Title

Choanoflagellate transcriptional regulation and the origin of animal cell types

Author

Coyle, Maxwell

Publication Date

2023

Choanoflagellate Transcriptional Regulation: Towards the Origin of Animal Cell Types

By

Maxwell Clark Coyle

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Nicole King, Chair
Professor Michael Eisen
Professor Iswar Hariharan
Professor Fyodor Urnov

Summer 2023

Abstract

Choanoflagellate Transcriptional Regulation: Towards the Origin of Animal Cell Types

by

Maxwell Clark Coyle

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Nicole King, Chair

Animal bodies develop through complex developmental pathways in which cells are programmed for particular fates and functions. Transcriptional regulation has been shown to be central to this process, but we know little about how transcriptional regulatory programs evolved along the animal stem lineage. Can we trace animal developmental programs to their unicellular, pre-animal roots? Which mechanistic aspects of transcriptional regulation are unique to animals and which are more deeply conserved? My doctoral research explored these questions through bioinformatic and genetic approaches in choanoflagellates, the closest living relatives of animals. Through a better understanding of transcription factors and cell type specification in these organisms, I strove to help us triangulate the transcriptional regulatory capacity of the common ancestor of animals and choanoflagellates, which lived hundreds of millions of years in the past.

Chapter 1 reviews how transcriptional regulation has evolved along the animal stem lineage. It has been frequently proposed that animal origins required the evolution of increasingly "complex" transcriptional regulation. I break this idea of complexity into specific mechanisms and trace what is known about the evolution of these mechanisms in animals and their closest living relatives.

In Chapter 2, I present an example of how functional interrogation of choanoflagellate transcriptional networks can help us better understand the ancient roles of specific transcription factors as well as the regulatory architecture of cell differentiation. I explored the function of the RFX family of transcription factors in choanoflagellates, identifying a particular sub-family (cRFXa) that has a functionally conserved role in regulating dozens of genes required for ciliogenesis. By generating genome-edited mutant strains, I show that cRFXa is essential for proper ciliogenesis in the model choanoflagellate *Salpingoeca rosetta*, and that this defect is coupled with the loss of full expression of dozens of highly conserved ciliary genes. Coupled with existing data from animals, this work shows that the RFX/ciliogenesis regulatory module dates before the divergence of animals and choanoflagellates. It also helps us to understand the regulatory changes that might have been required for the differentiation of ciliated and non-ciliated cells early in animal evolution.

Finally, in the Appendix I present work from early in my dissertation on a novel virus I helped to discover in *Entomophthora muscae*, a behavior-manipulating fungal pathogen of dipteran flies, including *Drosophila melanogaster*. We identified this virus through sequence analysis, including small RNA sequencing signatures generated by host Dicer processing, as well as through electron microscopy to directly visualize viral capsids in both cell-free extract and within fungal cells themselves.

Table of Contents

List of Figures and Tables

Chapter 1

Chapter 2

Appendix

Acknowledgements

First I would like to thank my tremendous advisor, Nicole King. Nicole has been steadfast and supportive throughout my entire graduate career. She has inspired me and taught me, delivering encouragement or constructive criticism at the right moments. Nicole showed me what it meant to hold myself to the highest standard as a person and a scientist, and her belief in me has helped me to overcome moments of self-doubt. I always leave a meeting with Nicole feeling energized and supported. Thank you, Nicole, for caring so much and for sharing your hard-earned wisdom with me.

I was blessed to find myself in a lab with so many kind and talented colleagues who each taught me important lessons. Flora Rutaganira was an example of how to combine rigorous and fearless science with a deeply compassionate and dependable presence. Thibaut Brunet taught me how to think precisely about evolution, how to pick out the most interesting questions, and how to have a goofy time in lab. David Booth showed me how to channel a deep passion for science and for people into daily practices. Alain Garcia de Las Bayonas demonstrated how to do science beautifully, grounded in a reverence for the natural world. I would also like to thank many more people who helped to make the King Lab a home for me: Michael Carver, Josean Reyes-Rivera, Jacob Steenwyk, Becca Arruda, Chrisa Staikou, Monika Sigg, Laura Wetzel, Ella Ireland, Tess Linden, Fredrick Leon, Ben Larson, and many more. I have also greatly appreciated the collaborative and kind spirit of the broader choanoflagellate and animal origins community.

Beyond the King Lab, I would like to thank many in the Berkeley biology community for making my time here fun and interesting, particularly the Fletcher Lab and the Tri/Tetra/Penta/Hexa-labs. I have learned from many Berkeley faculty, especially Mike Eisen, Fyodor Urnov, Iswar Hariharan, Elçin Ünal, Doug Koshland, and Barbara Meyer. I have been comforted and delighted by so many smart and compassionate friends: Dan Fines, Amanda Su, Yui Iwamoto, Rebecca Tarnopol, Chris Hoel, Victoria Blake, and more.

I would be remiss to not give a special acknowledgement to my cats, Pisco (Mr. Boots) and Ponce (The Little Cat). Human language doesn’t appropriately capture inter-species bonds, but I here display a series of slow blinks to show my appreciation.

I’m so incredibly grateful for the deep love of my sisters Emily and Rachael. The toughness, compassion, and heart of my mom Sherilyn are my foundation in life. During graduate school, we lost my father Rich to glioblastoma. Dad, you gave me the curiosity that has brought me to a career in science. I spent long periods of dissertation writing with a Mariners game on in the background and I felt you there with me.

Finally, Erika. The best thing that happened to me in graduate school was building a life with you. It is such a comfort and joy to know that whatever life serves up next, we will ride it out together.

Chapter 1

The evolution of transcriptional regulation in animal origins

Introduction

The morphological and behavioral complexity of animals is arguably unrivaled in the living world. While most eukaryotes exist as single cells or small, undifferentiated colonies (Brunet and King 2017), the multicellular bodies and diverse cell types of animals allow them to fly, swim, camouflage, echolocate, play, communicate, and perform tetrad dissections (Johnston and Mortimer 1959). Given the importance of transcriptional regulation in animal development and cellular differentiation, it has been hypothesized that the cellular diversity of animals and their unusual biology stemmed from the evolution of ever more complex transcriptional regulatory mechanisms (Levine and Tjian 2003; Sebé-Pedrós, Degnan, and Ruiz-Trillo 2017; Erwin 2020).

Here we revisit this hypothesis and examine its support in light of recent data. Because "complexity" does not have a single clear biological meaning, we focus on a set of empirical features that may contribute to transcriptional complexity, analyzing the phylogenetic distribution and functions of these features in animals and non-animals. In part 1 of this review, we aim to break down the idea of transcriptional complexity into several specific parameters and trace the evolution of these parameters along the stem lineage leading to modern animals. In part 2, we apply our conceptual framework to specific examples of transcriptional regulatory programs that share some conservation between animals and their closest relatives.

Why focus on transcription, given that it is but one of many possible modes of gene regulation? Functional perturbations of transcription factors (TFs) in animals demonstrate their powerful regulatory capacity, particularly in the orchestration of development and cellular differentiation. Transcription factor manipulations can turn fibroblasts into pluripotent stem cells (Takahashi et al. 2007), reprogram the antennae of flies into legs (Schneuwly, Klemenz, and Gehring 1987), and trigger devastating cancers (Muller and Vousden 2013). Modifications of transcriptional regulatory programs can lead to novelty and diversity in animal evolution, including the evolution of wing patterning in butterflies (Wallbank et al. 2016), skeletal changes in stickleback fish (Chan et al. 2010), and the loss of limbs in snakes (Kvon et al. 2016).

While other lineages (e.g. plants, fungi) have independently evolved high degrees of developmental complexity, to understand the role of transcriptional regulation in animal origins, the transcriptional apparatuses of animals must be compared with those of their closest relatives. In the last twenty years, molecular phylogenetic analyses have confirmed the proposition, based on morphological evidence, that choanoflagellates represent the closest living relatives of animals; together, choanoflagellates and animals form the choanozoan clade. Other groups of protists, including filastereans and ichthyosporeans have been shown to be closely related to choanozoans and therefore provide additional context for reconstructing the ancestry of transcriptional regulation on the animal stem lineage (Ruiz-Trillo et al. 2008; Shalchian-Tabrizi et al. 2008; Ros-Rocher et al. 2021).

Choanoflagellates are free-living aquatic microbial eukaryotes that eat bacteria and can form small colonies (Leadbeater 2015). Choanoflagellates transition between a diversity of cell types in response to environmental factors, such as nutrient availability (Levin and King 2013), spatial confinement (Brunet et al. 2021), and the presence of bacterial biomolecules (Alegado et al. 2012; Woznica et al. 2017). Some of these cell type transitions are regulated at the transcriptional level (Fairclough et al. 2013). Filastereans and ichthyosporeans also show temporally defined and environmentally responsive form of cellular differentiation (Suga and Ruiz-Trillo 2013; Sebé-Pedrós, Irimia, et al. 2013) (Figure 1.1). The choanoflagellate *Salpingoeca rosetta* and the filasterean *Pigoraptor chileana* have at least seven distinct cell types, but our knowledge of cell differentiation capabilities in these protistan animal relatives may be an underestimate given the impossibility of re-creating the diversity of ecological conditions in the lab (Figure 1.1). The upper bound of cell type numbers for protistan animal relatives overlaps with the lower bound of cell type numbers reported for basally branching animal groups like placozoans and sponges, which have as few as 6 morphologically distinguishable cell types (Figure 1.1). However, single-cell transcriptomics has prompted a reassessment of cell type numbers in animals, pushing the upper bound to 30 for some basal animal groups, 50 for cnidarians, and more than 200 for some bilaterian species (Figure 1.1). The transcriptional regulatory mechanisms behind animal development and differentiation likely have roots in an environmentally-responsive form of cellular differentiation still found in the closest living relatives of animals.

