

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

The gene content of diverse choanoflagellates illuminates animal origins

Permalink

<https://escholarship.org/uc/item/7xc2p94p>

Author

Richter, Daniel Joseph

Publication Date

2013

Peer reviewed|Thesis/dissertation

The Gene Content of Diverse Choanoflagellates Illuminates Animal Origins

by

Daniel Joseph Richter

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nicole King, Co-chair

Professor Michael B. Eisen, Co-chair

Professor W. Zacheus Cande

Professor Yun Song

Spring 2013

Abstract

The Gene Content of Diverse Choanoflagellates Illuminates Animal Origins

by

Daniel Joseph Richter

Doctor of Philosophy in Molecular and Cell Biology
and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Nicole King, Co-chair
Professor Michael Eisen, Co-chair

Fossil data indicate that the seminal events in animal evolution occurred over 600 million years ago, yet the first animals were likely microscopic and soft-bodied, providing little fossil evidence for their biology and morphology. To bridge the gap in our understanding of early animal evolution, we can compare the gene content of animals to that of choanoflagellates, a globally distributed group of microbial eukaryotes that are the closest living relatives of animals. Previously, genome sequencing of two species of choanoflagellates provided a valuable initial reconstruction of early animal biology, but the two sequenced species are relatively closely related, leaving the majority of choanoflagellate diversity unrepresented.

We sequenced the transcriptomes of 19 additional choanoflagellate species selected for their phylogenetic diversity, greatly increasing estimates of the gene content of the common ancestor of choanoflagellates and animals. We detect numerous genes in choanoflagellates that were previously found only in animals, including homologs of Toll-like receptors and the NF- κ B family of transcription factors, which are key components of animal innate immunity. In addition, we find genes encoding a hyaluronidase enzyme that is responsible for extracellular matrix modulation in animals and induces a change in cell number within multicelled colonies of the choanoflagellate *Salpingoeca napiformis*. We find that genes lost on the lineage leading to animals are significantly enriched for metabolic function. We identify genes shared exclusively among choanoflagellates and sponges that may underlie the biology of choanocytes, likely the earliest animal cell type. Finally, we demonstrate that Argonaute and Dicer, critical components of the RNAi gene silencing pathway in eukaryotes that were not detected in the first two choanoflagellate species to be sequenced, are both present in choanoflagellates.

Table of Contents

Chapter 1: Interpreting Genomic Data to Reconstruct the Evolution of Animals	1
1.1 Introduction.....	1
1.2 How Should Animal Origins Be Reconstructed?	1
1.2.1 Which organisms are the most appropriate to study in order to reconstruct animal origins?	1
1.2.2 What Is Inferred about the Cell and Organismal Biology of the Most Recent Common Ancestor of Animals?	2
1.2.3 What is Inferred about the Cell Biology of the Unicellular or Colonial Common Ancestor of Animals?	3
1.2.4 What Were the Organismal Changes on the Stem Lineage Leading to Animals?	3
1.3 Challenges in reconstructing animal evolution: what and how to compare?	4
1.4 The Current View of the Evolution of Animals from Genomic Comparisons.....	5
1.4.1 Genome-Scale Inferences from Early-Branching Animal Taxa and Comparisons to Choanoflagellates and Capsaspora	5
1.4.2 Genes, Gene Families and Other Characteristics Shared Exclusively by Animals.....	5
1.4.3 Genes Shared Exclusively Among Animals, Choanoflagellates, and Their Closest Relatives	6
1.4.4 Going Beyond Analyses Based on Identifying Shared Genes.....	8
1.5 Independent Evolution of Multicellularity: Similar to Or Different from Animals?.....	9
1.5.1 Is There Anything Special about the Evolution of Multicellularity in Animals?	9
1.5.2 Are Similar Types of Genes or Genomic Changes Involved in the Evolution of Multicellularity in Different Lineages?	9
1.5.3 What, If Anything, Differentiates Unicellular from Multicellular Lineages? The case of Volvox	10
1.5.4 Secondary Losses of Multicellularity	11
1.6 The Big Questions in Using Genomes to Study the Evolution of Multicellularity.....	12
Chapter 2: Reconstructing the Ancestry of the Animal Genome	13
2.1 Introduction.....	13
2.2 Results	14
2.2.1 The gene content of the common ancestor of animals and choanoflagellates.....	14
2.2.2 Novel choanoflagellate homologs of genes with roles in animal cell adhesion and signaling	14
2.2.3 Evidence for key components of the Notch pathway in choanoflagellates	15
2.2.4 The unicellular origins of animal innate immunity	16
2.2.5 Hyaluronidase and the evolution of animal extracellular matrix	17
2.2.6 Genes lost in the lineage leading to animals are overrepresented for metabolic functions ..	18
2.2.7 The biology of choanocytes, possibly the first cell type in animals	19
2.2.8 RNAi machinery is present in choanoflagellates	20
2.3 Discussion	21
Chapter 3: Culture, Sequencing, Transcriptome Assembly and Annotation of 19 Choanoflagellates	23
3.1 Origin of Cultures	23
3.2 Antibiotic Treatment and Optimization of Culture Conditions.....	24

3.3 Growth of Cultures in Large Batches in Preparation for RNA Isolation	25
3.4 RNA Isolation	26
3.5 Confirmation of Species Identity by DNA Sequencing	27
3.6 Test of PolyA Selection to Separate Choanoflagellate from Bacterial RNA	28
3.7 Library Preparation and mRNA Sequencing	30
3.8 Quality Trimming	30
3.9 <i>de novo</i> Transcript Assembly	31
3.10 Identification and Removal of Cross-Contamination.....	32
3.11 Prediction of Amino Acid Sequences from Assembled Transcripts	34
3.12 Estimation of Completeness of Predicted Proteins	35
3.13 Construction of Orthologous Groups of Proteins	36
3.14 Estimation of Ancestral Gene Content	37
3.15 Annotation of Predicted Proteins and Gene Ontology Enrichment.....	38
3.16 Measurement of Expression Levels	38
3.17 Attempts to Construct Phylogenetic Trees for Orthologous Gene Families	39
3.18 Hyaluronidase Treatment of <i>Salpingoeca napiformis</i> Cultures.....	40
Figures.....	41
References.....	72

Acknowledgements

I would like to thank Nicole King and Michael Eisen for their insight and inspiration. In addition, I would like to acknowledge all present and former members of both the King and Eisen labs for their support, for helping shape my project as it developed, and for enhancing my time as a graduate student. Finally, I'd like to thank my family and friends for being awesome.

Chapter 1: Interpreting Genomic Data to Reconstruct the Evolution of Animals

1.1 Introduction

Between roughly six hundred million and one billion years ago (Douzery et al. 2004; Hedges et al. 2004; Peterson et al. 2004; Parfrey et al. 2011), a unicellular or colonial eukaryote began an evolutionary transition that gave birth to the remarkable diversity of the animal lineage. What were the major events leading up to and during this evolutionary transition? We rephrase this question in a genomic context by examining evidence from early branching animals and their closest relatives. First, we introduce current hypotheses, largely based on ultrastructural observations, regarding the cell and organismal biology of the common ancestor of animals. Next, we summarize the flood of information that genomic analysis has brought to the understanding of the early events in animal evolution, and finally we interpret genomic analysis of animal evolution in the context of independent evolutionary transitions in other eukaryotic and bacterial lineages.

1.2 How Should Animal Origins Be Reconstructed?

1.2.1 Which organisms are the most appropriate to study in order to reconstruct animal origins?

The lack of fossil record for evolutionary transitions on the lineage leading to animals presents a problem due to the large amount of organismal change that occurred over a relatively short evolutionary time span. As an alternative, we can compare living taxa on both sides of the transition in order to produce a reconstruction. The two key nodes to study are the most recent common ancestor of animals and the most recent common ancestor of animals with their closest non-animal outgroup.

The earliest branching extant animals are the sponges, ctenophores, cnidarians, and *Trichoplax adhaerens*. A recent genome-wide analysis echoed the consensus view that sponges are the earliest-branching of these animals (Srivastava et al. 2010), although the issue of sponge paraphyly, and thus whether sponge-specific characteristics can be interpreted as having been present in the animal common ancestor, remains unresolved. Regardless of the branching order of the earliest animals, traits shared among more than one animal group, or between animals and their closest relatives, are likely to have been present in their most recent common ancestor.

Choanoflagellates, a group of unicellular and colonial microbial eukaryotes, are recognized as the closest relatives of animals. The close relationship between choanoflagellates and sponges was first hypothesized by Henry James-Clark in 1867 based on the shared cell morphology between choanoflagellates and choanocytes, a cell type in sponges (James-Clark 1867). Phylogenies based on sequence data have reinforced this idea by demonstrating that both animals and choanoflagellates are monophyletic groups, and that choanoflagellates are animals' closest known living relatives (Snell, Furlong, and Holland 2001; Lang et al. 2002; Philippe, Lartillot, and Brinkmann 2005; Steenkamp, Wright, and Baldauf 2006; Carr et al. 2008; Ruiz-Trillo et al. 2008; Shalchian-Tabrizi et al. 2008). In addition, several other members of the Holozoa, such as

the snail symbiont *Capsaspora owczarzaki* and the ichthyosporean *Sphaeroforma arctica*, are recognized as the closest relatives of animals and choanoflagellates (Ruiz-Trillo et al. 2008).

By comparing the biology of choanoflagellates and other non-animal Holozoa to that of animals, it is possible to produce a reconstruction of the common unicellular or colonial ancestor of animals. Three classes of traits are of special interest in interpreting animal evolution: first, features present only in the common unicellular or colonial ancestor of animals, to the exclusion of other outgroup species (such as fungi) are more likely to have played a role in animal evolution than are those that are ubiquitous in eukaryotes. Second, traits absent from the common unicellular or colonial ancestor with choanoflagellates but present in the common ancestor of animals evolved on the stem lineage leading to animals, and are immediate candidates to explain animal-specific biology. Third, features inferred to be present in the common ancestor of choanoflagellates and animals but which are entirely absent in animals are candidates for aspects of biology that were shed on the lineage leading to animals.

1.2.2 What Is Inferred about the Cell and Organismal Biology of the Most Recent Common Ancestor of Animals?

All animals are permanently multicellular, with the exception of sperm and egg cells participating in reproduction. Flagellated sperm and egg are thought to have been present in the most recent common ancestor of animals (Nichols, Dayel, and King 2009), although many animals are capable of asexual reproduction. Because the choanocyte cell morphology is shared by choanoflagellates and sponges, choanocyte-like cells are likely to have been one of the earliest animal cell types (Nichols, Dayel, and King 2009). This shared cell morphology leads to a fourth class of trait likely to have played a role in animal evolution: those that are shared exclusively among sponges and choanoflagellates are expected to be enriched for features unique to the biology of choanocytes.

Emerging evidence indicates that the common ancestor of animals may have possessed a rudimentary epithelium, although sponges appear to lack classically defined epithelial tissues (Nichols, Dayel, and King 2009). Evidence that all animals possess epithelia is drawn from the observation that sponges are capable of forming cell compartments and multiple types of cell layers that may be homologous structures to true epithelia (Tyler 2003; Leys, Nichols, and Adams 2009) and possess many of the molecular components involved in the development and maintenance of a polarized epithelium (Leys, Nichols, and Adams 2009; Fahey and Degnan 2010). In fact, it has been suggested that evidence for a polarized epithelium in the social amoeba *Dictyostelium discoideum* indicates that the epithelium is substantially evolutionarily older and was present in the common ancestor of social amoebae, fungi, and animals (Dickinson, Nelson, and Weis 2012).

Animals possess two or more germ layers, and it has been proposed that the developmental process of gastrulation is also a feature shared by all animals (Leys and Eerkes-Medrano 2005; Nielsen 2008), although it is unknown whether gastrulation is truly homologous among animals and not simply a common consequence of different modes of cell movement during development such as ingression, invagination and involution (Price and Patel 2004).

1.2.3 What is Inferred about the Cell Biology of the Unicellular or Colonial Common Ancestor of Animals?

Both animals and their closest relatives are capable of forming multicellular structures, although in the case of choanoflagellates, these structures are transient and undifferentiated (e.g. (Dayel et al. 2011)). Cell-cell attachment as a consequence of cell division is present in both animals and choanoflagellates (Fairclough, Dayel, and King 2010), and in some choanoflagellate species, cells within colonies are linked by fine intercellular bridges (Lapage 1925; Fjerdingstad 1961; Hibberd 1975; Leadbeater and Karpov 2000). Colony formation is also present in the ichthyosporean *S. arctica* (Jøstensen et al. 2002), but it is currently unknown whether the processes involved in multicellularity are homologous among these three lineages or whether they each evolved independently. Therefore, animals may have evolved either from a strictly unicellular ancestor or from one that was capable of undifferentiated colony formation.

The choanoflagellate cell type, with an apical flagellum surrounded by a collar of actin-filled microvilli was present in the common ancestor of animals. This cell type is responsible for capture and subsequent phagocytosis of bacteria in both sponges and choanoflagellates, strongly suggesting that bacterivory was present in the unicellular common ancestor of animals (NESCENT Consortium, in preparation). Stable associations with bacteria are most likely present in all animals, and in many cases bacteria can influence development or behavior of their animal host (e.g. (Koropatnick et al. 2004; Mazmanian et al. 2005)). Bacteria can also exert an influence on multicellular development in choanoflagellates, and it has been proposed that bacteria played a more prominent role in animal evolution than simply as prey (Alegado et al. 2012), and furthermore that the increased incidence of horizontal gene transfer as a consequence of bacterivory also played a role in animal evolution (Tucker 2012).

1.2.4 What Were the Organismal Changes on the Stem Lineage Leading to Animals?

A prominent hypothesis for the evolutionary transition to permanent multicellularity in animals was introduced by Ernst Haeckel in 1874 (Haeckel 1874) and later echoed by Libby Hyman in 1940 (Hyman 1940). The colony hypothesis contends that the earliest multicellular animals were undifferentiated colonial flagellates that evolved the capacity for cell differentiation and specialization. More recent interpretations extend the colony hypothesis to propose that mechanisms for cell-cell communication and adhesion were also involved (King 2004).

In addition to the evolution of mechanisms for cell differentiation, specialization, cell-cell communication and stable cell adhesion, what other changes occurred on the stem lineage leading to animals? It is likely that programmed development, differentiation of tissue layers and management of cell death within the organism (via apoptosis and autophagy) were animal innovations, although many of these may have evolved from existing mechanisms in a unicellular context. All animals are anisogamous, in that they have morphologically differentiated sperm and egg. The process of generating gametes and behavioral signals to begin reproduction may also have evolved on the stem lineage leading to animals. Choanoflagellates cannot yet be used as a comparison point for the evolution of sex in animals, as no evidence of sex has been documented in choanoflagellates, although they possess many of the conserved genes necessary for meiosis in eukaryotes (Carr, Leadbeater, and Baldauf 2010).

1.3 Challenges in reconstructing animal evolution: what and how to compare?

A direct and informative approach to understand the biology of an organism is to sequence its genome. What genomic aspects should be compared in order to reconstruct animal evolution? Over a time span approaching two billion years of independent evolution between extant animals and their closest living relatives, genomic characters whose underlying sequence is related to their function, and are therefore more likely to be conserved than non-functional characters, are better poised to provide useful points of comparison. Here, we summarize the advantages and disadvantages of two major types of features: protein domains and complete protein sequences. These two features can be used in tandem to provide a composite reconstruction of gene ancestry.

Protein domains are sequences of amino acids that can frequently be tied directly to a specific molecular function. Domains are generally better conserved than the proteins in which they are found, and thus are well suited to comparisons over long evolutionary distances. In addition, protein domains can be diagnostic for a specific function or class of genes, as in the case of the wnt domain, which is found only in animal Wnt signaling proteins, and the cadherin cytoplasmic domain, which is found only in animal classical cadherins. Protein domains also allow the reconstruction of the evolutionary history of the proteins in which they are found, as in the example of domain shuffling that occurred in the evolution of Hedgehog proteins in animals (Adamska, Matus, et al. 2007; King et al. 2008). The use of protein domain comparisons is not without drawbacks. Protein domains are generally represented in the form of profile hidden Markov models which are generated from alignments of proteins from multiple species known *a priori* to contain the domain. The choice of species represented in the alignment can therefore influence the diversity of species in which the domain might later be found, and in most databases, for example Pfam (Punta et al. 2012), alignments are generally overrepresented for the model organisms in which the function of the majority of protein domains has been investigated and in which the number of proteins containing each domain may have expanded. The latter may be the result of genome duplications occurring in later-branching animal taxa. The former is a natural consequence when a strong relationship between sequence and function is required, but can also bias against the inclusion of early-branching animals in protein domain construction, as they are less well studied at the molecular level. Although many protein domains contain diagnostic amino acid residues that do not change over evolutionary time spans, a domain constructed entirely from bilaterian sequences may not be appropriate for searching in animal outgroups. Another issue in the use of protein domains is that proteins of potential interest may lack protein domains, or they may contain protein domains that are ubiquitously distributed and thus are not diagnostic for homology identification (although they may nevertheless identify a specific molecular function).

Full protein sequences, in contrast, do not require the identification of constituent protein domains in order to identify orthology. Instead, orthology can be assessed by sequence comparison. While sequence comparison skirts many of the problems associated with the use of protein domains, it carries its own set of caveats. The long evolutionary distances involved may have scrambled protein sequences to a degree that makes it impossible to reliably detect homology. In the case where there is some homology present, phylogenetic methods may not have sufficient data to produce a well-supported relationship and existing orthology clustering

methods, while generally quite capable, still have room for improvement (e.g. (Chen et al. 2007)). Indeed, artifacts such as long-branch attraction can lead to inappropriate identification of orthology between animal and non-animal proteins (Mason, Stage, and Goldfarb 2009). Many orthologs, once identified, may lack information about protein function, although classification systems such as the Gene Ontology (Ashburner et al. 2000) can be used for broad characterization of function. Even in the absence of information about function, identification of orthology can be useful in understanding the scope of evolutionary innovation among lineages; for example, to pose questions such as: how many proteins were gained and how many were lost on the stem lineage leading to animals?

Other types of genomic features such as non-coding DNA and RNA are likely to have influenced animal evolution, but they may be more difficult to detect and to annotate depending on their level of evolutionary conservation. There has been tremendous recent interest in the study of various types of non-coding RNAs, and identifying types of RNAs shared between animals and their closest relatives will be an active area of research in the future, as tools for detection and classification such as Rfam are already available (Gardner et al. 2011). Analysis of genomic features can also be combined with biochemical information in order to study the evolution of biologically produced compounds in the fossil record, as has been done for sterols (Kodner et al. 2008).

1.4 The Current View of the Evolution of Animals from Genomic Comparisons

1.4.1 Genome-Scale Inferences from Early-Branching Animal Taxa and Comparisons to Choanoflagellates and Capsaspora

Genome-scale analyses of sponges have indicated that many of the key genes involved in animal development, cell type specification, adhesion, cell signaling, and innate immunity were present in the common ancestor of animals (Nichols et al. 2006; Srivastava et al. 2010). Which of these genes are specific to animals, and which were present before animals diverged from their common ancestor with choanoflagellates? To date, two choanoflagellate genomes have been sequenced, and an analysis of both genomes counted 9,755 genes present in the common ancestor of choanoflagellates and animals, with 1,235 genes gained and 718 genes lost on the lineage leading to animals (Fairclough et al. submitted). These two choanoflagellates may have secondarily lost genes present in their common ancestor with animals, and so study of additional choanoflagellates, and of additional holozoan outgroups (Ruiz-Trillo et al. 2007), will further refine these estimates in the future. Below, we discuss genes, gene families and other characteristics that are currently thought to be animal innovations, followed by those that are exclusively shared with choanoflagellates and *C. owczarzaki* and are therefore likely to have played a role in animal evolution.

1.4.2 Genes, Gene Families and Other Characteristics Shared Exclusively by Animals

Numerous classes of genes and gene families are likely to have evolved in the stem lineage leading to animals, including those involved in developmental signaling and transcription, cell

adhesion, immunity, programmed cell death and small RNA biology. Sponges are representatives of what are generally considered to be the earliest-branching animals, and thus genes shared between sponges and other animals are likely to have been present in their common ancestor.

The Wnt signaling pathway is involved in numerous developmental and other roles in animals, and is animal-specific (Adamska, Degnan, et al. 2007; Adamska et al. 2010). In addition, a number of developmental transcription factors are likely to have evolved on the lineage leading to animals. Several homeobox gene families such as Antennapedia, Pax and Sox are present in sponges and other animals (Gazave et al. 2008; Larroux et al. 2008), as are Tcf/Lef, Gli and Smads (Srivastava et al. 2010). Animal cell adhesion involves many genes of ancient origin, but classical cadherins are animal-specific (Srivastava et al. 2010; Nichols et al. 2012). In addition, certain combinations of domains involved in adhesion are animal-specific, such as the combination of immunoglobulin and fibronectin 3 domains (Srivastava et al. 2010). As discussed above, some molecular components of the epithelium appear to have evolved in the animal common ancestor (Fahey and Degnan 2010), including the Par, Discs Large, and Crumbs complexes (Srivastava et al. 2010).

Many features of the animal immune system are animal-specific, including interleukin receptors, immune adapters such as MyD88, immune effector systems, and viral defense factors such as MDA5-like RNA helicases (Gauthier, Du Pasquier, and Degnan 2010; Srivastava et al. 2010). Sponges lack neurons, but do possess a diverse set of GPCRs (Srivastava et al. 2010), and many genes involved in post-synaptic neuronal scaffolding (Sakarya et al. 2007) and neuronal cell type specification (Jackson et al. 2010), indicating that the ancestral metazoan already possessed many building blocks of the animal-specific nervous system. Initiator and effector caspases involved in programmed cell death are also animal innovations (Srivastava et al. 2010).

In most of the cases described above, the earliest animals are inferred to have had a relatively small complement of each class of gene, and diversification in later-branching animals generated the full diversity that we observe in animals today. Finally, we note that sequencing of closely related non-animals continues, and it is possible that some genes currently believed to be animal-specific have yet to be discovered in choanoflagellates or other outgroups within Holozoa.

1.4.3 Genes Shared Exclusively Among Animals, Choanoflagellates, and Their Closest Relatives

Of the roughly 10,000 genes that are estimated to have been present in the common ancestor of animals and choanoflagellates, those that were specifically gained in that common ancestor are of perhaps the most interest, as they are more likely than genes shared with other eukaryotes to have played a role in animal evolution. The past ten years have seen an explosion in information on the common ancestor of choanoflagellates and animals, largely driven by genome-scale sequencing on three choanoflagellate species, *Monosiga brevicollis*, *Monosiga ovata*, and *Salpingoeca rosetta* and the filasterian *C. owczarzaki* (Table 1.1). (We note that choanoflagellate species nomenclature was originally based on morphology rather than phylogeny, and current data indicates that *M. brevicollis* and *S. rosetta* are more closely related to one another than they are to *M. ovata* (Carr et al. 2008; Nitsche et al. 2011)). Most of these genes were detected by a combination of analysis of diagnostic protein domains and comparisons of orthology, but several genes have been further investigated for function *in vivo* and *in vitro*. The role of genes responsible for cell adhesion in animal evolution has been reviewed elsewhere (Abedin and King

2010) and continues to be a subject of active investigation (Fahey and Degnan 2012; Nichols et al. 2012). We summarize the most compelling findings for tyrosine kinases.

Tyrosine kinases play numerous signaling roles in animals, and are involved in differentiation, development, homeostasis, proliferation, aging, and other processes (for a review, see (Hunter 2009)). Each of the independent holozoan lineages (animals, choanoflagellates, and *C. owczarzaki*) are estimated to contain at least 100 distinct tyrosine kinase genes (King and Carroll 2001; King et al. 2008; Manning et al. 2008; Pincus et al. 2008; Suga et al. 2012), with the number of tyrosine kinases in *Monosiga brevicollis*, for example, exceeding the number found in some animals. Tyrosine kinases are not thought to be found outside of Holozoa, with the exception of a family of structurally unrelated genes found in bacteria (Lee and Jia 2009), although it has been suggested that they are present in divergent eukaryotic lineages (Judelson and Ah-Fong 2010).

Were these tyrosine kinase genes all present in the common ancestor of Holozoa, or are they the product of an independent diversification process in each lineage? Studies based on sequence analysis in five species of choanoflagellates and the filasterians *C. owczarzaki* and *Ministeria vibrans* have indicated that the majority of the expansion of the tyrosine kinase repertoire and the associated repertoire of tyrosine phosphatase and SH2 domain containing proteins appears to have occurred independently in each lineage (Manning et al. 2008; Pincus et al. 2008; Suga et al. 2008; Li, Scarlata, and Miller 2009; Liu et al. 2011; Suga et al. 2012). Interestingly, a similar expansion of tyrosine kinases was observed in the sponge *Amphimedon queenslandica*, in which the majority of genes are not shared with other animals, although some shared families such as EGF and Met are present (Srivastava et al. 2010).

Tyrosine kinases can be divided into two classes: receptor tyrosine kinases and tyrosine kinases lacking a predicted extracellular domain. Receptor tyrosine kinases in choanoflagellates and in *C. owczarzaki*, although they display domain architectures reminiscent of typical mammalian architectures, do not show a high level of sequence similarity to animal receptor tyrosine kinases (King and Carroll 2001; Manning et al. 2008; Suga et al. 2008; Miller 2012; Suga et al. 2012). Many cytoplasmic tyrosine kinases in choanoflagellates also appear to show relatively little orthology to animals by primary sequence or by domain architecture (King et al. 2008; Pincus et al. 2008), although a recent study including *C. owczarzaki* found that up to eight of ten conserved metazoan families of non-receptor tyrosine kinases were present in the common ancestor of Holozoa (Suga et al. 2012). Cytoplasmic tyrosine kinases reported to be conserved in Holozoa include Src, Abl, Tec, Csk (Manning et al. 2008; Ortutay et al. 2008), Fes, Fak, Shark/HTK16, and Jak (Suga et al. 2008; Suga et al. 2012). This indicates that a core set of cytoplasmic tyrosine kinases may have evolved in the common holozoan ancestor, and that the core set of genes was retained while each lineage independently diversified its complement of both receptor and cytoplasmic tyrosine kinases. Does conservation of core genes at the sequence level indicate conservation of function? This question has been investigated most rigorously in Src and its negative regulator Csk.

Src is a proto-oncogene first discovered as an endogenous gene in chicken with homology to Rous sarcoma virus (Stehelin et al. 1976). Homologs of Src in *M. brevicollis* (MbSrc1) and in *M. ovata* could functionally substitute for animal Src, and *M. brevicollis* MbSrc1 was found to have binding substrates and affinity similar to mammalian Src (Segawa et al. 2006; Li et al. 2008). In addition, *M. brevicollis* MbSrc1 and *C. owczarzaki* homologs CoSrc1 and CoSrc2 can be regulated by autophosphorylation, which is typical of animal tyrosine kinases (Li et al. 2008; Schultheiss et al. 2012). While these two aspects of Src function are shared with animals,

negative regulation by Csk is not. Both MbSrc1 and *M. ovata* Src were found to be phosphorylated by Csk, but neither phosphorylation resulted in negative regulation (Segawa et al. 2006; Li et al. 2008), while *C. owczarzaki* Csk was found not to phosphorylate either CoSrc1 or CoSrc2 (Schultheiss et al. 2012). These findings suggest that Src tyrosine kinase activity was ancestral in Holozoa, but that negative regulation by Csk was an animal-specific innovation. *M. brevicollis* has four identified homologs of Src, and another of the homologs, MbSrc4, was found not to functionally substitute for Src, although it was regulated by autophosphorylation (Li, Scarlata, and Miller 2009). Instead, it was found to have a role in membrane binding, with its C2 domain thought to be functionally convergent with the myristoylation signal of animal Src. These findings underscore a complex evolutionary history of tyrosine kinases in Holozoa, in which some gene families are shared by all extant lineages and thus likely to be ancestral. Other gene families appear to have expanded independently in different lineages, with a subset of those having rearranged similar building blocks to produce convergent functions.

What were the roles of tyrosine kinases in the common unicellular or colonial ancestor of animals? The phosphorylation activity of *M. brevicollis* tyrosine kinases can vary based on nutrient condition (King, Hittinger, and Carroll 2003), indicating that they may play a signaling role in choanoflagellates. Could the ancestral holozoan tyrosine kinase network have been responsible for receiving and interpreting environmental cues, and was the network co-opted in animals for intercellular communication (Miller 2012; Suga et al. 2012)? Few studies on the endogenous roles of tyrosine kinases in non-animal holozoa have been conducted, and an increased understanding of their role in a unicellular or colonial context is certain to clarify their contribution to animal evolution.

1.4.4 Going Beyond Analyses Based on Identifying Shared Genes

The presence or absence of genes in the common unicellular or colonial ancestor of animals is only the tip of the iceberg in attempting to understand the biology of animal evolution. For example, few analyses attempting to understand the nature of the genes lost leading to the animal lineage have been conducted. What other types of genetic or epigenetic changes besides gene gain and loss may have played a significant role? Gene regulation and the rewiring of transcriptional networks is likely to have occurred (Piasecki, Burghoorn, and Swoboda 2010; Young et al. 2011), although the long evolutionary distances involved and the labile nature of transcription factor binding sites make it more difficult to study changes at the DNA level. Regulation at the level of alternative splicing has become a more accessible phenomenon to study given the advent of recent sequencing technology (Westbrook 2011), and investigations of the roles of post-translational modification may follow. MicroRNAs are present in early-branching animals but not in *M. brevicollis* (Grimson et al. 2008), but the study of non-coding RNAs and their regulatory roles in the earliest animals is in its infancy. Could interactions with the environment have made a significant contribution to animal evolution? Bacteria are capable of modulating the cell biology and signaling of choanoflagellates (King, Hittinger, and Carroll 2003; Abedin 2010; Alegado et al. 2012), and further study in other unicellular Holozoa may illuminate the relationship of bacteria to animal evolution.

1.5 Independent Evolution of Multicellularity: Similar to Or Different from Animals?

1.5.1 Is There Anything Special about the Evolution of Multicellularity in Animals?

All animals are heterotrophic and reproduce by sperm and egg, but the shared animal characteristic that has attracted perhaps the most attention is their multicellularity. The evolution of multicellularity in animals is only one of many independent evolutionary events leading to multicellularity in both eukaryotes and bacteria (Bonner 2001; King 2004). Does the frequency with which multicellularity has arisen in different lineages have a relationship to how “difficult” it might be to evolve multicellularity? Indeed, some authors have argued that the evolutionary transition may be easy (Grosberg and Strathmann 2007). We take a neutral view on this question; we focus instead on reconstructing the events that may have taken place in different lineages. We note that there are also a number of related questions, such as: why have some multicellular lineages produced a significantly greater apparent diversity of species and forms than others? Are some multicellular forms more biologically predisposed to evolutionary innovation than others, or is chance the predominant factor? While these questions may be too difficult to answer given available data, we attempt to illuminate them by determining whether similar genomic changes have occurred in each independent lineage, or whether the evolution of multicellularity is genetically distinct each time it occurs.

Underlying these questions is the assumption that the evolution of multicellularity was a truly independent event in different lineages. Cell adhesion mechanisms in disparate multicellular groups appear to share little homology, especially in the underlying genes that are deployed (Abedin and King 2010). But, as discussed above, it has also been proposed that an epithelium was present in the common ancestor of animals and social amoebae, with the implication that multicellularity evolved only once in their common ancestor and was subsequently lost in several lineages (Dickinson, Nelson, and Weis 2012). We escape this potential complication by compiling evidence from the evolution of multicellularity in numerous disparate lineages in eukaryotes and in bacteria.

