Title
The Cellular and Molecular Basis of Multicellular Development in the Choanoflagellate Salpingoeca rosetta

Author
Fairclough, Stephen Robert

Publication Date
2011

Peer reviewed|Thesis/dissertation

The Cellular and Molecular Basis of Multicellular Development in the Choanoflagellate *Salpingoeca rosetta*

By
Stephen Robert Fairclough

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Molecular and Cellular Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Associate Professor Nicole King, Chair
Associate Professor Michael Eisen
Professor Nipam Patel
Professor N. Louise Glass

Fall 2011

The Cellular and Molecular Basis of Multicellular Development in the Choanoflagellate *Salpingoeca rosetta*

By Stephen Robert Fairclough

Abstract

The Cellular and Molecular Basis of Multicellular Development in the Choanoflagellate *Salpingoeca rosetta*

by

Stephen Robert Fairclough

Doctor of Philosophy in Molecular and Cellular Biology

University of California, Berkeley

Professor Nicole King, Chair

For more than a century and half choanoflagellates, the closest living relatives of animals, have fascinated evolutionary biologists. By characterizing the similarities and differences between choanoflagellates and animals, biologists have gained insight into their last common ancestor, the Urchoanimal. Understanding how choanoflagellates form multicelled colonies has important implications for understanding the cellular and evolutionary foundations of animal multicellularity. I determined that the colony forming choanoflagellate *Salpingoeca rosetta* forms colonies by cell division and not by cell aggregation. The observation that cytoplasmic bridges connect cells in *S. rosetta* colonies and other colonial choanoflagellates, as well as cells in sponges, suggests that this mechanism of colony formation may be ancestral within the choanoflagellate lineage and may have been present in the Urchoanimal as well. To understand the molecular mechanisms underlying *S. rosetta* colony formation and cellular differentiation, I analyzed the genome and transcriptome of *S. rosetta*. The comparison of choanoflagellate genomes with animal genomes has revealed that many of the genes involved in the basic mechanisms of cell adhesion, signaling, and differentiation that were previously thought to be unique to animals are present in choanoflagellates, indicating that these genes were present in the Urchoanimal. These insights also refine our understanding of genes that emerged on the stem lineage leading to the last common ancestor of all animals, the Urmetazoan. By characterizing patterns of gene expression in different cell types, in particular colonial cells, I determined that different cell types have unique patterns of gene expression suggesting there is transcriptional basis for the observed differentiation. By characterizing the set of genes that are expressed at higher levels in particular cell types, I discovered that septins, known regulators of cytokinesis, are over-expressed in colonial cells. I also found that subsets of the cadherin gene family, known for its role in intercellular adhesion and signaling, are over-expressed in colonial cells and attached cells. Taken together these data suggest that *S. rosetta* cells have a number of development trajectories, including multicellular colonies, whose biology involves homologs of animal genes operating in conserved and novel functions.

Table of Contents

Tables and Figures

Chapter 1 **The genetic basis of animal origins.**

Chapter 2 **Multicellular development in the choanoflagellate *S. rosetta.***

Chapter 3 **The genetic basis of multicellular development in the choanoflagellate *Salpingoeca rosetta.***

Chapter 1: The genetic basis of animal origins

Reconstructing the origin of animals

Reconstructing the evolutionary events that led to the origin of animals requires the understanding of the biology of ancient animals and their ancestors (1, 2). However, the fossil record, which often provides insight into the biology of more recent ancestors, has remained largely silent with regard to the biology of these organisms, presumably because of their antiquity, small size and composition. A complementary approach to understand ancient biology is to compare the biology of organisms alive today and infer what shared characteristics were present in their last common ancestor. For more than a century and a half choanoflagellates, the closest living relatives of animals (Fig. 1.1A), have fascinated evolutionary biologists because of their morphological and behavioral similarity to sponge feeding cells, choanocytes (3). By identifying their similarities and differences, biologists have gained understanding about the biology of their last common ancestor, the Urchoanimal (2, 3). The application of molecular biology and genomics to the study of choanoflagellates and early branching animal phyla is providing insights into the gene content of these ancient animal ancestors as well (4-6). The comparison of choanoflagellate genomes with animal genomes has revealed that many of the genes involved in the basic mechanisms of cell adhesion, signaling, and differentiation that were previously thought to be unique to animals are also present in choanoflagellates, indicating that these genes were present in the Urchoanimal (4, 7). These observations also refine our understanding of genes that emerged on the stem lineage leading to the last common ancestor of all animals, the Urmetazoan. By continuing to compile information about the biology of these two ancestors and the ancestors precending and following them we can reconstruct the foundational events that preceded the origin of animals and their diversification into the forms we see around us today.

Choanoflagellates are the closest living relatives of animals

Choanoflagellates are single-celled and colony-forming flagellates that have an ovoid or spherical cell body that is 3-10 μm in diameter and an apical flagellum surrounded by a collar of 30-40 microvilli (Fig. 1.1B) (8-10). Movement of the flagellum creates water currents that can propel free-swimming choanoflagellates through the water column or trap bacteria and detritus against the collar of microvilli where they are engulfed (Fig. 1.1B) (10, 11). This distinctive cell morphology and method of feeding are shared with the sponges and have historically been used as evidence for the close relationship between choanoflagellates and animals (3, 11). More recently, molecular phylogenetics and comparative genomics have provided robust support for this relationship. In addition, numerous independent analyses demonstrate that animals are a monophyletic group containing sponges, with choanoflagellates as their closest known living relatives (Fig. 1.1A) (5, 6, 12-15).

Competing hypotheses regarding the branch order of early animals

Despite the consistent signal placing of choanoflagellates as a sister group of animals, the phylogenetic relationships among early branching animal phyla are more contentious (16, 17). Data from the sequenced genomes of several early branching animal

lineages, such as the sponge *Amphimedon queenslandica,* the placozoan *Trichoplax adhaerens,* and the cnidarian *Nematostella vectensis*, as well as ESTs from *Oscarella carmela* and ctenophores, have yielded multiple conflicting animal phylogenies (Fig. 1.2) (5, 6, 15, 17-19). The fossil record and most molecular phylogenies, including one based on the recently sequenced *A. queenslandica* genome (6), suggest that sponges represent the earliest branching animals (Fig. 1.2A). Nonetheless, two recent phylogenetic analyses challenge this evolutionary scenario (Fig. 1.2B, C) (18, 19). In an analysis based on nuclear and mitochondrial genes from diverse animals, Schierwater *et al.* proposed that two lineages, one containing the Bilateria and another containing the Placozoa, Cnidaria, Ctenophora and Porifera (with Placozoa being the earliest branching lineage in this group) diverged soon after the origin of animals (Fig. 1.2B)(19). This topology, which has thus far not been independently supported, implies one of two scenarios (19). Either the two sister groups represent an extreme example of convergent evolution, with Cnidaria, Ctenophora, and Bilateria independently evolving body plans that include body axes, nervous systems, sensory organs, and other characteristics, or the ancestral animal was surprisingly complex and the Placozoa and Porifera have undergone extensive character loss. A similarly counterintuitive evolutionary scenario emerges from analyses performed by Dunn and colleagues (18), whose analyses based on ESTs from ctenophores and diverse other animals suggest that ctenophores are the earliest branching animals. This scenario, if correct, would imply that the last common ancestor of animals had morphological features such as nerves, muscles and a gut, and that these features were lost wholesale from Porifera and Placozoa, a hypothesis not supported by the fossil record. Alternatively, as with the tree in Fig. 1.2B, the ctenophores-first scenario for animal diversification requires that Ctenophora evolved nerves, muscles, and a gut independently of the other Eumetazoans. The notions that either Placozoa or Ctenophora are the earliest branching animal phyla likely result from two phenomena that particularly confound phylogenetic tree reconstruction for animals: rapid speciation and ancient branching (16, 17). Rapid speciation events spaced closely in time lead to short internal branches that are difficult to resolve because there is little phylogenetic signal. Ancient terminal branches tend to be long and replete with multiple substitutions occurring at the same position that mask phylogenetic signal.

In contrast to the ctenophores-early and placozoans-early hypotheses presented by Dunn *et al*. and Schierwater *et al*., Srivastava *et al.* and Philippe *et al*. 2009 independently reconstructed phylogenetic trees with similar branching patterns that indicate Porifera are the earliest branching animals (5, 6, 15, 17-19). The analysis by Srivastava *et al.* indicates that Porifera branch followed by Trichoplax and Cnidaria and Bilateria; however, because it does not include ctenophores, it does not address the issue of the Ctenophora branching. The analysis conducted by Philippe, *et al*. 2009 contains representatives of all the major relevant taxa, including ctenophores, and finds that Ctenophora branches sister to Cnidaria, which diverged from Bilateria after Placozoa and Porifera (15). In fact, in a follow-up analysis Philippe *et al.* 2011 found that just resolving conflicts among single-gene trees and concatenated trees in Schierwater *et al*. and Dunn *et al.* while retaining the original taxon sampling and inference methods is sufficient to eliminate all significant incongruence among the three recent studies (17).

In addition to the branching order of the Porifera, Placozoa, Cnidaria and Ctenophora, there is disagreement as to whether Porifera are monophyletic or paraphyletic. Determining whether Porifera are monophyletic or paraphyletic has important implications for understanding the biology of the first animals. If Porifera are paraphyletic it would imply that characters shared by all sponge lineages are ancestral for the Metazoa and that Eumetazoans are derived from animals with a sponge-like body plan. Alternatively, if Porifera are monophyletic, the morphology of the Urmetazoan remains mysterious, as Porifera may have evolved their body plan after the divergence of the Porifera and Eumetazoa lineages. Currently, there is a limited taxonomic diversity of poriferan sequence data available for phylogenetic reconstruction and multiple attempts to determine if Porifera are a unified group (15, 18, 20, 21) have resulted in conflicting results, although the most recent analysis indicates that Porifera are monophyletic (15).

In summary, the resolution of early branching within the animals is inherently difficult to determine. Nonetheless, resolution of these relationships will be essential to gain a complete understanding of the biology of the ancestral animal.

Epithelial cell adhesion

Intercellular adhesion is a fundamental characteristic of multicellularity. To understand the influence of cell-cell adhesion animal origins, it is necessary to reconstruct how adhesive mechanisms first evolved. One of the most iconic examples of intercellular adhesion in animals is in the epithelium. Eumetazoan epithelia are replete with special cell-cell junctions (including adherens, demosomal, gap, and tight/septate junctions) and cell-substrate junctions regulated by hemidesmosomes and focal adhesions that attach to a basal lamina (22). However, many of these characteristics are altered or absent from sponges, despite the fact that they have close lateral contacts as well as the barrier functions of epithelia (23). In fact, organisms as divergent from animals as the slime mold *Dictyostelium discoideum* can assume configurations in which cells have a polarized organization that defines a tissue boundary and can regulate the directional secretion of material, although the homology with animal epithelia is less than clear (24). Recent work has indicated that sponge epithelia meet the functional definition for epithelia and on-going examination of sponges continues to reveal morphological and functional similarity to bilaterian epithelia (23, 25). Furthermore, the sequencing of sponge genomes and transcriptomes as well as the sequencing of choanoflagellates and a related unicellular organism, *Capsaspora owczarzaki*, indicates that genes associated with epithelial structures and functions in Bilateria are present in sponges and, in some cases, arose prior the evolution of animals (6, 14, 26).

Of the animal-specific cadherins, classical cadherins are potentially the most relevant when attempting to identify genes that contributed to animal origins. Classical cadherins are transmembrane proteins that promote cell adhesion in epithelia (and other tissues) through homophilic binding to classical cadherins on neighboring cells. In addition, classical cadherins modulate the actin cytoskeleton through interactions with β-catenin, which dimerizes with α-catenin and binds to actin filaments. The classical cadherin/β-catenin complex, and the subsequent coordination of actin polymers leads to

the formation of an "actin belt" that extends from adherens junctions between epithelial cells and provides a permeability barrier between the apical and basal surfaces of the epithelial tissue layer. Like classical cadherins, β-catenin appears to be an animal innovation (27) (although there is a distant β-catenin relative, aardvark, in *D. discoideum*) (28). A yeast two-hybrid screen using β-catenin from the sponge *O. carmela* suggests that sponge β-catenin has the ability to interact with a sponge classical cadherin, indicating that the classical cadherin/β-catenin complex may be conserved in sponges, but awaits experimentation *in vivo* to determine if this interaction regulates the integrity and polarity of sponge epithelia (27).

In contrast to β-catenin and classical cadherins, α-catenin is broadly distributed in eukaryotes. Recent work by Dickinson *et al.* has characterized the function of an α-catenin homolog in *D. discoideum* (24). Although *D. discoideum* achieves multicellularity by aggregation rather than by cell division, epithelial polarity in *D. discoideum* requires the α-catenin homolog, which they found interacts with the b-catenin relative, aardvark. Although this interaction could be an example of convergent evolution, the observation that a catenin complex is essential for epithelial polarity in both *D. discoideum* and metazoans suggests that the catenin complex could be an ancient functional module. This complex may have functioned in cell polarity before the divergence of social amoebae and metazoans and was deployed to organize cell polarity in a multicellular context in both social amoebae and metazoans. While the insights into catenin function from *D. discoideum* are fascinating, the lack of cadherins, Wnt-signaling components, and homologs of animal tissue polarity proteins (e.g. PAR, Crumbs, and Scribble) in *D. discoideum* limit their value for elucidating how epithelial polarity in metazoans evolved (24). For this we will need further functional insights from other lineages such as the choanoflagellates and sponges.

Cell adhesion and tissue integrity in Eumetazoa are also mediated by interactions between epithelia and the basal lamina, a layer of extracellular matrix (ECM) comprised of collagen, laminin, fibronectin, proteoglycans and polysaccharides, that anchors epithelial cells and helps maintain cell polarity (29, 30). As with cell-cell adhesion, many of the molecules required for cell-substrate attachment emerged prior to or early in animal evolution. Integrins, which connect cells to the ECM, as well as the ECM proteins that comprise the basal lamina, have been detected in all major animal phyla, including some sponges, despite the fact that most Porifera lack an identifiable basal lamina (23, 30). In addition, integrin subunits and their predicted intracellular binding partners are found in the genome of a unicellular relative of animals, *C. owczarzaki*, and an integrin alpha domain has been reported in choanoflagellates (4, 26). As with many of the "epithelial" genes, the early emergence of integrins and ECM components indicates that these genes evolved prior to their function in the multicellular context of animals, suggesting that they had functions in the unicellular and simple colonial progenitors of animals. It is possible that modern-day epithelial polarity genes in animals had ancestral functions in the establishment cell polarity. Alternatively, ECM and adhesion genes may have mediated the interactions of cells with inert environmental substrates through the deposition of ECM, such as the choanoflagellate theca, but these hypothesis and others await experimentation.

Developmental signaling: organizing in space and time

A growing body of data from the study of the evolution of development indicates that genes involved in the regulation of development have a prominent role in morphological evolution. Seven intercellular signaling pathways are traditionally considered critical and unique to animal development: nuclear hormone receptor, WNT, TGF-b, Jak/STAT, Notch/Delta, Hedgehog, and receptor tyrosine kinases (RTK) (31). Conserved components of all of these pathways are expressed in sponges, indicating they were present in the last common ancestor of animals (6, 32, 33). In contrast, there is little evidence from the gene content of choanoaflagellates for the complete multi-component developmental signaling pathways typically thought to function in animals (4). While this could be due to our inability to detect homologs in choanoflagellates because of the deep divergence time, it seems likely that many of the components are simply not present in choanoflagellates. However, some signaling pathway components and signaling domains from animals are present in choanoflagellates, including RTKs, NL domains, the Hedgehog signal domain and the Hint domain (which together make up animal hedgehog proteins), the hedgehog receptor Patched, Dispatched, and STAT (4, 7). The presence of these components in choanoflagellates suggests that rudiments of some animal developmental signaling pathways were present in the last common ancestor of choanoflagellates and animals. Interestingly, the protein domains diagnostic of proteins in these pathways are frequently found in novel configurations compared to those in Bilateria, indicating that animal developmental signaling pathways were assembled (at least in part) by domain shuffling of preexisting domains. Alternatively, the fully assembled signaling pathways may have predated the origin of choanoflagellates, but their component proteins were shuffled within the choanoflagellate lineage (4, 34).

The evolutionary history of the Hedgehog signal domain is an excellent example of the importance of domain shuffling in the metazoan stem lineage. The Hedgehog signal domain is best known as a secreted ligand in the Hedgehog signaling pathway that regulates developmental patterning in bilaterians. In this context the Hedgehog ligand is composed of a N-terminal signaling domain that is autocatalytically cleaved from (and by) the C-terminal HINT domain (35). The sequencing of the *M. brevicollis* and *A. queenslandica* genomes revealed a new molecular context for this domain on the extracellular N-terminus of a cadherin molecule, dubbed Hedgling (4, 36). Both genomes also encode HINT domains in other proteins, indicating that these domains evolved independently. Although Hedgling and Hedgehog genes are absent from the Trichoplax genome, Hedgling, HINT domains and the classically described hedgehog protein are encoded by the genome of the cnidarian *Nematostella vectensis*, suggesting that the Hedgehog signal domain and the HINT domain were brought together by domain shuffling and that Hedgling was lost prior to the evolution of bilaterians (36). The expression of Hedgling in a developmentally regulated and localized manner in *A. queenslandica* highlights the need for experimental inquiry to characterize the functional significance of the bioinformatic observation. The functional characterization of the hedgehog-signal domain in choanoflagellates, Porifera and Cnidaria promises to inform our understanding of Hedgehog function in ancestral animals and beyond.

Tyrosine kinase signaling had long been considered a hallmark of intercellular communication, and potentially unique to animals (37). However, ESTs and subsequent genome sequencing of choanoflagellates revealed a remarkable number and diversity of tyrosine kinase signaling molecules (7). The choanoflagellate *M. brevicollis* has 128 tyrosine kinases, 38 tyrosine phosphatases, and 123 phosphotyrosine (pTyr)-binding SH2 proteins, all higher numbers than cataloged in any metazoan to date (4, 38, 39). In addition, the *M. brevicollis* tyrosine kinases display diverse combinations of signaling protein domains, including combinations of signaling domains that do not occur in metazoans. This uniquely divergent and elaborate signaling network in *M. brevicollis* potentially illuminates the early evolution of pTyr signaling.

Two choanoflagellate Src kinases, MbSrc1 and MbSrc4, have received experimental attention that begins to characterize the function of tyrosine kinase signaling in choanoflagellates and offer insights into their function in the last common ancestor of choanoflagellates and animals. Li and colleagues cloned and purified the non-receptor tyrosine kinase MbSrc1 from *M. brevicollis* and found that the individual Src homology 3 (SH3), SH2, and catalytic domains have similar functions to their mammalian counterparts (40). However, in contrast to mammalian c-Src, the SH2 and catalytic domains of MbSrc1 do not appear to be functionally coupled, suggesting that Src autoinhibition likely evolved more recently within the metazoan lineage.

Li and colleagues also cloned MbSrc4, which contains a lipid-binding C2 domain in the N-terminus of a protein with SH3-SH2-kinase domains and found that the enzyme is highly active as a tyrosine kinase and that the C2, SH3, and SH2 domains function to localize the kinase in a manner similar to mammalian Src-like kinases. The membrane-binding activity of the C2 domain functions similarly to the myristoylation signal of c-Src, suggesting that, like c-Src, MbSrc 4 interacts with membranes, but using mechanisms that arose through convergent evolution. When expressed in mammalian cells, full-length MbSrc4 displays low activity toward mammalian proteins, and it cannot functionally substitute for mammalian c-Src in a reporter gene assay. Removal of the MbSrc4 C2 domain leads to increased phosphorylation of cellular proteins. This suggests that in contrast to the related *M. brevicollis* Src-like kinase MbSrc1, MbSrc4 is not targeted properly to mammalian Src substrates, suggesting that the C2 domain plays a specific role in M. brevicollis signaling (40). Functional studies such as these have the power to illuminate evolutionary processes (such as domain shuffling) that contributed to animal origins.

Differentiation: divide (the work) and conquer

Gene regulation

New mechanisms for regulating gene expression were likely another important source of novelty during animal origins (41). Evo-devo, which has built its house on the characterization of cis-regulatory elements such as promoters, enhancers and other regulatory features, has yet to explore the cis-regulatory landscape of early branching animals. If the regulatory structures and networks of animals are critical to their

evolution and at certain levels canalized (42), investigating gene regulation in early branching animals may have important implications for reconstructing animal origins and for determining the "kernels" around which animal regulatory networks are built. Genome sequencing can reveal the gene content (and particularly transcription factor content) of poorly studied organisms and transcriptome sequencing and *in situs* can provide information about gene expression and localization (36 , 43 , 44 , 45). Finally, tools for manipulating gene function are currently being developed for diverse early branching animals. These, coupled with *in vitro* biochemical approaches (40, 46), open the door to future efforts to explore the regulatory landscape within which the first animals evolved.

Transcription factors

Transcriptional regulation is one of most crucial aspects of animal development, regulating cell fate, the cell cycle, patterning, proliferation, development, and differentiation. Most transcription factors that play important roles in bilaterian development originated before the divergence of extant animal phyla, although the sizes of these families in different lineages vary, e.g. with cnidarians having a transcription factor repertoire typically two to three times greater than that of Porifera and Placozoa (6, 47). The animal-specific transcription factors include members of the basic helix-loop-helix (bHLH) groups A and D, myocite enhancer factors 2 (Mef2), Fox, Sox, T-box, Ets, nuclear receptor (NR), Rel/nuclear factor-kappaB (NF-kappaB), basic-region leucine zipper (bZIP), and Smad families and a range of homeobox-containing classes, including ANTP, Prd-like, Pax, POU, LIM-HD, Six, and three-amino acid-loop extension (TALE) (41). The genome of *M. brevicollis* contains the standard set of transcription factors observed across eukaryotes as well as p53, Myc, and a putative Sox homolog but otherwise lacks the well-known animal transcription factors (4, 41, 48). However, some animal transcription factors (e.g. Churchill and LSF/Grainyhead (GRH)) that are absent from *M. brevicollis* are present in *C. owczarzaki*, indicating that some TFs previously thought to be animal specific evolved prior to the divergence of choanoflagellates and animals and were lost in the choanoflagellate lineage (41).

A small role for miRNAs in the evolution of animal development

Both miRNAs and piRNAs are inferred to have been present in stem Bilateria and deep sequencing of small RNAs from early branching animals has indicated that miRNAs and piRNAs have been present since the evolution of animals (21, 49). Because canonical miRNAs and piRNAs have not been detected in choanoflagellates, their evolution may have helped to usher in the era of multicellular animals. Animal miRNA evolution seems to have been very dynamic: pre-miRNAs of Porifera, Cnidaria and Bilateria have evolved distinct lengths and all miRNAs have been lost in *Trichoplax* (21, 49). In addition, none of the identified miRNAs have recognizable conservation between Porifera, Cnidaria and Bilateria and only a single *Nematostella vectensis* miRNAs has recognizable homology to a bilaterian miRNA. This could be either because it is, in fact, the only homolog of extant bilaterian miRNAs or because deep divergence of these organisms masks the common ancestry of their miRNAs. These observations of miRNAs mirror reports of miRNA–target interactions within the

nematode, fly and vertebrate lineages, where very few appear to be conserved throughout all three lineages (49). Although the mechanism of miRNA action was present in the Urmetazoan, the influence of miRNAs on early animal evolution may have been relatively labile.

