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.

Choanoflagellates are the closest known relatives of metazoans. To illuminate potential molecular mechanisms underlying the evolution of metazoan multicellularity, we sequenced and analyzed 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 signaling protein domains that are otherwise restricted to metazoans. 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.

Introduction

Choanoflagellates have long fascinated evolutionary biologists for their striking 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 homologs of metazoan signaling 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, Supp. Fig. S1 and Supp. Notes S3.1)[8, 13]. By virtue of their position on the tree of life, studies of choanoflagellates provide an unparalleled 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 play critical roles in oceanic carbon cycling and the microbial food web[17, 18].

Over 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 reconstruct the genomic changes attendant with the early evolution of metazoans, we sequenced the genome of the choanoflagellate *Monosiga brevicollis* and compared it with genomes from metazoans and other eukaryotes.

Gene structure and intron evolution

The ~ 41.6 million base pair (Mb) *M. brevicollis* genome contains approximately 9200 genes (Supp. Notes S1 and S2) and is comparable in size to genomes of filamentous fungi (~30-40 Mb) and other free-living unicellular eukaryotes (e.g., small diatoms at ~20-35 Mb[20] and Icthyosporeans at ~20-25 Mb[21]). Metazoan genomes are typically significantly larger, with few exceptions[22].

M. brevicollis genes have several distinguishing structural features (Table 1). While the *M. brevicollis* genome is compact, its genes are almost as intron rich as human genes (6.6 introns per *M. brevicollis* gene vs. 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 (Supp. Fig. S2 and Supp. Notes S3.3).

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 choanoflagellates and metazoans had genes at least as intron rich as those of modern choanoflagellates (Fig. 2, Supp. Figs. S3 and S4, Supp. Notes S3.3). 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 Metazoa[23].

Premetazoan history of protein domains and genes associated with metazoan multicellularity and development

The *M. brevicollis* genome provides unprecedented insight into the early evolution of metazoan genes. Pfam and SMART annotations of the *M. brevicollis* genome identify 78 protein domains that are exclusive to choanoflagellates and metazoans, only two of which have been previously reported in choanoflagellates (Supp. Table S4)[10]. Because genomic features shared by *M. brevicollis* and metazoans were likely 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 signaling and adhesion processes in metazoans, suggestive of 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 (albeit likely an incomplete one) at the complement of genes and protein domains that predate metazoan origins.

To further investigate the extent to which molecular components required for metazoan multicellularity evolved before the origin of metazoans, we performed targeted searches in the *M. brevicollis* genome and representative metazoan, fungal, and plant genomes for homologs of critical metazoan cell adhesion, cell signaling, 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, homologs 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. While 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 via endocytosis[25-27].

The genome of *M. brevicollis* also contains integrin-α and immunoglobulin (Ig) domains, cell adhesion domains formerly thought to be restricted to Metazoa. 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 Ig domain-containing proteins have both adhesive and immune functions. The *M. brevicollis* genome encodes a total of 5 Ig domains that show affinity for either 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 Ig domains (Supp. Table S7), suggesting that the radiation of the Ig 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 homologs of metazoan cell adhesion domains may act to mediate interactions between *M. brevicollis* and its extracellular environment.

ECM 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 (in which the first position is glycine and the second and third positions are frequently proline or hydroxyproline) in an arrangement similar to metazoan collagens[30]. Other ECM-associated domains previously known only from metazoans, and now *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 Supp. Table S4).

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 homologs 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 take note of the differences. Individual cell-adhesion and ECM-associated domains in the *M. brevicollis* genome often occur in unique arrangements and clear orthologs of specific metazoan adhesion proteins are rarely found. While 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] likely evolved after the divergence of the two lineages.

Domain shuffling in the evolution of metazoan intercellular signaling networks

Our analysis of the *M. brevicollis* genome reveals little evidence that metazoan-specific signaling pathways were present in the last common ancestor of choanoflagellates and metazoans. Many pathways are missing entirely and *M. brevicollis* genes with similarity to metazoan signaling machinery are largely found to share conserved domains without aligning across the full span of what are often complex multidomain proteins (e.g., EGF repeats are common to Notch but also to many other proteins; Supp. Table S8). Specifically, no receptors or ligands were identified from the NHR, WNT, and TGF-β signaling pathways. The only evidence of the Jak/STAT 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 signaling pathway, a signaling system important both for development and innate immunity in metazoans.