1.5.2 Are Similar Types of Genes or Genomic Changes Involved in the Evolution of Multicellularity in Different Lineages?

The brown algae *Ectocarpus siliculosus*, a stramenopile, represents an independent evolution of multicellularity from those that occurred in plants, animals, fungi, oomycetes and other groups. The recent publication of the genome sequence for *E. siliculosus* examined the question of whether these multicellular lineages showed similar trends in gene gain or loss (Cock et al. 2010). On average, there was a tendency for multicellular lineages to have gained more new gene families and lost fewer gene families, although the higher representation of sequenced genomes for multicellular versus unicellular species within some groups may make the comparison of gene gain and loss more difficult to interpret. There were no significant common trends detected for functional assignments of the gene families that were gained and lost in the multicellular lineages, but some features were shared among multicellular lineages. Eight protein domains, all with putative enzymatic function, were overrepresented in the multicellular versus unicellular lineages that were examined. In addition, seven transcription factor families were

overrepresented in multicellular lineages, including the MADS, LIM and SWI/SNF families.

Perhaps the most compelling commonality found among the multicellular lineages in comparison to their unicellular relatives was the tendency for the gain of receptor kinases. Three of the five multicellular groups that were compared appeared to have independently evolved membrane-spanning kinases: brown algae, plants and animals. Plants and brown algae lack tyrosine kinases, but do possess an abundance of serine/threonine kinases. It was previously shown that receptor kinases independently evolved in plants and in animals, with both signaling networks important for growth and development and also sharing convergent mechanistic similarity (Shiu and Bleecker 2001; Cock, Vanoosthuyse, and Gaude 2002; De Smet et al. 2009). A diverse set of kinases is also present in the independently evolved multicellular oomycetes (Judelson and Ah-Fong 2010) and the 133-member protein kinase family FunK1 is found only in multicellular fungi (Stajich et al. 2010). The social amoeba *Dictyostelium discoideum* has a large number of receptor serine/threonine kinases and receptor histidine kinases (Eichinger et al. 2005). Within bacteria, the genome of the multicellular cyanobacterium *Nostoc punctiforme* encodes roughly 300 sensory histidine kinase and serine/threonine kinase two-component signal transduction genes (Meeks et al. 2001). The discovery of a receptor kinase signaling network in brown algae strengthens the hypothesis that multicellular lineages deploy kinase signaling as an intra-organism mechanism of information transfer, as has been previously proposed (Lim and Pawson 2010).

Another similarity among multicellular lineages observed in the *E. siliculosus* genome paper was the presence of neurosecretory soluble NSF attachment protein receptors (SNAREs). SNAREs are a large protein superfamily, members of which mediate membrane fusion and are key components of nervous system function in animals (Jahn and Scheller 2006). SNARE proteins act in multi-protein complexes, one of which, the Munc18/syntaxin 1 complex, was shown to be shared between choanoflagellates and animals (Burkhardt et al. 2011). In plants, multicellular lineages have undergone expansion of SNARE proteins in comparison to their unicellular relatives (Sanderfoot 2007). The discovery of 20 SNARE proteins in *E. siliculosus* (Cock et al. 2010) indicates that membrane fusion may have been deployed repeatedly in the evolution of multicellularity in disparate lineages.

1.5.3 What, If Anything, Differentiates Unicellular from Multicellular Lineages? The case of *Volvox*

Within the plants, the genomes of the unicellular green alga *Chlamydomonas reinhardtii* and its multicellular relative *Volvox carteri* have both been sequenced. The comparison of *C. reinhardtii* and *V. carteri* is particularly compelling for two reasons: first, *V. carteri* is multicellular but only has two differentiated cell types (biflagellate somatic cells and reproductive cells), in which the somatic cells resemble *C. reinhardtii* (Kirk 2005). Second, the divergence of *V. carteri* and *C. reinhardtii* is more recent than the divergence of animals or land plants from their closest unicellular relatives (Herron et al. 2009). Therefore, the changes that have occurred since the divergence of *V. carteri* and *C. reinhardtii* from their common ancestor may be easier to detect and to understand than for other comparisons. The *V. carteri* genome was found to be 17% larger than the *C. reinhardtii* genome, primarily due to increased repeat content, and both the number of protein domains and the number of protein domain combinations was measured to be similar between the two species (Prochnik et al. 2010). Overall, only 32 groups

of orthologous genes differed in their presence or absence between the two species, but there was a significant trend for *V. carteri* to have more genes in each shared family than *C. reinhardtii*. In a more specific comparison of genes expected to be involved in the developmental process in *V. carteri*, only three major changes were observed, and the changes were in the size of gene families rather in their presence or absence. Two families of genes involved in extracellular matrix formation were greatly expanded in *V. carteri*. One of these two families, pherophorins, also have a *V. carteri*-specific role in sexual differentiation (Hallmann 2006). The third change observed an expansion of the cyclin D family in *V. carteri*, which may be involved in regulation of the cell cycle during development. The relative paucity of major changes in protein coding genes between *C. reinhardtii* and *V. carteri* indicates that ancestral genes may have acquired new roles in *V. carteri*, that the dramatic change in morphology and behavior required the evolution of relatively few, as yet unidentified genes, or that changes outside the protein coding repertoire may have been involved.

1.5.4 Secondary Losses of Multicellularity

In comparison with gains of multicellularity, losses of multicellularity are less well studied. Losses of multicellularity are proposed to have occurred in several lineages, with perhaps the best characterized case residing in the fungi (Bruns et al. 1992; Berbee and Taylor 1993; Liu, Whelen, and Hall 1999). Do secondarily unicellular lineages retain the primordial components necessary for the “reactivation” of multicellularity? Two studies in the unicellular fungus *Saccharomyces cerevisiae* found that selective constraints could cause cells to form aggregates within a small number of generations (Koschwanez, Foster, and Murray 2011; Ratcliff et al. 2012). A similar transition occurred in the presence of a predator in the green alga *Chlorella vulgaris* (Boraas, Seale, and Boxhorn 1998), which also appears to have lost multicellularity (Becker and Marin 2009). Is the transition to an undifferentiated aggregate of cells easier to achieve than a transition to a multicellular state with cell differentiation? The formation of aggregates may be the result of a simple switch to prevent cell separation following division, although one of the studies in yeast observed evidence for the evolution of a more complex life history (Ratcliff et al. 2012). The converse experiment has also been conducted, in which selective conditions could induce the loss of multicellularity in the social bacterium *Myxococcus xanthus* (Velicer, Kroos, and Lenski 1998). In both types of experiments, the relationship between laboratory conditions and the conditions that lead to the gain or the loss of multicellularity in the environment largely remain to be elucidated.

Cancer in animals has often been described as a loss of multicellularity. Indeed, some forms of cancer may actually become transmissible outside of their original host (Murgia et al. 2006), and some have even referred to cancer as the reverse of the evolutionary process that gave rise to animals (Davies and Lineweaver 2011). Many of the genes found to have evolved in the lineage leading to the most recent common ancestor of animals have also been implicated in tumor development (Srivastava et al. 2010). Do other multicellular lineages get cancer, or are animals uniquely predisposed due to their lack of cell wall and migratory cell types? Plant galls have been characterized as tumors (Riker 1958), although it has also been argued that plants do not get cancer (Doonan and Hunt 1996).

1.6 The Big Questions in Using Genomes to Study the Evolution of Multicellularity

What has been learned about the evolution of multicellularity from comparative genomics? In animals, many of the components of key pathways for cell adhesion, gene regulation and intercellular signaling were present in their unicellular or colonial common ancestor, and these components were then expanded and elaborated in animals. Other components appear to be animal innovations: transcription factors involved in body patterning, most aspects of the immune system, and many genes involved in nervous system function. A third class of genes evolved on the lineage leading to animals via modification by domain shuffling and other mechanisms to generate novel function. Independent instances of the evolution of multicellularity may each have involved a different path to the construction of an intercellular signaling network based on the shared mechanism of phosphorylation. Tantalizing clues to other pathways shared by multicellular lineages, such as those involving membrane fusion, will be investigated in the future.

Is it possible that genes that are conserved between unicellular organisms and their multicellular cousins could also have played a role in the evolution of multicellularity? This is undoubtedly true, although their relative importance in comparison with novel genes is an open question. A recent finding that plant actin can be functionally substituted by both animal and choanoflagellate actins (Kandasamy et al. 2012) is an initial observation on this subject. What about the role of gene loss in the evolution of multicellularity? Do our current hypotheses regarding evolutionary transitions to multicellularity underestimate the role of loss, or is there no biological correlate to the set of genes lost in different lineages? Loss, by its nature, can be more difficult to test, but determining the consequences of manipulating the function of genes lost in multicellular lineages in their closest unicellular relatives of multicellular lineages is an available avenue of investigation.

What is missing from prior analyses? In the coming years, the study of the evolution of multicellularity in disparate lineages will venture beyond the central dogma to explore the function of all sizes of non-coding RNAs, epigenetic phenomena and the influence of the environment. In addition, sequenced genomes have only recently become available for representatives of many pairs of unicellular and multicellular lineages. As the number of available genomes increases, and as tools for the measurement and manipulation of gene function are developed in these largely non-model organisms, hypotheses about specific genes can be tested. Functional inferences will be required to deepen the understanding of the evolution of multicellularity within lineages and the comparison among lineages. By using genomes as a tool to discover the gene content of extant organisms, and by examining the function of those genes in both a unicellular and multicellular context, we can begin to reconstruct the events leading to the diversification of many of the groups of organisms we see on earth today.

Chapter 2: Reconstructing the Ancestry of the Animal Genome

2.1 Introduction

Inferring the genome content of the last common ancestor of animals and choanoflagellates requires the appropriate choice of both choanoflagellate and animal species. To date, many animal genomes but only two choanoflagellate genomes have been sequenced: *Monosiga brevicollis* (King et al. 2008) and *Salpingoeca rosetta* (Fairclough et al. submitted). These two species are relatively closely related to one another within the choanoflagellates (Carr et al. 2008; Nitsche et al. 2011), leaving the bulk of choanoflagellate genomic diversity unexplored. If the clade containing these two choanoflagellates has experienced losses of genes that were present in the common ancestor of animals and choanoflagellates, then current estimates of the genomic complexity of that common ancestor are incomplete. Evidence that *M. brevicollis* and *S. rosetta* have experienced loss was discovered in the genome of *Capsaspora owczarzaki*, a filasterean snail symbiont that is one of the closest known relatives of animals and choanoflagellates. *C. owczarzaki* encodes several genes not previously found in choanoflagellates that were thought to have been animal specific, including the cell adhesion protein beta integrin (Sebé-Pedrós et al. 2010) and T-box and Runx transcription factor families (Sebé-Pedrós et al. 2011).

To gain a more complete picture of the genomic complexity of the first animals, we sequenced and assembled the expressed mRNA content from 19 additional species of choanoflagellates. Transcriptome sequencing is more accessible than genome sequencing for choanoflagellates because they are co-cultured with their bacterial prey; the fact that bacteria do not polyadenylate their mRNA can be used as a method of separating choanoflagellate from bacterial transcripts. The main disadvantage of mRNA sequencing is that genes that are not expressed cannot be sequenced, but one motivation behind our strategy of diverse choanoflagellate sampling was to overcome this potential shortcoming by virtue of collecting data from many species. Nevertheless, our transcriptome data set does not allow us to conclude that any particular gene is absent from a given species of choanoflagellate; therefore, when discussing gene loss, we focus instead on general patterns of loss in choanoflagellates.

Our transcriptome sequencing data and analyses have greatly increased estimates of the gene content of the common ancestor of choanoflagellates, and thus the common ancestor of choanoflagellates and animals. We detect numerous genes in choanoflagellates that were previously thought to have been specific to animals, including genes involved in animal innate immunity such as Toll-like receptors and members of the NF- κ B pathway, the Notch ligand Delta, and a hyaluronidase enzyme that is responsible for extracellular matrix degradation in animals. This hyaluronidase is specific to animals and choanoflagellates and is capable of causing a morphological change in multicelled colonies of the choanoflagellate *Salpingoeca napiformis*. In addition, for the first time, we provide a catalog of genes lost in the lineage leading to animals and summarize their potential functions. We also identify a group of genes found exclusively in choanoflagellates and in sponges that may be involved in the biology of the choanocyte cell type that was likely intimately involved in animal evolution. Finally, we demonstrate that Argonaute and Dicer, critical components of the RNAi pathway in eukaryotes,

are present in some choanoflagellates and were likely lost in the lineage leading to *M. brevicollis* and *S. rosetta*.

2.2 Results

2.2.1 The gene content of the common ancestor of animals and choanoflagellates

We obtained 20 clonal cultures of choanoflagellates, representing 19 species (Figure 2.1). For each culture, we optimized growth conditions, isolated and sequenced mRNA and produced a *de novo* transcriptome assembly. We compared the proteins predicted from our assemblies to the predicted proteins from the genome sequences of *M. brevicollis* and *S. rosetta*, 21 animal genomes selected for their phylogenetic diversity and 17 outgroup species representing both the closest known outgroups to animals and choanoflagellates and each currently recognized eukaryotic supergroup. From this comparison, we produced a set of orthologous gene families and analyzed their evolution in a phylogenetic context (see chapter 3 for a complete description of the methods for this chapter).

We infer that 19,520 gene families were present in the common unicellular or colonial ancestor of choanoflagellates and animals (Figure 2.1). Many gene families were both gained and lost in each evolutionary transition leading to and within animals. Gene loss and gain have overall played relatively equal roles along each branch. Within animals, the evolution of eumetazoa (Cnidaria and Bilateria) was accompanied by nearly twice as much gene gain as loss, whereas the lineage leading to Bilateria lost more gene families than it gained. We estimate that the most recent common ancestor of animals had 17,753 gene families, 2,742 of which were gained specifically in animals.

Of the 4,501 gene families we infer to have arisen on the lineage leading to the common ancestor of animals and choanoflagellates, many were previously thought to have been specific to animals, and many of these gene families play critical roles in animal development, cell adhesion, gene regulation, immunity and homeostasis (Table 2.1). These 4,501 gene families represent nearly one quarter of the total gene families present in the common ancestor of animals and choanoflagellates, and we highlight several examples in subsequent sections.

2.2.2 Novel choanoflagellate homologs of genes with roles in animal cell adhesion and signaling

A fundamental early event in animal history was the evolution of novel mechanisms for cell-cell adhesion and signaling, many of which are likely to be innovations based on a toolkit of existing genes (King 2004). We find new choanoflagellate orthologs in two classes of animal cell adhesion and signaling genes: cadherins and integrins. Diverse cadherins have been identified in choanoflagellates (Abedin and King 2008; Nichols et al. 2012), but two types of cadherins, Flamingo and Protocadherins, were previously thought to be animal-specific. The first described Flamingo cadherin (also known as starry night) is involved in the Frizzled-mediated establishment of planar polarity in the *Drosophila melanogaster* wing (Chae et al. 1999; Usui et al. 1999), and we find that predicted proteins from seven species of choanoflagellates are present in the same orthologous family as *D. melanogaster* Flamingo. In animals, Flamingo cadherins

are distinguished by the combination of two diagnostic domains (Pfam nomenclature is used throughout): GPS and 7tm_2, a seven-pass transmembrane domain. Many animal Flamingo cadherins also have HRM, EGF, and Laminin domains. Five choanoflagellate species have predicted proteins with identified GPS and 7tm_2 domains. *Salpingoeca urceolata* encodes a protein with Laminin_G_3, HRM, GPS and 7tm_2 domains and *Codosiga hollandis* a protein with Cadherin, GPS, and 7tm_2 domains, although no choanoflagellate species possesses the full combination of domains found in *D. melanogaster* Flamingo. Protocadherins are a large and diverse family of genes involved in animal development and cell adhesion (for a review, see (Frank and Kemler 2002)), and most, but not all, family members possess a diagnostic transmembrane Protocadherin domain paired with one or more extracellular Cadherin domains. Six species of choanoflagellates are predicted to encode proteins with Protocadherin domains, although none of these proteins also contains a Cadherin domain. The *Choanoeca perplexa* Protocadherin domain-containing protein also possesses an intracellular SH2 phosphorylated tyrosine binding domain, a combination not found in any other organism. Tyrosine kinase signaling networks greatly expanded in the common ancestor of choanoflagellates and animals (Manning et al. 2008; Pincus et al. 2008), and the Protocadherin domain in *C. perplexa* may have been co-opted as a tyrosine kinase signaling protein.

Integrins play critical roles in adhesion and cell signaling in animals, where they act as an alpha/beta heterodimer at the plasma membrane and as a link to downstream pathways involved in transcription, proliferation and motility (Hynes 2002). Both components are present in *C. owczarzaki* (Sebé-Pedrós et al. 2010), but a putative match to alpha integrin is the only reported evidence of either protein in choanoflagellates (King et al. 2008), implying that both integrins were present in the common ancestor of animals and choanoflagellates and at least beta integrin was lost. We find strong evidence for the diagnostic integrin beta domain in *Didymoecca costata*, but in no other species of choanoflagellate. In contrast, we find the beta integrin binding protein ICAP-1 (Chang et al. 1997) in all but two species of choanoflagellates, but not in *C. owczarzaki*. ICAP-1 has been proposed to act as a competitive inhibitor of beta integrin (Bouvard et al. 2003), although it is unlikely to play that role in the majority of choanoflagellate species which have ICAP-1 but lack beta integrin. These results indicate a complex evolutionary history for newly discovered genes in choanoflagellates that are involved in cell adhesion in animals. We infer that the building blocks of Flamingo, Protocadherin and integrins were present in the common ancestor of choanoflagellates and animals. In Flamingo cadherins, the same basic domain structure is found in animals and in choanoflagellates, but appears to have been diversified in the animal lineage. In contrast, in the case of Protocadherins, only the Protocadherin domain is shared between animals and choanoflagellates, and there is evidence that the protein structure and function may have diverged independently in both lineages. Both alpha and beta integrins and the beta integrin binding partner ICAP-1 appear to have been present in the common ancestor of animals and choanoflagellates, but the subsequent loss of beta integrin in nearly all choanoflagellates implies either that ICAP-1 did not have an ancestral role in beta integrin binding or that it was secondarily lost in choanoflagellates.

2.2.3 Evidence for key components of the Notch pathway in choanoflagellates

The Notch pathway plays a significant and conserved role in animal development, influencing proliferation, differentiation and apoptosis, among other processes (Artavanis-Tsakonas, Rand,

and Lake 1999). In animals, Notch and its ligand Delta/Serrate/Jagged (hereafter Delta) are transmembrane proteins that interact between adjacent cells. Notch is activated when bound by Delta, and a gamma secretase within the cytoplasm cleaves the intracellular domain of activated Notch. The activated intracellular domain translocates to the nucleus and binds CSL/Suppressor of Hairless, allowing Su(H) to activate downstream target genes. The core pathway, which has been found to interact with dozens of additional genes, is conserved in animals, and we also detect its constituents in choanoflagellates (Figure 2.3b). Su(H) appears to be present in all choanoflagellate species sequenced thus far. Three constituent domains of Notch in animals, EGF (epidermal growth factor), Notch and Ank (ankyrin), were previously found in different predicted proteins in *Monosiga brevicollis* (King et al. 2008), indicating that the protein domains, if not the Notch protein, were likely present in the common ancestor of animals and choanoflagellates. In our expanded choanoflagellate data set, we detect individual predicted proteins in *Salpingoeca napiformis* and in *Monosiga ovata* containing all three domains in the appropriate order and with a transmembrane domain in a similar location to that of animal Notch. While all choanoflagellate species encode predicted proteins with Notch domains, which in many cases are paired with EGF or Ank domains, these are the only two species with all three sets of domains in a single predicted protein, indicating that typical animal Notch domain organization was either lost in most choanoflagellates or independently evolved from the same building blocks in animals, *S. napiformis* and *M. ovata*. Similarly, we find that only one choanoflagellate species, *Salpingoeca gracilis*, encodes a predicted protein with both constituent domains of canonical animal Delta, MNNL and DSL, but that the DSL domain is present in six species of choanoflagellates (including *S. napiformis* but not *M. ovata*). Thus, conserved elements of the Notch pathway were likely present in the common ancestor of choanoflagellates and animals and subsequently lost in many species of choanoflagellates, although their function in the common ancestor and in choanoflagellates remains unknown.

2.2.4 The unicellular origins of animal innate immunity

Like the Notch pathway, the nuclear factor kappa B (NF- κ B) signaling pathway is involved in numerous organismal, developmental and cellular processes in animals. Here, we highlight its role in animal innate immunity in order to determine whether aspects of the animal immune system might have been present in their unicellular or colonial common ancestor with choanoflagellates. Previously, an NF- κ B ortholog and some other pathway members were detected in *C. owczarzaki* (Sebé-Pedrós et al. 2011; Gilmore and Wolenski 2012), but were thought to be lost in choanoflagellates, indicating that the common ancestor of animals, choanoflagellates and *C. owczarzaki* may have been capable of the transcriptional activation steps of the NF- κ B signaling cascade. What extracellular signals might initiate this signaling process? The Toll-like receptors (TLRs), a family of proteins which are responsible in animals for recognizing conserved aspects of diverse viral and bacterial pathogens and initiating an immune response, have thus far not been detected outside of animals, although a convergently evolved group of proteins is present in plants. In our data, we find evidence for the first TLR outside of animals, in addition to several conserved members of the TLR signal transduction pathway. Canonical animal TLRs are composed of an N-terminal signal peptide, one or more leucine rich repeat (LRR) domains, a transmembrane domain, and an intracellular Toll/interleukin 1 receptor/resistance (TIR) domain. There are five choanoflagellate species

encoding predicted proteins with TIR domains, and one species, *Acanthoecca spectabilis*, has a predicted protein exactly matching the canonical animal TLR domain organization. The *A. spectabilis* TLR is too distantly related to animal proteins to reliably detect orthology (as is the case for many choanoflagellate proteins due to the substantial amount of evolutionary history separating choanoflagellates and animals). Curiously, TLRs have not been found in sponges (Gauthier, Du Pasquier, and Degnan 2010), indicating that they may also be difficult to detect or that they have been secondarily lost in the sponge species that have been studied. We also find eight choanoflagellate species, including *A. spectabilis*, with the Rel homology domain (RHD) that is diagnostic for NF- κ B.

Could TLR stimulation leading to NF- κ B activation have been present in the common ancestor of animals and choanoflagellates? In animals, there are several pathways for this signaling process (reviewed in (Kawai and Akira 2007)), and we find evidence for two such pathways in choanoflagellates (Figure 2.3a). In the canonical activation pathway, different TLRs are bound by various sets of intracellular adapters including MyD88, which then initiate a signaling pathway via IRAK, TRAF and IKK α /IKK β that leads to the phosphorylation of I κ B and its degradation. Once I κ B is degraded, it releases NF- κ B, which then translocates to the nucleus to act as a transcription factor. By analyzing domain structure and protein family groups, we detect evidence for choanoflagellate predicted proteins encoding homologs of IRAK, TRAF, and I κ B, but we are unable to detect adapters such as MyD88 or the kinase IKK α /IKK β . We also find evidence in choanoflagellates for an alternative NF- κ B activation pathway via TRAM and IRF3, although the adapters TRIF and TBK1 do not appear to be present. None of these pathway components appear to be present outside Holozoa (animals, choanoflagellates and *C. owczarzaki*). Thus, the primordial components of the animal innate immune system are likely to have evolved on the lineage leading to Holozoa, with TLRs an innovation specific to animals and choanoflagellates. The absence of several critical members of the animal signaling pathway indicates that either the ancestral pathway involved fewer components or that the pathway was simplified and rewired in choanoflagellates, or some combination of the two.

2.2.5 Hyaluronidase and the evolution of animal extracellular matrix

Hyaluronic acid is a component of animal extracellular matrix (ECM), and is thereby involved in diverse animal physiological processes such as development, adhesion, differentiation, migration, proliferation, immune surveillance, inflammation and malignant transformation (for a review, see (Girish and Kemparaju 2007)). The hyaluronic acid molecule is a GlcA/GlcNAc polysaccharide polymer that can reach up to megadaltons in weight and plays a structural role in animal tissues and a signaling role via interaction with cell surface receptors. Hyaluronic acid intracellular synthesis at the plasma membrane is linked to its export, and nearly all choanoflagellates possess orthologs of the transmembrane synthesis genes HAS1, HAS2 and HAS3, although identification is complicated by the absence of diagnostic domains and the large number of biosynthetic genes in eukaryotes with similar domain architectures but different functions. In contrast, the major enzyme responsible for the modulation or degradation of hyaluronic acid, animal hyaluronidase, is identified by the diagnostic Glyco_hydro_56 domain, which is present in 13 species of choanoflagellates. Furthermore, all choanoflagellate species lacking the domain are predicted to encode proteins within the same orthologous gene family as animal hyaluronidase. Previously, this hyaluronidase had only been found within animals;

bacteria also produce hyaluronidases, but they act by a different catalytic mechanism than animal hyaluronidase. Bacterial hyaluronidases (often referred to as hyaluronic acid lyases) are implicated in pathogenesis; for example, they may be used to degrade extracellular matrix for host adhesion and colonization and as a source of nutrients (Hynes and Walton 2000). As such, small molecular weight hyaluronic acid is a potent immune activation agent in animals (Hynes and Walton 2000), whereas high molecular weight hyaluronic acid is anti-inflammatory (Toole 2004). In mammals, hyaluronidases in sperm interact with and degrade the egg extracellular matrix during sperm-egg fusion (Cherr, Yudin, and Overstreet 2001), and hyaluronidases are present in the venom of snakes, scorpions, bees, lizards, and numerous other animal species (Girish and Kemparaju 2007).

Choanoflagellates also possess extracellular matrix, and it is intriguingly found in the multicellular colonies formed by some species (Figure 2.4a). To investigate whether choanoflagellate extracellular matrix is homologous to animal extracellular matrix, we asked whether animal hyaluronidase is capable of modulating choanoflagellate colony morphology *in vivo*. We added purified animal testicular hyaluronidase to growing cultures of *Salpingoeca napiformis*. After 24 hours, we observed a striking effect: colonies in the hyaluronidase culture contained more cells per colony (Figure 2.4b). We confirmed this observation by counting the number of cells per colony in each treatment, and we discovered that the highest treatment concentration produced a culture with a significantly higher number of cells per colony than a vehicle control (Figure 2.4c), and that this did not appear to be the result of a simple increase in proliferation (data not shown). If the role of hyaluronidase is to break down the hyaluronic acid component of extracellular matrix, then the naïve expectation could be that choanoflagellate colonies should contain fewer cells. We observed the opposite effect, indicating that the physiological role of hyaluronidase in *S. napiformis* is more sophisticated than simply separating cells from one another within a colony. Nonetheless, the observation that hyaluronidase exerts an effect on choanoflagellate colony morphology via interaction with an extracellular matrix suggests that hyaluronidase may have played a role in the evolution of animals.

2.2.6 Genes lost in the lineage leading to animals are overrepresented for metabolic functions

Of the 19,520 gene families we estimate to have been present in the common ancestor of animals and choanoflagellates, 4,509 are not present in any animal genome we surveyed (see Figure 2.2). We explored the evolutionary history of these gene families by analyzing their gain and loss in a phylogenetic context (Figure 2.5a). Of the 4,509 gene families, 3,910 were present in the common ancestor of all the eukaryotic species included in our analysis, indicating that the bulk of these families are evolutionarily ancient. (We note that because early eukaryotic evolution is unresolved, any gene family present in representatives of more than one of the major eukaryotic supergroups is inferred to be present in the eukaryotic common ancestor.) Despite the observation that most families lost in animals are ancient, many of the gene families were lost independently in the fungal lineage and in the two major clades of choanoflagellates, craspedids and loricates, indicating that most of the gene families, while displaying some level of conservation within eukaryotes, are probably not involved in critical cellular activities present in all eukaryotes.

Identifying the function of gene families lost in animals that are not universally conserved within eukaryotes is complicated by the fact that they are likely to be less well studied than other

gene families. To produce a broad estimate of gene function, we identified 7,374 gene families associated with at least one Gene Ontology term within the 19,520 families present in the common ancestor of animals and choanoflagellates, of which 892 were annotated within the 4,509 gene families lost in animals. We found that animal losses were significantly enriched for several metabolic and molecular functions, including carbohydrate metabolism, carboxylic acid metabolism, glycoprotein biosynthesis, oxidoreductase activity and serine hydrolase activity (Figure 2.5b). Metabolic processes were significantly enriched overall ($p = 10^{-12}$): 683 of the 892 annotated gene families lost in animals were assigned a role in metabolism. A similar enrichment of metabolic function was also present in the subset of gene families specifically lost in animals and in no other lineage (data not shown). In contrast, genes lost in choanoflagellates were not enriched for metabolic processes. Instead, choanoflagellate losses were significantly enriched for transcription, DNA binding, and regulation of RNA metabolism, in accordance with the previous observation that *M. brevicollis* has streamlined its transcription machinery (King et al. 2008). Thus, since the divergence of choanoflagellates and animals from their common ancestor, animal metabolic processes have changed significantly, whereas those in choanoflagellates appear not to have changed to the same degree.

2.2.7 The biology of choanocytes, possibly the first cell type in animals

The cell morphology of choanoflagellates, consisting of an apical flagellum surrounded by a collar of microvilli, is also found in the choanocyte chambers of sponges (Figure 2.6a-c), suggesting that choanocytes may represent the earliest animal cell type (Nichols, Dayel, and King 2009). We analyzed our data set for the presence of gene families found only in sponges and choanoflagellates, as these are more likely than other genes to underlie the biology of choanocytes. Our comparison involved two genome-scale sponge data sets: predicted proteins from the genome sequence of *Amphimedon queenslandica* (Srivastava et al. 2010) and from transcriptome sequencing of *Oscarella* sp. (Nichols, Richter, et al. in preparation). Because *Oscarella* transcriptome data may be incomplete for genes that were not expressed at the time of tissue collection for RNA isolation, we focused initially on the set of 246 genes shared between *A. queenslandica* and at least one choanoflagellate (Figure 2.6d), using *Oscarella* data to confirm evidence for proteins of interest. We found that most of the 246 genes were lost in one of the four recognized clades of choanoflagellates, indicating that they were not likely to be critical components of choanocyte biology.

Among the gene families shared between sponges and many choanoflagellates, three gene families were of particular interest to us: 10396, 1066 and 7542 (Figure 2.6e-g; gene family identifiers are those produced by the prediction software we used). Gene family 10396 consists of predicted proteins encoding anywhere from three to nine leucine-rich repeat (LRR) domains, and is single copy in all but one species in our analysis. LRR domains are a structural motif often involved in protein-protein interactions, and are found in animal Toll-like receptors, as described above. Gene family 1066 contains predicted proteins with anywhere from one to 48 copies of the TIG domain, an immunoglobulin-like fold also implicated in protein-protein interaction (Bork et al. 1999). In contrast to gene family 10396, gene family 1066 has multiple members in all species we examined, with the exception of *A. queenslandica*. The presence of families of genes with repeated TIG and LRR domains is striking, but there were no other domains indicative of a more specific function in any of the predicted proteins within either family. To gauge the

physiological roles of these gene families in choanoflagellates and in sponges, we compared their expression level to that of all other predicted proteins. We find that the LRR domain-containing gene family 10936 is among the highest expressed genes within choanoflagellates and in *Oscarella*, whereas the TIG domain-containing gene family 1066 shows high variability of expression within choanoflagellates, varying anywhere from the 1st to the 98th percentile of gene expression (Figure 2.6e'-f'). We also examined expression data collected for different cell types in the colonial choanoflagellate *Salpingoeca rosetta* (Fairclough et al. submitted). Both gene families were significantly overexpressed in attached thecate cells (Figure 2.6h), indicating that they may be involved in the biology of attachment or in functions typically carried out by cells attached to a substrate. Furthermore, colonial cells of *S. rosetta* are the only sources of expression data in which protein family 10396 was not highly expressed. These initial data suggest that gene family 10396 is involved in a highly prevalent and conserved protein interaction process in choanoflagellates and in sponges, that 1066 may be involved in a similar but more variable process, and that both gene families are differentially regulated in different choanoflagellate life history stages.