Conclusions

The evolution of multicellularity is often cited as a "major transition" in evolution and it is thought to have been punctuated by the emergence of abundant novelty (50). However, the overall picture emerging from the study of opistikonts and early branching animals is more consistent with the model of evolution as a tinker (51). Multicellularity evolved numerous times (52), and in closely related unicellular and multicellular species, such as *Volvox carteri* and *Chlamydomonas reinhardtii* as well as *S. rosetta* and *M. brevicollis*, it has proven difficult to identify the causative gene content differences that explain the differences in their biology (46, 53). With the genomes of several relatives of animals now being sequenced (54), we find that traces of many of the mechanisms used by animals for cell adhesion, signaling and differentiation span the transition from unicellularity to multicellularity in the animal lineage, indicating that many of these domains and genes were functioning in unicellular contexts and co-opted to new functions in animals. As the field continues to refine the branching order of early animals and characterize the gene content of additional opsithokonts, we aim to reconstruct the evolutionary history of developmental processes in animals. By experimentally interrogating the functions of genes and domains in organisms whose ancestors bridged the transition to multicellularity, particularly choanoflagellates and sponges, we can reconstruct the evolution of animal genomes and determine how the evolution of gene functions contributed to the origin of animals.

Figure Legends

Figure 1.1 Choanoflagellates are the closest known living relatives of the Metazoa. **A,** Reconstructing the biology of last common ancestor of animals and choanoflagellates, i.e. the "Urchoanimal" (1), is critical for understanding the evolutionary innovations the occurred on the stem lineages leading to the last common ancestor of animals (2; the "Urmetazoan") and the last common ancestor of choanoflagellates. **B**, Choanoflagellate cell morphology is typified by an ovoid or spherical cell body that is 3-10 μm in diameter and an apical flagellum surrounded by a collar of 30-40 microvilli (bracket). Beating of the flagellum creates water currents (dashed arrows) that can propel free-swimming choanoflagellates through the water column or trap bacteria (grey shape) and detritus against the collar of microvilli.

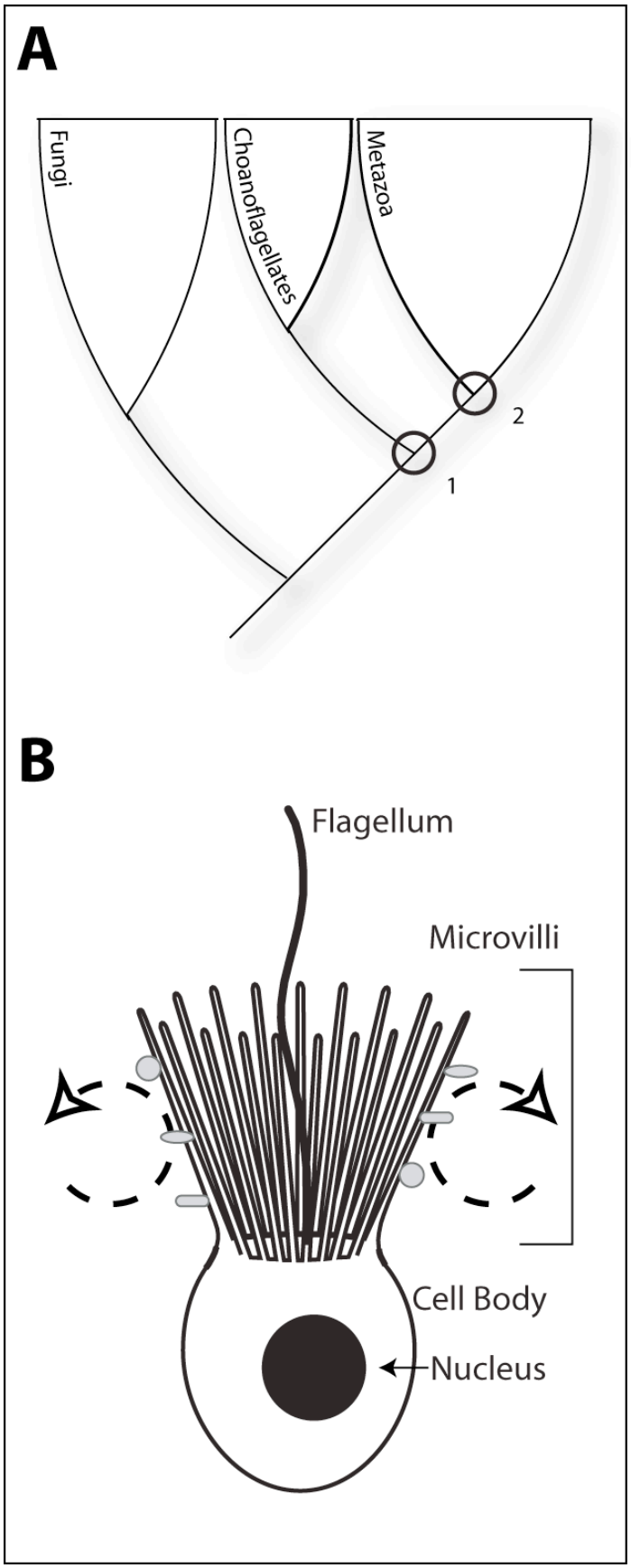

A
Fungi
Choanoflagellates
Metazoa
2
1
B
Flagellum
Microvilli
Cell Body
Nucleus

Figure 1.2 Proposed evolutionary relationships of among metazoans. Resolving the phylogenetic relationships of "early branching" animals is central reconstructing character states and understanding the biology of ancestral animals. **A**, Srivastava, *et al.* propose that the Porifera branch prior to the diversification of Eumetazoa. **B**, Schierwater, *et al.* propose that the Bilatera is sister to all other animals. **C**, Dunn, *et al.* propose that Ctenophora branch prior to the diversification of the other Metazoa. **D**, Philippe, *et al.* propose that Porifera branch prior to the diversification of Eumetazoa.

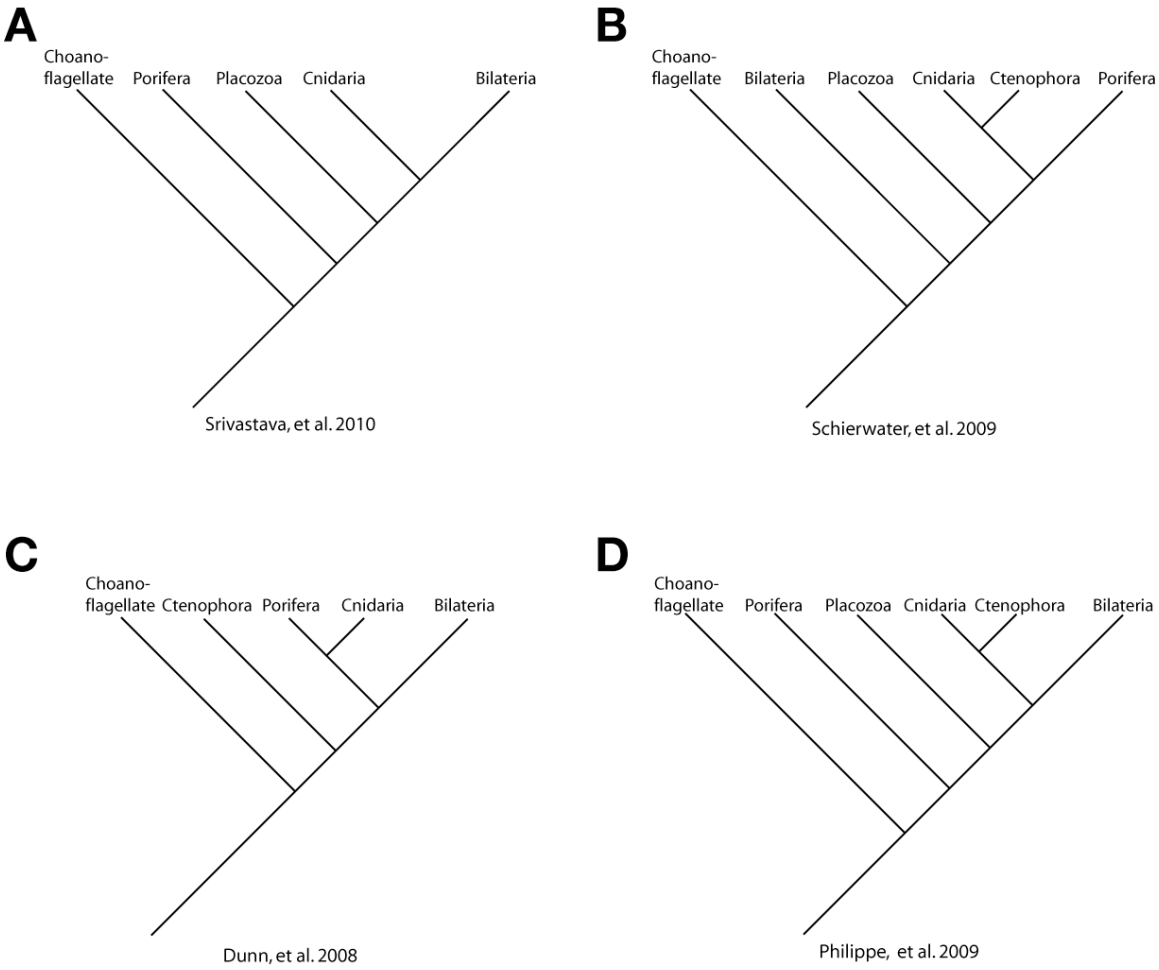


Figure 1.3 The phylogenetic distribution of important animal development genes. Despite a multicellular life stage being present in *S. rosetta* and absent in *M.* brevicollis, the developmental gene content between the two species is remarkably similar. *S. rosetta* has a fibronectin type 2 domain and a *bona fide* STAT protein, neither of which were identified in *M. brevicollis*. The phylogenetic tree above spieces indicates their evolutionary relationships and reveals that many of the animal development genes evolve after the divergence of choanoflagellates and animals. Speices included: *Homo sapien* (Hsap), *Drosophila melanogater* (Dmel), *Nematostella vectensis* (Nvec), *Amphimedon queenslandica* (Aque), *Salpingoeca rosetta* (Sros), *M. brevicollis* (Mbre), *Capsapora owczarzaki* (Cowc), *Neurospora crassa* (Ncra), *Saccharomyces cerevisiae* (Scre), *Dictyostelium discoideum* (Ddis).

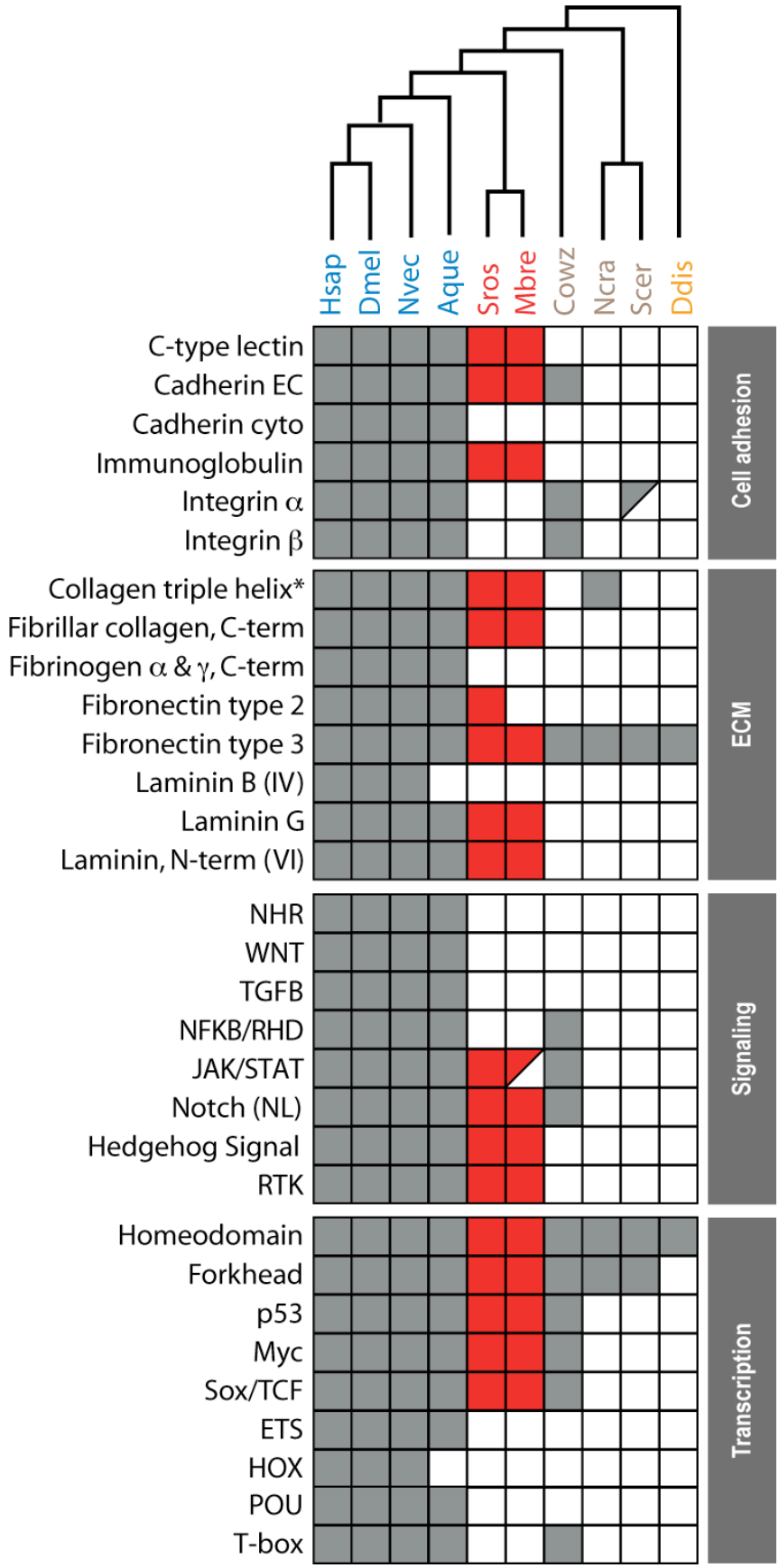
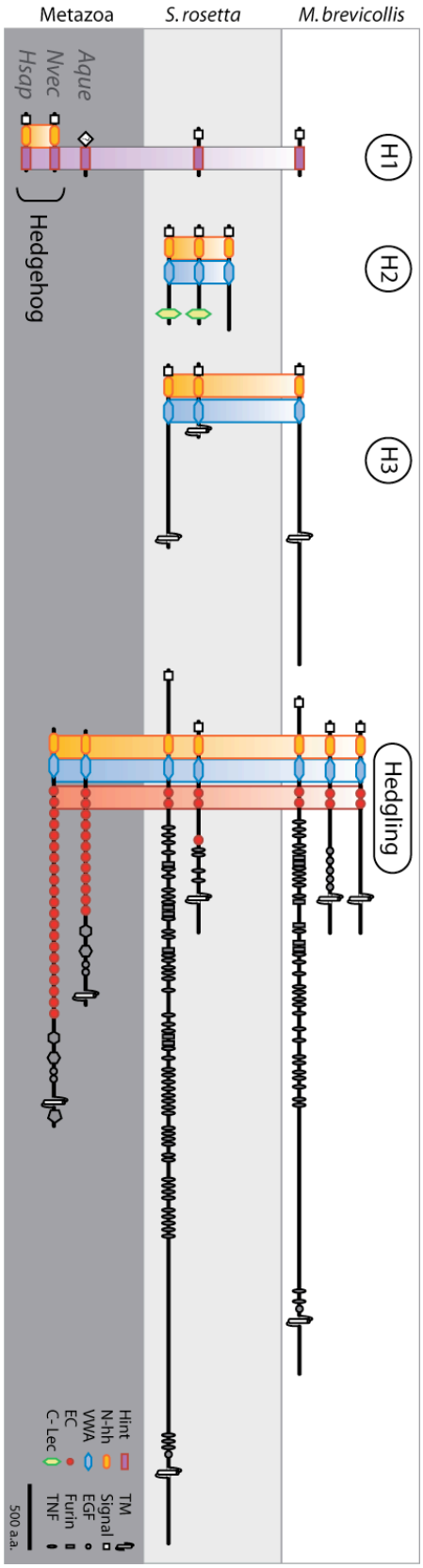

Hsap
Dmel
Nvec
Aque
Sros
Mbre
Cowz
Ncra
Scer
Ddis
C-type lectin
Cadherin EC
Cadherin cyto
Immunoglobulin
Integrin α
Integrin β
Cell adhesion
Collagen triple helix*
Fibrillar collagen, C-term
Fibrinogen α & γ, C-term
Fibronectin type 2
Fibronectin type 3
Laminin B (IV)
Laminin G
Laminin, N-term (VI)
ECM
NHR
WNT
TGFB
NFKB/RHD
JAK/STAT
Notch (NL)
Hedgehog Signal
RTK
Signaling
Homeodomain
Forkhead
p53
Myc
Sox/TCF
ETS
HOX
POU
T-box
Transcription

Figure 1.4 The diversity of choanoflagellate Hedgehog signal domain proteins. The Bilaterian Hedghog protein contains both an N-terminal Hedgehog signal domain with Hint domain (H1). The choanoflagellate proteome contains proteins that contain Hint domains and a number of Hedgehog signal domains in diverse and novel protein architectures. Both *M. brevicollis* and *S. rosetta* contain Hedgehog proteins associated with transmembrane domains (H3) a subset of which also contain cadherin domains (Hedglings). *S. rosetta* also contains three short peptides that contain a signal sequence and Hedgehog signal domain (H2), two of which also contain C-lectin domains.


Metazoa
S. rosetta
M. brevicollis
Aque
Nvec
Hsap
Hedgehog
H1
H2
H3
Hedgling
Hint
N-hh
VWA
EC
C-Lec
TM
Signal
EGF
Furin
TNF
500 a.a.

Chapter 2: Multicellular development in the choanoflagellate *S. rosetta*

The results presented here were published as part of the following paper:

Fairclough SR, Dayel MJ, and King N (2010) **Multicellular development in a choanoflagellate.** *Current Biology*. 20: R875-R876.

Little is known about how the first animals evolved from their single celled ancestors. Over 120 years ago, Haeckl proposed that animals evolved through "repeated self-division of [a] primary cell,"(55) an idea supported by the observation that all animals develop from a single cell (the zygote) through successive rounds of cell division (56). Nonetheless, there are multiple alternative hypotheses (57), including the formal possibility that multicellularity in the progenitor of animals occurred through cell aggregation, with embryogenesis by cell division being secondarily derived. The closest known relatives of animals, choanoflagellates, are emerging as a model system for testing specific hypotheses about animal origins (2, 13, 58, 59). Studying colony formation in choanoflagellates may provide a context for reconstructing the evolution of animal multicellularity. We find that the transition from single cells to multicelled colonies in the choanoflagellate *Salpingoeca rosetta* occurs by cell divison, with sister cells remaining stably attached.

While the life cycles of all choanoflagellates feature a prominent single-celled phase, many species are also capable of forming colonies of morphologically similar cells (3, 8, 60). Phylogenetics and the reconstruction of ancestral character states within the choanoflagellate group indicate that colony formation either evolved before the diversification of two of the three major choanoflagellate clades, or that it evolved multiple times independently (13). It is also possible that the last common ancestor of animals and choanoflagellates was capable of forming multicelled colonies (13). Thus, studies of the colony-forming choanoflagellate *S. rosetta* offer a unique opportunity to test hypotheses about the cell biology of colony formation and its potential relevance to the evolution of animal multicellularity.

Experimental Procedures

S. rosetta culture conditions

S. rosetta is a colonial choanoflagellate isolated from Hog Island, Virginia. *S. rosetta* is cultured with co-isolated bacteria at 25^{o}C in natural seawater infused with cereal grass media (61).

Production of a colony-free culture of *S. rosetta*

S. rosetta stock (ATCC 50818) was treated with a combination of antibiotics (chloramphenicol 68 µg ml^{-1}, ampicillin 50 µg ml^{-1}, streptomycin 50 µg ml^{-1}, erythromycin 50 µg ml^{-1}) and grown shaking in the absence of light. Every 48 h, the

culture was pelleted and resuspended in fresh media containing antibiotics. This procedure was repeated twice. The resultant choanoflagellate culture line, which contained antibiotic-resistant bacteria as food sources, was observed to lack colonies. The culture line was subsequently propagated over several weeks in antibiotic-free growth media and stocks were stored under liquid nitrogen. The culture line continues to lack colonies after several years of study.

Colony Induction

Colony-free cultures of *S. rosetta* cells were induced to form colonies by supplementation with the bacterium *Algoriphagus* sp. as a food source.

Video Microscopy

S. rosetta cultures were split 1:9 in fresh culture medium, inoculated with bacteria that induce colony development (*Algoriphagus* sp.) and incubated at 25^{o}C for 8 hours. Cells were mounted on a slide under a coverslip with #0 spacers. Phase images were captured through the 40X objective every 30s for 12 hours on a Leica DMI6000B Microscope equipped with a Leica DFC350 FX camera. Multiple movies were captured and the movie most clearly illustrating the process is presented in S1.

Inhibition of cell proliferation by aphidicolin

S. rosetta cultures were counted and split to a density of 1.25 X 10^5 cells per ml in 15 mL of fresh media and treated either with 51 μl of 5mg/ml aphidicolin in DMSO (15 mM) or with 51 μl of DMSO. Induced cultures were inoculated with 50 μl of stationary phase *Algoriphagus* sp. to induce colony development. After 36 hours, aphidicolin-treated cells were pelleted by centrifugation for 10 min at 3220x g. Half of the culture was resuspended in fresh media with *Algoriphagus* sp. and aphidicolin. The other half of the culture was resuspended in fresh media with *Algoriphagus* sp. and DMSO. Three biological replicates per treatment regime were scored for total cell number and number of cells in colonies every 12 hours.

