKEKSRVVLRQWY~~~TKNAYPSPREKRQLAEQTGLTTTQVSNWFKNRRQRDRAAE
PBX1_HUMAN         ARRKRRNFNKQATEILNEYFYSHLSNPYPSEEAKEELAKKCGITVSQVSNWFGNKRIRYKKNI
RenPBX             ITRTRPVLTRNSLKVLEEWYECHLDHPYPTASQVEWLAQVSSLNTEQVKKWFGNKRSRSKNTR
IRX2_HUMAN         DPAYRKNATRDATATLKAWLNEHRKNPYPTKGEKIMLAIITKMTLTQVSTWFANARRRLKKEN
RenIRO1            SAAGSITRRMRNTAVLVKWIEDHQSNPYPTKAEKQYLAYYSGMNMTQLSTWFANARRRIKKIG
RenIRO2            VQLASSRRRRRDATHLIEWLDLHQGNPYPTRVEKEQLVVISGMNFKQLNDWFANARRNIRKVG
RenIRX3            EKGSSSPGSWRNTDVLALWITEHLQLPYPGKVEKQYLCFYSNMSMKQVSTYFANARR~~~~~~
RenIRO4            CSNDMEARGSEGYKTSGEVVGAHQTNPYPTKAEKECLAECCGMSVKQLCTWFSNSRRQIRKLG
RenIRO5            YDSPRYKLTPERAIPLIKWFEEHKDHPYPSRHEKMLLCQSTQLTFTQVSTWFANARRRMKK~~
TGIF_HUMAN         KRRRRGNLPKESVQILRDWLYEHRYNAYPSEQEKALLSQQTHLSTLQVCNWFINARRRLLPDM
MEI1_HUMAN         RHKKRGIFPKVATNIMRAWLFQHLTHPYPSEEQKKQLAQDTGLTILQVNNWFINARRRIVQPM
RenMEIS            TGKKREKTSPASQKLLKEWLFSHSRCPYPTEDDKQNLCRMTGLSLQQLNNWFINARRRILPQK
MONOSIGA_MEIS1     SRHCTKRFASSSIDTLKEWLFAHTDRPYPTDQDKTELMQQTGLDLMQINNWFINARRRLLVKV
MONOSIGA_MEIS2     NTGGRNNMPHEVTSRLKEWFFAHTSHPYPSEQKKRELASQCDLTLQQINNWFINARRRLLNRP
A.nid4             NRRRRGNLPKPVTEILKAWFHAHLDHPYPSEEDKQMLMSRTGLTINQISNWFINARRRHLPAL
N.cra5             KNKRRGNLPKEVTEKLYAWLYGHLNHPYPTEDEKQKMMRETNMQMNQISNWFINARRRKVPLL
P.bla7             KKRRRGNLPREVTEFLKHWLIQHKAHPYPSEKEKGDLACRTGLTVNQISNWFINARRRILQPM
L.bic4             PQRKRGKLPKETTDYLKAWLHRHSDHPYPSEDEKKQLCHATGLSMSQVSNWMINARRRILAPA
N.cra6             ATKVNNRFSRESIKILKNWLSIHQKHPYPNDEEKEMLQKQTGLSKTQITGWLANARRRRGKVM
A.nid5             ARKSSSRLSREAVRILKAWLNDHSDHPYPTEEEKEELKLRTGLKRTQITNWLANARRRGKIRP
A.nid6             DSKESKQFVRKGARVLRDWFYQNEHCPYPSEEEKARLAAETGFSRQRISTWFANARRRHKQQK
```

Figure S11. Alignment of homeodomain sequences used for Mr. Bayes analysis. *Homo sapiens* homeodomain sequences were taken from the NCBI homeodomain resource. Sponge sequences are labeled with Ren and were found by BLAST of the *Reniera sp.* trace data from the NCBI trace archives. Fungal sequences were obtained from the Broad Institute (A.nid - *Aspergillus nidulans*; C.cin - *Coprinus cinerea*; C.neo - *Cryptococcus neoformans*; N.cra - *Neurospora crassa*; R.ory - *Rhizopus oryzae*) and JGI (A.nig - *Aspergillus niger*; L.bic - *Laccaria bicolor*; P.chr - *Phanerochaete chrysosporium*; P.bla - *Phycomyces blakesleeanus*).

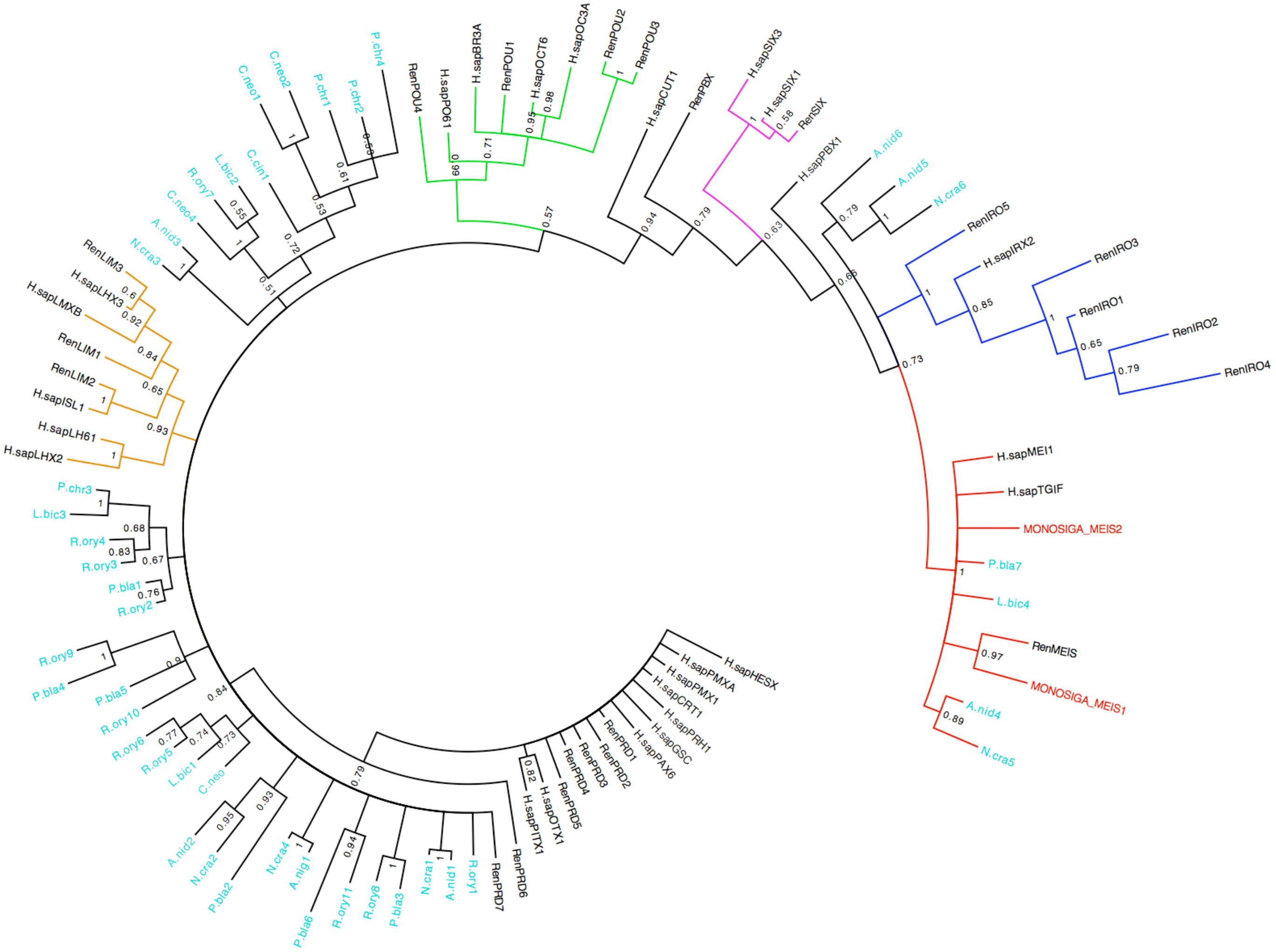


Figure S12. Phylogenetic relationships of representative human, sponge, and fungal homeodomains with the two *M. brevicollis* homeodomains. Analysis was done with Mr. Bayes [2, 3] run with mixed amino acid and inverse gamma settings for 3 million iterations with a burnin of 75,000. The Tree was made using FigTree (*Andrew Rambaut, http://tree.bio.ed.ac.uk/*). Fungal gene labels are in light blue and those from *M. brevicollis* are labeled in red. MEIS class clade is highlighted in red, IRO in dark blue, SIX in purple, POU in green, and LIM in orange.

Table S1. Genome sequencing summary.

Library IDs	Theoretical insert size	Actual insert size	Raw reads	Raw (untrimmed) sequence (Mb)	Passing reads	Quality and vector trimmed sequence (Mb)
AZSO	2-3 kb	3,061 +/- 525	7,620	8	6,599	5
BHUH	2-3 kb	2,365 +/- 355	295,882	314	262,757	185
BAFY	6-8 kb	6,593 +/- 1,284	7,680	8	5,457	4
BNUS	6-8 kb	7,059 +/- 1,769	242,175	235	226,029	165
BAFZ	35-40 kb	38,665 +/- 11,944	3,840	4	3,308	2
BIFH	35-40 kb	36,888 +/- 13,666	77,856	76	46,940	22
Total			635,053	645	551,090	383

Table S2. Supporting evidence for genes models.

Evidence	*M. brevicollis* v.1
Complete models (annotated start and stop codons)	8286 (90%)
Models with EST alignment	4186 (46%)
Models with nr alignment (e-value < 0.1)	7590 (83%)
Models with Swissprot alignment (e-value < 10^{-5})	5877 (64%)
Models with Pfam alignment (gathering threshold)	5160 (56%)

Table S5. Intron gain and loss as calculated by Csuros maximum likelihood.