Now that we have identified the relevant taxa and their patterns of cellular differentiation, we will break down the concept of transcriptional "complexity" into several specific parameters and compare these among taxonomic groups. First, we will look at evolutionary patterns in the number of TFs (Figure 1.2A), the generation of novel TF families (Figure 1.2B), and changes in DNA-binding specificity within these families (Figure 1.2C). Then we will consider the combinatorial action of TFs, including the evolution of protein-protein interactions (Figure 1.2D) and mechanisms of distal regulation (i.e. enhancers; Figure 1.2E).

Part 1: Measures and modes of transcriptional complexity

Number of TFs

TFs are proteins that regulate transcription by binding to DNA in a sequence-specific manner (Lambert et al. 2018). Any protein with these properties can be properly labeled as a TF, but most annotated TFs have been detected from genome assemblies by searching for genes that encode a DNA-binding domain (DBD) from a previously-characterized TF family (de Mendoza et al. 2013; Lambert et al. 2018). Recent years have seen a wave of genomes and transcriptomes from groups that are phylogenetically well positioned for reconstructing animal origins, including basally branching animals like sponges, ctenophores, and placozoans, as well as close animal relatives including choanoflagellates, filastereans, and ichthyosporeans (Srivastava et al. 2010; Ryan et al. 2013; Srivastava et al. 2008; King et al. 2008; Grau-Bové et al. 2017). By scanning these genomes, TF numbers can be mapped onto species trees, allowing comparisons among animals and non-animals and revealing changes in TF repertoires at important evolutionary nodes.

While many vertebrates encode more than 1000 TFs, animals in most phyla encode 300-700 predicted TFs, and the basally branching groups (placozoans, ctenophores and sponges) encode fewer than 300 (Figure 1.3A) (de Mendoza et al. 2013; Schmitz, Zimmer, and Bornberg-Bauer 2016). On the other hand, the closest animal relatives (choanoflagellates, ichythosporeans, and filastereans) generally encode less than 200 TFs, with exceptions like the ichthyosporean *Amoebidium parasiticum*, whose 345 TFs outpace the basal animal groups (Figure 1.3A) (de Mendoza et al. 2013). Therefore, while animals generally have more TFs than non-animals, the differences within the animal kingdom far exceed the gaps between close animal relatives and basally branching animals (Figure 1.3A).

The increased number of TFs in animals is due predominantly to expansions of TF families that predate animal origins (Figure 1.3A, Table 1.1) (de Mendoza et al. 2013), although contributions from novel, animal-specific TF families will be discussed in the next section. In the animal stem lineage, notable expansions occurred in the Homeodomain, C2H2 zinc finger and Forkhead families (de Mendoza and Sebé-Pedrós 2019). Within animals, lineage-specific expansions are also observed, including C2H2 zinc fingers in cephalopods and mammals (Albertin et al. 2015; Lambert et al. 2018), nuclear hormone receptors in nematodes (Schmitz, Zimmer, and Bornberg-Bauer 2016), and p53 in elephants (Abegglen et al. 2015).

Identifying transcription factors by DBD similarity can produce both false positives and false negatives. False positives occur when proteins encode a DBD but do not function as transcription factors, e.g. C2H2 zinc fingers involved exclusively in RNA binding (Joho et al. 1990) or homeodomains co-opted for ceramide synthase regulation (Mesika et al. 2007). But perhaps more problematic are false negatives: transcription factors that are not identified because their DBD is not in our library of known DBDs. To this point, 69 genes without a canonically known DBD have been identified as human TFs by the criteria of sequence-specific DNA binding (Lambert et al. 2018). It is difficult to know *a priori* how many TFs and DBD families have yet to be discovered in less well-studied organisms.

Novel TF families

TF families are primarily characterized by the DBDs they contain. The library of known DBDs, and therefore TF families, has increased over time as new families have been revealed through functional studies. Depending on the exact database analyzed, between 70 and 100 families of DBDs can be distinguished in eukaryotes (de Mendoza and Sebé-Pedrós 2019; Lambert et al. 2018; Finn et al. 2016). It is generally reasoned that each DBD family evolved once and therefore that the members of each family are more closely related to each other than to TFs of other families. One exception to this is the very short (13 amino acid) AT-hook domain, which can easily appear by convergent evolution and might function as a DBD in hundreds of genes that have not yet been identified as likely TFs (Aravind and Landsman 1998).

TF families novel to animals include nuclear hormone receptors, Ets, and MADF (Table 1.1) (de Mendoza and Sebé-Pedrós 2019). These all appear in the genomes of basally branching animals, indicating their origin in the animal stem linage (Srivastava et al. 2010; Ryan et al.

2013; Srivastava et al. 2008). How do new DBDs evolve? While many may evolve *de novo* from the mutation of non-coding regions or the mutation of coding regions without TF function, others may come from the duplication and rapid divergence of existing DBDs, as well as the domestication of DNA-binding transposase domains (Figure 1.2B) (Weirauch and Hughes 2011; de Mendoza and Sebé-Pedrós 2019; Babu et al. 2006; Aravind et al. 2005)

Animals are not the only group of eukaryotes to have evolved novel TF families, which are also seen in fungi (e.g. STE) and land plants (e.g. GRAS) (de Mendoza et al. 2013). Among protists, the IBD TF in *Trichomonas vaginalis* contains an evolutionarily unique DBD fold (Schumacher, Lau, and Johnson 2003), while apicomplexans have been shown to encode a novel family of TFs related to but highly divergent from AP2 TFs (Balaji et al. 2005). ApiAP2 TFs make up the majority of identifiable TFs in apicomplexans (de Mendoza et al. 2013). This means that uncharacterized, lineage-specific TF families may present a major challenge when it comes to assessing full TF repertoires. Given this possibility, as well as the fact that basal animals and close animal relatives remain poorly studied, it is not clear how confident we should be of the differences between both the number of TFs as well as the number of TF families in animals versus non-animals (Figure 1.3A).

How can we identify novel TF families? Unbiased genetic approaches, including forward genetics, are likely to be of great value. Structural similarity approaches, perhaps bolstered by tools like AlphaFold (Jumper et al. 2021), may help to find TF families with DBDs that are highly divergent, but still structurally related to, known families. One-hybrid screens can help identify TFs that bind to specific DNA sequences (Reece-Hoyes and Marian Walhout 2012), while proteomic surveys of nuclear and DNA-binding components can help to generate lists of candidate TFs (Tacheny et al. 2013).

Sequence specificity

Another metric by which a TF repertoire can expand its regulatory capacity is in its ability to recognize a wider range of DNA motifs (Figure 1.2C). Mutations affecting DNA binding can change the sequence specificity of a TF, so that homologous TFs in the same species or different species can recognize different DNA motifs (Baker, Tuch, and Johnson 2011).

How widespread is the evolution of novel DNA specificities among the TFs of animal and their closest relatives? TF specificities can be determined *in vitro* using techniques like protein binding microarrays (PBMs) and systematic evolution of ligands by exponential enrichment (SELEX) (Jolma and Taipale 2011). Both allow TF binding to be queried against a large library of short oligonucleotides. While these assays do not reproduce the *in vivo* complexity of a TF binding landscape, they can be performed in relatively high throughput and across any species with genomic data available. One study compared 242 orthologous TFs between mammals and *Drosophila* (separated by ~600 million years of evolution) and found almost perfect agreement in DNA binding specificity. However, other studies have shown that particular families, like C2H2 zinc fingers, can diverge rapidly in their specificities, even across the *Drosophila* clade (45 millions year of evolution) (Nadimpalli, Persikov, and Singh 2015). Further studies have confirmed that different TF families display very different rates of changes in DNA specificity. The C2H2 zinc fingers are relatively unique in their fast evolution, possibly due to their modular arrangement of tandem repeats (Wolfe, Nekludova, and Pabo 2000). Other families like Myb/SANT also diverge quickly (Lambert et al. 2019). However, many families – including Homeodomain, nuclear receptor, Ets, Sox/HMG, and Forkhead – show highly similar specificities even between mammals and cnidarians (Lambert et al. 2019).

Systematic surveys comparing TF specificities between animals and their closest relatives have not been published. The TFs for which binding data has been published in both animals and close animal relatives belong to families with highly conserved DNA specificities, like RFX (Coyle et al. 2023) and T-box (Sebé-Pedrós, Ariza-Cosano, et al. 2013). In these studies, the DNA-binding specificities were almost identical in animal and non-animal orthologs. For the fast-diverging families like C2H2 zinc fingers, the rapid rate of sequence evolution makes it difficult to assign orthologs between animals and non-animals, and therefore to assess the degree of DNA binding conservation, although it is expected to be low.

Given that the evolution of new DNA binding specificities is particularly enriched in just a few TF families, it is especially notable that one of these families, the C2H2 zinc fingers, have undergone large expansions multiple times in animals: first in the animal stem lineage, and again within various animal sub-groups, including mammals and cephalopods (Schmitz, Zimmer, and Bornberg-Bauer 2016; Albertin et al. 2015). In mammals, C2H2 zinc finger TFs commonly contain accessory KRAB domains that facilitate gene silencing and their recognition sites are enriched in different retrotransposons, suggesting that the diversification of this family is part of an intragenomic arms race (Najafabadi et al. 2015). However, it has also been shown that the protein-protein interactions of these C2H2 zinc fingers are almost as diverse as the sequences they recognize, suggesting that they have distinct functional roles and are likely sites of regulatory innovation (Schmitges et al. 2016; Imbeault, Helleboid, and Trono 2017).