Results

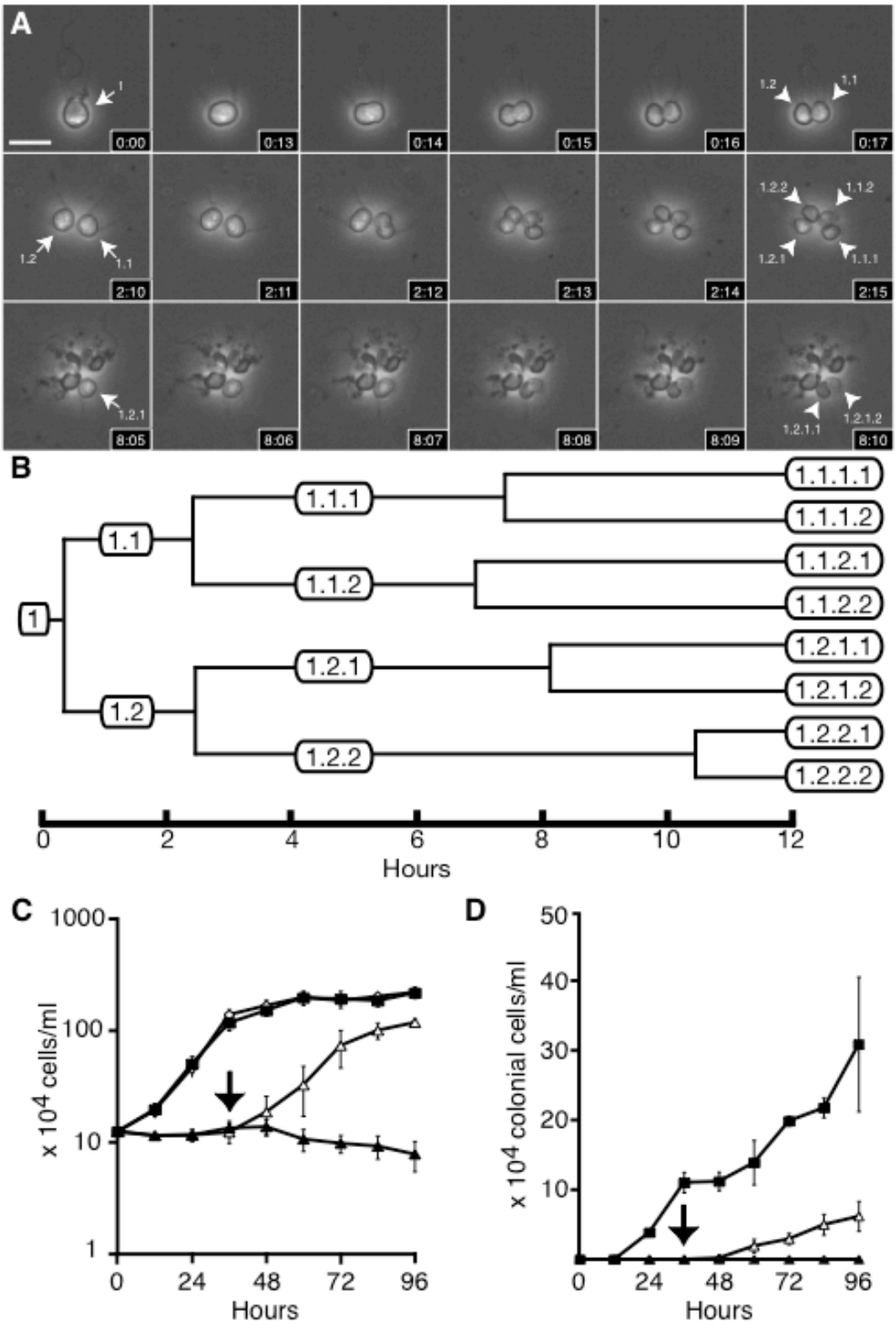
S. rosetta can exist as either single cells or rosette-shaped colonies that contain between 4 and ~50 cells arranged in closely packed spheres (Fig. 1A). To determine how colonies form, cultures of solitary *S. rosetta* cells were induced to form colonies by co-cultivation with the prey bacterium *Algoriphagus* sp. and monitored for at least 12 hours by time-lapse microscopy (see Supplemental Information). *S. rosetta* colonies were consistently observed to form through cell division and never by aggregation (Fig. 2.1A). Cell division during colony formation was asynchronous, suggesting that the cell cycle is not coordinated between sister cells in colonies (Fig. 2.1B).

Although direct observation demonstrated the centrality of cell division in colony formation and provided no evidence for cell aggregation, it is formally possible that *S. rosetta* colonies might form by aggregation at low frequency or under conditions that do not favor cell proliferation. In this case, colony formation through aggregation might be observed in cultures in which cell division is blocked. Therefore, we tested whether the cell cycle inhibitor, aphidicolin (62), can block cell proliferation in *S. rosetta* and thereby

block colony formation. In the presence of aphidicolin, *S. rosetta* cells fail to divide, yet continue to increase in size and otherwise appear to behave normally (Fig. 2.1C); upon removal of the drug, cell division resumes. To test whether colonies can form in the absence of cell division, *S. rosetta* cells were treated with either aphidicolin or DMSO (as a negative control) prior to induction of colony formation (Fig. 2.1D). DMSO-treated cultures developed colonies within 24 hours after induction, while cultures incubated with aphidicolin failed to form colonies, even after 96 hours of induction. Removal of aphidicolin from induced cultures after 36 hours of treatment permitted the development of colonies, demonstrating that the drug's effect was reversible and that the formation of colonies is dependent upon cell proliferation. Taken together, these findings demonstrate that rosette colonies form by cell division and not by cell aggregation.

Our finding that *S. rosetta* colonies develop through repeated cell division, coupled with the fact that development from a single cell is ubiquitous in animals, is consistent with the hypothesis that the last common ancestor of animals and choanoflagellates was capable of simple transitions to multicellularity. An important test of this hypothesis will be to determine whether colony formation is, indeed, ancestral within choanoflagellates and whether *S. rosetta* colony development is representative of an ancestral strategy for multicellular development. If so, the study of colony development in *S. rosetta* may provide mechanistic insights into early stages in the evolution of animal multicellularity and reveal the premetazoan function of developmental genes and their regulation.

Figure 2.1 *Salpingoeca rosetta* colonies develop through cell division, not aggregation. (A) Still images from a time-lapse movie show *S. rosetta* cells dividing (arrows) during colony development. In each case, the sister cells (arrowheads) remain attached. Extra-colonial cells were never observed to join a colony through aggregation. Scale bar represents 10 μm. The founder cell and its progeny were marked 1 - 1.2.1.2 to generate the cell pedigree in panel B. Time since start of movie (hours:minutes) is indicated in lower right of each panel. **(B)** The mapping of cell pedigree as a function of time, based on the time-lapse movie in panel A, shows that cells divide asynchronously during colony formation. **(C)** The cell cycle inhibitor aphidicolin prevents *S. rosetta* cell proliferation and the effect is reversed by removal of the drug. Aphidicolin (filled triangle) blocks cell proliferation relative to untreated cultures (in which colony formation is either induced (filled square) or uninduced (open square)). Removal of aphidicolin (arrow) allows the resumption of proliferation (open triangle). **(D)** Cell cycle inhibition prevents *S. rosetta* colony development. Relative to untreated *S. rosetta* cultures (filled square) in which the number of cells in colonies increases steadily after induction of colony development, induced cultures treated with aphidicolin (filled triangle) lack colonies. After removal of aphidicolin (arrow), the number of cells in colonies increases (open triangle), demonstrating that colony formation is dependent on cell proliferation. Error bars represent the standard error of the mean for each time point from triplicate cultures.


A
1
0:00
0:13
0:14
0:15
0:16
1.2
1.1
0:17
1.2
1.1
2:10
2:11
2:12
2:13
2:14
1.2.2
1.1.2
1.2.1
1.1.1
2:15
1.2.1
8:05
8:06
8:07
8:08
8:09
1.2.1.1
1.2.1.2
8:10
B
1
1.1
1.2
1.1.1
1.1.2
1.2.1
1.2.2
1.1.1.1
1.1.1.2
1.1.2.1
1.1.2.2
1.2.1.1
1.2.1.2
1.2.2.1
1.2.2.2
0
2
4
6
8
10
12
Hours
C
1000
100
10
1
x 10^4 cells/ml
0
24
48
72
96
Hours
D
50
40
30
20
10
x 10^4 colonial cells/ml
0
24
48
72
96
Hours

Chapter 3: The genetic basis of multicellular development in the choanoflagellate *Salpingoeca rosetta*

Introduction to *S. rosetta* biology

Animal multicellularity and development are rooted in basic mechanisms of cell adhesion, signaling and differentiation that were present in animal ancestors over 600 million years ago (1, 2). Understanding the unique evolution of animals, including the evolution of multicellularity and development, requires reconstructing the biology of early animals and their predecessors. Two ancestors central to understanding animal origins are the Urmetazoan and the Urchoanimal, the last common ancestor of animals and their closest living relatives, the choanoflagellates (Fig. 3.1A). By reconstructing the biology of the Urmetazoan and the Urchoanimal we can identify genetic changes coincident with the evolution of animals and identify characteristics that distinguish animals from other organisms.

Choanoflagellates are aquatic microeukaryotes with a distinctive cell morphology: an ovoid or spherical cell body 3-10 μm in diameter topped by a single apical flagellum that is surrounded by a collar of 30-40 microvilli (Fig. 3.2A). All choanoflagellates studied to date are obligate bacterivores. To feed, choanoflagellates beat their flagellum, creating water currents that can propel free-swimming choanoflagellates through the water column and trap bacteria and detritus against the collar of microvilli. The sequencing of the *Monosiga brevicollis* genome provided our first comprehensive look at a choanoflagllate genome and a glimpse into the minimal gene content of the Urchoanimal, indicating that it had more signaling and adhesive capacity than previously thought. However, unlike the seemingly strictly unicellular *M. brevicollis*, many choanoflagellates have the capacity to form multicelled colonies (10, 13, 60, 63, 64). Understanding the genetic basis for colony development may provide fundamental insights into the evolution of intercellular interactions along the metazoan and choanoflagellate stem lineages.

Salpingoeca rosetta is a colony forming choanoflagellate with at least five morphologically and behaviorally differentiated cell types (Fig. 3.2A) (10). We have identified three solitary cell types, slow swimming, fast swimming, and substrate attached (also known as 'thecate'), and two colonial forms, linear and rosette. The slow swimming cell displays the characteristic *S. rosetta* morphology and has a central role in the *S. rosetta* life history because it can differentiate into the each of the other cell types (Fig. 3.2A). To understand the gene content of a colonial choanoflagllate and how it is deployed during its life history, we sequenced the genome and transcriptome of the choanoflagellate, *Salpingoeca rosetta*. We find that these cell types have distinct transcriptional profiles with unique historical signatures.

Results

Genome structure and gene annotation

Sequenced to 33x average coverage with a combination of Sanger and 454 technology, the ~55.44 Mb *S. rosetta* genome was assembled into 154 scaffolds with an N50 of 1.52 Mb (Table S3.1). We identified 19 putative full-length chromosomes with predicted telomeres at each end and 24 scaffolds with a predicted telomere at a single end (Table S3.2). Therefore, *S. rosetta* has a minimum of 31 chromosomes. 11,629 genes were predicted *ab initio* using the Broad annotation pipeline trained with Sanger sequenced ESTs and refined with 21 Gb of transcriptome sequence data collected from diverse life history stages. 98% of predicted genes are supported by transcriptome sequence data.

The *S. rosetta* genome is 33% larger and contains 26% more genes than the genome of the solitary choanoflagellate *M. brevicollis* (4). Although smaller than most animal genomes, the number of protein coding genes in *S. rosetta* is comparable to animals with relatively low numbers of genes such as *Drosophila melanogaster* (13,767) (65). *S. rosetta* genes, which contain an average of 7.5 introns/gene, have comparable intron density to human genes (7.7 introns/gene) and greater intron density than *M. brevicollis* (6.6 introns/gene) and *C. owczarzaki* (3.8 introns/gene), a sister of choanoflagellates and animals.

Evolutionary assembly of animal and choanoflagellate genomes

Comparisons of the *S. rosetta* and *M. brevicollis* genomes with those of diverse animals and animal outgroups reveal the minimal gene content of long-extinct ancestors and the patterns of gene gain and loss that accompanied their evolution. Although the Urmetazoan and the Urchoanoimal genomes each contained at least 10,000 genes, ~10% of genes in the Urchoanimal were replaced with novel genes along the metazoan stem lineage. In contrast, the evolution of choanoflagellates from the Urchoanimal was characterized by extensive gene loss (~40% of the Urchoanimal genome) that was only weakly counteracted by gene gain. In addition, the *S. rosetta* and *M. brevicollis* lineages have each experienced massive gene gain, with 36% of the *S. rosetta* gene content and 33% *M. brevicollis* gene content is unique to each species, despite being the most closely related species in our analysis (13).

That the biology of animals is so different from that of other multicellular lineages can be attributed, in part, to the evolution of new genes with new functions. For example, the evolution of the Urchoanimal from the holozoan ancestor was marked by a disproportionate expansion in the number of genes associated with animal cell adhesion and cell-junction organization, including cadherins, PATJ (a component of adherens junctions) and vab-19, which is involved in epidermal attachment in animals (Fig. 3.1C, supplement), as well as genes involved in animal neuropeptide signaling and glycosphingolipid metabolism. The *S. rosetta* genome, like the genome of *M. brevicollis*, contains a diverse and abundant repertoire of tyrosine kinases (4, 7, 38-40, 48), although only a handful are orthologous with animal tyrosine kinases (Table S3.3-5). In the *S. rosetta* genome we have identified Eph tyrosine kinases (Supplement), which were previously thought to be unique to metazoans. *S. rosetta* also contains a distinctive class of receptor guanylyl cyclase (RGC) kinases (Table S3.4). Unlike in metazoans, in which all RGC kinases have a catalytically inactive kinase domain, 11 out of 12 detected RGC

kinases in *S. rosetta* are predicted to be catalytically active. Using the same annotation methods applied to the *S. rosetta* genome, we have also identified homologs of Eph and RGC kinases in *M. brevicollis*, providing further evidence that these genes were present in the Urchoanimal.

While the Urmetazoan genome was dominated by genes that were present in the Urchoanimal, it also contained over 1000 new genes that had evolved along the metazoan stem. These genes may have signified key innovations that facilitated animal origins. Among these new genes were those now associated with animal cell adhesion, including tetraspanins, δ-catenin and β-laminin, as well as genes involved in TGF-β and Wnt signaling. Both TGF-β-signaling and Wnt-signaling play critical roles in animal development, and their misregulation results in serious human diseases, including cancer. While both these pathway are well studied in eumetazoan models, little is known about the evolutionary origin of its components and their functions. Our top-down approach across diverse organisms has identified many key genes of these signaling pathways as innovations in the metazoan stem lineage. In fact many of the core components from these pathways (TGF-β, TGF-β receptors, Smads, Wnts, Wntless, TCF) were identified in all the animals in our analysis, underscoring their importance to animal biology.

Transcriptional insights into *S. rosetta* cell differentiation

The differences in the gene contents of *S. rosetta* and *M. brevicollis* likely reflect differences in their biology, including the ability of *S. rosetta* to form simple colonies. However, the extent of these differences makes a simple comparative genomic approach ineffective for identifying the molecular underpinnings of colony formation or cell differentiation in *S. rosetta*. Therefore, to gain insight into the deployment of the *S. rosetta* genome during its life history, we generated 21 Gb of RNA-seq data from eight samples, each derived from cultures enriched for one of four different *S. rosetta* cell types (Fig. 3.2A) that were fed with either mixed environmental bacteria, the colony-inducing bacterium *Algoriphagus machipongonensis*, or mixed environmental bacteria supplemented with *A. machipongonensis* (Fig. S3.1A) (10, 64). The presence of different bacteria in different cultures allowed us to determine whether transcriptional differences were associated with the cell type of the choanoflagellates or related to responses to the bacterial prey.

The transcriptional profiles from each of the eight samples clustered according to the cell types they contained rather than the bacteria that they had been fed (Fig. 3.2B), indicating that the dominant transcriptional signal was reflective of cell type. Therefore, these data have the potential to highlight genes required for the regulation of cell differentiation. Using three different approaches to identify sets of genes that are over-expressed in colonies or attached cells (the two cell types for which we collected data from multiple biological replicates), we identified 480 "colony genes" that are over-expressed in cells from linear and rosette colonies (Fig. 3.2C,E) and 1410 "attachment genes" that are over-expressed in solitary attached cells (Fig. 3.2D,E). A subset of genes, 83 for colonies and 601 for attached cells, were detected using all three analytical approaches, increasing our confidence about their potential contributions to *S. rosetta* cell differentiation (Fig. S3.2).

All of the members of one gene family, the septins, were consistently and significantly over-expressed in colonies (Fig. 3.3B). Septins are a family of GTPases found in fungi and animals, where their functions include the formation of diffusion barriers and stabilization of midbodies and ring canals (the cytoplasmic bridges that form between developing sperm cells) (66). The wholesale elevation of septin gene transcription during colony formation is striking because colony formation in *S. rosetta* occurs by incomplete cell division (64), following which neighboring cells remain physically linked by cytoplasmic bridges (Fig. 3.3A) (10). Interestingly, *S. rosetta* homologs of other midbody-associated proteins and septin regulators, such as Aurora kinase, Anillin, and Polo kinase, are also over-expressed in colonies, consistent with a role for septins in colony formation or maintenance (Fig. 3.3A).

The *S. rosetta* genome encodes four septins that we find are similar to septins 2, 6, 7, and 9 from humans. The *S. rosetta* septin homologs have highly conserved G1, G3 and G4 core GTPase domains that are defining characteristics of septins (Fig. S3.3) (67-69). In addition, a putative N-terminal polybasic region, which binds to membrane phospholipids (70, 71), is present in each *S. rosetta* septin, as is the Septin Unique Element, a conserved motif whose function remains unknown. The C-terminal coiled-coil domain that mediates septin-septin polymerization in a subset of septins is detected in *S. rosetta* septins 6 and 7 as expected, but not predicted in septin 2, possibly because of an assembly or annotation error in the C-terminus of this gene model. *S. rosetta* septin 9, like some of its homologs from metazoans and fungi, lacks a coiled-coil domain (Fig. S3.4). By threading *S. rosetta* septin homologs onto the crystal structure of the human septin filament (containing septins 2, 6 and 7), we find high levels of predicted structural conservation at critical binding interfaces, suggesting that *S. rosetta* septins may form functional filaments (Fig. 3.3C,D). Therefore, the concerted elevation of septin transcription in colonies, coupled with the predicted structural conservation of *S. rosetta* septins and the conserved roles of animal septins in regulating cytokinesis suggest that *S. rosetta* septins may play an important role in the regulation of colony development.

Transcriptional diversity of cadherins and hedgehog relatives

One family of genes thought to have had important roles in early animal evolution, the cadherins, are also found in choanoflagellates. The *S. rosetta* genome encodes 27 cadherin proteins, a number that is comparable to the complement of cadherins found in the genomes of many animals (including *D. melanogaster* and *C. intestinalis*) (4, 14, 27). Animal cadherins are best known for their roles in cell adhesion and intercellular signaling, and they can also act as docking sites for pathogenic bacteria (72-74). By extension from these known animal functions, choanoflagellate cadherins may facilitate cell adhesion in colonies, transduce signals, function in prey capture, or may act in choanoflagellate-specific processes, such as attachment to environmental substrates (Fig. 3.4A). Of the 27 cadherins present in the *S. rosetta* genome, eleven are over-expressed in substrate-attached cells where they may help regulate attached cell differentiation, potentially through functions related to signaling or environmental substrate attachment. A different set of ten cadherins appears to be specifically over-expressed (Fig. 3.4B) in colonies. Many of these colony-specific cadherins are

surprisingly short (from 565 - 8158 a.a.), arguably too short to be mediating direct adhesion between neighboring cells. Instead, they may act either as signaling molecules or by interacting with the loose ECM surrounding cells in colonies (Dayel 2011). Notably, there is no correlation between the expression of specific cadherins and the species of bacterial prey fed to different samples, suggesting either that different cadherins are not specific for particular species of bacterial prey or that cadherins with functions related to prey capture are not regulated at the transcriptional level.

Among the cadherins that are over-expressed in substrate-attached cells are two *S. rosetta* homologs of Hedgling. Hedgling proteins are of particular interest because, in addition to cadherin repeats and a transmembrane domain, they contain a hedgehog signaling domain at their extracellular N-terminus (4, 27, 44). In addition to the Hedgling homologs, the proteome of *S. rosetta* has five additional predicted proteins that each contain a hedgehog signal domain. In each case, the hedgehog signal domain is positioned adjacent to a VWA domain, a configuration also found in *A. queenslandica* and *N. vectensis*, suggesting that this is an ancient combination of domains. Four of the hedgehog signaling domain-containing proteins also have a predicted transmembrane domain and these proteins are all over expressed in solitary attached sells (Fig. 3.4C). The remaining three hedgehog domain-containing proteins are relatively small and each lack a transmembrane domain; these proteins are consistently over-expressed in colonial cells where it is plausible that they act as secreted ligands. The *S. rosetta* genome also encodes five proteins containing the Patched domain, which is diagnostic of animal hedgehog protein receptors. In *S. rosetta*, the Patched domain containing proteins show differential expression, with some over-expressed in colonies and others over-expressed in attached cells. Patched domain proteins have an ancient evolutionary history (75) and we predict that they may interact with one or more of the *S. rosetta* hedgehog domain containing proteins. If so, this interaction was likely present in the urchoanimal and may have provided the evolutionary basis for hedgehog signaling in animals.

Evolutionary signatures in the transcriptomes of differentiated cells

The genome of *S. rosetta* has a rich evolutionary history. We found that 58% of the *S. rosetta* genes originated prior to the evolution of the Holozoan ancestor, with the remaining genes originating in the stem lineages leading to the Urchoanimal (5%), the Urchoanoflagellate (6%), or *S. rosetta* (31%). We hypothesized that the transcriptional dynamics of genes necessary for *S. rosetta* cell differentiation may reflect their evolutionary history.

Within the set of genes over-expressed in substrate attached solitary cells, we find that genes uniquely shared with *M. brevicollis* are most enriched, followed by genes that are unique to *S. rosetta* or that originated along the Urchoanimal stem lineage (Fig. 3.5C). This enrichment of genes uniquely found in choanoflagellates may reflect a shared genetic basis for substrate attachment in *S. rosetta* and *M. brevicollis*, a cell behavior that is not directly shared with animals and is prominent in the life histories of both species. Substrate attachment genes unique to *S. rosetta* may pertain to ways in which the biology of the substrate-attached cell type, such as the construction of a theca, has been elaborated since the divergence with *M. brevicollis*.

In contrast with attached cells, the transcriptome of mature colonies (Fig. 3.5D) is dramatically enriched for genes unique to *S. rosetta*, with relatively little contribution from genes uniquely shared with *M. brevicollis*. This observation suggests that major aspects of *S. rosetta* colony biology evolved after the divergence from *M. brevicollis*. What is not clear from this data set is the nature of genes that contribute to the process by which colonies form from solitary swimming cells (10, 64). Therefore, we next analyzed the set of genes that are upregulated in solitary swimming cells and colonies (Fig. 3.5 F), as opposed to those that are either uniquely upregulated in colonies alone (Fig. 3.5E) or solitary swimming cells alone (Fig. 3.5G). In this dataset we find that genes shared by *S. rosetta* and animals are as enriched as those that are restricted to *S. rosetta*. Therefore, it is possible that the onset of colony development is regulated, in part, by genes that also contributed to animal evolution.

Discussion

Like a story that has been passed down through the generations, the *S. rosetta* genome provides a valuable, albeit incomplete, molecular record of premetazoan evolution. Nearly two-thirds of the genes in the Urchoanimal live on in the *S. rosetta* genome. Because these genes have continued to evolve since the divergence of the animal and choanoflagellate lineages, comparing their functions in choanoflagellates with their functions in animals may provide important new insights into the ancestral functions and evolution of critical animal genes.

The sequencing of choanoflagellate genomes has overturned our thinking on the evolution of several important animal gene families. For example, the genomes of *S. rosetta* and *M. brevicollis* have revealed that RTKs, genes essential for intercellular communication in animals as well as the regulation of development and cancer, evolved before the divergence of choanoflagellates and animals (4, 7). One of the largest families of RTKs in the animal kingdom, the Eph receptors, has now been detected in *S. rosetta* and *M. brevicollis*, indicating that these developmentally important signaling receptors were present in the Urchoanimal. Eph receptors are key regulators of the repulsion and adhesion of cells that underlie the establishment, maintenance, and remodeling of patterns of cellular organization in animals (76) and their discovery in *S. rosetta* lays the foundation for investigating both the core and ancestral functions of these important receptors.

Comparisons among animal and choanoflagellate genomes also reveal genes that are apparently restricted to animals and that may have played important roles in animal origins. For example, the ubiquity of the Wnt signaling pathway in animals, and its absence from all sequenced non-metazoan genomes (including *S. rosetta*), suggests that it evolved coincident with the evolution of animal epithelia, the most ancient animal tissue type. A critical function of the Wnt pathway in animals is the regulation of epithelial cell polarity; recent work in *A. queenslandica* suggests that Wnt signaling may also regulate cell polarity in sponges (36, 45). Therefore, the Urmetazoan likely had an active Wnt pathway whose evolution may have served as a key genetic innovation during the evolution of animals from the unicellular ancestors. The Wnt pathway is but one example of the diverse gene families that seemingly evolved along the metazoan stem

(Fig. 3.1B). Increased sampling of sponges, choanoflagellates and other related unicellular organisms will further refine our portrait of the minimal gene content of the Urmetazoan, as will refined predictive models that improve our ability to annotate these genomes.