Branch	Introns Gained	Introns Lost
Eukaryotic → *T. the*	64	157
Eukaryotic → Green plants ancestor	65	52
Green plants ancestor → *A. tha*	73	36
Green plants ancestor → *C. rei*	177	108
Eukaryotic → Opisthokont ancestor	56	23
Opisthokont → Basidomycete ancestor	75	126
Basidiomycete ancestor → *C. neo*	87	80
Basidiomycete ancestor → *P. chr*	32	42
Opisthokont → Holozoan ancestor	61	0
Holozoan ancestor → *M. bre*	69	167
Holozoan → Eumetazoan ancestor	135	23
Eumetazoan ancestor → *N. vec*	12	29
Eumetazoan → Bilaterian ancestor	30	13
Bilaterian ancestor→ *D. mel*	21	397
Bilaterian ancestor → *H. sap*	1	89

Branches shown on the tree in Figure 2 are indicated by the ancestor or extant species at the end of the branch and the ancestor at the last bifurcation. Intron gains and losses were calculated by the Csuros intronRates program[4] with no missing sites assumed and using an unrooted species tree. Holozoan ancestor denotes the ancestor of choanoflagellates and animals. Opisthokont ancestor denotes the ancestor of fungi and holozoans.

Table S4. Functional classification of domains unique to choanoflagellates and metazoans.

Cell Adhesion and Extracellular Matrix

Cadherin*
CUB
Ependymin
Fibrillar collagen C-terminal
HYR*
Kunitz/bovine pancreatic trypsin inhibitor*
Laminin G*
Laminin N-terminal
Reeler
Somatomedin B
Von Willebrand D*

Signal Transduction

Antistasin family
BTK motif
C1q*
CBL proto-oncogene N-term, domain 1
CBL proto-oncogene N-term, EF hand-like
CBL proto-oncogene N-term, SH2-like
ECSIT
Flotilin family
GoLoco motif
Heme NO binding associated
Hormone receptor
L27
Low-density lipoprotein receptor class A
Nine cysteines of family 3 GPCR
Pacifastin inhibitor (LCMII)
Phosphotyrosine binding (IRS-1 type)
Phosphotyrosine interaction (PTB/PID)
PI3-kinase family, p85-binding
Plexin
Raf-like ras-binding
Renin receptor-like protein
S-100/ICaBP type calcium binding
Seven transmembrane receptor, secretin family
SH3 domain-binding protein 5 (SH3BP5)
Spin/Ssty family
TNF (Tumor Necrosis Factor)

Cell Adhesion and Signal Transduction

Leucine rich repeat N-terminal
Immunoglobulin
Immunoglobulin c-2*
Immunoglobulin I-set*
Immunoglobulin V-set*

Transcriptional Control

Mbt repeat
p53 DNA-binding**
PET
STAT protein, DNA binding
Zinc finger, C2HC type

Cytoskeletal Associated

Nebulin repeat
Filament
Repeat in HS1/cortactin
Sarcoglycan complex subunit protein

Transporters/Channels

Dihydropyridine sensitive L-type calcium channel
Inward rectifier potassium channel
Organic anion transporter polypeptide (OATP)
Progressive ankylosis protein (ANKH)

Enzymes

Aspartyl/asparaginyl beta-hydroxylase
DNaseIc*
Cu_2 monooxygenase
Fzo-like conserved region
Galactose-3-O-sulfotransferase
Galactosyl transferase
Glycosyl hydrolase family 59*
Heparan sulfate 2-0-sulfotransferase*
N-acetylglucosaminyltransferase-IV conserved reg.
Phosphomevalonate kinase

Unknown

Assoc. with transcription factors and helicases
Domain of unknown function (DUF758)
Domain of unknown function (DUF837)
Fukutin-related
Hormone-sensitive lipase (HSL) N-terminus
MOFRL family*
N-terminal domain in C. elegans NRF-6
PHR
Protein of unknown function (DUF1241)
Selenoprotein S (SelS)
Translocon-associated protein, δ subunit precursor
Tropomyosin
Uncharacterized protein family (UPF0121)

*Present in bacteria **Partial domain present in *Zea mays* (Qi, 2003)

Table S5. Protein domains unique to choanoflagellates and other groups.

Domain Name	*Interpro ID*
Metazoa, Choanoflagellates, Fungi, and Dictyostelium	
Growth-Arrest-Specific Protein 2 Domain	IPR003108
Protein of unknown function (DUF1183)	IPR009567
Protein of unknown function (DUF1613)	IPR011671
Mss4 protein	IPR007515
UcrQ family	IPR004205
Diaphanous FH3 Domain	IPR010472
WSC domain	IPR002889
TAP C-terminal domain*	IPR005637
RasGAP C-terminus	IPR000593
GGL domain	IPR001770
Ras association (RalGDS/AF-6) domain	IPR000159
I/LWEQ domain	IPR002558
BTG family	IPR002087
Cysteine dioxygenase type I*	IPR010300
Fic protein family*	IPR003812
Fes/CIP4 homology domain (FCH)	IPR001060
GTPase-activator protein for Ras-like GTPase (Ras GAP)	IPR008936
RasGEF	IPR001895
RasGEF, N-terminal motif	IPR000651
Wiskott Aldrich syndrom homology region 2*	IPR003124
Alpha adaptin AP2, C-terminal domain	IPR003164
G-protein gamma like domain (GGL)	IPR001770
BTG domain	IPR002087
Metazoa, Choanoflagellates, and Fungi	
Arfaptin	IPR010504
ATP synthase D chain, mitochondrial (ATP5H)	IPR008689
Cation-dependent mannose-6-phosphate receptor	IPR000296
CP2 transcription factor family	IPR007604
CybS	IPR007992
Cytochrome c oxidase subunit Va	IPR003204
D-ala D-ala ligase C-terminus	IPR011095
Disintegrin	IPR001762
Dolichyl-phosphate-mannose-protein mannosyltransferase	IPR003342
Epoxide hydrolase N terminus	IPR010497
Forkhead domain	IPR001766
FRG1-like family	IPR010414
GDP/GTP exchange factor Sec2p	IPR009449
Golgi phosphoprotein 3 (GPP34)	IPR008628
HRDC (Helicase and RNase D C-terminal) domain	IPR002121
Inhibitor of Apoptosis domain	IPR001370
Microtubule associated	IPR012943
Peptidase C1-like family	IPR004134
Protein of unknown function (DUF1349)	IPR009784
Putative phosphatase regulatory subunit	IPR005036
Receptor L domain	IPR000494
RFX DNA-binding domain	IPR003150
SURF4 family	IPR002995
TEA/ATTS domain family	IPR000818
XPA protein C-terminus	IPR000465

XPA protein N-terminal	IPR000465

Metazoa, Choanoflagellates, and Dictyostelium

Tryptophan 2,3-dioxygenase*	IPR004981
DUF1632	IPR012435
Beta catenin interacting protein (ICAT)	IPR009428
DUF1394	IPR009828
RUN domain	IPR004012
Doublecortin	IPR003533
Translocon assoc. protein, gamma subunit	IPR009779
Hyaluronidase 2*	IPR013618
DUF1736	IPR013618
Fascin*	IPR010431
IRSp53/MIM homology domain (IMD)	IPR013606
Survival motor neuron protein (SMN)	IPR010304
Spectrin	IPR002017
Translocon-assoc protein, gamma subunit (TRAP-gamma)	IPR009779
Follistatin-N-terminal domain-like (FOLN)*	IPR003645

Metazoa and Choanoflagellates

Antistasin family	IPR004094
Aspartyl/asparaginyl beta-hydroxylase	IPR007803
Associated with TFs and helicases	IPR006576
BTK motif	IPR001562
C1q*	IPR001073
Cadherin*	IPR002126
CBL proto-oncogene N-term, domain 1	IPR003153
CBL proto-oncogene N-term, EF hand-like	IPR003153
CBL proto-oncogene N-term, SH2-like	IPR003153
Collagen triple helix	IPR000087
Cu_2 monooxygenase	IPR003153
CUB	IPR000859
Dihydropyridine sensitive L-type calcium channel	IPR000584
DNaseIc*	IPR008185
Domain of unknown function (DUF758)	IPR008477
Domain of unknown function (DUF837)	IPR008555
ECSIT	IPR010418
Ependymin	IPR001299
Fibrillar collagen C-terminal	IPR000885
Filament	IPR001664
Flotillin*	IPR004851
Fukutin-related	IPR009644
Fzo-like conserved region	IPR006884
Galactose-3-O-sulfotransferase	IPR009729
Galactosyl transferase	IPR002659
Glycosyl hydrolase family 59*	IPR001286
GoLoco motif	IPR003109
Heme NO binding associated	IPR011645
Heparan sulfate 2-0-sulfotransferase*	IPR007734
Hormone receptor	IPR000536
Hormone-sensitive lipase (HSL) N-terminus	IPR010468
HYR*	IPR003410
Immunoglobulin	IPR013151
Immunoglobulin c-2*	IPR003598
Immunoglobulin I-set*	IPR013098
Immunoglobulin V-set*	IPR013106

Integrin alpha	IPR013519
Inward rectifier potassium channel	IPR013521
Kunitz/bovine pancreatic trypsin inhibitor*	IPR002223
L27	IPR004172
Laminin G*	IPR001791
Laminin N-terminal	IPR008211
Leucine rich repeat N-terminal	IPR000372
Low-density lipoprotein receptor class A	IPR002172
Mbt repeat	IPR004092
MOFRL family*	IPR007835
N-AcetylglucosaminyltransferaseIV(GnT-IV) conserved region	IPR006759
Nebulin repeat	IPR013998
Nine cysteines of family 3 GPCR	IPR011500
NRF (N-ternminal domain in C. elegans NRF-6)	IPR006621
Organic anion transporter polypeptide (OATP)	IPR004156
p53 DNA-binding	IPR011615
Pacifastin inhibitor (LCMII)	IPR008037
PET	IPR010442
Phosphomevalonate kinase	IPR005919
Phosphotyrosine binding (IRS-1 type)	IPR013625
Phosphotyrosine interaction (PTB/PID)	IPR006020
PHR	IPR012983
PI3-kinase family, p85-binding	IPR003113
Plexin	IPR013548
Progressive ankylosis protein (ANKH)	IPR009887
Protein of unknown function (DUF1241)	IPR009652
Raf-like ras-binding	IPR003116
Reeler	IPR002861
Renin receptor-like protein	IPR012493
Repeat in HS1/cortactin	IPR003134
S-100/ICaBP type calcium binding	IPR013787
Sarcoglycan complex subunit protein	IPR006875
Selenoprotein S (SelS)	IPR009703
Seven transmembrane receptor, secretin family	IPR000832
SH3 domain-binding protein 5 (SH3BP5)	IPR007940
Somatomedin B	IPR001212
Spin/Ssty family	IPR003671
STAT protein, DNA binding	IPR013801
TNF (Tumor Necrosis Factor)	IPR006052
Translocon-associated protein, delta subunit precursor	IPR008855
Tropomyosin	IPR000533
Uncharacterized protein family (UPF0121)	IPR005344
Von willebrand D*	IPR001846
Zinc finger, C2HC type	IPR002515
Fungi and Choanoflagellates	IPR005109
Anp1	IPR005545
YCII-related domain*	IPR005545

*Present in bacteria

Table S6. Species included in comparative protein domain analysis.