The molecular activity of gene family 7542 was easier to identify: it is a non-receptor tyrosine kinase (Figure 2.6g). It is paired with a Calponin homology actin-binding domain (CH), a combination not detected in any protein outside of sponges and choanoflagellates. This tyrosine kinase gene family is highly expressed in almost all choanoflagellates and can be present in multiple copies within the same species, but is absent in *Oscarella* (Figure 2.6g'). Gene family 7542 was not significantly differentially regulated between cell types in *S. rosetta*. The conservation of this tyrosine kinase between choanoflagellates and animals stands in contrast to the divergence of domain architecture observed during the diversification of other tyrosine kinases in choanoflagellates and animals (Manning et al. 2008; Pincus et al. 2008) (although we note that the divergence of non-receptor tyrosine kinases is not as pronounced as that of receptor tyrosine kinases (Suga et al. 2012)), and indicates that its role may be critical to the biology of choanocytes. For each of these three gene families, sequence homology to genes outside the family was insufficient to permit the construction of a phylogenetic tree that would have allowed us to understand the relationship between the gene families and their closest neighbors.

2.2.8 RNAi machinery is present in choanoflagellates

The RNAi ribonucleases Argonaute and Dicer were previously believed to have been lost in choanoflagellates, based on their absence from the two sequenced choanoflagellate genomes (King et al. 2008), (Fairclough et al. submitted). In eukaryotes, Dicer generates small RNAs which associate with Argonaute family proteins to degrade or to repress the translation of mRNAs with sequences complementary to the small RNAs (Jinek and Doudna 2009). We find evidence based on diagnostic domains and on orthologous gene families for both Argonaute and Dicer in choanoflagellates, and we infer that both genes were likely lost in the common ancestor of *M. brevicollis* and *S. rosetta* (Figure 2.7). The pattern of loss of either Argonaute, Dicer, or both genes is prevalent in choanoflagellates. We infer that Argonaute was lost five independent times in choanoflagellates and that Dicer was lost seven times independently, although these could both be overestimates in the case that the genes were not expressed in our culture conditions. When Argonaute is present in a species, we find that it is highly expressed, and when Dicer and Argonaute are both present in the same species, Argonaute is always expressed at a

higher level than is Dicer. Curiously, we detect five species in which Argonaute is present and Dicer is absent or vice versa. These absences could reflect the presence of non-canonical RNAi genes or Dicer-independent RNAi pathways in choanoflagellates, as have been previously observed in budding yeast and in mice, respectively (Drinnenberg et al. 2009; Cheloufi et al. 2010). We also detect evidence in many choanoflagellates for homologs of a conserved Dicer-like helicase that is involved in Argonaute-mediated viral RNA silencing in the nematode *Caenorhabditis elegans* (Lu et al. 2009). These findings underscore the advantage of diverse phylogenetic sampling when using sequencing as a tool for gene discovery in choanoflagellates, and introduce RNAi as a possible approach to gene manipulation in choanoflagellates.

2.3 Discussion

Our results enhance current reconstructions of the gene content of the common unicellular or colonial ancestor of animals, in which many of the building blocks of animal gene families involved in adhesion, signaling, gene regulation, interaction with the extracellular matrix and innate immunity were already present. These building blocks can take the form of partial pathways that appear to have been elaborated in animals, as in the case of Toll-like receptor signaling or Notch, or they can occur as individual protein domains that were combined and diversified in the course of animal evolution, as we propose for the cadherin families we detected. In addition, none of the pathways that we have newly discovered in choanoflagellates exists in isolation in animals. It has been suggested that Flamingo cadherin links Frizzled to Notch in the eye of *Drosophila melanogaster* (Das, Reynolds-Kenneally, and Mlodzik 2002), and a flurry of recent evidence indicates that Notch is also significantly involved in the Toll-like receptor signaling pathway in animals (Palaga et al. 2008; Monsalve et al. 2009; Hindi et al. 2012; Zhang et al. 2012), and it is currently unknown whether these pathways are also linked in choanoflagellates.

We find that genes with known roles in animals can also modulate multicellular morphology in choanoflagellates, as in the case of hyaluronidase in colonies of *Salpingoeca napiformis*. Future studies will explore the relationship of this and similar findings to the biology and ecology of the common unicellular or colonial ancestor of animals and choanoflagellates. Indeed, as choanoflagellates are exposed directly to their surrounding environment, intercellular signaling or communication processes we discover to have been present in the common ancestor of animals and choanoflagellates suggest a link in communicating with the environment. For example, Flamingo cadherins are putative host receptors for the pathogenic bacterium *Streptococcus pneumoniae* (Blau et al. 2007), and it is known that the presence of bacteria or the signals they produce can modulate both gene expression and life history transitions in choanoflagellates (King, Hittinger, and Carroll 2003; Alegado et al. 2012). Genes we have newly identified as likely to represent specific functions in the biology of the choanocyte cell type may also be involved in interaction with the environment, either in the form of tyrosine kinase signaling pathways (Miller 2012), or through protein-protein binding interactions, which may be the case for the LRR domain and TIG domain-containing predicted proteins we discovered to be shared exclusively between sponges and choanoflagellates.

We have detected the RNAi components Argonaute and Dicer in choanoflagellates. The presence of RNAi machinery opens an avenue to gene manipulation in choanoflagellates, but its likely absence in many lineages poses questions regarding its biological role in choanoflagellates.

The genome of *M. brevicollis* contains very relatively few self-replicating repetitive elements (Martin Carr et al. 2008). Was the RNAi machinery lost after the successful elimination of retrotransposons from the lineage of choanoflagellates containing *M. brevicollis* and *S. rosetta*? Or was RNAi lost due to the acquisition of a beneficial virus, as in the case of the budding yeast *Saccharomyces cerevisiae* (Drinnenberg, Fink, and Bartel 2011)? Of the four most poorly growing choanoflagellate species in culture, three show evidence for Dicer expression, although this could be simple coincidence.

We have demonstrated that a sequencing approach based on the principle of maximizing phylogenetic diversity is an effective means to rapidly increase estimates of the gene content of the common ancestor of animals and choanoflagellates. We have analyzed this gene content to discover genes newly shared between choanoflagellates and animals and to begin to understand their functions in choanoflagellates, and we have highlighted the capability of several previously underappreciated classes of genes, those lost in animals and those shared exclusively between sponges and choanoflagellates, to illuminate the biology of animal origins.

Chapter 3: Culture, Sequencing, Transcriptome Assembly and Annotation of 19 Choanoflagellates

3.1 Origin of Cultures

We acquired the choanoflagellate cultures used for transcriptome sequencing either from the American Type Culture Collection (<http://www.atcc.org>), as a generous gift from Barry Leadbeater, University of Birmingham (UK), or by isolation from water samples (Table 3.1). For cultures we did not isolate, the original water sample source and collection date information were provided by the supplier. We note that the species names listed as provisional (“cf.”) are currently undergoing naming revision (Carr, Richter et al. in preparation).

We isolated two cultures, *Acanthoea spectabilis* (Virginia) and *Codosiga hollandis*, from water samples. *Acanthoea spectabilis* (Virginia) is a different isolate from *A. spectabilis* ATCC PRA-103, which was originally collected in Australia. We collected the water sample from which *Acanthoea spectabilis* was isolated on December 19, 2007 near Hog Island, Virginia (GPS coordinates: 37.472502, -75.816018) and we isolated *C. hollandis* from a sample collected on June 25, 2008 from Madeira, Portugal (GPS coordinates: 32.762222, -17.125833).

We began the isolation process by transferring the sample into a standard 100 millimeter by 15 millimeter petri dish. We scraped the sides of the original collection vessel with a cell scraper before transferring, since some choanoflagellate species can be found attached to substrates. We identified choanoflagellates using phase contrast microscopy on an inverted microscope using 10X to 40X objectives. We isolated individual cells from the petri dish by using a micromanipulator and a manual microinjector (Eppendorf PatchMan NP 2 and CellTram vario 5176 for *A. spectabilis*, and Sutter Instruments XenoWorks Micromanipulator and XenoWorks Analog Microinjector for *C. hollandis*). We pulled glass needles used for isolation from 1 millimeter diameter borosilicate glass (Science Products GmbH GB100-10) using a Flaming/Brown needle puller (Sutter Instruments P-87) with the following program: heat = 820, pull = 50, velocity = 140, time = 44. We polished and sterilized needles by passing them briefly through a low flame. We filled the manual microinjector with mineral oil prior to isolation and used a separate needle for each attempted isolation. To perform isolations, we withdrew the smallest possible volume containing the target choanoflagellate from the petri dish and transferred it to an individual culture flask containing appropriate growth medium (see Tables 3.3 and 3.4). In order to reduce the possibility of contamination during the isolation procedure, we generated sterile air flow across the microscope and micromanipulator apparatus using a HEPA-type air purifier (Holmes HAP412BN). Roughly 5-10% of individual choanoflagellate cells isolated in this manner survived the process to produce a clonal culture.

One culture, *Salpingoeca infusionum*, was contaminated by an unidentified biflagellated unicellular eukaryote, and we performed dilution to extinction in order to remove it from the culture. We scraped a growing culture in order to detach attached cells, and then diluted it at $1:10^6$ into each well of two 24-well plates containing 1 milliliter of culture medium. We left cells to grow for 7 days and then we scored the plates for the presence of *S. infusionum* and the contaminant. We found 4 of 48 wells containing only *S. infusionum* and bacteria, 1 well containing only the contaminant and bacteria, and the remaining 43 wells containing only

bacteria. We selected one of the four wells containing only *S. infusionum* for use in transcriptome sequencing.

3.2 Antibiotic Treatment and Optimization of Culture Conditions

Choanoflagellates are co-isolated with their bacterial prey, which serve as a food source. There may be up to several dozen species of bacteria that are initially co-isolated with each choanoflagellate culture. These bacteria compete both with the choanoflagellates and with one another for resources. In order to reduce the density of bacteria in each culture, and to reduce the occurrence of bacterial structures (such as biofilms) too large for choanoflagellates to phagocytose, we treated each culture with a panel of ten different antibiotics (Table 3.2). All antibiotics were obtained from Fisher Scientific with the exception of erythromycin, gentamicin, ofloxacin, and polymyxin B, which were obtained from Sigma Chemical. Antibiotic solutions were sterilized in a laminar flow hood by filtration through a 0.22 micrometer syringe filter (Fisher Scientific).

Because each culture contained a mixture of unidentified bacterial species, we could not predict *a priori* which treatments would be successful. Instead, we initially treated each culture with each antibiotic. We chose all initial treatments that decreased bacterial density and diversity and repeatedly passaged the cultures into new medium with fresh antibiotic until no further change in bacterial density or diversity was observed. We then re-treated each of these cultures with an additional round of all ten antibiotics, as the first round may have modified the composition of bacterial species in the culture and thus their response to antibiotics. We repeated rounds of treatment until no further reduction in bacterial density or diversity was observed. An example of pre- and post-antibiotic treatment cultures is shown for *C. hollandis* (Figure 3.1).

We tested different growth media for each choanoflagellate culture (Table 3.3) in order to maximize choanoflagellate cell density. A medium with too few nutrients does not allow for sufficient bacterial growth to serve as a choanoflagellate food source, but a medium that is too nutrient-rich will cause overgrowth of the bacteria at the expense of choanoflagellate density. We tested a range of concentrations for three types of nutrient sources: quinoa grains, proteose peptone/yeast extract, cereal grass (data not shown). We used filtered water when preparing all solutions, and as the base medium for freshwater species (Millipore Milli-Q). The base medium for marine species was 32.9 grams per liter of artificial sea water (Tropic Marin), which is equivalent to the salinity of sea water. We prepared stock solutions of 0.4% proteose peptone (w/v) (Sigma Chemical) and 0.08% yeast extract (w/v) (Becton Dickinson) and diluted them for use in growth medium at a ratio of 0.1 ml to 18.9 ml of artificial sea water. To prepare cereal grass media, we added cereal grass (also known as chlorophyll alfalfa powder, Basic Science Supplies) to autoclaved base medium and allowed it to steep until cool. Once cool, we removed large particles of cereal grass by repeated vacuum filtration through a ceramic Buchner funnel with a double layer of Grade 1 cellulose filter paper (Whatman). We sterilized all media with a 0.22 micrometer vacuum filter (Millipore Steritop). To prepare medium containing quinoa, we autoclaved organic quinoa grains and added them to the medium after filtration, with roughly 2 grains added per 25 cm² of culture vessel surface area. We found that small variations in the number of quinoa grains added did not significantly affect growth of choanoflagellate cultures (data not shown).

We noticed that cultures for two species, *C. hollandis* and *Salpingoeca amorphidum*, became acidic during growth, whereas all other cultures maintained a pH between 7 and 8. Cultures of *C.*

hollandis reached a pH of 5.5 and those of *S. amphoridium* reached a pH of 5. To test whether acidic pH was negatively affecting growth of the choanoflagellates, we prepared buffered medium with 50 millimolar HEPES (Fisher Scientific) adjusted to a pH of 7. We found that buffered medium ameliorated the growth of *S. amphoridium*, and so we adopted it as the standard growth medium for this culture. The choanoflagellates in cultures of *C. hollandis* died in the presence of buffered medium across a pH range we tested from 5 to 7 (data not shown), so we did not pursue buffering for *C. hollandis* cultures.

We performed all media preparation in a laminar flow hood. To reduce the possibility of cross-contamination among samples, we did not permit either antibiotic solutions or growth media within the laminar flow hood while any culture was also present within the hood, and we dispensed both antibiotics and media prior to the introduction of cultures into the hood.

Final nutrient content of each type of medium was measured by Flow Injection Analysis at the University of California, Santa Barbara Marine Science Institute (Table 3.3). Nutrient content was not determined for the 10% FCGM.3 medium.

In addition to antibiotic treatment and growth medium changes, we also tested a number of growth temperatures for each culture. Once we completed all testing, we arrived at a final set of optimal conditions (Table 3.4). We considered optimal conditions to be those that maximized choanoflagellate density and variety of cell types present, as we hypothesized that different cell types would yield the greatest diversity of transcripts for sequencing. We defined five generic cell types to describe choanoflagellates with varying methods of attachment and extracellular structures as follows: attached: attached to the bottom of the culture vessel or to a piece of floating debris, either directly, within a theca, within a lorica, or on a stalk. Slow swimmer: a swimming cell. Fast swimmer: a cell with a significantly reduced collar and swimming more quickly than a slow swimmer. Free-swimming colonial: in a colony swimming in the water column. Attached colonial: in a colony attached to a stalk. Passively suspended: suspended in the water column, either naked or within a lorica. See (Dayel et al. 2011) for further discussion on choanoflagellate cell types and, as examples, (Leadbeater 2008) and (Leadbeater and Cheng 2010) for descriptions of both types of loricate choanoflagellates (nudiform and tectiform).

3.3 Growth of Cultures in Large Batches in Preparation for RNA Isolation

We routinely grew choanoflagellates in 25 cm² angled neck cell culture flasks with 0.2 micrometer vented caps (Corning) containing 25 milliliters of medium. We performed all cell culture work in a laminar flow hood. To reduce the possibility of cross-contamination among samples, we only worked with a single culture at a time within the laminar flow hood, and we cleaned the hood thoroughly with 70% ethanol before each culture was brought into the hood. To grow up large batches of cells, we used different growth vessels, volumes, growth durations and centrifugation times as appropriate to each culture (Table 3.5). We used Delong neck Pyrex glass culture flasks (Corning), 150 millimeter plastic tissue culture dishes (Becton Dickinson), and 75 cm² angled neck cell culture flasks with 0.2 micrometer vented caps (Corning).

We harvested cells differently depending on the cell types present in the culture. For cultures with 5 percent or fewer attached cells, we pipetted out the medium without scraping, as this reduced the amount of bacteria collected. For cultures containing between 5 and 90 percent attached cells, we harvested choanoflagellates by scraping a single plate and collecting all of the growth medium. For cultures with 90 percent or greater attached cells, we combined multiple

plates as follows: first, we removed the culture medium from the first plate by pipetting. This procedure left a small film of medium remaining on the plate, which we expected to capture a substantial fraction of swimming cells, as these were generally found near the surface of the plate (data not shown). Once we removed the culture medium, we added 50 milliliters of either artificial sea water or filtered water, as appropriate, scraped cells from the plate, removed the medium from the subsequent plate, transferred all 50 milliliters to the next plate, and continued until all plates were combined. We chose this method of combining in order to minimize the number of bacteria collected (as they dominate the water column), to maximize the choanoflagellate cells available for harvesting, and to reduce the amount of growth medium present, as we found that even small amounts of residual growth medium in cell pellets from some cultures could contain significant concentrations of RNA-degrading agents (data not shown). To eliminate quinoa grains or large bacterial debris, some cultures were filtered with a 40 micrometer cell strainer (Fisher) during collection.

We spun harvested cultures in 50 milliliter conical tubes at 3220 x g in a refrigerated centrifuge at 4 degrees Celsius. When we harvested more than 50 milliliters for a culture, we split the culture into different conical tubes, centrifuged, and removed all but 2.5 milliliters of supernatant from each tube. We then resuspended each pellet and combined them into a single 50 milliliter conical tube, then centrifuged again with the same conditions.

After the final centrifugation, we removed the first 47.5 milliliters of supernatant by pipetting, and we removed the last 2.5 milliliters by aspiration. We flash froze cell pellets in liquid nitrogen and stored them at -80 degrees Celsius.

To reduce the possibility of cross-contamination among samples, we centrifuged each culture independently, we discarded disposable plastic ware after each use, and we cleaned all other material (bench tops, pipettes, etc.) with ELIMINase (Decon Laboratories) before harvesting each culture.

3.4 RNA Isolation

We performed total RNA isolation from cell pellets with the RNAqueous kit (Ambion). We used the manufacturer's protocol with one modification: we doubled the amount of lysis buffer used, as we found that in some cases this increased yield and decreased RNA degradation (data not shown). We performed both optional steps in the protocol before filter binding: to eliminate unlysed debris (which was largely bacterial in origin), we spun the lysate for 3 minutes at top speed at 1 degree Celsius in a refrigerated bench top centrifuge, and to reduce the viscosity of the supernatant from this spin, we passed it through a 25 gauge syringe needle several times. We used the minimum suggested volumes of elution solution in two elution steps (40 microliters and 10 microliters). We reserved 2 microliters of eluate for quantification and for subsequent PCR on genomic DNA (which is also isolated by the RNAqueous kit). We measured RNA concentration using a spectrophotometer (NanoDrop ND-1000).

For all species except *C. hollandis*, we immediately proceeded to digest genomic DNA using the TURBO DNA-free kit (Ambion). We followed the manufacturer's protocol with a 30 minute incubation. After digestion, we removed DNase enzyme with DNase Inactivation Reagent, except for *Salpingoeca amphoridium*. We found that the *S. amphoridium* culture was incompatible with the DNase inactivation reagent (data not shown), so we instead removed DNase by extracting with two volumes of pH 8 phenol:chloroform:isoamyl alcohol, removing

residual phenol with two volumes of chloroform:isoamyl alcohol, and precipitating with 0.3 M sodium acetate pH 5.2 (all three from Sigma), 25 micrograms of GlycoBlue (Ambion) as a carrier and two volumes of 100% ethanol. We washed the pellet in 70% ethanol and resuspended in 50 microliters of nuclease-free water. We also observed that the RNA from *Didymoeca costata* appeared to be contaminated with protein after DNase removal with DNase Inactivation Reagent (data not shown), so we performed a pH 8 phenol:chloroform extraction as just described to remove the excess protein.

For *C. hollandis*, we observed that the total RNA degraded significantly in the presence of DNase buffer (Figure 3.2), so we were unable to perform DNase digestion. Instead, we performed three successive rounds of extraction with pH 4.5 phenol:chloroform:isoamyl alcohol (Sigma), as DNA should not be soluble in the aqueous phase of acid phenol:chloroform. After three rounds of extraction, we performed the chloroform:isoamyl and precipitation steps as described above.

To reduce the possibility of cross-contamination among samples, we performed RNA isolation and DNase digestion for a single culture at a time, we used disposable materials when possible, and we cleaned all other materials (bench tops, centrifuges, dry baths, pipettes, etc.) thoroughly with ELIMINase before use.

To evaluate total RNA quality, we ran all samples on Bioanalyzer 2100 RNA Pico chips (Agilent). We considered high-quality total RNA to fulfill four criteria: (1) four distinct ribosomal RNA peaks (16S and 23S for bacteria, 18S and 28S for choanoflagellates), (2) low signal in all other regions, as a non-ribosomal signal is likely evidence of degradation, (3) at least a 1:1 ratio of 28S ribosomal area to 18S ribosomal area, as 28S ribosomal RNA is likely to degrade more easily than is 18S ribosomal RNA, and (4) an RNA Integrity Number, as measured by the Bioanalyzer instrument, of 7 or greater (Schroeder et al. 2006). (We note that the Bioanalyzer software could not calculate RIN for several cultures.) If we were not able to obtain high-quality total RNA after the first attempt for any culture, we repeated cell growth, centrifugation and total RNA isolation up to a maximum of 5 times. Based on quality measured in Bioanalyzer traces, we selected the best available total RNA sample for each culture to use for transcriptome sequencing (Figure 3.3).

3.5 Confirmation of Species Identity by DNA Sequencing

In order to confirm the species identification of our cultures, we attempted to sequence the 18S small subunit ribosomal DNA locus for each culture. For most cultures, we were able to take advantage of the fact that the RNAqueous kit also isolates genomic DNA, and so we reserved 2 microliters from each RNA isolation for subsequent PCR of genomic DNA. For other cultures, we instead isolated genomic DNA from separate cell pellets that were collected as described above (Table 3.6).

For cultures whose genomic DNA was isolated from a separate cell pellet, we used the same procedures for cell collection as for the cell pellets collected for RNA isolation (see Table 3.5). We performed a standard phenol:chloroform DNA extraction, as follows: we resuspended pellets in lysis buffer (10 mM Tris (Fisher), 0.1 mM EDTA, 0.0005% SDS (both Sigma), pH 8), followed by treatment with 20 milligrams per milliliter RNase A (Sigma) and 100 milligrams per milliliter Proteinase K (Fisher). We extracted DNA with two volumes of pH 8 phenol:chloroform:isoamyl alcohol and repeated the extraction until no protein was present at the

interphase, then removed residual phenol with two volumes of chloroform:isoamyl alcohol and precipitated with 1 M ammonium acetate (all three from Sigma), 25 micrograms of GlycoBlue (Ambion) as a carrier and two volumes of 100% ethanol. We washed the pellet in 70% ethanol and resuspended in nuclease-free water.

We amplified the small subunit ribosomal locus from genomic DNA using PCR with Taq polymerase (New England Biolabs), the proofreading enzyme Pfu (Agilent) and universal eukaryotic ribosomal primers 18S_1F (5' AACCTGGTTGATCCTGCCAGT 3') and 18S_1528R (5' TGATCCTTCTGCAGGTTACCTAC 3') (Medlin et al. 1988) using the following program: 94° C for 3 minutes, 35 cycles of: 94° C for 30 seconds, 52° C for 30 seconds, 72° C for 3 minutes 30 seconds, followed by a final extension at 72° C for 10 minutes. We either separated PCR products by gel electrophoresis and extracted bands of the appropriate size with the QIAquick Gel Extraction Kit, or we purified them with the QIAquick PCR Purification Kit (both from Qiagen).

Some choanoflagellate small subunit ribosomal DNA loci were 2 to 3 kilobases in length, and so we cloned these with the TOPO TA Cloning vector (Invitrogen) before sequencing. We submitted shorter sequences for direct sequencing of the purified PCR product (Table 3.6). Sequencing was performed at the UC Berkeley DNA Sequencing Facility (Berkeley, California).

For cultures cloned into the TOPO TA Cloning vector, we sequenced using two standard primers within the vector sequence: M13F (5' GTAAACGACGGCCAGTG 3') and M13R (5' CAGGAAACAGCTATGACC 3'). For cultures with longer loci, we designed the pan-choanoflagellate 18S sequencing primers 18S_564F (5' AATTCCAGCTCCAATAGC 3') and 18S_1205R (5' ATGTCTGGACCTGGTGAG 3'). For some species, we also required additional primers. For *Salpingoeca gracilis*, we bypassed an internal polynucleotide repeat by sequencing with the following three primers: Sg_18S_1 (5' CCTTCAGACGAATCATCGAG 3'), Sg_18S_2 (5' TGAGAACAAACCGCCAAAG 3') and Sg_18S_3 (5' AATGCCTGCTTTGAACACTC 3'). For *Salpingoeca napiformis*, which had a particularly long locus, we used the additional primers 18S_528F (5' CGGTAATTCCAGCTG 3'), 18S_1220F (5' GCGATAGTTGGGTGTGGAAG 3'), 18S_1200R (5' GGGCATCACAGACCTG 3') and 18S_1871R (5' TCCTGACACTTCCGGTTTC 3').

We base called sequence reads using phred version 0.0210425.c (Ewing et al. 1998; Ewing and Green 1998) with the '-trim_alt' option and all other parameter values set to their defaults, and aligned them using FSA version 1.15.0 (Bradley et al. 2009) with default parameter values.

In all cases, if there was 18S sequence available in GenBank for a given species, it matched the sequence we generated.

3.6 Test of PolyA Selection to Separate Choanoflagellate from Bacterial RNA

The standard preparation protocol for Illumina mRNA sequencing involves the use of a poly-dT sequence bound to beads in order to select polyadenylated mRNA. In the case where only eukaryotic mRNA is being sequenced, this step serves to purify mRNA away from other types of RNA such as rRNA and tRNA. In the case of choanoflagellates that are co-cultured with their bacterial food source, the polyA selection step also serves to separate choanoflagellate mRNA from bacterial mRNA, which is not polyadenylated. Because the amount of bacterial RNA isolated from a choanoflagellate culture often exceeds the amount of eukaryotic RNA by one to several orders of magnitude, we reasoned that the standard polyA selection step might not be

sufficient to remove bacterial RNA in the case of choanoflagellates.

To test this hypothesis, we performed an experiment on *Salpingoeca rosetta* Px1, a culture containing a species of choanoflagellate grown with a single species of bacterium, *Algoriphagus machipongonensis*. Because the genome sequences of both species have been determined (Alegado et al. 2011), (Fairclough et al. submitted), we could identify the origin of sequenced RNA using a straightforward mapping procedure.

S. rosetta Px1 cells were cultivated as described previously (Fairclough, Dayel, and King 2010). We isolated *S. rosetta* Px1 cells from 50 milliliters of a three day old culture in a 150 millimeter tissue culture dish. We removed attached cells from the plate surface by scraping. We performed centrifugation (with a 10 minute spin), RNA isolation, DNase digestion and total RNA quality assessment as described above.

The standard Illumina TruSeq v2 mRNA preparation protocol involves two rounds of polyA selection with a single set of poly-dT-coated beads. We compared this against a modified protocol that repeats the polyA selection steps, for a total of four rounds of polyA selection with two sets of beads. For all other aspects of library preparation, we followed the manufacturer's protocol. Libraries were quantified by qPCR (Kapa Biosystems) and sequenced on a HiSeq 2000 machine (Illumina) at the Vincent J. Coates Genomics Sequencing Laboratory at the California Institute for Quantitative Biosciences (Berkeley, CA).

A total of 16,970,914 single-end 50 base pair reads were generated for the mRNA prepared with two rounds of polyA selection, and 17,182,953 were generated for the mRNA prepared with four rounds of selection. We mapped reads to the *S. rosetta* genome and to the *A. machipongonensis* genome using BWA version 0.6.1 (Li and Durbin 2009) with default parameter values and we determined read counts mapping to each genome using SAMtools version 0.1.18 (Li et al. 2009) with default parameter values. In addition, we used the *S. rosetta* genome browser hosted at the Broad Institute (http://www.broadinstitute.org/annotation/genome/multicellularity_project/FeatureSearch.html) to query the position of the 5S, 18S and 28S ribosomal RNA loci. We chose one ribosomal locus on supercontig 1.8 and downloaded the associated sequence for each (5S: 1900295-1900408, 18S: 1914756-1916850 and 28S: 1917502-1923701) to use as a mapping target.

The number of reads mapping to the non-ribosomal portion of the *S. rosetta* genome did not differ significantly between the two data sets: 12,737,031 reads mapped for the two round data, and 12,585,647 for the four round data. Similarly, 10,509,262 reads from the two round data mapped to *S. rosetta* transcripts (which were also downloaded from the Broad Institute web site), and 10,181,522 for the four round data.

The four round method removed roughly an order of magnitude more non-polyadenylated RNA than the two round method (Figure 3.4). We also recognized the possibility that two additional rounds of polyA selection could lead to lower coverage of the 5' ends of transcripts, as polyA selection is performed at the 3' end and individual RNA molecules could suffer breakage due to pipetting or heating during the selection process. To test this, we plotted transcript coverage from both experiments against transcript length, which indicated that only a subtle loss of 5' end coverage occurred (Figure 3.5a).

We did not normalize read counts due to the possibility that the significantly different amount of ribosomal RNA reads that we expected to be present in each data set could introduce a normalization bias, but the results did not change using normalized read counts (data not shown). We also observed that the four round data set, while it consisted of more total reads, had a slightly lower overall read quality. To address this, we tested whether a difference in overall read

quality between the two data sets could account for the difference in read mapping by randomly resampling the two round data set to contain the same number of Phred-like quality 20 or quality 30 bases as the four round data set, but neither resampling affected the results (data not shown).

We also tested whether transcript assembly quality would suffer in the four round data set by assembling both data sets *de novo* with Trinity release 2012-03-17 (Grabherr et al. 2011) with default parameter values and mapping the resulting contigs to the *S. rosetta* genome using BLAT version 34 (Kent 2002) with default parameter values, but we observed no significant difference between the two data sets (Figure 3.5b).

Given the superior ability of four rounds of polyA selection to remove contaminating bacterial RNA with little to no loss of transcript coverage, we adopted this methodology for subsequent transcriptome sequencing.

3.7 Library Preparation and mRNA Sequencing

We attempted to begin the TruSeq v2 mRNA library preparation protocol (Illumina) with approximately 2 micrograms of total RNA per culture. While it would have been possible to begin with less total RNA, we reasoned that starting with a relatively large amount would insure against the possibility that we might lose coverage due to degraded RNA (which could occur during the RNA isolation process or during the library preparation process). We produced a rough estimate of the amount of choanoflagellate total RNA for each sample by estimating the ratio of choanoflagellate ribosomal RNA peaks to those of the bacteria in our Bioanalyzer traces and multiplying the resulting fraction by the total amount of RNA present in the sample (Table 3.7).

We performed the TruSeq v2 mRNA protocol for total RNA from each culture with four rounds of polyA selection (instead of the standard two) and two further modifications: first, we performed the AMPure XP (Agencourt) bead clean-up step following the PCR enrichment step one additional time in order to further remove adapters, and second, we eluted with 1.5 fewer microliters in all bead elution steps, in order to reduce volume loss during the protocol. Because we performed the TruSeq protocol on 5 RNA samples at a time, and because the samples were later multiplexed into groups of 6 or 7 per sequencing lane, we were concerned about the possibility of cross-contamination during both processes. To allow us to detect evidence of cross-contamination if it were to occur, we ensured that the groupings for sample preparation and sequencing were distinct (Table 3.7).

We estimated each library's concentration using the Qubit dsDNA HS Assay (Invitrogen) and validated their quality and determined their fragment size distribution with a Bioanalyzer 2100 High Sensitivity DNA chip (Agilent).

Libraries were quantified by qPCR (Kapa Biosystems) and sequenced on a HiSeq 2000 machine (Illumina) at the Vincent J. Coates Genomics Sequencing Laboratory at the California Institute for Quantitative Biosciences (Berkeley, CA).