Protein-protein interactions

Transcription factors participate in a variety of protein-protein interactions (PPIs): with co-activators and co-repressors, with the pre-initiation complex machinery, with other TFs, with histones, and with chromatin readers and writers. These interactions provide a rich substrate for the evolution of transcriptional regulatory mechanisms (Figure 1.2D).

Many PPIs are mediated by regions outside of the DNA-binding domain, known as accessory domains. The domain architecture of a TF refers to the number, order, and orientation of accessory domains relative to a DBD. Novel domain architectures have been shown to evolve rapidly in many classes of animal TFs (Schmitz, Zimmer, and Bornberg-Bauer 2016), and may provide the main sources of innovation for those families for which DNA-binding specificity remains relatively unchanged. The number of domain architectures scales logarithmically with the size of a TF family and many TF family expansions are preceded by the innovation of a novel domain architecture (Schmitz, Zimmer, and Bornberg-Bauer 2016). While some protein-protein interactions can be traced to well-conserved and identifiable interfaces, others require regions that may be short, unstructured, and highly divergent in sequence space (Plevin, Mills, and Ikura 2005). The presence of these small and labile interaction interfaces, as well as the complex and non-linear interaction networks among PPIs, makes it nearly impossible to accurately predict protein-protein interactions of TFs solely from genomic sequence. Therefore, comparative genomics can only take us so far when it comes to assessing which TF repertoires encode more combinatorial complexity than others with respect to PPIs.

Some TF families bind DNA as obligate dimers, with the capacity to form homodimers, heterodimers, or both, e.g. the bZIP family (Rodríguez-Martínez et al. 2017). For some bZIP heterodimers, the DNA binding preferences of the heterodimer is a concatenation of the preferences of each binding partner, while in other cases emergent and unpredictable DNA binding specificities result (Rodríguez-Martínez et al. 2017). The bZIPs are also the only TF family where a systematic attempt has been made to compare PPIs between animal and non-animal repertoires. By testing the *in vitro* binding affinities of hundreds of potential bZIP heterodimers across a range of taxa, it was shown that animal bZIPs former denser interaction networks, i.e. the proportion of possible heterodimers that can function together is greater in animals than in yeast or choanoflagellates (Reinke et al. 2013). This may have allowed the animal bZIP network to become more complex, even in the absence of large changes in TF family size or the DNA binding specificity of individual TFs.

Beyond direct interaction, there are other modes through which TFs may cooperate with one another. For instance, clusters of TF binding sites at enhancers are proposed to allow TFs to evict nucleosomes and maintain open chromatin by mass action, even in the absence of direct binding to one another (F. Reiter, Wienerroither, and Stark 2017). Since enhancer-mediated regulation is such a hallmark of animal transcription (Levine and Tjian 2003), and proposed to be an animal innovation (Erwin 2020), we will now discuss this particular mechanism directly.

Distal enhancers

In animals, the transcription level of a gene can be affected by regulatory sequences called enhancers, which can be separated from their target core promoters by dozens to millions of bases (Levine 2010; Banerji, Rusconi, and Schaffner 1981). Particular attention has been paid to “distal” enhancers (located in regions far from the core promoter, although no consistent threshold has been defined for “far”) on developmentally controlled genes, where multiple enhancers controlling the same downstream gene can be activated in different developmental contexts (Levine, Cattoglio, and Tjian 2014; Chan et al. 2010). This has led to the hypothesis that the evolution of distal enhancers was essential for the evolution of animal developmental complexity (Erwin 2020; Levine 2010). By this argument, complex development required some genes to receive more regulatory inputs than could be accommodated by regions close to the core promoter. Supporting this, transcriptional regulation driven by distal enhancers has been well-characterized in several bilaterian model systems – human cells, mice, zebrafish, and *Drosophila*.

Enhancer sequences evolve rapidly, losing and gaining TF binding sites, with the result that it can be difficult to identify homologous enhancer regions even within groups like drosophilids (Hare et al. 2008) and mammals (Villar et al. 2015). Therefore, to look for enhancers outside of bilaterians, we cannot rely on simple sequence similarity metrics the way we might for coding regions.

Despite this limitation, analysis of chromatin features has suggested that basally branching animals might utilize distal enhancers (Figure 1.4A). In bilaterians, enhancers are associated with chromatin marks such as H3K27Ac and H3K4me1, or depositors of these marks such as p300 (an acetyltransferase) (Visel et al. 2009). In the cnidarian *Nematostella vectensis*, many p300 peaks can be identified more than 300 bases from transcription start sites and these peaks show both H3K4me1 and H3K27Ac enrichment (Schwaiger et al. 2014). About 75% of tested enhancers were validated in reporter assays. In the sponge *Amphimedon queenslandica*, patches of H3K4me1 enrichment more than 200 bases from transcription start sites identified several putative regulatory sites, although no functional validation is currently possible in this organism (Figure 1.4A) (Gaiti et al. 2017).

Despite this pioneering work, the scope of distal regulation implicated by this data remains murky. The distance thresholds that were used to identify regulatory elements (200 bp, 300 bp) are well within the functional range of cis-regulatory elements in *S. cerevisiae*, which is not typically understood to have distal regulation (Dobi and Winston 2007). Second, the functional relevance of these chromatin marks is uncertain. For instance, in both flies and mouse embryonic stem cells, genome-wide loss of H3K4me1 methylation has only minor phenotypic and gene-regulatory consequences (Rickels et al. 2017; Dorighi et al. 2017). Finally, using an scRNAseq dataset for *Amphimedon queenslandica*, patterns of gene expression can be well-predicted by promoter proximal elements alone (Sebé-Pedrós, Chomsky, et al. 2018). This challenges the idea that the involvement of distal enhancers is strictly necessary for all animal development.

Whether the closest animal relatives use distal enhancers is unknown. In *Capsaspora owczarzaki,* chromatin accessible sites distal to transcription start sites (defined as 800 bp) did not show enrichment of H3K4me1 over H3K4me3 and were smaller than similar sites found in animals (Sebé-Pedrós et al. 2016). However, no sites could be functionally tested at the time and the limitations of using H3K4me1 as an enhancer proxy have been discussed. Finally, since *Capsaspora*, like choanoflagellates and other close animal relatives, has a compact genome, finding enhancers far from promoters limits the search space. This limitation is unjustified since the literature on distal enhancers strongly supports the prevalence of promoter-proximal sequences acting as distal enhancers at other promoter regions, e.g. the SV40 enhancer (Banerji, Rusconi, and Schaffner 1981). This dual promoter-enhancer function may be highly prevalent as suggested by the STARR-seq high-throughput assay in *Drosophila* cells (Zabidi et al. 2015).

How might we detect whether animal relatives make use of distal transcriptional regulation? Ultimately, we will need to look for functional validation. Distal enhancers, if they do exist in these organisms, might likely derive from introns or promoters of nearby genes (Figure 1.4B). The presence of p300 in introns, particularly overlapping with likely TF motifs, could provide a list of possible candidates. Evolutionary analysis of intron gains and losses point to a massive intronization event in the last common ancestor of animals and choanoflagellates (Grau-Bové et al. 2017). As genome editing becomes more efficient in choanoflagellates and other animal relatives, high throughput assays like STARR-seq may also provide lists of putative enhancers.

Part 2: Case studies for pre-animal transcriptional networks and their modification in animals

Genome sequencing of choanoflagellates and other animal relatives revealed that many TFs essential for animal development and cell type differentiation are also present in protistan relatives of animals, including p53, Runx, Myc, T-box, RFX, and NF-κB TFs (Table 1.1) (Sebé-Pedrós et al. 2011; de Mendoza et al. 2013). The presence of animal developmental TFs in non-animals revealed that the origin of animal developmental gene regulation was not simply due to the evolution of novel "developmental" genes. Rather, many animal developmental genes were likely co-opted from functions they previously served in a unicellular, non-animal context.

In this section, we compare specific TF regulatory programs that operate in animals and close animal relatives. These models exemplify how transcriptional regulation has evolved in the animal stem lineage.

Myc:Max

Together, the Myc and Max transcription factors regulate a broad diversity of cell fates in animals, including division, differentiation, and apoptosis (Eilers and Eisenman 2008). Myc:Max heterodimers promote cell division partially by activating suites of genes required for ribosome biogenesis, a rate-limiting step in cell proliferation (Eilers and Eisenman 2008; van Riggelen, Yetil, and Felsher 2010). The two TFs are members of the bHLH family and dimerize through their leucine zippers. Upon heterodimerization, the pair can bind to a palindromic DNA motif called the E-box (CACGTG) (Figure 1.5A). Because Myc requires Max for heterodimerization and binding, Myc activity can be indirectly inhibited by the sequestration of Max, either when Max forms homodimers or heterodimerizes with Mad or Mnt, two additional members of the Myc/Max bHLH sub-family (Figure 1.5A) (Grandori et al. 2000).

Myc and Max are encoded by close animal relatives, including choanoflagellates and *Capsaspora owczarzaki*. In both organisms, the role of Myc:Max in regulating ribosome biogenesis is likely conserved (Figure 1.5A). E-boxes are found in the promoter regions of conserved ribosome biogenesis genes in animals and the choanoflagellate *Monosiga brevicollis*, but not in *Saccharomyces cerevisiae* (Brown, Cole, and Erives 2008). The set of genes for which both animal and choanoflagellates homologs contain an E-box consists almost entirely of ribosome biogenesis components (Brown, Cole, and Erives 2008). Moreover, *M. brevicollis* Myc and Max can heterodimerize and bind to E-boxes *in vitro* (Young et al. 2011). However, while sequence-specific Myc:Max binding appears to be conserved between animals and choanoflagellates, E-box presence alone is not sufficient to predict Myc-driven regulation, as both Max homodimers and even bHLH family members outside of the Myc sub-family can bind these same E-boxes (K. A. Robinson and Lopes 2000). In *Capsaspora,* E-boxes are also enriched in the promoters of ribosome biogenesis genes and these regions show chromatin signatures of activation in the proliferative stage (Sebé-Pedrós et al. 2016).

