The unicellular ancestry of the proto-oncogene Myc

Permalink
https://escholarship.org/uc/item/5hr0154z

Author
Young, Susan Louise

Publication Date
2010

Peer reviewed|Thesis/dissertation
The unicellular ancestry of the proto-oncogene Myc

by

Susan Louise Young

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

Committee in charge:

Professor Nicole King, chair
Professor Michael B. Eisen
Professor N. Louise Glass
Professor Michael Levine

Spring 2010
Abstract

The unicellular ancestry of the proto-oncogene Myc

by Susan Louise Young

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Nicole King, Chair

The origin of metazoans required the evolution of mechanisms of cell-cell adhesion, coordination and communication among neighboring cells, and the establishment of differentiated cell types. Did the molecular building blocks of metazoan multicellularity exist in their single celled ancestors, or are they unique metazoan innovations? To address this question we sequenced and analyzed the genome of the unicellular marine choanoflagellate Monosiga brevicollis. Choanoflagellates, a phylum of flagellated unicellular and colonial eukaryotes found in diverse aqueous habitats around the globe, are among the closest living relatives of metazoans.

The roughly 46 million base pair genome of M. brevicollis contains approximately 9,200 unexpectedly intron-rich genes, including a number of genes that encode cell adhesion and signaling protein domains that are otherwise restricted to metazoans. The physical linkages among these domains often differ between M. brevicollis and metazoans suggesting that abundant domain shuffling followed the separation and subsequent diversification of the choanoflagellate and metazoan lineages. Metazoans also have a richer diversity of transcription factors than does M. brevicollis, indicating that the evolution of early metazoans may have involved an increase in the sophistication of transcriptional regulation. Nonetheless, a few metazoan-type transcription factors were identified in M. brevicollis: members of the p53, Myc, and Sox/TCF transcription factor families.

Myc is a developmentally critical transcription factor that plays roles in the most fundamental of cellular processes: cell growth, proliferation, and death. Investigating the function of Myc in choanoflagellates promises to delineate the role of Myc before the origin of animals and may inform how the strict regulation of cell life and death in metazoans arose from a unicellular context. Here, we demonstrate M. brevicollis Myc heterodimerizes with M. brevicollis MAX and localizes to the nucleus and cytoplasm of choanoflagellate cells in varying intensity. We further show that the tyrosine kinase (TK) inhibitor genistein reduces the expression of MbMyc, suggesting that TK signaling regulates MbMyc. Because metazoans Mycs are also known to be regulated by TK signaling, we hypothesize that an emergent network of TK signaling and transcriptional regulation was present in the
unicellular ancestor of animals.
Chapter 1: The Origin of Animals

Over 650 million years ago, before the supercontinent Pangea had formed, before plants had moved from lakes to land, the first multicellular animal evolved from a single-celled ancestor. The history of all animals, from giant squids to garden slugs, from corals to cats, can be traced back to a common single-celled, or unicellular, ancestor. The consequence? All animal body plans were built upon a common genetic foundation; a foundation that, in many ways, directed their subsequent evolution and development. To best understand where we come from, and what makes animals unique among Earth's creatures, we must understand the unicellular context from which animals arose. Recent insights from the study of choanoflagellates, a charismatic group of single-celled protozoa that are closely related to animals, have brought us closer to understanding the unicellular nature of animal origins.

For the first two billion years of life on Earth, unicellular organisms ruled. Indeed, unicellular life still predominates on Earth, both in terms of biomass and number of species. However, for many of us, it is the multicellular organisms that define our ecosystems and seem most familiar. There are over a dozen multicellular groups on Earth today, including plants, some fungi, multicellular algae, and of course, animals. You might expect that all these groups share a common origin and that the original multicellular life form later diversified into redwoods, mushrooms, and crows. However, genetic and cellular evidence clearly show that each of these groups made the transition from unicellularity to multicellularity independently (Baldauf 2003).

Because all major multicellular groups evolved independently, their biology and lifestyles were uniquely shaped by those of their unicellular ancestors. For example, land plants evolved from an ancestor with chloroplasts and a cell wall; today plants gather energy by photosynthesis and their cells are generally non-motile (Becker and Marin 2009; Kenrick and Crane 1997). Although each origin of multicellularity was independent, all of these 15 or so major transitions required innovations in the ways cells stick together, communicate with one another, and differentiate their functions for the greater benefit of the whole organism. No doubt, unicellular organisms have their own tools for adhesion, communication, and differentiation that they use to mate, find food and avoid predation. What’s more, the toolset at a given unicellular organism’s disposal can vary greatly in the distinct branches of life. Therefore, we must focus our gaze on the most recent unicellular ancestors of animals to understand the basic cellular context from which animals arose.

While the last unicellular progenitor of animals is long extinct and left no fossil record behind, we can look to living organisms to reconstruct our unicellular ancestry. The choanoflagellates are the closest living relatives of animals and in these microorganisms are clues to the biology of the organism that founded the animal kingdom (Lang et al. 2002; King 2004). Choanoflagellates are unicellular aquatic organisms that resemble modified sperm (Leadbeater 2000), with a ring of finger-like projections, the “collar,” that surrounds the flagellum and is used to capture bacteria prey (Box 2). The flagellum propels the choanoflagellate through water or, if the choanoflagellate has settled on a surface, it can use its flagellum to generate a current that draws bacteria into its collar. There are at least 150
species of choanoflagellates and they can be found in both salt water and fresh water, all across the globe. Chances are, if you live near a stream, lake or ocean, some of your unicellular cousins, the choanoflagellates, are in it.

The first hints that choanoflagellates might be closely related to animals came from studies by 19th century cell biologists Henry James Clark and William Saville-Kent (Saville-Kent 1880). With new microscope technology that allowed biologists a better view into the sub-millimeter world, they observed that the unique cell shape of the choanoflagellate was remarkably similar to that of feeding cells of the oldest group of animals, the sponges. In many senses, sponges are the simplest type of animals. They do not have a nervous system, a digestive tract, or even true tissues; instead, they constantly circulate water through channels in their body to bring nutrients in and take waste out. Sponges feed on microscopic bacteria floating in the water and they trap their microscopic prey on their collared-cells, just like choanoflagellates.

More than 650 MYA, the genetic lines of choanoflagellates and animals split, each taking the clay of their common ancestor's genome and molding it into two new branches of life. By identifying the genetic features that are shared between choanoflagellates and animals, we can better understand what their common ancestor's genome held. In the last few years, such comparisons have been facilitated by researchers who have uncovered the genetic features of basal animals like the sponge and the sea anemone, as well as the choanoflagellate, deepening our understanding of animal origins and animal evolution (Srivastava et al. 2008; Putnam et al. 2007; Nichols et al. 2006; King et al. 2008; Chapman et al. 2010). Prior to these important sequencing projects, there was a huge gap in our knowledge of the gene content of early branching animal groups. We can now see that there is a good deal of conservation in the genetic toolkit of all animals, no matter their shape or stature.

Animals evolved from an organism whose own genome had been shaped by billions of years of unicellular life. To understand the specific unicellular context from which animals evolved, we can compare the genetic content of diverse animals to their unicellular relatives. Because they fulfill the basic requirements of multicellularity, the biological phenomena of cell-cell adhesion, cell-cell communication, and the regulation of cell differentiation were probably the most critical to the evolution of multicellularity. Therefore, the steps that led to multicellularity included co-option of unicellular adhesion, communication, and differentiation tools into new and perhaps more elaborate functions with the additional innovation of new genes and gene functions, to create Earth's most impressive branch of life, the animals.

**Multicellular need: Cell-cell adhesion**

Single-celled organisms can adhere to one another as well as to surfaces. For example, baker's yeast can join with other members of its species to mate. Adhesion proteins on the outer surfaces of yeast cells interact and enable the cells to fuse (Cross et al. 1988). Some unicellular eukaryotes also adhere to inert surfaces; for example, pathogenic yeasts such as *Candida albicans* adhere to medical devices and form drug-resistant biofilms (Kojic and Darouiche 2004).
While adhesion is used in special situations in unicellular eukaryotes, it is a basic requirement of multicellular life. Cell adhesion proteins are important for tissue structure and the ability to create water-tight barriers like our skin. Other adhesion proteins make up a large part of our connective tissues by serving as linker molecules, providing structural support to the animal cells. Animals have dozens of specialized adhesion genes that are only found in animal genomes. It's unlikely that all of these animal-type adhesion proteins evolved en masse in the genome of the first animals and so a closer look at animal origins was necessary to resolve when and how animal adhesion genes first evolved.

In 2000, one of us (King) began taming the choanoflagellate into a workable research organism. Modern biology had moved far beyond the simple light microscopes used by Clark and Seville-Kent. Molecular biology had greatly advanced our knowledge of the inner mechanics of cells and the emerging Age of Genomics was beginning to expose the wealth of knowledge hidden in the genomes of life on earth. King sampled some of the genes in the choanoflagellate Monosiga brevicollis (a solitary choanoflagellate isolated from an underwater cave in Bermuda) and found animal-type adhesion molecules called cadherins. Cadherins are long adhesive proteins that ensure cells in animal tissues are tightly bound together.

It was not initially clear what a protein responsible for keeping cells stuck together in an animal might be doing in an organism that never sticks to other members of its species. However, a graduate student in the King Lab demonstrated that two cadherins are localized to the choanoflagellate collar – suggesting that they might be used in detecting or capturing bacteria for the choanoflagellate's dinner (Abedin and King 2008). Cadherins are also used by pathogenic bacteria to sneak themselves into animal cells, providing independent evidence that bacteria can bind to cadherins (Mengaud et al. 1996). Together, these facts suggest that an ancestral role of the cadherin gene family might have been to bind to bacteria. At some point early in animal evolution, then, these proteins evolved a new function – to bind animal cells together. The phenomenon of adaptation of existing proteins for a new function, “co-option,” is one mechanism by which proteins evolve.

**Multicellular need: Cell-cell communication**

One of the challenges for multicellular organisms is to coordinate and integrate the functions of different cells so they do not work at cross purposes. The language spoken by cells, be they unicellular organisms or different cells in your body, is that of molecular signaling pathways, in which small proteins secreted by one cell travel to a receptor on another cell where they stimulate a cascade of specific molecular activity. There are dozens of signaling pathways in eukaryotes and each uses different types of molecules to carry their messages, creating specific signaling without the worry of crossed lines of communication. Specific receptors on the outer surface of cells receive the messages and communicate the information from the outside of the cell to the inner control center, the nucleus. The outside-to-inside communication often relies on chemical modifications performed by the receptor protein, which starts a cascade of protein-to-protein changes, passing the signal down like a baton in a relay race.
Cell signaling between unicellular organisms can occur between members of the same species, like the signaling between different mating types in yeast, or it can occur between members of different species. For example, members of the *Plasmodium* family, the unicellular parasite that causes malaria, synchronize their maturation in response to melatonin produced by their host. This ensures that the blood of their victims contains the highest concentration of mature parasites at the time their carrier, the mosquito, is likely to bite its victim (Srinivasan et al. 2009).

Animals have their own types of signaling pathways. Seven critical developmental signaling pathways give animal cells the information that they need to build an adult animal from an embryo (Gerhart 1999). Despite the outcomes of developmental programs across animal diversity, all of the seven major developmental signaling pathways seem to be present in all major animal lineages. Evolutionary biologists theorize that genes that guide the growth of developing animals also played important roles during the evolution of animal body plans. For example, the genes that are important for organizing the architecture of modern animal skeletons are thought to have also been important for the evolution of skeletal patterning. Because animals seemed to be the only organisms that had the seven important developmental signaling pathways, they were thought to be key innovations for the origin of animals.

Receptor tyrosine kinases, for example, are a family of proteins that were thought to be unique to animals and biologists postulated that this family was an important innovation for animal origins. Receptor tyrosine kinases are situated at the outer edge of cells and can recognize signals from other cells. Many of the signals recognized by receptor tyrosine kinases are growth factors that instruct cells to grow and divide. However, when we sampled genes from the choanoflagellate *M. brevicollis*, we found receptor tyrosine kinases in its genome. What's more, when we measured choanoflagellate tyrosine kinase activity by first starving cells and then adding back nutrient-rich food (a well-known assay developed in animal systems), we found that the choanoflagellate tyrosine kinases also responded to the influx of food with a flurry of activity, just like animal cells. Thus, ancestral tyrosine kinase proteins may have worked to monitor nutrient availability in the environment.

A complete picture of the genetic sequence of *M. brevicollis* came from the Monosiga Genome Project carried out by us and other members of the King Lab at the University of California at Berkeley in collaboration with Prof. Rokhsar and colleagues at the Joint Genome Institute in Walnut Creek, California. With the complete genome in hand and help from collaborators, we could see that there are actually more tyrosine kinase genes in a unicellular choanoflagellate than there are in humans (Manning et al. 2008). The sequenced genome of *M. brevicollis* has corroborated and extended the earlier finding that some animal gene families with strong connections to cellular interactions and development are present in choanoflagellates and thus were present in their unicellular common ancestor.

**Multicellular Need: Differentiation**

A great advance of multicellular life is the ability to have different cells performing different functions. This allows each cell to perform a different job and, as a whole, the organism benefits from the division of labor. Multicellular organisms accomplish this with
molecular systems that precisely regulate which genes are active in each cell. Animals possess many kinds of specialized proteins called transcription factors that control which genes are turned on, or "expressed," and when. A cell in your heart, for example, expresses a different set of genes than a cell in your liver, and this is largely due to the activity of transcription factors.

Unicellular organisms can change their gene expression program in response to environmental factors. For example, when the protist Naegleria gruberi encounters a nutrient-poor environment, it switches from a slow-moving amoeboid form into a fast-swimming flagellated form that is more adept at finding a better feeding ground. Naegleria accomplishes this switch by activating and inactivating certain genes that govern cell shape (Fulton 1977). Thus, this unicellular organism is capable of having different cell shapes over its lifespan by temporally differentiating its gene expression.

The functions of each cell in a multicellular organism become integrated and interdependent through both temporal and spatial differentiation of gene expression. Often, the differentiation of cells represents the intersection of cell-cell signaling with the regulation of gene expression: communication between cells travels form the surface of the receiving cell inward to the nucleus, where genes can be activated or inactivated in response. In general, animals possess larger and more complex gene regulatory mechanisms than their simple unicellular relatives (Levine and Tjian 2003). Likewise, many of the specialized transcription factors found in animals are not found in the M. brevicollis genome. These observations suggest that much of the complex gene regulation activities in animals evolved after choanoflagellates and early animals split ways. This raises the possibility that innovations in gene regulatory systems were one of the key steps in the origin of animals.

Two exceptions to this finding are two families of genes, p53 and Myc, that often cause cancer in humans when they aren't functioning properly. This is in part due to the important roles these genes play in the life, death, and proliferation of animal cells. A standing question is whether choanoflagellate p53 and Myc play similar roles in choanoflagellates. Understanding their function in choanoflagellates will give us insight into what they were doing in the last unicellular ancestor of animals. And this insight will expand our understanding of why and how these processes sometimes fail in humans. Both of these gene families have been studied by computational biologists who can use the information encoded in genome sequences to predict the function of and influences on genes. A study headed by Prof. Aurora Nedelcu of the University of New Brunswick found that selective pressures beyond those found in vertebrates (like an increase in lifespan) are relevant to the p53 gene family's history and a study by Prof. Albert Erives and colleagues at Dartmouth College predicts that choanoflagellate Myc will regulate genes important for protein production, a very critical part of cell growth, a function of the Myc gene family in animals (Brown, Cole, and Erives 2008; Nedelcu and Tan 2007; Eilers and Eisenman 2008).

**Domain shuffling: a mechanism for protein evolution**

With new genomic sequences from early animal groups and the choanoflagellate, we have access to more data points along the path of animal evolution, allowing us to trace the step-wise changes of evolutionarily important gene families. Piecing together the evolution
of the hedgehog signaling pathway, composed of the signaling protein Hedgehog and its many protein binding partners, could show us how a complex and developmentally critical signaling pathway evolved in animals. Hedgehog signaling is required for the development of many different tissues in vertebrates and insects (the pathway is named after a mutation in flies which disrupts development and causes fly embryos to appear stubbly and hairy) (Nüsslein-Volhard and Wieschaus 1980; Riddle and Tabin 1999). The signaling molecule, Hedgehog, travels from one cell to another, regulating the fate and behavior of cells that receive it. Hedgehog proteins in animals have two particular stretches of amino acids that each serve a different function. Parts of proteins that act as modular units, like these stretches do, are called “domains.” One domain in Hedgehog acts like scissors to cut itself free from the rest of the protein. When the scissor-like domain is gone, the other domain, the true signaling part of the protein, becomes active and can go activate other cells.

Protein domains can be identified on the basis of their amino acid sequence alone. When members of the King Lab examined the *M. brevicollis* genome, we were able to identify both the scissor-like domain and the signaling domain of Hedgehog proteins. However, instead of finding both domains in the same protein, the domains were found in two different proteins in choanoflagellates. The evolutionary path between the unicellular Hedgehog-like protein and that of animals was further elucidated by the sea anemone and sponge genomes. Genetic footprints lead to the conclusion that after the animals and choanoflagellates diverged, a process known as domain shuffling brought the scissor and signaling domains together into a new protein, the ancestral Hedgehog.

Domain shuffling, a molecular phenomenon by which the modular units of ancestral proteins are rearranged into new proteins with new functions, played an important role in the evolution of animal signaling systems. Because each domain can function as an individual unit, pre-existing protein units can be combined to create new functions. Thus, it has been suggested that modular protein evolution by domain shuffling contributed significantly to animal evolution. This phenomenon is also seen in the evolution of the tyrosine kinases. Both choanoflagellates and animals have lots of tyrosine kinases but a surprising discovery of the *M. brevicollis* genome was that only a small subset of its many tyrosine kinases have domain combinations identical to those of animals. The striking disparity of domain combinations suggests that receptor tyrosine kinase signaling existed in a nascent form in the common ancestor of choanoflagellates and animals, and was subsequently specialized and elaborated upon in each lineage.

**Future Directions and the bigger question**

In the next few years, we expect a great expansion in our knowledge of the genomes and cellular biology of choanoflagellates and the earliest branching animals. As we learn more about the unicellular functions of gene families that are important to the multicellular lifestyle of animals, we will shed light are the fundamental functions of these gene families. The biochemical and regulatory properties of genes that are critical for animal development (and that are often medically-relevant for humans) are currently active areas of study in choanoflagellates. As greater numbers of labs continue to turn to the choanoflagellate and other unicellular relatives of animals as new model organisms, we expect great strides in the community's understanding of their biology and thus the biology of the unicellular ancestor.
of animals.

An understanding of how animal multicellularity evolved leaves us with bigger questions: why did it evolve and what can we learn from examining other transitions to multicellularity? The fact that multicellularity evolved multiple times in a wide variety of evolutionary lineages (a situation unlikely to arise purely by accident) implies that multicellularity can endow some organisms with an evolutionary advantage (Rokas 2008). Such advantages might include avoiding predation by outgrowing predators, ease of reproduction and dispersal, or more efficient food consumption. As our understanding of the transition to multicellularity in various parts of the tree of life continues to grow, we hope that unifying features of these major transition will become clear.
Chapter 1 Figures
Box 1: Multiple origins of multicellularity and their nearest neighbors
Box 2: Choanoflagellate colonies: are prelude to multicellularity?
Box 1: Multiple origins of multicellularity and their nearest neighbors

Multicellularity has evolved at least 15 times on Earth. Three multicellular groups, the multicellular fungi, plants, and animals, have evolved large, architecturally complex body plans and figure most prominently in our lives (Rokas 2008). Fossils of multicellular fungi reveal this group to have made the transition to multicellularity first, around 1.4 BYA (Butterfield 2005). Multicellular plants and animals are both thought to have originated around 500 MYA (Kenrick and Crane 1997; Rokas 2008).