3.8 Quality Trimming

Between 23,189,370 and 64,292,051 paired-end 100 base pair sequence reads were generated per library (Table 3.8). Sequencing group 1 (see Table 3.7) was run twice due to an issue with

another lane on the same sequencing chip.

We assessed quality of the data set using FastQC version 0.10.1 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). While overall quality of the data set was high, we observed that the first 10 bases of each library displayed a skewed base content with no concomitant drop in Phred-like quality (Figure 3.6a, b). We had previously observed a phenomenon wherein the first 10% of bases within reads were the source of a disproportionate number of sequencing errors without a corresponding drop in quality (Gerhold et al. 2011). These two observations led us to excise the first 10 bases of each read in subsequent steps.

We performed read trimming using Trimmomatic version 0.22 (Lohse et al. 2012) and implemented three separate filters: (1) removal of TruSeq adapter sequence, (2) removal of the first 10 bases of each read, and (3) trimming low quality bases from the ends of each read. To accomplish this, we ran Trimmomatic in three phases. In the first phase, we clipped palindromic adapters using the directive ILLUMINACLIP:2:40:15, we removed the first ten bases using HEADCROP:10 and we discarded resulting reads fewer than 25 bases with MINLEN:25. This resulted in two data sets: one set with reads whose mate pair remained in the set, and the other composed of forward reads whose reverse pair was removed due to adapter contamination (no reverse unpaired reads remained, as they would have been removed by the adapter clipping). The second phase operated on the remaining paired data set (which represented the overwhelming majority of reads). We clipped simple adapters using the directive ILLUMINACLIP:2:40:15, we imposed a minimum Phred-like quality cutoff on the first ten and last ten bases using LEADING:10 and TRAILING:10, subjected the read to a minimum sliding window quality using SLIDINGWINDOW:4:20 and discarded resulting reads shorter than 25 bases with MINLEN:25. The third phase operated on the unpaired forward reads from the first phase, and implemented the same directives as the second phase. We chose a minimum read length of 25 because that is the k-mer length for the *de novo* assembly program we used, Trinity, and so reads shorter than 25 bases would not be included in assemblies. In all adapter clipping operations, we used adapter sequences appropriate to the index used for multiplexed sequencing (see Table 3.7). When clipping palindromic adapters, we used as input the sequence of the TruSeq Universal Adapter and the reverse complement of the index-specific adapter. When clipping simple adapters, we used as input the sequence of both universal and specific adapters and their reverse complements. These read trimming steps produced a data set with overall higher quality (Figure 3.6c, d). To confirm that trimming had no deleterious effect on our ability to reconstruct full length transcripts, we also ran our assembly pipeline on the pre-trimmed reads for each species. We found no significant difference in the length of assembled transcripts: there were only roughly 5-10% more contigs of length 500 or greater in the assemblies from pre-trimmed reads, nearly all of which we estimate to be the result of spuriously predicted alternative splice forms due to sequencing error (data not shown).

3.9 *de novo* Transcript Assembly

We performed *de novo* transcript assembly for each species with Trinity release 2012-10-05 (Grabherr et al. 2011). To produce input to Trinity, we concatenated all files resulting from Trimmomatic into a single file using Fastool version 0.1.3 (<https://github.com/fstrozzi/Fastool>) with the '--illumina-trinity' option. Because we produced a single file using Fastool, we specified the '--run_as_paired' option to Trinity. When running Trinity, we set all other parameters to their

default values except for the minimum contig length, which we specified to be 150 bases rather than the default of 200 bases. We chose this value because we did not want to exclude assembled transcript fragments of lengths 150-199, which might produce predicted proteins of up to length 50-66 amino acids, from subsequent analyses. We based the minimum value of 50 amino acids on an analysis of protein domain lengths in the Pfam version 26.0 database (Punta et al. 2012). We analyzed the average match length (the 'average_length' field of the 'pfamA' table) for all Pfam domains and found that of the 5,571 protein domains present in any animal species, 408 (7%) had a length from 50-66 amino acids. Thus, we chose a minimum contig length of 150 to allow the possibility of including transcript fragments with these domains. There were also 356 protein domains with average length 33-49 and 205 domains with average length shorter than 33, but we found that reducing the minimum Trinity contig length below 150 produced a number of contigs that was untenable for subsequent analyses (data not shown).

3.10 Identification and Removal of Cross-Contamination

Although we took numerous precautions to guard against the possibility of cross-contamination among species in our data set, we nevertheless examined the assembled Trinity contigs for evidence of cross-contamination. We ran the *blastn* program from the BLAST package version 2.2.26 (Altschul et al. 1997) with an expectation value of 1×10^{-10} to match contigs from each species against all other species. In order to separate a contaminating sequence from a truly homologous one, we were forced to make heuristic decisions in choosing cutoffs for two properties of the BLAST hits: percent match and match length. In examining the BLAST hits, we reasoned that perfect matches between two species should be the result of cross-contamination, as any two choanoflagellate species are distantly related to one another at the nucleotide level (M Carr et al. 2008; Nitsche et al. 2011). We also reasoned that highly conserved genes would be nearly identical but not perfect matches between any two species. To determine whether we could separate these two categories of matches, we examined the percent match distribution of all cross-species BLAST hits (Figure 3.7a). We found that the majority of putative cross-contaminated matches were perfect. However, sequencing errors are more likely to affect the final sequence of a cross-contaminated contig rather than a non-contaminated contig, as only a small number of reads should have found their way into the contaminated assembly. Therefore, percent identities slightly less than 100% are expected for cross-contaminated contigs. We found that there appeared to be a separation between the number of BLAST hits at less than 96% percent identity and the number at 96% or greater, and so we chose a minimum of 96% identity as our first cutoff. We explored match length in a similar manner (Figure 3.7b). The distribution of match lengths was more difficult to interpret. As 90 bases was the maximum length of our quality-trimmed reads and we would expect cross-contamination to occur on a read-by-read basis, we chose a minimum match length of 90 bases as our second cutoff.

Using these two cutoffs, we calculated the number of matches between all pairs of species (Figure 3.8a). We found that cross-contamination did occur, but only within each sequencing lane, because species that were part of same library preparation group (the other likely source of cross-contamination) but not part of the same sequencing group did not experience significant levels of cross-contamination, and vice versa. Cross-contamination within a sequencing lane is a known phenomenon, and it is estimated to cause incorrect assignment of roughly 0.5% of index

pairs within a sequencing lane (Kircher, Sawyer, and Meyer 2012). For most applications, such as estimating expression differences by mapping reads to a sequenced genome, this does not cause a significant issue, but in our case essentially all reads assigned to an incorrect index will become part of the resulting *de novo* assembly.

Our discovery that cross-contamination occurred almost exclusively within a sequencing lane largely exonerated all of our other protocols as potential sources of contamination, but it was still necessary for us to attempt to eliminate the contamination that did occur within each sequencing lane. One previously suggested method is to examine the quality of index reads (Kircher, Sawyer, and Meyer 2012), and so we explored this option first. The underlying hypothesis is that inappropriate index assignments may be caused by errors in the separate sequencing reads produced for each fragment that determine which adapter was used during library preparation. We implemented the previously proposed cutoff of discarding any sequence read whose index read had a single base below a specified Phred-like quality (Kircher, Sawyer, and Meyer 2012). In addition, we implemented an alternative cutoff to discard sequence reads whose index reads fell below a specified average quality.

To test the minimum quality cutoff and average quality cutoff strategies, we focused on the six species multiplexed together in group 1 (see Table 3.7). We applied a quality cutoff to all of the index reads, assigned them to adapters based on their six base pair sequence, and produced separate *de novo* assemblies for each using Trinity as described above. Within each cutoff strategy, we varied two parameters: first, the value of the Phred-like quality cutoff, and second, whether we accepted only exact matches to each index, or whether we accepted a single mismatch (the default Illumina protocol is to accept a single mismatch in the index). For the minimum quality cutoff strategy with exact matches, we tested Phred-like qualities of 0 (that is, only requiring an exact match), 15, 20, 25, and 30, and a minimum quality of 20 when allowing a single mismatch. For the average quality cutoff strategy with exact matches, we tested qualities of 15, 20, 25 and 30, and qualities of 25 and 30 when allowing a single mismatch. We were not able to produce a combination of parameters that eliminated cross-contamination, as assayed by the all versus all BLAST strategy described above. We found that even the most aggressive strategy, a minimum quality cutoff of 30 with an exact index match, still allowed 386 contaminated contigs while discarding over 20% of the data (Figure 3.9). Therefore, we did not use the index quality cutoff method to remove cross-contamination from our sequence data.

Instead, we reasoned that if cross-contamination occurred due to the misreading of a small number of index reads that produced a set of assembled cross-contaminated contigs, there should be a large discrepancy between the number of sequence reads mapping to the contigs that originated from the species that was the source of contamination versus the species that was contaminated. To test this hypothesis, we aligned all reads within a sequencing lane to all of the contigs produced within the lane. While implementing the alignment process, we noticed that there were a small number of contigs with ten thousand or more reads mapping from all species sequenced within the same lane. These contigs all had relatively short tandem repeats, and we observed that each of our sequence data sets also contained a relatively large number of reads (on the order of 10^5) passing our quality trimming filters but also containing short tandem repeats. As these tandem repeats would interfere with our measurement of the source of contamination, we masked contigs with Tandem Repeats Finder version 4.04 (Benson 1999), with the following parameter values: match = 2, mismatch = 7, indel penalty = 7, match probability = 80, mismatch probability = 10, min score = 30, max period = 24. All parameter values were taken from the default values on the Tandem Repeats Finder web site, except for the minimum score to report

and the maximum period of the tandem repeat. We chose 24 as the maximum period of the tandem repeat, as we wanted it to be smaller than the k-mer size of 25 used in Trinity, and we chose 30 as the minimum score to report, as this would be the score of a 24-base long sequence containing tandem repeats with 2 bases not corresponding perfectly to the predicted repeat pattern (a score of 2 times 22 matches minus a score of 7 times 2 mismatches or indels).

We mapped reads to masked contigs using the Burroughs-Wheeler Aligner, BWA, version 0.6.2 (Li and Durbin 2009) and we determined read mapping counts using SAMtools version 0.1.18 (Li et al. 2009). When running BWA, we required perfect read matches using '-n 0 -k 0' in order to ensure that only truly cross-contaminated reads, and not reads from highly conserved genes in another species, would map. All other BWA parameter values were set to their defaults. Using the read mappings, for each putatively cross-contaminated contig we identified using the BLAST strategy described above, we counted the number of reads mapping from the species that produced the contig and compared it to the maximum number of reads mapping from any single other species within the same sequencing lane. We chose the maximum value because we did not know *a priori* the source of contamination within the lane. Using the ratio of these two read counts, we found that we could identify the true species of origin for nearly all of the putatively cross-contaminated contigs (Figure 3.10). This observation also confirmed the high specificity of the BLAST match cutoffs we applied to identify cross-contaminated contigs: 97% of contigs we identified as putatively cross-contaminated had at least an order of magnitude difference in the number of source species versus contaminant species reads mapping.

To eliminate cross-contaminated contigs, we removed all contigs identified as putatively cross-contaminated that did not have at least 10 times as many reads mapping from the species that produced the contig as from any other species, with one exception: if a contig had at least 10,000 reads mapping from the species that produced the contig, we did not discard it, regardless of read mapping ratio. We applied this exception because we observed that the most highly expressed transcripts (for example, alpha tubulin and elongation factor 1 alpha) also tended to be the most conserved, and thus the read mapping ratio was often close to 1 for these contigs. The decontamination process resulted in the removal of between 0.9% and 5.4% of contigs for each species (Table 3.9) and produced a data set with substantially lower levels of remaining contigs identified as cross-contaminated (Figure 3.8b). Contigs remained identified as cross-contaminated for two reasons: first, some contigs identified as cross-contaminated within a sequencing lane survived removal because they had at least 10,000 reads mapping, and second, cross-contaminants observed between lanes were explicable due to biological reasons. For example, we sequenced two isolates of *Stephanoeca diplocostata*, so hits between the two assembled transcriptomes are expected. In addition, matches between lanes could be *bona fide* conserved sequences between the species (as was the case for several dozen transcripts in the tectiform loricates, *A. spectabilis*, *Helgoeca nana* and *Savilla micropora*). Finally, some of the same (as yet unidentified) bacterial species were present in some pairs of cultures, and small amounts of their RNA survived the polyA selection process (because they contained a stretch of adenosines) and were sequenced and assembled.

3.11 Prediction of Amino Acid Sequences from Assembled Transcripts

We used Transdecoder release 2012-08-15 (<http://transdecoder.sourceforge.net/>) to predict amino acid sequences from decontaminated assembled transcripts, with a minimum protein

sequence length of 50 (see discussion in 'de novo Transcript Assembly' above) and all other parameter values set to their defaults. We noticed that many of the resulting predicted proteins coming from different contigs within a species were identical. Furthermore, we also observed many contigs whose predicted proteins were a subset of the predicted proteins from another contig. For example, contig 1 could have predicted proteins A and B, and contig 2 could have two predicted proteins exactly matching A and B, and a third predicted protein C. We removed both types of redundancy (exact matches and subsets) from the data set of predicted proteins, and we also removed the contigs from which they were predicted. This process eliminated between 1.4% and 19.2% of contigs from the set of decontaminated contigs within each species (Table 3.9). We also noted a correlation between number of contigs and the percentage of contigs identified as redundant ($R^2 = 0.31$). We speculate that the source of the redundancy was due to the Trinity assembler: as part of the assembly process, incompatible sequences for the same contig are separated into putative alternative splice forms. Incompatible sequences can arise from true alternative splice forms or from sequence errors, and we suspect that the sources of redundant predicted proteins were enriched for the latter case. This does not explain all of the cases, however, as the proportion of redundant contigs was only slightly correlated with the mean quality of the initial data set ($R^2 = 0.11$).

Assemblies for three cultures had a substantially higher number of transcripts remaining after the removal of redundancy than the others: *C. hollandis* and both *S. diplocostata* isolates. In the case of *C. hollandis*, the high number of contigs resulted from incomplete removal of bacterial genomic sequence with stretches of adenosines, as we were not able to perform a DNase digestion on the RNA from this culture (see 'RNA Isolation' above). For *S. diplocostata*, we were able to eliminate the possibility that there was a high level of polymorphism present within each culture that caused the assembler to produce two contigs for each haplotype (data not shown). In addition, after the removal of nearly identical proteins from each data set (Table 3.9; see 'Construction of Orthologous Groups of Proteins to Estimate Ancestral Gene Content' below), the two *S. diplocostata* assemblies had a number of remaining predicted proteins close to the median for all species, so we infer that the additional contigs are an artifact of some part of the *de novo* assembly process.

3.12 Estimation of Completeness of Predicted Proteins

To estimate the completeness of each of the predicted protein sets, we compared our data to a set of 458 core eukaryotic proteins from CEGMA (Parra, Bradnam, and Korf 2007). We used hmmscan from the HMMER 3.0 package (<http://hmmer.janelia.org>) with default parameter values to search each profile HMM from CEGMA against our decontaminated non-redundant predicted protein sets. In addition, for comparison, we also searched the predicted protein sets of the two previously sequenced choanoflagellate genomes, *Monosiga brevicollis* (King et al. 2008) and *Salpingoeca rosetta* (Fairclough et al. submitted). After examining the distribution of expectation values from the resulting hits (data not shown), we chose a conservative expectation value of 1×10^{-20} as a maximum to retain hits.

Our predicted protein sets were remarkably complete. By choosing only the best matching predicted protein within each assembly (by expectation value) to each core eukaryotic protein profile, we found that all of our assemblies matched 70% or greater of core protein profiles over at least 80% of their length (as measured by the proportion of the profile aligned by hmmscan).

Furthermore, we found that the level of coverage estimated from our transcriptomes was generally at least as complete as the two previously sequenced choanoflagellate genomes, indicating that our lack of perfect coverage of core eukaryotic proteins was likely the result of failure to detect homology rather than the result of incomplete reconstruction of transcript sequences (Figure 3.11a). To test whether lack of coverage of some core eukaryotic proteins was due to a fragmentary assembly in which the same core eukaryotic protein could be represented by two or more contigs in the assembly, we computed the same coverage metrics for all matching predicted proteins within each assembly, rather than just the best match. We then measured the improvement over having used only the best match, and we found that, for all assemblies, there was less than a 10% improvement in coverage of core proteins over at least 80% of their length (Figure 3.11b).

To measure the level of transcript fusions within our assembled data set, we also performed a comparison to CEGMA. We counted the number of cases in which two or more core eukaryotic proteins had their best match to the same predicted protein. We found the rate of putative fusions to be very low, with all species having 5 or fewer fusions out of the 458 protein set, and 15 having 2 or fewer. In comparison, *M. brevicollis* and *S. rosetta* each had three putative fusions (data not shown).

There are two slight caveats to our estimates using CEGMA: first, highly conserved proteins are generally more likely to be highly expressed and therefore easier to assemble. Second, the average length of the core eukaryotic protein profile HMMs in CEGMA is 436 amino acids whereas the average length of predicted proteins in the complete genome of *Monosiga brevicollis*, for example, is 599 amino acids, so our measurements may slightly underestimate the completeness of longer proteins that might be expected to be present. Nonetheless, we believe our estimates of completeness are robust enough that our predicted protein sets are an accurate representation of the proteins present in each of the cultures that we sequenced.

3.13 Construction of Orthologous Groups of Proteins

In order to produce a set of orthologous groups of proteins for comparison, we chose a set of representative animals and outgroup species (Table 3.10). We used all available choanoflagellate data, including the predicted protein sets from all 19 species we sequenced and the complete predicted protein sets from the *Monosiga brevicollis* (King et al. 2008) and *Salpingoeca rosetta* genomes (Fairclough et al. submitted). We found that the two *Stephanoeca diplocostata* isolates that we sequenced contained essentially equivalent predicted protein sets (data not shown), so we only used the Australian isolate for comparison. We chose 21 representative animals to compare to the 21 species of choanoflagellates we used. We included all 5 available early branching non-bilaterian animal species with sequenced genomes. In addition, in order to include more than one sponge genome, we included proteins predicted from the transcriptome sequence of the homoscleromorph sponge *Oscarella sp.* (Nichols, Richter et al. in preparation), which was assembled and annotated using a pipeline identical to the one we used here. For the remaining 15 animal genomes, we attempted to select 5 lophotrochozoans, 5 ecdysozoans, and 5 deuterostomes. There were only 2 non-mollusk lophotrochozoans available, so we chose 3 lophotrochozoans, 6 ecdysozoa, and 6 deuterostomes, attempting to maximize the phylogenetic diversity of the species we chose. We included 17 outgroups in our analysis: the two species with sequenced genomes that are the closest relatives of choanoflagellates and animals (*Capsaspora owczarzaki*

and *Sphaeroforma arctica*), five representative fungi, the apusozoan *Thecamonas trahens*, the amoebozoan *Dictyostelium discoideum*, and 8 species representing all other major eukaryotic lineages.

Before beginning our analysis, we observed that within each of the species' predicted protein sets from our transcriptome assemblies, there were a number very similar, but not identical proteins. As most of these were nearly identical along their entire length, we hypothesized that the bulk of these cases were assembly artifacts rather than biologically meaningful predicted proteins. Nearly identical predicted proteins can pose problems for construction of orthologous groups, both because they may create spurious species-specific groups and because they slow down subsequent computational steps, which depend on the number of matches found by BLAST. To avoid these problems, we removed nearly identical proteins within each species from our data set using CD-HIT version 4.5.4 (Li and Godzik 2006) with the following parameter values: -t 0 -G 0 -c 0.9 -A 40 -U 50. We experimented with a wide range of parameter values for CD-HIT and found that they had limited effect (no more than 5-10% of eventual protein data set size). Nevertheless, we chose a set of parameters that we felt would identify the type of redundancy we were attempting to remove. The most common type of redundancy we observed was two predicted proteins that were nearly identical over their entire length (as few as 40 amino acids but generally greater) but where a small number of amino acids (generally 25 or less) differed at either the amino or carboxy terminus or both. Therefore, we chose parameters designed to match this observation: a similarity of 90% or greater (-c 0.9), measured within the region of alignment of the two proteins (-G 0), of minimum 40 amino acids in length (-A 40) with a minimum of 50 amino acids (25 for each terminus) excluded from the alignment (-U 50). In addition, we set the tolerance to zero (-t 0) because we were not limited by run time.

Running CD-HIT reduced the size of each predicted protein set by between 13% and 77% (Table 3.9). If the nearly identical sequences that were eliminated actually represented different alternative splice isoforms rather than assembly artifacts, inferences about their orthology should not be strongly affected, because CD-HIT chooses the longest representative within each group of similar sequences.

We used the OrthoMCL algorithm version 2.0.3 (Li, Stoeckert, and Roos 2003) to construct orthologous groups of proteins using all recommended parameter values. To perform an all against all homology search, we used blastp from the BLAST package version 2.2.26 (Altschul et al. 1997) with an expectation value of 1×10^{-5} , as recommended by OrthoMCL. We performed all of the database steps of the OrthoMCL algorithm as suggested. We determined orthologous groups with MCL-edge version 12-068 (Enright, Van Dongen, and Ouzounis 2002; van Dongen and Abreu-Goodger 2012) using an inflation parameter of 1.5, as recommended by OrthoMCL. We tested a range of parameter values for the blastp expectation value and the MCL inflation parameter and found that they did not have a strong effect (less than 5-10%) on our estimate of the gene content of the common ancestor of choanoflagellates and animals (see 'Estimation of Ancestral Gene Content' below).

3.14 Estimation of Ancestral Gene Content

To estimate gene content, gene gain and gene loss at each ancestral node, we applied Dollo parsimony to the orthologous gene clusters we generated. We transformed the orthologous clusters into presence or absence data for each cluster (present for species within the cluster,

absent for species not in the cluster), then used the dollop program from the Phylip package version 3.69 (Felsenstein 1989) to infer ancestral states with the 'Print states at all nodes of tree' option. We randomized the input order of species once using a random seed, and all other parameter values were set to their defaults.

The tree reconstructed by Dollo parsimony closely matched the species tree (data not shown). Differences from the species tree did not affect the ancestral reconstruction of the nodes we were most interested in, the common ancestor of choanoflagellates, the common ancestor of animals and the common ancestor of choanoflagellates and animals, as all three of these groupings were reconstructed correctly.

Assuming that orthologous gene families are constructed perfectly, Dollo parsimony is the appropriate ancestral reconstruction method to use, as all genes within a family are orthologous and are thus related by common ancestry. However, no available orthologous gene family construction method is perfect. In particular, inappropriate orthology assignments between distantly related species will cause the Dollo parsimony method to assign an ancient origin to the gene family with multiple secondary losses. We recognized that this problem might be present in our data set. We examined the distribution of the number of predicted losses for each gene family (data not shown), but we did not observe an obvious separation between genes with few inferred losses (true orthologs) and genes with many losses (false orthologs), although some gene families were inferred to contain many secondary losses. As more projects in the future rely on orthologous groups of genes to construct ancestral gene content, this general issue may need to be addressed.

3.15 Annotation of Predicted Proteins and Gene Ontology Enrichment

We predicted protein domains for all sets of proteins used in this study using the Pfam version 26.0 database and the pfam_scan.pl tool revision 2010-06-08 (Punta et al. 2012), which uses hmmscan from the HMMER 3.0 package (<http://hmmer.janelia.org>). All Pfam searches were performed with default parameter values, including the gathering threshold for expectation values provided for each domain.

We predicted signal peptides and transmembrane domains using Phobius version 1.01 (Käll, Krogh, and Sonnhammer 2004) with default parameter values.

We annotated protein function using the PANTHER Classification System (Thomas et al. 2003). We used the PANTHER HMM library version 7.2 (Mi et al. 2010) and the PANTHER HMM Scoring tool version 1.03 (Thomas et al. 2006) with default parameter values. We accepted PANTHER HMM hits with the recommended strict expectation value cutoff of 10^{-23} . We used PANTHER-provided files to associate PANTHER HMM hits with Gene Ontology terms (Ashburner et al. 2000). For orthologous gene families produced by OrthoMCL, we accepted all annotations for which half or more of the proteins within the family shared the same PANTHER HMM hit.

To measure enrichment of Gene Ontology terms, we used Ontologizer version 2.1 (Robinson et al. 2004; Grossmann et al. 2007) with the Parent-Child-Union method of comparison, a Bonferroni correction for multiple testing, and an adjusted p-value cutoff of 0.05.

3.16 Measurement of Expression Levels

In order to measure the number of reads that went into the construction of each tandem repeat masked contig in our decontaminated data set, we calculated reads per kilobase per million reads (RPKM) by mapping reads with BWA version 0.6.2 (Li and Durbin 2009) and producing counts using SAMtools version 0.1.18 (Li et al. 2009). When running BWA, we required perfect read matches using '-n 0 -k 0'. Because there were up to 200 predicted alternative splice forms for some contigs, we allowed up to 200 equally best hits to be reported by the same program with the '-n 200' option and we extracted all hits using the xa2multi.pl script provided by the BWA package. We counted reads mapping to each contig using the idxstats program within SAMtools and we calculated RPKM as $(10^9 * \text{number of reads mapping in contig}) / (\text{contig length} * \text{total reads mapping for all contigs})$.

3.17 Attempts to Construct Phylogenetic Trees for Orthologous Gene Families

For several orthologous gene families, we attempted to construct phylogenetic trees of the protein sequences within the family and closely related sequences in order to understand the evolutionary relationships within and between families. We attempted to construct trees for three gene families found only in choanoflagellates and sponges: the LRR domain-containing family 10396, the TIG domain-containing family 1066, and the tyrosine kinase family 7542. We also attempted to construct a tree for the gene family containing choanoflagellate Dicer sequences.

For each family, we began by producing a multiple alignment using FSA version 1.15.0 (Bradley et al. 2009) with the '--fast' option and all other parameter values set to their defaults. We also tested alignments using MUSCLE version 3.8.31 (Edgar 2004) with default parameter values, but this produced no difference in final results when compared to FSA. We trimmed alignments to conserved positions using Gblocks 0.91b (Castresana 2000) with Allowed Gap Positions set to 'half' and the most lenient possible values for other parameters. The alignment for gene family 1066 had no conserved positions remaining after Gblocks, and so we did not proceed with this family.

For the remaining three families, we searched for the most closely related outgroup sequences in our data set using the HMMER 3.0 package (<http://hmmer.janelia.org>). We built models for each family using hmmbuild with default parameter values, then searched all of the proteomes used in orthologous gene family construction with hmmsearch, accepting the top hit for each species. We then built a new alignment including closely related sequences for each gene family using FSA and trimmed using Gblocks as above. We constructed phylogenetic trees for each alignment using MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). In each case, we allowed MrBayes to estimate the fixed rate model (aamodelpr=mixed), specified gamma-distributed rate variation among sites with a proportion of invariable sites (rates=invgamma) and ran for one million generations with a burn in of 250,000 generations. All other parameter values were set to their defaults. For all three families, the two chains converged but the resulting trees were almost entirely unresolved, with no more than 2 or 3 species estimated to be more closely related to one another than to other species, even those that were predicted to be within the orthologous gene family (data not shown).

We reasoned that because Dicer is a multi-domain protein, it might be possible to construct an alternative alignment based on concatenated individual alignments of each domain, and this alignment method might provide a better reconstruction of homologous positions than a general aligner such as FSA or MUSCLE. The canonical Dicer protein consists of six domains in the Pfam database: DEAD, Helicase_C, dsRNA_bind, PAZ and two Ribonuclease_3 domains. We

downloaded HMMs for each domain from the Pfam version 26.0 database (Punta et al. 2012) and searched them individually against the Dicer orthologous gene family using hmmsearch. We used hmmalign to align each resulting hit below the Pfam gathering threshold individually for each domain. In the case of Ribonuclease_3, we aligned the first and second hits (by position) within each protein separately. We concatenated all six domain alignments in order and constructed a phylogenetic tree using MrBayes as above. The resulting phylogenetic tree was also unresolved.

3.18 Hyaluronidase Treatment of *Salpingoeca napiformis* Cultures

To test whether hyaluronidase could have an effect on choanoflagellate morphology by altering the structure of the extracellular matrix found in colonies, we treated *Salpingoeca napiformis*, a freshwater colony-forming choanoflagellate, with increasing concentrations of bovine testicular hyaluronidase (type IV-S, Sigma H4272), which is expected to have the same or similar enzymatic activity to predicted choanoflagellate hyaluronidase. We began by splitting actively growing cultures at 1:5 into new medium and allowing the cultures to grow for 24 hours. We added the following concentrations of hyaluronidase to separate cultures: none, 50, 100, 300 and 500 micrograms per milliliter. After 24 hours in the presence of hyaluronidase, we collected images with phase contrast microscopy and counted the number of cells per colony for 10 microliter aliquots of each treatment using a Bright-Line Hemacytometer (Hausser Scientific). All counts were performed blinded to the identity of each sample, and we counted 30 colonies per sample.

Figures

Gene or Class of Genes	Species
<i>Tyrosine kinases</i>	
Tyrosine kinases and receptor tyrosine kinases	<i>M. brevicollis</i>
Src	<i>M. brevicollis</i> , <i>M. ovata</i> , <i>C. owczarzaki</i>
Convergent evolution in kinase signaling properties	<i>M. brevicollis</i>
in <i>Monosiga ovata</i> , <i>Stephanoeca diplocostata</i> , <i>Codosiga gracilis</i>	<i>M. ovata</i>
PTB domain substrate targeting	<i>M. brevicollis</i>
SH2 domain containing proteins	<i>M. brevicollis</i>
Tec family	<i>M. brevicollis</i>
in <i>Capsaspora owczarzaki</i> , <i>Ministeria vibrans</i>	<i>C. owczarzaki</i> , <i>M. vibrans</i>
<i>Cell adhesion / cell signaling</i>	
Cadherins	<i>M. brevicollis</i> , <i>S. rosetta</i>
Laminin	<i>M. brevicollis</i>
Collagen	<i>M. brevicollis</i>
C-type lectin	<i>M. brevicollis</i>
Integrins	<i>C. owczarzaki</i> , <i>M. brevicollis</i>
Hedgehog, hedgling	<i>M. ovata</i> , <i>M. brevicollis</i>
membrane-associated guanylate kinases (MAGUK)	<i>C. owczarzaki</i>
<i>Ciliary genes</i>	
RFX transcription factors	<i>M. brevicollis</i>
<i>Gene regulation</i>	
Myc	<i>M. brevicollis</i>
T-box, Runt	<i>C. owczarzaki</i>
p53/63/73	<i>M. brevicollis</i>
<i>Growth</i>	
Hippo	<i>C. owczarzaki</i>
Pak	<i>M. ovata</i>
Teneurins	<i>M. brevicollis</i>
Nme genes	<i>M. brevicollis</i>
<i>Immunity</i>	
Dispanins	<i>M. brevicollis</i>
NF-kB	<i>C. owczarzaki</i>
Immunoglobulin	<i>M. brevicollis</i>
<i>Apoptosis</i>	
mitochondrial carrier homolog 2	<i>M. brevicollis</i>
<i>Genes involved in neuronal function in animals</i>	
Secretion	<i>M. brevicollis</i>
Calcium channels	<i>M. brevicollis</i>
Sodium channels	<i>M. brevicollis</i>
Post-synaptic density	<i>M. brevicollis</i>
<i>Ribosomal biogenesis</i>	
Rbm19/Mrd1 RNA binding domains	<i>M. brevicollis</i>
<i>Gene fusions</i>	
G6PD	<i>M. brevicollis</i>
RJL family of small GTPases in flagellar function	<i>M. brevicollis</i>
<i>Other cases</i>	
Androglobin	<i>M. brevicollis</i>
Coronins: actin-dependent, roles in cytokinesis, motility, phagocytosis	<i>M. brevicollis</i>
Arginine kinase isoforms	<i>M. brevicollis</i>
Ubiquitin ligases	<i>M. brevicollis</i>
Cytoskeletal proteins: spectrin, ankyrin, 4.1 and adducin	<i>M. brevicollis</i>
Oligoadenylate synthetase	<i>M. brevicollis</i>
NADPH oxidases	<i>M. brevicollis</i>

Table 1.1 Genes shared among animals and their closest relatives.