The apparent conservation of Myc:Max regulation of ribosome biogenesis in choanozoans and *Capsaspora* suggests that the Myc:Max network regulated ribosome biogenesis in the unicellular progenitors of animals. This in itself represents an evolutionary change, as this sub-family of TFs is not found in the vast majority of eukaryotic diversity (Young et al. 2011). It is an example of how an ancient DBD family, the bHLH TFs, can diversify through the establishment of new sub-families that co-opt ancient regulatory functions, such as the regulation of ribosome biogenesis. It is possible that the consolidation of ribosome biogenesis control under Myc:Max regulation opened new possibilities for increasingly complex regulation, as the network of homodimers and heterodimers within this sub-family allows for many possible inputs to influence the essential decision of whether to undergo cell division.

Myc functionality in animals goes beyond regulating ribosome biogenesis (Figure 1.5A). This elaboration likely stems from its combinatorial action with other genes beyond Max and may involve novel types of protein-protein interactions. Outside of the DBD, vertebrate Myc proteins contain four other domains, only two of which are conserved in non-animal relatives (Young et al. 2011). One of these domains, Myc homology box IV, is only found in vertebrates and regulates apoptosis (Cowling et al. 2006). The Myc example shows how a core TF regulatory mechanism evolved in the unicellular ancestors of animals and was later expanded to play diverse roles in animal development. This expansion of possible functions derived from the combinatorial power of distinct heterodimers to regulate Myc:Max activity as well as the evolution of other types of protein-protein interactions.

RFX

RFX (regulatory factor X) TFs regulate ciliogenesis in a wide diversity of animals, from vertebrates to *Drosophila* to *C. elegans* (Quigley and Kintner 2017; Dubruille et al. 2002; Swoboda, Adler, and Thomas 2000). Cilia are produced by many animal cell types, including sperm, most epithelial cells, and numerous cells of sensory function (photoreceptors, olfactory neurons) (J. F. Reiter and Leroux 2017). Ciliogenesis requires the complex orchestration of hundreds of genes, and the coordinated transcription of this set must occur in the proper cell types at the right developmental time points (Choksi et al. 2014). Animal RFX TFs regulate ciliogenesis target genes by binding to the recognition site GTTRCY (Figure 1.5A) (Jolma et al. 2013). RFX can bind as a monomer or a dimer, in which case the recognition site consists of a palindrome of inverted half-sites (Reith et al. 1990, 1994; Gajiwala et al. 2000). Notably, RFX TFs are not found in most eukaryotes (many of which bear cilia), being restricted to opisthokonts and amoebozoans, which together form the Amorphea clade in many modern eukaryotic phylogenies (Swoboda, Adler, and Thomas 2000; Adl et al. 2012; Coyle et al. 2023). Before this year, the only published functional data on RFX function outside of animals was in ascomycete fungi, which use RFX TFs to regulate DNA damage repair and the cell cycle (Hao et al. 2009; Wu and McLeod 1995; Huang, Zhou, and Elledge 1998).

My recent work explored the function of RFX TFs in the choanoflagellate *Salpingoeca rosetta*, revealing a conserved (and therefore pre-animal) regulatory link between RFX and ciliogenesis genes (Coyle et al. 2023). This study also showed that the RFX TF family expanded from one to three members on the choanozoan stem lineage, which may have coincided with its acquisition of a role in regulating ciliogenesis (Coyle et al. 2023). Interestingly, one specific sub-family that resulted from this duplication is responsible for almost all published reports of RFX regulating ciliogenesis, in both animals and now choanoflagellates (Swoboda, Adler, and Thomas 2000; Chung et al. 2012; Coyle et al. 2023). The sequence specificity of RFX TFs is among the most highly conserved of known TFs, being almost identical in fungi, choanoflagellates, and animals, and between different RFX sub-families (Jolma et al. 2013; Badis et al. 2008; Coyle et al. 2023). This is supported by the almost invariant conservation of DNA-contacting residues in RFX DBDs (Piasecki, Burghoorn, and Swoboda 2010; Coyle et al. 2023).

The RFX example illustrates how animal transcriptional regulation can be shaped by the emergence and subsequent expansion of novel DBD classes. Cilia are ancient eukaryotic organelles whose biogenesis was likely regulated by other transcriptional mechanisms before the appearance of the RFX family (Carvalho-Santos et al. 2011). The RFX TF family appeared in the ancestors of opisthokonts and amoebozoans, and only later (in the choanozoan stem) did RFX adopt control over ciliogenesis. The three ancient choanozoan RFX sub-families underwent additional expansions in vertebrates, further partitioning functions. For instance, RFX2 in mammals specifically controls ciliary gene expression in spermatogenesis (Kistler et al. 2015), while RFX3 controls ciliary gene expression in other tissues (Bonnafe et al. 2004). RFX1, on the other hand, is embryonic lethal in mice (Feng, Xu, and Zuo 2009), and may have retained a cell cycle function that may be as ancient as opisthokonts, given the role of RFX in cell cycle regulation in fungi (Bugeja, Hynes, and Andrianopoulos 2010) and the growth defect observed in an RFX knockout in choanoflagellates (Coyle et al. 2023).

While the DNA-binding specificity of RFX TFs has remained unchanged over more than a billion years of evolution, the ability of RFX to function as a monomer or a dimer may allow plasticity in regulatory evolution. *S. rosetta* RFX appears to function almost entirely through monomeric sites (Coyle et al. 2023), while animal RFX ChIP-seq motifs often contain a “strong” and “weak” half-site, which may represent the cumulative signature of monomeric and dimeric binding sites (Figure 1.5A) (Lemeille et al. 2020). Furthermore, different RFX sub-family members can heterodimerize, which may allow further points of regulatory control. RFX often co-regulates its ciliary gene targets with another TF, FoxJ1 (Choksi et al. 2014). Vertebrate RFX and FoxJ1 have been shown to physically associate (Quigley and Kintner 2017), but it is still unknown how old this binding interaction is or more generally how RFX protein-protein interactions compare between animals and animal relatives (Figure 1.5B). Finally, despite the prevalence of enhancer-mediated regulation in animals, most RFX binding sites in animal ciliary genes are located close to transcription start sites (Sugiaman-Trapman et al. 2018). This may be a consequence of the ancient and highly conserved nature of this regulatory program, which may date to a time before distal transcriptional regulation was common.

Conclusion

The studies of Myc and RFX show how TFs coordinate gene expression to enable cellular functions in close animal relatives. Notably, these relatives all exhibit complex life histories with several functionally distinct cell types (Suga and Ruiz-Trillo 2013; Sebé-Pedrós, Irimia, et al. 2013; Dayel et al. 2011; Alegado et al. 2012). RNA sequencing experiments have shown these cell types to be transcriptionally distinct, with numerous TFs differentially expressed (Fairclough et al. 2013; Sebé-Pedrós et al. 2016). Some of these cell types are part of the sexual cycle (gametes, spores), while others form in response to environmental cues or stressors (colonies, aggregates, cysts, dispersal forms) (Figure 1.5B). Overall, we are assembling a picture of a unicellular ancestor of animals that could regulate its gene expression and cellular phenotype to perform multiple functions: amoeboid and/or flagellar-driven motility, digestion, secretion, sex, and cell division. Some regulatory modules (RFX and ciliogenesis, Myc:Max and ribosome biogenesis) were likely already in place. It is likely that cell differentiation preceded animal origins as part of temporally defined and environmentally-responsive programs (Zakhvatkin 1949; Mikhailov et al. 2009). As this unicellular ancestor evolved into a complex multicellular animal, several modifications were made to its transcriptional regulatory apparatus.

Animal evolution likely did involve an increase in the total number of transcription factors, as revealed by comparing animal TF repertoires to those of their closest relatives (de Mendoza et al. 2013; de Mendoza and Sebé-Pedrós 2019). We do recommend some caution around this claim, given the likelihood of undiscovered TF families, particularly in understudied lineages like protistan eukaryotes. It is also notable that basally branching animals like sponges and ctenophores have TF repertoires close in number to those of animal relatives. While novel TF families do appear in animals, the bulk of the increased TF repertoire in animals comes from the expansion of more ancient TF families (de Mendoza et al. 2013; de Mendoza and Sebé-Pedrós 2019).

For some families, like C2H2 zinc fingers and Myb/SANT TFs, the specific DNA sequences they recognize have been shown to diverge quickly in evolutionary time (Nadimpalli, Persikov, and Singh 2015; Lambert et al. 2019). Therefore, as these families expand, they can create more complex transcriptional networks by virtue of creating a wider vocabulary of genomic recognition sequences. This might explain the success of C2H2 zinc fingers, having undergone expansions in the stem lineage of animals, as well as within animal clades like mammals and cephalopods (Albertin et al. 2015; Schmitz, Zimmer, and Bornberg-Bauer 2016; Lambert et al. 2018).