Dictyostelium	
Dictyostelium discoideum	*Dictyostelium discoideum AX4*
Fungi	
Aspergillus fumigatus	*Candida glabrata*
Cryptococcus neoformans	*Encephalitozoon cuniculi*
Eremothecium gossypii	*Kluyveromyces lactis*
Saccharomyces cerevisiae	*Schizosaccharomyces pombe*
Yarrowia lipolytica	
Metazoa	
Anopheles gambiae	*Apis mellifera*
Bos Taurus	*Caenorhabditis elegans*
Canis familiaris	*Ciona intestinalis*
Danio rerio	*Drosophila melanogaster*
Gallus gallus	*Homo sapiens*
Macaca mulatta	*Monodelphis domestica*
Mus musculus	*Pan troglodytes*
Rattus norvegicus	*Takifugu rubripes*
Tetraodon nigroviridis	*Xenopus tropicalis*
Unicellular eukaryotes	
Cryptosporidium hominis	*Cyanidioschyzon merolae*
Debaryomyces hansenii	*Giardia lamblia*
Monosiga brevicollis	*Plasmodium falciparum*
Thalassiosira pseudonana	

Genomes of these species were used in the initial analysis of the phylogenetic distribution of *M. brevicollis* protein domains. The phylogenetic distributions of domains classified by this analysis as unique to choanoflagellates and another phylogenetic group were manually annotated using the Pfam and SMART online databases.

Table S7. Immunoglobulin domains are restricted to choanoflagellates and metazoans.

	Metazoa			Choanoflagellates	Fungi		Dictyostelia	Plants
	Hsap	*Cint*	*Dmel*	*Mbre*	*Ccin*	*Ncra*	*Ddis*	*Atha*
Immunoglobulin*	1502	144	503	5	0	0	0	0

*Total number of immunoglobulin (Ig)-type domains (Ig, Ig-like, Ig c1-set, Ig subtype 2, Ig v-set) predicted by SMART.

Table S8. Intercellular signaling pathways across phyla.

		Animals				Choanozoa	Fungi				Amoebazoa	Plant
Pathway	Component	*Hsap*	*Cint*	*Dmel*	*Nvec*	*Mbre*	*Rory*	*Ncra*	*Scer*	*Ccin*	*Ddis*	*Atha*
NHR												
	ROR	●	●	●	⊙	○	○	○	○	○	○	○
	Hnf4	●	●	●	⊙	○	○	○	○	○	○	○
	Err	●	●	●	⊙	○	○	○	○	○	○	○
WNT												
	Wnt	●	●	●	●	○	○	○	○	○	○	○
	Fzd	●	●	●	●	○	○	○	○	○	●	○
	Dsh	●	●	●	●	○	○	○	○	○	○	○
TGFβ												
	ALK	●	●	●	●	○	○	○	○	○	○	○
	TGFβr	●	●	●	●	○	○	○	○	○	○	○
	Smad	●	●	●	●	○	○	○	○	○	○	○
NFKβ/Toll												
	NFKβ	●	●	●	●	○	○	○	○	○	○	○
	Tlr	●	⊙	●	●	○	○	○	○	○	○	○
	Tollip	●	●	○	●	●	○	○	○	○	○	○
JAK/STAT												
	Jak	●	●	●	⊙	○	○	○	○	○	○	○
	Stat	●	●	●	⊙	⊙	○	○	○	○	⊙	○
Notch												
	Notch	●	●	●	●	⊙	○	○	○	○	○	○
	Delta	●	●	●	⊙	○	○	○	○	○	○	○
	Presenilin	●	●	●	●	●	●	○	○	○	●	●
	Furin	●	●	●	●	⊙	○	○	○	○	○	○
	TACE	●	●	●	●	⊙	○	○	○	○	○	○
Hedgehog												
	Ptc	●	●	●	●	●	○	⊙	⊙	○	⊙	⊙
	Hh	●	●	●	●	⊙	○	○	○	○	○	○
	Smo	●	●	●	●	○	○	○	○	○	⊙	○
	Fu	●	●	●	●	●	○	○	○	○	●	○
RTK												
	Rtk	●	●	●	●	●	○	○	○	○	○	○

A filled circle (●) indicates presence of a homolog with strong similarity. A partially filled circle (⊙) indicates a gene with partial similarity (e.g. contains some but not all domains diagnostic of that protein). An open circle (○) indicates no homologs found. ROR, Retinoid-related orphan receptors ; Hnf4, Hepatocyte nuclear factor 4 ; ERR, Estrogen-Related Receptor; Fzd, Frizzled; DSH Disheveled; ALK, Activin-Like Kinase *TGFβ*r, *TGFβ receptor;* SMAD, SMA/MAD Mothers Against Decapentaplegic; Tlr, Toll-like receptor; Jak, Janus Kinase; Stat, ; DSL, Delta Serrate Lag-2, Ptc, Patched; Hh, Hedgehog; Smo, Smoothened; Fu, Fused; Sufu, Suppressor of Fused, Rtk, Receptor Tyrosine Kinase.

Table S9. *M. brevicollis* presents a key intermediate in the evolution of MAPK signaling.

		Animal		*Choanoflagellate*	Fungi		Dictyostelia
Kinase		*H.sap*	*N.vec*	*M.bre*	*S.cer*	*N.cra*	*D.dis*
MAPKKK	MEKK1	●	●	●			
	MEKK2	●	●	●			
	MTK1(MEKK4)	●	●				
	ASK (MEKK5-7)	●	●	●			
	MEKK15	●	●				●
	Mos	●	●				
	Raf	●	●				
	LZK (MEKK12-13)	●	●	●			
	MLK (MEKK9-11)	●	●	●			
	TAO	●	●	●			
	UNCLASSIFIABLE		●	●	●	●	●
MAPKK	MKK1	●	●	●	●	●	●
	MKK5	●	●	●			
	MKK3	●	●				
	MKK4	●	●				
	TOPK	●	●	●			
	UNCLASSIFIABLE			●	●	●	
MAPK	ERK	●	●	●	●	●	●
	ERK5	●	●	●			
	p38	●	●	●	●	●	
	JNK	●	●				
	ERK3	●	●				
	ERK7	●	●	●			●
	NMO	●	●				
	UNCLASSIFIABLE				●	●	

Sequence analysis of the three tiers of kinases from the MAPK module in metazoans (human, sea anemone (*Nvec*; *Nematostella vectensis*), choanoflagellate (*M. brevicollis*), fungi (*S.cer*: *Saccharomyces cerevisiae*; *N.cra*: *Neurospora crassa*) and slime mold (*Dictyostelium discoideum*) shows the emergence of MAPK modules in choanoflagellates and lower meatzoans. Kinase subfamilies on the left are from the classification given at kinase.com, based on human kinases.

Table S10. Basal transcription factors present in *M. brevicollis*.

Basal Machinery		*H. sap*	*D. mel*	*M. bre*	*S. cer*
RNA polymerse II	Rpb1	●	●	●	●
	Rpb2	●	●	●	●
	Rpb3	●	●	●	●
	Rpb4	●	●	⊙	●
	Rpb5	●	●	●	●
	Rpb6	●	●	●	●
	Rpb7	●	●	●	●
	Rpb8	●	●	●	●
	Rpb9	●	●	●	●
	Rpb10	●	●	●	●
	Rpb11	●	●	●	●
	Rpb12	●	●	●	●
	TBP	●	●	●	●
	TBP 2	○	○	●	○
	TFIIA -L	●	●	●	●
	TFIIA -S	●	●	⊙	●
	TFIIB	●	●	●	●
	TFIIE-L	●	●	●	●
	TFIIE-S	●	●	●	●
	TFIIF-L	●	●	○	●
	TFIIF-S	●	●	●	●
TFIIH	XPB	●	●	●	●
	XPD	●	●	●	●
	p62	●	●	○	●
	p52	●	●	○	●
	p44	●	●	●	●
	p34	●	●	○	●
	cdk7	●	●	●	●
	cyclin H	●	●	●	●
	Mat1	●	●	●	●
	p8	●	●	○	
Co-activators		·	·		
	PC4	●	●	●	
TFIID	TAF1	●	●	●	●
	TAF2	●	●	●	●
	TAF3	●	●	○	●
	TAF4	●	●	○	●
	TAF5	●	●	●	●
	TAF6	●	●	○	●

	TAF7	●	●	○	●
	TAF8	●	●	○	●
	TAF9	●	●	●	●
	TAF10	●	●	●	●
	TAF11	●	●	○	●
	TAF12	●	●	○	●
Mediator	MED1	●	●	○	●
	MED2	○	○	○	●
	MED3	○	○	○	●
	MED4	●	●	○	●
	MED5	○	○	○	●
	MED6	●	●	●	●
	MED7	●	●	●	●
	MED8	●	●	●	●
	MED9	●	●	○	●
	MED10	●	●	○	●
	MED11	●	●	○	●
	MED12	●	●	○	●
	MED13	●	●	○	●
	MED14	●	●	○	●
	MED15	●	●	○	●
	MED16	●	●	○	●
	MED17	●	●	○	●
	MED18	●	●	○	●
	MED19	●	●	○	●
	MED20	●	●	○	●
	MED21	●	●	●	●
	MED22	●	●	○	●
	MED23	●	●	○	○
	MED24	●	●	○	○
	MED25	●	●	○	○
	MED26	●	●	○	○
	MED27	●	●	○	○
	MED28	●	●	○	○
	MED29	●	●	○	○
	MED30	●	●	○	○
	MED31	●	●	○	●
Chromatin Transactions		·	·	·	·
	CBP(p300)	●	●	⊙	○
	GCN5	●	●	●	●
	ISWI	●	●	●	●
	SWI/SNF	●	●	●	●
	Osa	●	●	⊙	
Elongation factors					
	TFIIS	●	●	●	●
	PAF-1	●	●	●	●

	NELF	●	●	●	
	DSIF	●	●	●	●

Key: ● - present, ⊙ - weak alignment but present, ○ - absent or unidentifiable. Species abbreviations: *H. sap* - *Homo sapiens*, *D. mel* - *Drosophila melanogaster*, *M. bre* - *Monosiga brevicollis*, *S. cer* - *Saccharomyces cerevisiae*.