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

Homologs of *hedgehog*, *dispatched*, and *patched* genes are also present; however, there is no evidence for *smoothened* nor its defining frizzled domain. In metazoans, Hedgehog consists of an amino-terminal signaling domain and carboxy terminal Hedgehog/Intein (HINT) domain responsible for autocatalytically cleaving the protein. In one *M. brevicollis* Hedgehog-like protein, a Hedgehog amino-terminal signaling domain is found at the amino terminus of a large transmembrane protein that, instead of a HINT domain, includes von Willebrand A, cadherin, TNFR, Furin, and EGF domains. Similar proteins are found in the sponge *Amphimedon queenslandica* and 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 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, *Hoglet*, was identified in

the distantly related *Monosiga ovata*[12], supporting the idea that isolated signaling components were present in the last common ancestor of choanoflagellates and metazoans.

Phospho-tyrosine signaling machinery: Divergent use of a common toolkit

Phospho-tyrosine (pTyr) based signaling was considered unique to metazoans until its recent observation in choanoflagellates[9, 11]. The key domains involved in pTyr signaling are found in abundance in the *M. brevicollis* genome: tyrosine kinase (TK) 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) (Supp. Fig. S7). In contrast, these domains are rare in non-metazoans; for example *S. cerevisiae* has no TKs, only three PTP domains, and a single SH2 domain. These findings support a model in which the full set of pTyr signaling machinery evolved prior to the separation of the choanoflagellate and metazoan lineages.

Although pTyr signaling machinery is present in metazoans and choanoflagellates, the mode of usage in *M. brevicollis* may be distinct from 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 TK, PTP, and SH2 domains are distinct from those seen in any metazoan genome (Fig. 5 and Supp. Notes S3.7). In contrast, for other sets of common signaling domains (those involved in phospho-Ser/Thr, Ras GTP, and Rho GTP

signaling) the majority of observed combinations are shared between *M. brevicollis* and metazoans. These observations are consistent with a simple model in which phospho-Ser/Thr, Ras GTP, and Rho GTP signaling were more fully elaborated prior to 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 signaling 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 (Supp. Table S10) 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, homologs of the largest subunit of TFIIF and several subunits of TFIIH are apparently lacking from the genome and the EST collection (Supp. Fig. S8), reminiscent of the absence of several basal factors from the *Giardia lamblia* genome, and 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 remodeling complexes (Supp. Fig S9 and Supp. Notes S3.8).

Perhaps not surprisingly, *M. brevicollis* possesses most of the ubiquitous families of eukaryotic transcription factors (Supp. Fig. S10). The majority of the predicted transcription factors are zinc-coordinating; approximately 44% are

C2H2-type zinc fingers. Eight proteins (5% of a total 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 prior to the origin of metazoans, the "typical", or non-TALE, homeodomains containing ~60 amino acids and the TALE-class homeodomains containing 63+ amino acid homeodomains[39]. The *M. brevicollis* genome encodes only two homeodomain proteins, both of which group with the MEIS sub-class of TALE homeodomains (Supp. Fig. S12). 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 *A. queenslandica* (demosponge) genome sequence reads, indicating that this characteristic metazoan gene family likely 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. Homologs of the metazoan p53, Myc and Sox/TCF families were identified, while many transcription factor families associated with metazoan patterning and development (ETS, HOX, NHR, POU and T-box) appear to be absent (Fig. 3).

Discussion

Choanoflagellates, sponges, and other metazoans last shared a unicellular common ancestor in the late Precambrian, 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 signaling targets. In contrast, components of most other intercellular signaling 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 survive. There are, for example, no other known microbial eukaryotes that possess any of the eight developmental signaling 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 signaling 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 signaling domains in *M. brevicollis* that appear in unique combinations relative to animals. For phosphotyrosine based signaling in particular, the striking 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 upon in each lineage.

Given the limited transcription factor diversity in *M. brevicollis*, it is striking that the genome encodes representatives of the otherwise metazoan specific p53, Myc, and Sox/TCF transcription factor families. These transcription factors may have played 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. While *M. brevicollis* is strictly unicellular, other choanoflagellates facultatively form colonies, and the modulation of these associations by cell signaling,

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 intrinsic and extrinsic factors underlying metazoan origins.

Methods summary

All analyses described here were performed on Version 1.0 of the genome sequence. Details of analyses not described below 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* (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.