However, the majority of TF families retain similar DNA binding specificities across long evolutionary periods and appear to generate increased complexity through networks of protein-protein interactions. This is illustrated by the bZIPs, where it has been shown that even without a great expansion of gene family size, animal bZIPs show increased heterodimerization capabilities and therefore increased combinatorial possibilities (Reinke et al. 2013). It is also illustrated by examples like Myc, which has likely undergone an expansion of function from regulating ribosome biogenesis to regulating diverse aspects of cell proliferation, differentiation, and apoptosis, all while maintaining a conserved DNA binding specificity (Eilers and Eisenman 2008; Young et al. 2011; Brown, Cole, and Erives 2008). The essential nature of PPIs in the evolution of transcriptional networks presents a challenge, as these cannot be assayed as systematically and efficiently as DNA binding preferences *in vitro*. The development of transgenic tools should help with the identification of *in vivo* interacting partners for TFs (Booth and King 2020; Phillips et al. 2022; Kożyczkowska et al. 2021).

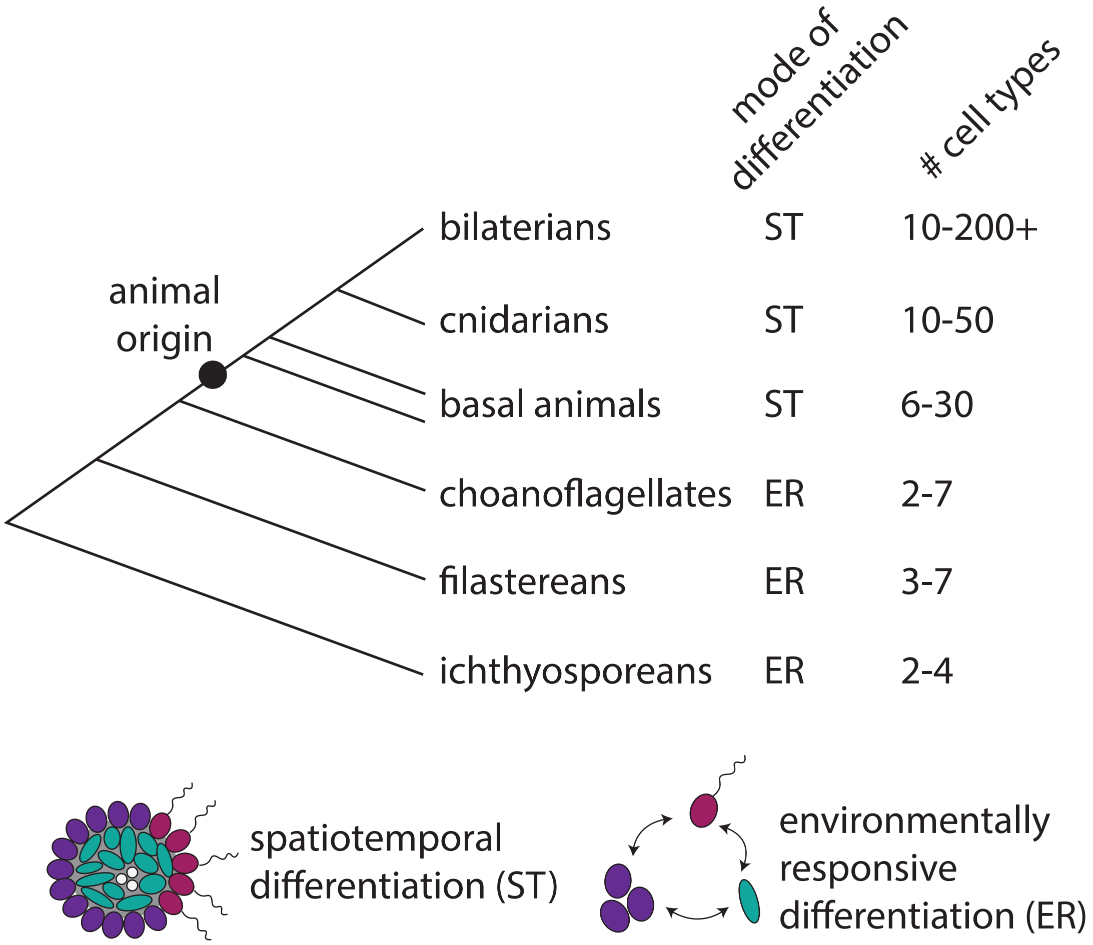
The role of enhancers in animal origins is unclear. Given the lack of functional studies in sponges, ctenophores, choanoflagellates, and other close animal relatives, there is no hard evidence suggesting the prevalence, or lack thereof, of distal regulation in these key lineages. However, there is increasing support for distal regulation in cnidarians, with numerous functionally validated enhancers in *Nematostella vectensis* (Schwaiger et al. 2014; Sebé-Pedrós, Saudemont, et al. 2018). Therefore the mechanism of distal enhancer regulation may be more important for animal diversification than for animal origins.

In this review we have attempted to address the specific contributions of one hypothesized driver of animal origins: an increase in the complexity of transcriptional regulation. We have done this through identifying parameters of transcriptional regulatory networks that contribute to complexity, and then comparing what is known about these parameters in both animals and their closest living relatives. Moving forward, we advocate specifically for (1) more conceptual development around the biological meaning of “complexity” and how it is generated by specific molecular mechanisms, as well as (2) more functional studies in taxa key to understanding animal origins, particularly basally branching animals and protistan animal relatives.

Figures and Tables

Figure 1.1. Multicellular development and spatiotemporal cell differentiation evolved in the animal clade.

All animals display spatiotemporal cellular differentiation as part of development, while all close animal relatives display environmentally-responsive temporal cell differentiation. The range provided for the number of cell types in each lineage combines data from morphology and transcriptomics, with the morphological estimate representing the lower bound and single cell RNA sequencing data representing an upper bound, although the upper bound for bilaterians is particularly unclear. Cell type number ranges were drawn from the following references: bilaterians (Valentine, Collins, and Porter Meyer 1994; Bell and Mooers 1997; Plass et al. 2018; Cao et al. 2019; Tabula Muris Consortium et al. 2018; Hulett et al. 2023), cnidarians (Chapman 1974; Bell and Mooers 1997; Sebé-Pedrós, Saudemont, et al. 2018; Siebert et al. 2019; Levy et al. 2021), basal animals (Simpson 1984; Bell and Mooers 1997; Smith et al. 2014; Sebé-Pedrós, Chomsky, et al. 2018; Musser et al. 2021), choanoflagellates (Dayel et al. 2011; Levin and King 2013; Leadbeater 2015; Brunet et al. 2021), filastereans (Sebé-Pedrós, Irimia, et al. 2013; Tikhonenkov et al. 2020), and ichthyosporeans (Suga and Ruiz-Trillo 2013). It is hypothesized that both the transition from temporal to spatiotemporal cell differentiation as well as the overall increase in number of cell types required an increasingly complex apparatus for transcriptional regulation (Levine and Tjian 2003; Sebé-Pedrós, Degnan, and Ruiz-Trillo 2017; Erwin 2020).



	mode of differentiation	# cell types
bilaterians	ST	10-200+
cnidarians	ST	10-50
basal animals	ST	6-30
choanoflagellates	ER	2-7
filastereans	ER	3-7
ichthyosporeans	ER	2-4

animal origin
spatiotemporal differentiation (ST)
environmentally responsive differentiation (ER)

Figure 1.2. Evolutionary modes for transcriptional regulation

(A) The number of transcription factors encoded by an organism's genome can increase (or decrease) over evolutionary time. Increases can be due to expansion of existing TF families (green and blue DBDs) or the appearance of new TF families (red DBD) (de Mendoza et al. 2013; de Mendoza and Sebé-Pedrós 2019).

(B) New transcription factor families, characterized by novel DBD folds, may arise in various ways, including the duplication and rapid divergence of existing DBDs where little sequence homology is retained. For instance, a number of eukaryotic DBDs utilize helix-turn-helix structural motifs for DNA binding and some of these may be evolutionary related (Aravind et al. 2005; Weirauch and Hughes 2011). New TF families may also arise *de novo* from non-coding sequence (or from coding sequence without an initial DBD function) or from transposases (Weirauch and Hughes 2011; de Mendoza and Sebé-Pedrós 2019; Babu et al. 2006).

(C) Transcription factors can change their DNA-binding specificity to recognize new sequences. This is most commonly due to mutations in the DBD, particularly in residues that contact DNA (Lambert et al. 2019). Some TF families, most notably the C2H2 zinc fingers, often contain tandem repeated DBDs, and in this family novel DNA specificities arise from both the expansion (or contraction) of tandem repeats, often combined with substitution mutations in the repeats themselves (Najafabadi et al. 2015).

(D) Transcription factor functions frequently evolve through changes in protein-protein interactions. These mutations are often located in domains outside of the DBD that mediate these interactions. Domain acquisition, loss, rearrangement, or mutation can all affect the PPIs available for a given TF (Schmitz, Zimmer, and Bornberg-Bauer 2016). Some regions that mediate PPIs are very small and degenerate, making their prediction difficult and their evolutionary appearance or disappearance rapid (Plevin, Mills, and Ikura 2005). Another type of PPI affecting TF function is dimerization. Many families (bZIPs, bHLHs, RFX) can form functional units by both homodimerization and heterodimerization, and mutations within the dimerization domains can change the number and function of possible dimers (Rodríguez-Martínez et al. 2017; Grandori et al. 2000).

(E) Transcriptional regulation can become more complex by involving more TFs in the regulation of a given target gene, and distal enhancers provide a mechanism for increasing the number of regulatory inputs (Levine 2010). Distal enhancers often contain clusters of TF binding sites, and TF cooperativity at these sites (which may be mediated by direct binding or indirect mechanisms) is essential for their function (F. Reiter, Wienerroither, and Stark 2017).