Table S11: Number of *M. brevicollis* protein models containing transcription factor family specific domains.

Transcription Factor Family	Pfam Domain Id	No. protein models containing domain
BolA-like	PF01722	1
Cold-shock DBD	PF00313	1
HTH	PF01381	1
PC4	PF02229	1
PAH	PF02671	1
STAT DBD	PF02864	1
Tubby-like	PF01167	1
Homeobox	PF00046	2
HSF DBD	PF00447	2
p53 DBD	PF00870	2
RFX DBD	PF02257	2
ZnF NF-X1	PF01422	2
E2F TDP DBD	PF02319	3
MADS/SRF type	PF00319	4
FH	PF00250	8
bZIP	PF07716, PF00170	12
HLH	PF00010	13
Myb DBD	PF00249	14
ZnF CCCH	PF00642	16
ZnF C2H2	PF00096	68
Total:		**155**

bZip: basic-leucine zipper; DBD: DNA binding domain; E2f-TDP: E2F/DP (dimerizaton partner) family winged-helix DNA-binding domain; FH: forkhead; Hbx: homeobox; HLH: helix-loop-helix; HSF: heat shock factor; HTH: helix-turn-helix; PAH: paired amphipathic helix; RFX: regulatory factor X; SRF: serum response factor; STAT: signal transducer and activator of transcription; ZnF: Zinc finger.

Supplementary Notes

S1. Genome sequencing and assembly

S1.1 Pilot sequencing efforts. The bacterivorous lifestyle of choanoflagellates and the lack of robust axenic cultures presented a challenge for the production of a high quality genome sequence and assembly. Pilot sequencing from total genomic DNA preparations (containing both bacterial and *M. brevicollis* DNA) revealed that over 80% of the DNA was bacterial, meaning that coverage of the choanoflagellate genome would be insufficient for a quality assembly. We therefore employed two strategies for dealing with bacterial contamination prior to sequencing: (1) reduction of bacterial diversity in cultures and (2) separation of bacterial and choanoflagellate DNA after DNA isolation. Using physical separation techniques combined with antibiotic treatments, a culture line with only a single contaminating bacterial species, *Flavobacterium* sp, was developed. The GC content of *Flavobacterium* (33%) is sufficiently different from that of *M. brevicollis* (55%) to allow separation of the two genomes over a CsCl gradient. *M. brevicollis* genomic DNA isolated in this manner was used to construct replicate libraries containing inserts of 2-3 kb, 6-8 kb, and 35-40 kb, each of which was used for paired end shotgun sequencing. The estimated fractions of bacterial clones in the main libraries (BHUH, BIFH, BNUS) ranged from 3% - 12% and sequences from these clones assembled almost entirely into a single 4.2 Mb scaffold, presumably representing the full genome of *Flavobacterium* sp.

S1.2 Generation of a monoxenic *M. brevicollis* culture, MX1. *M. brevicollis* (ATCC 50154) grown with mixed bacteria was propagated at 25°C in ATCC 1525, growth media prepared by infusing seawater with Ward's Cereal Grass Media (Ward's Natural Science) until the culture reached stationary growth (four days). To reduce the bacterial diversity, the culture was treated with 50ug/mL streptomycin, 50 ug/mL kanamycin, and 12.5 ug/mL chloramphenicol, supplemented with γ-irradiated *Enterobacter aerogenes,* and then cultured in the dark with gentle shaking for 48 hours. The culture was split and the antibiotic treatment was repeated four additional times. The antibiotic-treated culture was pelleted at 4K rpm, 20 min, 15°C and cultured for 48 hours in antibiotic free ATCC 1525 media, during which there was no apparent bacterial proliferation. Cells from an isolated colony of *Flavobacterium sp.* were then added to the culture to support choanoflagellate growth. The culture was further sterilized via a U-tube technique of migration-dilution adapted from Claff, 1940[5]. Briefly, 15mL of culture were concentrated by centrifugation at 6k rpm for 10 min at 25°, and then resuspended in 5mL of ATCC 1525 media. The concentrated culture was placed in the first well of a six well plate, which was connected by three sterile glass U-shaped tubes to the adjacent well filled with fresh ATCC 1525 media. After 48 hours, the culture in the second well was supplemented with cells from a colony of *Flavobacterium sp.* The resulting culture, MX1, was shown to be monoxenic by PCR amplification, cloning and sequencing of multiple independent bacterial 16S rRNA clones using the following primer set: 5'- AGA GTT TGA TCC TGG CTC AG-3' and 5'-ACC TTG TTA CGR CTT-3', modified from Weisburg et. al, 1991[6]. All clones were identical and related to 16S sequences from bacteria in genus *Flavobacterium*. Members of this genus have GC contents ranging from 31.6%-50.0% [7]**.**

S1.3 Isolation of *M. brevicollis* genomic DNA. *M. brevicollis* MX1 was grown to a density of 10^7cells/mL in ATCC 1525 media and 750mL of culture was pelleted by two rounds of centrifugation at 10K rpm for 30 min at 4^{o}C. Cell pellets were frozen at –80^{o}C and ground to a fine powder under liquid N_2. *M. brevicollis* genomic DNA (at this point contaminated with *Flavobacterium sp.* genomic DNA) was isolated with the Puregene® DNA purification system (Gentra Systems). The *M. brevicollis* genomic DNA was separated from the contaminating *Flavobacterium sp.* DNA via CsCl density gradient ultracentrifugation. Briefly, 2280ug of contaminated genomic DNA was centrifuged to equilibrium (65K rpm for 40hrs) on six gradients of 1.69g/mL CsCl, in the presence of 40ug/mL of the dye Hoechst 33258 (Molecular Probes). The lower of two resulting bands in each gradient was recovered and the DNA was separated from the Hoechst dye by five extractions with NaCl-saturated n-butanol. The CsCl was dialyzed out of the DNA solution through Spectra/Por® MWCO 8000 dialysis tubing (Spectrum Laboratories, Inc.) over 50 hours at 4^o C. The purified *M. brevicollis* genomic DNA was rescued from the dialysis tubing and then ethanol precipitated using Pellet Paint® Co-precipitant (Novagen). The final yield was 24ug of purified *M. brevicollis* genomic DNA, representing a 1% recovery from the initial amount of contaminated genomic DNA. This process was repeated to obtain a sufficient amount of choanoflagellate genomic DNA to build the DNA libraries necessary for sequencing.

S1.4 Genome assembly and validation. The initial data set was derived from 6 whole-genome shotgun (WGS) libraries: two with theoretical insert sizes of 2-3 KB, two with theoretical insert sizes of 6-8 KB, and two with theoretical insert sizes of 35-40 KB (Table S1). The reads were screened for vector using Cross_match (http://www.phrap.org/phredphrap/phrap.html), then trimmed for vector and quality[8]. Reads shorter than 100 bases after trimming were excluded.

The data was assembled using release 2.9.2 of Jazz, a WGS assembler developed at the JGI[8-10]. A word size of 13 was used for seeding alignments between reads. The unhashability threshold was set to 40, preventing 13-mers present in the data set in more than 40 copies from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together sequences that are more than about 97% identical. The genome size and sequence depth were initially estimated to be 50 MB and 8.0, respectively.

S1.5 Assembly analysis and quality control. The initial assembly contained 47.4 MB of scaffold sequence, of which 3.7 MB (7.8%) was gaps. There were a total of 1,151 scaffolds, with a scaffold N/L50 of 13/1.10 MB, and a contig N/L50 of 220/52.4 KB. (N50 is the number of pieces (scaffolds or contigs) that account for 50% of the assembly; L50 is the minimum length of these pieces). The assembly was then filtered to remove short and redundant scaffolds:

- Short scaffolds were defined as those with < 1 KB total length.
- Redundant scaffolds were defined as those with < 5 KB total length, where > 80% matched a scaffold that was > 5 KB total length in a single, BLAT-determined alignment (Kent 2002), at any % ID.

After excluding redundant and short scaffolds, there remained 46.0 MB of scaffold sequence, of which 3.4 MB (7.4%) was gaps. The filtered assembly contained 232 scaffolds, with a scaffold N/L50 of 12/1.13 MB, and a contig N/L50 of 210/53.3 KB. The sequence depth derived from the assembly was 8.45 ± 0.09.

There were 107,459 reads that were not placed in the assembly for various reasons, 13,215 of which were excluded due to quality/vector trimming. Of the remaining 94,244 unplaced reads, the overwhelming majority (~95%) had GC contents that suggested they were part of the *M.brevicollis* genome. The unplaced reads whose mean GC contents were greater than 40% contained roughly 14 MB of trimmed sequence. If this sequence were at the same depth as the rest of genome, it would correspond to roughly 1.7 MB of genome, and so could account for at most about half of the gap sequence. The remainder of the gaps could consist of uncloned segments of the genome, the short/redundant scaffolds, mis-estimates of the gap sizes, or other mis-assembly-related issues.

To estimate the completeness of the original assembly (i.e. including short and redundant scaffolds), a set of 29,246 *M. brevicollis* ESTs was BLAT-aligned to the unassembled trimmed data set, as well as the original assembly itself[11]. 28,821 ESTs (98.5%) were more than 80% covered by the raw sequence data, 29,053 (99.3%) were more than 50% covered, and 29,139 (99.6%) were more than 20% covered. By way of comparison, of the 29,019 ESTs (99.2%) that had BLAT alignments to the original assembly, 28,387 (97.1%) were more than 80% covered by scaffold alignments, 28,866 (98.7%) were more than 50%, and 28,987 (99.1%) were more than 20% covered.

The mitochondrial genome was available before the assembly was run[12] and was used to identify the corresponding organelle scaffolds. There were three such scaffolds (scaffold IDs 243, 254, and 558) in the released assembly. These scaffolds were excluded from the subsequent genome annotation.