The unicellular contexts in which each multicellular lineage evolved was unique and influenced the evolutionary path of each group. Evolutionary biologists study the nearest living unicellular relatives of each multicellular lineage to learn about the specific unicellular context from which each lineage evolved. The chytrids are a group of aquatic...
unicellular fungi that diverged early from the rest of the fungi. Chytrids are best known by one pathogenic species that is currently contributing to rapid declines in amphibian populations worldwide (Berger et al. 1998; James et al. 2006). The unicellular fresh-water green algae are the closest relatives of the land plants. The green algae are a diverse group of organisms with many different cell features and lifestyles, but all have chloroplasts, the subcellular structures in which photosynthesis takes place.
Box 2: Choanoflagellate colonies: a prelude to multicellularity?

Some species of choanoflagellates exist only as solitary cells (like *M. brevicollis*) while others can form aggregates or “colonies.” Choanoflagellate colonies can take on many different shapes, from round rosettes to branching colonies. Unlike true multicellular organisms, the different cells in choanoflagellate colonies are not thought to take on differentiated functions and do not become interdependent on one another. Nonetheless, some scientists suspect that a colonial state was a stepping-stone to the first true multicellular organism. Photo credits: Mark Dayel.
Chapter 1 Bibliography


Chapter 2: The genome of the choanoflagellate *Monosiga brevicollis* and the evolution of animal multicellularity

Overview

Choanoflagellates have long fascinated evolutionary biologists for their striking similarity to the “feeding cells” (choanocytes) of sponges and the possibility that they might represent the closest living relatives of metazoans (James-Clark 1868; Saville-Kent 1880). Over the past decade or so, evidence supporting this relationship has accumulated from phylogenetic analyses of nuclear and mitochondrial genes (Steenkamp, Wright, and Baldauf 2006; Medina et al. 2003; Lang et al. 2002), comparative genomics between the mitochondrial genomes of choanoflagellates, sponges, and other metazoans (Burger et al. 2003; Lavrov et al. 2005), and the finding that choanoflagellates express homologs of metazoan signaling and adhesion genes (King and Carroll 2001; King, Hittinger, and Carroll 2003; Segawa et al. 2006; Snell et al. 2006). Furthermore, species-rich phylogenetic analyses demonstrate that choanoflagellates are not derived from metazoans, but instead represent a distinct lineage that evolved before the origin and diversification of metazoans (Figure 1a, Figure A2.1) (Lavrov et al. 2005; Rokas, Krüger, and Carroll 2005). By virtue of their position on the tree of life, studies of choanoflagellates provide an unparalleled window into the nature of the unicellular and colonial progenitors of metazoans (King 2004).

Choanoflagellates are abundant and globally distributed microbial eukaryotes found in marine and freshwater environments (Buck and Garrison 1988; Thomsen and Larsen 1992). Like sponge choanocytes, each cell bears an apical flagellum surrounded by a distinctive collar of actin-filled microvilli, with which choanoflagellates trap bacteria and detritus (Figure 1b). Using this highly effective means of prey capture, choanoflagellates link bacteria to higher trophic levels and thus play critical roles in oceanic carbon cycling and the microbial food web (Leadbeater 2000; Boenigk and Arndt 2002).

Over 125 choanoflagellate species have been identified and all species have a unicellular life history stage. Some can also form simple colonies of equipotent cells, although these differ substantially from the obligate associations of differentiated cells in metazoans (Leadbeater 1983). Studies of basal metazoans indicate that the ancestral metazoan was multicellular and had differentiated cell types, an epithelium, a body plan, and regulated development including gastrulation. In contrast, the last common ancestor of choanoflagellates and metazoans was likely unicellular or possibly capable of forming simple colonies, underscoring the abundant biological innovation that accompanied metazoan origins.

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

Methods and Materials

*Immunofluorescence Staining of M. brevicollis.*
M. brevicollis cells that were grown shaking at 120 rpm to a density between 10^6 and 10^7 cells/mL were fixed by adding formaldehyde to a final concentration of 4%. We then applied approximately 0.5 mL of the fixed culture to poly-L-lysine coated coverslips and incubated for 30 minutes. After gently washing the coverslips 4 times with PEM (100 mM PIPES, pH 6.9, 1 mM EGTA, 0.1 mL MgSO₄) we blocked and permeabilized the cells for 30 minutes with blocker (PEM/1% BSA/0.3% TritonX-100) and subsequently replaced the blocker with E7 beta-tubulin primary antibodies diluted in blocker (Developmental Studies Hybridoma Bank). After incubating the cells with the antibodies for 16 hours at 4°C, we washed the coverslips 4 times with blocker, applied fluorescein conjugated donkey anti-mouse IgG (H+L) (Jackson Laboratories) secondary antibodies and incubated for 1 hr in the dark, subsequently washing 4 times with PEM. To visualize F-actin, we incubated the cells with 6 U/ mL rhodamine phalloidin (Molecular Probes) diluted in PEM. To the rhodamine phalloidin-PEM, we added DAPI at a concentration of 10 ng/ mL to visualize the DNA. We applied this mixture to the slides and incubated for 25 min in the dark. We then washed the coverslips 3 times with PEM and mounted them onto slides using 10 µL ProLong Gold antifade reagent (Molecular Probes). All steps were performed at room temperature unless specified otherwise. We took all images using a Leica DMI6000 B inverted compound microscope and Leica DFC350 FX camera at 100X magnification using oil immersion.

Separation of choanoflagellate and bacterial DNA

Using physical separation techniques combined with antibiotic treatments, a culture line with only a single bacterial food source, Flavobacterium sp., was developed. The GC content of Flavobacterium (33%) is sufficiently different from that of M. brevicollis (55%) to allow separation of the two genomes over a CsCl gradient.

M. brevicollis MX1 was grown to a density of 10^7 cells/ mL in ATCC 1525 media and 750 mL of culture was pelleted by two rounds of centrifugation at 10k rpm for 30 min at 4°C. Cell pellets were frozen at −80°C and ground to a fine powder under liquid N₂. M. brevicollis genomic DNA (at this point contaminated with Flavobacterium sp. genomic DNA) was isolated with the Puregene DNA purification system (Gentra Systems). The M. brevicollis genomic DNA was separated from the contaminating Flavobacterium sp. DNA via CsCl density gradient ultracentrifugation. Briefly, 2280 ug of contaminated genomic DNA was centrifuged to equilibrium (65K rpm for 40hrs) on six gradients of 1.69 g/ mL CsCl, in the presence of 40 µg/ mL of the dye Hoechst 33258 (Molecular Probes). The lower of two resulting bands in each gradient was recovered and the DNA was separated from the Hoechst dye by five extractions with NaCl-saturated n-butanol (Figure A2.2). The CsCl was dialyzed out of the DNA solution through Spectra/Por MWCO 8000 dialysis tubing (Spectrum Laboratories, Inc.) over 50 hours at 4°C. The purified M. brevicollis genomic DNA was rescued from the dialysis tubing and then ethanol precipitated using Pellet Paint Co-precipitant (Novagen). The final yield was 24 µg of purified M. brevicollis genomic DNA, representing a 1% recovery from the initial amount of contaminated genomic DNA. This process was repeated to obtain a sufficient amount of choanoflagellate genomic DNA to build the DNA libraries necessary for sequencing. M. brevicollis genomic DNA isolated in this manner was used to construct replicate libraries containing inserts of 2-3 kb, 6-8 kb, and 35-40 kb, each of which was used for paired end shotgun sequencing.
Genome sequencing, assembly, and validation

The ~ 41.6 Mb draft sequence of the *M. brevicollis* genome was generated from ~8.5-fold redundant paired-end whole genome shotgun sequence coverage. The initial data set was derived from 6 whole-genome shotgun (WGS) libraries: two with theoretical insert sizes of 2-3 kb, two with theoretical insert sizes of 6-8 kb, and two with theoretical insert sizes of 35-40 kb (Table A2.1).

The reads were screened for vector using Cross_match (http://www.phrap.org/phredphrap/phrap.html), then trimmed for vector and quality (Chapman 2004). Reads shorter than 100 bases after trimming were excluded. The data was assembled using release 2.9.2 of Jazz, a WGS assembler developed at the JGI (Aparicio et al. 2002; Chapman 2004; Nik Putnam 2004). A word size of 13 was used for seeding alignments between reads. The unhashability threshold was set to 40, preventing 13-mers present in the data set in more than 40 copies from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together sequences that are more than about 97% identical. The estimated fractions of bacterial clones in the main libraries ranged from 3% - 12% and sequences from these clones assembled almost entirely into a single 4.2 Mb scaffold, presumably representing the full genome of *Flavobacterium* sp. The genome size and sequence depth were initially estimated to be 50 MB and 8.0, respectively. Completeness of the draft genome was assessed by capturing ~98.5% of sequenced ESTs.

Gene prediction and annotation

Before gene prediction, the 218 scaffolds were masked using RepeatMasker (http://www.repeatmasker.org/) and a custom repeat library of 108 putative transposable elements, which are available on the *M. brevicollis* genome portal downloads page. After masking, a variety of gene prediction programs were deployed, based on a variety of methods. These were 1) the *ab initio* method FGENESH (Salamov and Solovyev 2000) (Softberry Inc., NY, USA), 2) the homology-based methods FGENESH+ (Salamov and Solovyev 2000) (Softberry Inc., NY, USA) and GeneWise (Birney, Clamp, and Durbin 2004) seeded by BLASTx alignments against sequences of all opisthokont entries in the GenBank nonredundant protein database as of May 2006, and 3) mappings of EST cluster consensus sequences from *M. brevicollis* produced using EST_map (Softberry Inc., NY, USA). EST clusters were assembled using single link clustering at 98% identity. Both the JGI ESTs and ESTs from ChoanoBase (http://mcb.berkeley.edu/labs/king/blast/) were used to assemble clusters.

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

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

Though many genes were predicted by ab initio methods, the gene catalog is supported by other evidence (Table A2.2). 90% of the predicted genes are complete models in the sense of having start and stop codons, 83% of the gene catalog aligns with proteins in the GenBank nr database (e-value < 0.1) and 56% of the predicted genes possess Pfam domains. Furthermore, 46% of the gene catalog is consistent with the ESTs collected from exponentially growing M. brevicollis.

All predicted gene models were annotated for protein function using domain prediction tool InterProScan (Quevillon et al. 2005) and hardware-accelerated double-affine Smith-Waterman alignments (http://www.timelogic.com) against Swiss-Prot (Boeckmann et al. 2003), KEGG (Kanehisa et al. 2006), KOG (Koonin et al. 2004). Then KEGG hits were used to map EC numbers, and EC, Interpro, and Swiss-Prot hits were used to map Gene Ontology (GO) terms (Ashburner et al. 2000). In addition we ran SignalP (Bendtsen et al. 2004) and TMHMM (Krogh et al. 2001) for analysis of protein localization.

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

Intron analysis

Homologs of 473 highly conserved genes from M. brevicollis and representative eukaryotes were aligned to reveal the position and phylogenetic distribution of 1989 highly reliable intron splice sites at 1054 conserved positions. The evolutionary history of introns in orthologous genes was estimated using Dollo parsimony, Roy-Gilbert maximum likelihood, and Csuros maximum likelihood (Csuros 2005; Rogozin et al. 2003; Roy and Gilbert 2005).

Analysis of signaling, adhesion and transcription factor families.

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

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

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

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

Protein identification numbers for M. brevicollis and metazoan signalling homologs

The following M brevicollis protein models were identified as homologs of metazoan

For the study of Notch and Hedgehog evolution, the following *M. brevicollis* protein models were used: (JGI protein identification numbers): *Mbrev N1*: 29255; *Mbrev N2*: 26647, *Mbrev N3*: 27644, *Mbrev H1*: 28599, *Mbrev H2*: 33852. The following metazoan protein sequence were used: (NCBI accession numbers): *Nvec Notch*: 20239, *Nvec Hh*: 241466, *Nvec Hedgling*: 200640, *Hsap Notch*: NP_060087.2, *Hsap Hh*: NP_00184.1

**Results**

**Gene structure and intron evolution**

The ~41.6 million base pair (Mb) *M. brevicollis* genome contains approximately 9200 genes and is comparable in size to genomes of filamentous fungi (~30-40 Mb) and other free-living unicellular eukaryotes (e.g., small diatoms at ~20-35 Mb and Ichthyosporeans at ~20-25 Mb) (Armbrust et al. 2004; Ruiz-Trillo et al. 2006). Metazoan genomes are typically significantly larger, with few exceptions (Seo et al. 2001).

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

Comparisons of intron positions in a set of conserved genes from *M. brevicollis*, diverse metazoans, and representative intron-rich fungi, plants, and a ciliate reveal that the last common ancestor of choanoflagellates and metazoans had genes at least as intron rich as those of modern choanoflagellates (Figure 2). Notably, these analyses reveal that the eumetazoan ancestor contained a substantially higher density of introns than the last common ancestor of choanoflagellates and metazoans. This is consistent with a proliferation of introns during the early evolution of Metazoa (Sullivan, Reitzel, and Finnerty 2006).

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

The *M. brevicollis* genome provides unprecedented insight into the early evolution of metazoan genes. Pfam and SMART annotations of the *M. brevicollis* genome identify 78 protein domains that are exclusive to choanoflagellates and metazoans, only two of which have been previously reported in choanoflagellates (Table A2.4) (King, Hittinger, and Carroll 2003). Because genomic features shared by *M. brevicollis* and metazoans were likely present in their last common ancestor, this study extends the evolutionary history of a cohort of important protein domains to the premetazoan era. Many of these domains are central to cell signaling and adhesion processes in metazoans, suggestive of a role in the origin of multicellularity. In contrast, metazoan genomic features that are missing from the *M. brevicollis* genome may have evolved within the metazoan lineage, or may have existed in the last common ancestor with choanoflagellates and were subsequently lost on the stem leading to *M. brevicollis*. Presumably there are many genomic features that evolved in the
metazoan lineage, and the *M. brevicollis* genome provides our first glimpse (albeit likely an incomplete one) at the complement of genes and protein domains that predate metazoan origins.

To further investigate the extent to which molecular components required for metazoan multicellularity evolved before the origin of metazoans, we performed targeted searches in the *M. brevicollis* genome and representative metazoan, fungal, and plant genomes for homologs of critical cell adhesion, cell signalling, and transcription factor protein families.

An abundance of cell adhesion domains

A critical step in the transition to multicellularity was the evolution of mechanisms for stable cell adhesion. *M. brevicollis* encodes a diverse array of cell adhesion and extracellular matrix (ECM) protein domains previously thought to be restricted to metazoans (Figure 3). At least 23 *M. brevicollis* genes encode one or more cadherin domains, homologs of which are required for cell sorting and adhesion during metazoan embryogenesis (Halbleib and Nelson 2006), and 12 genes encode C-type lectins, 2 of which are transmembrane proteins. While soluble C-type lectins have functions ranging from pathogen recognition to ECM organization, transmembrane C-type lectins mediate specific adhesive activities such as contact between leukocytes and vascular endothelial cells, cell recognition, and molecular uptake via endocytosis (Gupta and Surolia 2007; Yamaguchi 2000; Zelensky and Gready 2005).

The genome of *M. brevicollis* also contains integrin-alpha and immunoglobulin (Ig) domains, cell adhesion domains formerly thought to be restricted to Metazoa. In metazoans, integrin-alpha and integrin-beta domain-containing proteins heterodimerize before binding to ECM proteins such as collagen (Akiyama 1996). We find that *M. brevicollis* has at least 17 integrin-alpha domain-containing proteins, but no integrin-beta domains. Metazoan Ig domain-containing proteins have both adhesive and immune functions. The *M. brevicollis* genome encodes a total of 5 Ig domains that show affinity for either the I-set, V-set or C2-set subfamilies, but not the vertebrate-specific C1-set subfamily. In contrast to *M. brevicollis*, metazoan genomes possess from ~150 to ~1,500 Ig domains (Table A2.6), suggesting that the radiation of the Ig superfamily occurred after the divergence of choanoflagellates and metazoans.

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

ECM associated protein domains

As the targets of many adhesion receptors, the question of whether metazoan-type ECM proteins and domains evolved before or after the transition to multicellularity is of
great interest. In metazoans, collagens are ECM proteins that polymerize to form a major component of the basement membrane of epithelia and have been invoked as a potential “key innovation” during the transition to multicellularity (Erwin 1993). We find five collagen domain-encoding genes in the *M. brevicollis* genome, two of which encode the diagnostic Gly-X-Y repetitive sequence motif (in which the first position is glycine and the second and third positions are frequently proline or hydroxyproline) in an arrangement similar to metazoan collagens (van der Rest and Garrone 1991). Other ECM-associated domains known only from metazoans and, now, *M. brevicollis*, include laminin domains (an important class that contributes to the basement membrane), the reeler domain (found in the neuronal ECM protein reelin), and the ependymin domain (an extracellular glycoprotein found in cerebrospinal fluid; Figure 3 and Table A2.4) (Tissir and Goffinet 2003; Suárez-Castillo and García-Arrarás 2007).

The discovery of putatively secreted ECM proteins in a free-living choanoflagellate suggests that elements of the metazoan ECM evolved in contact with the external environment before being sequestered within an epithelium. Although some choanoflagellates secrete extracellular structures or adhere to form colonial assemblages, *M. brevicollis* is not known to do so (Leadbeater 1983; Leadbeater 1979; Leadbeater 1994). Instead, these ECM protein homologs in *M. brevicollis* may mediate an analogous process such as substrate attachment.

Against the backdrop of abundant conservation of cell adhesion and ECM domains among the genomes of *M. brevicollis* and metazoans, it is important to take note of the differences. Individual cell-adhesion and ECM-associated domains in the *M. brevicollis* genome often occur in unique arrangements and clear orthologs of specific metazoan adhesion proteins are rarely found. While the domains associated with metazoan adhesion and ECM proteins were present in the ancestor of choanoflagellates and metazoans, the canonical metazoan adhesion protein architectures (Hutter et al. 2000) likely evolved after the divergence of the two lineages.

Domain shuffling in the evolution of metazoan intercellular signaling networks

Our analysis of the *M. brevicollis* genome reveals little evidence that metazoan-specific signaling systems were present in the last common ancestor of choanoflagellates and metazoans. Many pathways are missing entirely and *M. brevicollis* genes with similarity to metazoan signaling machinery are largely found to share conserved domains without aligning across the full span of what are often complex multidomain proteins (e.g., EGF repeats are common to Notch but also to many other proteins; Table A2.7). Specifically, no receptors or ligands were identified from the NHR, WNT, and TGF-beta signaling pathways. The only evidence of the Jak/STAT pathway is an apparent STAT-like gene that encodes a partial SH2 domain. Convincing evidence is also lacking for the Toll signaling pathway, a signaling system important both for development and innate immunity in metazoans.