Functional genomics in extant taxa such as *S. rosetta* may help illuminate the molecular and cell biology of our long extinct ancestors. Through the transcriptional profiling of different *S. rosetta* cell types we aimed to identify potential regulators of cell differentiation, while also gaining insights into the biology of the Urchoanimal. Septins, a family of GTPases found in both fungi and animals, were significantly upregulated in colonies. Septins in animals have critical roles in cytokinesis, as diffusion barriers in primary cilia and in the formation of midbodies and ring canals (77, 78). The expression dynamics of the *S. rosetta* septins and their conserved roles in the regulation of cytokinesis suggest that they may contribute to *S. rosetta* colony development by forming and stabilizing the cytoplasmic bridges observed between colonial cells (10). If so, this phenomenon would highlight the role of cooption in the evolution of one example of simple multicellularity, that is the colonies of *S. rosetta*, and highlight the value of transcriptomics for rapidly interrogating the genetics of cell differentiation and devlopment in non-model organisms.

The sequencing of the *S. rosetta* transcriptome also allowed us to test hypothesess about the functions of animal genes in choanoflagellates. Cadherins are present in all animals and have key roles in regulating animal cell adhesion and signaling. Given their ubiquity and importance it is possible this family of proteins contributed to animal origins. The transcriptional profiling of *S. rosetta* cadherins revealed that, unlike septins, different cadherins have different expression patterns, with a subset being over-expressed in substrate-attached cells and while others are over-expressed in colonial cells. Of particular interest are the Hedglings, Cadherins that contain an N-terminal Hedgehog domain, as they are likely to be involved in signaling and may shed light on the evolution of the Hedgehog signaling pathway. Our analyses of the *S. rosetta* and *M. brevicollis* genomes have revealed several additional Hedgehog signal domain containing proteins (27) that lack either cadherin or HINT domains, but contain additional protein domains that expand their potential functions and complicate efforts to reconstruct the evolutionary history of the hedgehog domain (45). The three *S. rosetta* Hedgehog domain-containing proteins that lack a transmembrane domain are robustly over-expressed in colonial cells; these proteins also contain signal peptides and, like animal Hedgehog proteins, may act as secreted ligands. If these proteins also interact with homologs of the hedgehog receptor, Patched, we may infer that this interaction was present in the Urchoanoflagellate and potentiated the origin of the Hedgehog signaling pathway.

The genome of *S. rosetta* has a rich evolutionary history replete with an abundance of signaling, adhesion and differentiation genes including the examples we have highlighted with similarity to genes important in animals. The transcriptional profiling of the *S. rosetta* genome provides a comprehensive look at the deployment of its gene content over the course of its life history. This mapping of gene expression to cell biology begins to bring into focus a functional understanding of choanoflagellate gene

content and will hopefully allow us to gaze through time at ancestral function.

Material and Methods

Salpingoeca rosetta culture conditions

S. rosetta is a colonial choanoflagellate isolated from Hog Island, Virginia. *S. rosetta* is cultured with co-isolated bacteria at 25^{o}C in natural seawater infused with cereal grass media (79).

Production of a monoxenic culture of *S. rosetta*

A monoxenic culture of *S. rosetta* was generated in which the sole source of bacteria is *A. machipongonensis*. The undefined population of environmental bacteria in the ATCC 50818 culture was replaced through the following: the culture was treated with a combination of multiple antibiotics (ofloxacin 10 µg ml-1, kanamycin 50 µg ml-1, streptomycin 50 µg ml-1), serially diluted to further reduce the diversity of bacteria associated with the choanoflagellate culture, sorted by choanoflagellate cell size on a DAKO Cytomation MoFlo High Speed Cell Sorter (Carpenteria, CA), and finally supplemented with the colony-inducing bacterium *A. machipongonensis*. The resultant choanoflagellate culture line was propagated over several weeks in antibiotic-free growth media. Bacterial monoxenicity was assessed by plating on modified Zobell medium (Carlucci and Pramer, 1957) and restriction fragment length polymorphism analysis. When split daily for 1–2 weeks, this monoxenic culture produces cultures dominated by rosette colonies.

Isolation of *S. rosetta* genomic DNA

Genomic DNA was isolated from the monoxenic culture of *S. rosetta* grown only with *A. machipongonensis*. Genomic DNA was harvest and the *S. rosetta* and *A. machipongonensis* DNA separated on a CsCl gradient as described for the genome sequencing of *M. brevicollis* (4).

Genome Sequencing

Purified *S. rosetta* genomic DNA was sequenced with 454 and Sanger Whole Genome Shotgun methodology.

454 sequencing

Purified *S. rosetta* genomic DNA was shattered into small fragments (~0.6kb or~3kb) and isolated by electrophoresis. 0.6kb fragments were tailed with 454 sequencing adapters while 3kb fragments were circularized on a biotinylated linker, sheared, and fragments containing biotinylated linker retrieved and tailed with 454 sequencing adapters. Adapter ligated fragments were sequenced from one end, creating fragment or paired reads.

Sanger sequencing

Purified *S. rosetta* genomic DNA was shattered into small fragments (~4 kb or ~40 kb)

and captured in a vector and cloned. Captured fragments was sequenced from both ends to create paired reads.

Genome Assembly

Fragments generated from 454 data were first assembled with Newbler (http://454.com/products-solutions/analysis-tools/gsdenovo-assembler.asp). Assembly of assembled 454 and Sanger WGS sequence was carried out using the April 2009 version of HybridAssemble module of the Arachne assembler (http://www.broadinstitute.org/crd/wiki/index.php/HybridAssemble). The assembly was then manually modified to close additional gaps and break misassembled joins using Arachne tools.

Genome Annotation

Protein-coding genes were initially annotated using a combination of *ab initio* predictions (GeneMark.hmm-ES, AUGUSTUS, GlimmerHMM), protein sequence homology based evidence (blast, GeneWise), and transcript structures built from ESTs using PASA package. The package EVM (EVidenceModeler) was used to build gene models from all available input evidence. The obtained gene models were further improved by incorporating RNAseq data from eight different conditions using PASA and inchworm pipelines to get a final gene set.

TK annotation

Manual annotations for the *S. rosetta* kinases were made through BLAST, multiple sequence alignments, hidden markov models, presence or absence of accessory domains and phylogenic trees. *S. rosetta* kinases were compared to 9 previously annotated kinomes including 6 metazoans, 1 choanoflagellate, 1 fungus and 1 plant

Septin characterization

The final gene predictions for the *S. rosetta* genome included 5 septin domain containing proteins (PTSG_04106, PTSG_06009, PTSG_07215, PTSG_04363 and PTSG_04364) as predicted by Pfam (80). Upon further examination PTSG_04363 and PTSG_04364 were merged and this new gene model (PTSG_ 04364*) was further verified by sequencing directly from a *S. rosetta* cDNA library using specifically designed primers (5'TCAACGAAACGATTTCAAGC and 5'GTGGTCCGAGTTGTCGACTT). Conserved septin specific residues including the N-terminus polybasic region were identified manually while coiled-coil domains were predicted using the COILS program implementing the default settings (81). Sequences with average probabilities below 0.8 were not considered to have coiled-coil domains.

Septin homology mapping

Each *S. rosetta* septin was homology modeled using Loopp version 4.0 available through the University of Texas (http://clsb.ices.utexas.edu/loopp/web/) (82). Individual *S. rosetta* septin homology models were loaded into MacPymol (83) and similar residues determined using NCBI BLAST alignment and colored red. Each structure was then aligned to the crystal structure of the human septin filament 2QAG.PDB.

Phylogenetic analyses

The four *S. rosetta* septin sequences were added to Pan *et al*'s (84) alignment in order to establish putative gene homology assignments. Sequences were aligned using the Clustal Omega multiple sequence alignment program (85) and variable sequence regions were systematically removed using Gblocks (86). A maximum likelihood analysis was done on the resulting alignment of 183 amino acid characters using PHYML (87). The WAG substitution model was implemented with a mixed model of rate heterogeneity and 4 rate categories where the fraction of invariable sites and the gamma distribution parameter alpha were estimated from the data set. Bootstrap support (100 replicates) was estimated for the single resulting tree topology (Fig. S3.5).

Reconstructing gene gain and loss in Holozoa

To characterize how gene content changed during evolution of the Holozoa (descendents of the last common ancestor of *C. owczarzaki* and Metazoa) we identified orthologous genes from 34 sequenced genomes using OrthoMCL and mapped their origin using Dollo parsimony (Fig. 3.1A) (88, 89). Using this map we estimated the minimal gene content in the holozoan ancestor and estimated the extent of gene gain and loss along the subsequently diverging lineages.

Cell type enrichment

Solitary swimming cells were isolated from the supernatant fraction of cultures grown in the presence of mixed bacteria, but not *A. machipongonensis* (64).

Solitary attached cells cell cultures were collected from cultures by removing the supernatant, washing three times with 10mL of culture media and removing the attached cells from the plate surface with a plastic cell lifter.

Cultures consisting primarily of chain colonies were generated by diluting 2 mL of cells from the supernatant of solitary swimming cells into 15 mL fresh medium every day for 1–2 weeks.

Cultures consisting primarily of rosette colonies were produced using two different strategies. In the first approach, a culture solitary swimming cells was inoculated with live *A. machipongonensis* bacteria, which induces the development of rosette colonies that became the dominant form in the culture within 2 days. Rosette colonies were also isolated from cultures grown exclusively with live *A. machipongonensis.*

RNA-seq

Total RNA was isolated from *S. rosetta* cultures using the Qiagen RNAeasy kit followed by four consecutive rounds oligo-dt hybridization, washing, and elution with Qiagen Oligotex kit to purify mRNA. Purified mRNA was treated with Turbo DNA-free (Ambion) per manufactures recommendation. The integrity of mRNA was assessed using an Agilent 2100 Bioanalyzer and quantified using RNA Quant-it assay for the Qubit Fluorometer (Invitrogen).

We created dUTP second-strand libraries starting from 200 ng of Turbo DNase treated, polyA-enriched RNA, using a previously described method (90) with the following modifications. We fragmented RNA in 1x fragmentation buffer (Affymetrix) at 80°C for 4 min, purified and concentrated it to 6 µl after ethanol precipitation. In

addition, we added an 8-base barcode to each library to enable pooling of these libraries. The adaptor ligation step was done with 1.2 µl of index adaptor mix and 4000 cohesive end units of T4 DNA Ligase (New England Biolabs) overnight at 16°C in a final volume of 20 µl. Finally, we generated libraries with an insert size ranging from 225 to 425 bp.

Identification of differentially expressed genes

Pairwise comparison

Colonial/attached: Read count was compared between samples (CoA1/AtA, CoA2/AtA, CoAB/AtAB, CoAB/AtB, ChB/AtA, ChB/AtAB, ChB/AtB) using edgeR installed under Bioconductor v2.8. A gene was considered differentially expressed between colonial cells and attached cells if it was significantly differentially expressed (p-value <0.05) in at least three comparisons and a fold change greater than 1.5 in the remaining comparisons.

Colonial/swimming: Read count was compared between samples (CoA1/SwB, CoA2/SwB, CoAB/SwB, ChB/SwB) using edgeR installed under Bioconductor v2.8. A gene was considered differentially expressed between colonial cells and swimming cells if it was significantly differentially expressed (p-value <0.05) in at least two comparisons and had a fold change greater than 1.5 in the remaining comparisons.

Attached/swimming: Read count was compared between samples (AtA/SwB, AtAB/SwB, AtB/SwB) using edgeR installed under Bioconductor v2.8. A gene was considered differentially expressed between attached cells and swimming cells if it was significantly differentially expressed (p-value <0.05) in at least one comparison and had a fold change greater than 1.5 in the remaining comparisons.

Group comparison

Read count was compared between groups of samples using edgeR installed under Bioconductor v2.8 and considered differentially expressed for p-value<0.05. Colony/Attach: (CoA1, CoA2, CoAB, ChB Vs AtA, AtB, AtAB); Colony/Swim: (CoA1, CoA2, CoAB, ChB Vs SwB); Attach/Swim: (AtA, AtB, AtAB Vs SwB)

Iterative Heatmap clustering

FPKM values were log2(FPKM) transformed, quantile normalize, and filtered requiring Max(log2(FPKM)) -Min(log2(FPKM))>2. The filtered gene sets clustered hierarchically with gplots installed undere R v2.12 and sub-clusters were manually selected and re-clustered (Fig. S3.2). Sub-sampling and re-clustering was preformed 4 times resulting in the gene set presented in (Fig. 3.2 C-E).

OrthoMCL

Protein sets for 34 genomes, including 9 fungi, 11 animals and several others, were first filtered by removing short proteins less than 30 AA and shorter isoforms encoded from each gene (i.e., only the longest protein product was retained for each gene). Then an All-Vs-All blast (blastp, evalue cutoff 1e-5) was run on the filtered proteins. The blast output was loaded into a MySQL database, and we ran the

OrthoMCL2 pipeline to build ortholog groups with default parameters.

Enrichment Analysis

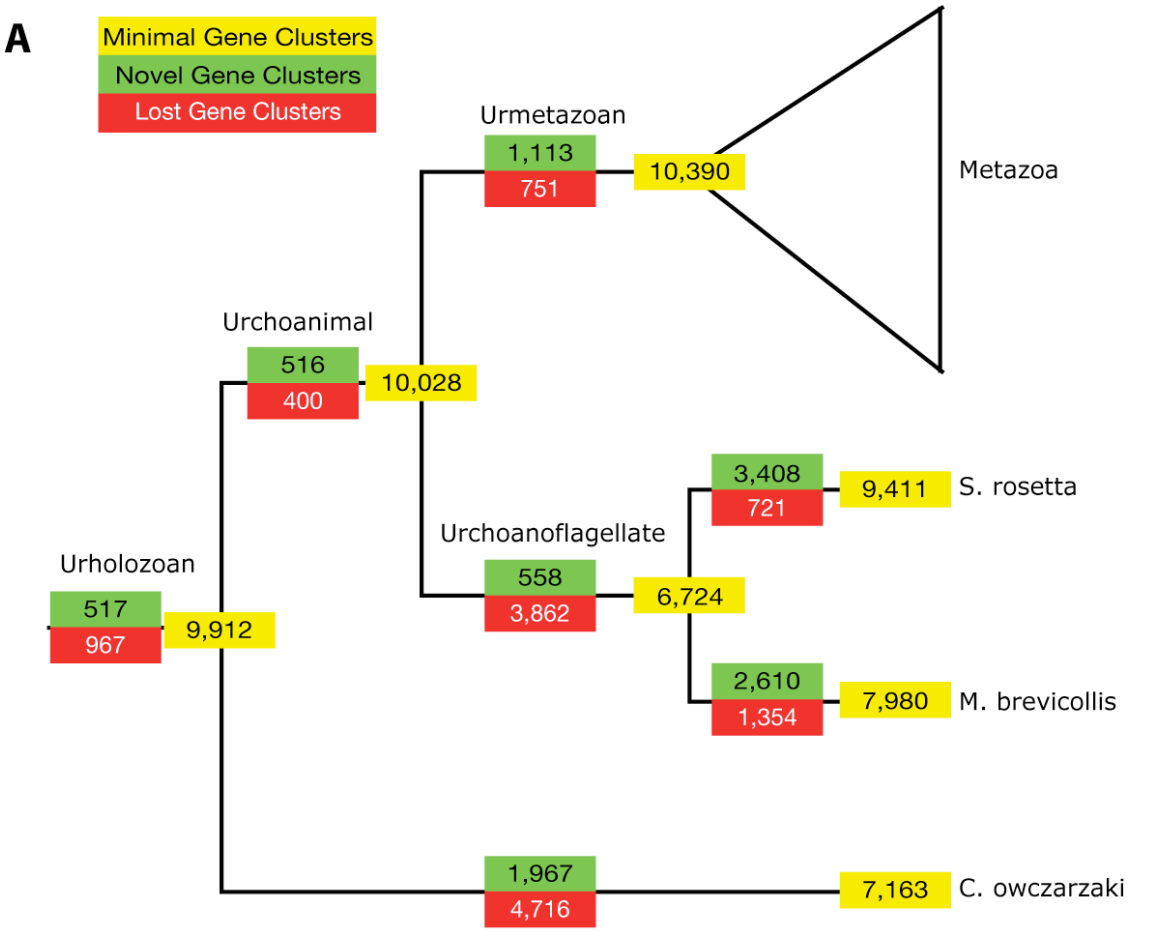
OrthoMCL clusters were annotated as ancient, metazoan, choanoflagellate or *S. rosetta*-specific based on the most distant member of the cluster. The relative frequencies of phylogenic annotations were calculated for the entire *S. rosetta* genome. Expression clusters were tested for phylogenic enrichment by comparing their annotation counts to frequencies for the entire genome. Annotation counts were assumed to follow a multinomial distribution, which was validated through a Monte Carlo simulation.

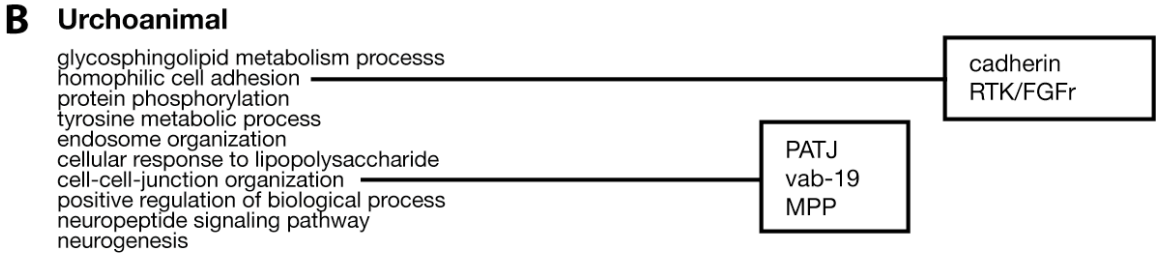
A jackknifing analysis was run to test the sensitivity of phylogenic enrichment to the species included. 10,000 trials were run, each with a random set of species. *S. rosetta* and *M. brevicollis* were included in all trials. Each of the 32 remaining species had an 80% probability of being included in any given trial. The OrthoMCL2 algorithm was rerun for each species set to generate new clusters. Annotation frequencies were re-calculated for the entire genome and the expression clusters were tested for phylogenic enrichment.

The MCL algorithm was run an addition 19 times to test the sensitivity of the results to the inflation parameter of the MCL algorithm. Values for inflation ranged from 1.1 to 3. All 34 species were included.

Figures

Figure 3.1 Holozoan gene gain and loss. A. Overall gene family number remained relatively stable with 15% turn-over from the Urholozoan to the Urmetazoan. In contrast, the *C. owczaraki* and choanoflagellates are marked by 40% gene loss. **B.** Gene gain on the stem leading to the Urchoanimal is enriched cell adhesion and junction genes. The Urmetazoan stem is enriched for developmental signaling and adhesion genes.


A
Minimal Gene Clusters
Novel Gene Clusters
Lost Gene Clusters
Urmetazoan
1,113
751
10,390
Metazoa
Urchoanimal
516
400
10,028
3,408
721
9,411
S. rosetta
Urchoanoflagellate
558
3,862
6,724
Urholozoan
517
967
9,912
2,610
1,354
7,980
M. brevicollis
1,967
4,716
7,163
C. owczarzaki


B
Urchoanimal
glycosphingolipid metabolism processs
homophilic cell adhesion
protein phosphorylation
tyrosine metabolic process
endosome organization
cellular response to lipopolysaccharide
cell-cell-junction organization
positive regulation of biological process
neuropeptide signaling pathway
neurogenesis
cadherin
RTK/FGFr
PATJ
vab-19
MPP

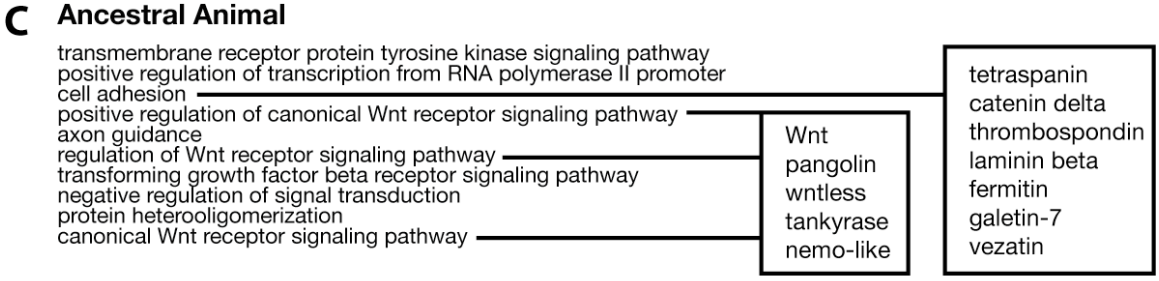

C
Ancestral Animal
transmembrane receptor protein tyrosine kinase signaling pathway
positive regulation of transcription from RNA polymerase II promoter
cell adhesion
positive regulation of canonical Wnt receptor signaling pathway
axon guidance
regulation of Wnt receptor signaling pathway
transforming growth factor beta receptor signaling pathway
negative regulation of signal transduction
protein heterooligomerization
canonical Wnt receptor signaling pathway
Wnt
pangolin
wntless
tankyrase
nemo-like
tetraspanin
catenin delta
thrombospondin
laminin beta
fermitin
galetin-7
vezatin

Figure 3.2 Gene expression correlates with cell type. A, The *S. rosetta* life cycle has at least five morphologically and behaviorally differentiated cell types: solitary cells attached to a substrate (At), solitary swimming cells (Sw), solitary fast swimming cells (Fs), linear colonies (LC), and rosette colonies (RC). **B**, Similar cell types have similar gene expression. Values at a given node are the correlation value of least correlated samples. Brackets indicate experimentally paired samples. **C**,**D**, Three methods, comparison of experimentally paired samples, comparison of groups of similar cell types, and iterative hierarchical clustering, were used to identify over-expressed in colonial cells (**C**) and attached cells (**D**). **E**, Expression, by sample, of genes identified as differential expressed in either colonial or attached cells by the union of methods.