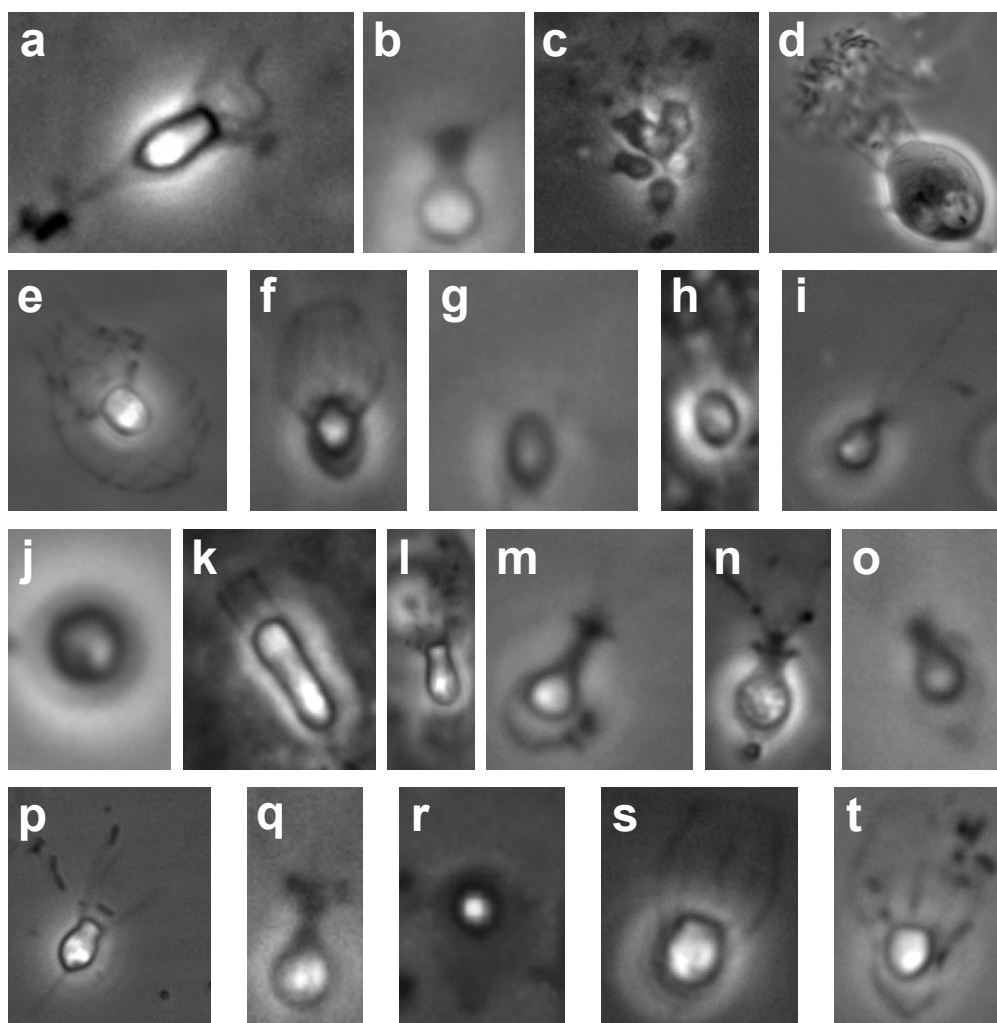


Figure 2.1. Phase contrast light microscopy images of the choanoflagellate cultures sequenced in this study. Species are ordered alphabetically. (a) *Acanthoecca spectabilis* (Virginia) (b) *Choanoeca perplexa* (c) *Codosiga gracilis* (d) *Codosiga hollandis* (e) *Diaphanoeca grandis* (f) *Didymoeca costata* (g) *Helgoeca nana* (h) *Monosiga gracilis* (i) *Monosiga ovata* (j) *Salpingoecca amphoridium* (k) *Salpingoecca gracilis* (l) *Salpingoecca infusionum* (m) *Salpingoecca minuta* (n) *Salpingoecca napiformis* (o) *Salpingoecca pyxidium* (p) *Salpingoecca roanokei* (q) *Salpingoecca urceolata* (r) *Savillea micropora* (s) *Stephanoeca diplocostata* (Australia) (t) *Stephanoeca diplocostata* (France). Images are not to scale.

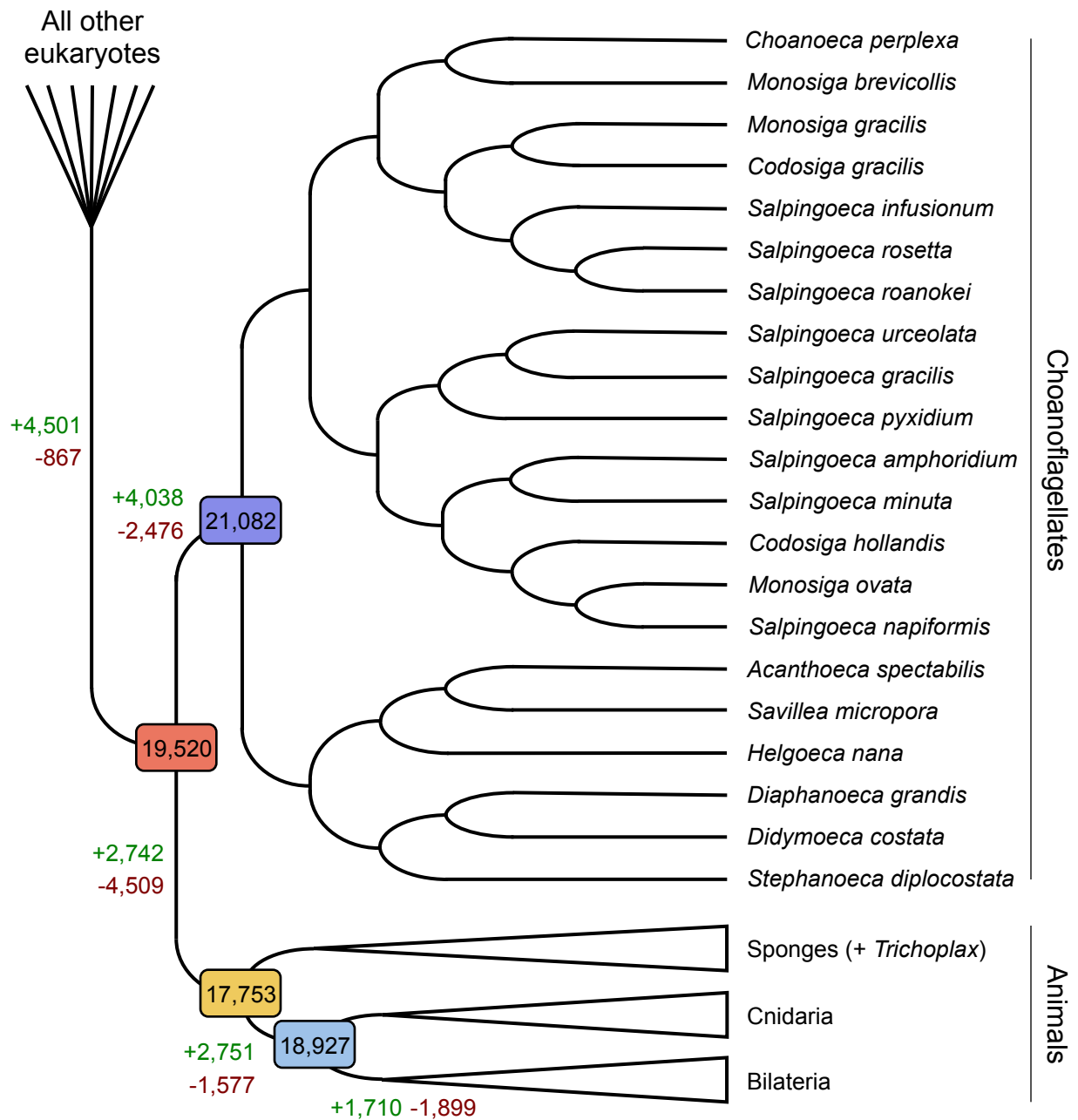


Figure 2.2. Gene gain and loss superimposed on a phylogenetic tree of choanoflagellates and animals. Branch lengths are not to scale, although a similar amount of sequence evolution appears to have occurred within the animal and choanoflagellate lineages (Carr et al. 2008). All choanoflagellates represented, with the exception of *Monosiga brevicollis* and *Salpingoeca rosetta*, were sequenced as part of this study. Tree is adapted from (Nitsche et al. 2008) and (Carr, Richter et al. in preparation). Estimated ancestral gene content is shown at each node of interest, gene gains on a lineage are marked in green, and losses are marked in red. Gains and losses are not shown for non-bilaterian animals, as there are 3 or fewer species in each group. Substantial gene gain and loss has occurred within each lineage, with bursts of innovation accompanying each major transition in animal evolution.

Class	Gene	Diagnostic Domains	Number Choanos with Domains	Number Choanos in OrthoMCL Group	Number Animals in OrthoMCL Group
<i>Cadherin</i>	Flamingo	HRM, GPS, 7tm_2	5	7	18
	Protocadherin	Protocadherin	6	1	11
<i>Integrin</i>	Integrin beta	signalp, Integrin_beta, tm	1	0	-
	Integrin beta binding protein	ICAP-1_inte_bdg	19	19	8
<i>Notch</i>	Notch	EGF, Notch, tm, Ank	21	9	21
	Delta/Serrate/Jagged	signalp, MNNL, DSL, tm	6	7	1
	CSL / Su(H)	LAG1-DNAbind, BTB, TIG	21	21	19
<i>Immune response</i>	Toll-like receptor	signalp, LRR, tm, TIR	5	0	-
	NF-kB	RHD	8	0	-
<i>Hyaluronidase</i>	Hyaluronidase	Glyco_hydro_56	13	21	21
<i>BRCA2</i>	BRCA2	BRCA-2_helical, BRCA-2_OB1, Tower, BRCA-2_OB3	21	18	17
<i>TGFb</i>	TGF beta receptor	TGF_beta_GS	1	0	-
	TGF beta	signalp, TGF_beta_propeptide	4	7	1

Table 2.1. Key animal genes newly discovered to be shared with choanoflagellates. Diagnostic domains are either Pfam domains or signal peptide/transmembrane domains predicted by Phobius (signalp, tm). Domains listed are diagnostic for each gene in animals and are shared with choanoflagellates (not all diagnostic domains are shown). The number of choanoflagellate species with predicted proteins containing one or more of the diagnostic domains is shown. The final two columns are based on OrthoMCL computation of gene families. Some choanoflagellate proteins contain appropriate diagnostic domains but are not in the same family as the corresponding animal protein (as may occur for multi-domain proteins). Conversely, some choanoflagellate proteins may not pass the significance threshold for Pfam domains, but are still in the same family as determined by OrthoMCL.

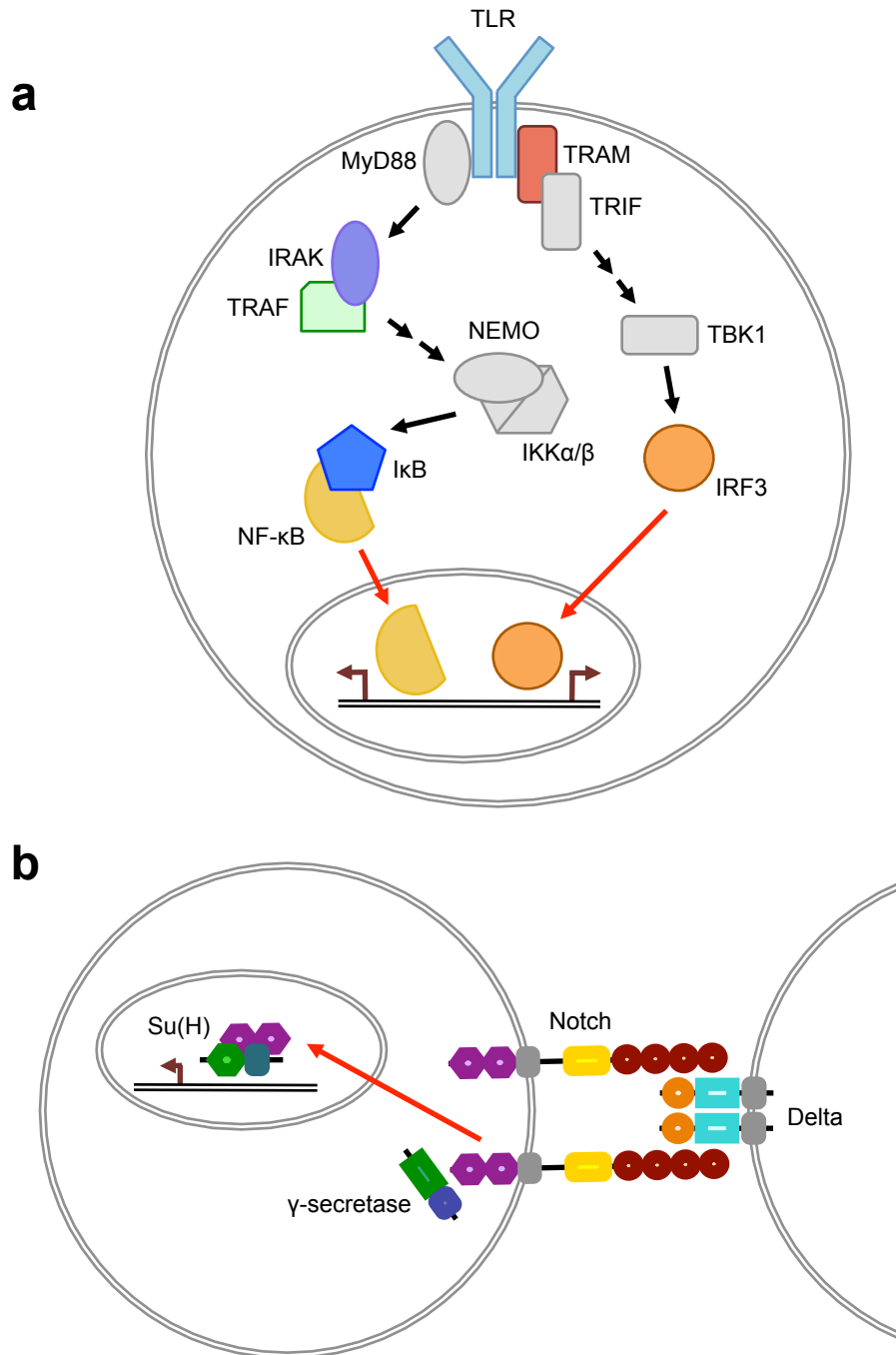


Figure 2.3. Components of two key animal pathways are shared by choanoflagellates: immune response and Notch signaling. (a) A simplified version of the Toll-like receptor pathway in animals, adapted from (Kawai and Akira 2007; Gilmore 2012). Colored components are found in choanoflagellates, grey components are not. Most of the key proteins are found in choanoflagellates, but the adapters MyD88, TRIF and TBK1 and the pathway members NEMO and IKKα/β are not. (b) The critical components of the Notch signaling pathway are present in choanoflagellates, including Notch, Delta and Su(H). γ-secretase, which cleaves the intracellular domain of Notch to allow it to translocate to the nucleus, is also found in choanoflagellates.

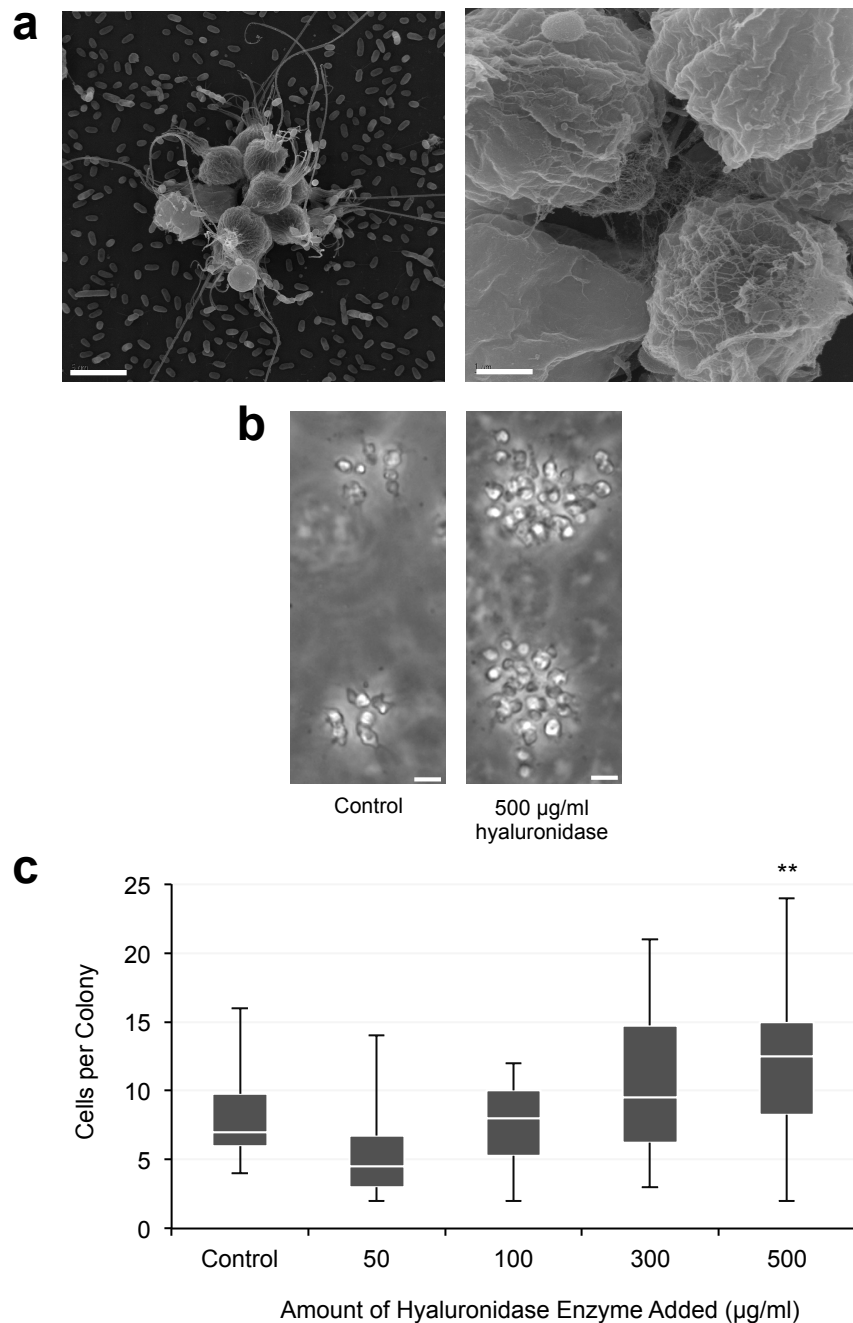


Figure 2.4. Hyaluronidase, an enzyme that degrades extracellular matrix in animals, increases the number of cells per colony in *Salpingoeca napiformis*. (a) SEM of *S. napiformis* colonies, showing extracellular matrix between cells within the colony, scale = 5 µm (left), 1 µm (right). (b) Localization of fluorescein-labeled hyaluronic acid in *S. napiformis*. (b) Representative phase-contrast light microscopy of *S. napiformis* culture 24 hours after treatment with vehicle control (water) or 500 µg/ml hyaluronidase enzyme, scale bars 10 µm. (c) Addition of hyaluronidase causes an increase in the number of cells per colony. Boxes are 25th/75th percentile, bars are min and max values, and white bar is median. ** different from control, $p = 0.01$, two-sample Kolmogorov-Smirnov test.

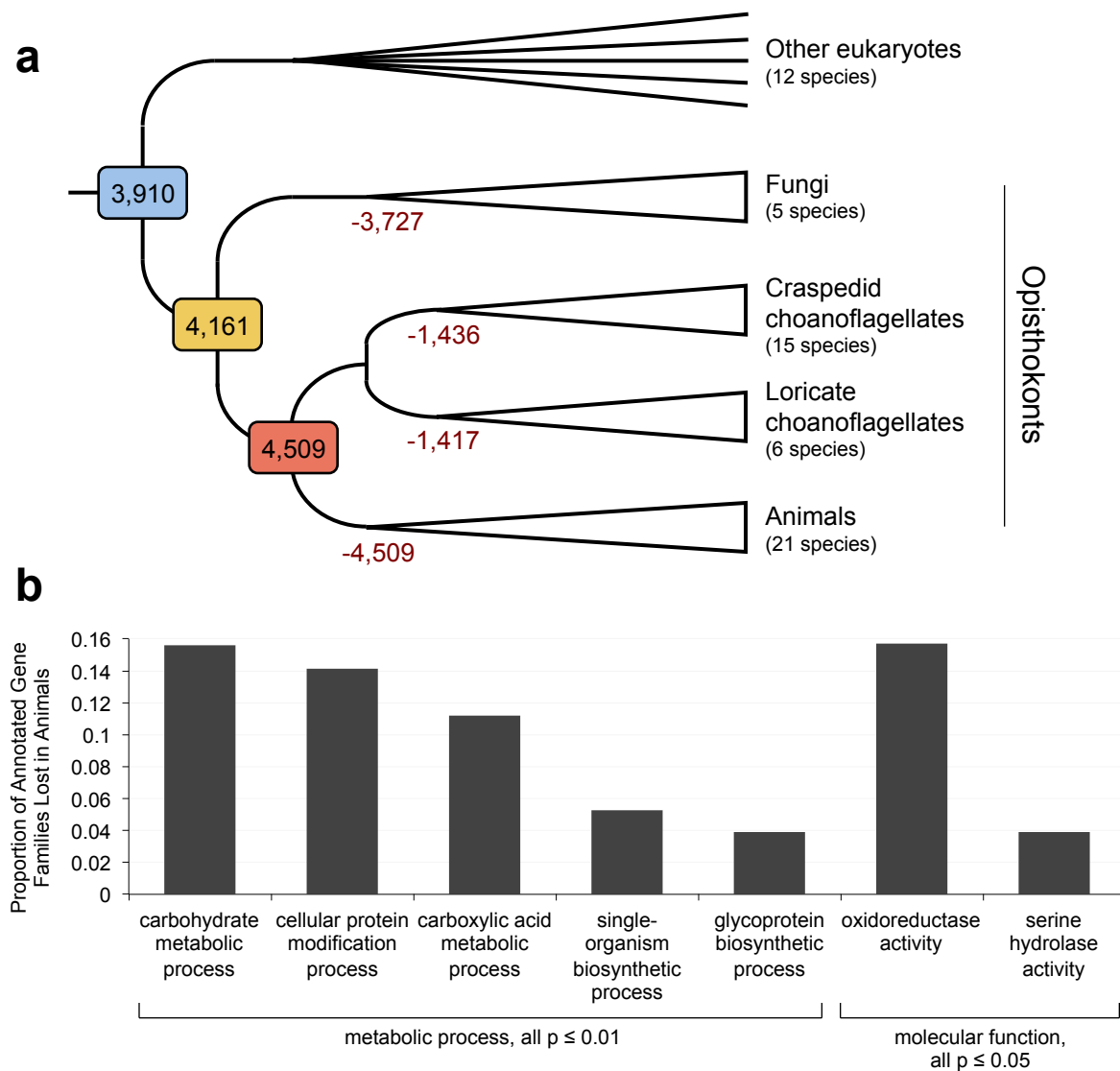


Figure 2.5. Genes lost in animals are significantly overrepresented for metabolic functions. (a) Phylogenetic distribution of gene families lost in animals. Branch lengths are not to scale. The majority of families lost in animals are inferred to have been present in the ancestral eukaryote, but many of these families have also been lost in other Opisthokont lineages. (b) Gene Ontology metabolic processes or molecular functions significantly overrepresented in the 4,509 gene families lost in animals, when compared with the 19,520 gene families present in their common ancestor with choanoflagellates. Several metabolic processes, including glycoprotein biosynthesis and carbohydrate metabolism, are significantly overrepresented.

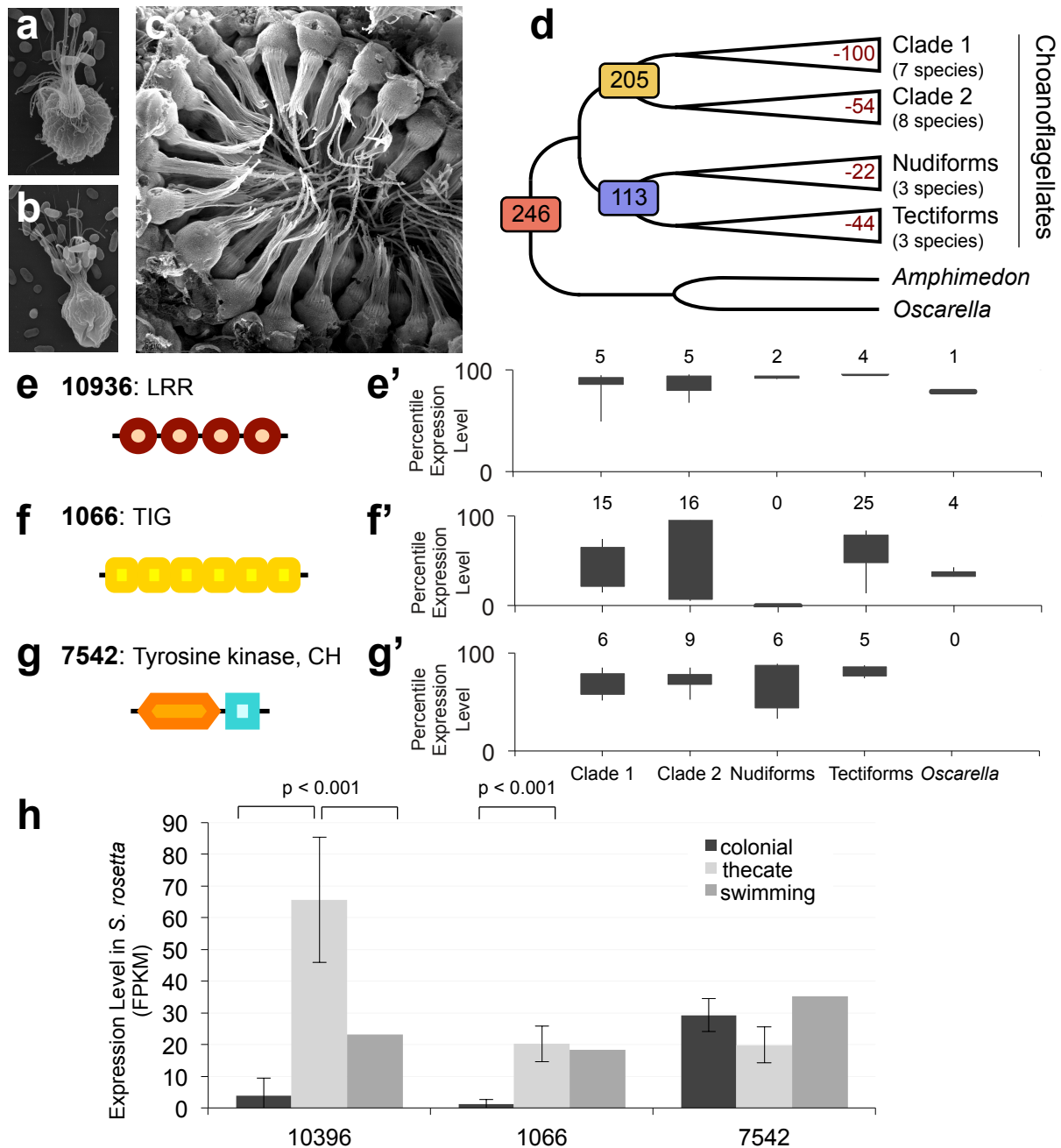


Figure 2.6. Gene families found exclusively in choanoflagellates and sponges. (a) - (c) SEM images. (a), (b) Top and side view of *Salpingoeca napiformis*. (c) Choanocyte chamber of the sponge *Oscarella*, whose cells share ultrastructural affinity with choanoflagellates. (d) Tree showing the number of gene families shared exclusively between choanoflagellates and sponges. 246 families are present in *Amphimedon* and at least one choanoflagellate, but many of those are lost within choanoflagellates. (e) - (g) Representative domain structure and (e') - (g') expression level (as a percentile within each species, box 25th/75th percentiles, lines 10th/90th percentiles) for three shared gene families. Size of gene family within each clade is above each plot. (e) Leucine-rich repeat containing protein. (f) TIG repeat (an immunoglobulin-like fold) containing protein. (g) A tyrosine kinase whose domain architecture is unique to animals and choanoflagellates. (h) Expression level in different cell types of *Salpingoeca rosetta* for the three gene families. Error bars are +/- s.d. if more than one replicate was measured. 10396 and 1066 are significantly upregulated in thecate cells, an attached cell type.

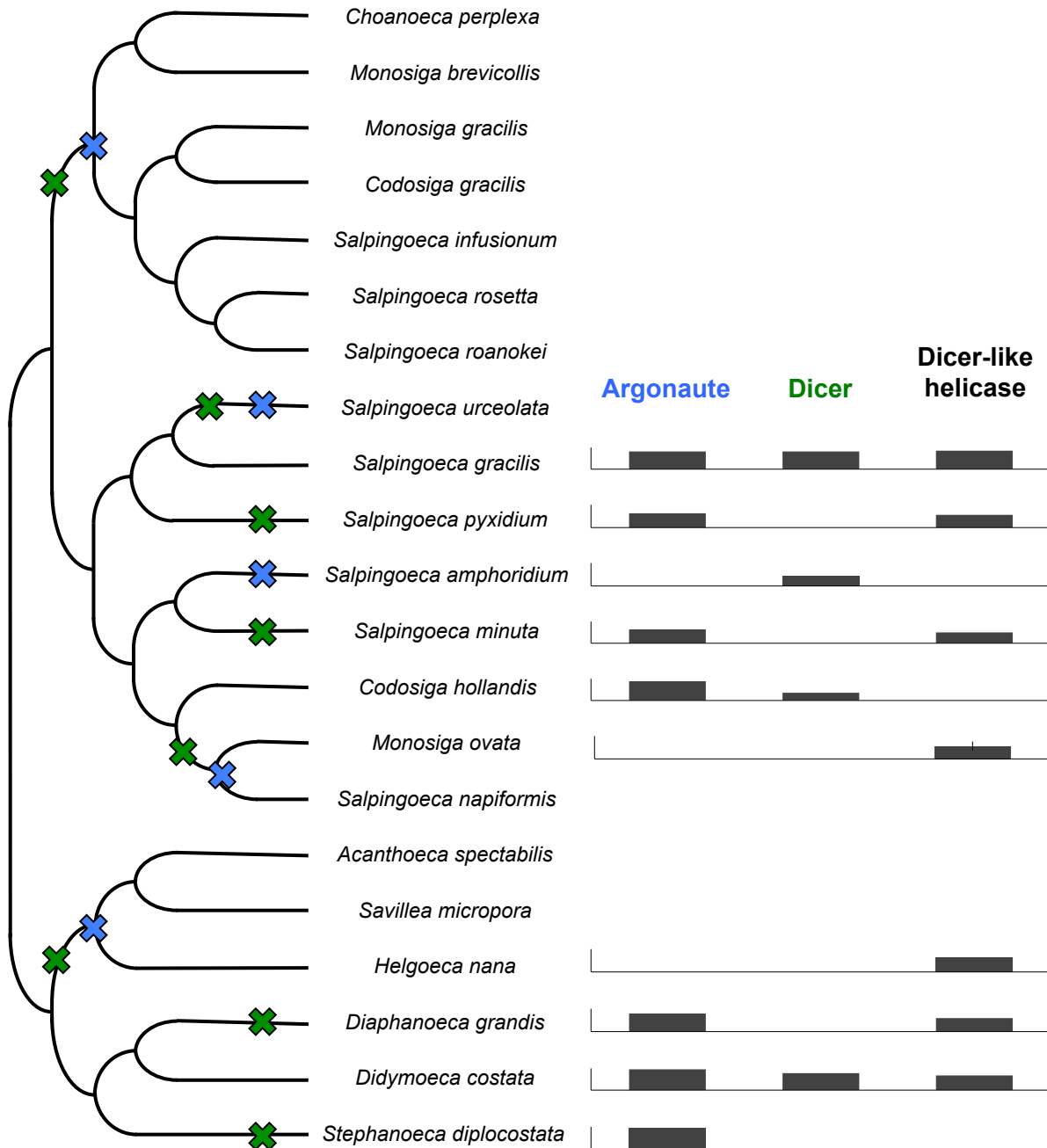


Figure 2.7. RNAi components are present in choanoflagellates, but have been lost multiple times in different lineages. A phylogenetic tree of choanoflagellates (adapted from (Nitsche et al. 2011) and (Carr, Richter et al. in preparation)) with inferred losses of Argonaute (blue) and Dicer (green). For each species with at least one of Argonaute, Dicer, or Dicer-like helicase (a gene implicated in antiviral RNAi in animals), the expression level of that gene is shown as a percentile rank within each species (error bars s.d. for *M. ovata*, which has two Dicer-like helicases). Species lacking a gene have no expression value. Overall, there have been multiple losses of both Argonaute and Dicer, and RNAi machinery appears to be entirely absent in the clade containing the two previously sequenced species, *M. brevicollis* and *S. rosetta*. In the species with Argonaute, it is highly expressed, and Dicer always shows a lower relative expression level.