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

Homologs of *hedgehog* (encoding the diagnostic HINT domain), *dispatched*, and *patched* genes are also present; however, there is no evidence for *smoothed* nor its defining frizzled domain. Furthermore, a Hedgehog N-terminal peptide domain is found at the amino terminus of a large transmembrane domain protein that, instead of a HINT domain, includes von Willebrand A, cadherin, TNFR, Furin, and EGF domains. Similar proteins are found in the sponge *Amphimedon queenslandica* and the cnidarian *Nematostella vectensis*, revealing that the *M. brevicollis* genome captures an ancestral arrangement of protein domains rather than representing a lineage-specific domain shuffling event (Adamska et al. 2007). Another *M. brevicollis* Hedgehog-like protein contains a HINT domain, a key region involved in autocatalytic processing of Hedgehog (Figure 4B). The identification of a *hedgehog*-like gene in a choanoflagellate is not without precedent. A distinct HINT domain-containing protein, *Hoglet*, was identified in the distantly related *Monosiga ovata*, supporting the idea that isolated signaling components were present in the last common ancestor of choanoflagellates and metazoans (Snell et al. 2006).

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

Phospho-tyrosine (pTyr)-based signaling was considered unique to metazoans until its recent observation in choanoflagellates (King and Carroll 2001; Segawa et al. 2006). The key domains involved in pTyr signaling are found in abundance in the *M. brevicollis* genome: tyrosine kinase (TK) domains that phosphorylate tyrosine (~120 occurrences), pTyr-specific phosphatases (PTP) that remove the phosphate modification (~30), and SH2 domains that bind pTyr-containing peptides (~80). In contrast, these domains are rare outside of Metazoa; for example *S. cerevisiae* has no TKs, only three PTP domains, and a single SH2 domain. These findings support a model in which the full set of pTyr signaling machinery evolved prior to the separation of the choanoflagellate and metazoan lineages.

Although pTyr signaling machinery is present in metazoans and choanoflagellates, the mode of usage in *M. brevicollis* may be distinct from metazoans. A simple metric for the usage of a particular domain is the range of domain types with which it is found in combination (Letunic et al. 2006). In the *M. brevicollis* genome, more than half of the observed pairwise domain combinations involving TK, PTP, and SH2 domains are distinct from those seen in any metazoan genome (Figure 5 and Figure A2.7). In contrast, for other sets of common signaling domains (those involved in phospho-Ser/Thr, Ras GTP, and Rho GTP signaling) the majority of observed combinations are shared between *M. brevicollis* and metazoans. These observations are consistent with a simple model in which phospho-Ser/Thr, Ras GTP, and Rho GTP signaling were more fully elaborated prior to the branching of the choanoflagellate and metazoan lineages (consistent with the presence of these systems in other eukaryotes, including fungi, *Dictyostelium*, and plants). In contrast, simple pTyr signaling may have emerged in the common ancestor and diverged radically between choanoflagellates and metazoans.

*Streamlined transcriptional regulation*

The core transcriptional apparatus of *M. brevicollis* is, in many ways, typical of most
eukaryotes examined to date (Table A2.8) including for example, all 12 RNA polymerase II subunits and most of the transcription elongation factors (TFIIS, NELF, PAF, DSIF, and P-TEFb, but not elongin). However, homologs of the largest subunit of TFIIF and several subunits of TFIIFH are apparently lacking from the genome and the EST collection (Figure A2.4), reminiscent of the absence of several basal factors from the Giardia lamblia genome, and suggesting alternative strategies for interacting with core promoter elements (Best et al. 2004). Similarly, only a limited number of general co-activators are identifiable in M. brevicollis, including the components of several chromatin-remodeling complexes (Figure A2.4).

Perhaps not surprisingly, M. brevicollis possesses most of the ubiquitous families of eukaryotic transcription factors (Figure A2.5). The majority of the predicted transcription factors are zinc-coordinating; approximately 44% are C2H2-type zinc fingers. Eight proteins (5% of a total 155 predicted transcription factors) are forkhead transcription factors, otherwise known only from metazoans and fungi.

The homeodomain transcription factors are an ancient protein family found in all known eukaryotes. At least two major superfamilies of homeodomain proteins evolved prior to the origin of metazoans, the "typical", or non-TALE, homeodomains containing ~60 amino acids and the TALE-class homeodomains containing 63+ amino acid homeodomains (Derelle et al. 2007). The M. brevicollis genome encodes only two homeodomain proteins, both of which group with the MEIS sub-class of TALE homeodomains (Figure A2.6). Apparently, genes encoding non-TALE homeodomain proteins have been lost in the lineage leading to M. brevicollis. Bona fide HOX class homeobox genes -- a subclass of the non-TALE superclass -- are absent from both M. brevicollis and the A. queenslandica (demosponge) genome sequence reads, indicating that this characteristic metazoan gene family likely emerged along the stem leading to eumetazoans (Larroux et al. 2007).

However, M. brevicollis does possess a subset of the transcription factor families thought to be specific to metazoans. While many transcription factor families associated with metazoan patterning and development (ETS, HOX, NHR, POU and T-box) appear to be absent (Figure 3), homologs of the metazoan p53, Myc and Sox/TCF families were identified.

**Discussion**

Choanoflagellates, sponges, and other metazoans last shared a unicellular common ancestor in the late Precambrian, more than 600 million years ago (Knoll 2005; Peterson and Butterfield 2005). Although the origin of metazoans was a pivotal event in life’s history, little is known about the genetic underpinnings of the requisite transition to multicellularity. Comparisons of modern genomes provide our most direct insights into the ancient genomic conditions from which metazoans emerged. By comparing choanoflagellate and metazoan genomes we infer that their common ancestor had intron-rich genes, some of which encoded protein domains characteristically associated with cell adhesion and the ECM in animals.

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

The mechanism of invention of new genes on the metazoan stem, and their integration to create the emergent network of cell signaling and transcriptional regulation fundamental to metazoan biology, remains mysterious. Domain shuffling is implicated by the presence of essential metazoan signaling domains in *M. brevicollis* that appear in unique combinations relative to animals (Ekman, Björklund, and Elofsson 2007; Tordai et al. 2005). For phosphotyrosine-based signaling in particular, the striking divergence of domain combinations suggests that this mode of cellular interaction existed in a nascent form in the common choanoflagellate-metazoan ancestor, and was subsequently specialized and elaborated upon in each lineage.

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

The *M. brevicollis* sequence opens the door to genome-enabled studies of choanoflagellates, a diverse group of microbial eukaryotes that are important in their own right as bacterial predators in both marine and freshwater ecosystems. While *M. brevicollis* is strictly unicellular, other choanoflagellates facultatively form colonies, and the modulation of these associations by cell signaling, adhesion, transcriptional regulation, and environmental influences is poorly understood. An improved understanding of choanoflagellate ecology, in conjunction with biogeochemical studies of the Precambrian, may shed light on the intrinsic and extrinsic factors that led to the origin of metazoan multicellularity.
Chapter 2 Tables and Figures

Table 1: Features of the *M. brevicollis* genome compared to genomes from metazoans, fungi, *Dictyostelium*, and *Arabidopsis*

Figure 1: Introduction to the choanoflagellate *Monosiga brevicollis*
Figure 2: Steady gain in intron abundance preceded the origin and diversification of metazoans
Figure 3: Phylogenetic distribution of metazoan-type cell adhesion and sequence-specific transcription factor families
Figure 4: Domain shuffling and the evolution of Notch and Hedgehog
Figure 5: Divergent usage of protein domains involved in phospho-tyrosine-based signaling between *M. brevicollis* and metazoans
Table 1. Features of the *M. brevicollis* genome compared to genomes from metazoans, fungi, *Dictyostelium*, and *Arabidopsis*.

<table>
<thead>
<tr>
<th></th>
<th>Animals</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Hsap</em></td>
<td><em>Cint</em></td>
<td><em>Dmel</em></td>
<td><em>Nvec</em></td>
<td><em>Mbre</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genome size (Mbp)</td>
<td>2,900</td>
<td>160</td>
<td>180</td>
<td>356</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number of genes</td>
<td>23,224</td>
<td>15,852</td>
<td>14,601</td>
<td>27,273</td>
<td>9,196</td>
<td>13,544</td>
<td>9,826</td>
<td>13,607</td>
<td>27,273</td>
<td></td>
</tr>
<tr>
<td>Mean gene size (bp)</td>
<td>27,000</td>
<td>4,585</td>
<td>5,247</td>
<td>4,496</td>
<td>3,004</td>
<td>1,679</td>
<td>1,528</td>
<td>1,756</td>
<td>2,287</td>
<td></td>
</tr>
<tr>
<td>Mean intron density (introns/gene)</td>
<td>7.7</td>
<td>6.8</td>
<td>4.9</td>
<td>4.3</td>
<td>6.6</td>
<td>4.4</td>
<td>1.8</td>
<td>1.9</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>Mean intron length (bp)</td>
<td>3,365</td>
<td>477</td>
<td>1,192</td>
<td>801</td>
<td>174</td>
<td>75</td>
<td>136</td>
<td>146</td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>Percent Intergenic (%)</td>
<td>78</td>
<td>55</td>
<td>57</td>
<td>66</td>
<td>33</td>
<td>46</td>
<td>56</td>
<td>29</td>
<td>50</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Introduction to the choanoflagellate *Monosiga brevicollis*. (A) The close phylogenetic affinity between choanoflagellates and metazoans highlights the value of the *M. brevicollis* genome for investigations into metazoan origins, the biology of the last common ancestor of metazoans (filled circle) and the biology of the last common ancestor of choanoflagellates and metazoans (open circle). Genomes from species shown with their four-letter abbreviation were used for protein domain comparisons in this study: Human (*Homo sapiens*; Hsap), Ascidian (*Ciona intestinalis*; Cint), *Drosophila* (*Drosophila melanogaster*; Dmel), Cnidarian (*Nematostella vectensis*; Nvec), *M. brevicollis* (Mbrev), Zygomycete (*Rhizopus oryzae*; Rory), Basidiomycete (*Coprinus cinereus*; Ccin), Ascomycete (*Neurospora crassa*; Nera), Hemiascomycete (*Saccharomyces cerevisiae*; Scer), Slime mold (*Dictyostelium discoideum*; Ddis), and *Arabidopsis* (*Arabidopsis thaliana*; Atha). (B, C, D) Choanoflagellate cells bear a single apical flagellum (arrow, B) and an
apical collar of actin-filled microvilli (bracket, C). (D) An overlay of beta-tubulin (green), polymerized actin (red) and DNA localization (blue) reveals the position of the flagellum within the collar of microvilli. Scale bar = 2 µm.
Figure 2. Steady gain in intron abundance preceded the origin and diversification of metazoans. Ancestral intron content, intron gains and intron losses were inferred by the Csuros maximum likelihood method from a sample of 1,054 intron positions in 473 highly conserved genes in representative metazoans (humans, *Drosophila melanogaster*, and *Nematostella vectensis*), *Monosiga brevicollis*, intron-rich fungi (*Cryptococcus neoformans* and *Phanerochaete chrysosporium*), plants and green algae (*Arabidopsis thaliana* and *Chlamydomonas reinhardtii*), and a ciliate (*Tetrahymena thermophila*) (Csuros 2005). Branches with more gain than loss are blue, those with more loss than gain are red, and those with comparable amounts are black. The inferred or observed number of introns present in ancestral and extant species are indicated next to proportionally sized circles. As in Figure 1, the last common ancestor of metazoans and the last common ancestor of choanoflagellates and metazoans are represented by a filled circle and an open circle, respectively.
Figure 3. Phylogenetic distribution of metazoan-type cell adhesion domains and sequence-specific transcription factor families. *M. brevicollis* possesses diverse adhesion and ECM domains previously thought to be unique to metazoans. In contrast, many metazoan sequence specific transcription factors are absent from the *M. brevicollis* gene catalog. For adhesion and ECM domains, a filled box indicates a domain identified by both SMART and Pfam, a half-filled box indicates a domain identified by either SMART or Pfam, and an open box indicates a domain that is not encoded by the current set of gene models ((Letunic et al. 2006; Bateman et al. 2004). Presence (filled box) or absence (empty box) of transcription factor families was determined by reciprocal BLAST and SMART/Pfam domain annotations. Species names follow convention from Figure 1.
**Figure 4. Domain shuffling and the evolution of Notch and Hedgehog.** Analysis of the draft gene set reveals that *M. brevicollis* possesses protein domains characteristic of metazoan Notch (A) and Hedgehog (Hh) proteins (B), some of which were previously thought to be unique to metazoans. The presence of these domains in disparate peptides in *M. brevicollis* suggests that domain shuffling has occurred in these proteins since the separation of the choanoflagellate and metazoan lineages.
Figure 5. Divergent usage of protein domains involved in phospho-tyrosine based signaling between *M. brevicollis* and metazoans. A metric for functional usage of a domain within a genome is the number of other domains with which it co-occurs in a single protein. Numbers of pairwise domain combinations are indicated for classes of signaling domains involved in Ras, Rho, phospho-Ser/Thr, and pTyr signaling. In cases where a domain combination occurs multiple times within an individual protein or genome, it is only counted once. All combinations observed in *M. brevicollis* are indicated either as those that
are only observed in the *M. brevicollis* genome (magenta), or those that are observed both in *M. brevicollis* and metazoan genomes (grey). P-Tyr signaling domains in *M. brevicollis* are unique in that the majority of their observed pairwise domain combinations are distinct from those observed in metazoans.
Chapter 2 Bibliography


Neuroscience 4, no. 6 (June): 496-505. doi:10.1038/nrn1113.
Chapter 2 Appendix 1

Section 1: Supplemental Figures
Figure A2.1: Choanoflagellates are a close outgroup of Metazoa.
Figure A2.2: Separation of *M. brevicollis* genomic DNA from *Flavobacterium sp.* DNA via CsCl density gradient ultracentrifugation.
Figure A2.3: Domains significantly over-represented in choanoflagellates.
Figure A2.4: Diagrams of metazoan general transcription factors and coactivators.
Figure A2.5: Relative abundance of transcription factor families in *M. brevicollis*.
Figure A2.6: Alignment of homeodomain sequences.
Figure A2.7: Tyrosine kinase protein architectures in *M. brevicollis*.

Section 2: Supplemental Tables
Table A2.1: Genome sequencing summary.
Table A2.2: Supporting evidence for gene models.
Table A2.3: Functional classification of domains unique to choanoflagellates and metazoans.
Table A2.4: Protein domains unique to choanoflagellate and other groups.
Table A2.5: Species included in initial protein domain analysis.
Table A2.6: Immunoglobulin domains are restricted to choanoflagellates and animals.
Table A2.7: Intercellular signaling pathways across phyla.
Table A2.8: Basal transcription factors present in *M. brevicollis*.
Table A2.9: Number of *M. brevicollis* protein models containing transcription factor family-specific domains.

Section 3: Additional Notes and Methods
A3.1 Phylogenetic Analysis
A3.2 Gene structure statistics
A3.3 Protein domain content of *M. brevicollis*
A3.4 TATA-binding proteins and transcription elongation factors.

Section 4: Chapter 2 Appendix 1 Bibliography
**Figure A2.1. Choanoflagellates are a close outgroup of Metazoa.** The phylogenetic analysis of 50 genes shows that *M. brevicollis* is placed outside metazoans (including poriferans and cnidarians), and justifies its choice for comparative genomic investigations into
the transition from a unicellular to the multicellular metazoan lifestyle. (A) The tree with the highest likelihood in the maximum likelihood analyses is shown. Choanoflagellates (B) Bootstrap support values for all branches shown in A are shown. For each branch, the bootstrap support values from the maximum likelihood and maximum parsimony are shown, respectively.
Figure A2.2: Separation of *M. brevicollis* genomic DNA from *Flavobacterium sp.* genomic DNA via CsCl density gradient ultracentrifugation. The GC content of the prey species *Flavobacterium* sp. (33%) is sufficiently different from that of *M. brevicollis* (55%) to allow separation of the two genomes over a CsCl gradient. DNA purified by this method was used to build the 3-kb, 8-kb, and 40-kb genomic libraries for sequencing by the Joint Genome Institute. Photo credit: Nipam Patel.
B.

![Bar Chart](image_url)

- **Ankyrin**: H. sapiens 14, M. brevicollis 177, S. pombe 240
- **Tyrosine protein kinase**: H. sapiens 13, M. brevicollis 204, S. pombe 101
- **SH2 motif**: H. sapiens 1, M. brevicollis 101, S. pombe 0
- **PDZ/DHR/GLGF**: H. sapiens 0, M. brevicollis 133, S. pombe 59
- **EGF-like**: H. sapiens 0, M. brevicollis 158, S. pombe 40
- **Protein Kinase**: H. sapiens 13, M. brevicollis 315, S. pombe 103
- **Leucine-rich repeat**: H. sapiens 10, M. brevicollis 242, S. pombe 76
- **FG-GAP**: H. sapiens 0, M. brevicollis 35, S. pombe 24
- **Poly(ADP-ribose) polymerase, catalytic region**: H. sapiens 0, M. brevicollis 35, S. pombe 17
- **EGF, extracellular**: H. sapiens 0, M. brevicollis 34, S. pombe 0

**Number of Proteins Containing Specified Domain**
Figure A2.3. Domains significantly over-represented in choanoflagellates. Significantly over-represented domains in the choanoflagellate genome were identified by comparing the occurrence of PFAM domains excluding repeats (one hit per protein) in *M. brevicollis* to the human (panel A) and *S. pombe* (panel B) genomes. The ten most significantly over-represented domains from each comparison as determined by a Chi-squared test are shown, with the most significantly over-represented domain shown at the top of the graphs. The number of proteins containing each domain is indicated.
Figure A2.4. Diagrams of metazoan general transcription factors and coactivators.
Blue indicates subunits found in *M. brevicollis*; yellow indicates a subunit not found in *M. brevicollis*; and red indicates a possible homolog in *M. brevicollis*. A. Diagram of TFIIH. B. Diagram of TFIIID. C. Diagram of Mediator.
Figure A2.5. Relative abundance of transcription factor families in *M. brevicollis*. Of 155 protein models containing transcription factor associated domains, the percentage of protein models containing the indicated family specific domain is shown. bZip: basic-leucine zipper; E2F-TDP: E2F/DP (dimerization partner) family winged-helix DNA-binding domain; FH: forkhead; Hbx: homeobox; HLH: helix-loop-helix; HTH: helix-turn-helix; ZnF: zinc finger.
<table>
<thead>
<tr>
<th></th>
<th>Xla Meis1</th>
<th>Mbr hbx1</th>
<th>Mmu Pbx</th>
<th>Mbr hbx2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STGDIDDPDFDKKERKKRNGQFPVATN----LMRAWLFCETLPYFEEQKKCOLAQDTGLTLQVNNNFINARRRTVQP</td>
<td>HTCTDPFRPKAARRPSPKSRQHTKPPASSIDTDREWLFAETDNPYPLQKXTELQOQTGLDLQOINWFINARRRLIK</td>
<td>LDLTSLLDNEQKSKNKBRGQLPKHATN---LMRSWLEFQQLNHPYEISDEKQQAADQNLIQLQVNNNFINARRRLLQP</td>
<td>ASMDTAQLQSTRASNTGCRRNXTLEVTSPRKLREWASKTHYSEOKKRELAPLCLLTLQOINWFINARRRLKAR</td>
</tr>
</tbody>
</table>
Figure A2.6 Both homeobox genes in *M. brevicollis* are TALE class.