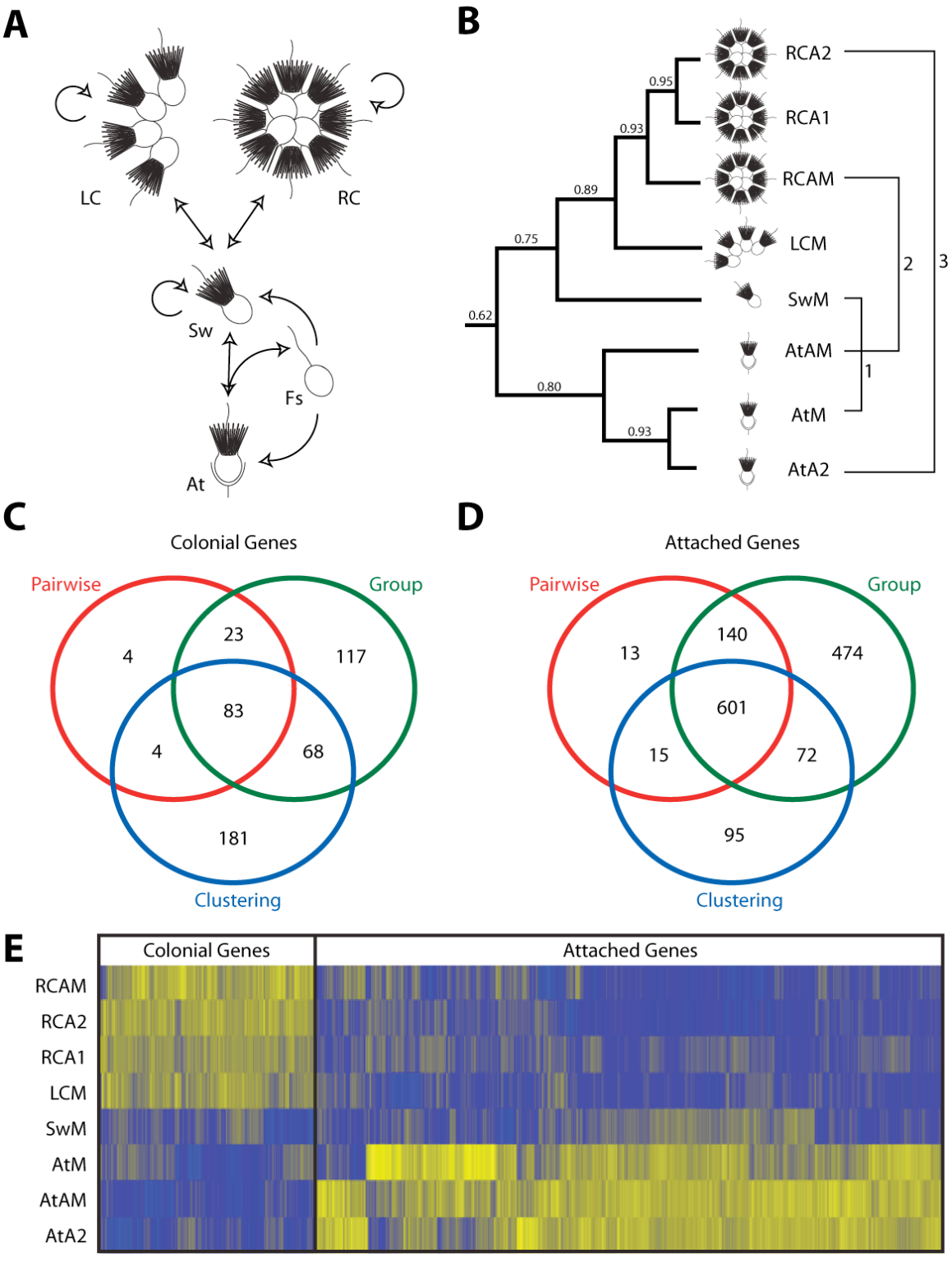

A
LC
RC
Sw
Fs
At
B
RCA2
RCA1
RCAM
LCM
SwM
AtAM
AtM
AtA2
0.95
0.93
0.89
0.75
0.62
0.80
0.93
1
2
3
C
Colonial Genes
Pairwise
Group
Clustering
4
23
117
83
4
68
181
D
Attached Genes
Pairwise
Group
Clustering
13
140
474
601
15
72
95
E
Colonial Genes
Attached Genes
RCAM
RCA2
RCA1
LCM
SwM
AtM
AtAM
AtA2

Figure 3.3 Septins are over expressed in colonial cells. A, colonial cells are attached to neighboring cells by cytoplasmic bridges trisected by electron dense bands. Photo Credit: Mark Dayel **B**, *S. rosetta* septin sequences mapped to crystallized septin filament indicating sequence similarity. **C**, the expression of septins and selected midbody genes in colonial and non-colonial cell types.

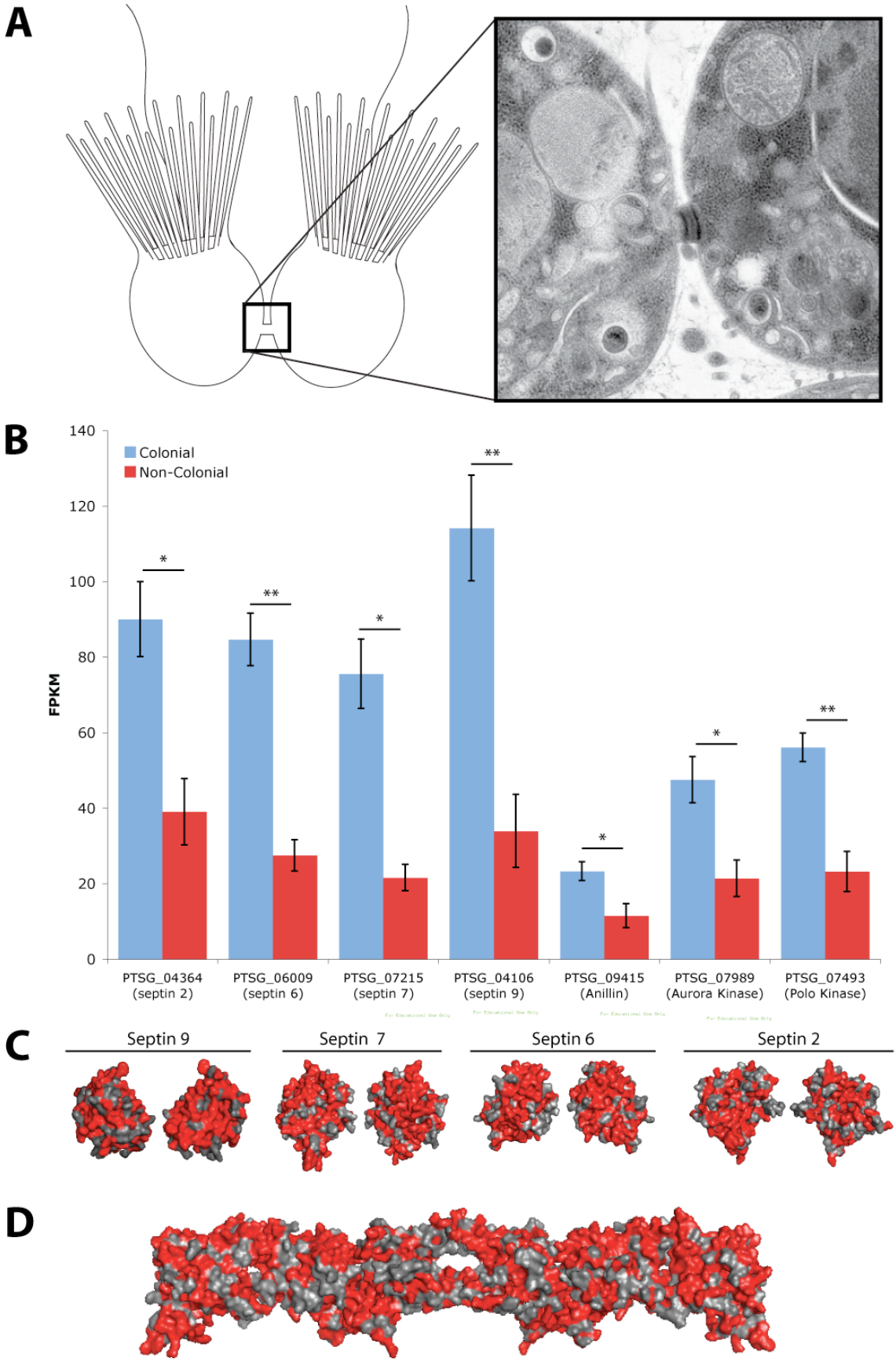

A
B
Colonial
Non-Colonial
FPKM
140
120
100
80
60
40
20
0
*
**
*
**
*
*
**
PTSG_04364 (septin 2)
PTSG_06009 (septin 6)
PTSG_07215 (septin 7)
PTSG_04106 (septin 9)
PTSG_09415 (Anillin)
PTSG_07989 (Aurora Kinase)
PTSG_07493 (Polo Kinase)
C
Septin 9
Septin 7
Septin 6
Septin 2
D

Figure 3.4 Cadherins are over expressed in attached and colonial cells. A, Choanoflagellate cadherins have been hypothesized to function in substrate attachment, prey capture, signaling and intercellular attachment. **B**, Expression of cadherins is consistent with function in substrate attachment and intercellular attachment, but there is no correlation between bacterial prey and cadherin expression. **C**, genes containing the Hedgehog signaling domain and a transmembrane domain are over expressed in solitary attached sells. Genes containing the Hedgehog signaling domain without a transmembrane domain are over expressed in colonial cells. **D**, Patched domain containing proteins are over expressed in colonial and attached cells.

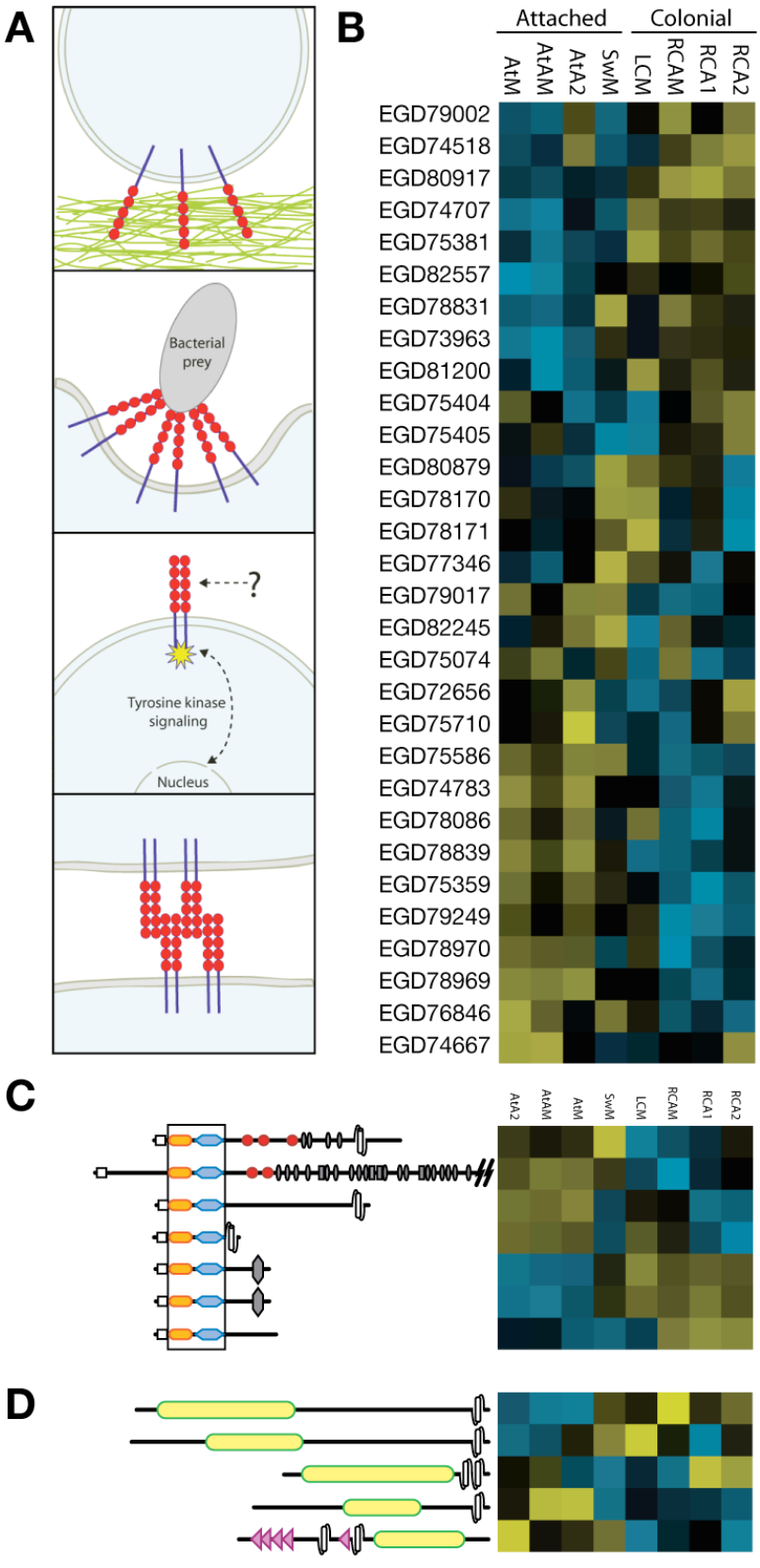

A
Bacterial prey
?
Tyrosine kinase signaling
Nucleus
B
Attached
Colonial
AtM
AtAM
AtA2
SwM
LCM
RCAM
RCA1
RCA2
EGD79002
EGD74518
EGD80917
EGD74707
EGD75381
EGD82557
EGD78831
EGD73963
EGD81200
EGD75404
EGD75405
EGD80879
EGD78170
EGD78171
EGD77346
EGD79017
EGD82245
EGD75074
EGD72656
EGD75710
EGD75586
EGD74783
EGD78086
EGD78839
EGD75359
EGD79249
EGD78970
EGD78969
EGD76846
EGD74667
C
AtA2
AtAM
AtM
SwM
LCM
RCAM
RCA1
RCA2
D

Figure 3.5. Phylogenetic origin of differentially expressed genes. A-C, The phylogenetic origin of the *S. rosetta* genome (**A)** and genes over expressed in colonial cells (**B**) or attached cells (**C**). **D,E**, Percent enrichment of genes by phylogenetic origin by cell type. *$p<0.05$, **$p<0.01$

Union

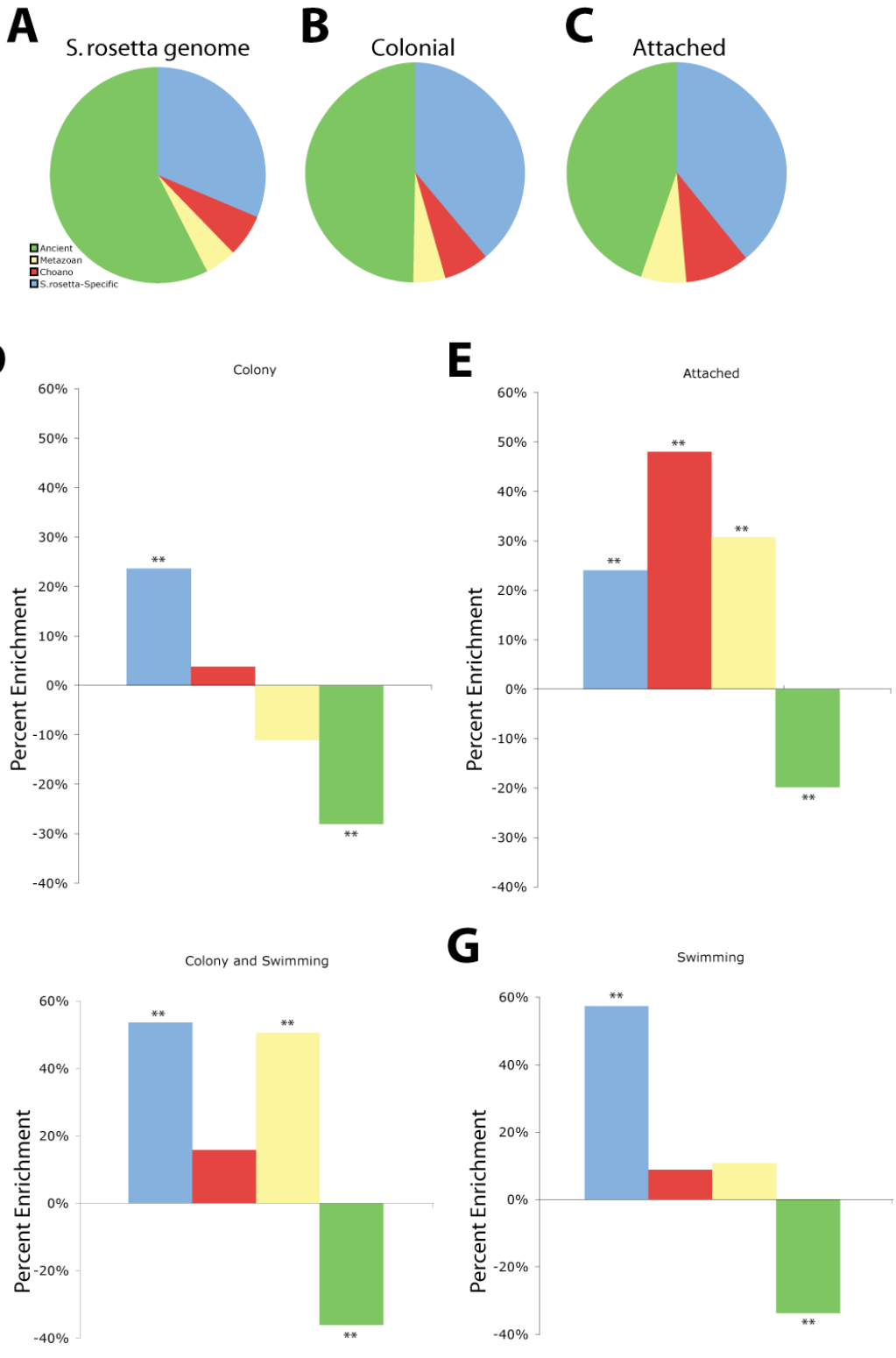

A
S. rosetta genome
Ancient
Metazoan
Choano
S.rosetta-Specific
B
Colonial
C
Attached
D
Colony
Percent Enrichment
60%
50%
40%
30%
20%
10%
0%
-10%
-20%
-30%
-40%
**
E
Attached
F
Colony and Swimming
G
Swimming

Table S3.1 *S. rosetta* and *M. brevicollis* genome statistics.

Category for Comparison	*S. rosetta*	*M. brevicollis*
Assembly Statistics		
Number of Scaffolds	154	218
Scaffold N50	1.52 Mb	1.07 Mb
% Q40	98.51	N/A
Genome		
Size	55.44 Mb	41.63 Mb
Genome coverage	33.07x	8.39x
% GC Content	56.01 %	54.89%
Genes		
Number of genes	11,629	9,171
Median gene length	3,220	2,183.00
% Genes with EST support	98%	46%
Exons		
Number of exons	100,147	69,682
Median exon length	117	124
Introns		
Intergenic	10,312	9,282
Number of introns	88,594	60,636
Mean intron length (bp)	255	174
Mean introns per gene	7.5	6.6

Table S3.2 Telomeres predicted in the *S. rosetta* genome.

Table 1. Proterospongia telomeres. (is 4kb mate pair) *matches both ends.
** matches at 300 kb from end – assembly problem?
43 found with 40 kb mate pairs (checked the top 250 matches – not all useful)
Total of 62 telomeres longer than 20kb.
And total of 19 full-length chromosomes and 24 one-ended chromosomes that need to be joined.

Super-contig	Size (kbp)	End	TTAGGG Assembled	40kb Mates	Missing kb	Sc15.3 Genes match nearby
1	2,682	5’	No	1	40	Why not? Mostly missing?
1	2,682	3’				
2	2,480	5’	No	4	10	Yes
2	2,480	3’	No	3(1)	**	Yes
3	2,356	5’	Yes	-	0	Yes
3	2,356	3’	No	-	-	Yes
4	2,372	5’				
4	2,372	3’	No	-	-	Yes
5	2,237	5’	No	(1*)	-	Yes
5	2,237	3’	No	(1*)	-	Yes
6	2,078	5’				
6	2,078	3’	No	4	10	Yes
7	2,024	5’	No	4	30	Yes
7	2,024	3’	No	-	-	Yes
8	1,951	5’	found with 16.3			
8	1,951	3’	Yes	13(4)	0	Yes
9	1,890	5’	No	9	15	Yes
9	1,890	3’	No	-	-	Yes
10	1,852	5’	No	-	-	Yes
10	1,852	3’	No	8	5	Yes
11	1,685	5’				
11	1,685	3’	No	-	-	Yes
12	1,669	5’	No	1	35	Why not? Matches 8.3??
12	1,669	3’	No	9	15	Maybe
13	1,638	5’				
13	1,638	3’	No	16	10	Yes
14	1,475	5’	No	-	-	Yes
14	1,475	3’	No	-	-	Yes
15	1,520	5’	No	-	-	Yes
15	1,520	3’	No	11	5	Yes
16	1,370	5’				
16	1,370	3’	No	3	15	Why Not? Shared with 8.5
17	1,374	5’	No	7(1)	10	Yes
17	1,374	3’	No	-	-	Yes
18	1,385	5’	No	1	5	Yes
18	1,385	3’	Yes	5	0	Yes
19	1,260	5’	No	3	30	Yes
19	1,260	3’				

20	1,278	5’	No	-	-	Yes
20	1,278	3’	No	-	-	Yes
21	1,216	5’	No	2(1)	5	Yes
21	1,216	3’				
22	1,286	5’	No	3	10	Yes
22	1,286	3’	No	3(1)	10	Yes
23	1,192	5’	Yes	9	0	Yes
23	1,192	3’				
24	1,067	5’	No	2	10	Yes
24	1,067	3’	No	10	15	Yes
25	983	5’	No	2	5	Yes
25	983	3’				
26	897	5’				
26	897	3’	No	1	5	Yes
27	927	5’	No	1	30	Yes
27	927	3’	No	-	-	Yes
28	911	5’	No	1	10	Yes
28	911	3’				
29	916	5’	Yes	7	0	Yes
29	916	3’	No	6	10	Yes
30	789	5’	No	-	-	Yes
30	789	3’				
31	796	5’	No	3	15	Yes
31	796	3’	No	3	10	Yes
32	786	5’	No	-	-	Yes
32	786	3’	No	1	20	Yes
33	740	5’	No	4	15	Yes
33	740	3’	No	1	35	Yes
34	739	5’	No	1	10	Yes
34	739	3’	No	1	35	Why not? Similar to 36.5?
35	753	5’	No	7	5	Yes
35	753	3’				
36	741	5’	No	2	30	Yes
36	741	3’				
37	725	5’	Yes	-	0	Yes
37	725	3’	No	-	-	Yes
38	740	5’				
38	740	3’	No	-	-	Yes
39	427	5’	No	-	-	Yes
40	447	3’	No	2	5	Yes
41	388	3’	No	-	-	Yes
42	358	3’	No	2	35	Yes
43	321	5’	No	-	-	Yes
44	82	5’	No	-	-	Yes
45	36	3’	No	9	5	Why not? Unique!
46	16	5’	No	-	-	Yes

65	7	5’	No	-	-	Yes
101	5	5’	No	-	-	Yes
132	4	3’	No	-	-	Yes
135	5	3’	No	-	-	Yes

Table S3.3 Genomes used for comparative genomics

Organism	Genome Source
H. sapiens	ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/
C. intestinalis	ftp://ftp.ensembl.org/pub/current/fasta/ciona_intestinalis/pep/
S. purpuratus	ftp://ftp.ncbi.nih.gov/genomes/Strongylocentrotus_purpuratus/protein/
D. melanogaster	ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.33_FB2011_01/
C. elegans	ftp://ftp.wormbase.org/pub/wormbase/genomes/c_elegans/sequences/protein/
L. gigantea	http://genome.jgi-psf.org/Lotgi1/Lotgi1.download.ftp.html
H. magnipapillata	ftp://ftp.ncbi.nih.gov/genomes/Hydra_magnipapillata/protein/
N. vectensis	http://genome.jgi-psf.org/Nemve1/Nemve1.download.ftp.html
T. adhaerens	http://genome.jgi-psf.org/Triad1/Triad1.download.ftp.html
A. gambiae	ftp://ftp.ncbi.nih.gov/genbank/genomes/Anopheles_gambiae
A. queenslandica	ftp://ftp.jgi-psf.org/pub/JGI_data/Amphimedon_queenslandica/annotation/
M. brevicollis	http://genome.jgi-psf.org/Monbr1/Monbr1.download.ftp.html
S. rosetta	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
C. owczarzaki	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
S. punctatus	http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiDownloads.html
R. oryzae	http://www.broadinstitute.org/annotation/genome/rhizopus_oryzae/MultiDownloads.html
C. cinereus	http://www.broadinstitute.org/annotation/genome/coprinus_cinereus/MultiDownloads.html
C. neoformans	ftp://ftp.ncbi.nih.gov/genomes/Fungi/Cryptococcus_neoformans_var_JEC21_uid10698/
N. crassa	http://www.broadinstitute.org/annotation/genome/neurospora/MultiDownloads.html
S. cerevisae	http://downloads.yeastgenome.org/sequence/genomic_sequence/orf_protein/
S. pombe	http://www.broadinstitute.org/annotation/genome/schizosaccharomyces_group/MultiDownloads.html
A. macrogynus	http://www.broadinstitute.org/annotation/genome/multicellularity_project.3/MultiDownloads.html
B. dendrobatidis	http://www.broadinstitute.org/annotation/genome/batrachochytrium_dendrobatidis/MultiDownloads.html
A. thaliana	http://www.plantgdb.org/XGDB/phplib/download.php?GDB=At
C. reinhardtii	http://genome.jgi-psf.org/Chlre4/Chlre4.download.ftp.html
V. carteri	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v6.0/Vcarteri/annotation/
E. siliculosus	https://bioinformatics.psb.ugent.be/gdb/ectocarpus/
P. patens	ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v6.0/Ppatens/annotation/
T. pseudonana	http://genome.jgi-psf.org/Thaps3/Thaps3.download.ftp.html
E. huxleyi	http://genome.jgi-psf.org/Emihu1/Emihu1.download.ftp.html
T. thermophila	ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_thermophila/annotation_dbs/final_release_oct2008/
P. tetraurelia	http://www.genoscope.cns.fr/externe/Download/Projets/Projet_FN/data/annotation/
N. gruberi	http://genome.jgi-psf.org/Naegr1/Naegr1.download.ftp.html
D. discoideum	http://dictybase.org/Downloads/

Table S3.4 *S. rosetta* kinome. The *S. rosetta* kinome contains 469 kinases. 93 of these are tyrosine kinases (TKs) and 376 are Serine / Threonine (Ser/Thr) kinases. The Ser/Thr are well preserved between *M. brevicollis* and *S. rosetta*. In contrast, the TKs are quickly evolving with divergent sequences and large numbers of gains and losses.