Genome sequencing, assembly, and validation

The ~ 41.6 Mb draft sequence of the *M. brevicollis* genome was generated from ~8.5-fold redundant paired-end whole genome shotgun sequence coverage (Supp. Notes S1.4 and Supp. Table S1). Sequence data derived from six whole-genome shotgun (WGS) libraries was assembled using release 2.9.2 of the WGS assembler Jazz. Completeness of the draft genome was assessed by capturing ~98.5% of sequenced ESTs.

Gene prediction and annotation

9,196 gene models were predicted and annotated using the JGI Annotation Pipeline (Supp. Notes S2). The assembly and annotation data are available from JGI Genome Portal at http://www.jgi.doe.gov/Mbrevicollis and from DBJ/EMBL/GenBank under the project accession ABFJ00000000. Additional choanoflagellate genome resources are listed in Supp. Notes S5.

Intron analysis

Homologs of 473 highly conserved genes from *M. brevicollis* and eight representative eukaryotes were aligned to reveal the position and phylogenetic distribution of 1989 highly reliable intron splice sites at 1054 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 signaling, adhesion and transcription factor protein domains

Gene models containing metazoan signaling, adhesion, and transcription factor domains were identified using text and Interpro domain ID searches of the Joint Genome Institute (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 (Supp. Notes S3.5).

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 four-letter abbreviation were used for protein domain comparisons in this study: Human (Homo sapiens; Hsap), Ascidian (Ciona intestinalis; Cint), *Drosophila* (Drosophila melanogaster; Dmel), Cnidarian (Nematostella vectensis; Nvec), *M. brevicollis* (Mbrev), Zygomycete (Rhizopus oryzae; Rory), Basidiomycete (Coprinus cinereus; Ccin), Ascomycete (Neurospora crassa; Ncra), Hemiascomycete (Saccharomyces cerevisiae; Scer), Slime mold (Dictyostelium discoidum; Ddis), and *Arabidopsis* (Arabidopsis thaliana; Atha). (**B, C, D**) Choanoflagellate cells bear a single apical flagellum (arrow, **B**) and an apical collar of actin-filled microvilli (bracket, **C**). (**D**) An overlay of α-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, *Drosophila*

melanogaster, and *Nematostella vectensis*), *Monosiga brevicollis*, intron-rich fungi (*Cryptococcus neoformans A* and *Phanerochaete chrysosporium*), plants and green algae (*Arabadopsis thaliana* and *Chlamydomnas 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 are black. The inferred or observed number of introns present in ancestral and extant species are 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.

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. In contrast, many metazoan sequence specific transcription factors are absent from the *M. brevicollis* gene catalog. 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. Presence (filled box) or absence (empty box) of transcription factor families was determined by reciprocal BLAST and SMART/Pfam domain annotations (Supp. Notes S3.5). Species names follow the convention from Fig. 1.

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 (A) and Hedgehog (Hh) proteins (B), some of which 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. See Supp. Notes S3.6 for protein accession numbers and Supp. Fig. S6 for identification of all displayed protein domains.

Figure 5. Divergent usage of protein domains involved in phospho-tyrosine based signaling 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 signaling domains involved in Ras, Rho, phospho-Ser/Thr, and pTyr signaling. In cases where 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 those that are observed both in *M. brevicollis* and metazoan genomes (grey). P-Tyr signaling domains in *M. brevicollis* are unique in that the majority of their observed pairwise domain combinations are distinct from those observed in metazoans.

References

1. **James-Clark, H. On the spongiae ciliatae as infusoria flagellata; or observations on the structure, animality, and relationship of *Leucosolenia botryoides*. Annals and Magazine of Natural History 1, 133-142; 188-215; 250-264 (1868).**
2. **Saville Kent, W. A Manual of the Infusoria (David 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. International Journal of Astrobiology 2, 203-211 (2003).**
5. **Philippe, H. et al. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol Biol Evol 21, 1740-52 (2004).**
6. **Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W. & Burger, G. The closest unicellular relatives of animals. Curr Biol 12, 1773-8. (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 U S A 100, 892-7 (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, 1231-9 (2005).**
9. **King, N. & Carroll, S. B. A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. Proc Natl Acad Sci U S A 98, 15032-7. (2001).**
10. **King, N., Hittinger, C. T. & Carroll, S. B. Evolution of key cell signaling and adhesion protein families predates animal origins. Science 301, 361-3. (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 U S A 103, 12021-6 (2006).**
12. **Snell, E. A. et al. An unusual choanoflagellate protein released by Hedgehog autocatalytic processing. Proc Biol Sci 273, 401-7 (2006).**
13. **Rokas, A., Kruger, D. & Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. Science 310, 1933-8 (2005).**
14. **King, N. The unicellular ancestry of animal development. Dev Cell 7, 313-25 (2004).**
15. **Buck, K. R. & Garrison, D. L. Distribution and Abundance of Choanoflagellates (Acanthoecidae) across the Ice-Edge Zone in the Weddell Sea, Antarctica. Marine Biology 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 Biology 12, 53-63 (1992).**