Two homeobox genes were identified in *M. brevicollis*. Both appear to be TALE (three amino acid loop extension)-class homeobox genes. The proline-tyrosine-proline feature is characteristic of TALE homeodomains. Xla: *Xenopus laevis*; Mbr: *Monosiga brevicollis*; Mmu: *Mus musculus*; hbx: homeobox. Homeodomain boundaries as in Bürglin, 1997 (Bürglin 1997).
Figure A2.7: A diversity of protein domain architectures in *M. brevicollis* tyrosine kinases. The majority of tyrosine kinases in *M. brevicollis* have unique domain architectures or novel domain combinations. Domains surrounded by a box are not found with metazoan tyrosine kinases. Domain name abbreviations are those used in SMART and/or Pfam (Letunic et al. 2006; Bateman et al. 2004). Domains marked with an asterisk are only predicted by one of the two domain prediction programs used although both contain the domain in their database.
Table A2.1 Genome sequencing summary.

<table>
<thead>
<tr>
<th>Library IDs</th>
<th>Theoretical insert size</th>
<th>Actual insert size</th>
<th>Raw reads</th>
<th>Raw (untrimmed) sequence (Mb)</th>
<th>Passing reads</th>
<th>Quality and vector trimmed sequence (Mb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AZSO</td>
<td>2-3 kb</td>
<td>3,061 +/- 525</td>
<td>7,620</td>
<td>8</td>
<td>6,599</td>
<td>5</td>
</tr>
<tr>
<td>BHUH</td>
<td>2-3 kb</td>
<td>2,365 +/- 355</td>
<td>295,882</td>
<td>314</td>
<td>262,757</td>
<td>185</td>
</tr>
<tr>
<td>BAFY</td>
<td>6-8 kb</td>
<td>6,593 +/- 1,284</td>
<td>7,680</td>
<td>8</td>
<td>5,457</td>
<td>4</td>
</tr>
<tr>
<td>BNUS</td>
<td>6-8 kb</td>
<td>7,059 +/- 1,769</td>
<td>242,175</td>
<td>235</td>
<td>226,029</td>
<td>165</td>
</tr>
<tr>
<td>BAFZ</td>
<td>35-40 kb</td>
<td>38,665 +/- 11,944</td>
<td>3,840</td>
<td>4</td>
<td>3,308</td>
<td>2</td>
</tr>
<tr>
<td>BIFH</td>
<td>35-40 kb</td>
<td>36,888 +/- 13,666</td>
<td>77,856</td>
<td>76</td>
<td>46,940</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>635,053</td>
<td>645</td>
<td>551,090</td>
<td>383</td>
</tr>
</tbody>
</table>
Table A2.2. Supporting evidence for genes models.

<table>
<thead>
<tr>
<th>Evidence</th>
<th>M. brevicollis v.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete models (annotated start and stop codons)</td>
<td>8286 (90%)</td>
</tr>
<tr>
<td>Models with EST alignment</td>
<td>4186 (46%)</td>
</tr>
<tr>
<td>Models with nr alignment (e-value &lt; 0.1)</td>
<td>7590 (83%)</td>
</tr>
<tr>
<td>Models with Swissprot alignment (e-value &lt; 10^{-5})</td>
<td>5877 (64%)</td>
</tr>
<tr>
<td>Models with Pfam alignment (gathering threshold)</td>
<td>5160 (56%)</td>
</tr>
</tbody>
</table>
Table A2.3 Functional classification of domains unique to choanoflagellates and metazoans.

<table>
<thead>
<tr>
<th><strong>Cell Adhesion and Extracellular Matrix</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cadherin*</td>
<td>Laminin G*</td>
</tr>
<tr>
<td>CUB</td>
<td>Laminin N-terminal</td>
</tr>
<tr>
<td>Ependymin</td>
<td>Reeler</td>
</tr>
<tr>
<td>Fibrillar collagen C-terminal</td>
<td>Somatomedin B</td>
</tr>
<tr>
<td>HYR*</td>
<td>Von Willebrand D*</td>
</tr>
<tr>
<td>Kunitz/bovine pancreatic trypsin inhibitor*</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th><strong>Cell Adhesion and Signal Transduction</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Leucine rich repeat N-terminal</td>
<td>Immunoglobulin I-set*</td>
</tr>
<tr>
<td>Immunoglobulin</td>
<td>Immunoglobulin V-set*</td>
</tr>
<tr>
<td>Immunoglobulin c-2*</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Transcriptional Control</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mib repeat</td>
<td>STAT protein, DNA binding</td>
</tr>
<tr>
<td>p53 DNA-binding**</td>
<td>Zinc finger, C2HC type</td>
</tr>
<tr>
<td>PET</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Cytoskeletal Associated</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nebulin repeat</td>
<td>Repeat in HS1/cortactin</td>
</tr>
<tr>
<td>Filament</td>
<td>Sarcoglycan complex subunit protein</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Transporters/Channels</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dihydropyridine sensitive L-type calcium channel</td>
<td>Organic anion transporter polypeptide (OATP)</td>
</tr>
<tr>
<td>Inward rectifier potassium channel</td>
<td>Progressive ankylosis protein (ANKH)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Enzymes</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspartyl/asparaginyl beta-hydroxylase</td>
<td>Galactosyl transferase</td>
</tr>
<tr>
<td>DNase C*</td>
<td>Glycosyl hydrolase family 59*</td>
</tr>
<tr>
<td>Cu2 monoxygenase</td>
<td>Heparan sulfate 2-O-sulfotransferase*</td>
</tr>
<tr>
<td>Fz2-like conserved region</td>
<td>N-acetylglucosaminyltransferase-IV conserved reg.</td>
</tr>
<tr>
<td>Galactose-3-O-sulfotransferase</td>
<td>Phosphomevalonate kinase</td>
</tr>
</tbody>
</table>

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

Present in bacteria Partial domain present in Zea mays (Qi, 2003)
Table A2.4. Protein domains unique to choanoflagellates and other groups.
<table>
<thead>
<tr>
<th>Domain Name</th>
<th>Interpro ID</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metazoa, Choanoflagellates, Fungi, and Dictyostelium</strong></td>
<td></td>
</tr>
<tr>
<td>Growth-Arrest-Specific Protein 2 Domain</td>
<td>IPR003108</td>
</tr>
<tr>
<td>Protein of unknown function (DUF1183)</td>
<td>IPR009567</td>
</tr>
<tr>
<td>Protein of unknown function (DUF1613)</td>
<td>IPR011671</td>
</tr>
<tr>
<td>Mss4 protein</td>
<td>IPR007515</td>
</tr>
<tr>
<td>UcrQ family</td>
<td>IPR04205</td>
</tr>
<tr>
<td>Diaphanous FH3 Domain</td>
<td>IPR010472</td>
</tr>
<tr>
<td>WSC domain</td>
<td>IPR002889</td>
</tr>
<tr>
<td>TAP C-terminal domain*</td>
<td>IPR005637</td>
</tr>
<tr>
<td>RasGAP C-terminus</td>
<td>IPR000593</td>
</tr>
<tr>
<td>GGL domain</td>
<td>IPR001770</td>
</tr>
<tr>
<td>Ras association (RalGDS/AF-6) domain</td>
<td>IPR000159</td>
</tr>
<tr>
<td>ILW domain</td>
<td>IPR002558</td>
</tr>
<tr>
<td>BTG family</td>
<td>IPR002087</td>
</tr>
<tr>
<td>Cysteine dioxygenase type I*</td>
<td>IPR010300</td>
</tr>
<tr>
<td>Fic protein family*</td>
<td>IPR003812</td>
</tr>
<tr>
<td>Fes/CIP4 homology domain (FCH)</td>
<td>IPR001060</td>
</tr>
<tr>
<td>GTPase-activator protein for Ras-like GTPase (Ras GAP)</td>
<td>IPR008936</td>
</tr>
<tr>
<td>RasGEF</td>
<td>IPR001895</td>
</tr>
<tr>
<td>RasGEF, N-terminal motif</td>
<td>IPR000651</td>
</tr>
<tr>
<td>Wiskott Aldrich syndrom homology region 2*</td>
<td>IPR003124</td>
</tr>
<tr>
<td>Alpha adaptin AP2, C-terminal domain</td>
<td>IPR003164</td>
</tr>
<tr>
<td>C-protein gamma like domain (GGL)</td>
<td>IPR001770</td>
</tr>
<tr>
<td>BTG domain</td>
<td>IPR002087</td>
</tr>
<tr>
<td><strong>Metazoa, Choanoflagellates, and Fungi</strong></td>
<td></td>
</tr>
<tr>
<td>Arfaptin</td>
<td>IPR010504</td>
</tr>
<tr>
<td>ATP synthase D chain, mitochondrial (ATP5H)</td>
<td>IPR008689</td>
</tr>
<tr>
<td>Cation-dependent mannose-6-phosphate receptor</td>
<td>IPR000296</td>
</tr>
<tr>
<td>CP2 transcription factor family</td>
<td>IPR007604</td>
</tr>
<tr>
<td>CybS</td>
<td>IPR007992</td>
</tr>
<tr>
<td>Cytochrome c oxidase subunit Va</td>
<td>IPR003204</td>
</tr>
<tr>
<td>D-ala D-ala ligase C-terminus</td>
<td>IPR011095</td>
</tr>
<tr>
<td>Disintegrin</td>
<td>IPR001762</td>
</tr>
<tr>
<td>Dolichyl-phosphate-mannose-protein mannosyltransferase</td>
<td>IPR003342</td>
</tr>
<tr>
<td>Epoxide hydrolase N terminus</td>
<td>IPR010497</td>
</tr>
<tr>
<td>Forkhead domain</td>
<td>IPR001766</td>
</tr>
<tr>
<td>FRG1-like family</td>
<td>IPR010414</td>
</tr>
<tr>
<td>GDP/GTP exchange factor Sec2p</td>
<td>IPR009449</td>
</tr>
<tr>
<td>Golgi phosphoprotein 3 (GPP34)</td>
<td>IPR008628</td>
</tr>
<tr>
<td>HRDC (Helicase and RNase D C-terminal) domain</td>
<td>IPR002121</td>
</tr>
<tr>
<td>Inhibitor of Apoptosis domain</td>
<td>IPR001370</td>
</tr>
<tr>
<td>Microtubule associated</td>
<td>IPR012943</td>
</tr>
<tr>
<td>Peptidase C1-like family</td>
<td>IPR004134</td>
</tr>
<tr>
<td>Protein of unknown function (DUF1349)</td>
<td>IPR009784</td>
</tr>
<tr>
<td>Putative phosphatase regulatory subunit</td>
<td>IPR005036</td>
</tr>
<tr>
<td>Receptor L domain</td>
<td>IPR000494</td>
</tr>
<tr>
<td>RFX DNA-binding domain</td>
<td>IPR003150</td>
</tr>
<tr>
<td>SURF4 family</td>
<td>IPR002995</td>
</tr>
<tr>
<td>TEA/ATTTS domain family</td>
<td>IPR000818</td>
</tr>
<tr>
<td>XPA protein C-terminus</td>
<td>IPR000465</td>
</tr>
<tr>
<td>XPA protein N-terminal</td>
<td>IPR000465</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------</td>
</tr>
<tr>
<td><strong>Metazoa, Choanoflagellates, and Dictyostelium</strong></td>
<td></td>
</tr>
<tr>
<td>Tryptophan 2,3-dioxygenase*</td>
<td>IPR004981</td>
</tr>
<tr>
<td>DUF1632</td>
<td>IPR012435</td>
</tr>
<tr>
<td>Beta catenin interacting protein (ICAT)</td>
<td>IPR009428</td>
</tr>
<tr>
<td>DUF1394</td>
<td>IPR009828</td>
</tr>
<tr>
<td>RUN domain</td>
<td>IPR004012</td>
</tr>
<tr>
<td>Doublecortin</td>
<td>IPR003533</td>
</tr>
<tr>
<td>Translocon assoc. protein, gamma subunit</td>
<td>IPR009779</td>
</tr>
<tr>
<td>Hyaluronidase 2*</td>
<td>IPR013618</td>
</tr>
<tr>
<td>DUF1736</td>
<td>IPR013618</td>
</tr>
<tr>
<td>Fascin*</td>
<td>IPR010431</td>
</tr>
<tr>
<td>IRSp53/MIM homology domain (IMD)</td>
<td>IPR013606</td>
</tr>
<tr>
<td>Survival motor neuron protein (SMN)</td>
<td>IPR010304</td>
</tr>
<tr>
<td>Spectrin</td>
<td>IPR002017</td>
</tr>
<tr>
<td>Translocon-assoc protein, gamma subunit (TRAP-gamma)</td>
<td>IPR009779</td>
</tr>
<tr>
<td>Follistatin-N-terminal domain-like (FOLN)*</td>
<td>IPR003645</td>
</tr>
<tr>
<td><strong>Metazoa and Choanoflagellates</strong></td>
<td></td>
</tr>
<tr>
<td>Antistasin family</td>
<td>IPR004094</td>
</tr>
<tr>
<td>Aspartyl/asparaginyl beta-hydroxylase</td>
<td>IPR007803</td>
</tr>
<tr>
<td>Associated with TFs and helicases</td>
<td>IPR006576</td>
</tr>
<tr>
<td>BTK motif</td>
<td>IPR001562</td>
</tr>
<tr>
<td>C1q*</td>
<td>IPR001073</td>
</tr>
<tr>
<td>Cadherin*</td>
<td>IPR002126</td>
</tr>
<tr>
<td>CBL proto-oncogene N-term, domain 1</td>
<td>IPR003153</td>
</tr>
<tr>
<td>CBL proto-oncogene N-term, EF hand-like</td>
<td>IPR003153</td>
</tr>
<tr>
<td>CBL proto-oncogene N-term, SH2-like</td>
<td>IPR003153</td>
</tr>
<tr>
<td>Collagen triple helix</td>
<td>IPR000887</td>
</tr>
<tr>
<td>Cu$_2$ monoxygenase</td>
<td>IPR0003153</td>
</tr>
<tr>
<td>CUB</td>
<td>IPR000859</td>
</tr>
<tr>
<td>Dihydropyridine sensitive L-type calcium channel</td>
<td>IPR000584</td>
</tr>
<tr>
<td>DNaselc*</td>
<td>IPR008185</td>
</tr>
<tr>
<td>Domain of unknown function (DUF758)</td>
<td>IPR008477</td>
</tr>
<tr>
<td>Domain of unknown function (DUF837)</td>
<td>IPR008555</td>
</tr>
<tr>
<td>ECSIT</td>
<td>IPR010418</td>
</tr>
<tr>
<td>Ependymin</td>
<td>IPR001299</td>
</tr>
<tr>
<td>Fibrillar collagen C-terminal</td>
<td>IPR000885</td>
</tr>
<tr>
<td>Filament</td>
<td>IPR001664</td>
</tr>
<tr>
<td>Fliotillin*</td>
<td>IPR004851</td>
</tr>
<tr>
<td>Fukutin-related</td>
<td>IPR009644</td>
</tr>
<tr>
<td>Frz-like conserved region</td>
<td>IPR006884</td>
</tr>
<tr>
<td>Galactose-3-O-sulfotransferase</td>
<td>IPR009729</td>
</tr>
<tr>
<td>Galactosyl transferase</td>
<td>IPR002659</td>
</tr>
<tr>
<td>Glycosyl hydrolase family 59*</td>
<td>IPR001286</td>
</tr>
<tr>
<td>GoLoco motif</td>
<td>IPR003109</td>
</tr>
<tr>
<td>Heme NO binding associated</td>
<td>IPR011645</td>
</tr>
<tr>
<td>Heparan sulfate 2-0-sulfotransferase*</td>
<td>IPR007734</td>
</tr>
<tr>
<td>Hormone receptor</td>
<td>IPR000536</td>
</tr>
<tr>
<td>Hormone-sensitive lipase (HSL) N-terminus</td>
<td>IPR010468</td>
</tr>
<tr>
<td>HYR*</td>
<td>IPR003410</td>
</tr>
<tr>
<td>Immunoglobulin</td>
<td>IPR013151</td>
</tr>
<tr>
<td>Immunoglobulin c-2*</td>
<td>IPR003598</td>
</tr>
<tr>
<td>Immunoglobulin I-set*</td>
<td>IPR013098</td>
</tr>
<tr>
<td>Immunoglobulin V-set*</td>
<td>IPR013106</td>
</tr>
</tbody>
</table>
Integrin alpha \hspace{1cm} IPR013519
Inward rectifier potassium channel \hspace{1cm} IPR013521
Kunitz/bovine pancreatic trypsin inhibitor* \hspace{1cm} IPR002223
L27 \hspace{1cm} IPR004172
Laminin G* \hspace{1cm} IPR001791
Laminin N-terminal \hspace{1cm} IPR008211
Leucine rich repeat N-terminal \hspace{1cm} IPR000372
Low-density lipoprotein receptor class A \hspace{1cm} IPR002172
Mb repeat \hspace{1cm} IPR004092
MOFRL family* \hspace{1cm} IPR007835
N-AcetylglicosaminyltransferaseIV(GnT-IV) conserved region \hspace{1cm} IPR006759
Nebulin repeat \hspace{1cm} IPR013998
Nine cysteines of family 3 GPCR \hspace{1cm} IPR011500
NRF (N-terminal domain in C. elegans NRF-6) \hspace{1cm} IPR006621
Organic anion transporter polypeptide (OATP) \hspace{1cm} IPR004156
p53 DNA-binding \hspace{1cm} IPR011615
Pacifastin inhibitor (LCMII) \hspace{1cm} IPR008037
PET \hspace{1cm} IPR010442
Phosphoethanolamine kinase \hspace{1cm} IPR005919
Phosphotyrosine binding (IRS-1 type) \hspace{1cm} IPR013625
Phosphotyrosine interaction (PTB/PID) \hspace{1cm} IPR006020
PHR \hspace{1cm} IPR012983
PI3-kinase family, p85-binding \hspace{1cm} IPR003113
Plexin \hspace{1cm} IPR013548
Progressive ankylosis protein (ANKH) \hspace{1cm} IPR009887
Protein of unknown function (DUF1241) \hspace{1cm} IPR009652
Raf-like ras-binding \hspace{1cm} IPR003116
Reeler \hspace{1cm} IPR002861
Renin receptor-like protein \hspace{1cm} IPR012493
Repeate in HS1/cortactin \hspace{1cm} IPR003134
S-100/IeBP type calcium binding \hspace{1cm} IPR013787
Sarcoglycan complex subunit protein \hspace{1cm} IPR006875
Selenoprotein S (SelS) \hspace{1cm} IPR009703
Seven transmembrane receptor, secretin family \hspace{1cm} IPR000832
SH3 domain-binding protein 5 (SH3BP5) \hspace{1cm} IPR007940
Somatomedin B \hspace{1cm} IPR001212
Spin/Ssty family \hspace{1cm} IPR003671
STAT protein, DNA binding \hspace{1cm} IPR013801
TNF (Tumor Necrosis Factor) \hspace{1cm} IPR006052
Translocon-associated protein, delta subunit precursor \hspace{1cm} IPR008855
Tropomyosin \hspace{1cm} IPR000533
Uncharacterized protein family (UPF0121) \hspace{1cm} IPR005344
Von willebrand D* \hspace{1cm} IPR001846
Zinc finger, C2HC type \hspace{1cm} IPR002515

**Fungi and Choanoflagellates** \hspace{1cm} IPR005109
Anp1 \hspace{1cm} IPR005545
YCII-related domain* \hspace{1cm} IPR005545

*Present in bacteria
Table A2.5. Species included in comparative protein domain analysis.