ID	Group	Family	Subfamily
PTSG_08832T0	AGC	Akt	
PTSG_09660T0	AGC	Akt	
PTSG_11435T0	AGC	Akt	
PTSG_12484T0	AGC	Akt	
PTSG_11108T0	AGC	DMPK	CRIK
PTSG_11974T0	AGC	DMPK	GEK
PTSG_12757T0	AGC	DMPK	ROCK
PTSG_08590T0	AGC	GRK	BARK
PTSG_00259T0	AGC	MAST	MAST
PTSG_05215T0	AGC	MAST	MAST
PTSG_04961T0	AGC	NDR	
PTSG_09122T0	AGC	NDR	
PTSG_03674T0	AGC	PDK1	
PTSG_04201T0	AGC	PKA	
PTSG_03113T0	AGC	PKA	
PTSG_10123T0	AGC	PKC	Alpha
PTSG_06585T0	AGC	PKC	Alpha
PTSG_09611T0	AGC	PKC	Alpha
PTSG_10688T0	AGC	PKC	Alpha
PTSG_11451T0	AGC	PKC	Alpha
PTSG_01655T0	AGC	PKC	Eta
PTSG_11595T0	AGC	PKC	Eta
PTSG_11129T0	AGC	PKC	Iota
PTSG_10105T0	AGC	PKG	
PTSG_09011T0	AGC	RSK	MSK
PTSG_12221T0	AGC	RSK	p70
PTSG_02853T0	AGC	RSK	RSK
PTSG_04567T0	AGC	RSK	RSK
PTSG_13231T0	AGC	RSK	RSK
PTSG_04602T0	AGC	SGK	
PTSG_06704T0	AGC		
PTSG_01310T0	Atypical	ABC1	ABC1-B
PTSG_00116T0	Atypical	ABC1	ABC1-B
PTSG_06100T0	Atypical	ABC1	ABC1-B
PTSG_10756T0	Atypical	ABC1	ABC1-B

PTSG_08305T0	Atypical	ABC1	ABC1-C
PTSG_02098T0	Atypical	Alpha	ChaK
PTSG_03801T0	Atypical	Alpha	VWL
PTSG_05057T0	Atypical	Brd	
PTSG_13143T0	Atypical	Brd	
PTSG_01725T0	Atypical	HisK	
PTSG_02738T0	Atypical	HisK	
PTSG_08798T0	Atypical	HisK	
PTSG_00063T0	Atypical	PDHK	PDHK
PTSG_00845T0	Atypical	PDHK	PDHK
PTSG_08628T0	Atypical	PDHK	PDHK
PTSG_11608T0	Atypical	PDHK	PDHK
PTSG_12720T0	Atypical	PIKK	ATM
PTSG_00446T0	Atypical	PIKK	ATR
PTSG_05114T0	Atypical	PIKK	ATR
PTSG_09301T0	Atypical	PIKK	ATR
PTSG_11029T0	Atypical	PIKK	ATR
PTSG_11835T0	Atypical	PIKK	ATR
PTSG_12633T0	Atypical	PIKK	ATR
PTSG_12885T0	Atypical	PIKK	ATR
PTSG_11836T0	Atypical	PIKK	DNAPK
PTSG_02075T0	Atypical	PIKK	FRAP
PTSG_11127T0	Atypical	PIKK	FRAP
PTSG_06227T0	Atypical	PIKK	SMG1
PTSG_09381T0	Atypical	PIKK	
PTSG_03292T0	Atypical	RIO	RIO1
PTSG_04684T0	Atypical	RIO	RIO2
PTSG_10152T0	Atypical	RIO	RIO3
PTSG_07849T0	CAMK	CAMK-Unique	
PTSG_08004T0	CAMK	CAMK-Unique	
PTSG_08939T0	CAMK	CAMK-Unique	
PTSG_00163T0	CAMK	CAMK1	
PTSG_09430T0	CAMK	CAMK1	
PTSG_10090T0	CAMK	CAMK2	
PTSG_01778T0	CAMK	CAMKL	AMPK
PTSG_04405T0	CAMK	CAMKL	AMPK
PTSG_05194T0	CAMK	CAMKL	AMPK
PTSG_07354T0	CAMK	CAMKL	AMPK

PTSG_09966T0	CAMK	CAMKL	AMPK
PTSG_12391T0	CAMK	CAMKL	AMPK
PTSG_00462T0	CAMK	CAMKL	CHK1
PTSG_07039T0	CAMK	CAMKL	kin4
PTSG_08374T0	CAMK	CAMKL	LKB
PTSG_02343T0	CAMK	CAMKL	MARK
PTSG_10175T0	CAMK	CAMKL	MARK
PTSG_11471T0	CAMK	CAMKL	NIM1
PTSG_07878T0	CAMK	CAMKL	PASK
PTSG_12259T0	CAMK	CAMKL	QIK
PTSG_02285T0	CAMK	CAMKL	SNRK
PTSG_07665T0	CAMK	DCAMKL	
PTSG_10114T0	CAMK	DCAMKL	
PTSG_03159T0	CAMK	MAPKAPK	MAPKAPK
PTSG_03766T0	CAMK	MAPKAPK	MAPKAPK
PTSG_07730T0	CAMK	MAPKAPK	MAPKAPK
PTSG_13004T0	CAMK	MAPKAPK	MNK
PTSG_08777T0	CAMK	RAD53	
PTSG_04457T0	CAMK	TSSK	
PTSG_12416T0	CAMK		
PTSG_07885T0	CK1	CK1	CK1-D
PTSG_06973T0	CK1	CK1	CK1-D
PTSG_11580T0	CK1	CK1	CK1-G
PTSG_05555T0	CK1	TTBK	
PTSG_03252T0	CK1	VRK	
PTSG_01028T0	CMGC	CDK	CCRK
PTSG_03213T0	CMGC	CDK	CDC2
PTSG_03214T0	CMGC	CDK	CDC2
PTSG_03867T0	CMGC	CDK	CDC2
PTSG_08420T0	CMGC	CDK	CDC2
PTSG_00030T0	CMGC	CDK	CDC5
PTSG_08323T0	CMGC	CDK	CDK10
PTSG_10371T0	CMGC	CDK	CDK7
PTSG_00138T0	CMGC	CDK	CDK8
PTSG_00805T0	CMGC	CDK	CRK7
PTSG_07137T0	CMGC	CDK	CRK7
PTSG_10173T0	CMGC	CDK	CRK7
PTSG_00747T0	CMGC	CDK	PCTAIRE
PTSG_06662T0	CMGC	CDK	PITSLRE
PTSG_05370T0	CMGC	CDKL	

PTSG_06706T0	CMGC	CDKL	
PTSG_04001T0	CMGC	CK2	
PTSG_01985T0	CMGC	CLK	
PTSG_12163T0	CMGC	DYRK	DYRK1
PTSG_08399T0	CMGC	DYRK	DYRK2
PTSG_12767T0	CMGC	DYRK	DYRK2
PTSG_00232T0	CMGC	DYRK	PRP4
PTSG_05204T0	CMGC	DYRK	PRP4
PTSG_07485T0	CMGC	GSK	
PTSG_09505T0	CMGC	GSK	
PTSG_07429T0	CMGC	MAPK	ERK1
PTSG_11082T0	CMGC	MAPK	ERK1
PTSG_08125T0	CMGC	MAPK	ERK5
PTSG_02788T0	CMGC	MAPK	ERK7
PTSG_11071T0	CMGC	MAPK	ERK7
PTSG_03419T0	CMGC	MAPK	p38
PTSG_06575T0	CMGC	MAPK	p38
PTSG_13049T0	CMGC	MAPK	p38
PTSG_09237T0	CMGC	RCK	MAK
PTSG_01609T0	CMGC	RCK	MOK
PTSG_11567T0	CMGC	SRPK	
PTSG_07989T0	Other	Aur	
PTSG_09352T0	Other	BUB	
PTSG_07309T0	Other	Bud32	
PTSG_06601T0	Other	CAMKK	Meta
PTSG_00392T0	Other	CDC7	
PTSG_06705T0	Other	Haspin	
PTSG_12653T0	Other	IRE	
PTSG_05619T0	Other	NAK	BIKE
PTSG_01962T0	Other	NAK	GAK
PTSG_02001T0	Other	NAK	MPSK
PTSG_00355T0	Other	NEK	NEK1
PTSG_04657T0	Other	NEK	NEK1
PTSG_05317T0	Other	NEK	NEK1
PTSG_05717T0	Other	NEK	NEK1
PTSG_06021T0	Other	NEK	NEK1
PTSG_09423T0	Other	NEK	NEK1
PTSG_10545T0	Other	NEK	NEK1
PTSG_11844T0	Other	NEK	NEK1
PTSG_00129T0	Other	NEK	NEK10

PTSG_09647T0	Other	NEK	NEK2
PTSG_12012T0	Other	NEK	NEK4
PTSG_07490T0	Other	NEK	NEK6
PTSG_01201T0	Other	NEK	NEK8
PTSG_02940T0	Other	NEK	NEK8
PTSG_08458T0	Other	NEK	NEK8
PTSG_03788T0	Other	NEK	
PTSG_12064T0	Other	NEK	
PTSG_12377T0	Other	NEK	
PTSG_12714T0	Other	NFK1	
PTSG_08264T0	Other	NRBP	
PTSG_01828T0	Other	Other-Unique	
PTSG_04254T0	Other	Other-Unique	
PTSG_07158T0	Other	Other-Unique	
PTSG_10216T0	Other	Other-Unique	
PTSG_10496T0	Other	Other-Unique	
PTSG_10419T0	Other	PEK	GCN2
PTSG_07493T0	Other	PLK	PLK1
PTSG_09382T0	Other	PLK	SAK
PTSG_02395T0	Other	SCY1	
PTSG_04183T0	Other	SCY1	
PTSG_06380T0	Other	SgK071	
PTSG_00066T0	Other	Slob	
PTSG_08642T0	Other	Slob	
PTSG_06925T0	Other	TBCK	
PTSG_04707T0	Other	TLK	
PTSG_01122T0	Other	TTK	
PTSG_11852T0	Other	ULK	Fused
PTSG_00458T0	Other	ULK	ULK
PTSG_01376T0	Other	ULK	ULK
PTSG_07407T0	Other	ULK	ULK
PTSG_07822T0	Other	ULK	ULK
PTSG_11321T0	Other	VPS15	
PTSG_01499T0	Other	WEE	
PTSG_03206T0	Other	WEE	
PTSG_05844T0	Other	WEE	

PTSG_07404T0	Other	WNK	
PTSG_00142T0	RGC	RGC	
PTSG_05043T0	RGC	RGC	
PTSG_05045T0	RGC	RGC	
PTSG_05341T0	RGC	RGC	
PTSG_06107T0	RGC	RGC	
PTSG_08165T0	RGC	RGC	
PTSG_08175T0	RGC	RGC	
PTSG_09225T0	RGC	RGC	
PTSG_10344T0	RGC	RGC	
PTSG_10440T0	RGC	RGC	
PTSG_10870T0	RGC	RGC	
PTSG_12646T0	RGC	RGC	
PTSG_12647T0	RGC	RGC	
PTSG_12648T0	RGC	RGC	
PTSG_11019T0	STE	STE11	ASK
PTSG_00201T0	STE	STE11	MEKK2
PTSG_05701T0	STE	STE11	MEKK2
PTSG_10004T0	STE	STE11	MEKK2
PTSG_13011T0	STE	STE11	STE11-Unclassified
PTSG_12589T0	STE	STE20	FRAY
PTSG_11972T0	STE	STE20	KHS
PTSG_10999T0	STE	STE20	MSN
PTSG_10780T0	STE	STE20	MST
PTSG_11624T0	STE	STE20	NinaC
PTSG_01477T0	STE	STE20	PAKA
PTSG_07364T0	STE	STE20	SLK
PTSG_09125T0	STE	STE20	SLK
PTSG_09127T0	STE	STE20	SLK
PTSG_09185T0	STE	STE20	SLK
PTSG_10162T0	STE	STE20	SLK
PTSG_08880T0	STE	STE20	STE20-Unclassified
PTSG_03289T0	STE	STE20	STLK
PTSG_03937T0	STE	STE20	TAO
PTSG_01120T0	STE	STE20	YSK
PTSG_09942T0	STE	STE20	YSK
PTSG_04378T0	STE	STE7	MEK1
PTSG_04559T0	STE	STE7	MEK1
PTSG_09379T0	STE	STE7	MEK4
PTSG_04499T0	STE	STE7	MEK5

PTSG_09802T0	STE	STE7	STE7-Unclassified
PTSG_12707T0	STE	STE7	STE7-Unclassified
PTSG_00346T0	TK	CTKA	
PTSG_05383T0	TK	CTKB	
PTSG_05589T0	TK	CTKB	
PTSG_06770T0	TK	Eph	
PTSG_07290T0	TK	FAK	
PTSG_05525T0	TK	FYTK	
PTSG_07629T0	TK	HMTK	
PTSG_03356T0	TK	HMTK	
PTSG_01992T0	TK	HMTK	
PTSG_02132T0	TK	HMTK	
PTSG_02847T0	TK	HMTK	
PTSG_03111T0	TK	HMTK	
PTSG_03234T0	TK	HMTK	
PTSG_03439T0	TK	HMTK	
PTSG_07139T0	TK	HMTK	
PTSG_07140T0	TK	HMTK	
PTSG_07725T0	TK	HMTK	
PTSG_08747T0	TK	HMTK	
PTSG_09926T0	TK	HMTK	
PTSG_10301T0	TK	HMTK	
PTSG_12301T0	TK	HMTK	
PTSG_11547T0	TK	RTKB	
PTSG_01974T0	TK	RTKC	
PTSG_02760T0	TK	RTKC	
PTSG_04584T0	TK	RTKC	
PTSG_05128T0	TK	RTKC	
PTSG_05491T0	TK	RTKC	
PTSG_10477T0	TK	RTKC	
PTSG_12191T0	TK	RTKC	
PTSG_13042T0	TK	RTKC	
PTSG_07959T0	TK	RTKD	
PTSG_01333T0	TK	RTKE	
PTSG_07501T0	TK	RTKH	
PTSG_04246T0	TK	RTKJ	
PTSG_04263T0	TK	RTKJ	
PTSG_00058T0	TK	RTKN	
PTSG_00092T0	TK	RTKN	
PTSG_04221T0	TK	RTKN	

PTSG_04223T0	TK	RTKN	
PTSG_05453T0	TK	RTKN	
PTSG_10316T0	TK	RTKN	
PTSG_00498T0	TK	RTKO	
PTSG_02705T0	TK	RTKO	
PTSG_12775T0	TK	RTKO	
PTSG_00744T0	TK	RTKP	
PTSG_04469T0	TK	RTKP	
PTSG_07743T0	TK	RTKP	
PTSG_05332T0	TK	RTKQ	
PTSG_11342T0	TK	RTKQ	
PTSG_08572T0	TK	SFK-Abl	
PTSG_02633T0	TK	SFK-Abl	
PTSG_06415T0	TK	SFK-Abl	
PTSG_08192T0	TK	SFK-Csk	
PTSG_11483T0	TK	SFK-Src	
PTSG_04679T0	TK	SFK-Src	
PTSG_10481T0	TK	SFK-Src	
PTSG_07054T0	TK	SFK-Src	
PTSG_11563T0	TK	SFK-Tec	
PTSG_10626T0	TK	Syk	
PTSG_01646T0	TK	UTK	UTKB
PTSG_04570T0	TK	UTK	UTKB
PTSG_10109T0	TK	UTK	UTKC
PTSG_11314T0	TK	UTK	UTKC
PTSG_08430T0	TK	UTK	UTKE
PTSG_11059T0	TK	UTK	UTKG
PTSG_01193T0	TK	UTK	UTKH
PTSG_07321T0	TK	UTK	UTKI
PTSG_07465T0	TK	UTK	UTKJ
PTSG_10793T1	TK	UTK	UTKK
PTSG_13074T0	TK	UTK	UTKL
PTSG_12045T0	TK	UTK	UTKM
PTSG_00723T0	TK	UTK	
PTSG_00753T0	TK	UTK	
PTSG_01744T0	TK	UTK	
PTSG_04631T0	TK	UTK	
PTSG_05587T0	TK	UTK	
PTSG_05722T0	TK	UTK	
PTSG_06501T0	TK	UTK	

PTSG_07220T0	TK	UTK
PTSG_07484T0	TK	UTK
PTSG_07744T0	TK	UTK
PTSG_07839T0	TK	UTK
PTSG_08416T0	TK	UTK
PTSG_08663T0	TK	UTK
PTSG_09784T0	TK	UTK
PTSG_09785T0	TK	UTK
PTSG_10873T0	TK	UTK
PTSG_11064T0	TK	UTK
PTSG_12437T0	TK	UTK
PTSG_12766T0	TK	UTK
PTSG_13138T0	TK	UTK
PTSG_07550T0	TK	UTK
PTSG_08210T0	TKL	A
PTSG_09158T0	TKL	A
PTSG_11443T0	TKL	A
PTSG_00070T0	TKL	A
PTSG_00107T0	TKL	A
PTSG_00233T0	TKL	A
PTSG_00320T0	TKL	A
PTSG_01420T0	TKL	A
PTSG_01856T0	TKL	A
PTSG_02595T0	TKL	A
PTSG_02604T0	TKL	A
PTSG_02677T0	TKL	A
PTSG_02796T0	TKL	A
PTSG_02831T0	TKL	A
PTSG_02955T0	TKL	A
PTSG_03474T0	TKL	A
PTSG_04279T0	TKL	A
PTSG_05182T0	TKL	A
PTSG_05193T0	TKL	A
PTSG_05302T0	TKL	A
PTSG_05412T0	TKL	A
PTSG_05421T0	TKL	A
PTSG_05988T0	TKL	A
PTSG_06012T0	TKL	A
PTSG_06013T0	TKL	A
PTSG_06029T0	TKL	A

PTSG_06052T0	TKL	A
PTSG_06775T0	TKL	A
PTSG_06961T0	TKL	A
PTSG_07401T0	TKL	A
PTSG_07808T0	TKL	A
PTSG_07821T0	TKL	A
PTSG_07902T0	TKL	A
PTSG_08146T0	TKL	A
PTSG_08245T0	TKL	A
PTSG_08405T0	TKL	A
PTSG_08407T0	TKL	A
PTSG_08428T0	TKL	A
PTSG_08601T0	TKL	A
PTSG_08692T0	TKL	A
PTSG_08736T0	TKL	A
PTSG_08820T0	TKL	A
PTSG_08998T0	TKL	A
PTSG_09035T0	TKL	A
PTSG_09124T0	TKL	A
PTSG_09126T0	TKL	A
PTSG_09128T0	TKL	A
PTSG_09145T0	TKL	A
PTSG_09230T0	TKL	A
PTSG_09231T0	TKL	A
PTSG_09257T0	TKL	A
PTSG_09258T0	TKL	A
PTSG_09261T0	TKL	A
PTSG_09263T0	TKL	A
PTSG_09278T0	TKL	A
PTSG_09726T0	TKL	A
PTSG_10000T0	TKL	A
PTSG_10027T0	TKL	A
PTSG_10042T0	TKL	A
PTSG_10089T0	TKL	A
PTSG_10163T0	TKL	A
PTSG_10189T0	TKL	A
PTSG_10193T0	TKL	A
PTSG_10227T0	TKL	A
PTSG_10236T0	TKL	A
PTSG_10239T0	TKL	A

PTSG_10248T0	TKL	A
PTSG_10258T0	TKL	A
PTSG_10262T0	TKL	A
PTSG_10353T0	TKL	A
PTSG_10383T0	TKL	A
PTSG_10385T0	TKL	A
PTSG_10390T0	TKL	A
PTSG_10604T0	TKL	A
PTSG_10894T0	TKL	A
PTSG_10925T0	TKL	A
PTSG_11141T0	TKL	A
PTSG_11146T0	TKL	A
PTSG_11265T0	TKL	A
PTSG_11439T0	TKL	A
PTSG_11906T0	TKL	A
PTSG_11911T0	TKL	A
PTSG_12661T0	TKL	A
PTSG_12860T0	TKL	A
PTSG_12981T0	TKL	A
PTSG_00871T0	TKL	A
PTSG_05564T0	TKL	B
PTSG_02897T0	TKL	B
PTSG_02993T0	TKL	B
PTSG_02218T0	TKL	B
PTSG_02727T0	TKL	B
PTSG_03870T0	TKL	B
PTSG_04430T0	TKL	B
PTSG_04444T0	TKL	B
PTSG_04482T0	TKL	B
PTSG_07307T0	TKL	B
PTSG_07945T0	TKL	B
PTSG_07954T0	TKL	B
PTSG_08208T0	TKL	B
PTSG_08613T0	TKL	B
PTSG_08621T0	TKL	B
PTSG_09988T0	TKL	B
PTSG_11075T0	TKL	B
PTSG_11122T0	TKL	B
PTSG_12821T0	TKL	B
PTSG_00302T0	TKL	Fer

PTSG_03858T0	TKL	LISK	LIMK
PTSG_07236T0	TKL	LISK	LIMK
PTSG_08837T0	TKL	LISK	TESK
PTSG_04723T0	TKL	LRRK	
PTSG_10622T0	TKL	MLK	HH498
PTSG_04629T0	TKL	MLK	LZK
PTSG_01588T0	TKL	MLK	MLK
PTSG_08045T0	TKL	MLK	MLK
PTSG_12761T0	TKL	MLK	RAF
PTSG_10603T0	TKL	MLK	Shk
PTSG_03936T0	TKL	MLK	TAK1
PTSG_00902T0	TKL	MLK	ZAK
PTSG_02937T0	TKL	MLK	
PTSG_12445T0	TKL	RTKC	Shk
PTSG_00098T0	TKL	TKL-Unique	
PTSG_00743T0	TKL	TKL-Unique	
PTSG_04616T0	TKL	TKL-Unique	
PTSG_04724T0	TKL	TKL-Unique	
PTSG_06299T0	TKL	TKL-Unique	
PTSG_08355T0	TKL	TKL-Unique	
PTSG_10417T0	TKL	TKL-Unique	
PTSG_11993T0	TKL	TKL-Unique	
PTSG_02857T0	TKL	TKL-Unique	
PTSG_07786T0	TKL	TKL-Unique	
PTSG_02019T0	TKL	TKL-Unique	
PTSG_07470T0	TKL	TKL-Unique	
PTSG_04927T0	TKL		
PTSG_08085T0	TKL		
PTSG_02276T0	TKL		
PTSG_03128T0	TKL		
PTSG_04241T0	TKL		
PTSG_07038T0	TKL		
PTSG_10165T0	TKL		
PTSG_13073T0	TKL		
PTSG_01401T0	TKL		
PTSG_08656T0	TKL		
PTSG_10914T0	TKL		

Table S3.5 *S. rosetta* tyrosine kinases. The *S. rosetta* tyrosine kinases were separated into 30 families based on kinase domain sequence similarity, accessory domains and phylogenic trees. Below is a table of the counts of kinases in each family.