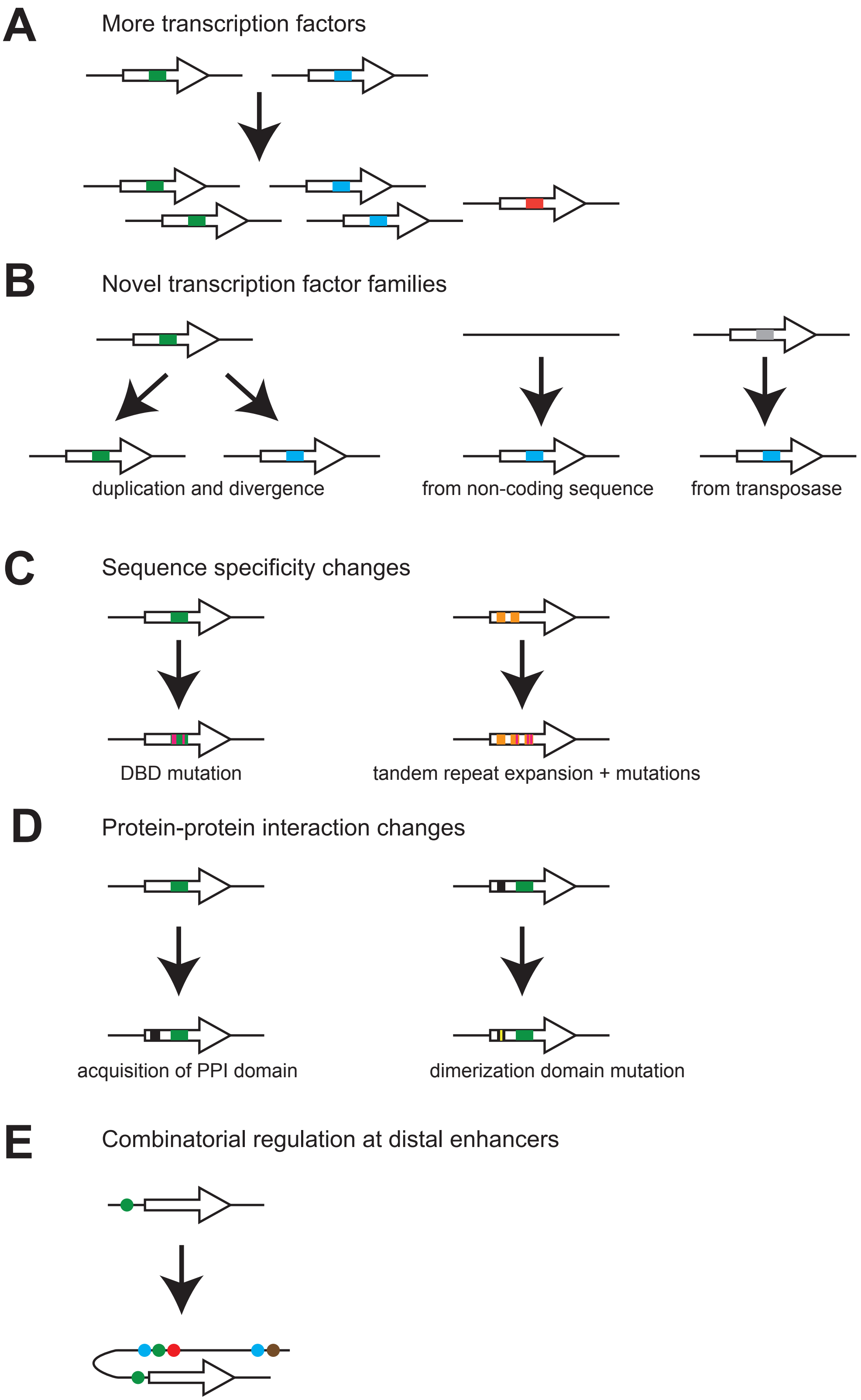

A
More transcription factors
B
Novel transcription factor families
duplication and divergence
from non-coding sequence
from transposase
C
Sequence specificity changes
DBD mutation
tandem repeat expansion + mutations
D
Protein-protein interaction changes
acquisition of PPI domain
dimerization domain mutation
E
Combinatorial regulation at distal enhancers

Figure 1.3. The expansion and diversification of TF repertoires.

(A) Animals on average encode more transcription factors than their closest relatives. Bilaterians and cnidarians encode more TFs than basal animals, while vertebrates (represented here by *Homo sapiens* and *Danio rerio*) encode more than most other bilaterians. Increases in TF number is driven more by the expansion of pre-animal TF families (gray) than by the appearance of animal-specific TF families (blue). Data taken from a comprehensive survey of eukaryotic TF distribution (de Mendoza et al. 2013). Animal-specific families defined by (de Mendoza et al. 2013) are CUT, DM, Ets, GCM, IRF, MADF_DNA_bdg, MH1, TF_AP-2, TSC22, and zf-C4.

(B) Animals have increased the combinatorial possibilities of bZIP TFs by increasing the proportion of functional heterodimers formed by different bZIP TFs. Data taken from (Reinke et al. 2013) in which *in vitro* binding assays are used to assess functional heterodimers. Binding assay data from 21 °C was used for display here, although binding assays were also performed at 4 °C and 37 °C. Taxon diagrams from phylopic.org, dedicated to the public domain under a CC0 1.0 Universal Public Domain Dedication license.

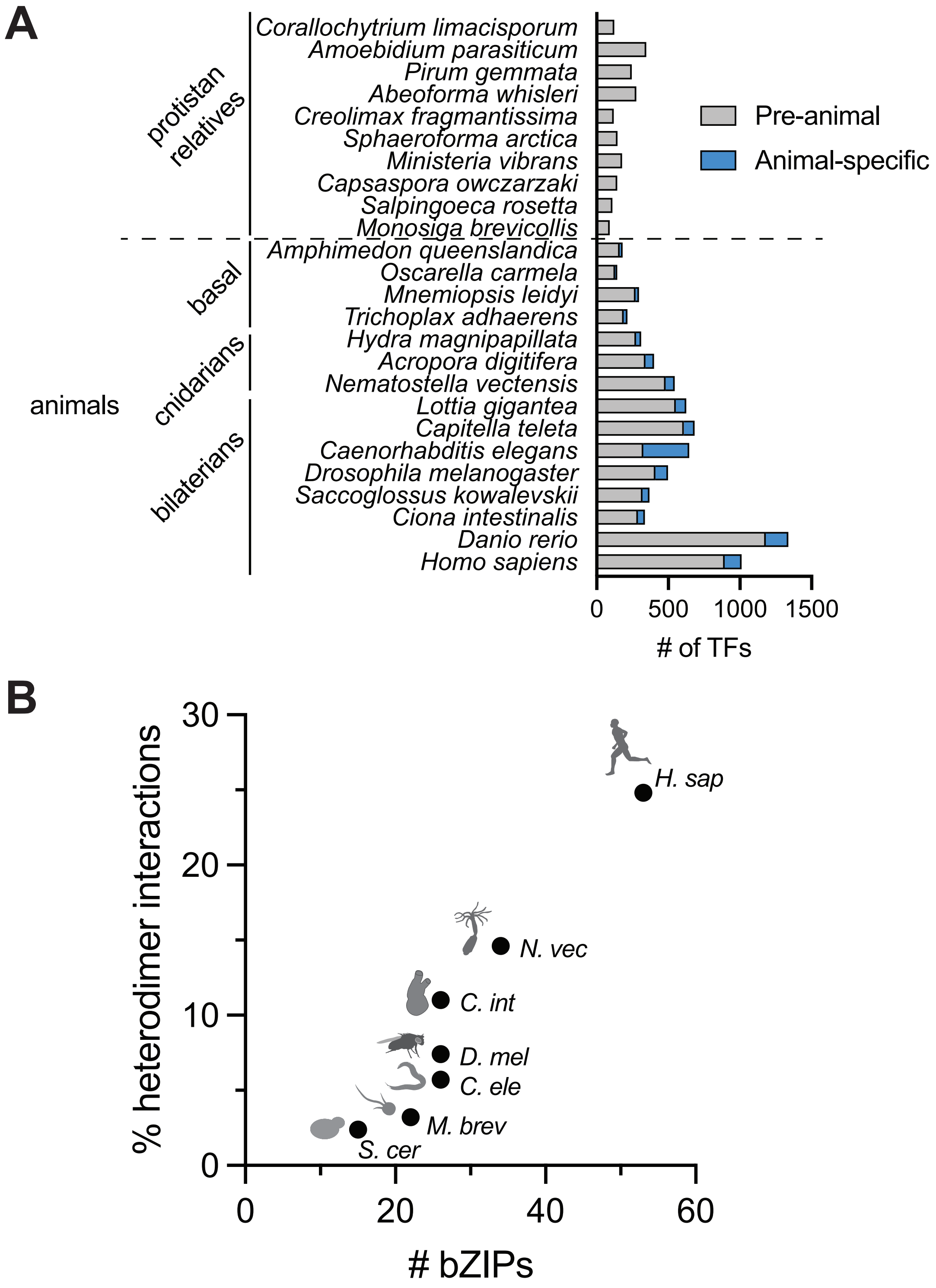

A
protistan relatives
Corallochytrium limacisporum
Amoebidium parasiticum
Pirum gemmata
Abeoforma whisleri
Creolimax fragmantissima
Sphaeroforma arctica
Ministeria vibrans
Capsaspora owczarzaki
Salpingoeca rosetta
Monosiga brevicollis
animals
basal
Amphimedon queenslandica
Oscarella carmela
Mnemiopsis leidyi
Trichoplax adhaerens
cnidarians
Hydra magnipapillata
Acropora digitifera
Nematostella vectensis
bilaterians
Lottia gigantea
Capitella teleta
Caenorhabditis elegans
Drosophila melanogaster
Saccoglossus kowalevskii
Ciona intestinalis
Danio rerio
Homo sapiens
Pre-animal
Animal-specific
0
500
1000
1500
# of TFs
B
% heterodimer interactions
30
20
10
0
H. sap
N. vec
C. int
D. mel
C. ele
M. brev
S. cer
0
20
40
60
# bZIPs

Figure 1.4. The presence of enhancers in close animal relatives remains ambiguous.