<table>
<thead>
<tr>
<th>Fungi</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspergillus fumigatus</td>
<td>Candida glabrata</td>
</tr>
<tr>
<td>Cryptococcus neoformans</td>
<td>Encephalitozoon cuniculi</td>
</tr>
<tr>
<td>Erethoptecium gossypii</td>
<td>Kluveromyces lactis</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>Schizosaccharomyces pombe</td>
</tr>
<tr>
<td>Yarrowia lipolecta</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metazoa</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Anopheles gambiae</td>
<td>Apis mellifera</td>
</tr>
<tr>
<td>Bos Taurus</td>
<td>Caenorhabditis elegans</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>Ciona intestinalis</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Macaca mulatta</td>
<td>Monodelphis domestica</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>Pan troglodytes</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>Takifugu rubripes</td>
</tr>
<tr>
<td>Tetraodon nigroviridis</td>
<td>Xenopus tropicalis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unicellular eukaryotes</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cryptosporidium hominis</td>
<td>Cyanidioschyzon merolae</td>
</tr>
<tr>
<td>Debaryomyces hansenii</td>
<td>Giardia lamblia</td>
</tr>
<tr>
<td>Monosiga brevicollis</td>
<td>Plasmodium falciparum</td>
</tr>
<tr>
<td>Thalassiosira pseudonana</td>
<td></td>
</tr>
</tbody>
</table>

Genomes of these species were used in the initial analysis of the phylogenetic distribution of *M. brevicollis* protein domains. The phylogenetic distributions of domains classified by this analysis as unique to choanoflagellates and another phylogenetic group were manually annotated using the Pfam and SMART online databases.
Table A2.6 Immunoglobulin domains are restricted to choanoflagellates and metazoans.

<table>
<thead>
<tr>
<th>Immunoglobulin*</th>
<th>Metazoa</th>
<th>Choanoflagellates</th>
<th>Fungi</th>
<th>Diclyostelia</th>
<th>Plants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hsap</td>
<td>Cint</td>
<td>Dmel</td>
<td>Mbre</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1502</td>
<td>144</td>
<td>503</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ccin</td>
<td>Noctra</td>
<td>Noctra</td>
<td>Ddis</td>
<td>Atha</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Total number of immunoglobulin (Ig)-type domains (Ig, Ig-like, Ig c1-set, Ig subtype 2, Ig v-set) predicted by SMART.
Table A2.7 Intercellular signaling pathways across phyla.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Component</th>
<th>Animals</th>
<th>Choanozoa</th>
<th>Fungi</th>
<th>Amoebozoa</th>
<th>Plant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><em>Hsap</em></td>
<td><em>Cint</em></td>
<td><em>Dmel</em></td>
<td><em>Nvec</em></td>
<td><em>Mbre</em></td>
</tr>
<tr>
<td>NHR</td>
<td><em>ROR</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Hnf4</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Err</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>WNT</td>
<td><em>Wnt</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Fzd</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Dsh</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>TGFβ</td>
<td><em>ALK</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>TGFβr</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Smad</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>NFKβ/Toll</td>
<td><em>NFKβ</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Tlr</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Tollip</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>JAK/STAT</td>
<td><em>Jak</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Stat</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Notch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Notch</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Delta</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Presenilin</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>Furin</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td></td>
<td><em>TACE</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Hedgehog</td>
<td><em>Ptc</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td><em>Hh</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td><em>Smo</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td><em>Fu</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>RTK</td>
<td><em>Rtk</em></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

A filled circle (●) indicates presence of a homolog with strong similarity. A partially filled circle (○) indicates a gene with partial similarity (e.g., contains some but not all domains diagnostic of that protein). An open circle (◻) indicates no homologs found. ROR, Retinoid-related orphan receptors; Hnf4, Hepatocyte nuclear factor 4; ERR, Estrogen-Related Receptor; Fzd, Frizzled; DSH, Disheveled; ALK, Activin-Like Kinase TGFβr, TGFβ receptor; SMAD, SMA/MAD Mothers Against Decapentaplegic; Tlr, Toll-like receptor; Jak, Janus Kinase; Stat, Dsl, Delta Serrate Lag-2, Ptc, Patched; Hh, Hedgehog; Smo, Smoothened; Fu, Fused; Sufu, Suppressor of Fused, Rtk, Receptor Tyrosine Kinase.
Table A2.8 Basal transcription factors present in *M. brevicollis*.

<table>
<thead>
<tr>
<th>Basal Machinery</th>
<th><em>H. sap</em></th>
<th><em>D. mel</em></th>
<th><em>M. bre</em></th>
<th><em>S. cer</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RNA polymerase II</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rpb1</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb2</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb3</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb4</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb5</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb6</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb7</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb8</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb9</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb10</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb11</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Rpb12</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TBP</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TBP2</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>TFIIA-L</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TFIIA-S</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>TFIIB</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TFIIE-L</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TFIIE-S</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TFIIF-L</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>TFIIF-S</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td><strong>TFIIB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XPB</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>XPD</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>p62</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>p52</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>p44</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>p34</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>cdk7</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>cyclin H</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Mat1</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>p8</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td><strong>Co-activators</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC4</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TAF1</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TAF2</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TAF3</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>TAF4</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>TAF5</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TAF6</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>Mediator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>TAF7</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>TAF8</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>TAF9</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TAF10</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>TAF11</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>TAF12</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED1</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED2</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED3</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED4</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED5</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED6</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>MED7</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>MED8</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>MED9</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED10</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED11</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED12</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED13</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED14</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED15</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED16</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED17</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED18</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED19</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED20</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED21</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>MED22</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED23</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED24</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED25</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED26</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED27</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED28</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED29</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED30</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>MED31</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
</tr>
<tr>
<td>Chromatin Transactions</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CBP(p300)</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>GCN5</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>ISWI</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>SWI/SNF</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Osa</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Elongation factors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFIIS</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>PAF-1</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td></td>
<td>NELF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>DSIF</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A2.9: Number of *M. brevicollis* protein models containing transcription factor family specific domains.

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

bZip: basic-leucine zipper; DBD: DNA binding domain; E2f-TDP: E2F/DP (dimerization partner) family winged-helix DNA-binding domain; FH: forhead; Hbx: homeobox; HLH: helix-loop-helix; HSF: heat shock factor; HTH: helix-turn-helix; PAH: paired amphipathic helix; RFX: regulatory factor X; SRF: serum response factor; STAT: signal transducer and activator of transcription; ZnF: Zinc finger.
Section 3: Additional Methods and Notes

A3.1 Phylogenetic Analysis.

A previously published 32-species, 50-gene data matrix containing metazoan, choanoflagellate and fungal species was updated with the orthologous genes from the *M. brevicollis* genome (Rokas, Krüger, and Carroll 2005). Additionally, the corresponding orthologous genes from a fungus (*Rhizopus oryzae*, phylum Zygomycota), a plant (*Arabidopsis thaliana*), and two protists (*Entamoeba histolytica* and *Dictyostelium discoideum*) were added to increase taxonomic diversity in the data matrix. Orthology was established by the reciprocal best BLAST hit criterion (Koonin 2005). Specifically, each gene from each of the additional species was considered a true ortholog if it was the best reciprocal BLAST hit with the corresponding gene in *Homo sapiens*.

All analyses were performed on the amino acid sequences. Genes were aligned with CLUSTALW (Thompson, Higgins, and Gibson 1994). Indels and areas of uncertain alignment were excluded from further analysis. Phylogenies were estimated using maximum likelihood (ML) and maximum parsimony (MP), using PHYML and PAUP*, respectively (Figure A2.1) (Guindon and Gascuel 2003; Swofford 2002). Support was assessed using bootstrap re-sampling with 100 replicates. For ML, the model of amino acid evolution utilized was estimated by PROTEST and enforced in all subsequent analyses (Abascal, Zardoya, and Posada 2005). The best-fit model for the 50-gene data matrix was WAG, with rate heterogeneity among sites (value of the gamma shape parameter alpha = 0.87) and a proportion of sites set to be invariable (value = 0.16) (Whelan and Goldman 2001). MP analyses were performed with all sites equally weighted and with tree-bisection-reconnection branch swapping. Data matrices and trees are available from the authors on request.

A3.2 Gene structure statistics.

*M. brevicollis* gene structure statistics are based on the JGI filtered models gene set. The gene structure statistics for other species were found on their respective genome browser websites: *N. vectensis*: http://genome.jgi-psf.org/Nemve1/Nemve1.home.html; *C. intestinalis*: http://genome.jgi-psf.org/Cioin2/Cioin2.home.html; *N. crassa*: http://www.broad.mit.edu/annotation/genome/neurospora/; *C. cinereus*: http://www.broad.mit.edu/annotation/genome/coprinus_cinereus; *D. discoideum*: http://dictybase.org) with the exception of *A. thaliana*, for which gene structure statistics were taken from a comparative genome paper (Town et al. 2006). Many of the *N. vectensis* gene models in the current release are incomplete, so the statistics given are based on a set of over 1,000 genes whose structures are known from full length mRNA (N. Putnam, personal communications). The estimated gene number was taken from the *Nematostella vectensis* genome paper (Putnam et al. 2007).

A3.3 Protein domain content of *M. brevicollis*.

The protein domain content of the *M. brevicollis* genome was annotated using Pfam v20 and SMART v5.1 with standard cutoff values (Bateman et al. 2004; Finn et al. 2006; Letunic et al. 2006). Two protein sets were annotated, the Monbr1_all_proteins.fasta (with
completely identical proteins removed) and the Monbr1_best_proteins.fasta. All the analysis described in the text used the Monbr1_best_proteins.fasta set.

The analysis of the phylogenetic distribution of protein domains found in *M. brevicollis* included the species listed in Table A2.5. To identify domains found exclusively in choanoflagellates and other phylogenetic groups, lists were generated using the Pfam and SMART annotations of these genomes. The lists of Pfam and SMART domains were combined using Interpro ID numbers to eliminate overlap. The phylogenetic distribution of each domain thought to be unique to a *M. brevicollis* and a given phylogenetic group was then checked by hand using the SMART and Pfam databases online. This ensured that the domain did not occur in a species outside that grouping that was not included in the initial automated analysis. The functions of domains identified as unique to *M. brevicollis* and metazoans were hand-curated.

Over and under-represented protein domains in *M. brevicollis* as compared to humans and *S. pombe* were also identified. This analysis was done using SMART’s genomic mode, to avoid over-counting domains due to redundant protein sets. Domains predicted by both SMART and Pfam were included and combined using Interpro ID numbers. The number of times each domain occurred in *M. brevicollis* was compared to its occurrence in *S. pombe* and humans. Significantly different numbers of domains were identified by the Chi-square test and ranked by their p-value. The top 200 significantly over and under represented domains were identified. Two sets of comparisons were made, the first of which counted each domain only once per protein and the second of which counted all occurrences of each domain. The top ten over-represented domains as compared to humans and *S. pombe* are shown in Figure A2.3.

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

**A3.4 TATA-binding proteins and transcription elongation factors.**

*M. brevicollis* possesses a second TATA-binding-protein (TBP) family member, suggesting a choanoflagellate-specific gene duplication that may be associated with gene regulatory diversity. In contrast to the initiation machinery, most of the known eukaryotic transcription elongation factors (TFIIS, NELF, PAF, DSIF, and P-TEFb, but not elongin) have clear homologs in the *M. brevicollis* genome.
Section 4: Chapter 2 Appendix 1 Bibliography


Chapter 3: Analysis of MbMyc

Introduction

The ability to tightly regulate cell growth and proliferation is crucial to the development and maintenance of the multicellular structures of metazoans. The transcription factor Myc alters the proliferative state of cells in response to signals from the extracellular environment and has been shown to regulate cell growth and proliferation, differentiation, and apoptosis (Grandori et al. 2000). Myc has been identified in all major animal lineages and thus was present in the founding ancestor of animals. Choanoflagellates are the closest living unicellular relatives of metazoans and the recently sequenced genome of Monosiga brevicollis has revealed the presence of Myc-family transcription factor homologs (King et al. 2008; Brown, Cole, and Erives 2008). Investigating the function of Myc in choanoflagellates promises to delineate the role of Myc before the origin of metazoans and may inform how the strict regulation of cell life and death in metazoans arose from a unicellular context.

Myc is a basic-helix-loop-helix leucine-zipper (bHLH-LZ) transcription factor and its function is shaped by interactions with closely related bHLH-LZ proteins. Heterodimerization of Myc with the bHLH-LZ protein MAX is required for binding to the canonical Myc target sequence, the E-box (Blackwood and Eisenman 1991; Blackwood, Lüscher, and Eisenman 1992). Myc activity can be antagonized indirectly by the dimerization of MAX with other bHLH-LZ transcription factors, Mxd and Mnt, thereby reducing the availability of MAX for dimerization with Myc. Myc function is also shaped by its sub-cellular localization. During differentiation and loss of proliferation, Myc is localized to the cytoplasm whereas cells in proliferative states generally exhibit localization of Myc to the nucleus, where it can perform its transcriptional regulatory activities (Vriz et al. 1992; Wakamatsu et al. 1993; Wang et al. 1997; Rumio et al. 2000; Craig et al. 1993). Myc imparts its control over cell growth and proliferation by regulating the expression of target genes required for ribosome biogenesis, cell-cycle regulation, and nucleotide and amino acid synthesis genes (Bello-Fernandez, Packham, and Cleveland 1993; Orian 2003; Bouchard et al. 2001; Hermeking et al. 2000).

Myc expression is regulated by several signal transduction pathways including tyrosine kinase (TK) signaling (Kelly et al. 1983; Barone and Courtneidge 1995). Signaling by secreted growth factors leads to an increase in c-myc mRNA levels by activating a tyrosine kinase pathway that includes the cellular tyrosine kinase Sre (Barone and Courtneidge 1995; Blanchard et al. 1985; Dean et al. 1986). Experiments in metazoan cancer cell models demonstrating that exposure to the tyrosine kinase inhibitor genistein results in down-regulation of Myc expression in a dose-dependent manner provide additional support for the connection between TK signaling and Myc regulation (Brown, Jolly, and Wei 1998; Jagadeesh, Kyo, and Banerjee 2006).

Metazoan Myc proteins are characterized by a C-terminal bHLH-LZ domain and a conserved set of N-terminal "Myc homology box" motifs that can be used to identify Myc homologs. The Myc homology boxes mediate transactivation and repression by Myc, often via interactions with chromatin modifying complexes (Cowling and Michael D Cole 2006;
McMahon et al. 1998; McMahon, Wood, and M D Cole 2000; Herbst et al. 2005; Kurland and Tansey 2008; Spotts et al. 1997; Stone et al. 1987; Kato et al. 1990). Myc homology boxes II and III (MBII and MBIII) are the most widely conserved of the Myc homology boxes and are found in vertebrate, insect, and cnidarian Myc proteins (Brough et al. 1995; Gallant 2006; Gallant et al. 1996; Hartl et al. 2010). Myc homology boxes I and IV (MBI and MBIV) are not found in all Myc proteins, although the absence of MBI in some lineages may represent a loss of the domain during species divergence (Hartl et al. 2010; Cowling et al. 2006).

Myc, MAX, and Mxd-like proteins have been found in the genomes of every major clade of metazoans for which there is genomic sequence available, including bilaterians like the mouse and the fly, cnidarians, a placozoan, and a poriferan (Gallant et al. 1996; Hartl et al. 2010; Richards et al. 2008). Human and mouse c-Myc and Drosophila dMyc are the best studied of the Myc homologs and the conservation of function between these homologs, as demonstrated by overlapping target gene sets and the ability of dMyc to substitute for c-Myc, corroborates the structural conservation observed in the domain architecture of Mycs from diverse animal phyla (Bouchard et al. 2001; Bello-Fernandez, Packham, and Cleveland 1993; Hermeking et al. 2000; Trumpp et al. 2001; Orian 2003). Myc-family proteins have only been identified in one group of non-metazoans, the choanoflagellates (King et al. 2008; Brown, Cole, and Erives 2008). The limited phylogenetic distribution of Myc-family proteins suggests that these critical regulators of metazoan development and metabolism evolved after the divergence of the fungal and metazoan lineages. Furthermore, because M. brevicollis possesses a rich and complex tyrosine kinome that includes Src-family tyrosine kinases, it is possible that the regulation of Myc by TK signaling also evolved in close proximity to the origin of metazoan multicellularity (Li et al. 2008; Manning et al. 2008; King et al. 2008; Pincus et al. 2008).

We previously reported a candidate homolog of Myc in the genome of M. brevicollis and others have reported a M. brevicollis homolog of MAX (King et al. 2008; Brown, Cole, and Erives 2008). Here we strengthen previous arguments of homology with detailed domain analyses and provide the first functional analyses of this unicellular homolog of a proto-oncogene. We demonstrate conservation of dimerization activity, cellular localization, and response to disruptions in tyrosine kinase signaling between choanoflagellate and metazoan Myc genes. Furthermore, we identify Myc, Max and Mxd-like homologs in Capsaspora owczarzaki (formerly known as Nuclearia sp.), a unicellular eukaryote that is sister to the choanoflagellate and metazoan clade. These results push the origin of the Myc, MAX, Mxd-like network into the unicellular world and demonstrate that functional characteristics of Myc proteins predate the divergence of the choanoflagellate and metazoan lineages.

Materials and Methods

Identification of Myc, MAX, and Mxd/Mnt homologs in unicellular relatives of Metazoa

The genomes of the choanoflagellate Salpingoea rosetta (formerly Proterospongia sp.) and the unicellular opisthokont Capsaspora owczarzaki were queried for homologs of Myc, MAX, and Mxd/Mnt homologs by blastn searches (Altschul et al.
1997) in the Broad Genome browser (http://www.broadinstitute.org/annotation/genome/multicellularity_project/GenomesIndex.html) using sequences of diverse metazoan Myc family members. The sequences used to query the unicellular genomes can be identified with the following NCBI accession numbers or JGI protein identification numbers: Myc: Homo sapiens c-myc P01106.1, Drosophila melanogaster diminutive NP 525062.2, Nematostella vectensis Nem32 XP 001640859.1, Trichoplax adherens Myc XP_002113957; MAX: H. sapiens MAX NP 002373.3, D. melanogaster MAX NP 649097.1; N. vectensis Nem36 XP 001641425.1; T. adherens MAX XP_002107861; Mxd-like: H. sapiens Mxd-1 NP 002348.1; Gallus gallus Mxd-like NP 001034399.1; Danio rerio Mxd-1 XP 698607.2; D. melanogaster Mnt, isoform H NP 570071.3.

Reciprocal best blast searches were used to validate the best hits (Tatusov, Koonin, and Lipman 1997). Sequences are listed in Appendix 1 Section 3. Unlike the genome of S. rosetta, which possessed only a homolog of MAX, the genome of C. owczarzaki contained homologs of Myc, MAX, and Mxd-like proteins.