Family	Receptor?	Metazoan?	S. rosetta	Monosiga	In Both
Abl	N	Y	3	3	3
Csk	N	Y	1	1	1
CTKA	N	N	1	2	1
CTKB	N	N	2	2	2
FAK	N	Y	1	0	0
FYTK	N	N	1	2	1
HMTK	N	N	15	15	15
Src	N	Y	4	4	4
Syk	N	Y	1	1	1
Tec	N	Y	1	1	1
Eph	Y	Y	1	2	1
FGTK	Y	N	0	14	0
LRTK	Y	N	0	5	0
RTKA	Y	N	0	9	0
RTKB	Y	N	1	9	1
RTKC	Y	N	8	8	2
RTKD	Y	N	1	4	1
RTKE	Y	N	1	6	1
RTKF	Y	N	0	3	0
RTKG	Y	N	0	2	0
RTKH	Y	N	1	2	1
RTKJ	Y	N	2	2	1
RTKK	Y	N	0	2	0
RTKL	Y	N	0	4	0
RTKM	Y	N	0	2	0
RTKN	Y	N	6	1	1
RTKO	Y	N	3	0	0
RTKP	Y	N	3	0	0
RTKQ	Y	N	2	0	0
UTK			34	27	10
Total:			**93**	**133**	**48**

Figure S3.1: Transcriptional profiling experimental design. A. Each cell type from the *S. rosetta* life history sampled was grown with different feeder bacteria. Cell types are listed on the left and bacterial prey on the right. Experimental samples are labeled with cell type and bacterial prey. To identify gene over expressed in cell types we compared samples from the same conditions (pairwise) and by cell type (grouping) **C.** Pairwise: To control of environmental variation, samples grown in the same conditions were compared to identify differentially expressed gene. **D.** Grouping: To increase statistical power we grouped samples of similar cell types

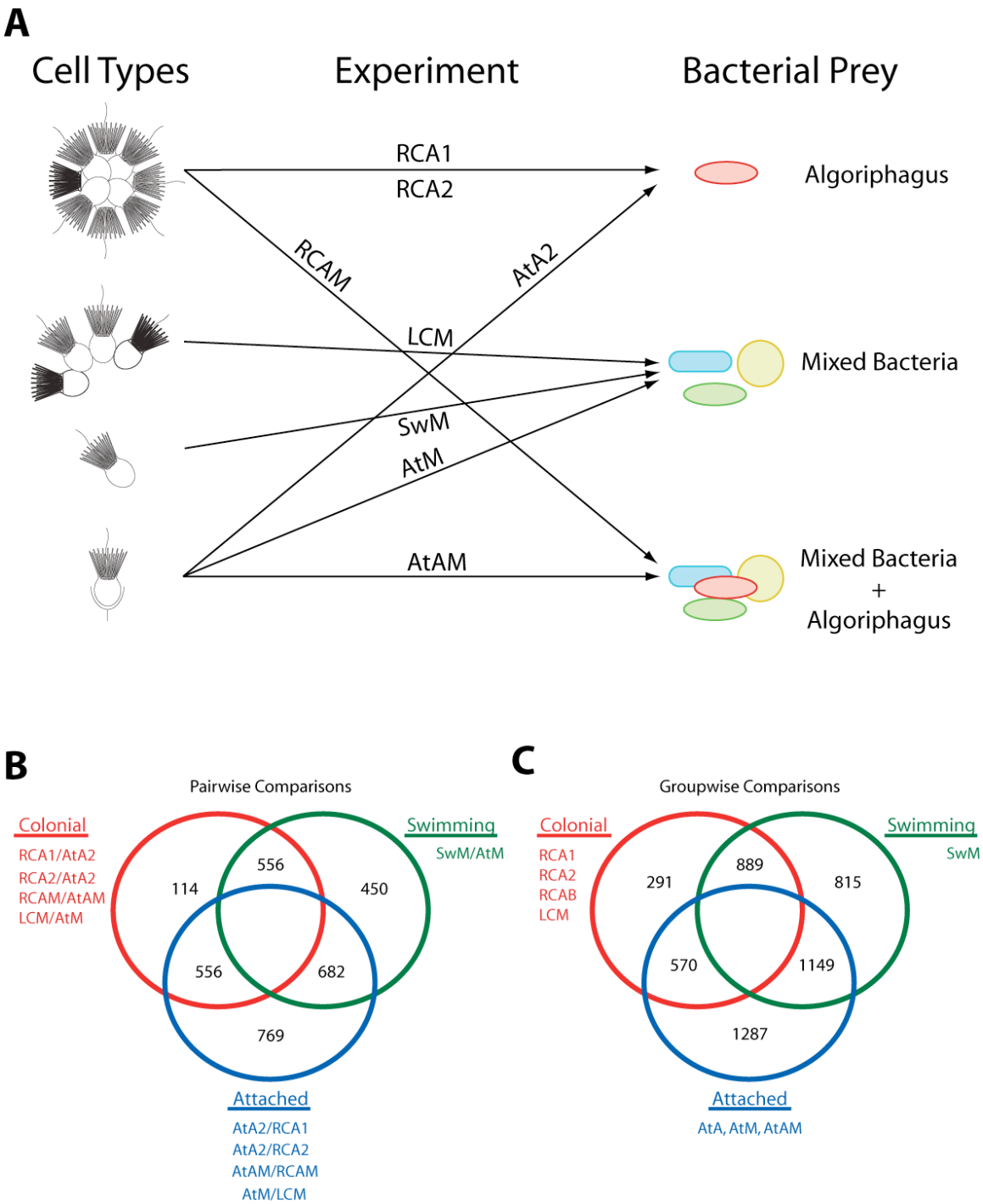

A
Cell Types
Experiment
Bacterial Prey
RCA1
RCA2
RCAM
AtA2
LCM
SwM
AtM
AtAM
Algoriphagus
Mixed Bacteria
Mixed Bacteria
+
Algoriphagus
B
Pairwise Comparisons
Colonial
RCA1/AtA2
RCA2/AtA2
RCAM/AtAM
LCM/AtM
Swimming
SwM/AtM
Attached
AtA2/RCA1
AtA2/RCA2
AtAM/RCAM
AtM/LCM
114
556
450
556
682
769
C
Groupwise Comparisons
Colonial
RCA1
RCA2
RCAB
LCM
Swimming
SwM
Attached
AtA, AtM, AtAM
291
889
815
570
1149
1287

Figure S3.2: Differentially expressed identified by iterative clustering. FPKM values were log2(FPKM) transformed, quantile normalize, and filtered requiring Max(log2(FPKM)) -Min(log2(FPKM))>2. The filtered genes set clustered hierarchically with gplots installed under R v2.12 and sub-clusters were manually selected and re-clustered (supplement Figure 2). Sub-sampling and re-clustering was preformed 4 times resulting in the final gene set presented in Figure 2 C-E.

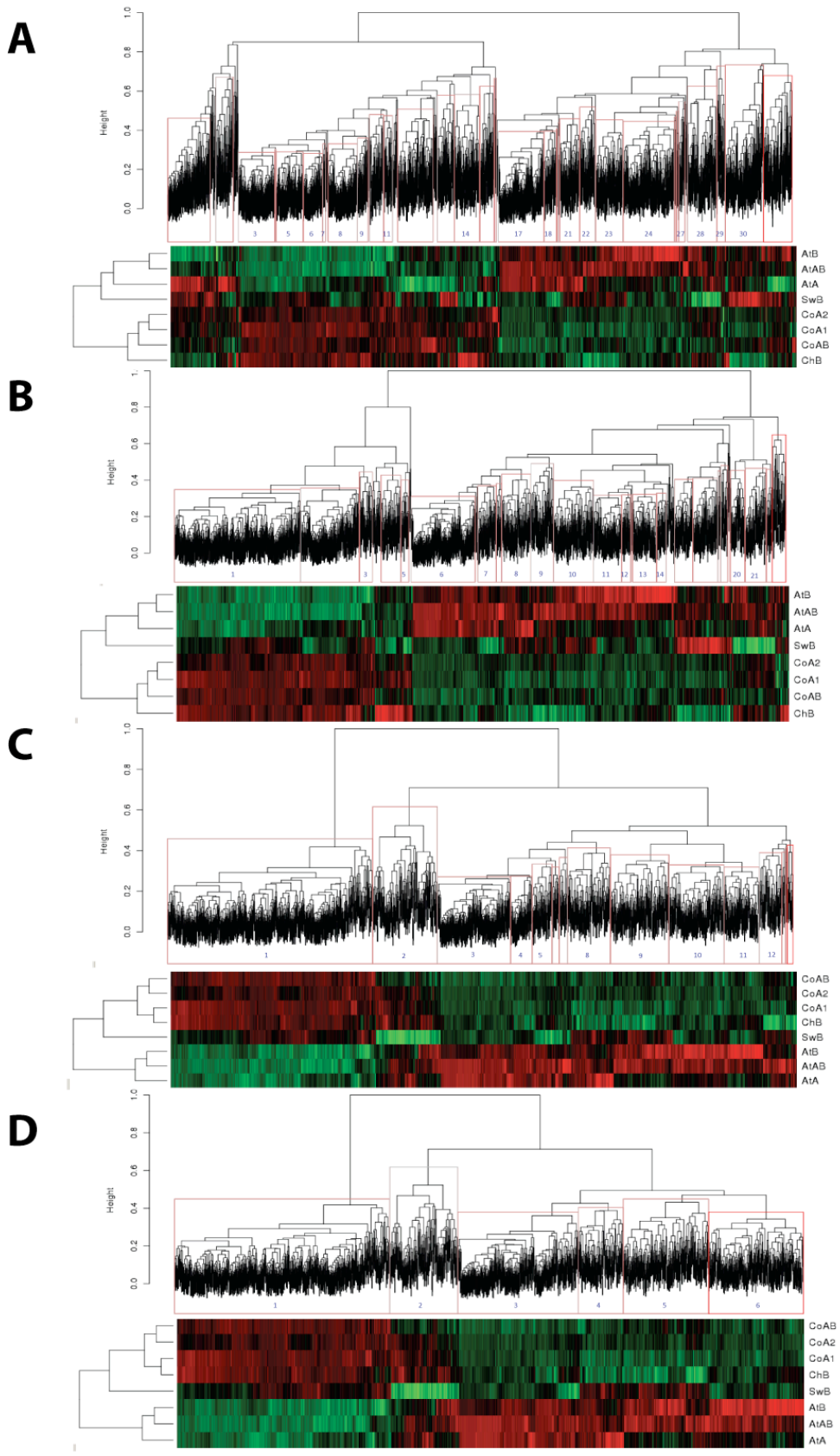

A
Height
AtB
AtAB
AtA
SwB
CoA2
CoA1
CoAB
ChB
B
Height
AtB
AtAB
AtA
SwB
CoA2
CoA1
CoAB
ChB
C
Height
CoAB
CoA2
CoA1
ChB
SwB
AtB
AtAB
AtA
D
Height
CoAB
CoA2
CoA1
ChB
SwB
AtB
AtAB
AtA

Figure S3.3 Characteristic septin organization. A typical septin containing the conserved GTP_CDC binding domain with three conserved GTP binding motifs: G1, G3 & G4. The N-terminal polybasic region is shown in black and the septin unique element is indicated in grey. The predicted coiled-coil domain, which is present in many septins, is indicated at the C-terminus.

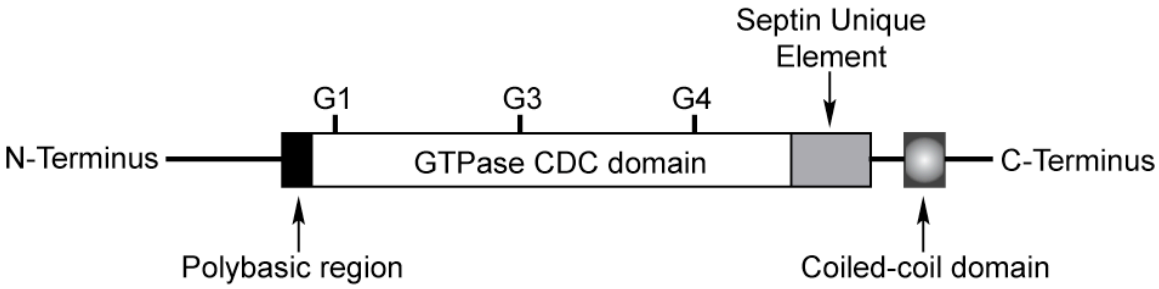
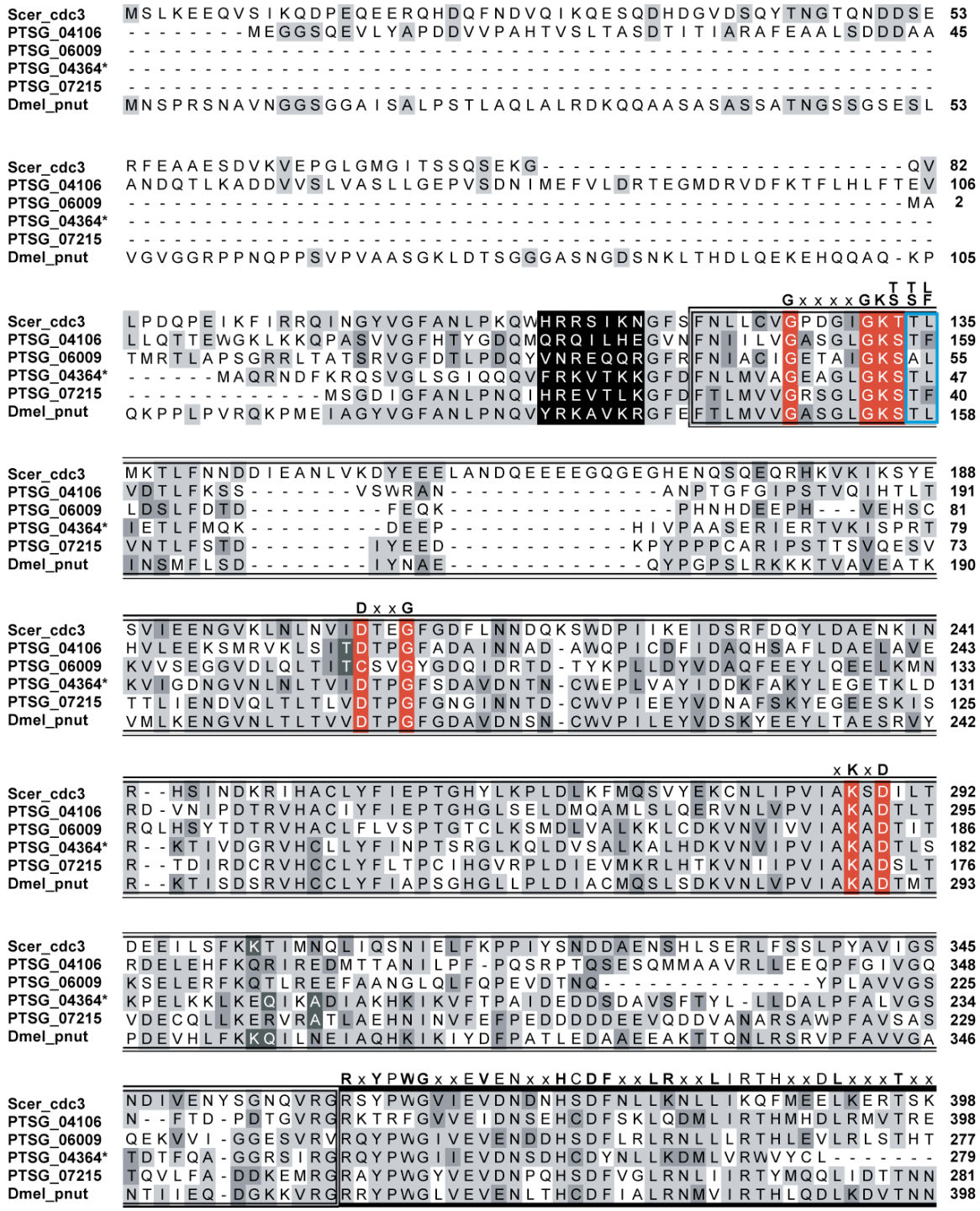


Figure S3.4 Septin sequence conservation. The amino-acid sequence alignment is shown for the four *S. rosetta* septins in addition to the *Drosophila melanogaster* and *Saccharomyces cerevisiae* septins Pnut and Cdc3, respectively. Numbers at the end of each line indicate amino acid positions. Conservation across the alignment is shown with grey shading. The approximate position of the polybasic region is indicated in black at the N-terminus while predicted coiled-coil domains are underlined in purple at the C-terminus. The GTP_CDC binding domain is in the double-bound box with the conserved motifs G1 (GxxxxGK[ST]), G2 (DxxG), and G3 (xKxD) indicated above the corresponding amino acid sequence that is highlighted in red. Two additional residues that are highly conserved across spetins are highlighted in the blue box directly following the G1 binding motif. The septin unique element (91) is in the single-bound box with the residues that are conserved in at least 50% of the septin sequence examined by (92) noted above the corresponding amino acid sequence. Bold typeface corresponds to residues that are conserved across at least 75% of these septins (92).


Scer_cdc3
PTSG_04106
PTSG_06009
PTSG_04364*
PTSG_07215
Dmel_pnut
G x x x x G K S S F
D x x G
x K x D
R x Y P W G x x E V E N x x H C D F x x L R x x L I R T H x x D L x x x T x x

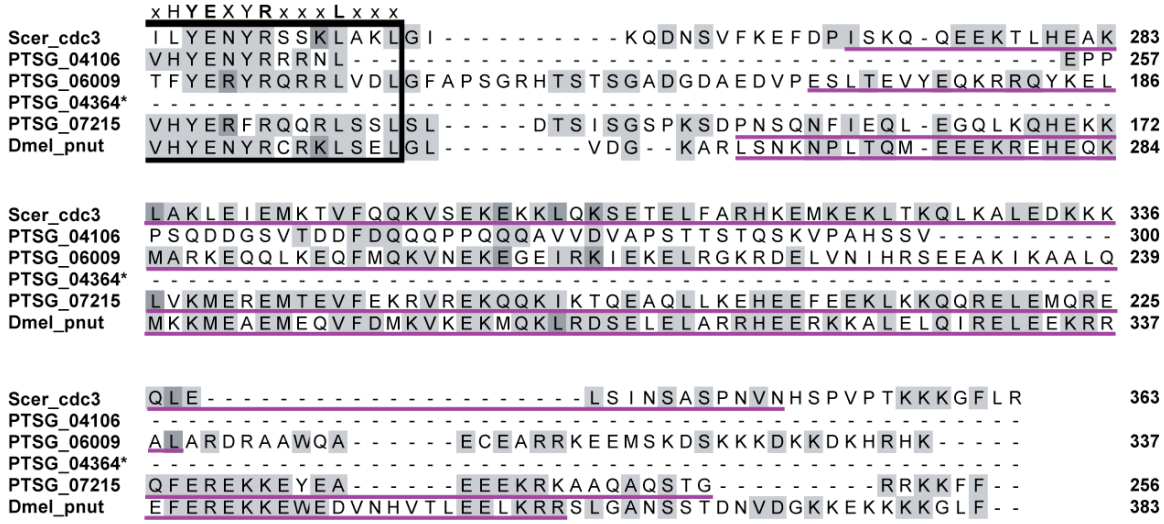

xHYExYRxxxLxxx
Scer_cdc3 283
PTSG_04106 257
PTSG_06009 186
PTSG_04364*
PTSG_07215 172
Dmel_pnut 284
Scer_cdc3 336
PTSG_04106 300
PTSG_06009 239
PTSG_04364*
PTSG_07215 225
Dmel_pnut 337
Scer_cdc3 363
PTSG_04106
PTSG_06009 337
PTSG_04364*
PTSG_07215 256
Dmel_pnut 383

Figure S3.5 Septin gene family phylogeny. The four *S. rosetta* amino acid sequences were added to Momany et al's (2008) published dataset using the same gene identifiers [see Table 1 in Momany et al's (2008)]. The sequences were aligned using Clustal Omega and highly variable regions were removed using Gblocks to avoid systematic errors that can results from spurious homology assignments that are characteristic of theses regions. The resulting topology includes many previously supported clades (92). The EcuSep2 (*Encephalitozoon cuniculi*) sequence is a notable exception as is the division of Group 2B. The major groupings are indicated to the right of each corresponding clade and the four *S. rosetta* septins are highlighted in bold with red arrows.


Group 3
(CDC11)
Group 5
(ASPE)
Group 4
(CDC12)
Group 2B
(M_III & IV)
Group 1 (A/B)
Group 1B
(M_II)
Group 1A
(CDC10+M_I)
Group 2B
(M_III & IV)
Group 2A
(CDC3)
PTSG_07215
PTSG_04364
PTSG_06009
PTSG_04106
0.0
0.0
0.0

Table S3.6 Phylogenetic enrichment by method and cell type.