(A) The presence of enhancers in different animal and non-animal taxa is indicated, drawing from different lines of evidence, including the presence of genomic regions with enhancer-associated chromatin marks (“chromatin”), the demonstration of enhancer function with reporter assays or through effects on endogenous gene expression (“function”) as well as the conservation of sequence features with validated enhancers (“conservation”) in other lineages. Black circles indicate reports of enhancers meeting these criteria, white circles indicate the lack of these criteria being met (where tested), and gray indicates the lack of assessment. Note that functional tests of enhancers have not been performed in basal animals or close animal relatives. References: bilaterian chromatin (Visel et al. 2009), cnidarian chromatin (Schwaiger et al. 2014), basal animal chromatin (Gaiti et al. 2017), filasterean chromatin (Sebé-Pedrós et al. 2016), bilaterian function (Levine 2010), cnidarian function (Schwaiger et al. 2014), bilaterian conservation (Villar et al. 2015), basal animal conservation (Wong et al. 2020).

(B) An example gene (green transcription start site), and the location of various regulatory elements in its surrounding neighborhood. The orange (500) and purple (200) promoter-proximal elements are incapable of initiating enhancer-like activation in an orientation-independent manner. All elements have open chromatin and H3K27Ac. Shown below are examples of regions picked out as enhancer elements by different chromatin-based profiling strategies. The red (1800) element is never identified due to its proximity to the TSS of a neighboring gene. The orange element may be misidentified as having distal enhancer properties if a short window is used.

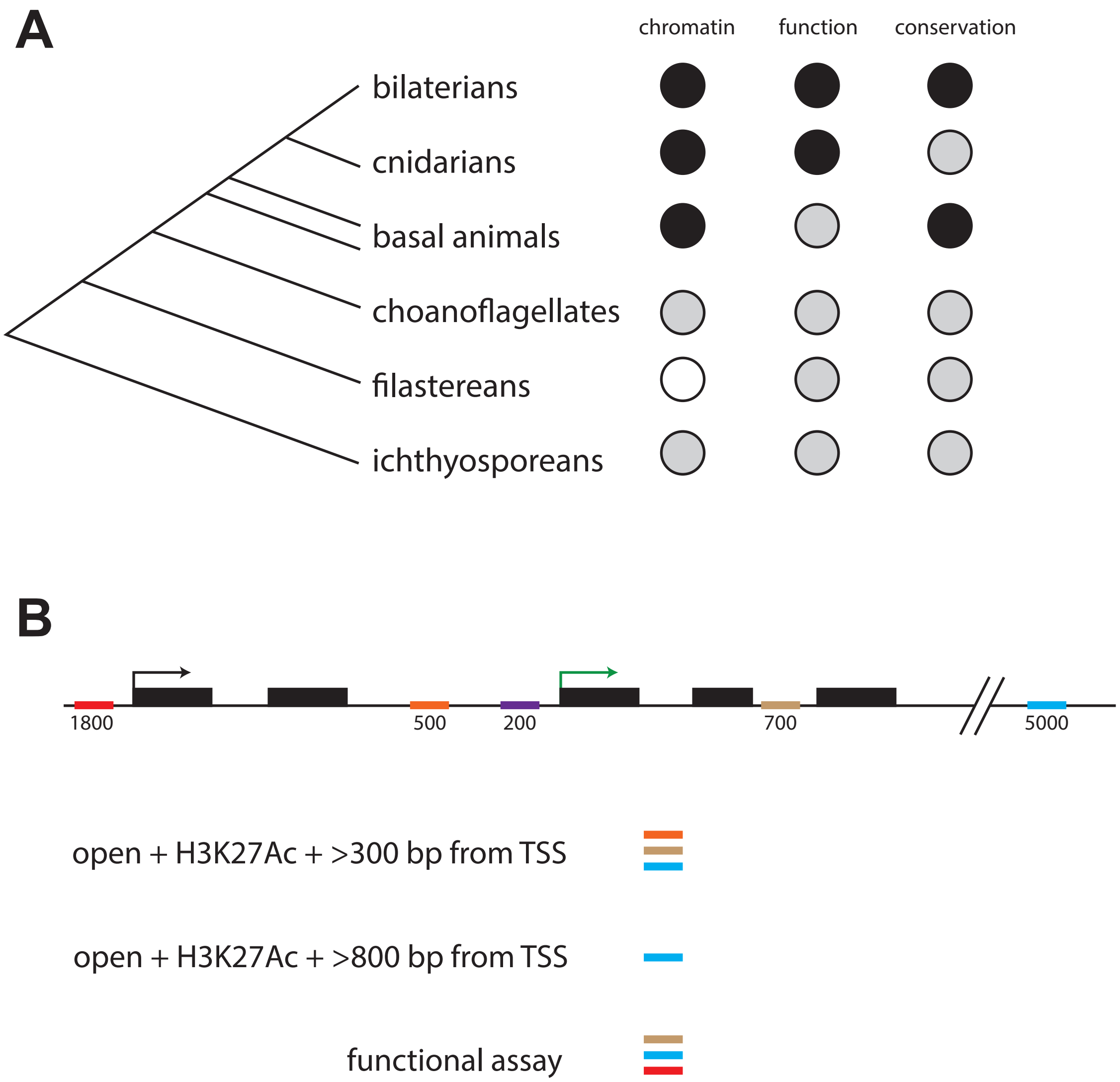

A
chromatin
function
conservation
bilaterians
cnidarians
basal animals
choanoflagellates
filastereans
ichthyosporeans
B
1800
500
200
700
5000
open + H3K27Ac + >300 bp from TSS
open + H3K27Ac + >800 bp from TSS
functional assay

Figure 1.5. The pre-animal roots of transcriptional networks and cellular differentiation.

(A) Transcriptional regulatory modules with pre-animal roots. Bioinformatic (Brown, Cole, and Erives 2008) and biochemical (Young et al. 2011) evidence in choanoflagellates supports an ancient role for Myc:Max in regulating ribosome biogenesis genes (van Riggelen, Yetil, and Felsher 2010). In the cellular cartoon, ribosomes are shown in blue. Dashed arrows indicate animal-specific functions of Myc:Max (Eilers and Eisenman 2008). Bioinformatic and functional evidence shows that choanoflagellate RFX regulates ciliogenesis (Coyle et al. 2023), arguing for the pre-animal ancestry of this regulatory network. However, RFX homodimers and physical binding to FoxJ1 may represent animal modifications (Quigley and Kintner 2017; Coyle et al. 2023).

(B) A gallery of cell types possibly present in the protistan ancestor of animals. Modern-day choanoflagellates have filter feeders, gametes, cysts, clonal colonies, attached cells, and amoebas (Dayel et al. 2011; Levin and King 2013; Woznica et al. 2017; Brunet et al. 2021; Leadbeater 2015). Modern-day filastereans have aggregates, cysts, attached cells, and amoebas (Sebé-Pedrós, Irimia, et al. 2013; Tikhonenkov et al. 2020). Modern-day ichthyosporeans have coenocytes and amoebas (Suga and Ruiz-Trillo 2013).