To identify additional contaminant scaffolds, a “kitchen-sink” megablast against the NCBI nt database was performed (using the following parameters: -D 2 -z 1e9 -F "m D" -b 100 -v 100 -p 90 -e 1e-10). The resulting alignments were partitioned by top-level NCBI taxonomic classification: Archaea, Bacteria, Eukaryota, Viroids, Viruses, Other, and Unclassified. The last four were grouped together as “Non-Cellular”, while Archaea and Bacteria were lumped together as ”Prokaryotic”. Each scaffold was then tentatively classified based on the distribution of its hits between these three larger categories. Scaffolds with only Eukaryota hits, or no alignments at all, were assumed to be part of the main genome. Scaffolds with some (or all) of their alignments in the other categories had those hits manually examined to determine how reliable they were likely to be. Low-quality hits, or ones to sequences that were probably mislabeled in NCBI, were discounted, and the scaffolds were reclassified based on the remaining ones.

Six scaffolds had various types of non-cellular alignments. Examination of these alignments revealed that four of these scaffolds were almost certainly part of the main genome, due to the nature of the hits themselves, and extensive additional alignments to *M.brevicollis* ESTs. One of the scaffolds (scaffold ID 58) was confirmed as non-cellular material, as it was entirely covered by high % ID alignments to various types of cloning vector. The final scaffold in this set (scaffold ID 170) was tagged as mis-assembled, as it was a chimera of sequences that aligned (on one side) to cloning vectors and *E.coli*, and on the other to eukaryotic sequences. The non-cellular and mis-assembled scaffolds were excluded from the subsequent genome annotation.

Five scaffolds had a combination of eukaryotic and prokaryotic BLAST hits. Examination of the details of these alignments, along with hits to the *M.brevicollis* ESTs, indicated that four of the five (scaffold IDs 16, 31, 43, and 49) were probably part of the main genome. The fifth (scaffold ID 243) was separately determined to be part of the mitochondrion; see above for details.

Two scaffolds had only prokaryotic hits to the NCBI nt database. Examination of the alignments, and the fact that their GC contents were consistent with the known low-GC prokaryotic contaminant, indicated that they were true prokaryotic scaffolds. One of these scaffolds (scaffold ID 1) was 4.2 MB in length and, as mentioned above, likely represents almost the entire genome of the prokaryotic contaminant.

Finally, seven additional scaffolds (scaffold IDs 56, 62, 99, 171, 221, 233, and 460), while not having any BLAST hits to the NCBI nt database, had GC contents consistent with the known prokaryotic contaminant. Five of these scaffolds (62, 99, 171, 221, and 460) had no BLAT alignments to the *M.brevicollis* ESTs, and so were immediately moved into the prokaryotic contaminant category. The other two scaffolds had some EST alignments (scaffold 56: 75 EST alignments; scaffold 233: 9 EST alignments). However, as even the largest confirmed prokaryotic scaffold had seven EST alignments, the remaining two low-GC scaffolds were moved into the prokaryotic category as well. All of the prokaryotic scaffolds were excluded from the subsequent genome annotation. After the removal of these and the other scaffolds mentioned above, 218 putative nuclear scaffolds remained.

S1.6 No detectable single nucleotide polymorphism in *M. brevicollis*. To characterize the level of variation in the population isolate of *M. brevicollis* that was used for sequencing, we searched for single nucleotide polymorphisms (SNPs) among the whole-genome shotgun (WGS) and expressed sequence tag (EST) reads generated by the sequencing project. Raw sequencing reads were trimmed for vector and quality as described above (S1.4 Genome assembly and validation), leaving 551,090 WGS reads and 29,246 reads available for comparison. To determine the overlapping positions that could be used for SNP detection, we aligned trimmed reads against the JGI *M. brevicollis* genome assembly v1.0 using BLAT v. 32[11] with default parameters. A total of 495,647 WGS reads and 28,997 EST reads were successfully mapped to genomic scaffolds. We applied two filters to eliminate incorrect read alignments. First, to ensure unique alignments, we only accepted the best alignment for a read if the ratio between the BLAT score of the second highest scoring alignment and the BLAT score of the highest scoring alignment was no greater than 0.8. Second, we required that paired end reads from the same insert align on the opposite strand to the same genomic scaffold, and within the insert size of the library from which the reads were sequenced. After this filtering step, 388,890 WGS reads and 20,934 EST reads remained for SNP detection.

To produce tractable sets of reads for multiple sequence alignment, we divided the genome into 5 kilobase segments, and produced alignments for each segment using all passing reads either partially or fully included in the segment. Repetitive regions of the genome that have been incorrectly collapsed by the assembly process would cause spurious SNPs to be detected, as reads from two different regions of the genome would be included and aligned within the same segment. To eliminate such segments from consideration, we counted the number of reads mapped by BLAT within each segment with greater than 300 matches to the segment, including all alignments from all trimmed reads, as the uniqueness criterion may have eliminated reads from potentially repetitive regions. More than 90% of segments contained between 0 and 100 reads, and we rejected segments containing 100 or more reads (the average number of reads in a rejected segment was 747). We created multiple sequence alignments for passing segments using MAP[13], with a match score of 1, a mismatch score of -2, a gap open cost of 4, a gap extension cost of 3, and a gap limit of 5. To remove alignment artifacts caused by simple repetitive sequence, we did not consider bases within regions detected by Tandem Repeats Finder version 4.00[14], run with the default parameters. We eliminated low quality regions within reads by applying the quality criteria of the Neighbourhood Quality Standard [15, 16]. Any positions with at least two different alleles passing NQS(25, 20) were considered to be putative SNPs. Using our technique, it is also possible to discover insertions or deletions among WGS and EST reads. However, such differences are significantly more likely to be artifacts of alignment or incorrect base calling, and so we chose to focus our initial variation discovery efforts on SNPs.

We discovered 6,313 putative SNPs among the combined WGS and EST reads, or roughly one SNP per 6,595 sequenced bases. However, the distribution of putative SNP positions in the genome was highly non-uniform, with 4,585 of the putative SNPs within 100 bases of each other. While it is possible that this distribution of SNPs is caused by inhomogeneity in mutation rate or exists due to the action of positive or negative selection, the simplest explanation is that the SNPs within 100 bases of each other are artifacts of over-collapsed regions within the genome assembly that were able to escape our filtering process. Manual examination of 20 randomly selected segments containing two or more SNPs within 100 bases of each other confirmed that all such segments were the result of comparison between two different genomic regions. After eliminating such segments from consideration, only 1,478 putative SNPs remained. In addition, none of these putative SNP positions had more than one read carrying the alternate allele, implying either that all putative SNPs were artifacts of the cloning and sequencing process or that they were present at very low allele frequencies. Manual examination of 20 randomly selected SNPs from the remaining 1,478 putative SNPs revealed 9 of the SNPs to be errors made by the base caller. To investigate the remaining 11 randomly selected SNPs that were not base calling errors, we designed PCR amplicons of roughly 650 bases in length flanking each of the SNPs, and performed PCR followed by sequencing for each amplicon in 4 separate populations of *M. brevicollis*. None of the putative SNP positions was polymorphic in any of the sequenced populations, and no detectable variation was present at any other position within the amplified regions. Thus, our results are consistent with a lack of single nucleotide polymorphism in the sequenced isolate of *M. brevicollis*, although it is formally possible that there is extremely rare variation that our methodology was unable to detect.

S1.7 Mode of reproduction and ploidy of *M. brevicollis* remain unknown. We could not use the lack of variation detected in *Monosiga* to infer ploidy or to determine mode of reproduction. Two strong population bottlenecks occurred in the demographic history of the sequenced culture: one at the initial isolation of Monosiga and another during the preparation of a monoxenic strain for sequencing (Supp. Notes S1.2). These bottlenecks may have reduced the population size to two or fewer individuals, and were sufficient to obscure any signal in variation that could have been used to make inferences regarding ploidy or sex. Although our lab cultures were rapidly expanded following both bottlenecks, they retained a small effective population size[17]. Therefore, genetic drift could have quickly eliminated variation completely in either a haploid or a diploid population, given that the relative difference in rate of reduction of heterozygosity is only two-fold [18].

S2. Joint Genome Institute (JGI) annotation of the genome. The JGI annotation pipeline takes multiple inputs (scaffolds, repeats, and ESTs) and produces annotated gene models and other features that are deposited in a database. The data can be accessed by the public through the JGI *M. brevicollis* genome portal at http:www.jgi.doe.gov/Mbrevicollis.

Before gene prediction, the 218 scaffolds were masked using RepeatMasker (http://www.repeatmasker.org/) and a custom repeat library of 108 putative transposable elements, which are available on the *M. brevicollis* genome portal downloads page. After masking, a variety of gene prediction programs were deployed, based on a variety of methods. These were 1) the *ab initio* method FGENESH[19] (Softberry Inc., NY, USA), the homology-based methods FGENESH+[19] (Softberry Inc., NY, USA) and GeneWise[20] seeded by BLASTx alignments against sequences of all opisthokont entries in the GenBank nonredundant protein database as of May 2006, and 3) mappings of EST cluster consensus sequences from *M. brevicollis* produced using EST_map (Softberry Inc., NY, USA). EST clusters were assembled using single link clustering at 98% identity. Both the JGI ESTs and ESTs from ChoanoBase (http://mcb.berkeley.edu/labs/king/blast/) were used to assemble clusters.

GeneWise models were completed by using scaffold data to find in frame upstream start and downstream stop codons. EST clusters were used to extend, verify, and complete the predicted gene models using custom scripts (estExt, I. Grigoriev, unpublished). The resulting set of models was then filtered for the “best” models, based on criteria of completeness, length, EST support, and homology support, to produce a non-redundant representative set. This representative set was subject to community-wide manual curation and comparative genomics studies.

9196 non-redundant gene predictions constitute release 1.0. The majority of these genes (87%) were predicted by the *ab initio* method FGENESH using a parameterization based on *M. brevicollis* full-length mRNAs and EST cluster consensus sequences that appeared to contain a full open reading frame. Only 13% of gene structure models were predicted using homology-based methods, specifically FGENESH+ and GeneWise using peptides from GenBank to seed the non-redundant database (Supp. Table S1). When possible, these predictions were corrected and/or extended using ESTs. A small number of gene models (< 1%) were predicted based only on clusters of overlapping ESTs that consistently aligned to the genome and had substantial open reading frames.

Though many genes were predicted by *ab initio* methods, the gene catalog is supported by other evidence (Supp. Table S2). 90% of the predicted genes are complete models in the sense of having start and stop codons, 83% of the gene catalog aligns with proteins in the GenBank nr database (e-value < 0.1) and 56% of the predicted genes possess Pfam domains. Furthermore, 46% of the gene catalog is consistent with the ESTs collected from exponentially growing *M. brevicollis*.