Identification of Myc, MAX, and Mxd/Mnt homologs in diverse metazoans

N. vectensis and T. adherens Mxd-like homologs were identified by tblastn searches in their respective JGI genome browsers using the Mxd-like protein sequences listed in the previous section: http://genome.jgi-psf.org/Nemve1 and http://genome.jgi-psf.org/Triad1. Myc and Mxd homologs in Oscarella carmela ESTs had been reported as unpublished observations (Simionato et al. 2007), but had not been explicitly described. Using the Myc, MAX, and Mxd-like protein sequences listed in the previous section, O. carmela homologs were identified by tblastn searches in EST and RNA-Seq datasets (Nichols et al. 2006 and unpublished data). Reciprocal best blast searches were used to validate the best hits. Sequences are listed in Supplementary Material Section 3.

Modeling of three-dimensional organization of MbMyc and MbMAX conservation

The human Myc-MAX-DNA co-complex crystal structure (1NKP) was obtained from the RCSB Protein Data Bank (www.pdb.org) (Nair and Stephen K Burley 2003). Residues 353-434 of c-Myc from human (accession P01106) were aligned with residues 279-383 of MbMyc in ClustalW (Thompson, Higgins, and Gibson 1994). Residues that were conserved or similar as defined by the Gonnet matrix were highlighted in pink on the 1NKP structure in MacPyMol (http://www.pymol.org) (Gonnet, Cohen, and Benner 1992). Residues 22-103 of human MAX (accession NP_002373) were aligned with residues 35-119 of MbMAX. Conserved and similar residues were highlighted in yellow on the 1NKP structure as above.

In-vitro protein-protein interaction assay

MBP epitope-tagged MbMyc and MBP were synthesized in BL21 cells and purified over amylose columns. GST epitope-tagged MbMAX and GST were synthesized in BL21 cells and bound to glutathione-sepharose columns. Purified MBP-MbMyc or MBP were allowed to interact with the protein-coated glutathione columns in a binding buffer containing 10 mM Tris-Hcl, 150 mM KCL, 0.5 mM EDTA, 0.1% Triton-X 100, 12.5% glycerol, and 1 mM DTT. The unbound fraction (“flow through”) was saved and after
washed the beads extensively with 1X PBS, the bound proteins were eluted by boiling in SDS-PAGE loading buffer (25 mM Tris-HCl, 1% SDS, 10% glycerol, and 1% 2-mercaptoethanol). The eluent and the flow-through were separated by SDS-polyacrylamide gel electrophoresis and subjected to Western blotting by transferring the separated proteins onto a nitrocellulose membrane, and probing with a 1:500 dilution of the anti-MbMyc antibody diluted in 5% powdered milk in TBST (50mM Tris-HCl, 150mM NaCl, 0.1% Tween-20).

**anti-MbMyc antibody development**

We generated rabbit anti-MbMyc polyclonal antibodies by PCR amplifying the entire coding region of MbMyc from *M. brevicollis* cDNA using the following primers:

1) Myc Antibody Forward: 5’ CCGGAATTCATGAGCTCGTTCTATGC, which contains an EcoRI restriction site (underlined)

2) Myc Antibody pMal Reverse: 5’ GCCGGATCCAGGTTTTACGAAGGTGGCC, which contains a BamHI restriction site (underlined) as well as a stop codon (in bold).

3) Myc Antibody pGex Reverse: 5’ CCGGTGACCCAGTTTTACGAAGGTGGCC, which contains a SalI restriction site (underlined) as well as a stop codon (in bold).

We cloned the synthesized DNA into pGEX-6P-1 (Amersham Biosciences, Piscataway, NJ) and pMAL-c2X (New England BioLabs, Ipswich, MA) expression vectors using the engineered restriction sites to generate two fusion proteins: MbMyc N-terminally tagged with glutathione-S-transferase (GST) and the same peptide tagged N-terminally with maltose-binding protein (MBP). We transformed *Escherichia coli* BL21 cells to express each construct and purified the fusion proteins using affinity chromatography. Antibodies were raised in rabbits against MBP-MbMyc antigen (Covance Research Products Inc., Princeton, NJ) and we affinity purified anti-MbMyc antibodies from the rabbit serum using GST-MbMyc fusion protein.

**MbMyc immunofluorescence staining and imaging**

For Figures 4 and A3, *M. brevicollis* cultured with *Flavobacterium sp.* NK107 was grown to 1.7 x 10^6 cells per mL in choanoflagellate growth medium. Cells were pelleted by spinning for 10 minutes at 500 x g and then resuspended in a small volume of artificial seawater from which 1 mL aliquots were dispensed across 3 x 1.5 mL tubes. Formaldehyde was added to a final concentration of 4% and cells were pipetted onto poly-L-lysine coated glass coverslips to fix as they settled on the slip for 20 minutes. The 4% formaldehyde fixative was removed and the cells were fixed in 6% acetone in artificial seawater for 5 minutes. The slips were then washed with PEM (100 mM PIPES, pH 6.9, 1 mM EGTA, 0.1 mM MgSO4) and then blocked in PEM/1% BSA/0.3% Triton-X 100) for 30 minutes. The block was removed and replaced with a 1:100 dilution of MbMyc antibody in blocker and incubated at 4°C overnight. The slips were then washed three times with blocker and a 1:400 dilution of Alexa Fluor 488 goat anti-rabbit IgG (H+L) (Molecular Probes, Carlsbad, CA) secondary antibody was added and incubated for 1 hr at room temperature. The slips were then washed three times with blocker and then incubated with a 5U/100 µL rhodamine-phalloidin (Invitrogen), 10 ng/µL DAPI in PEM solution for 20 minutes. Finally, the slips were washed twice with PEM and then were mounted to glass slides in Prolong Gold with DAPI (Invitrogen) and sealed with clear nail polish.

76
Images were taken at 16-bit depth with a 100X oil-immersion objective on an inverted Leica microscope as Z-stacks. The best 1-6 slices of each stack were Z-projected with Maximum intensity and converted to 8-bit image depth in ImageJ (Rasband, W.S., ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, http://rsb.info.nih.gov/ij/, 1997-2009). The 8-bit projections were then contrast adjusted and then color and DIC overlays were cropped to regions of interest in Adobe Photoshop CS (Adobe Systems Inc, San Jose, CA).

Effect of genistein treatment on MbMyc transcript levels

*M. brevicollis* co-cultured with *Enterobacter aerogenes* was grown to a density of 1.9 x 10^6 cells/mL in natural choanoflagellate growth medium. Fourteen mLs of culture was dispensed into 10 cm plates and allowed to settle for 1.5 hrs. One mL of drug dissolved in natural choanoflagellate medium was added to a final concentration of 25 µM genistein in ethanol, or equal volumes of ethanol. Each time point was done in triplicate and 0-hr time points were taken before treatments were added. After 24 hrs of treatment, the 48-hour and 72-hour cultures were spun down at 4k rpm for 10 minutes at 25°C and resuspended in fresh media/drug mixtures and returned to the same 10 cm plate. This process was repeated at 48 hours for the 72-hour time point. At each time point, the plates were scraped, cell culture was transferred to a 15 mL conical and spun down at 4k rpm for 10 minutes at 4°C. Cell pellets were resuspended in cold 1X PBS, inverted gently, and spun again at 4k rpm for 10 min at 4°C. The washed cell pellet was then resuspended in RLT buffer (Qiagen RNeasy Midi Kit), homogenized by triturating 10 times through an 18 gauge needle and then frozen at –80°C.

The total RNA in each cell homogenate was isolated using the Qiagen RNeasy Midi Kit. Total RNA was concentrated in YM-100 microcon columns and then reverse-transcribed into cDNA with oligo-dT priming. Real-time PCR was performed with ABI Syberygreen master mix on an ABI 7300 Real-Time System (Applied Biosystems, Foster City, CA) using the following primers:

1. MbMyc QPCR F: 5' TCG TTC TGT GCC ATG TCC AT
2. MbMyc QPCR R: 5' GGA GTG AGC GCA AAA AGT GTT
3. GAPDH QPCR F: 5' TCG AGT CCA CTG GTG TCT TCA
4. GAPDH QPCR R: 5' CGT TGA CAC CCA TCA CAA ACA

Data was processed in Microsoft Excel (Microsoft, Redmond, WA) and Myc values were normalized to an internal control, glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Transcript input amounts were calculated based on standard curves performed for each primer set in each real-time PCR run. Standard curves were based on a titration of input amounts for each primer set in which 5 dilutions of cDNA (2 ng, 5 ng, 10 ng, 20 ng and 40 ng) were used as template; each dilution was tested in triplicate. Log input amounts were plotted on the x-axis and mean C_t (threshold cycle) values for each dilution plotted on the y axis, and the resulting linear graph was used to calculate the regression line (R^2 always greater than 0.94), from which relative sample input amounts were calculated.
Results

Analysis of Myc and MAX conservation in unicellular relatives of metazoans

To investigate whether the Myc- and MAX-like sequences previously identified in the genome of *M. brevicollis* (hereafter, MbMyc and MbMAX) are *bona fide* homologs of metazoan Myc and MAX, we conducted detailed studies of the domain architectures and the first functional studies of these proteins. Notably, the bHLH-LZ region of MbMyc, like c-Myc and other metazoan Myc proteins, is at the extreme C-terminus of the protein (Figure 1a) (L J Penn et al. 1990; Gallant et al. 1996). Furthermore, like metazoan MAX proteins, MbMAX is much smaller than MbMyc (136 amino acids vs. 399 amino acids; for comparison, c-Myc contains 434 amino acids and human MAX contains 160 amino acids).

While homologs of the Mxd-like transcription factors were not found in *M. brevicollis*, a search of the *C. owczarzaki* genome reveals homologs of Myc, MAX and Mxd. Key features of Myc and MAX proteins, including residues in the basic region important for recognition of the E-box, hydrophobic residues in the leucine zipper required for heterodimerization, and overall protein-domain architecture, are conserved in the *M. brevicollis* and *C. owczarzaki* homologs (Figure 1) (Nair and Stephen K Burley 2003; Landschulz, Johnson, and McKnight 1988; Ferré-D’Amaré et al. 1993).

The strongest conservation amongst metazoan Myc and MAX protein sequences can be found in their bHLH regions. The bHLH-LZ regions of MbMyc and MbMAX share 67% and 84% amino acid similarity, respectively, with the bHLH-LZ regions from human Myc and MAX (King et al. 2008). MbMyc and MbMAX bHLH-LZ regions share more similarity with metazoan Myc and MAX proteins than with non-Myc/MAX bHLH genes, represented by human MyoD (Figure 1b, see also (Brown, Cole, and Erives 2008). In addition, diagnostic residues that distinguish Myc and MAX from other bHLH transcription factor proteins (Atchley and Fernandes 2005) are found in MbMyc and MbMAX, providing further support for their homology (Figure A3.1).

To explore whether MbMyc and *C. owczarzaki* Myc contain any of the previously characterized Myc homology boxes (MBI-MBIV) that lie within the N-terminal transactivation domain, alignments were made in ClustalW with iteratively smaller segments of the N-terminal region to identify the best matches for these domains (Figure 1b). Both MBI and MBIV are absent from *Drosophila* and neither MBI nor MBIV is detected in MbMyc or *C. owczarzaki* Myc (Cowling et al. 2006; Gallant 2006). In contrast with MBI and MBIV, conservation of the MBIID domain extends from vertebrates to sponges and to a lesser degree to *M. brevicollis* and *C. owczarzaki* (Figure 1c). Like MBII, the MBIID domain is conserved from vertebrates to sponges to *C. owczarzaki* (Figure 1d). While MbMyc lacks the clear conservation of the MBIID domain seen in *C. owczarzaki* Myc (Fig. 1c), the fact that this domain is conserved between diverse metazoans and *C. owczarzaki* suggests that it was an ancestral feature that was lost in the *M. brevicollis* lineage.
By demonstrating that major diagnostic features of Myc are conserved in its unicellular homologs and that Myc's obligate binding partner and at least one of its antagonists are also present in the genomes of unicellular relatives of metazoans, we suggest that this regulatory network evolved prior to Metazoa. Our analyses of the M. brevicollis and C. owczarzaki genomes extend the origin of Myc and other members of the Myc network to well before the origin of animals (Figure 2).

Conservation of heterodimerization and nuclear localization in MbMyc

Defining the functions of MbMyc will be valuable for reconstructing the ancestral role of the Myc protein family in the lineage leading to Metazoa. We therefore tested whether MbMyc behaves like metazoan Myc proteins in two critical respects: heterodimerization with MAX and dynamic localization to the nucleus. In metazoans, Myc must heterodimerize with MAX in order to bind DNA and regulate the transcription of E-box-containing target genes (Amati et al. 1992; Blackwood and Eisenman 1991; Prendergast, Lawe, and Ziff 1991).

The sequence conservation of MbMyc and MbMAX mapped onto a Myc-MAX-DNA co-complex structure predicts that the two proteins have the capacity to form heterodimers on DNA (Figure 3a) (PDB 1NKP, Burley and Nair, 2003). This sequence-based prediction is upheld by a pull-down assay in which column-bound MbMAX (tagged with GST) retains MbMyc (tagged MBP), presumably through heterodimerization (Figure 3b). The use of different tags on MbMAX and MbMyc, as well as a negative control with the unfused GST tag incubated with MBP-MbMyc, eliminate the concern that off-target interactions contributed to the binding between MbMAX and MbMyc. Therefore, structure/function predictions and in vitro biochemical assays indicate that MbMyc and MbMAX form heterodimers.

A second important feature of metazoan Myc function is its dynamic localization, with the majority of the Myc protein pool localizing to nuclei except when cells undergo mitosis, differentiation, or loss of proliferative ability, in which cases Myc redistributes into the cytoplasm (Abrams, Rohrschneider, and Eisenman 1982; Donner, Greiser-Wilke, and Moelling 1982; Hann et al. 1983; Alitalo et al. 1983; Winqvist, Saksela, and Alitalo 1984; Vriz et al. 1992; Wakamatsu et al. 1993; Rumio et al. 2000; Persson and Leder 1984; Eisenman et al. 1985; Wang et al. 1997; Craig et al. 1993). Staining M. brevicollis cells with a polyclonal antibody generated against MbMyc reveals a similarly dynamic pattern of MbMyc localization. In most cells, MbMyc is concentrated in the nucleus (co-localizing with DAPI staining), while also being distributed diffusely throughout the cell body, including the flagellum (Figures 4 and A3.4). The ratio of nuclear to cytoplasmic MbMyc varies from cell to cell with a minority of cells containing primarily cytoplasmic Myc localization and little detectable nuclear localization.

Thus, two features of metazoan Myc function, heterodimerization with MAX and dynamic localization to the nucleus, are conserved in MbMyc and were likely characteristic of Myc in the last common ancestor of choanoflagellates and metazoans.
Tyrosine kinase signaling regulates MbMyc transcript levels

Another important question about the ancestral role of the Myc protein family is whether Myc's response to external signals of proliferation and growth is an ancestral function of this proto-oncogene. Because choanoflagellates possess both Myc-family homologs and TK signaling, it is possible that a TK-Myc network is a conserved feature of the Myc gene. To investigate whether MbMyc is regulated by tyrosine kinase signaling, we measured MbMyc transcript levels following disruption of tyrosine kinase signaling with the general TK inhibitor genistein.

MbMyc transcript levels in *M. brevicollis* cultures treated with 25 μM genistein fell significantly below MbMyc transcript levels in cultures treated with the control solvent alone (Figure 5). This result demonstrates that the effect of genistein on MbMyc expression parallels its effect on metazoan Myc expression. Furthermore, this result is consistent with the hypothesis that, like metazoan Myc, MbMyc is regulated by tyrosine kinase signaling.

Discussion

Our results demonstrate that the choanoflagellate *M. brevicollis* possesses proteins that are similar, both at the sequence level and functionally, to the Myc and MAX transcription factors found in metazoans. We also identify Myc, MAX, and Mxd-like homologs in the unicellular eukaryote *C. owczarzaki*, an independent opisthokont lineage that is most closely related to choanoflagellates and metazoans (Ruiz-Trillo et al. 2004).

Based on our finding that MbMyc expression is reduced by exposure to the TK inhibitor genistein, we hypothesize that an emergent network of tyrosine kinase signaling and transcriptional regulation was present in the unicellular ancestor of metazoans. Genistein has also been shown to inhibit proliferation of both choanoflagellates and metazoan cells, which may indicate that not only did a TK-Myc network exist in the last common ancestor of choanoflagellate and metazoans, but additionally, this network may have been a critical interpreter of external proliferation signals (Fotsis et al. 1995; King, Hittinger, and Carroll 2003). Interestingly, Src-family tyrosine kinases, the only family of tyrosine kinases shared between choanoflagellates and metazoans, have been shown to regulate *c-myc* expression (Manning et al. 2008; Barone and Courtneidge 1995; Blake et al. 2000). Whether the Src-family kinases of *M. brevicollis* regulate MbMyc expression will require further examination.

The most pressing question stemming from this work is the nature of MbMyc target genes. Myc proteins in bilaterians regulate an amazing diversity of cellular processes, many of which are fundamental to cellular physiology. To determine whether the early repertoire of Myc functions was similarly grand or perhaps was more limited in scope, will require a thorough investigation of the function of Myc proteins in early metazoans and their unicellular relatives. The large repertoire of bilaterian Myc regulation may have arisen from a simpler form of the network in unicellular organisms that was elaborated in the lineage leading to metazoans. To shed light on the repertoire of unicellular Myc target genes warrants investigation into the direct targets of MbMyc. Given the connection between cell
growth and proliferation, TKs and Myc, it is formally possible that MbMyc targets are related to mitogenesis. Indeed, based on an enrichment of E-boxes in choanoflagellate ribosome biogenesis gene promoters, MbMyc has been predicted to activate ribosome biogenesis genes (Brown, Cole, and Erives 2008).