Species	Str	Source	ATCC Number	Description		Freshwater /marine	Original Water Sample Source, if known (in the US unless specified)	Year
<i>Monosiga cf. ovata</i>		Barry Leadbeater	50635	naked	craspedid	freshwater	freshwater pond, Yaroslavl, Russia	1979
<i>Codosiga cf. gracilis</i>		ATCC	50454	stalked	craspedid	marine	tide pool on the salt marsh at Freiston Shore, Lincolnshire, England	1972
<i>Monosiga cf. gracilis</i>		ATCC	50964	stalked	craspedid	marine	salt marsh sediment, Cattleshed Marsh, VA	2000
<i>Codosiga cf. hollandis</i>		this study	dep.	stalked	craspedid	freshwater	waterfall at Levada do Risco, 1000m above sea level, Madeira, Portugal	2008
<i>Salpingoeca cf. amphoridium</i>		ATCC	50788	thecate	craspedid	freshwater		
<i>Salpingoeca urceolata</i>		Barry Leadbeater	50560	thecate	craspedid	marine	contaminant in Porphyra sp. culture, Wellington, New Zealand	1987
<i>Salpingoeca cf. minuta</i>		Barry Leadbeater	50938	thecate	craspedid	marine	wet salt marsh sediment, Hog Island, VA	2001
<i>Salpingoeca cf. pyxidium</i>		Barry Leadbeater	50929	thecate	craspedid	marine	salt marsh sediment, Hog Island, VA	1999
<i>Choanoeca perplexa</i>		ATCC	50453	thecate	craspedid	marine	Atlantic Ocean off Plymouth, England	1977
<i>Salpingoeca infusionum</i>		ATCC	50559	thecate	craspedid	marine		
<i>Salpingoeca cf. roanokei</i>		Barry Leadbeater	50931	thecate	craspedid	marine	dry salt marsh sediment, Hog Island, VA	2000
<i>Salpingoeca cf. gracilis</i>		ATCC	50959	thecate	craspedid	marine	surface of an old microbial mat of a flooded salt pan, Great Marsh, DE	
<i>Salpingoeca cf. napiformis</i>		ATCC	50153	thecate	craspedid	freshwater	Biomphalaria glabrata aquarium culture, Rockville, MD	1985
<i>Acanthoeca spectabilis</i>	VA	this study	dep.	nudiform	loricate	marine	shallow lagoon, Hog Island, VA	2007
<i>Helgoeca nana</i>		Barry Leadbeater	50073	nudiform	loricate	marine	Lower Narragansett Bay, RI	1981
<i>Savillea cf. micropora</i>		Barry Leadbeater	dep.	nudiform	loricate	marine	Four-Mile Bridge, Pontrhydybont, Anglesey, Wales	2000
<i>Diaphanoeca grandis</i>		Barry Leadbeater	50111	tectiform	loricate	marine	Narragansett Bay, RI	1982
<i>Stephanoeca diplocostata</i>	AU	Barry Leadbeater	dep.	tectiform	loricate	marine	Swan Bay, Queenscliff, Victoria, Australia	1996
<i>Stephanoeca diplocostata</i>	FR	Barry Leadbeater	50456	tectiform	loricate	marine	rockpool at Le Cabellou, South Brittany, France	1976
<i>Didymoeca costata</i>		Barry Leadbeater	dep.	tectiform	loricate	marine	Swan Bay, Queenscliff, Victoria, Australia	1996

dep. indicates species that have been deposited to ATCC but do not yet have ATCC Numbers

Table 3.1. Origin of choanoflagellate cultures sequenced in this study

Antibiotic	Abbreviation	Vehicle	Final Concentration (mg/ml)
ampicillin	A	water	100
chloramphenicol	C	100% ethanol	170
erythromycin	E	100% ethanol	50
gentamicin	G	water	50
kanamycin	K	water	50
ofloxacin	O	0.1M sodium hydroxide	10
polymyxin B	P	water	50
rifampicin	R	methanol	25
streptomycin	S	water	50
tetracycline	T	70% ethanol	50

Table 3.2. Antibiotic treatments tested for each culture

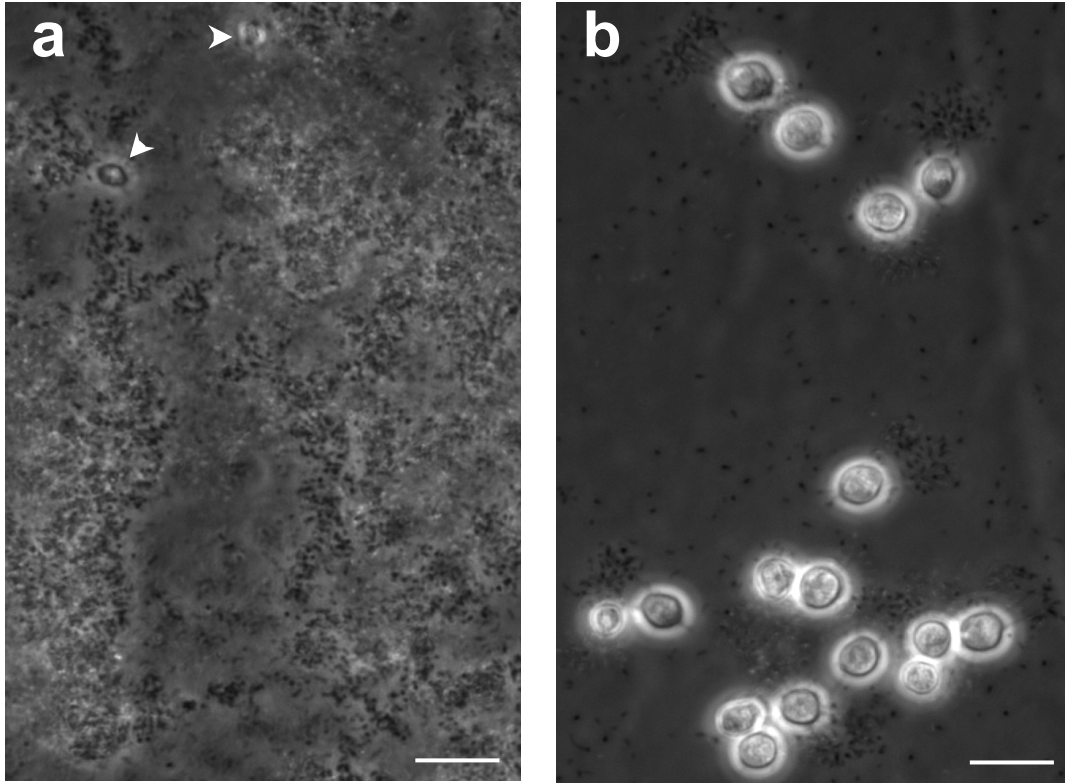


Figure 3.1. Successful antibiotic treatment of *Codosiga hollandis*. (a) a culture where the density of bacteria (here found in a biofilm) have overwhelmed the choanoflagellates (white arrowheads), (b) a successful antibiotic treatment yields a culture with more choanoflagellates and fewer bacteria. The choanoflagellates are also larger in the treated culture. Both images are phase contrast microscopy, scale bar 10 micrometers.

Medium	Base	Nutrient(s) Added	Phosphate (μmol)	Silicate (μmol)	Nitrite (μmol)	Nitrite + Nitrate (μmol)	Ammonia (μmol)
ASW + quinoa	artificial sea water	quinoa grains	4.87	-	-	4.32	11.53
OE	artificial sea water	0.21 mg/L proteose peptone, 0.04 mg/L yeast extract	2.05	-	-	3.12	6.52
10% CGM.3	artificial sea water	0.5 g/L cereal grass	11.73	3.61	0.15	5.30	9.02
25% CGM.3	artificial sea water	1.25 g/L cereal grass	28.28	10.00	0.27	11.22	16.35
10% FCGM.3	filtered water	0.5 g/L cereal grass	n.d.	n.d.	n.d.	n.d.	n.d.
25% FCGM.3	filtered water	1.25 g/L cereal grass	42.29	5.09	0.23	11.29	16.39

Values marked with '-' were below the lower limit of detection (1 mmol for silicate, 0.1 mmol for nitrite), n.d. not determined

Table 3.3. Growth media and nutrient analysis

Species	Medium	Temp (C)	Antibiotic Treatment (in order)	Approximate Percentage of Cell Types					
				Attached, single cell	Slow swim	Fast swim	Colonial, swim	Colonial, attached	Passively suspended
<i>Acanthoecca spectabilis</i> (VA)	10% CGM.3	10	S, R	90	0	10	0	0	0
<i>Choanoeca perplexa</i>	OE	15	S	95	5	0	0	0	0
<i>Codosiga gracilis</i>	OE	15	S, K	50	0	0	0	50	0
<i>Codosiga hollandis</i>	10% FCGM.3	25	O	95	0	0	0	5	0
<i>Diaphanoeca grandis</i>	ASW + quinoa	10	S	0	0	0	0	0	100
<i>Didymoeca costata</i>	OE	15	G, K, S	0	0	0	0	0	100
<i>Helgoeca nana</i>	ASW + quinoa	4	K	90	0	10	0	0	0
<i>Monosiga gracilis</i>	10% CGM.3	25	S	80	0	0	0	20	0
<i>Monosiga ovata</i>	25% FCGM.3	25	O	15	10	0	0	0	75
<i>Salpingoeca amphoridium</i>	10% FCGM.3 + HEPES	15	S, O	100	0	0	0	0	0
<i>Salpingoeca gracilis</i>	25% CGM.3	25	none	100	0	0	0	0	0
<i>Salpingoeca infusionum</i>	10% CGM.3	25	K, O	80	20	0	0	0	0
<i>Salpingoeca minuta</i>	OE	15	none	95	5	0	0	0	0
<i>Salpingoeca napiformis</i>	25% FCGM.3	25	C, A, E	5	15	20	60	0	0
<i>Salpingoeca pyxidium</i>	OE	15	none	95	5	0	0	0	0
<i>Salpingoeca roanokei</i>	OE	15	none	40	0	0	0	60	0
<i>Salpingoeca urceolata</i>	ASW + quinoa	15	K	95	0	0	0	0	5
<i>Savillea micropora</i>	ASW + quinoa	15	S	100	0	0	0	0	0
<i>Stephanoeca diplocostata</i> (AU)	OE	15	none	0	0	0	0	0	100
<i>Stephanoeca diplocostata</i> (FR)	OE	15	none	0	0	0	0	0	100

Table 3.4. Growth media, successful antibiotic treatments and cell types for each culture

Species	Growth Vessel	Growth Volume Per Vessel (ml)	Number of Vessels	Filtration During Collection?	Growth Duration	Ratio of Culture:Fresh Medium to Seed Growth	Centrifuge Time (min)
<i>Acanthoecca spectabilis</i> (VA)	150 mm dish	50	5	no	3 weeks	1:10	20
<i>Choanoeca perplexa</i>	150 mm dish	50	5	no	1 week	1:25	20
<i>Codosiga gracilis</i>	150 mm dish	50	5	no	3 weeks	1:25	20
<i>Codosiga hollandis</i>	150 mm dish	50	10	yes	2.5 weeks	1:25	20
<i>Diaphanoeca grandis</i>	4 L glass flask	1000	1	yes	6 weeks	1:10	60
<i>Didymoeca costata</i>	150 mm dish	50	5	yes	5 weeks	1:17	40
<i>Helgoeca nana</i>	150 mm dish	50	10	yes	2 weeks	1:10	20
<i>Monosiga gracilis</i>	150 mm dish	50	1	no	3 days	1:25	20
<i>Monosiga ovata</i>	150 mm dish	50	5	no	5 days	1:10	20
<i>Salpingoeca amphoridium</i>	150 mm dish	50	5	no	4 weeks	1:17	20
<i>Salpingoeca gracilis</i>	75 cm ² flask	60	5	yes	7 weeks	1:20	20
<i>Salpingoeca infusionum</i>	150 mm dish	50	1	no	5 days	1:25	20
<i>Salpingoeca minuta</i>	150 mm dish	50	5	no	4 weeks	1:25	20
<i>Salpingoeca napiformis</i>	1 L glass flask	200	1	no	4 days	1:10	20
<i>Salpingoeca pyxidium</i>	150 mm dish	50	5	no	1 week	1:25	20
<i>Salpingoeca roanokei</i>	150 mm dish	50	5	no	3 weeks	1:25	20
<i>Salpingoeca urceolata</i>	150 mm dish	50	5	yes	5 weeks	1:25	20
<i>Savillea micropora</i>	150 mm dish	50	10	yes	3 weeks	1:25	20
<i>Stephanoeca diplocostata</i> (AU)	2 L glass flask	500	2	yes	4 weeks	1:5	40
<i>Stephanoeca diplocostata</i> (FR)	2 L glass flask	500	2	yes	4 weeks	1:5	40

Table 3.5. Culture growth in large batches for harvesting

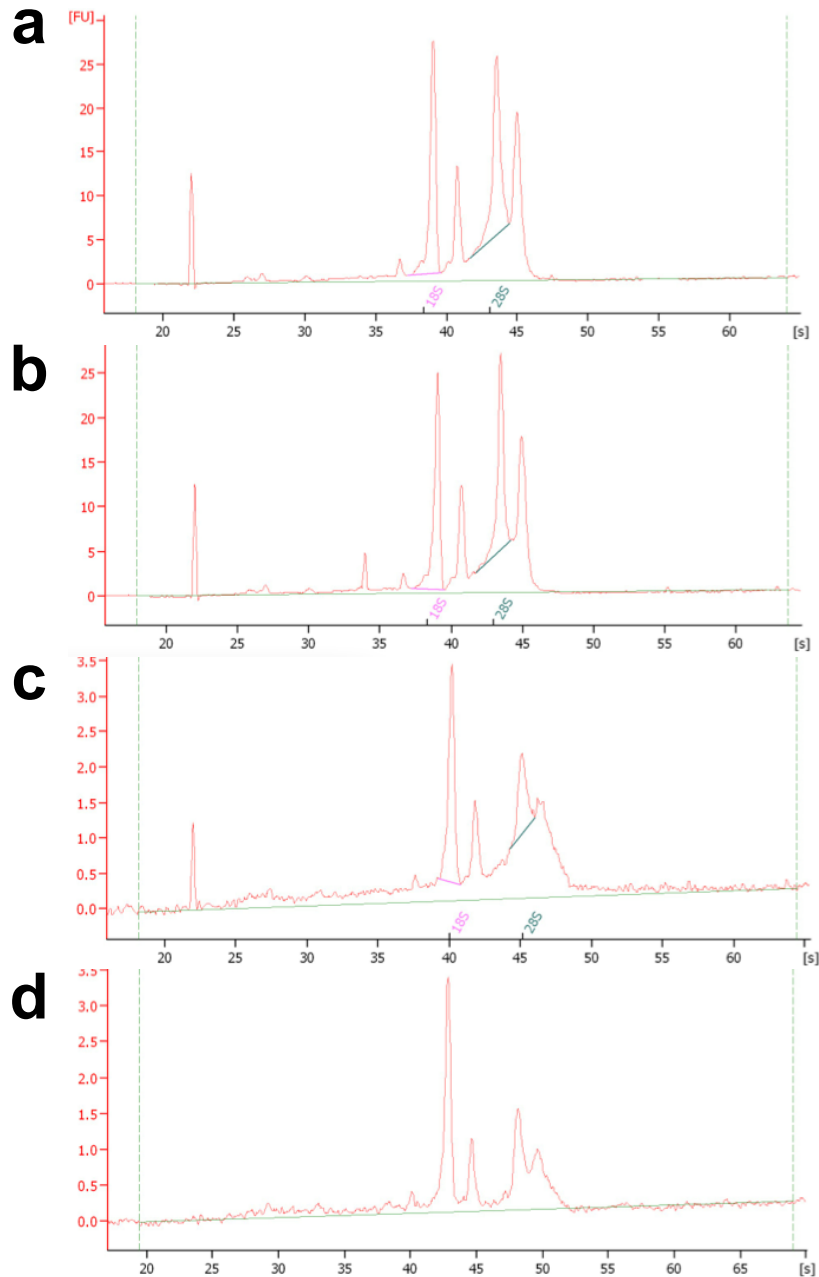


Figure 3.2. DNase buffer degrades *Codosiga hollandis* total RNA. (a) *C. hollandis* total RNA after isolation, displaying four sharp ribosomal peaks (16S and 23S for bacteria, 18S and 28S for *C. hollandis*) (b) *C. hollandis* total RNA after heating for 30 minutes at 37 degrees Celsius shows little evidence of degradation (c) *C. hollandis* total RNA is degraded after addition of buffer (but not DNase) and heating for 30 minutes at 37 degrees Celsius. The fluorescence axis is 10 times lower than in the non-degraded samples (d) *C. hollandis* total RNA after addition of both DNase buffer and DNase and heating for 30 minutes at 37 degrees Celsius. The bulk of the degradation appears to occur in the presence of the buffer, and addition of the DNase enzyme does not have a marked effect on RNA quality.

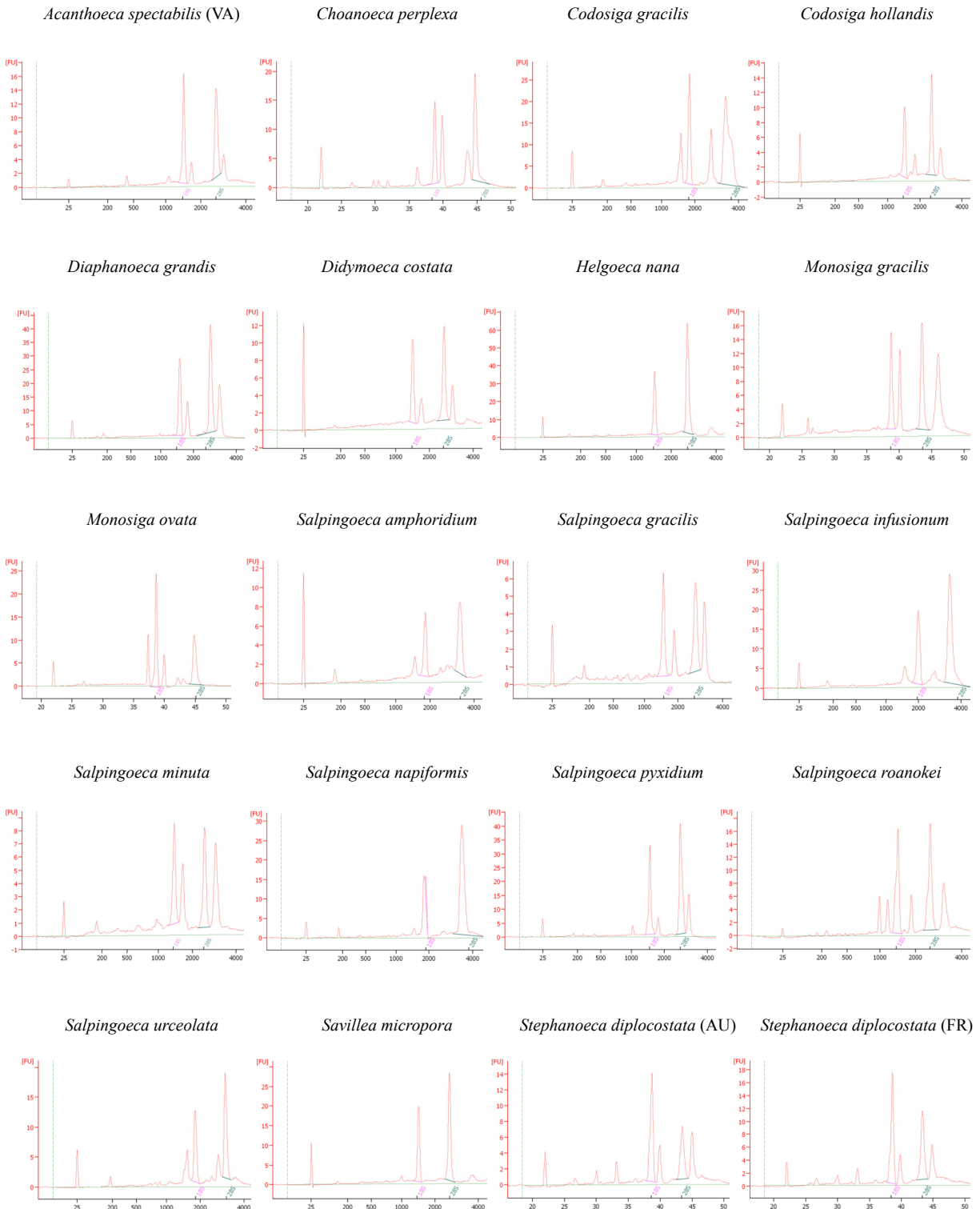


Figure 3.3. Bioanalyzer Pico RNA traces of total RNA used in transcriptome sequencing. Source culture for total RNA isolation is identified above each Bioanalyzer trace.

Species	DNA Source	PCR Cleanup	Sequencing Template	Estimated Locus Size (bp)	Primers Used
<i>Acanthoea spectabilis</i> (VA)	phenol:chloroform	gel extraction	direct	1694	18S_1F, 18S_1528R
<i>Choanoeca perplexa</i>	RNAqueous	gel extraction	direct	1699	18S_1F, 18S_564F, 18S_1205R, 18S_1528R
<i>Codosiga gracilis</i>	phenol:chloroform	gel extraction	TOPO	2567	M13F, 18S_564F, 18S_1205R, M13R
<i>Codosiga hollandis</i>					
<i>Diaphanoeca grandis</i>	RNAqueous	gel extraction	direct	1713	18S_1F, 18S_1528R
<i>Didymoeca costata</i>					
<i>Helgoeca nana</i>	RNAqueous	gel extraction	direct	1695	18S_1F, 18S_1528R
<i>Monosiga gracilis</i>	RNAqueous	gel extraction	TOPO	1815	M13F, 18S_564F, 18S_1205R, M13R
<i>Monosiga ovata</i>	phenol:chloroform	PCR cleanup	direct	1689	18S_1F, 18S_1528R
<i>Salpingoeca amphoridium</i>	phenol:chloroform	PCR cleanup	TOPO	2266	M13F, 18S_564F, 18S_1205R, M13R
<i>Salpingoeca gracilis</i>	RNAqueous	gel extraction	TOPO	2571	M13F, 18S_564F, Sg_18S_1, Sg_18S_2, Sg_18S_3, M13R
<i>Salpingoeca infusionum</i>	RNAqueous	gel extraction	TOPO	1981	M13F, 18S_564F, 18S_1205R, M13R
<i>Salpingoeca minuta</i>					
<i>Salpingoeca napiformis</i>	phenol:chloroform	PCR cleanup	TOPO	3114	18S_1F, 18S_528F, 18S_1220F, 18S_1871R, 18S_1200R, 18S_1528R
<i>Salpingoeca pyxidium</i>	phenol:chloroform	gel extraction	direct	1726	18S_1F, 18S_1528R
<i>Salpingoeca roanokei</i>	RNAqueous	gel extraction	TOPO	1946	M13F, 18S_564F, 18S_1205R, M13R
<i>Salpingoeca urceolata</i>	phenol:chloroform	gel extraction	direct	1708	18S_1F, 18S_1528R
<i>Savillea micropora</i>					
<i>Stephanoeca diplocostata</i> (AU)	phenol:chloroform	gel extraction	direct	1713	18S_1F, 18S_564F, 18S_1205R, 18S_1528R
<i>Stephanoeca diplocostata</i> (FR)	phenol:chloroform	gel extraction	direct	1715	18S_1F, 18S_564F, 18S_1205R, 18S_1528R

Table 3.6. Genomic DNA sequencing of the small subunit ribosomal locus

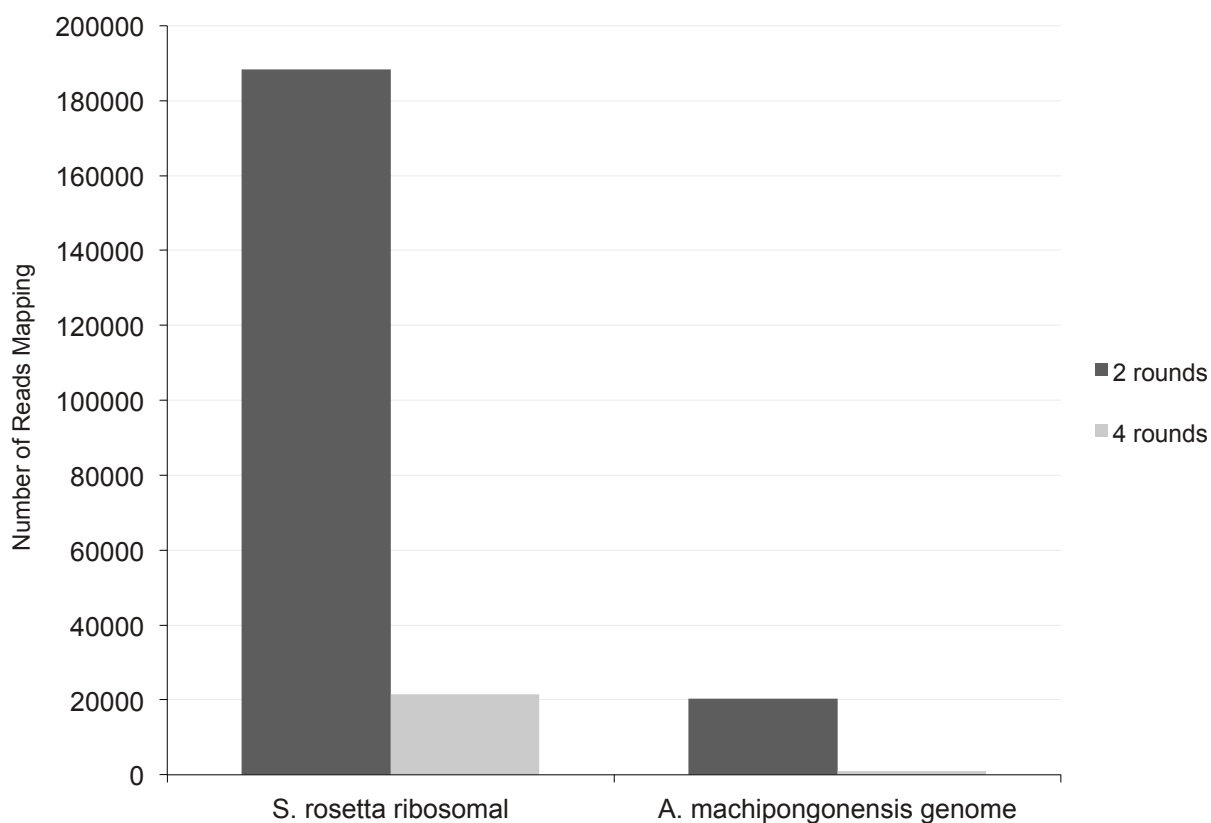


Figure 3.4: An additional two rounds of polyA selection remove an order of magnitude more non-polyadenylated RNA. Read counts for the two data sets mapped against the *S. rosetta* ribosomal locus and against the *A. machipongonensis* genome. Both sets of counts are an order of magnitude lower with four rounds rather than two rounds of polyA selection.

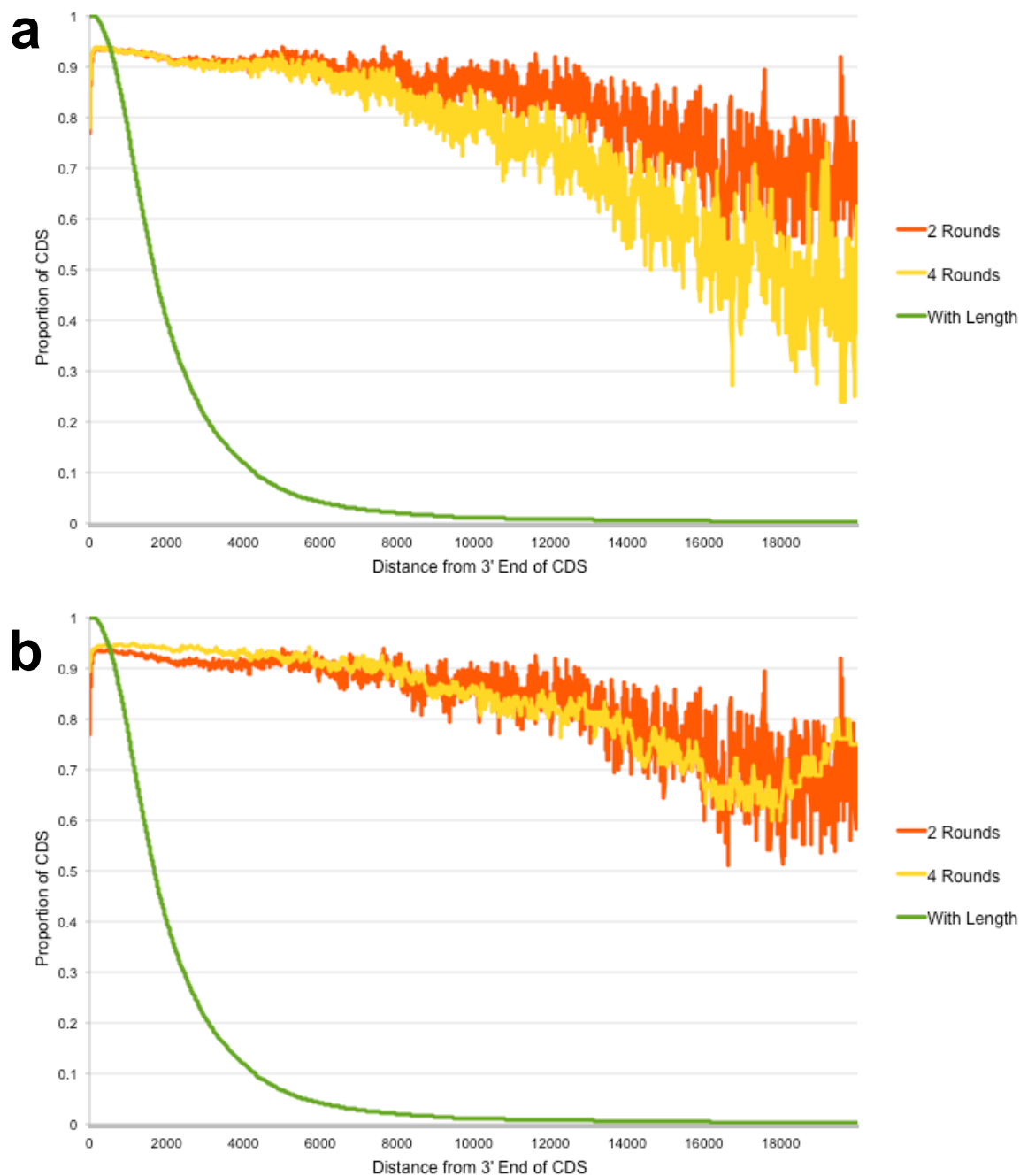


Figure 3.5: An additional two rounds of polyA selection do not cause significant loss of read coverage at the 5' ends of transcripts. (a) The proportion of transcripts with at least one read mapping at each given distance from their 3' end for two rounds and four rounds of polyA selection. Also shown is the proportion of total transcripts in the *S. rosetta* genome at each length. There is a difference between the two round and four round coverage beginning at roughly 10,000 bases from the 3' end of transcripts, but only a small proportion of transcripts in the *S. rosetta* genome is 10,000 bases or longer. (b) The same plot for *de novo* assemblies of both data sets using Trinity.