All predicted gene models were annotated for protein function using domain prediction tool InterProScan[21] and hardware-accelerated double-affine Smith-Waterman alignments (http://www.timelogic.com) against Swiss-Prot[22], KEGG[23], KOG[24]. Then KEGG hits were used to map EC numbers, and EC, Interpro, and Swiss-Prot hits were used to map Gene Ontology (GO) terms[25]. In addition we ran SignalP[26] and TMHMM[27] for analysis of protein localization.

We predicted that 2,030 proteins (22%) possess a leader peptide, 2,100 proteins (23%) possess at least one transmembrane domain, and 1,132 (12%) possess both. We assigned 1,843 distinct GO terms to 4,834 proteins (53%) using EC-to-GO, Swiss-Prot-to-GO, and InterPro-to-GO mappings (http://www.geneontology.org/GO.indices.shtml). We also assigned 1,952 proteins (21%) to KEGG pathways, with a total of 640 distinct EC numbers. The top 4 most populated KEGG pathways are amino acid, complex carbohydrate, carbohydrate, and complex lipid metabolism (436, 387, 289, and 377 proteins, respectively). The complex carbohydrate metabolism pathway includes nearly 200 proteins devoted to the KEGG map starch and sucrose metabolism (MAP00500). Finally, we assigned 6883 proteins (75%) to 3389 KOGs.

S3. Analysis with an evolutionary perspective

S3.1 Phylogenetic Analysis. A previously published 32-species, 50-gene data matrix[28] containing metazoan, choanoflagellate and fungal species was updated with the orthologous genes from the *M. brevicollis* genome. Additionally, the corresponding orthologous genes from a fungus (*Rhizopus oryzae*, phylum Zygomycota), a plant (*Arabidopsis thaliana*), and two protists (*Entamoeba histolytica* and *Dictyostelium discoideum*) were added to increase taxonomic diversity in the data matrix. Orthology was established by the reciprocal best BLAST hit criterion[29]. Specifically, each gene from each of the additional species was considered a true ortholog if it was the best reciprocal BLAST hit with the corresponding gene in *Homo sapiens*.

All analyses were performed on the amino acid sequences. Genes were aligned with CLUSTALW[30]. Indels and areas of uncertain alignment were excluded from further analysis. Phylogenies were estimated using maximum likelihood (ML) and maximum parsimony (MP), using PHYML[31] and PAUP*[32], respectively (Supp. Fig. S1). Support was assessed using bootstrap re-sampling with 100 replicates (Supp. Fig. S1). For ML, the model of amino acid evolution utilized was estimated by PROTTEST[33] and enforced in all subsequent analyses. The best-fit model for the 50-gene data matrix was WAG[34], with rate heterogeneity among sites (value of the gamma shape parameter alpha = 0.87) and a proportion of sites set to be invariable (value = 0.16). MP analyses were performed with all sites equally weighted and with tree-bisection-reconnection branch swapping. Data matrices and trees are available from the authors on request.

S3.2 Gene structure statistics. *M. brevicollis* gene structure statistics are based on the JGI filtered models gene set. The gene structure statistics for other species were found on their respective genome browser websites:

N. vectensis: http://genome.jgipsf.org/Nemve1 /Nemve1.home.html;
C. intestinalis: http://genome.jgi-psf.org/Cioin2/Cioin2.home.html;
N. crassa: http://www.broad.mit.edu/annotation/genome/neurospora/;
C. cinereus: http://www.broad.mit.edu/annotation/genome/coprinus_cinereus;
D. discoideum: http://dictybase.org

An exception was *A. thaliana*, for which gene structure statistics were taken from a comparative genome paper[35]. Many of the *N. vectensis* gene models in the current release are incomplete (N. Putnam, personal communications), so the statistics given are based on a set of over 1,000 genes whose structures are known from full length mRNA. The estimated gene number was taken from the *Nematostella vectensis* genome paper[36].

S3.3 Intron evolution. To study intron loss and gain in orthologous genes in multiple species, we aligned *M. brevicollis* genes to human (ENSEMBL models release 26.35.1), *Drosophila melanogaster* (BDGP4 ENSEMBL model release 41), *Nematostella vectensis* (JGI v1.0), *Phanerochaete chrysosporium* (JGI v2.0), *Cryptococcus neoformans A* (Broad Institute v3.0), *Arabidopsis thaliana* (TIGR release 5), *Chlamydomonas reinhardtii* (JGI v3.0), and *Tetrahymena thermophila* (TIGR, 2005) genes. In 473 cases, a human gene was found to have a mutual best hit to a gene from each of the other nine species, forming a tentative cluster of orthologous genes to be studied further. We also analyzed introns positions from a subset of these species: *Arabadopsis thaliana*, *Cryptococcus neoformans A*, *M. brevicollis*, *N. vectensis*, and *H. sapiens*. This allowed us to analyze a larger number of intron positions than was possible with the nine species data set. In this subset, 538 human genes had mutual best blast hits to each of the other species. Notably, the average numbers of introns per gene in this set of highly conserved genes was different from the average numbers of introns per gene for the entire genomes (12.4 vs. 7.7 introns/gene in humans, 11.7 vs. 5.8 in *N. vectensis*, 8.8 vs. 6.6 introns/gene in *M. brevicollis*, 6.5 vs. 5.3 in *C. neoformans*, and 8.8 vs. 4.4 in *A. thaliana*).

Gene models are often incomplete in the 5' ends and may have poorly determined splice sites, so we restrict our analysis to regions of highly conserved peptides in the orthologs of all five species. The independent identification of such regions in multiple species provides strong evidence for the accuracy of the gene models in these regions. We built multiple alignments of the orthologous clusters using ClustalW and identified gap-free blocks flanked by fully conserved amino acids. We then identified the annotated splice sites within these regions for all the species, with the additional requirements that 1) none of the peptides have a gap in the alignment closer than 3 amino acids from the splice site and 2) no two different peptides have splice sites at different positions closer than 4 amino acids. Empirically, these requirements are necessary to avoid spurious detection of intron gains and losses due to ambiguities in either the multiple alignment or the gene models' splice sites. Finally, we required that at least 5 amino acids out of 10 in the flanking regions of the splice sites be either fully conserved or have strong functional similarity among all species. In the set of genes from all nine species 1,989 intron splice sites at 1,054 highly reliable positions were identified by these requirements. In the five species set 3,847 intron splice sites at 2,121 conserved positions were identified. Presence or absence of introns at these positions across the two sets of taxa was used to build binary character matrices.

Several methods have been developed to infer the evolutionary history of introns in orthologous genes. To gain a comprehensive view of the possible scenarios of intron evolution in *M. brevicollis* and early metazoans, we used three methods; Dollo parsimony, Roy-Gilbert maximum likelihood, and Csuros maximum likelihood. The results of the Csuros maximum likelihood analysis for the nine species set of introns is shown in Figure 2 in the main text and Supp. Table S5. The results of the other analyses for the nine species set are shown in Supp. Figure S3 and the results for the five species set of introns are shown in Supp. Figure S4. Though the different models infer varying amounts of intron loss and gain for various branches, all three models and both data sets indicate that the ancestor of choanoflagellates and metazoans was as or more intron rich than *M. brevicollis.* Additionally, all models infer a significant gain of introns between the ancestor of metazoans and choanoflagellates and the eumetazoan ancestor, followed by little if any net intron gain within metazoans.

Dollo parsimony assumes that introns appearing at the same positions in orthologous genes were gained only once and then subsequently lost in as many lineages necessary to fit the observed phylogenetic pattern[37]. The ancestral state in all cases is a gene without introns. Intron gain and loss events were mapped onto an established species tree using PAUP 4.0b10[32].

The Roy-Gilbert maximum likelihood method calculates intron loss rates and incorporates them into the estimation of ancestral intron contents[38]. This method was applied to the current data set using a PERL implementation written and made available by Jason Stajich and Scott Roy[39]. This method requires an outgroup to infer ancestral intron states, so no inference is made for the most basal node.

The Csuros maximum likelihood method is a probabilistic model that estimates ancestral intron states and intron gain and loss rates for each branch[40]. This method was applied to the current data set using the Java application intronRates.jar made publicly available by the author (http://www.iro.umontreal.ca/~csuros/introns/). This model can also infer a number of "all zero" columns, or introns that were present in an ancestral state but lost in all extant taxa. The results shown here assume that there were no such "all zero" columns, but including "all zero" columns in the model does not dramatically change the results for this data set. This method works best with an unrooted tree, as shown.

From an analysis of all predicted introns in the *M. brevicollis* genome, we observed that its introns are on average shorter than introns found in metazoans. The distribution of *M. brevicollis* intron lengths shows that there are few extremely long introns (Supp. Fig. S2A). To determine how this difference manifests itself in introns found in orthologous positions in *M. brevicollis* and metazoans, we examined 419 introns from the set of orthologous introns described above that are found in *M. brevicollis* and humans (Supp. Fig. S2B). The average length of these introns in *M. brevicollis* is 132 base pairs as compared to 3,438 base pairs in humans, and the length distributions are significantly different between the two species (Kolmogorov-Smirnov comparison test, D = 0.815, p < 0.01).

S3.4 Protein domain content of *M. brevicollis*. The protein domain content of the *M. brevicollis* genome was annotated using Pfam v20[41, 42] and SMART v5.1[43] with standard cutoff values. Two protein sets were annotated, the Monbr1_all_proteins.fasta (with completely identical proteins removed) and the Monbr1_best_proteins.fasta. All the analysis described in the text used the Monbr1_best_proteins.fasta set.

The initial analysis of the phylogenetic distribution of protein domains found in *M. brevicollis* included the species listed in Supp. Table S6. To identify domains found exclusively in choanoflagellates and other phylogenetic groups, lists were generated using the Pfam and SMART annotations of these genomes. The lists of Pfam and SMART domains were combined using Interpro ID numbers to eliminate overlap. The phylogenetic distribution of each domain thought to be unique to *M. brevicollis* and a given phylogentic group was then checked by hand using the SMART and Pfam databases online in order to include additional species distribution information. The functions of domains identified as unique to *M. brevicollis* and metazoans were hand-curated.