17. **Hartmut Arndt, D. D., Brigitte Auer, Ernst-Josef Cleven, T. G., Markus Weitere and & Mylnikov, A. P. in The Flagellates (ed. Leadbeater, B. s. C. G., 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, 465-80 (2002).**
19. **Leadbeater, B. S. C. Life-history and ultrastructure of a new marine species of *Proterospongia* (Choanoflagellida). Journal of the Marine Biological Association U.K. 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, 379-84 (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, 219-29 (2006).**
24. **Halbleib, J. M. & Nelson, W. J. Cadherins in development: cell adhesion, sorting, and tissue morphogenesis. Genes Dev 20, 3199-214 (2006).**
25. **Gupta, G. & Surolia, A. Collectins: sentinels of innate immunity. Bioessays 29, 452-64 (2007).**
26. **Yamaguchi, Y. Lecticans: organizers of the brain extracellular matrix. Cell Mol Life Sci 57, 276-89 (2000).**
27. **Zelensky, A. N. & Gready, J. E. The C-type lectin-like domain superfamily. Febs J 272, 6179-217 (2005).**
28. **Akiyama, S. K. Integrins in cell adhesion and signaling. Hum Cell 9, 181-6 (1996).**
29. **Erwin, D. H. The Origin of Metazoan Development - a Paleobiological Perspective. Biological Journal of the Linnean Society 50, 255-274 (1993).**
30. **van der Rest, M. & Garrone, R. Collagen family of proteins. Faseb J 5, 2814-23 (1991).**
31. **Tissir, F. & Goffinet, A. M. Reelin and brain development. Nat 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 Biol Sci 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. European Journal of Protistology 30, 111-124 (1994).**
35. **Hutter, H. et al. Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. Science 287, 989-94 (2000).**

36. **Adamska, M. et al. The evolutionary origin of hedgehog proteins. Curr Biol 17, R836-7 (2007).**
37. **Letunic, I. et al. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34, D257-60 (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, 1537-47 (2004).**
39. **Derelle, R., Lopez, P., Le Guyader, H. & Manuel, M. Homeodomain proteins belong to the ancestral molecular toolkit of eukaryotes. Evol Dev 9, 212-9 (2007).**
40. **Larroux, C. et al. The NK homeobox gene cluster predates the origin of Hox genes. Curr Biol 17, 706-10 (2007).**
41. **Knoll, A. H. Life on a Young Planet (Princeton University Press, Princeton, NJ, 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 U S A 102, 9547-52 (2005).**
43. **Ekman, D., Bjorklund, A. K. & Elofsson, A. Quantification of the Elevated Rate of Domain Rearrangements in Metazoa. J Mol Biol (2007).**
44. **Tordai, H., Nagy, A., Farkas, K., Banyai, L. & Patthy, L. Modules, multidomain proteins and organismic complexity. Febs J 272, 5064-78 (2005).**
45. **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).**
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, 1512-7 (2003).**
47. **Roy, S. W. & Gilbert, W. Complex early genes. Proc Natl Acad Sci U S A 102, 1986-91 (2005).**
48. **Bateman, A. et al. The Pfam protein families database. Nucleic Acids Res 32, D138-41 (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 Richard Melmon. The Rokhsar group is supported by the Gordon and Betty Moore Foundation and Richard Melmon. We thank Jason Stajich, Philip Johnson, and Richard Lusk for helpful discussions, Emily Hare, Eric Meltzer and Kazutoyo Osoegawa for technical advice, Emina Begovic for assistance with Figure 1, Mark Dayel and Nipam Patel for critical reading of the manuscript and Sean 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. King and D. Rokhsar are co-senior authors.

Author Information The genome assembly and annotation data are deposited at DBJ/EMBL/GenBank under the project accession ABFJ00000000. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. Correspondence and requests for materials should be addressed to N.K. (nking@berkeley.edu) or D.R. (dsrokhsar@lbl.gov)