	Attached									
	Intersection					Union				
	Observed	Expected	%	% Enrichment	P	Observed	Expected	%	% Enrichment	P
S.rosetta-Specific	241	189.4122658	28%	27%	7.35999E-06	551	444.3781942	27%	24%	1.64584E-09
Choano	71	38.58635009	8%	84%	7.24164E-07	134	90.52704429	7%	48%	7.09309E-06
Metazoan	41	30.97146508	5%	32%	0.060484499	95	72.66183986	5%	31%	0.008113076
Ancient	248	342.0299191	29%	-27%	1.911E-14	630	802.4329216	31%	-21%	3.98327E-20

	Colony									
	Intersection					Union				
	Observed	Expected	%	% Enrichment	P	Observed	Expected	%	% Enrichment	P
S.rosetta-Specific	36	26.15843271	31%	38%	0.01341906	187	151.2776831	26%	24%	0.000468367
Choano	9	5.328896934	8%	69%	0.074186735	32	30.81771721	4%	4%	0.735707997
Metazoan	3	4.27725724	3%	-30%	0.769370144	22	24.73594549	3%	-11%	0.66219415
Ancient	35	47.23541312	30%	-26%	0.013258776	239	273.1686542	33%	-13%	0.002003081

	Colony+Swim									
	Intersection					Union				
	Observed	Expected	%	% Enrichment	P	Observed	Expected	%	% Enrichment	P
S.rosetta-Specific	148	92.02725724	38%	61%	5.16409E-12	443	288.3730835	35%	54%	-1.35447E-13
Choano	18	18.74744463	5%	-4%	0.992938888	68	58.74627342	5%	16%	0.189601131
Metazoan	30	15.04770017	8%	99%	0.000246831	71	47.15289608	6%	51%	0.000616911
Ancient	96	166.177598	25%	-42%	3.43336E-16	333	520.727747	27%	-36%	2.39348E-35

	Swim									
	Intersection					Union				
	Observed	Expected	%	% Enrichment	P	Observed	Expected	%	% Enrichment	P
S.rosetta-Specific	155	90.1362862	40%	72%	2.53131E-14	426	270.7240204	36%	57%	-1.50546E-13
Choano	15	18.36222317	4%	-18%	0.511270486	60	55.15087308	5%	9%	0.451020203
Metazoan	10	14.73850085	3%	-32%	0.255758077	49	44.26703578	4%	11%	0.414556467
Ancient	106	162.7629898	27%	-35%	3.41355E-11	324	488.8580707	27%	-34%	1.99525E-29

	Colony+Attached									
	Intersection					Union				
	Observed	Expected	%	% Enrichment	P	Observed	Expected	%	% Enrichment	P
S.rosetta-Specific	6	12.92163543	8%	-54%	0.02322746	124	195.0851789	12%	-36%	2.06449E-10
Choano	3	2.632346678	4%	14%	0.536500404	28	39.74201448	3%	-30%	0.05606001
Metazoan	0	2.11286201	0%	-100%	0.228526779	23	31.89906303	2%	-28%	0.116463777
Ancient	32	23.33315588	44%	37%	0.002680134	444	352.2737436	42%	26%	3.19744E-14

	Attached+Swim									
	Intersection					Union				
	Observed	Expected	%	% Enrichment	P	Observed	Expected	%	% Enrichment	P
S.rosetta-Specific	133	79.10562181	41%	68%	1.23856E-12	557	370.3151618	35%	50%	0
Choano	31	16.11509796	10%	92%	0.000362559	130	75.43920358	8%	72%	2.06925E-09
Metazoan	13	12.93483816	4%	1%	0.839148313	71	60.55153322	4%	17%	0.154001219
Ancient	74	142.8444421	23%	-48%	3.00067E-18	417	668.6941014	26%	-38%	3.9686E-49

Figure S3.6 Phylogenetic enrichment is robust to species composition. To test the sensitivity of phylogenic enrichment to the species included we ran a jackknifing analysis with 10,000 trials each with a random set of species. *S. rosetta* and *M. brevicollis* were included in all trials and each of the 32 remaining species had an 80% probability of inclusion in any given trial. The OrthoMCL2 algorithm was rerun for trial to generate new clusters. Annotation frequencies were re-calculated for the entire genome and the expression clusters were tested for phylogenic enrichment.

Union

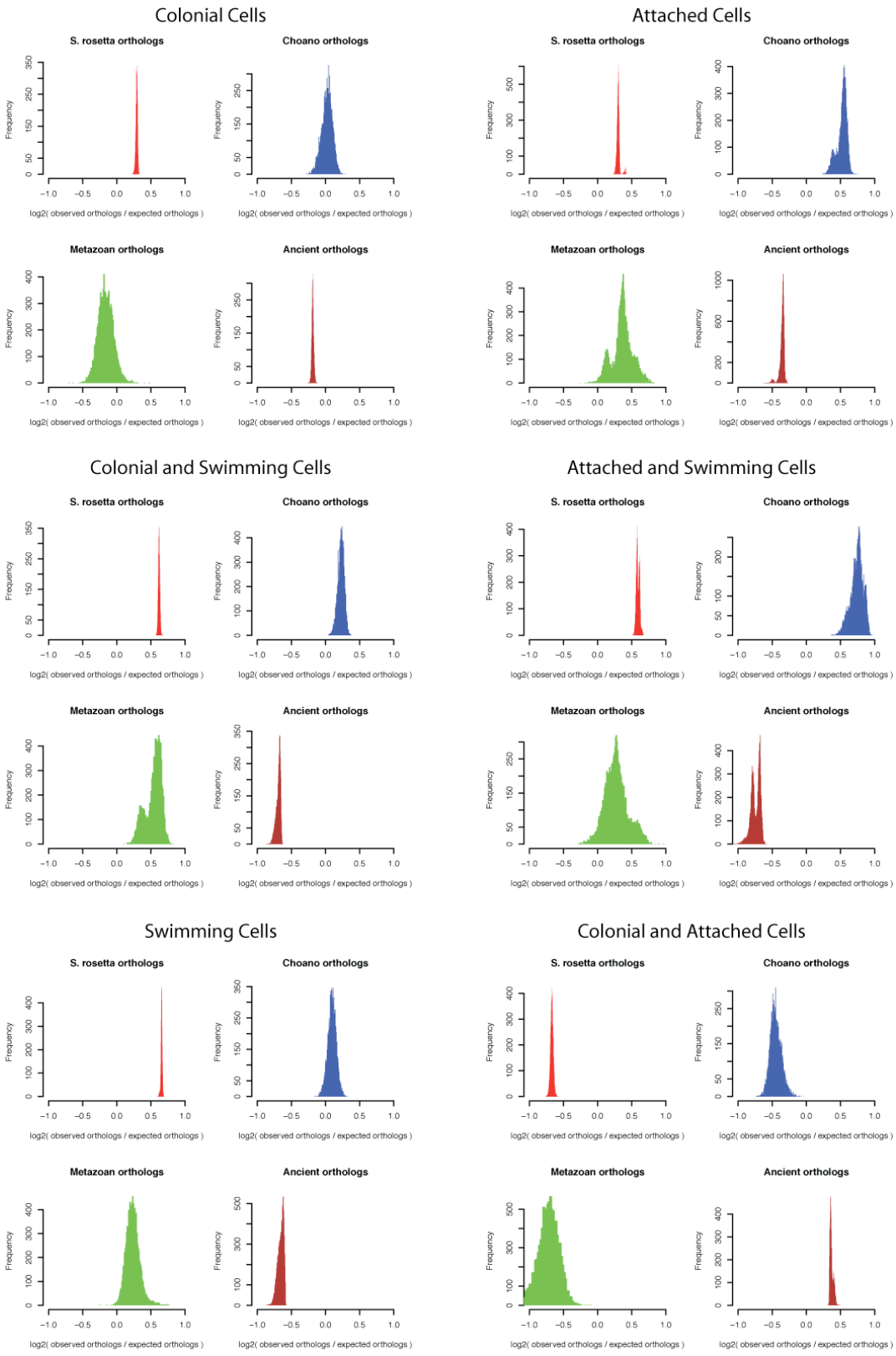
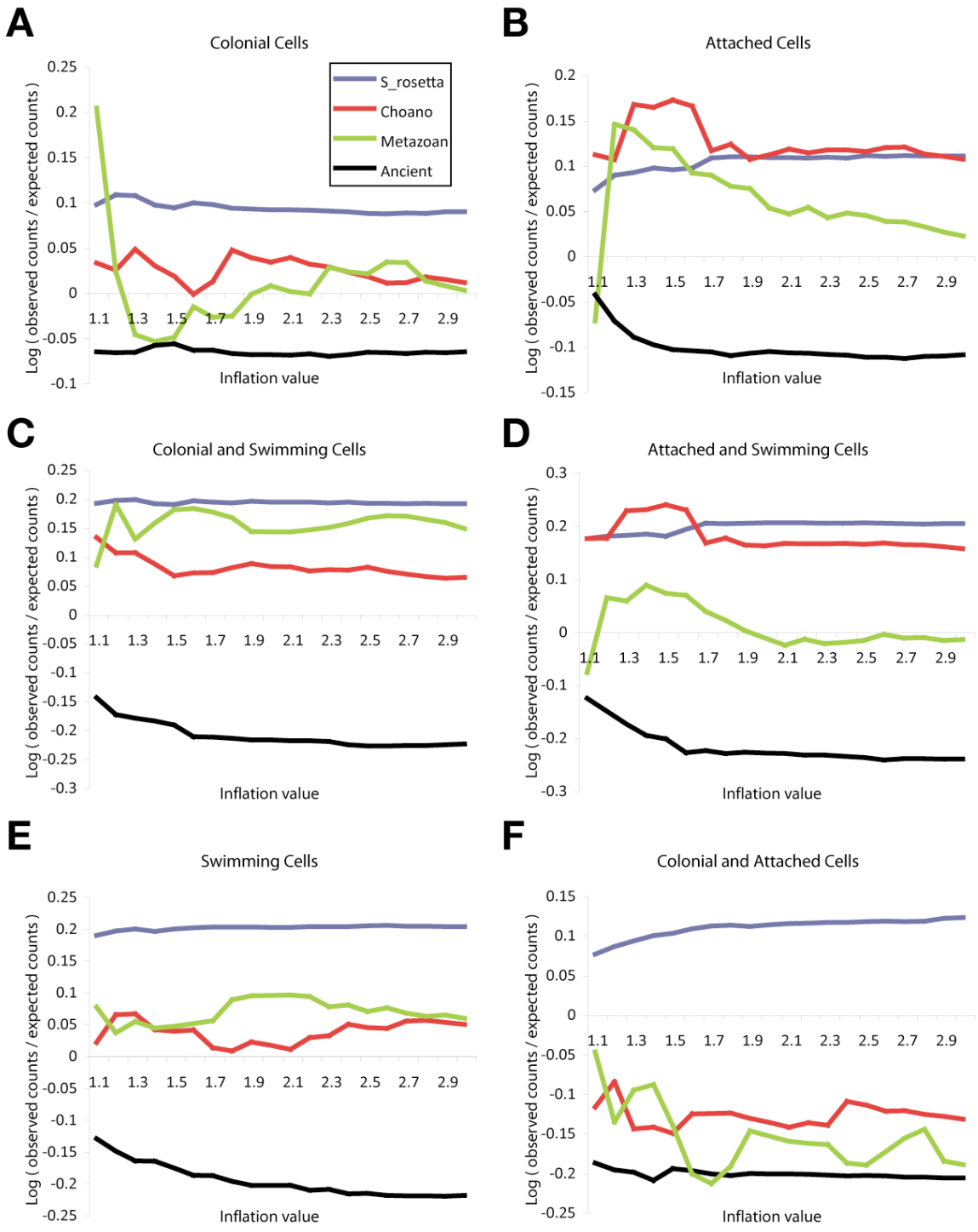

Colonial Cells
S. rosetta orthologs
Choano orthologs
Metazoan orthologs
Ancient orthologs
Attached Cells
S. rosetta orthologs
Choano orthologs
Metazoan orthologs
Ancient orthologs
Colonial and Swimming Cells
S. rosetta orthologs
Choano orthologs
Metazoan orthologs
Ancient orthologs
Attached and Swimming Cells
S. rosetta orthologs
Choano orthologs
Metazoan orthologs
Ancient orthologs
Swimming Cells
S. rosetta orthologs
Choano orthologs
Metazoan orthologs
Ancient orthologs
Colonial and Attached Cells
S. rosetta orthologs
Choano orthologs
Metazoan orthologs
Ancient orthologs
Frequency
log2(observed orthologs / expected orthologs)

Figure S3.7 Phylogenetic enrichment is robust to species inflation MCL value. To test the sensitivity of phylogenic enrichment to the MCL inflation parameter, we ran the MCL algorithm an addition 19 times with values from 1.1 to 3. All 34 species were included.

Union



1. Knoll, A. H. & Carroll, S. B. (1999) *Science* **284,** 2129-2137.
2. King, N. (2004) *Dev Cell* **7,** 313-325.
3. Kent, W. S. (1880) *A manual of the infusoria : including a description of all known flagellate, ciliate and tentaculiferous protozoa: British and foreign and an account of the organisation and affinities of the sponges* (David Bogue, London).
4. King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I.*, et al.* (2008) *Nature* **451,** 783-788.
5. Srivastava, M., Begovic, E., Chapman, J., Putnam, N. H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M. L.*, et al.* (2008) *Nature* **454,** 955-U919.
6. Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E., Mitros, T., Richards, G. S., Conaco, C., Dacre, M., Hellsten, U.*, et al.* (2010) *Nature* **466,** 720-726.
7. King, N., Hittinger, C. T., & Carroll, S. B. (2003) *Science* **301,** 361-363.
8. Leadbeater, B. & Thomsen, H. (2000) *Lawrence : Society of Protozoologists.***,** 14-38.
9. Leadbeater, B. (2008) in *Protistology*, pp. 256-267.
10. Dayel, M. J., Alegado, R. A., Fairclough, S. R., Levin, T. C., Nichols, S. A., Mcdonald, K., & King, N. (2011) in *Developmental Biology*, pp. 73-82.
11. Leadbeater, B. & Kelly, M. (2001) in *Water & Atmosphere*, pp. 9-11.
12. Lavrov, D. V. (2007) *Integr Comp Biol* **47,** 734-743.
13. Carr, M., Leadbeater, B. S., Hassan, R., Nelson, M., & Baldauf, S. L. (2008) *Proc Natl Acad Sci U S A* **105,** 16641-16646.
14. Abedin, M. & King, N. (2008) *Science* **319,** 946-948.
15. Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., & Quéinnec, E. (2009) in *Current Biology*, pp. 706-712.
16. Rokas, A., King, N., Finnerty, J., & Carroll, S. B. (2003) *Evol Dev* **5,** 346-359.
17. Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011) in *Plos Biol*, p. e1000602.
18. Dunn, C., Hejnol, A., Matus, D., Pang, K., Browne, W., Smith, S., Seaver, E., Rouse, G., Obst, M., Edgecombe, G.*, et al.* (2008) in *Nature*, pp. 745-749.
19. Schierwater, B., Eitel, M., Jakob, W., Osigus, H.-J., Hadrys, H., Dellaporta, S. L., Kolokotronis, S.-O., & Desalle, R. (2009) in *Plos Biol*, p. e20.
20. Lavrov, D. V., Wang, X., & Kelly, M. (2008) in *Molecular Phylogenetics and Evolution*, pp. 111-124.
21. Sperling, E. A., Peterson, K. J., & Pisani, D. (2009) *Molecular biology and evolution* **26,** 2261-2274.
22. Tyler, S. (2003) *Integr Comp Biol* **43,** 55-63.
23. Leys, S. P., Nichols, S. A., & Adams, E. D. M. (2009) in *Integr Comp Biol*, pp. 167-177.
24. Dickinson, Daniel J., Weis, William I., & Nelson, W. J. (2011) in *Developmental Cell*, pp. 32-34.
25. Adams, E. D. M., Goss, G. G., & Leys, S. P. (2010) in *PLoS ONE*, p. e15040.

26. Sebe-Pedros, A., Roger, A. J., Lang, F. B., King, N., & Ruiz-Trillo, I. (2010) *Proceedings of the National Academy of Sciences of the United States of America* **107,** 10142-10147.
27. Nichols, S. A., Roberts, B. W., Richter, D. J., Fairclough, S. R., & King, N. (2011) *In prep*.
28. Harwood, A. J. (2008) *Methods in molecular biology (Clifton, N.J* **469,** 21-32.
29. Nichols, S. A., Mark J, D., & King, N. (2008) in *Animal Evolution: Genes, Genomes, Fossils and Trees*.
30. Abedin, M. & King, N. (2010) in *Trends in cell biology*, pp. 734-742.
31. Gerhart, J. (1999) *Teratology* **60,** 226-239.
32. Nichols, S., Dirks, W., Pearse, J., & King, N. (2006) in *Proceedings of the National Academy of Sciences*, pp. 12451-12456.
33. Bridgham, J. T., Eick, G. N., Larroux, C., Deshpande, K., Harms, M. J., Gauthier, M. E. A., Ortlund, E. A., Degnan, B. M., & Thornton, J. W. (2010) in *Plos Biol*, p. e1000497.
34. Gazave, E., Lapebie, P., Richards, G. S., Brunet, F., Ereskovsky, A. V., Degnan, B. M., Borchiellini, C., Vervoort, M., & Renard, E. (2009) *BMC Evol Biol* **9,** 249.
35. Lee, J. J., Ekker, S. C., von Kessler, D. P., Porter, J. A., Sun, B. I., & Beachy, P. A. (1994) in *Science* , pp. 1528-1537.
36. Adamska, M., Matus, D. Q., Adamski, M., Green, K., Rokhsar, D. S., Martindale, M. Q., & Degnan, B. M. (2007) *Curr Biol* **17,** R836-837.
37. Darnell, J. E., Jr. (1997) *Proc Natl Acad Sci U S A* **94,** 11767-11769.
38. Manning, G., Young, S. L., Miller, W. T., & Zhai, Y. (2008) *Proc Natl Acad Sci U S A* **105,** 9674-9679.
39. Pincus, D., Letunic, I., Bork, P., & Lim, W. (2008) in *P Natl Acad Sci Usa*.
40. Li, W., Scarlata, S., & Miller, W. T. (2009) in *Biochemistry*, pp. 5180-5186.
41. Sebe-Pedros, A., De Mendoza, A., Lang, B. F., Degnan, B. M., & Ruiz-Trillo, I. (2011) in *Molecular biology and evolution*, pp. 1241-1254.
42. Davidson, E. H. & Erwin, D. H. (2006) in *Science* , pp. 796-800.
43. Extavour, C. G., Pang, K., Matus, D. Q., & Martindale, M. Q. (2005) in *Evol Dev*, pp. 201-215.
44. Richards, G., Simionato, E., Perron, M., Adamska, M., Vervoort, M., & Degnan, B. (2008) in *Current Biology*, pp. 1156-1161.
45. Adamska, M., Larroux, C., Adamski, M., Green, K., Lovas, E., Koop, D., Richards, G. S., Zwafink, C., & Degnan, B. M. (2010) in *Evolution & Development*, pp. 494-518.
46. Fairclough, S., Chen, Z., Kramer, E., Zhen, Q., Young, S., Robertson, H., Begovic, E., Westbrook, M. J., Richter, D. J., Manning, G.*, et al.* (2011) **In Prep**.
47. Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V. V.*, et al.* (2007) *Science* **317,** 86-94.
48. Young, S. L., Diolaiti, D., Conacci-Sorrell, M., Ruiz-Trillo, I., Eisenman, R. N., & King, N. (2011) in *Molecular biology and evolution*, pp. 2961-2971.
49. Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degnan, B. M., Rokhsar, D. S., & Bartel, D. P. (2008) *Nature* **455,** 1193-1197.

50. John Maynard Smith, E. S. (1995).
51. Jacob, F. (1977) *Science* **196,** 1161-1166.
52. Grosberg, R. K. & Strathmann, R. R. (2007) *Annual Review of Ecology Evolution and Systematics* **38,** 621-654.
53. Prochnik, S. E., Umen, J., Nedelcu, A. M., Hallmann, A., Miller, S. M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., Fritz-Laylin, L. K.*, et al.* (2010) in *Science* , pp. 223-226.
54. Ruiz-Trillo, I., Burger, G., Holland, P., King, N., Lang, F., Roger, A., & Gray, M. (2007) *Trends in Genetics* **23,** 113-118.
55. Haeckel, E. (1874) *Quarterly Journal of Microscopical Science* **14,** 142-165.
56. Wolpert, L. & Szathmary, E. (2002) *Nature* **420,** 745.
57. Mikhailov, K. V., Konstantinova, A. V., Nikitin, M. A., Troshin, P. V., Rusin, L. Y., Lyubetsky, V. A., Panchin, Y. V., Mylnikov, A. P., Moroz, L. L., Kumar, S.*, et al.* (2009) *Bioessays* **31,** 758-768.
58. Steenkamp, E. T., Wright, J., & Baldauf, S. L. (2006) *Molecular biology and evolution* **23,** 93-106.
59. King, N., Young, S. L., Abedin, M., Carr, M., & Leadbeater, B. S. (2009) *Cold Spring Harb Protoc* **2009,** pdb emo116.
60. Leadbeater, B. S. C. (1983) *Journal of the Marine Biological Association U.K.* **63,** 135-160.
61. King, N., Young, S. L., Abedin, M., Carr, M., & Leadbeater, B. S. (2009) *Cold Spring Harb Protoc* **2009,** pdb prot5147.
62. Ikegami, S., Taguchi, T., Ohashi, M., Oguro, M., Nagano, H., & Mano, Y. (1978) *Nature* **275,** 458-460.
63. King, N. (2005) *Curr Biol* **15,** R113-114.
64. Fairclough, S. R., Dayel, M. J., & King, N. (2010) *Current Biology* **20,** R875-R876.
65. Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W.*, et al.* (2000) *Science* **287,** 2204-2215.
66. Ginzberg, R. D., Morales, E. A., Spray, D. C., & Bennett, M. V. (1985) *Cell Tissue Res* **239,** 477-484.
67. Field, C. M., al-Awar, O., Rosenblatt, J., Wong, M. L., Alberts, B., & Mitchison, T. J. (1996) *The Journal of cell biology* **133,** 605-616.
68. Field, C. M. & Kellogg, D. (1999) *Trends in cell biology* **9,** 387-394.
69. Mendoza, M., Hyman, A. A., & Glotzer, M. (2002) *Curr Biol* **12,** 1858-1863.
70. Casamayor, A. & Snyder, M. (2003) *Molecular and cellular biology* **23,** 2762-2777.
71. Zhang, J., Kong, C., Xie, H., McPherson, P. S., Grinstein, S., & Trimble, W. S. (1999) *Curr Biol* **9,** 1458-1467.
72. Blau, K., Portnoi, M., Shagan, M., Kaganovich, A., Rom, S., Kafka, D., Chalifa Caspi, V., Porgador, A., Givon-Lavi, N., Gershoni, J. M.*, et al.* (2007) *The Journal of infectious diseases* **195,** 1828-1837.
73. Boyle, E. C. & Finlay, B. B. (2003) *Current opinion in cell biology* **15,** 633-639.
74. Mengaud, J., Ohayon, H., Gounon, P., Mege, R. M., & Cossart, P. (1996) *Cell* **84,** 923-932.

75. Beachy, P. A., Hymowitz, S. G., Lazarus, R. A., Leahy, D. J., & Siebold, C. (2010) *Genes & development* **24,** 2001-2012.
76. Wilkinson, D. G. (2000) *International review of cytology* **196,** 177-244.
77. Hu, Q., Milenkovic, L., Jin, H., Scott, M. P., Nachury, M. V., Spiliotis, E. T., & Nelson, W. J. *Science* **329,** 436-439.
78. Toure, A., Rode, B., Hunnicutt, G. R., Escalier, D., & Gacon, G. (2011) *Biological chemistry* **392,** 799-803.
79. King, N. & Carroll, S. B. (2001) *Proc Natl Acad Sci U S A* **98,** 15032-15037.
80. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., & Sonnhammer, E. L. (2002) *Nucleic acids research* **30,** 276-280.
81. Lupas, A., Van Dyke, M., & Stock, J. (1991) *Science* **252,** 1162-1164.
82. Loopp (2011).
83. Schrodinger, LLC (2010).
84. Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., Terry, A., Salamov, A., Fritz-Laylin, L. K., Marechal-Drouard, L.*, et al.* (2007) *Science* **318,** 245-250.
85. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al. Molecular systems biology* **7,** 539.
86. Castresana, J. (2000) *Molecular biology and evolution* **17,** 540-552.
87. Guindon, S., Lethiec, F., Duroux, P., & Gascuel, O. (2005) *Nucleic acids research* **33,** W557-559.
88. Li, L., Stoeckert, C. J., Jr., & Roos, D. S. (2003) *Genome research* **13,** 2178-2189.
89. Chen, F., Mackey, A. J., Stoeckert, C. J., Jr., & Roos, D. S. (2006) *Nucleic acids research* **34,** D363-368.
90. Djupedal, I., Kos-Braun, I. C., Mosher, R. A., Soderholm, N., Simmer, F., Hardcastle, T. J., Fender, A., Heidrich, N., Kagansky, A., Bayne, E.*, et al.* (2009) *The EMBO journal* **28,** 3832-3844.
91. Versele, M. & Thorner, J. (2004) *The Journal of cell biology* **164,** 701-715.
92. Momany, M., Pan, F., & Malmberg, R. L. (2008) in *The Septins* (John Wiley & Sons, Ltd), pp. 35-45.