Many of the domains found exclusively in metazoans and *M. brevicollis* are involved in cell signaling and adhesions in metazoans (Supp. Table S4). For example, Bruton's tyrosine kinase motif [44], which is involved in the regulation of cell proliferation through tyrosine kinase signaling in metazoans is also found in *M. brevicollis*. The *M. brevicollis* genome contains additional domains involved in tyrosine kinase signaling in metazoans, including the phosphotyrosine binding domain (PTB/PID) and the SH3 domain binding protein 5 domain. The *M. brevicollis* genome also encodes metazoan specific domains associated with the extracellular matrix (ECM). These include the reeler domain (found in the neuronal ECM protein reelin[45]), the ependymin domain (an extracellular glycoprotein found in cerebrospinal fluid[46]), and the somatomedin B domain (found in the blood plasma ECM protein vitronectin[47]). Evidence for these protein domains in choanoflagellates, each of which were previously known only in metazoans, extends their evolutionary history to the last common holozoan ancestor, and raises questions about their ancestral functions.

Over and under-represented protein domains in *M. brevicollis* as compared to humans and *S. pombe* were also identified. This analysis was done using SMART's genomic mode, to avoid over-counting domains due to redundant protein sets. Domains predicted by both SMART and Pfam were included and combined using Interpro ID numbers. The number of times each domain occurred in *M. brevicollis* was compared to its occurrence in *S. pombe* and humans. Significantly different numbers of domains were identified by the Chi-square test and ranked by their p-value. The top 200 significantly over and under represented domains were identified. Two sets of comparisons were made, the first of which counted each domain only once per protein and the second of which counted all occurrences of each domain. The top ten over-represented domains, when eacg domain is counted once per protein, as compared to humans and *S. pombe* are shown in Supp. Fig. S5.

Domains that are over-represented in *M. brevicollis* compared to humans include the FG-GAP domain (Interpro ID IPR013517) and the hyaline repeat, or HYR, domain (Interpro ID IPR003410). The FG-GAP domain, a domain that is found in the extracellular portion of transmembrane proteins (e.g. α-integrins) and that mediates interactions with the ECM[48], occurs in 35 proteins in the *M. brevicollis* genome and only 24 proteins in the human genome. The hyaline repeat (HYR) occurs in 13 proteins in *M. brevicollis* as compared to only three proteins in humans. This predominantly extracellular domain is found in the human glycoprotein hyaline and the sea urchin protein hyalin, which forms an extracellular scaffold around the developing sea urchin embryo[49]. Notably, the five most significantly over-represented domains in M. brevicollis relative to *S. pombe* -- ankyrin (IPR002110), SH2 (IPR000980), tyrosine protein kinase (IPR001245), PDZ (IPR001478) and EGF-like (IPR006209) domains -- are important in numerous metazoan signaling pathways. EGF domains are particularly prominent in metazoan multidomain proteins involved in cell signaling[50].

The SMART and Pfam annotations of the *M. brevicollis* genome, as well as the complete results of the analysis of over and under represented domains, can be found online at http://smart.embl.de/Monosigia/index.html.

S3.5 Analysis of signaling, adhesion and transcription factor families. Text and Interpro domain ID searches using the Joint Genome Institute (JGI) *M. brevicollis v1.0* genome browser (http://shake.jgi-psf.org/Monbr1/Monbr1.home.html) were performed to examine the predicted protein models for annotations in categories related to adhesion, signaling, and transcriptional regulation. The online Pfam and SMART tools were used to confirm the presence of domains present in their respective databases. A model was said to contain the domain if both tools identified that domain, except in cases where the domain was not in either the SMART or Pfam database. In these cases, presence predicted by either SMART or Pfam was considered sufficient.

tBLASTn was used to search for members of the transcription factor families listed in Figure 3. All hits with an e-value less than 1 were examined by a reciprocal BLAST search against the NCBI nr (non-redundant) protein database. Those protein models that had reciprocal BLAST hits belonging to the specific transcription factor family were further examined by the Pfam and SMART queries described above if family specific DNA-binding domains were available. Some protein models were further examined if Pfam and SMART did not contain domains specific to the DNA binding domains of the families. The categorization of MbMyc was confirmed by a reciprocal BLAST search against the NCBI nr protein database in which the best defined hits (e.g. not “hypothetical protein”) were all to Myc transcription factors. The *M. brevicollis* Sox transcription factor, found in a tBLASTn search using animal Sox protein sequences, was confirmed by a reciprocal BLAST search against the NCBI nr protein database in which the best defined hits were all to Sox transcription factors.

The presence of specific proteins or domains in *H. sapiens* and *D. melanogaster* was determined by text search in Homologene and Entrez (NCBI). Domains were identified in *C. intestinalis* and *N. vectensis* using the JGI *Nematostella vectensis v1.0* and *Ciona intestinalis v2.0* genome browsers (*N. vectensis*: http://genome.jgi-psf.org/Nemve1/Nemve1.home.html; *C. intestinalis*: http://genome.jgi-psf.org/Cioin2/Cioin2.home.html). Specific proteins and domains in *S. cerevisiae and D. discoideum* were identified by text search and GO on their respective genome browsers (http://www.yeastgenome.org and http://dictybase.org). Specific proteins and domains in the *R. oryzae*, *N. crassa*, and *C. cinereus* genomes were identified by text and BLAST searches of the Broad Institute’s genome browsers (*R. oryzae*: http://www.broad.mit.edu/annotation/genome/rhizopus_oryzae/Home.html, *N. crassa*: http://www.broad.mit.edu/annotation/genome/neurospora/Home.html, *C. cinereus*: http://www.broad.mit.edu/annotation/genome/coprinus_cinereus/Home.html). Domains in the *A. Thaliana* genome were identified by BLASTp searches performed on the *Arabidopsis thaliana* Integrated Database (http://atidb.org/cgi-perl/gbrowse/atibrowse).

S3.6 Protein identification numbers for *M. brevicollis* and metazoan signalling homologs. The following *M brevicollis* protein models were identified as homologs of metazoan signaling proteins (JGI protein identification numbers): *Mbrev Tollip*: 38093; *Mbrev STAT*-like: 44371; *Mbrev Notch*-like: 26647; *Mbrev Presenilin*: 29512; *Mbrev Furin-like*: 14515; *Mbrev TACE*-like: 22277; *Mbrev Patched*: 38011, 36995, 36866; *Mbrev Hedgehog*-like: 33852, 36484, 28599; *Mbrev Fused*: 29411.

For the study of Notch and Hedgehog evolution, the following *M. brevicollis* protein models were used: (JGI protein identification numbers): *Mbrev* N1 29255; *Mbrev*

N2 26647, *Mbrev* N3 27644, *Mbrev* H1 28599, *Mbrev* H2 33852. The following metazoan protein sequence were used: (NCBI accession numbers): *Nvec* Notch 20239, *Nvec* Hh 241466, *Nvec Hedgling* 200640, *Hsap* Notch NP_060087.2, *Hsap* Hh NP_00184.1

S3.7 Phospho-tyrosine signaling machinery. We used the SMART domain prediction algorithm to assign domain architectures to the proteins in the *M. brevicollis* filtered gene set (filtered SMART set). Within this set we identified all pairwise domain combinations, i.e. the set of domains that appear in the same protein as a TyrKc domain, PTPc domain, or a SH2 domain (Fig. 5). We also performed the pairwise domain analysis for metazoans and non-metazoans (fungi, amoebae, etc.) using the SMART genomic database. Along with the pairwise domain analysis we sorted the filtered set, the metazoan set and the non-metazoan set based on domain architecture of complete proteins using the SMART domain architecture inquiry tool.

S3.8 TATA-binding proteins and transcription elongation factors. *M. brevicollis* possesses a second TATA-binding-protein (TBP) family member, suggesting a choanoflagellate-specific gene duplication that may be associated with gene regulatory diversity. In contrast to the initiation machinery, most of the known eukaryotic transcription elongation factors (TFIIS, NELF, PAF, DSIF, and P-TEFb, but not elongin) have clear homologs in the *M. brevicollis* genome.

S3.9 MAPK signaling. Eukaryotic cells contain multiple mitogen-activated protein kinase (MAPK) cascades that are activated by external stimuli and that produce distinct physiological responses. The core of MAPK signaling is a signature three-kinase module (MAPKKK→MAPKK→MAPK) that is conserved from yeast to human[51]. The simple fungal system contrasts with the multiple distinct MAPK pathways in metazoans used to control a larger array of cellular processes. By exploring the MAPK cascade kinases of *M. brevicollis*, we found an unexpectedly early emergence of one MAPK pathway, and potentially new or unstudied variations in the coupling of these pathways.

The canonical Erk MAPK pathway (Mkk1→Erk) is conserved throughout eukaryotes (Supp. Table S9). The functionally distinct Erk5 cascade, (Mekk2→ Mkk5→Erk5), was previously found only in deuterostomes[52], but is now seen in *M. brevicollis*, as well as the primitive metazoan *Nematostella vectensis*, strongly suggesting an ancient origin followed by loss in both insect and nematode lineages[53]. The evolution of this pathway is intriguing because the three-tiered cascade emerges intact in choanoflagellates with no clear kinase homologs or intermediates in fungi. We do not know the function of Erk5 signaling in choanoflagellates, but in mammals the Erk5 pathway is primarily activated by stress stimuli, and can also be activated by traditional Erk stimuli such as nerve growth factor (NGF)[54]. Erk5 can also be directly activated by PI3 Kinase downstream of the Insulin Receptor.

In contrast with the finding of an intact Erk5 pathway, partial pathway evolution is exemplified by stress-activated p38 MAPK signaling in *M. brevicollis*. A functionally p38-like MAPK is present in yeast (Hog1) and there are at least three clear p38 genes in *M. brevicollis*. These contain the conserved TxY activation phosphorylation site but *M. brevicollis* lacks their canonical activators, Mkk3/Mkk4. This suggests an alternative upstream kinase of which the best candidate is the dual-specificity kinase TOPK (PBK), which in humans is known to activate p38. This suggests that TOPK might be the original p38 activator and that the Mkk3/Mkk4 kinases evolved more recently within Metazoa. Further evidence for the partial evolution of p38 signaling in choanoflagellates can be found at the MAPKKK level: *M. brevicollis* contains genes not found in fungi encoding apoptosis specific kinase (Ask1), Tao2 and multiple members of the mixed-lineage kinase (MLK) family, kinases that are known to at least partially activate p38 signaling in mammals[55-57].