Species	Library Prep Group	Seq Group	TruSeq Adapter	Total RNA Used (µg)	Est. Proportion of Choano Total RNA	Est. Amount of Choano Total RNA Used (µg)	Bioanalyzer RNA Integrity Number (if calculated)
<i>Acanthoecca spectabilis</i> (VA)	2	1	2	7.6	0.10	0.8	8.3
<i>Choanoeca perplexa</i>	4	3	11	4.0	0.50	2.0	8.4
<i>Codosiga gracilis</i>	1	2	5	3.3	0.60	2.0	N/A
<i>Codosiga hollandis</i>	4	3	27	1.6	0.30	0.5	8.3
<i>Diaphanoeca grandis</i>	2	1	1	6.7	0.30	2.0	9.2
<i>Didymoeca costata</i>	4	3	10	1.3	0.25	0.3	7.9
<i>Helgoeca nana</i>	1	1	8	36.2	0.02	0.7	9.3
<i>Monosiga gracilis</i>	3	2	15	3.4	0.40	1.3	7.4
<i>Monosiga ovata</i>	1	3	19	7.1	0.20	1.4	N/A
<i>Salpingoeca amphoridium</i>	4	3	3	1.0	0.80	0.8	N/A
<i>Salpingoeca gracilis</i>	3	2	16	1.1	0.35	0.4	7.2
<i>Salpingoeca infusionum</i>	2	1	12	2.5	0.80	2.0	N/A
<i>Salpingoeca minuta</i>	3	2	6	4.2	0.40	1.7	7.4
<i>Salpingoeca napiformis</i>	2	1	18	2.4	0.85	2.0	N/A
<i>Salpingoeca pyxidium</i>	1	2	9	5.0	0.20	1.0	9
<i>Salpingoeca roanokei</i>	4	3	13	4.3	0.25	1.1	N/A
<i>Salpingoeca urceolata</i>	2	1	4	1.8	0.75	1.3	N/A
<i>Savillea micropora</i>	3	2	14	10.5	0.02	0.2	8.8
<i>Stephanoeca diplocostata</i> (AU)	3	2	21	2.3	0.20	0.5	7.8
<i>Stephanoeca diplocostata</i> (FR)	1	3	7	2.6	0.25	0.7	8.2

Table 3.7. Amount of RNA used for each sample and groupings for preparation and sequencing

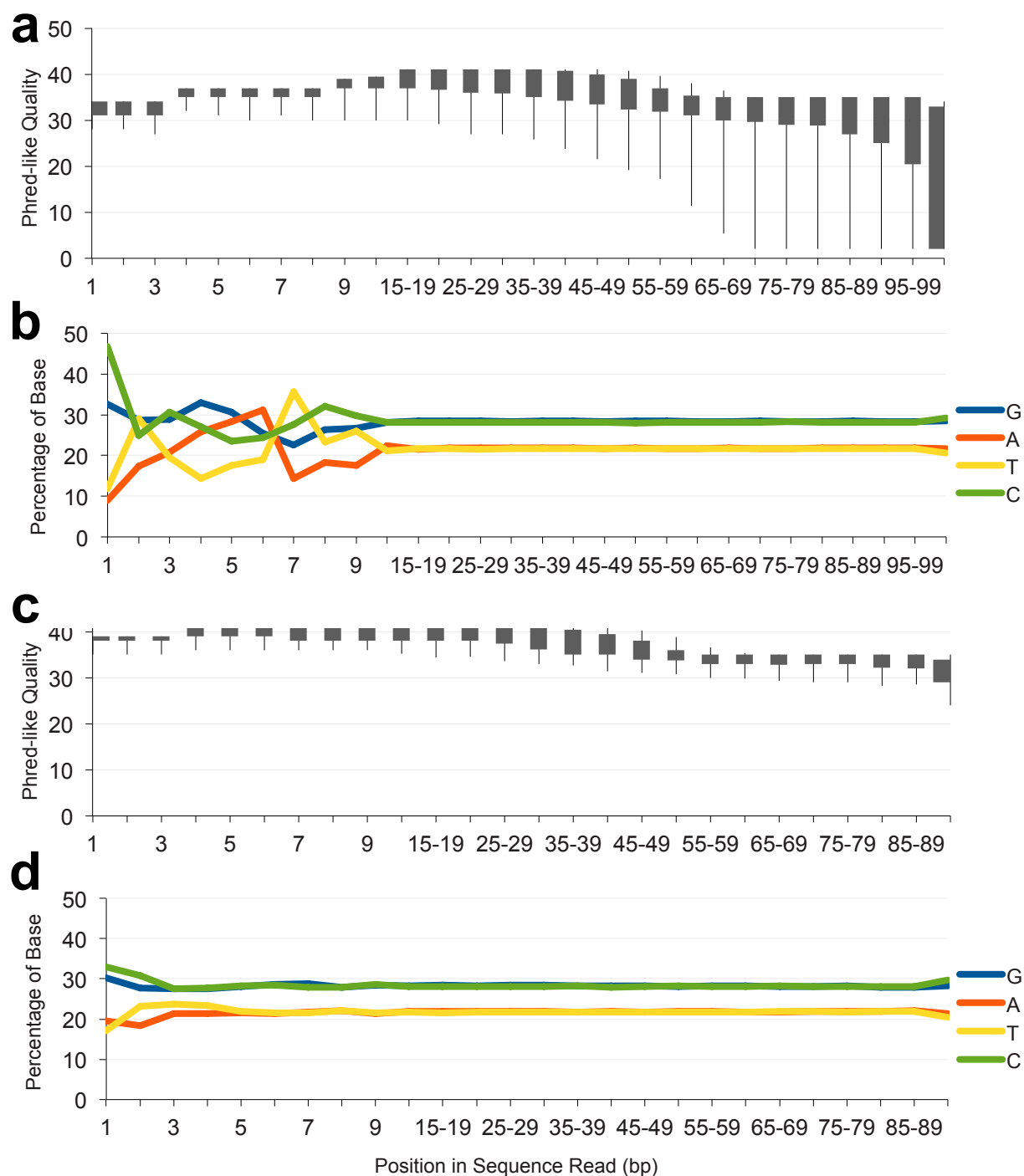


Figure 3.6. Quality distribution and skewed base content for a representative sequenced library. All data are for forward reads of *Savillea micropora*, the median species in terms of average Phred-like sequence quality. (a) Distribution of base quality over length. The bar indicates lower and upper quartiles, and the lines indicate 10th and 90th percentiles. (b) Per base sequence content, showing a skew in the first 10 bases. (c) Base quality for the same reads after trimming (including the removal of the first 10 bases). (d) Per base sequence content after trimming.

Species	Before Trimming				After Trimming		
	Read Pairs	% of Phred-like Q30 Bases	Mean Quality Score (Phred-like)	Total Bases (Gb)	Read Pairs	Single Reads (Gb)	Total Bases
<i>Acanthoea spectabilis</i> (VA)	59,109,470	82	33	11.8	50,788,323	4,875,512	8.5
<i>Choanoeca perplexa</i>	26,208,173	88	35	5.2	22,383,573	2,050,010	3.9
<i>Codosiga gracilis</i>	26,704,737	90	35	5.3	23,273,424	1,915,663	4.1
<i>Codosiga hollandis</i>	35,448,171	86	34	7.1	29,429,801	3,142,561	5.1
<i>Diaphanoeca grandis</i>	47,902,988	85	34	9.6	42,233,441	3,377,492	7.2
<i>Didymoeca costata</i>	23,189,370	90	35	4.6	20,146,116	1,637,662	3.6
<i>Helgoeca nana</i>	58,673,144	81	32	11.7	49,781,791	5,100,155	8.3
<i>Monosiga gracilis</i>	31,588,205	90	35	6.3	27,438,200	2,235,150	4.9
<i>Monosiga ovata</i>	27,246,933	84	33	5.4	22,266,888	2,686,321	3.8
<i>Salpingoeca amphoridium</i>	30,661,113	84	33	6.1	25,167,313	3,037,242	4.4
<i>Salpingoeca gracilis</i>	28,860,561	87	34	5.8	24,310,578	2,500,297	4.2
<i>Salpingoeca infusionum</i>	54,460,558	86	34	10.9	48,343,034	3,624,495	8.3
<i>Salpingoeca minuta</i>	26,880,591	84	33	5.4	21,890,021	2,803,179	3.8
<i>Salpingoeca napiformis</i>	64,254,414	81	32	12.9	54,481,059	5,715,512	9.0
<i>Salpingoeca pyxidium</i>	28,832,395	85	33	5.8	23,692,561	2,801,690	4.1
<i>Salpingoeca roanokei</i>	31,320,875	88	34	6.3	26,553,099	2,591,937	4.7
<i>Salpingoeca urceolata</i>	64,292,051	81	32	12.9	54,692,229	5,742,516	9.1
<i>Savillea micropora</i>	25,823,530	86	34	5.2	21,617,389	2,205,618	3.8
<i>Stephanoeca diplocostata</i> (AU)	28,367,867	88	34	5.7	23,919,736	2,228,549	4.2
<i>Stephanoeca diplocostata</i> (FR)	29,340,174	87	34	5.9	24,740,926	2,479,524	4.3

Table 3.8. Read counts and quality trimming

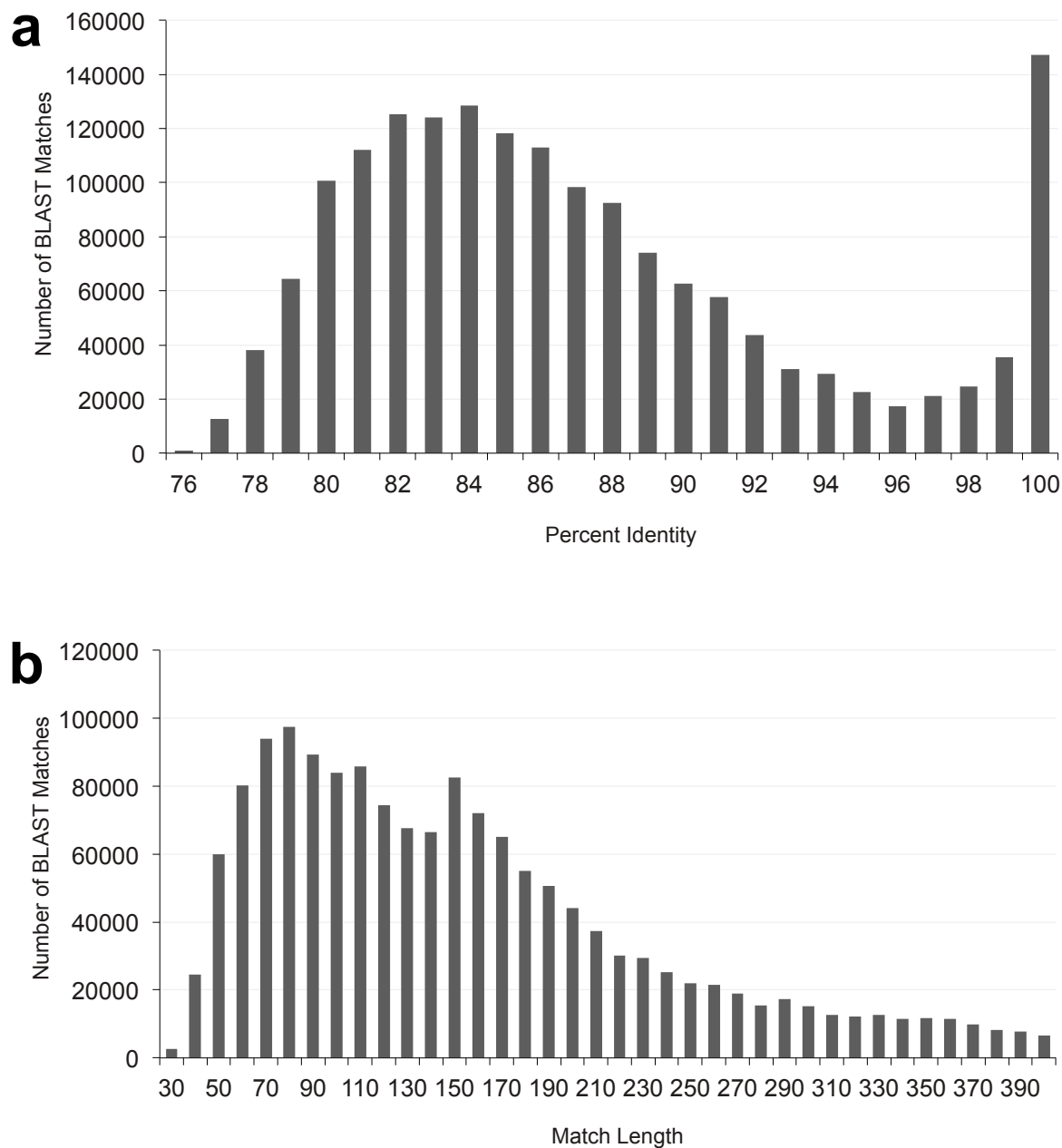


Figure 3.7. Percent match and match length distribution for cross-species BLAST matches. (a) Percent identity distribution for all BLAST matches between any two species. The distribution appears to be a mix of homologous transcripts conserved between species, centered at roughly 84 percent identity, and cross-contaminated transcripts. (b) BLAST hit match length distribution for the same matches. The peak occurring at 150 bases represents the minimum contig length from the *de novo* assembly process.

a

	A. spe	D. gra	H. nan	S. inf	S. nap	S. urc	C. gra	M. gra	S. dip (A	S. gra	S. mic	S. min	S. pyx	C. hol	C. per	D. cos	M. ova	S. amp	S. dip (F	S. roa
A. spe	63,522	1,016	1,169	3,922	1,265	1,159	0	1	2	1	56	1	2	2	1	4	4	0	1	2
D. gra	989	78,944	877	870	1,047	1,025	5	2	7	4	12	2	3	5	2	7	0	6	10	5
H. nan	1,166	921	76,920	2,125	1,076	1,059	32	16	15	35	36	18	65	20	5	15	46	7	54	16
S. inf	3,034	903	2,189	68,675	1,046	1,030	6	4	5	11	9	6	10	1	6	7	4	4	2	3
S. nap	1,206	1,025	1,010	991	44,158	1,244	2	3	2	7	6	5	2	3	6	3	8	10	0	0
S. urc	1,349	1,241	1,307	1,203	1,568	85,507	30	21	135	63	308	25	81	29	27	144	10	31	165	19
C. gra	0	8	25	6	2	12	40,427	1,026	833	530	765	829	867	23	40	36	37	23	72	24
M. gra	2	8	6	7	5	15	963	26,775	1,282	668	969	917	1,036	28	30	31	27	22	71	109
S. dip (A	4	11	18	6	3	112	823	1,381	134,699	671	1,662	841	1,161	226	117	293	202	209	115,490	250
S. gra	3	5	37	16	9	72	543	734	682	70,960	610	688	858	38	27	53	62	42	88	47
S. mic	66	15	27	7	5	211	717	1,007	1,394	541	53,829	620	807	34	29	159	35	41	421	34
S. min	1	2	6	11	5	21	782	985	825	786	682	62,268	1,041	30	25	27	29	30	50	533
S. pyx	3	5	56	10	1	71	923	1,157	1,316	809	832	946	36,752	39	30	36	73	36	133	43
C. hol	4	8	12	2	2	14	22	28	86	32	37	23	36	136,419	1,751	2,003	1,503	2,102	1,724	2,243
C. per	1	3	5	4	5	15	39	32	47	29	29	22	30	1,164	48,741	834	854	935	775	944
D. cos	4	6	15	4	2	124	38	28	200	44	194	26	37	1,166	621	64,990	614	835	1,031	662
M. ova	8	0	42	6	8	12	38	26	86	47	39	26	76	835	725	665	41,823	686	635	808
S. amp	0	3	9	3	18	27	24	20	66	40	37	33	36	1,280	835	913	796	52,954	747	824
S. dip (F	12	18	56	9	0	129	165	131	134,950	132	659	150	269	1,102	654	958	792	767	156,933	766
S. roa	4	4	5	8	0	10	24	95	91	42	34	600	37	1,176	820	739	888	792	808	58,943

b

	A. spe	D. gra	H. nan	S. inf	S. nap	S. urc	C. gra	M. gra	S. dip (A	S. gra	S. mic	S. min	S. pyx	C. hol	C. per	D. cos	M. ova	S. amp	S. dip (F	S. roa
A. spe	54,503	0	12	2	1	1	0	0	1	1	56	0	1	1	0	2	1	0	1	1
D. gra	1	74,357	3	0	1	1	0	0	2	0	2	1	0	2	0	3	0	4	0	0
H. nan	17	2	70,332	0	0	16	5	2	3	13	30	3	35	4	4	5	5	2	20	3
S. inf	3	0	0	58,778	1	1	1	0	1	2	1	3	1	0	1	3	0	1	0	0
S. nap	5	0	0	2	40,894	1	0	0	2	2	2	2	2	1	0	2	0	4	0	0
S. urc	4	1	12	2	0	77,368	2	1	10	20	259	6	44	2	7	42	0	7	57	8
C. gra	0	0	5	1	0	2	35,981	1	0	0	0	2	0	0	19	5	0	3	1	1
M. gra	0	0	2	0	0	1	1	24,187	10	2	3	1	5	2	0	0	0	0	0	2
S. dip (A	3	6	4	3	3	17	0	0	110,026	3	3	6	5	6	2	44	5	0	92,942	14
S. gra	1	0	4	5	1	25	0	3	1	66,729	1	1	15	5	1	5	1	2	2	20
S. mic	66	2	23	1	1	188	0	2	2	1	51,900	0	7	0	4	57	0	3	263	3
S. min	0	1	3	8	3	7	0	1	5	3	0	58,038	3	2	0	6	0	3	1	457
S. pyx	1	0	30	2	1	46	0	2	3	13	5	2	33,156	1	1	6	7	2	11	3
C. hol	1	1	5	0	1	2	0	2	1	4	0	2	1	118,951	9	2	2	2	0	0
C. per	0	0	5	1	0	5	18	0	1	1	4	0	1	1	41,409	0	0	2	1	0
D. cos	2	2	5	2	1	40	4	0	36	5	58	4	6	1	0	54,877	0	0	3	3
M. ova	1	0	5	0	0	0	0	0	1	1	0	0	7	0	1	0	38,589	2	0	1
S. amp	0	2	2	1	2	3	1	0	0	2	3	1	6	3	3	0	3	48,643	0	1
S. dip (F	1	0	17	0	0	55	1	0	104,527	2	347	1	17	0	0	3	0	0	125,282	0
S. roa	1	0	3	0	0	2	1	2	6	20	2	516	3	0	1	4	0	0	0	53,268

Figure 3.8. Number of BLAST matches between all pairs of species. Species are sorted into sequencing groups, each of which was run within a lane, in order to illustrate that cross-contamination occurred almost entirely within each lane. Due to space constraints, the genus name and the first three letters of each species name are shown. (a) Number of cross-contaminated contigs, as estimated by BLAST, before cross-contamination was removed and (b) after removal. A small number of hits remains after removal of cross-contamination. These are either contigs that were not removed due to high expression levels, contigs matching between the two isolates of *Stephanoea diplocostata*, bona fide homologous matches between closely-related species or due to the same species of bacteria being present in different choanoflagellate cultures.

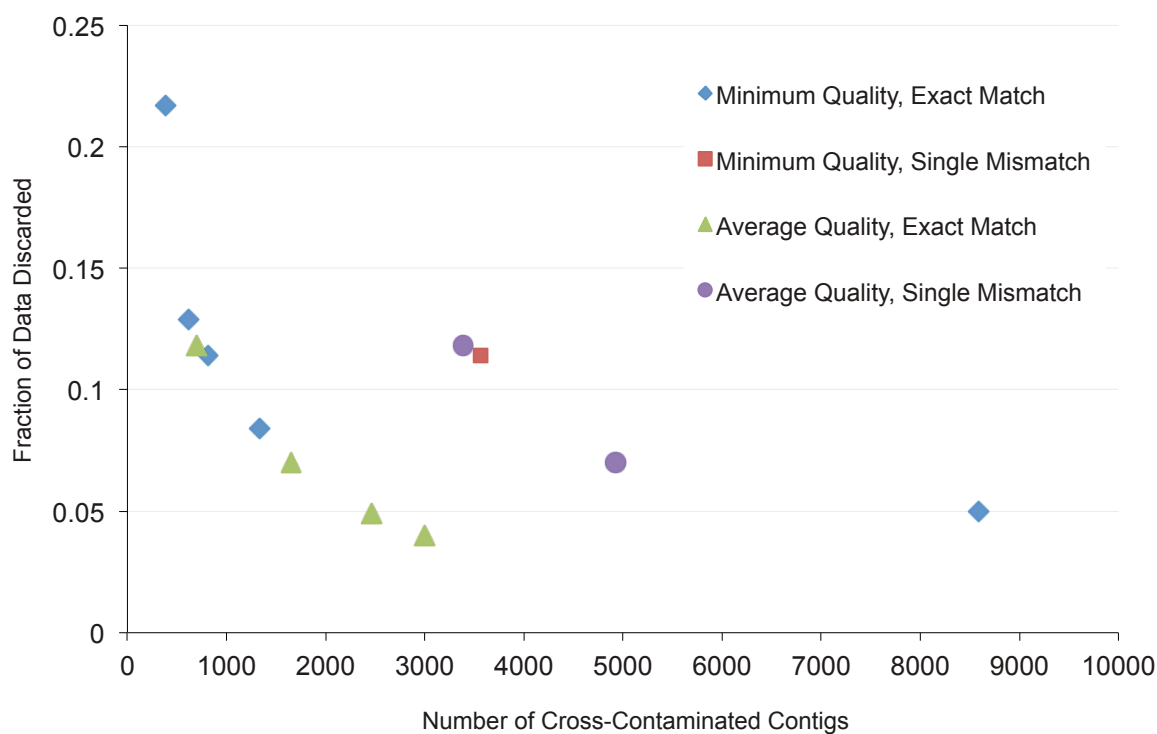


Figure 3.9. Using index quality does not remove all cross-contamination. For each cutoff strategy and requirement for exact matches or a single mismatch in the index read, the number of cross-contaminated contigs remaining after the cutoff is applied and a *de novo* assembly is performed versus the fraction of reads discarded by the cutoff. No strategy is capable of removing all cross-contaminated contigs, and strategies with more stringent quality cutoffs discard substantial proportions of the data.

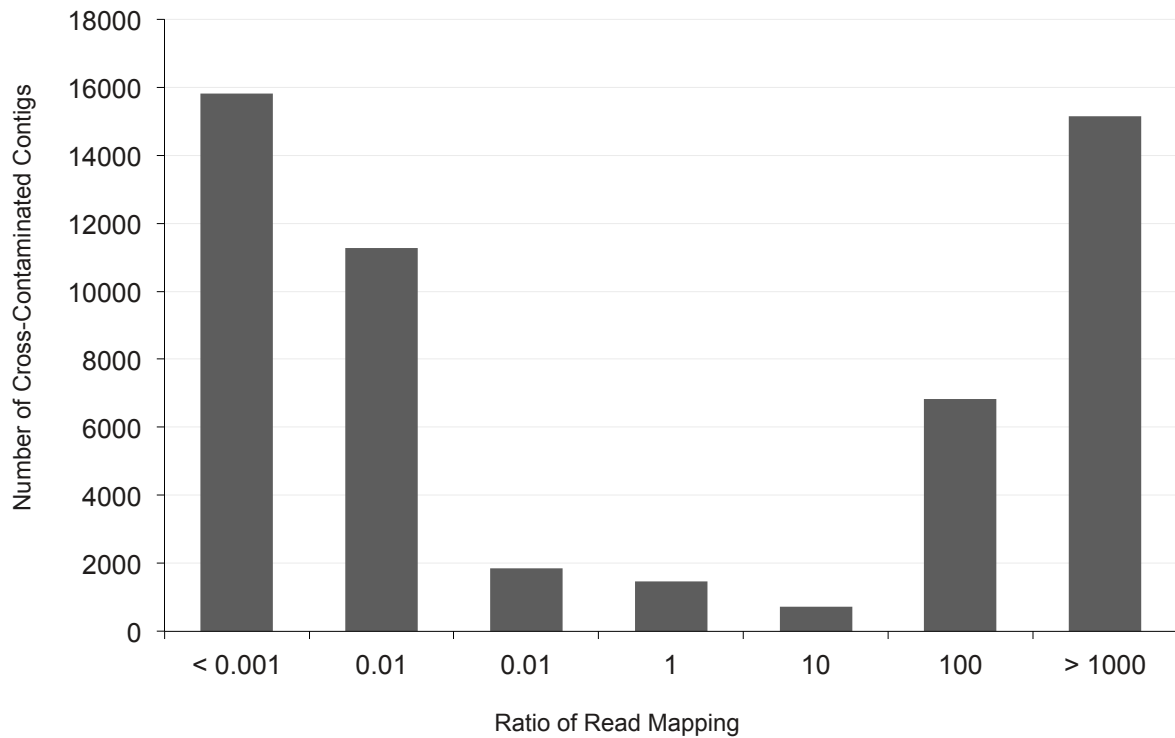


Figure 3.10. The original source of nearly all cross-contaminated contigs can be identified by read mapping. The distribution of read mapping ratios for all putatively cross-contaminated contigs. A ratio below 1 indicates that more reads mapped to a contig from another species within the same lane, whereas a ratio above 1 indicates that more reads mapped to a contig from the source species than from any other species within the same lane. Nearly all (97%) cross-contaminated contigs have a ratio much higher or lower than 1, indicating that their origin can be easily identified. We discarded the remaining 3% of contigs whose origin we could not identify.

Species	Assembled Contigs	Contigs After Decontamination	Contigs After Removal of Redundancy	Predicted Proteins	Predicted Proteins after CD-HIT
<i>Acanthoeca spectabilis</i> (VA)	64,475	62,619	55,456	69,587	37,378
<i>Choanoeca perplexa</i>	48,865	47,807	41,533	57,922	23,074
<i>Codosiga gracilis</i>	42,258	41,101	37,809	48,821	22,743
<i>Codosiga hollandis</i>	138,360	137,136	120,889	157,674	55,898
<i>Diaphanoeca grandis</i>	89,947	88,784	85,340	60,898	50,148
<i>Didymoeca costata</i>	65,525	64,194	55,412	72,423	31,402
<i>Helgoeca nana</i>	77,478	75,778	70,890	61,675	38,092
<i>Monosiga gracilis</i>	27,911	27,037	25,320	28,155	18,763
<i>Monosiga ovata</i>	42,315	41,018	39,081	36,177	25,195
<i>Salpingoeca amphoridium</i>	54,465	53,046	50,153	38,834	30,162
<i>Salpingoeca gracilis</i>	76,402	75,091	72,165	61,548	37,092
<i>Salpingoeca infusioformis</i>	69,624	65,831	59,726	64,524	26,803
<i>Salpingoeca minuta</i>	64,501	63,314	60,269	56,880	34,749
<i>Salpingoeca napiformis</i>	45,588	44,320	42,324	35,126	23,240
<i>Salpingoeca pyxidium</i>	37,025	35,378	33,429	41,236	22,289
<i>Salpingoeca roanokei</i>	62,206	60,723	56,529	50,823	30,026
<i>Salpingoeca urceolata</i>	88,283	86,969	80,144	71,782	34,616
<i>Savillea micropora</i>	54,159	53,006	52,230	38,802	33,699
<i>Stephanoeca diplocostata</i> (AU)	135,372	132,793	110,698	116,721	34,275
<i>Stephanoeca diplocostata</i> (FR)	157,624	156,051	125,972	153,521	35,319

Table 3.9. *de novo* assembly and protein prediction statistics

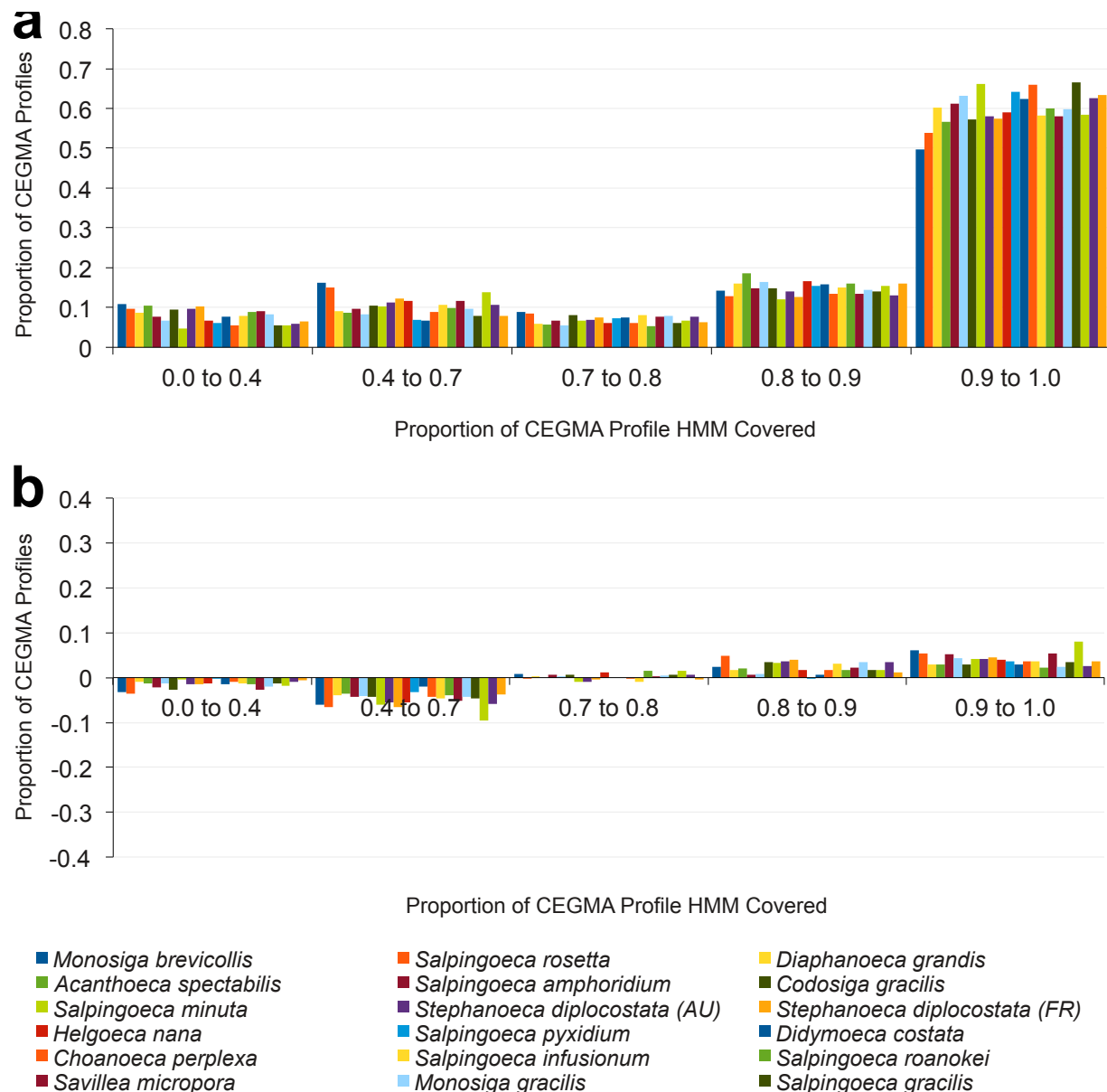


Figure 3.11. The predicted protein sets are remarkably complete for all species, as measured by comparison to core eukaryotic proteins in CEGMA. (a) For each core eukaryotic protein, we measured the proportion of that protein covered by the best match within the predicted proteins for each assembly. The distribution of coverage levels within each assembly is shown. The first two species are the two previously sequenced choanoflagellate genomes, *Monosiga brevicollis* and *Salpingoeca rosetta*. Coverage levels for all transcriptomes are at least as high as for the two complete genome sequences. (b) The change in coverage levels when all matches from a set of predicted proteins are used, not just the best match. The y axis is shown at the same scale as in (a). There is only a minor increase in coverage level (less than 0.1) from allowing all matches, indicating that the transcripts producing predicted proteins that match CEGMA profiles are not fragmentary.