Decades of detailed studies on the Myc transcription factor have not only enlightened biologists as to the complex biology of this proto-oncogene, but have also uncovered a myriad of wide-reaching biological phenomenon. For example, studies on Myc were the first to expose oncogenic activation by insertional mutagenesis, chromosomal translocation and gene amplification; the regulation of gene expression at the level of transcriptional elongation; and translation-independent and dependent mRNA decay mechanisms (Meyer and Linda Z. Penn 2008). Perhaps yet another biological phenomenon Myc researchers will uncover is the manner in which a unicellular protein can be co-opted into functions critical for multicellular life.
Chapter 3 Figures
Figure 1: Myc and MAX bHLH-LZ regions from metazoans and their close relatives.
Figure 2: Phylogenetic distribution of Myc, MAX, and Mxd-like family members.
Figure 3: MbMyc interacts with MbMAX.
Figure 4: MbMyc distribution is often concentrated in the region of the nucleus.
Figure 5: The TK inhibitor genistein reduces MbMyc expression levels.
Figure 1: Myc and MAX bHLH-LZ regions from metazoans and their close relatives.
A) Schematic showing relative protein size and conserved domain positions in c-myc and MbMyc. B) An alignment of the basic helix-loop-helix leucine zipper region of Myc and MAX from human (Homo sapiens), fly (Drosophila melanogaster), anemone (Nematostella vectensis), placazoan (Trichoplax adherens) and sponge (Oscarella carmela), a solitary choanoflagellate (Monosiga brevicollis), and an amoeboid symbiotic opisthokont (Capsaspora owczarzaki). Light blue highlights mark conserved residues unique to the Myc or MAX proteins and gray highlights mark residues conserved with a non-Myc or MAX bHLH family member (MyoD). Heptadic repeats of hydrophobic residues in the leucine zipper are bold and underlined and marked by an arrowhead above the sequence. The asterisks above the alignments indicate residues involved in sequence-specific recognition of the E-box DNA sequence (Nair and Burley, 2003). C) Alignments of Myc box II and D) Myc box III (also known as the “Acidic region” from the same species described in 1b. Green and purple highlights mark conserved and similar residues in Myc box II and Myc box III, respectively.
**Figure 2: Phylogenetic distribution of Myc, MAX, and Mxd-like family members.** Myc, MAX, and Mxd-like (Mxd or Mnt) proteins are found in metazoans and some of their unicellular relatives, but are not found in yeast or non-opisthokont lineages. The presence of at least one member of the Myc, MAX, and Mxd-like protein families is indicated by a colored box. The presence or absence of Myc, MAX, and Mxd-like in the species shown...
was determined by literature searchers when data were available or reciprocal best BLAST analysis when homology had not been previously assessed (see Methods). choanos = choanoflagellates. Full species names: *Saccharomyces cerevisiae; Capsaspora owczarzaki; Salpingoeca rosetta; Monosiga brevicollis; Oscarella carmela, Trichoplax adherens; Nematostella vectensis, Caenorhabditis elegans; Drosophila melanogaster; Mus musculus; Homo sapiens.*
<table>
<thead>
<tr>
<th>Mol. weight markers (kDa)</th>
<th>150</th>
<th>100</th>
<th>75</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bait</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GST</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GST-MbMAX</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B

MBP-MbMyc

A
**Figure 3: MbMyc interacts with MbMAX.** A) Similar and conserved residues between MbMyc and human c-myc (pink) and MbMAX and human MAX (yellow) are highlighted on the x-ray structure of the bHLHLZ domains of human Myc-MAX heterodimer bound to an E-box-containing DNA oligo. B) GST-MbMAX and GST, immobilized on glutathione sepharose beads, were used as bait to test for interaction with MBP-MbMyc. Bound proteins were analyzed by Western blot (shown), probed with MbMyc antibody. Only GST-MbMAX exhibited an interaction with MBP-MbMyc. Cropped blot is shown to improve conciseness of image; full-length blot is presented in Figure A3.2.
Figure 4: MbMyc distribution is often concentrated in the region of the nucleus. A) DIC image of *M. brevicollis* cells and their prey bacterium. The position of one bacterium is marked with an asterisk in each panel for reference. B) DAPI staining marks the nuclei of the choanoflagellate cells (triangles) as well as the DNA of the choanoflagellate prey bacteria. C) MbMyc staining shows that MbMyc is present throughout the cell body, including the flagellum, but is concentrated in the region of the nucleus. D) An overlay of panels B and C (DAPI in red and MbMyc in green) with phalloidin stain highlighting the collar (blue). The cell in the lower right corner is missing its flagellum, likely a result of mechanical disruption during the staining process.
**Figure 5:** The TK inhibitor genistein reduces MbMyc expression levels. In contrast with cultures grown in the presence of a control solvent (ethanol, open circles), the amount of MbMyc transcript of cultures treated with 25 μM of the TK inhibitor genistein in ethanol (filled circles) is significantly reduced. Error bars indicate standard error of the mean of three biological replicates.
Chapter 3 Bibliography


and differentiation through induction of apoptosis and regulation of tyrosine kinase activity and N-myc expression. *Carcinogenesis* 19, no. 6 (June): 991-997.


Schwinkendorf, D, and P Gallant. 2009. The conserved Myc box 2 and Myc box 3 regions


Chapter 3 Appendix 1

Section 1: Additional Tables and Figures
Table A3.1: Supporting information for Figure 2

Figure A3.1: Signature motifs of Myc and MAX proteins compared to *M. brevicollis* sequences
Figure A3.2: MbMyc interacts with MbMAX, full-length blot
Figure A3.3: MbMyc staining is specific
Figure A3.4: Second example of MbMyc distribution in cells

Section 2: Additional Methods for Figures A3.1 and A3.3
1. Signature motif analysis
2. MbMyc antibody competition assay

Section 3: Sequences supporting alignments and phylogenetic distribution of Myc, MAX, and Mxd-like proteins (Figures 1, 2, 3, and A3.1)
1. Myc sequences
2. MAX sequences
3. Mxd-like sequences
4. MyoD sequences

Section 4: Chapter 3 Appendix 1 Bibliography
## Table A3.1. Supporting references for Figure 2

<table>
<thead>
<tr>
<th>Species</th>
<th>Myc present?</th>
<th>MAX present?</th>
<th>Mxd-like present?</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saccaromyces cerevisiae</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Reviewed in (Gallant 2006)</td>
</tr>
<tr>
<td>Capsaspora owczarzaki</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>This study</td>
</tr>
<tr>
<td>Salpingoeca rosetta</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>This study</td>
</tr>
<tr>
<td>M. brevicollis</td>
<td>Yes&lt;sup&gt;1,2&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;2&lt;/sup&gt;</td>
<td>No&lt;sup&gt;3&lt;/sup&gt;</td>
<td>1: (King et al. 2008); 2: (Brown, Cole, and Erives 2008); 3: This study</td>
</tr>
<tr>
<td>Oscarellla carmella</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>This study</td>
</tr>
<tr>
<td>Trichoplax adherens</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>(Srivastava et al. 2008) and this study</td>
</tr>
<tr>
<td>Nematostella vectensis</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>(Simionato et al. 2007)</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>(Yuan et al. 1998)</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>Yes&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;2&lt;/sup&gt;</td>
<td>1: (Gallant et al. 1996; 2: Loo et al. 2005)</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>Yes&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;3&lt;/sup&gt;</td>
<td>1: (Battey 1983); 2: (Prendergast, Lawe, and Ziff 1991); 3: (Chin et al. 1995)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>Yes</td>
<td>Yes&lt;sup&gt;2&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;3&lt;/sup&gt;</td>
<td>1: reviewed by (Varmus 1984); 2: (Blackwood and R N Eisenman 1991); 3: (Ayer, Kretzner, and R N Eisenman 1993)</td>
</tr>
</tbody>
</table>
### Amino acid number in signature motif

<table>
<thead>
<tr>
<th>Myc signature motif</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
</tr>
</thead>
</table>

14 of 18 informative residues fit signature motif = 78%

| MAX signature motif | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 |
|---------------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| MbMAX               | K | R | A | X | H | N | X | L | E | R | X | R | R | X | D | H | I | K | X | X | F | X | X | L | K | R | D | T | I | P |

18 of 18 informative residues fit signature motif = 100%
Figure A3.1 Signature motifs of Myc and Max proteins compared to *M. brevicollis* sequences. The signature motif comprises the basic and helix-1 sequence components of the bHLH domain as described by Atchley and Fernandes, 2005. Residues highlighted in blue indicate a match between the *M. brevicollis* sequence and the signature motif, showing strong match for MbMyc and a perfect match for MbMAX. Sites in gray (e.g. position 4) are variable positions at which no amino acid has a frequency of >50%. Figure adapted from Atchley and Fernandes, 2005.
Figure A3.2: MBP-MbMyc interacts with GST-MbMAX, full-length blot. Full-length blot from which Figure 3 was cropped. GST-MbMAX and GST, immobilized on glutathione sepharose beads, were used as bait to test for interaction with MBP-MbMyc. Bound proteins were analyzed by Western blot (shown), probed with MbMyc antibody. Only GST-MbMAX exhibited an interaction with MBP-MbMyc. “flo thru” = flow through.
<table>
<thead>
<tr>
<th>Competition Type</th>
<th>DIC</th>
<th>MbMyc</th>
<th>DAPI</th>
<th>Merge</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 µg BSA</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>20 µg GST-MbMyc</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>5 µg GST-MbMyc</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>0.5 µg GST-MbMyc</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
<tr>
<td>No competition</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
<td>![Image]</td>
</tr>
</tbody>
</table>
Figure A3.3: MbMyc detection by indirect immunofluorescence is specific. A competition experiment in which the anti-MbMyc antibody is preabsorbed with 0 µg, 0.5 µg, 5µg, or 20 µg of GST-MbMyc or 20 µg of BSA shows that the nuclear and cell body staining exhibited by this antibody is specific to MbMyc.
Figure A3.4: MbMyc distribution varies between cells. A) DIC image of *M. brevicollis* cells and their prey bacterium. The position of one bacterium is marked with an asterisk in each panel for reference. B) DAPI staining marks the nuclei of the choanoflagellate cells (triangles) as well as the DNA of the choanoflagellate prey bacteria. C) MbMyc staining shows that MbMyc is present throughout the cell body, including the flagellum, but is concentrated in the nucleus in some cells. In other cells, the distribution is more uniform throughout the cell body. D) An overlay of panels B and C (DAPI in red and MbMyc in green) with phallodin stain (blue) highlighting the collar.
Section 2: Additional Methods for Figures A3.1 and A3.3

2.1: Myc and MAX signature motif comparisons

To find the choanoflagellate residues that correspond to the signature sequence for Myc proteins as established by Atchley and Fernandes, 2005, MbMyc was aligned to Myc proteins from chicken (P01109), human (NP002458), and fly (NP525062.2), using ClustalW (Thompson, Higgins, and Gibson 1994; Atchley and Fernandes 2005). The signature motif comprises the basic and helix-1 sequence components of the bHLH domain. Residues highlighted in blue indicate a match between the M. brevicollis sequence and the signature motif, showing a strong match for MbMyc and a perfect match for MbMAX. Sites in gray (e.g. position 4) are variable positions at which no amino acid has a frequency of >50%.

2.2: MbMyc competition assay for IMF specificity

M. brevicollis cultured with Flavobacterium sp. NK107 was grown to 4.9 x 10^6 cells per mL in choanoflagellate growth medium. Formaldehyde was added to a final concentration of 4% and cells were incubated at room temperature for 15 minutes and then pelleted by spinning at 500 x g for 5 minutes at room temperature. The fixed cells were resuspended in PEM (100 mM PIPES, pH 6.9, 1 mM EGTA, 0.1 mM MgSO4) and then pipetted onto poly-L-lysine-coated coverslips and allowed to settle for 30 minutes. The slips were then washed three times with PEM and then incubated in blocker (PEM/1%BSA/0.3% TritonX-100) overnight at 4°C. Anti-MbMyc antibody was preabsorbed with 0.5 μg, 5 μg, and 20 μg of GST-MbMyc and 20 μg BSA overnight at 4°C. Blocker was removed and then preabsorbed antibody mixes were applied for 1.5 hrs at room temperature. Slips were washed three times with blocker and then a 1:300 dilution of Alexa Fluor 488 goat anti-rabbit IgG (H+L) (Molecular Probes, Carlsbad, CA) secondary antibody and incubated for 1 hr in the dark, subsequently washing twice with blocker and 1x with PEM. To visualize DNA and F-actin, we incubated the cells with 20U/ 100 μL rhodamine phalloidin (Molecular Probes, Carlsbad, CA) and 10 ng/ mL DAPI diluted in PEM for 20 minutes in the dark. Finally, the slips were washed twice with PEM and mounted onto slides using 10 μL of Vectashield with DAPI reagent (Vector Laboratories, Burlingame, CA).
Section 3: Sequences supporting alignments and phylogenetic distribution of Myc, MAX, and Mxd-like proteins (Figure 1, 2, and A3.1)

3.1 Myc Sequences

>**H. sapiens** c-myc P01106.1
MPLNVSTNRNYLDYDGSQVPQYFYCDEEEENFYQQQQQLQPPAPSEDIIWKKFELL
PTPPLSPRRSGLCSPSYVAVTFSLRGDNDDGGSQGFSATDQLEMVTELLGGMVNZ
QSFICDPPDDEIYKIIIQQDMCWGSGSAAAAKLVEKLSYAARDKDSGSPNPARGHS
VCSTSSLQLDLSAAASECIDPSVVFYPFLNDSSPSKSCASQSSASSFSAPSDDSLSSTES
SPQGSPQPLVHLHEEPTTSDEESEQEDEEDIDEITUVEKQARKSVPAGHSSKPH
SPPLVLKRRCHVSTQHNYAAPPSTRKDYPAARKRKLDSVRVLQIQSNRKC
TSPRSSDTTEENVKRTRHNLERQRRNLKRSFFAFRALQPELENKEAPKVTVLLKAT
AYILSVQAAEEOQKLISEEDLLRKRREQLKHKLEQLRNSCA

>**D. melanogaster** diminutive NP_525062.2
MALYRSDPYSIMDDQLFSNISIFMDNDLDDMDKLLSSSTIQSDLKIEDMESVFQD
YDLEEDMKPEIRNIDCMWPAAMSSCLTSNGNGIESGNSAASSYETGAVALMVG
STNLYSAQYRQSQQSTTDNTQSNQHQHVNVSAENMPVIIKKEALADLYTVQCKRLRSGG
DKKSQIQQDEVHLPPIPGGSSLRRKRNQDIIRKSGELSGDSKYYQRPDTPHSLTDEVA
SEFRHNVLRACLVMGSNNSTLNGSDVNYIKIQRELNTGDPLPVPYIPINDVL
DVLNQHSNSTGGQQLNQLDEQQQAIDIAITGRNTVDSPTTGSDSDSDGEGPLN
FDLRRHRTSNGNSASITTNNNSNKNNKLKNNSNGMLMMHIDHTYTRCNDM
VDDGPNLTPSDDEEDVVSYTDKLPTNSCHLMLGALTGFQMAHKSIDHMKQKP
RYNNFNLITYPASSPVSANSRYPSSPISTPYQNCASSASPSYPLVSDSSVSSSSS
SSQFQIITSSSNKGRKRSSLKDPGLLISISSSSVYLPGVNNKVTSSSSMSKSRGK
KVGTSSGNTPSIIEQDQVDVAMDRNWKQRRSQGATSTSSSSVHRKDFVLGFDEADTIE
KRNNQHENMERQRRIGLKLLEALKKIPTIRKDERAPKVTNREAAKLCLQITLTQEEK
ELSMQQRLLSLQLQMRDQTALASYEMLNESRVSUG

>**N. vectensis** Nem32 XP_001640859.1
MTLVAEHLMDTFSDFSLPSPLFDFPEDGFMKKSMTSIEEDIMSDYFPPTPP
ISPGCSSIAEIDPGVPQCPDDEDDNFVAAEKEKSYFQENFDKIDLMCMWANG
AFENADKKHTAAAOQKVNTSQLKIRLCCQITPPAESTARRVSVDPAIEFPF
ADNSTDGPKLHGEDSDSEEEVDVIESAHKEGVEKAPTDADMVEPETDPILSPS
SDASTFEDFDKTETLGTGKSRKSSKYSHELSYKQGRKSSKTSNSGSGSDS
SEYETRATHNLERKRRNDLKLKFQLRDRVAELKDNERPAPKVSNLREAAKLCLQ
ITLTQEEEK

>**T. adherens** myc PID57914 XP_002113957
MAVHAEEAFSMLDFEPEYGSYMGEDDEDDNISCLDIMPTPLSPARQQYITTDSSN
YLADKLQLVQTDNLDIDSMNOTNNNGKSRLLIQDCMWANGCETDKK
NLYVNTVSAFDTPCATPPRAAEIFSTSDCVDIAPVFTYTLSDQGQQQQFVEAQSD
SEEEIDVVTVEKPNKRLSSLIELQQHVKTEDLQSPRTKRAKPSQISTKGEACSPKG
GGLSVKPDIDNVDVRATHNLERKRRNLDRYQFTRQIDIPDLEDNERAPKVNLKSTEYI
KFLKEEESKLMKETERERRKALLAKIDILSKREN

>**O. carmela** scaffold9677_GENSCAN_predicted_peptide_2
MEVEGGEALAAPSGGDGKTAKGKETEDTKDSPPGDVPAPVHKGEKGRAPIESEAS
KEAGEERSFRDDVMSRTCRASAMRNYPYYEHSCIRKHQLSLLHTLPPK
HPDVPVIRPHSAAALTNTAVGLLQLPRTKARFCFLATLCTTHILANANAVLSHCALTS
AFFDVFSAVKVARQAANYGPAVDWIKPKDIFLDMGNGSGLTAQEVTTSD
TTVTGKDSQTQSETAVSALSFPANGNEKEQVATSSDLCPVAPLVDHSHYAAP
VPTPPPAPCSVSESEEIDVDNQRAKKRAJHVAHAPRALASTRPAPPVRVTQLPP
PAKRPRVASAPACRSVSPAARSTTPRARIQLGRRTSTIKSSGGKQRRQSYPRR
RGGPGRSVNGQDESTDTEVDAGRTHMVNERMRDDGQRNMFGVRLCVPDIKDNE
KSPKVLKATECVMKLTADEERLRELEYEQQRFRLKFLLTALNQQSERLQ

>M. brevicolli MbMyc XP_001747958
MSSFCAMSMSLDHSLPGMKNFSFSLPGQHFLRSLDLTTTNSHTDRLRPSVPDE
DDPFEFGARISASLRELNSVMDEILGVSPEEKTPSSVTLHDDGFSGFAADPDSPEDN
LSFLVHNRDPLGLSATAEELDANPLTPFDNPFPFPFTFHQQDDHNPS
AKELLSPFDLDHMNNGNGSYFLECNGNVPEAGDILAADVVFVQAGETDSAY
NSSPASPPKHPSPRLDGSPLRKNGRNFDDGRKVKVPSPKASSANRISNNVRRVS
TTSTTSTPTRTNQYRGMSQRDAHAERERRVQLRENFEALRAEVPRLRDADKAA
TLQILRENAIKYRIKREDEQLLEEQAQLIAINDSRLKKATVHDDDSVPPPLAPP

>C. owczarzaki Myc_g599.t1
MLAALDDALSDQADAASELASAAALAADAGDHVLLTAAQLDAFSSHHPEDHTLPWD
DTALRSISAGMDALANLPDPPAAAQAWEHSHADLFPOPLLSHHDLWAGLNN
DDMDSDMHMLGVSAAADAMDSDDLAFMQHHDEASHALGALAA
LANSDPSNQHHHHQAYSSPAQSGLVLTLKKTQHKSHKSTPSTPTVPTTRRKSTRLTR
ITAPATSPSKAQRMKRM2AAAAAALARANTLSGVSVDMMDEDVEELEGEDTETRTL
KVEPALAQAQLAQQQQQFSGSYSAAGTPLASPMARHRGGGAATAAHLGVPLAQFQG
SSGTPQRPSRPAHDEDEDEEIDVVDGDESDDDTAAQLPQRRPADSAANGRGPVRS
NLVIPQPAFASPAAPIFAQAXASSYAAAAGSTGLTRPMTMNLLNPQRTQSAPSSPQPSP
SAVATAARKKNEQVTPSLMVSTPKRASVENVAAAAAAA1ATTTA
AIAAGGMSTPTGTNPRSMPSATSMSPSTMPSSTLNPSTPSPTQVFNEDDSMGEDD
ATDSVAKRRFHNVLERRRRDLNKSFIMLQKIPPELVDTSAAAKVHLRKAAYIERS
LCDREVELEAAKARALREHDELVARLQALRGSR

3.2 MAX Sequences

>H. sapiens MAX NP_002373.3
MSDNDDIEDEEQPRFQAADKRAHNNALERKRRDHKDSFHLRDSVPSCQGE
KASRAQILKATEYIQMHTQIDDRLKRQNLLEQQVRALEKARSSAQL
QTNYPPSNDLXYN2AKG7ISAFGDSGDSSESEEPQPSRKKLRMEAS