Finally, the choanoflagellate and *Nematostella* genome data reinforce the metazoan-specificity of Jnk signaling. No members of the Jnk MAPK family can be found in fungi or choanoflagellates, and the Jnk activators, Mkk4 and Mkk7, are also missing. Interestingly, many of the MAPKKKs that activate the p38 pathway and the Jnk pathway in mammals are present in *M. brevicollis*. Since Jnk MAPK is most closely related to p38, one hypothesis is that Jnk evolved from a duplication event of p38, and co-opted the upstream components already in place for p38 signaling. Outside of the Jnk pathway, the MAPKs Erk3 and NMO, and the Erk activators Raf and Mos also appear to be exclusive to metazoans.

In summary, MAPK signaling in choanoflagellates is intermediate in complexity between fungi and animals. While *M. brevicollis* lacks some of the hallmarks of metazoan signaling, including p38 activators and the Jnk MAPKs, it has more versatility compared to the fungal MAPK networks, including a full Erk5 cascade and a doubling of the number of MAPKKKs, suggesting a greater diversity of upstream signals and environmental inputs. Future study of the functions of *M. brevicollis* MAPK components will provide an important bridge between the findings from MAPK studies in yeasts and metazoans, and will provide insights into the ancestry and elaboration of the MAPK pathway in animal evolution.

S4. Immunofluorescence Staining of *M. brevicollis.* We fixed *M. brevicollis* cells that were grown shaking at 120 rpm to a density between 10^6 and 10^7 cells/ mL by adding formaldehyde to a final concentration of 4%. We then applied approximately 0.5 mL of the fixed culture to poly-L-lysine coated coverslips and incubated for 30 minutes. After gently washing the coverslips 4 times with PEM (100 mM PIPES, pH 6.9, 1 mM EGTA, 0.1 mL $MgSO_4$) we blocked and permeabilized the cells for 30 minutes with blocker (PEM/1% BSA/0.3% TritonX-100) and subsequently replaced the blocker with E7 β-tubulin primary antibodies diluted in blocker (Developmental Studies Hybridoma Bank). After incubating the cells with the antibodies for 16 hours at 4^o C, we washed the coverslips 4 times with blocker, applied fluorescein conjugated donkey α-mouse IgG (H+L) (Jackson Laboratories) secondary antibodies and incubated for 1 hr in the dark, subsequently washing 4 times with PEM. To visualize F-actin, we incubated the cells with 6 U/ mL rhodamine phalloidin (Molecular Probes) diluted in PEM. To the rhodamine phalloidin-PEM, we added DAPI at a concentration of 10 ng/ mL to visualize the DNA. We applied this mixture to the slides and incubated for 25 minutes in the dark. We then washed the coverslips 3 times with PEM and mounted them onto slides using 10 μl ProLong Gold antifade reagent (Molecular Probes). All steps were performed at room temperature unless specified otherwise. We took all images using a Leica DMI6000

B inverted compound microscope and Leica DFC350 FX camera at 100X magnification using oil immersion.

S5. Tools for choanoflagellate genomics.
M. brevicollis JGI genome portal:
http://genome.jgi-psf.org/Monbr1/Monbr1.home.html
A browser that contains automated and manual gene models and annotations for M. brevicollis. Gene sets and scaffolds can be downloaded.
SMART annotation of *M. brevicollis*:
http://smart.embl.de/Monosigia/
SMART protein domain predictions and protein domain architectures for M. brevicollis.
Metazome:
http://www.metazome.net/
A multi-taxon tool for comparative genomics.
Choanobase:
http://mcb.berkeley.edu/labs/king/blast/
ESTs from the choanoflagellate M. brevicollis and Proterospongia sp.
Taxonomically Broad EST Database:
http://amoebidia.bcm.umontreal.ca/pepdb/searches/organism.php?orgID=MN
ESTs from the choanoflagellates Monosiga ovata and M. brevicollis.

References

1. Guillebault, D. et al. A new class of transcription initiation factors, intermediate between TATA box-binding proteins (TBPs) and TBP-like factors (TLFs), is present in the marine unicellular organism, the dinoflagellate Crypthecodinium cohnii. J Biol Chem 277, 40881-6 (2002).
2. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574 (2003).
3. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-5 (2001).
4. Csuros, M. in Proceedings of the Comparative Genomics: RECOMB 2005 International Workshop; Dublin, Ireland. (ed. McLysaght A, H. D.) 47-60 (Springer-Verlag, Berlin, 2005).
5. Claff, C. L. A migration-dilution apparatus for the sterilization of protozoa. . Physiol. Zool. 13, 334-341 (1940).
6. Weisburg, W. G., Barnes S.M., Pelletier D.A., and Lane D.J. 16S rDNA amplification for phylogenetic study. J Bacteriol. 173, 697-703 (1991).
7. Sottile, M. I., Baldwin, J. N. & Weaver, R. E. Deoxyribonucleic acid hybridization studies on Flavobacterium meningosepticum. Appl Microbiol 26, 535-9 (1973).
8. Chapman, J. A. in Physics (University of California, Berkeley, Berkeley, 2004).
9. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297, 1301-10 (2002).

10. Putnam, N. H. in Physics (University of California, Berkeley, Berkeley, 2004).

11. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).

12. Bullerwell, C. E., Gray, M.W. Evolution of the mitochondrial genome: protist connections to animals, fungi and plants. Current Opinion in Microbiology 7, 528--534 (2004).

13. Huang, X. On global sequence alignment. Comput Appl Biosci 10, 227-35 (1994).

14. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573-80 (1999).

15. Altshuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407, 513-6 (2000).

16. Mullikin, J. C. et al. An SNP map of human chromosome 22. Nature 407, 516-20 (2000).

17. Hartl, D. L. & Clark, A. G. Principle of Population Genetics (Sinauer Associates, Inc., Sunderland, MA, 1997).

18. Fisher, R. A. The Genetical Theory of Natural Selection (Clarendon, Oxford, 1930).

19. Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in Drosophila genomic DNA. Genome Res. 10, 516-22 (2000).

20. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. Genome Research 14, 988-995 (2004).

21. Quevillon, E. et al. InterProScan: protein domains identifier. Nucleic Acids Res 33, W116-20 (2005).

22. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31, 365-70 (2003).

23. Kanehisa M, G. S., Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Genome Biology 5, R7 (2006).

24. Koonin, E. V. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biology 5, R7 (2004).

25. Ashburner, M. et al. Gene ontology: tool for the unification of biology. . Nature Genetics 25, 25-9 (2000).

26. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology 340, 783-795 (2004).

27. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of Molecular Biology 305, 567-580 (2001).

28. Rokas, A., Kruger, D. & Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. Science 310, 1933-8 (2005).

29. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39, 309-38 (2005).

30. Thompson, J. D., Higgins, D. G. & Gibson, T. J. Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673-4680 (1994).

31. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696-704 (2003).
32. Swofford, D. L. (Sinauer, Sunderland, MA, 2002).
33. Abascal, F., Zardoya, R. & Posada, D. Prottest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104-2105 (2005).
34. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18, 691-9 (2001).
35. Town, C. et al. Comparative genomics of Brassica oleracea and Arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell 18, 1348-59 (2006).
36. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317, 86-94 (2007).
37. Kondrashov, F. & Koonin, E. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. Trends in Genetics 19, 115-119 (2003).
38. Roy, S. W., Gilbert, W. Complex early genes. Proceedings of the National Academy of Sciences 102, 1986-1991 (2005).
39. Stajich, J. E., Dietrich, F. S. & Roy, S. W. Comparative genomic analysis of fungal genomes reveals intron rich ancestor. (2007).
40. Csuros, M. in Proceedings of the Comparative Genomics: RECOMB 2005 International Workshop (ed. McLysaght, A., Huson, D.) 47-60 (Berlin: Springer-Verlag, Dublin, Ireland, 2005).
41. Finn, R. D. et al. Pfam: clans, web tools and services. Nucleic Acids Res 34, D247-51 (2006).
42. Bateman, A. et al. The Pfam protein families database. Nucleic Acids Res 32, D138-41 (2004).
43. Letunic, I. et al. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34, D257-60 (2006).
44. Lindvall, J. M. et al. Bruton's tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling. Immunol Rev 203, 200-15 (2005).
45. Tissir, F. & Goffinet, A. M. Reelin and brain development. Nat Rev Neurosci 4, 496-505 (2003).
46. Suarez-Castillo, E. C. & Garcia-Arraras, J. E. Molecular evolution of the ependymin protein family: a necessary update. BMC Evol Biol 7, 23 (2007).
47. Schvartz, I., Seger, D. & Shaltiel, S. Vitronectin. Int J Biochem Cell Biol 31, 539-44 (1999).
48. Baneres, J. L., Roquet, F., Martin, A. & Parello, J. A minimized human integrin alpha(5)beta(1) that retains ligand recognition. J Biol Chem 275, 5888-903 (2000).
49. Wessel, G. M., Berg, L., Adelson, D. L., Cannon, G. & McClay, D. R. A molecular analysis of hyalin--a substrate for cell adhesion in the hyaline layer of the sea urchin embryo. Dev Biol 193, 115-26 (1998).
50. Tordai, H., Nagy, A., Farkas, K., Banyai, L. & Patthy, L. Modules, multidomain proteins and organismic complexity. Febs J 272, 5064-78 (2005).

51. Widmann, C., Gibson, S., Jarpe, M. B. & Johnson, G. L. Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human. Physiol Rev 79, 143-80 (1999).
52. Bradham, C. A. et al. The sea urchin kinome: a first look. Dev Biol 300, 180-93 (2006).
53. Manning, G., Plowman, G. D., Hunter, T. & Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci 27, 514-20 (2002).
54. Nishimoto, S. & Nishida, E. MAPK signalling: ERK5 versus ERK1/2. EMBO Rep 7, 782-6 (2006).
55. Chen, Z. & Cobb, M. H. Regulation of stress-responsive mitogen-activated protein (MAP) kinase pathways by TAO2. J Biol Chem 276, 16070-5 (2001).
56. Gallo, K. A. & Johnson, G. L. Mixed-lineage kinase control of JNK and p38 MAPK pathways. Nat Rev Mol Cell Biol 3, 663-72 (2002).
57. Matsukawa, J., Matsuzawa, A., Takeda, K. & Ichijo, H. The ASK1-MAP kinase cascades in mammalian stress response. J Biochem (Tokyo) 136, 261-5 (2004).