Clade	Common name (if any)	Species Name	Version
Choanoflagellates			
Choanoflagellatea		<i>Monosiga brevicollis</i>	1.0
Choanoflagellatea		<i>Salpingoeca rosetta</i>	1
Animals, non-bilateria			
Porifera	Homoscleromorph sponge	<i>Oscarella sp</i>	
Porifera	Demosponge	<i>Amphimedon queenslandica</i>	1.0
Cnidaria	Freshwater cnidarian	<i>Hydra magnipapillata</i>	2.0
Cnidaria	Sea anemone	<i>Nematostella vectensis</i>	1.0
Cnidaria	Coral	<i>Acropora digitifera</i>	1.0.1
Placozoa	Trichoplax	<i>Trichoplax adhaerens</i>	1.0
Animals, deuterostomes			
Urochordata	Sea squirt	<i>Ciona intestinalis</i>	KH.68
Gnathostomata	Three-spined stickleback	<i>Gasterosteus aculeatus</i>	BROADS1
Gnathostomata	House mouse	<i>Mus musculus</i>	38.1
Cephalochordata	Lancelet	<i>Branchiostoma floridae</i>	1.0
Echinodermata	Purple sea urchin	<i>Strongylocentrotus purpuratus</i>	3.1
Hemichordata	Acorn worm	<i>Saccoglossus kowalevskii</i>	3.0
Animals, ecdysozoa			
Nematoda	Round worm	<i>Caenorhabditis elegans</i>	WS233
Nematoda	Round worm	<i>Pristionchus pacificus</i>	WS233
Crustacea	Water flea	<i>Daphnia pulex</i>	1.0
Hexapoda	Fruit fly	<i>Drosophila melanogaster</i>	5.47
Myriapoda	Centipede	<i>Strigamia maritima</i>	1.0
Chelicerata	Deer tick	<i>Ixodes scapularis</i>	1.2
Animals, lophotrochozoa			
Mollusca	Limpet	<i>Lottia gigantea</i>	1.0
Annelida	Polychaete worm	<i>Capitella teleta</i>	1.0
Annelida	Leech	<i>Helobdella robusta</i>	1.0
Holozoa			
Filasteria		<i>Capsaspora owczarzaki</i>	2
Ichthyosporea		<i>Sphaeroforma arctica</i>	1
Fungi			
Basidiomycete	Mushroom	<i>Coprinopsis cinerea</i>	3
Zygomycete	Filamentous fungus	<i>Rhizopus oryzae</i>	3
Ascomycete	Fission yeast	<i>Schizosaccharomyces pombe</i>	13.19
Chytrid		<i>Spizellomyces punctatus</i>	1
Ascomycete		<i>Yarrowia lipolytica</i>	3
Apusozoa			
Apusozoa		<i>Thecamonas trahens</i>	1
Amoebozoa			
Amoebozoa	Social amoeba	<i>Dicytostelium discoideum</i>	10-08-2012
Archaeplastids			
Plants	Thale cress	<i>Arabidopsis thaliana</i>	TAIR10
Plants	Moss	<i>Physcomitrella patens</i>	1.1
Chlorophytes		<i>Chlamydomonas reinhardtii</i>	5.3
Excavates			
Heterolobosea		<i>Naegleria gruberi</i>	1.0
Haptophytes			
Haptophytes	Coccolithophore	<i>Emiliania huxleyi</i>	1.0
Alveolates			
Ciliates		<i>Tetrahymena thermophila</i>	oct2008
Rhizaria			
Chlorarachniophytes		<i>Bigelowiella natans</i>	1.0
Stramenopiles			
Stramenochromes	Centric diatom	<i>Thalassiosira pseudonana</i>	3.0

Table 3.10. Genomes used for orthology comparison

References

- Abedin M, King N. 2008. The premetazoan ancestry of cadherins. *Science* (New York, N.Y.) 319:946–8.
- Abedin M, King N. 2010. Diverse evolutionary paths to cell adhesion. *Trends in cell biology* 20:734–42.
- Abedin M. 2010. Cadherin evolution and the origin of animals. :92.
- Adamska M, Degnan SM, Green KM, Adamski M, Craigie A, Larroux C, Degnan BM. 2007. Wnt and TGF-beta expression in the sponge *Amphimedon queenslandica* and the origin of metazoan embryonic patterning. *PloS one* 2:e1031.
- Adamska M, Larroux C, Adamski M, Green K, Lovas E, Koop D, Richards GS, Zwafink C, Degnan BM. 2010. Structure and expression of conserved Wnt pathway components in the demosponge *Amphimedon queenslandica*. *Evolution & development* 12:494–518.
- Adamska M, Matus DQ, Adamski M, Green K, Rokhsar DS, Martindale MQ, Degnan BM. 2007. The evolutionary origin of hedgehog proteins. *Current biology* : CB 17:R836–7.
- Alegado RA, Brown LW, Cao S, Dermenjian RK, Zuzow R, Fairclough SR, Clardy J, King N. 2012. A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. *eLife* 1:e00013.
- Alegado RA, Ferriera S, Nusbaum C, Young SK, Zeng Q, Imamovic A, Fairclough SR, King N. 2011. Complete genome sequence of *Algoriphagus* sp. PR1, bacterial prey of a colony-forming choanoflagellate. *Journal of bacteriology* 193:1485–6.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25:3389–402.
- Artavanis-Tsakonas S, Rand MD, Lake RJ. 1999. Notch signaling: cell fate control and signal integration in development. *Science* (New York, N.Y.) 284:770–6.
- Ashburner M, Ball CA, Blake JA, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25:25–9.
- Becker B, Marin B. 2009. Streptophyte algae and the origin of embryophytes. *Annals of botany* 103:999–1004.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 27:573–80.
- Berbee ML, Taylor JW. 1993. Dating the evolutionary radiations of the true fungi. *Canadian Journal of Botany* 71:1114–1127.

- Blau K, Portnoi M, Shagan M, et al. 2007. Flamingo cadherin: a putative host receptor for *Streptococcus pneumoniae*. *The Journal of infectious diseases* 195:1828–37.
- Bonner J. 2001. *First Signals: the Evolution of Multicellular Development*. Princeton, New Jersey: Princeton University Press
- Boraas ME, Seale DB, Boxhorn JE. 1998. Phagotrophy by a flagellate selects for colonial prey: A possible origin of multicellularity. *Evolutionary Ecology* 12:153–164.
- Bork P, Doerks T, Springer TA, Snel B. 1999. Domains in plexins: links to integrins and transcription factors. *Trends in biochemical sciences* 24:261–3.
- Bouvard D, Vignoud L, Dupé-Manet S, Abed N, Fournier H-N, Vincent-Monegat C, Retta SF, Fassler R, Block MR. 2003. Disruption of focal adhesions by integrin cytoplasmic domain-associated protein-1 alpha. *The Journal of biological chemistry* 278:6567–74.
- Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L. 2009. Fast statistical alignment. *PLoS computational biology* 5:e1000392.
- Bruns TD, Vilgalys R, Barns SM, et al. 1992. Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. *Molecular phylogenetics and evolution* 1:231–41.
- Burkhardt P, Stegmann CM, Cooper B, Kloepper TH, Imig C, Varoqueaux F, Wahl MC, Fasshauer D. 2011. Primordial neurosecretory apparatus identified in the choanoflagellate *Monosiga brevicollis*. *Proceedings of the National Academy of Sciences of the United States of America* 108:15264–9.
- Carr M, Leadbeater BSC, Baldauf SL. 2010. Conserved meiotic genes point to sex in the choanoflagellates. *The Journal of eukaryotic microbiology* 57:56–62.
- Carr M, Leadbeater BSC, Hassan R, Nelson M, Baldauf SL. 2008. Molecular phylogeny of choanoflagellates, the sister group to Metazoa. *Proceedings of the National Academy of Sciences of the United States of America* 105:16641–6.
- Carr M, Nelson M, Leadbeater BSC, Baldauf SL. 2008. Three families of LTR retrotransposons are present in the genome of the choanoflagellate *Monosiga brevicollis*. *Protist* 159:579–90.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* 17:540–52.
- Chae J, Kim MJ, Goo JH, Collier S, Gubb D, Charlton J, Adler PN, Park WJ. 1999. The *Drosophila* tissue polarity gene *starry night* encodes a member of the protocadherin family. *Development (Cambridge, England)* 126:5421–9.
- Chang DD, Wong C, Smith H, Liu J. 1997. ICAP-1, a novel beta1 integrin cytoplasmic domain-associated protein, binds to a conserved and functionally important NPXY sequence motif of beta1 integrin. *The Journal of cell biology* 138:1149–57.

- Cheloufi S, Dos Santos CO, Chong MMW, Hannon GJ. 2010. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature* 465:584–9.
- Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PloS one* 2:e383.
- Cherr GN, Yudin AI, Overstreet JW. 2001. The dual functions of GPI-anchored PH-20: hyaluronidase and intracellular signaling. *Matrix biology : journal of the International Society for Matrix Biology* 20:515–25.
- Cock JM, Sterck L, Rouzé P, et al. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617–21.
- Cock JM, Vanoosthuyse V, Gaude T. 2002. Receptor kinase signalling in plants and animals: distinct molecular systems with mechanistic similarities. *Current opinion in cell biology* 14:230–6.
- Das G, Reynolds-Kenneally J, Mlodzik M. 2002. The atypical cadherin Flamingo links Frizzled and Notch signaling in planar polarity establishment in the *Drosophila* eye. *Developmental cell* 2:655–66.
- Davies PCW, Lineweaver CH. 2011. Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. *Physical biology* 8:015001.
- Dayel MJ, Alegado RA, Fairclough SR, Levin TC, Nichols SA, McDonald K, King N. 2011. Cell differentiation and morphogenesis in the colony-forming choanoflagellate *Salpingoeca rosetta*. *Developmental biology* 357:73–82.
- Dickinson DJ, Nelson WJ, Weis WI. 2012. An epithelial tissue in *Dictyostelium* challenges the traditional origin of metazoan multicellularity. *BioEssays : news and reviews in molecular, cellular and developmental biology* 34:833–40.
- Van Dongen S, Abreu-Goodger C. 2012. Using MCL to extract clusters from networks. *Methods in molecular biology (Clifton, N.J.)* 804:281–95.
- Doonan J, Hunt T. 1996. Cell cycle. Why don't plants get cancer? *Nature* 380:481–2.
- Douzery EJP, Snell EA, Baptiste E, Delsuc F, Philippe H. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proceedings of the National Academy of Sciences of the United States of America* 101:15386–91.
- Drinnenberg IA, Fink GR, Bartel DP. 2011. Compatibility with killer explains the rise of RNAi-deficient fungi. *Science (New York, N.Y.)* 333:1592.
- Drinnenberg IA, Weinberg DE, Xie KT, Mower JP, Wolfe KH, Fink GR, Bartel DP. 2009. RNAi in budding yeast. *Science (New York, N.Y.)* 326:544–50.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32:1792–7.

- Eichinger L, Pachebat JA, Glöckner G, et al. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* 435:43–57.
- Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research* 30:1575–84.
- Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* 8:186–94.
- Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* 8:175–85.
- Fahey B, Degnan BM. 2010. Origin of animal epithelia: insights from the sponge genome. *Evolution & development* 12:601–17.
- Fahey B, Degnan BM. 2012. Origin and evolution of laminin gene family diversity. *Molecular biology and evolution* 29:1823–36.
- Fairclough SR, Dayel MJ, King N. 2010. Multicellular development in a choanoflagellate. *Current biology* : CB 20:R875–6.
- Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164–166.
- Fjordingstad EJ. 1961. Ultrastructure of the collar of the choanoflagellate *Codonosiga botrytis* (Ehrenb.). *Zeitschrift für Zellforschung und mikroskopische Anatomie* (Vienna, Austria : 1948) 54:499–510.
- Frank M, Kemler R. 2002. Protocadherins. *Current opinion in cell biology* 14:557–62.
- Gardner PP, Daub J, Tate J, et al. 2011. Rfam: Wikipedia, clans and the “decimal” release. *Nucleic acids research* 39:D141–5.
- Gauthier MEA, Du Pasquier L, Degnan BM. 2010. The genome of the sponge *Amphimedon queenslandica* provides new perspectives into the origin of Toll-like and interleukin 1 receptor pathways. *Evolution & development* 12:519–33.
- Gazave E, Lapébie P, Renard E, Bézac C, Boury-Esnault N, Vacelet J, Pérez T, Manuel M, Borchellini C. 2008. NK homeobox genes with choanocyte-specific expression in homoscleromorph sponges. *Development genes and evolution* 218:479–89.
- Gerhold AR, Richter DJ, Yu AS, Hariharan IK. 2011. Identification and characterization of genes required for compensatory growth in *Drosophila*. *Genetics* 189:1309–26.
- Gilmore TD, Wolenski FS. 2012. NF- κ B: where did it come from and why? *Immunological reviews* 246:14–35.
- Girish KS, Kemparaju K. 2007. The magic glue hyaluronan and its eraser hyaluronidase: a biological overview. *Life sciences* 80:1921–43.

- Grabherr MG, Haas BJ, Yassour M, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29:644–52.
- Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455:1193–7.
- Grosberg RK, Strathmann RR. 2007. The Evolution of Multicellularity: A Minor Major Transition? *Annual Review of Ecology, Evolution, and Systematics* 38:621–654.
- Grossmann S, Bauer S, Robinson PN, Vingron M. 2007. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics (Oxford, England)* 23:3024–31.
- Haeckel E. 1874. The gastrea-theory, the phylogenetic classification of the animal kingdom and the homology of the germ-lamelle. *Quarterly Journal of Microscopical Science* 14:142–165.
- Hallmann A. 2006. The pherophorins: common, versatile building blocks in the evolution of extracellular matrix architecture in Volvocales. *The Plant journal : for cell and molecular biology* 45:292–307.
- Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC evolutionary biology* 4:2.
- Herron MD, Hackett JD, Aylward FO, Michod RE. 2009. Triassic origin and early radiation of multicellular volvocine algae. *Proceedings of the National Academy of Sciences of the United States of America* 106:3254–8.
- Hibberd DJ. 1975. Observations on the ultrastructure of the choanoflagellate *Codosiga botrytis* (Ehr.) Saville-Kent with special reference to the flagellar apparatus. *Journal of cell science* 17:191–219.
- Hindi SM, Paul PK, Dahiya S, Mishra V, Bhatnagar S, Kuang S, Choi Y, Kumar A. 2012. Reciprocal Interaction between TRAF6 and Notch Signaling Regulates Adult Myofiber Regeneration upon Injury. *Molecular and cellular biology* 32:4833–45.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)* 17:754–5.
- Hunter T. 2009. Tyrosine phosphorylation: thirty years and counting. *Current opinion in cell biology* 21:140–6.
- Hyman LH. 1940. *The Invertebrates: Protozoa through Ctenophora*. New York: McGraw-Hill
- Hynes RO. 2002. Integrins: bidirectional, allosteric signaling machines. *Cell* 110:673–87.
- Hynes WL, Walton SL. 2000. Hyaluronidases of Gram-positive bacteria. *FEMS microbiology letters* 183:201–7.

- Jackson DJ, Meyer NP, Seaver E, et al. 2010. Developmental expression of COE across the Metazoa supports a conserved role in neuronal cell-type specification and mesodermal development. *Development genes and evolution* 220:221–34.
- Jahn R, Scheller RH. 2006. SNAREs--engines for membrane fusion. *Nature reviews. Molecular cell biology* 7:631–43.
- James-Clark H. 1867. Conclusive proofs on the animality of the ciliate sponges, and their affinities with the Infusoria Flagellata. *Memoirs of the Boston Society of Natural History* 1:305–340.
- Jinek M, Doudna JA. 2009. A three-dimensional view of the molecular machinery of RNA interference. *Nature* 457:405–12.
- Judelson HS, Ah-Fong AM V. 2010. The kinome of *Phytophthora infestans* reveals oomycete-specific innovations and links to other taxonomic groups. *BMC genomics* 11:700.
- Jøstensen J-P, Sperstad S, Johansen S, Landfald B. 2002. Molecular-phylogenetic, structural and biochemical features of a cold-adapted, marine ichthyosporean near the animal-fungal divergence, described from in vitro cultures. *European Journal of Protistology* 38:93–104.
- Kandasamy MK, McKinney EC, Roy E, Meagher RB. 2012. Plant vegetative and animal cytoplasmic actins share functional competence for spatial development with protists. *The Plant cell* 24:2041–57.
- Kawai T, Akira S. 2007. Signaling to NF-kappaB by Toll-like receptors. *Trends in molecular medicine* 13:460–9.
- Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome research* 12:656–64.
- King N, Carroll SB. 2001. A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. *Proceedings of the National Academy of Sciences of the United States of America* 98:15032–7.
- King N, Hittinger CT, Carroll SB. 2003. Evolution of key cell signaling and adhesion protein families predates animal origins. *Science (New York, N.Y.)* 301:361–3.
- King N, Westbrook MJ, Young SL, et al. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451:783–8.
- King N. 2004. The unicellular ancestry of animal development. *Developmental cell* 7:313–25.
- Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research* 40:e3.
- Kirk DL. 2005. A twelve-step program for evolving multicellularity and a division of labor. *BioEssays : news and reviews in molecular, cellular and developmental biology* 27:299–310.

- Kodner RB, Summons RE, Pearson A, King N, Knoll AH. 2008. Sterols in a unicellular relative of the metazoans. *Proceedings of the National Academy of Sciences of the United States of America* 105:9897–902.
- Koropatnick TA, Engle JT, Apicella MA, Stabb E V, Goldman WE, McFall-Ngai MJ. 2004. Microbial factor-mediated development in a host-bacterial mutualism. *Science (New York, N.Y.)* 306:1186–8.
- Koschwanetz JH, Foster KR, Murray AW. 2011. Sucrose utilization in budding yeast as a model for the origin of undifferentiated multicellularity. *PLoS biology* 9:e1001122.
- Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology* 338:1027–36.
- Lang BF, O’Kelly C, Nerad T, Gray MW, Burger G. 2002. The closest unicellular relatives of animals. *Current biology : CB* 12:1773–8.
- Lapage G. 1925. Notes on the choanoflagellate *Codosiga botrytis* Ehrenberg. *Quarterly Journal of Microscopical Science* 69:471–508.
- Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM. 2008. Genesis and expansion of metazoan transcription factor gene classes. *Molecular biology and evolution* 25:980–96.
- Leadbeater BS, Karpov SA. 2000. Cyst formation in a freshwater strain of the choanoflagellate *Desmarella moniliformis* Kent. *The Journal of eukaryotic microbiology* 47:433–9.
- Leadbeater BSC, Cheng R. 2010. Costal strip production and lorica assembly in the large tectiform choanoflagellate *Diaphanoeca grandis* Ellis. *European journal of protistology* 46:96–110.
- Leadbeater BSC. 2008. Choanoflagellate lorica construction and assembly: the nudiform condition. I. *Savillea* species. *Protist* 159:259–68.
- Lee DC, Jia Z. 2009. Emerging structural insights into bacterial tyrosine kinases. *Trends in biochemical sciences* 34:351–7.
- Leys SP, Eerkes-Medrano D. 2005. Gastrulation in Calcareous Sponges: In Search of Haeckel’s *Gastrea*. *Integrative and comparative biology* 45:342–51.
- Leys SP, Nichols SA, Adams EDM. 2009. Epithelia and integration in sponges. *Integrative and comparative biology* 49:167–77.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25:1754–60.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25:2078–9.

- Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13:2178–89.
- Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)* 22:1658–9.
- Li W, Scarlata S, Miller WT. 2009. Evidence for convergent evolution in the signaling properties of a choanoflagellate tyrosine kinase. *Biochemistry* 48:5180–6.
- Li W, Young SL, King N, Miller WT. 2008. Signaling properties of a non-metazoan Src kinase and the evolutionary history of Src negative regulation. *The Journal of biological chemistry* 283:15491–501.
- Lim WA, Pawson T. 2010. Phosphotyrosine signaling: evolving a new cellular communication system. *Cell* 142:661–7.
- Liu BA, Shah E, Jablonowski K, Stergachis A, Engelmann B, Nash PD. 2011. The SH2 domain-containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes. *Science signaling* 4:ra83.
- Liu YJ, Whelen S, Hall BD. 1999. Phylogenetic relationships among ascomycetes: evidence from an RNA polymerase II subunit. *Molecular biology and evolution* 16:1799–808.
- Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B. 2012. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic acids research* 40:W622–7.
- Lu R, Yigit E, Li W-X, Ding S-W. 2009. An RIG-I-Like RNA helicase mediates antiviral RNAi downstream of viral siRNA biogenesis in *Caenorhabditis elegans*. *PLoS pathogens* 5:e1000286.
- Manning G, Young SL, Miller WT, Zhai Y. 2008. The protist, *Monosiga brevicollis*, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. *Proceedings of the National Academy of Sciences of the United States of America* 105:9674–9.
- Mason DA, Stage DE, Goldfarb DS. 2009. Evolution of the metazoan-specific importin alpha gene family. *Journal of molecular evolution* 68:351–65.
- Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL. 2005. An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell* 122:107–18.
- Medlin L, Elwood HJ, Stickel S, Sogin ML. 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* 71:491–9.
- Meeks JC, Elhai J, Thiel T, Potts M, Larimer F, Lamerdin J, Predki P, Atlas R. 2001. An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. *Photosynthesis research* 70:85–106.

- Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic acids research* 38:D204–10.
- Miller WT. 2012. Tyrosine kinase signaling and the emergence of multicellularity. *Biochimica et biophysica acta* 1823:1053–7.
- Monsalve E, Ruiz-García A, Baladrón V, et al. 2009. Notch1 upregulates LPS-induced macrophage activation by increasing NF-kappaB activity. *European journal of immunology* 39:2556–70.
- Murgia C, Pritchard JK, Kim SY, Fassati A, Weiss RA. 2006. Clonal origin and evolution of a transmissible cancer. *Cell* 126:477–87.
- Nichols SA, Dayel MJ, King N. 2009. Genomic, phylogenetic and cell biological insights into metazoan origins. In: Telford MJ, Littlewood DTJ, editors. *Animal Evolution: Genomes, Fossils and Trees*. Oxford: Oxford University Press. p. 24–32.
- Nichols SA, Dirks W, Pearse JS, King N. 2006. Early evolution of animal cell signaling and adhesion genes. *Proceedings of the National Academy of Sciences of the United States of America* 103:12451–6.
- Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N. 2012. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/ β -catenin complex. *Proceedings of the National Academy of Sciences of the United States of America* 109:13046–51.
- Nielsen C. 2008. Six major steps in animal evolution: are we derived sponge larvae? *Evolution & development* 10:241–57.
- Nitsche F, Carr M, Arndt H, Leadbeater BSC. 2011. Higher level taxonomy and molecular phylogenetics of the Choanoflagellata. *The Journal of eukaryotic microbiology* 58:452–62.
- Ortutay C, Nore BF, Vihinen M, Smith CIE. 2008. Phylogeny of Tec family kinases identification of a premetazoan origin of Btk, Bmx, Itk, Tec, Txk, and the Btk regulator SH3BP5. *Advances in genetics* 64:51–80.
- Palaga T, Buranaruk C, Rengpipat S, Fauq AH, Golde TE, Kaufmann SHE, Osborne BA. 2008. Notch signaling is activated by TLR stimulation and regulates macrophage functions. *European journal of immunology* 38:174–83.
- Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences of the United States of America* 108:13624–9.
- Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)* 23:1061–7.

- Peterson KJ, Lyons JB, Nowak KS, Takacs CM, Wargo MJ, McPeck MA. 2004. Estimating metazoan divergence times with a molecular clock. *Proceedings of the National Academy of Sciences of the United States of America* 101:6536–41.
- Philippe H, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular biology and evolution* 22:1246–53.
- Piasecki BP, Burghoorn J, Swoboda P. 2010. Regulatory Factor X (RFX)-mediated transcriptional rewiring of ciliary genes in animals. *Proceedings of the National Academy of Sciences of the United States of America* 107:12969–74.
- Pincus D, Letunic I, Bork P, Lim WA. 2008. Evolution of the phospho-tyrosine signaling machinery in premetazoan lineages. *Proceedings of the National Academy of Sciences of the United States of America* 105:9680–4.
- Price AL, Patel NH. 2004. The evolution of gastrulation: cellular and molecular aspects. In: Stern CD, editor. *Gastrulation: from Cells to Embryo*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. p. 695–702.
- Prochnik SE, Umen J, Nedelcu AM, et al. 2010. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science (New York, N.Y.)* 329:223–6.
- Punta M, Coggill PC, Eberhardt RY, et al. 2012. The Pfam protein families database. *Nucleic acids research* 40:D290–301.
- Ratcliff WC, Denison RF, Borrello M, Travisano M. 2012. Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences of the United States of America* 109:1595–600.
- Riker AJ. 1958. PLANT TUMORS: Introduction. *Proceedings of the National Academy of Sciences of the United States of America* 44:338–9.
- Robinson PN, Wollstein A, Böhme U, Beattie B. 2004. Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology. *Bioinformatics (Oxford, England)* 20:979–81.
- Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)* 19:1572–4.
- Ruiz-Trillo I, Burger G, Holland PWH, King N, Lang BF, Roger AJ, Gray MW. 2007. The origins of multicellularity: a multi-taxon genome initiative. *Trends in genetics : TIG* 23:113–8.
- Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008. A phylogenomic investigation into the origin of metazoa. *Molecular biology and evolution* 25:664–72.
- Sakarya O, Armstrong KA, Adamska M, Adamski M, Wang I-F, Tidor B, Degnan BM, Oakley TH, Kosik KS. 2007. A post-synaptic scaffold at the origin of the animal kingdom. *PloS one* 2:e506.

- Sanderfoot A. 2007. Increases in the number of SNARE genes parallels the rise of multicellularity among the green plants. *Plant physiology* 144:6–17.
- Schroeder A, Mueller O, Stocker S, et al. 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology* 7:3.
- Schultheiss KP, Suga H, Ruiz-Trillo I, Miller WT. 2012. Lack of Csk-Mediated Negative Regulation in a Unicellular Src Kinase. *Biochemistry*.
- Sebé-Pedrós A, De Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I. 2011. Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. *Molecular biology and evolution* 28:1241–54.
- Sebé-Pedrós A, Roger AJ, Lang FB, King N, Ruiz-Trillo I. 2010. Ancient origin of the integrin-mediated adhesion and signaling machinery. *Proceedings of the National Academy of Sciences of the United States of America* 107:10142–7.
- Segawa Y, Suga H, Iwabe N, Oneyama C, Akagi T, Miyata T, Okada M. 2006. Functional development of Src tyrosine kinases during evolution from a unicellular ancestor to multicellular animals. *Proceedings of the National Academy of Sciences of the United States of America* 103:12021–6.
- Shalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, Jakobsen KS, Cavalier-Smith T. 2008. Multigene phylogeny of choanozoa and the origin of animals. *PloS one* 3:e2098.
- Shiu SH, Bleecker AB. 2001. Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases. *Proceedings of the National Academy of Sciences of the United States of America* 98:10763–8.
- De Smet I, Voss U, Jürgens G, Beeckman T. 2009. Receptor-like kinases shape the plant. *Nature cell biology* 11:1166–73.
- Snell EA, Furlong RF, Holland PW. 2001. Hsp70 sequences indicate that choanoflagellates are closely related to animals. *Current biology : CB* 11:967–70.
- Srivastava M, Simakov O, Chapman J, et al. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* 466:720–6.
- Stajich JE, Wilke SK, Ahrén D, et al. 2010. Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proceedings of the National Academy of Sciences of the United States of America* 107:11889–94.
- Steenkamp ET, Wright J, Baldauf SL. 2006. The protistan origins of animals and fungi. *Molecular biology and evolution* 23:93–106.
- Stehelin D, Varmus HE, Bishop JM, Vogt PK. 1976. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* 260:170–3.

- Suga H, Dacre M, De Mendoza A, Shalchian-Tabrizi K, Manning G, Ruiz-Trillo I. 2012. Genomic survey of premetazoans shows deep conservation of cytoplasmic tyrosine kinases and multiple radiations of receptor tyrosine kinases. *Science signaling* 5:ra35.
- Suga H, Sasaki G, Kuma K-I, Nishiyori H, Hirose N, Su Z-H, Iwabe N, Miyata T. 2008. Ancient divergence of animal protein tyrosine kinase genes demonstrated by a gene family tree including choanoflagellate genes. *FEBS letters* 582:815–8.
- Thomas PD, Kejariwal A, Campbell MJ, et al. 2003. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic acids research* 31:334–41.
- Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B. 2006. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic acids research* 34:W645–50.
- Toole BP. 2004. Hyaluronan: from extracellular glue to pericellular cue. *Nature reviews. Cancer* 4:528–39.
- Tucker RP. 2012. Horizontal Gene Transfer in Choanoflagellates. *Journal of experimental zoology. Part B, Molecular and developmental evolution*:1–9.
- Tyler S. 2003. Epithelium--the primary building block for metazoan complexity. *Integrative and comparative biology* 43:55–63.
- Usui T, Shima Y, Shimada Y, Hirano S, Burgess RW, Schwarz TL, Takeichi M, Uemura T. 1999. Flamingo, a seven-pass transmembrane cadherin, regulates planar cell polarity under the control of Frizzled. *Cell* 98:585–95.
- Velicer GJ, Kroos L, Lenski RE. 1998. Loss of social behaviors by *myxococcus xanthus* during evolution in an unstructured habitat. *Proceedings of the National Academy of Sciences of the United States of America* 95:12376–80.
- Westbrook MJ. 2011. Introns and alternative splicing in choanoflagellates. :130.
- Young SL, Diolaiti D, Conacci-Sorrell M, Ruiz-Trillo I, Eisenman RN, King N. 2011. Premetazoan ancestry of the Myc-Max network. *Molecular biology and evolution* 28:2961–71.
- Zhang Q, Wang C, Liu Z, Liu X, Han C, Cao X, Li N. 2012. Notch signal suppresses Toll-like receptor-triggered inflammatory responses in macrophages by inhibiting extracellular signal-regulated kinase 1/2-mediated nuclear factor κ B activation. *The Journal of biological chemistry* 287:6208–17.