>D. melanogaster MAX NP_649097.1
MSMSDDDRDIDIESDEGDSDTGLGSSRHNTNFTQAEKRAHNNALERRRRDHDK
ESFTNLRLEAVPTLKGAKSRAQILKKTTECQTMRRKISENQDIEEKIRQNNIIAKQI
QALESSNGDQFSEFLSDEEVGSEEADDEDLDQDFSRRKKMKTFHA

>N. vectensis Nem36 XP_001641425.1
ADKRAHNNALERKRRDHKDSFHLRDSISLQGEKASRAQILNKATDYIQFMRRK
NHSHQTDIDDLKRQNLILDQQGMYWV

>T. adherens MAX_PID 51792
MSDEKYLVDIDDDNGDSDKSTSLQADKRAHHNALEKRRDHDKCIDCFGLR
DSVPTLQGEKASRAQILNATDYIQFMKQKNQNHQSDIEDIRKENYQLEQLKTLQ
RTNNTGTATSENIDSSTTTTNSGRTRRNKAKRELQSGNDQKTDTKKVKE

>O. carmela Contig 20508
MSDFDDDRDIDVDENFDDKRRHNALEKRRRDNIDCFTTLRDAVPSMAGEKA
SRAQVLNKRTEYIELSRKTEDHERIDELKRQNdSLQQQIESLERSRGNRVSDE
RNILNQIQAETLEDPDRPRQEN

>M. brevicollis MbMAX XP_001746778
MFCYHLLSLKFRALTSSMDIVPLGLFGARSGKADKRAHHNALEKRRDHDKS
FTMLRDHTPNSINGEKVQSARQLNATDYIQYMKRNQAHHQAEEMEELRNQNASL
QMOSQMQRHKSEEDGKPHVFO

>C. owczarzaki CoMAX MAX_8021_g
MEENTISQGEDYSDMDVPDSGRHVKAHHNALEKRRDHKEFKNELRDT
VPISGDKASRSLILNEFIVTMTQRNTAHEAIDAIRKQNETLRLQILDLENHS

3.3 Mxd-like sequences

>H. sapiens MAX dimerization protein 1 NP_002348.1
MAAAVRMNQIQLLEAADDYRDERRDEAEHYASMLPYNNKDRDLRKRNKSSKN
NSSSRSTHNMERKNNRAHLRLCEKLKGLVPLGPESSRHTTLSLUTKALKHIDKLED
CDRAVHQQDQLQREQQRHLKQRLKGPLGERRMDSGTVSSRSSDREEDVDVES
TDYLTGDLWSSSSVSDDERGSMQSLGSDEGYSSTSIIKIFKQDISHKACLGL

>G. gallus LOC426299 NP_001034399.1
MAAVGDDGSGVGSISGLILAEAEYLERRRERTGTPPPFRHLPPPAARSRCVPTE
AEHYASLLPPAGARRERIPRPGPKGRGSGRGGRSTHNMERKNNRAHLRLCEKL
KVLVPLGPESSKHTTLSLMMKAKLHIIKLEDYDRALHIEQQLQREQQRHLKQLEK
LG3RVTDSVGASVSSSDSDDRGETCSTLDSRWSH

>D. rerio XP_698607.2
MTAIEKLQMLIEAAAYLDERREREAEHYASLPFTSNKRDGLKRKISKKNCSRR
STHNMERKNNRAHLRLCEKLKSLVPLGPESSNRHTTLSLMMRAKEHJKLEDSEJK
AQHTIDQLQREQQRHLRRRLCEQLQVERTRMSMSTISSDKSDDQQEVDVDEGTD
YLLGDLEWSTSSSVSDDERGSLRSSCSDEGYSASSSRLRANSQENSLVACSL

>D. melanogaster Mnt, isoform H NP_570071.3
MSGIGGIGVLEAAEYLDRERERAEHYASLPFTSNKRDGLKRKISKKNCSRR
STHNMERKNNRAHLRLCEKLKSKLVPGLPSNMRHTTLSLMMRAKEHJKILEDSEK
AQHTIDQLQREQQRHLRRRLCEQLQVERTRMSMSTISSDKSDDQQEVDVDEGTD
YLLGDLEWSTSSSVSDDERGSLRSSCSDEGYSASSSRLRANSQENSLVACSL

LLAAATSATATATATATATAAATASNPAKRKLSSIAEIPSFLVVSDAKGVYSAHT

108
>N. venteresisc Mxd-like Xp_001636672
MCKSPDGCFPRTRSPRLVLFLFHQFVILPTGQHAVFVFNIDDRRRVSESADALRRR
RPQGAGTRETHNKLEKNNRAHLKCECFDVLKREVTLEDKTSLNLNLRSALKHIQIL
KQEREYENYEQGLQLNKRARERDALQFLIAERKLPNLVSKVVPQGDAKPPS
CTTTATAASCPVSSTQTPSPSPEGVCTCSTCSENGAVSTCMTDQIEENHASSD
NDTKITYNEWSKLEDNDDEKTEDGEHDLHEDERIEIDEVQGDVIGENNHV
MVQEVIESQVDQCDTVLKGSTTPYANAGRRKRRSADKREEDSDYDGDITKEEP
KTQIACARVSS

>T. adherens Xp_002117396
MTLNIQLANRSTTHQLEKRARRAHLRRTCMESRLLSSVPALNKGKATTLLSLEDGARDYI
EALKTSASKTVQTEKLKREQLRKRISDLEQVNALGGGQYVQPKWASKDNEE
ENIEVDVSSEDDCALLSFGSGDGAADNRI

>O. carmela Mxd-like
Rndl_Full_Contig_13123MSLHTLLEAAHFVDTDSDEETKKNKQQLSLVLTAP
PELDOIQVGSKVESSGAPATSGPAAPSRSSRRQRAQIVGAGVTRETHNQLEKNR
RAOLKTYFENLQOLPLGLWDKTSNLKSGALDCIEELKRDREHETNQL
RRQKRMRKQEQLKNRLRANRILREHSGGGETMLQDDHVLSTTSTS
SVTELQPAAYASSISGKKWHSHIPEPQPGSSNTTQASVVVNCSPDPDPLV
S DQNLES

>C. owczarzaki CAOG_03883_1
MAALLQAAELLQREERAAQHFMQGSGGMMPPPPAPSNQRFAPGAPMPMASPQPH
HPMMMHAPHPGAPMDPGYHAGYQHPHTHNPARTARSQSPPPPQLAYSSLGFHAAP
PRTMTPSQQMPAPMPHHQFQHPQMYPVAPPANHMYMPPGTMQGSSPGKSLGPITGA
AARRRRRNSMRSFDSYEDVYEDVEDVEDEDDTDEGYNDDDDQFQVSAR
RTPPAPARSGSRTMSSTFSSESGRNNSEGFPQQSPSYPGPFAAHGHPATMSPASNSNG
SSGATSSGLKANNSSGTSGLPDAASPQGSSGRGSKRSVKREQQHEHMNKMN
RAHLKTFCEEQLAPGLDKMAKPSVIALQHAKQFIELTHYQAQQAIDRLQQLN
VWCCSKLDAETHPVPAAYLQQPMQGKAPQPQHMHPAGEENGYYESHHHHSHH
HEHEGHQDMESSTSSSTTPSDKSTDVVERLDNSESPEPAHGTFTSGRSTGRQRE

3.4 MyoD sequence
> H. sapiens MyoD CAA40000.1
MELLSPPLRDVDLTAPODGSLCSFATTDDFYDDPCFDSPDLRFEDLDPRLMHVGALL
KPEEHSFPAAVHPAPGAREDEHVRAPSGHHQAGRCLLWACKACKRKTTTNADRR
KAATMRERRRLSKVNEAFETLKRCTSSNPNQRLPKVEILRNAIRYIEGLQALLRDQD
AAPPGAAAFYAPGPLPPGRGGEHYSGDSDASSPRNSCSDGMDYSGPPSGARRRNC
YEGAYYNEAPSEPRPGKSAAVSSLDYLSSIVERISTESPAAPLLADVPSESPRRQE
AAAPSEGESSGDPTQSPDAAPQCPAGANPNPIYQVL

110
Section 4: Chapter 3 Appendix 1 Bibliography


Chapter 3 Appendix 2: Preliminary evidence for MbMyc autoregulation

Section 1: Preliminary evidence for MbMyc binding to the MbMyc coding locus
1. Chromatin Immunoprecipitation of MbMyc
2. Electrophoretic mobility shift assays to test MbMyc binding to the MbMyc locus
3. Table and Figures to support sections 1.1 and 1.2

Table A3.2: Summary of MbMyc ChIP-Seq run data

Figure A3.5: Enrichment of MbMyc binding at the MbMyc locus
Figure A3.6: MbMyc and MbMAX bind to a 213bp region of the MbMyc coding locus
Figure A3.7: MbMyc interacts with 50bp segments of the MbMyc coding locus

Section 2: Sequences of regions described in Appendix 2
1. MbMyc ChIP peak sequence
2. 213 bp EMSA probe
3. 50bp EMSA probes

Section 3: Chapter 3 Appendix 2 Bibliography
Section 1: Chromatin Immunoprecipitation of MbMyc

One of the ways I attempted to address MbMyc's function was to identify the subset of *M. brevicollis* genes regulated by MbMyc. Broadly speaking, animal Mycs positively regulate genes involved in cell growth, ribosome biogenesis, protein synthesis, and metabolism and negatively regulate gene involved in cell cycle arrest, cell adhesion, and cell-cell communication (Dang et al 2006; Eilers and Eisenman, 2008). To identify which genes in *M. brevicollis* are regulated by MbMyc, I attempted to develop the ChIP-Seq technique (Chromatin Immunoprecipitation followed by high-throughput sequencing) in *M. brevicollis*.

My first attempt at this technique was plagued by an abundance of contaminating sequences in the final Illumina libraries (Table A3.2). The majority of the Illumina-sequencing reads that resulted from my MbMyc-ChIP library were neither *M. brevicollis* DNA nor *Flavobacter* DNA (*Flavobacter* sp. NK107 is the feeder bacteria of the choanoflagellate culture used in this ChIP experiment). I used the Velvet program (Zerbino and Birney. 2008) to build contigs from the short Illumina sequences. Twenty-six contigs were obtained from the Myc ChIP-Seq reads and 1171 contigs were built from the Input DNA reads. I then used BLAST to compare the sequences against the *Flavobacter* sp. NK107 genome (which was obtained as the largest scaffold from the *M. brevicollis* genome project) and the NCBI nr database. Of the 26 contigs built from MbMyc ChIP-Seq reads, 25 were identified as cloning vector by BLAST (one contig contained only guanine residues). Of the 450 Input DNA contigs I inspected, 439 were identified as *Flavobacter* sequences, 2 were identified as cloning vector sequence, and 9 did not have a clear match. Of the three contigs built from the preimmune ChIP-Seq reads, one was identified as *M. brevicollis* and the other two consisted of single nucleotide repeats.

This contamination meant that insufficient choanoflagellate sequences were obtained to analyze MbMyc's binding on a genome-wide scale. However, I was able to detect one region of enriched binding (Figure A3.5). I used the histogram tool of Excel (Microsoft, Redmond, WA) to look for any enriched areas across the 20 largest scaffolds of the *M. brevicollis* genome within the reads that did match to the *M brevicollis* genome (this initial mapping is included as part of the Illumina sequencing service). Only one region of the genome showed an enrichment of the MbMyc ChIP-Seq data. This peak lies within the MbMyc gene locus itself (Scaffold 19:443500-444650). No enrichment was detected in the Input DNA reads nor the preimmune ChIP-Seq reads at this locus. Only one other peak of enrichment was detected out of the three Illumina datasets. A peak of enrichment was identified on Scaffold 17 in the Input DNA sequence set. This peak lies in between positions 400000-402000, which corresponds to the Actin gene locus. Other preliminary data from the King lab suggests that this locus may contain duplications of the actin gene, which may explain why this peak of enrichment was seen.

Interestingly, Myc has been shown to negatively autoregulate its own transcription and the mechanisms of this regulation occurs through Myc protein binding within the Myc coding region (Facchini et al, 1997). In animal cells, high levels of Myc protein lead to the down-regulation of Myc transcript (Adams and Cory 1985; Cleveland et al. 1988; Grignani
et al. 1990; Lombardi, Newcomb, and Dalla-Favera 1987; Penn et al. 1990). This autoregulation is conserved between mammals and flies (Goodliffe, Wieschaus, and Cole 2005). The precise mechanism of this autoregulation is unknown, but requires direct binding of Myc to its own promoter which lies within its coding region (Facchini et al. 1997).

The presence of feeder bacteria in choanoflagellate cultures raises problems with many biochemical techniques involving *M. brevicollis*. It may be that an abundance of bacterial DNA prevents specific isolation of protein-DNA complexes with the anti-MbMyc antibody. It may also be that I was attempting to perform ChIP at inappropriate times in the choanoflagellate culture's growth phase. Shiue et al 2009 showed that c-Myc binding is only seen in serum-stimulated Rat1 fibroblast cells and not in serum-starved cells (Shiue, Berkson, and Wright 2009).
1.2: Electrophoretic mobility shift assays to test MbMyc binding to the MbMyc locus

The limited results of our ChIP-Seq experiment are so far consistent with MbMyc possessing functional similarity with animal Mycs. To further test this interaction, I attempted Electrophoretic Mobility Shift Assays (EMSAs) with recombinant GST-MbMyc and MBP-MbMAX.

I tested a 213 bp region of the 1150 bp Myc peak to start. The 213 bp region was PCR amplified from *M. brevicollis* genomic DNA and labeled with [α-32P] dCTP by random priming reaction with Klenow fragment (NEB) and purified with the Qiagen Nucleotide Removal kit. I tested the ability of 250 ng Myc, 250 ng MAX, and 250 ng Myc and 250 ng MAX to interact with this rather long probe. I saw an interaction between the presumed GST-MbMyc-MBP-MbMAX heterodimer with a subset of the peak seen in the ChIP-Seq experiment (Figure A3.6) (An attempt to use a probe encompassing the entire 1150bp peak was not successful.)

To identify a more specific locus for MbMyc-MbMAX binding in what seemed to be a positive result with the 213 bp probe, I used five double-stranded oligos that tiled across the 213 bp region in 50 bp windows. The five tiling subpeaks were ordered as single-stranded oligos from IDT and then labeled with gamma-32-ATP by T4 PNK (NEB). Complementary labeled strands were then combined, purified by Qiagen Nucleotide Removal kit, and then annealed by heating at 65°C for 10 minutes and letting the samples come slowly down to room temperature by leaving them in the heatblock after it had been turned off. Four of the five shorter probes seemed to be interacting with GST-MbMyc alone, and not MBP-MbMAX or the GST-MbMyc-MBP-MbMAX heterodimer (Figure A3.7). These interesting but confusing results require further exploration. See Appendix 2 Section 2 for all EMSA probe sequences.
Section 1.3: Table and Figures to support Sections 1.1 and 1.2

<table>
<thead>
<tr>
<th>Library</th>
<th>M. brevicollis sequences</th>
<th>Flavobacter sequences</th>
<th>Other sequences</th>
<th>Total reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>MbMyc ChIP</td>
<td>12 %</td>
<td>6 %</td>
<td>80 %</td>
<td>5,342,257</td>
</tr>
<tr>
<td>Input DNA</td>
<td>8 %</td>
<td>70 %</td>
<td>12 %</td>
<td>9,042,634</td>
</tr>
</tbody>
</table>

Table A3.2: Summary of MbMyc ChIP-Seq run data. The ChIP-Seq run was fraught with contaminating sequences that did not match the choanoflagellate genome nor the genome of the prey bacterium, *Flavobacter* sp. The contaminating sequences are largely comprised of cloning vector fragments, which I suspect contaminated the samples during preparation of the Illumina libraries.
Figure A3.5: Enrichment of MbMyc binding at the MbMyc locus. An imperfect Chip-Seq experiment revealed an enrichment of MbMyc binding at the MbMyc coding locus, suggesting that MbMyc may autoregulate its own expression.
Figure A: Enrichment of MbMyc binding (reads span 1150bp)

Figure B: MbMyc coding region, 213bp EMSA probe

Table: GST-MbMyc, MBP-MbMAX cold probe

- GST-Myc: MBP-MAX: DNA complex
- non-specific binding
Figure A3.6: MbMyc and MbMAX bind to a 213bp region of the MbMyc coding locus. A) Schematic of the MbMyc genomic locus. A ChIP-Seq experiment performed with an antibody against MbMyc identified a 1150 bp region of enriched MbMyc binding at the MbMyc genomic locus. An EMSA probe was designed to a 213 bp portion of the 1150 bp enrichment. b) An EMSA shows the GST-MbMyc: MBP-MAX heterodimer interacts with the 213 bp portion of the MbMyc locus enrichment. Non-specific binding is not competed away by cold probe.
Figure A3.7: MbMyc interacts with 50bp segments of the MbMyc coding locus. A) Schematic of dsOligos used to tile across the larger 213 bp probe shown in Figure A3.6. B) An EMSA shows that GST-Myc interacts with dsOligos 1-4 but does not interact with dsOligo 5.
Appendix 2 Section 2: Sequences of regions described in Section 1

2.1: Sequence of the MbMyc ChIP-Seq peak of enrichment
>scaffold_19|443500|444650 (1151 bp)
GGTGCCGCAAGGAGGAGGCACCTGACTGTCATCGTATGACCGACACAGTGGC
TTTCTTGCGTACGTGTTGATGAGCTGCTGCTCCTCGGAGAGTT
TGTGTTTCTTTCTTCATCGCCAGTGGCTTGTATGATCGTGTTGCCTCGCGAAGGATCTGTAAGGGTGAGGCACTGCTCAGCGTCGCTGAGGGAACCTCGGCGCGAAGAGCCTCAAGTTCTCCCTCAGCTGC

2.2: MbMyc peak EMSA 213bp EMSA probe
>scaffold_19|443527|443739 (213 bp)
CTGTCATCGTCATCGTGAGGGCTTCTCTTCTTTCTGCTAGACTGCTGCTGGTTGAGCTG
TGAGTGAGCCTTCTCCTCGAGGAGTTGTTTCTCTTCATCACGGATGCGCTTGGATTCTGATGACAGTTCTGATG

2.3: 50bp subpeak tiling oligos
>Myc_subpeak_1
CTGTCATCGTCATCGTGACGACAGTGCGCTTCTTCTGCTAGACTGCTGCCTGGATTCTGATGACAGT

>Myc_subpeak_2
GTGAGTGCGAGTGGCTTCTCTCGAGGAGTTGTTTCTTTCTGCTTACACGGTAGGCGCTTGGATTCTGATGACAGT

>Myc_subpeak_3
TCTTCTCAGCAGGGATGGCGCTTCTTGAGTCTATGCGTGCTGAGGGCTTGGATTCTGATGACAGT

>Myc_subpeak_4
CTGTCGAGAACCGCGCGGCGCAAGAGCCTCACAAAGGTTCGGGAGCGCTGAGGGCTTGGATTCTGATGACAGT

>Myc_subpeak_5
AATGAGGGAACCTCGGCGCGAAGAGCCTCAAAAGGTTCGGGAGCGCTGAGGGCTTGGATTCTGATGACAGT

123
Section 3: Chapter 3 Appendix 2 Bibliography