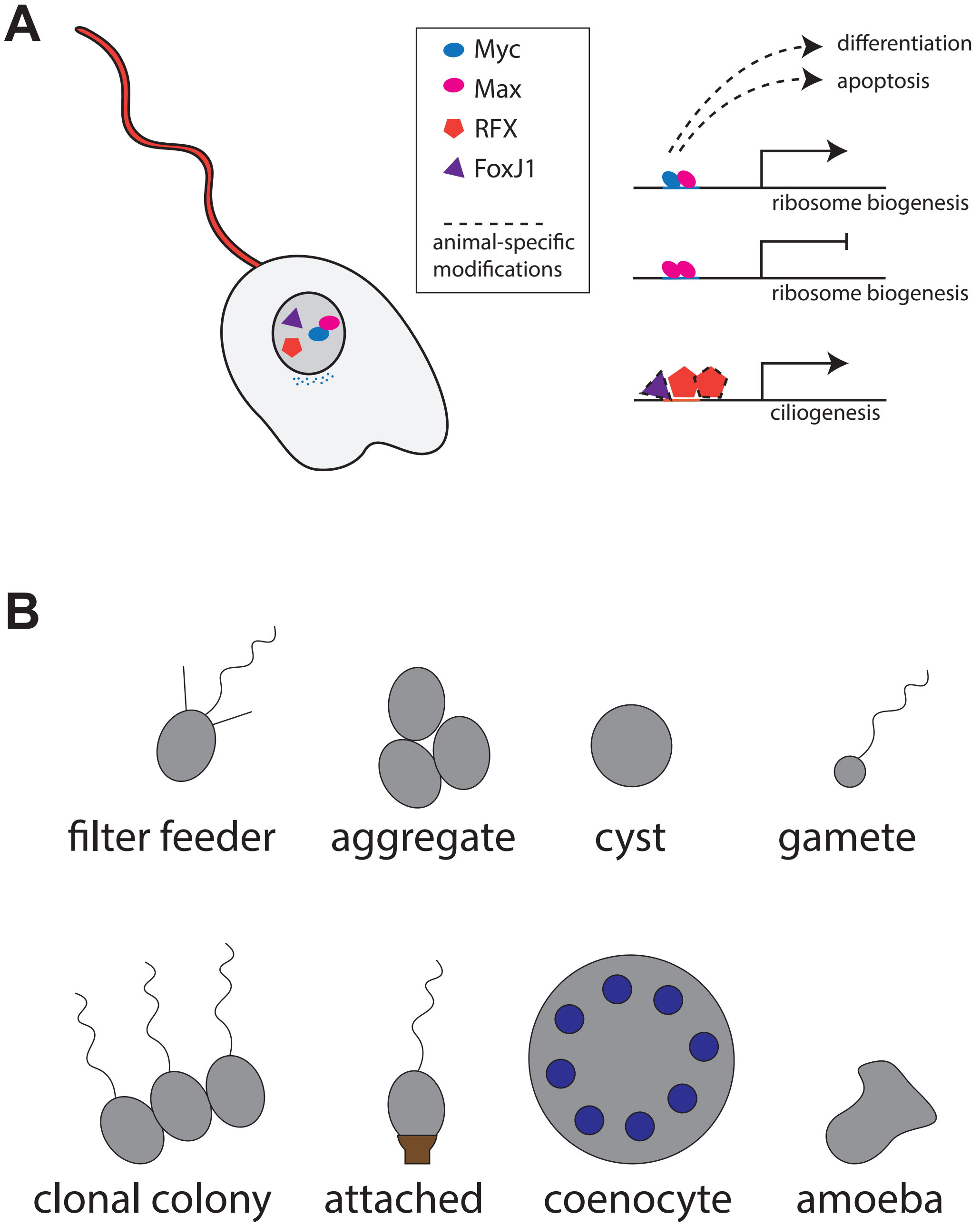

A
Myc
Max
RFX
FoxJ1
animal-specific modifications
differentiation
apoptosis
ribosome biogenesis
ribosome biogenesis
ciliogenesis
B
filter feeder
aggregate
cyst
gamete
clonal colony
attached
coenocyte
amoeba

Figure 1.6. Genetic tools aid the identification of protein-DNA and protein-protein interactions.

The determination of TF DNA binding preferences and protein interaction partners in systems with and without genetic tools available. Without genetic tools, the possibility of using antibody-based approaches depends on the ability to find or generate a quality antibody, which may be impossible, particularly in emerging or non-model systems. Other approaches, such as protein binding microarrays (PBM) or systematic evolution of ligands by exponential enrichment (SELEX) may be used to identify TF-DNA interactions (Jolma and Taipale 2011), while yeast-2-hybrid (Y2H) can be used to identify protein-protein interactions. When genetic tools are available, the generation of an epitope-tagged TF of interest permits experimental possibilities compatible with well-validated epitope antibodies.

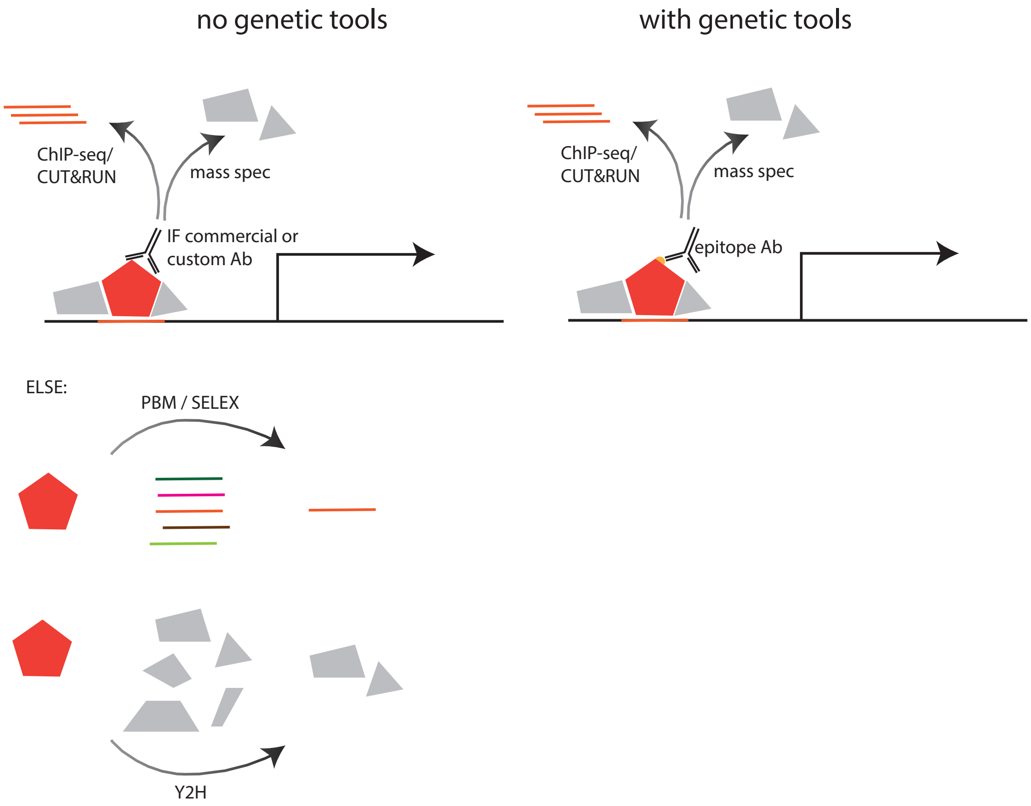

no genetic tools
with genetic tools
ChIP-seq/
CUT&RUN
mass spec
IF commercial or
custom Ab
ChIP-seq/
CUT&RUN
mass spec
epitope Ab
ELSE:
PBM / SELEX
Y2H

Table 1.1. The timing of new TF families, new TF sub-families, and TF family expansions on the animal stem lineage.

Data collated from broad surveys of eukaryotic TF distribution (de Mendoza et al. 2013; de Mendoza and Sebé-Pedrós 2019; Weirauch and Hughes 2011) as well as more focused analyses on the evolution of particular TF families and sub-families (Young et al. 2011; Larroux et al. 2008; Coyle et al. 2023). Choanozoa includes animals and choanoflagellates, while Holozoa includes Choanozoa, filastereans, ichthyosporeans, and corallochytreans. Opisthokonta includes Holozoa and Holomycota, which consists of fungi and their close relatives. Amorphea includes Opisthokonta and Amoebozoa (Adl et al. 2012).

TABLE 1.1	NOVEL TF FAMILIES	NOVEL TF SUB-FAMILIES	TF FAMILY EXPANSIONS	REFERENCES
EUKARYOTES	C2H2 zinc finger, Homeodomain, bZIP, bHLH, Myb, GATA, HMG-box, ARID, E2F, HSF, MADS			(de Mendoza and Sebé-Pedrós 2019; Weirauch and Hughes 2011)
AMORPHEA	RFX, CSL, STAT			(de Mendoza and Sebé-Pedrós 2019)
OPISTHOKONTA	T-box, p53, NFκB, Forkhead*		bHLH	(de Mendoza et al. 2013; de Mendoza and Sebé-Pedrós 2019; Larroux et al. 2008; Coyle et al. 2023)
HOLOZOA	Runx	Myc/Max (bHLH), Maf (bZIP)		(Young et al. 2011; de Mendoza and Sebé-Pedrós 2019; de Mendoza et al. 2013)
CHOANOZOA		RFXa,b,c (RFX); FoxJ, FoxN (Forkhead)		(Larroux et al. 2008; Coyle et al. 2023)
ANIMAL	Ets, MADF, Nuclear hormone receptor	bHLH A,C-F types (bHLH); POU, ANTP, LIM-HD, NK, Six (homeodomain); Sox (HMG-box)	C2H2 zinc finger, Homeodomain, Forkhead, T-box	(Weirauch and Hughes 2011; de Mendoza et al. 2013; Degnan et al. 2009)

* The evolutionary explanation behind the phylogenetic distribution of the Forkhead domain remains unclear. See (Coyle et al. 2023) for a discussion of how to interpret the sparse distribution of Forkhead domains outside of opisthokonts.

Chapter 2

An RFX transcription factor regulates ciliogenesis in the closest living relatives of animals

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

Coyle, M. C. *et al.* An RFX transcription factor regulates ciliogenesis in the closest living relatives of animals. *Current Biology.* (2023) doi:10.1016/j.cub.2023.07.022

Files S1-S8 and Videos S1-S4 are available in the published online version.

Summary

Cilia allowed our protistan ancestors to sense and explore their environment, avoid predation, and capture bacterial prey (Fritz-Laylin 2020; Nielsen 2008; Bloodgood 2010). Regulated ciliogenesis was likely critical for early animal evolution (Margulis 1992; Buss 1988; Nielsen 2008; Brunet and King 2017) and, in modern animals, deploying cilia in the right cells at the right time is crucial for development and physiology. Two transcription factors, RFX and FoxJ1, coordinate ciliogenesis in animals (Choksi et al. 2014; Chung et al. 2012; Yu et al. 2008) but are absent from the genomes of many other ciliated eukaryotes, raising the question of how the regulation of ciliogenesis in animals evolved (Piasecki, Burghoorn, and Swoboda 2010; Chu, Baillie, and Chen 2010). By comparing the genomes of animals with those of their closest living relatives, the choanoflagellates, we found that the genome of their last common ancestor encoded at least three RFX paralogs and a FoxJ1 homolog. Disruption of the RFX homolog *cRFXa* in the model choanoflagellate *Salpingoeca rosetta* resulted in delayed cell proliferation and aberrant ciliogenesis, marked by the collapse and resorption of nascent cilia. In *cRFXa* mutants, ciliogenesis genes and *foxJ1* were significantly down-regulated. Moreover, the promoters of *S. rosetta* ciliary genes are enriched for DNA motifs matching those bound by the cRFXa protein *in vitro*. These findings suggest that an ancestral *cRFXa* homolog coordinated ciliogenesis in the progenitors of animals and choanoflagellates and that the selective deployment of the RFX regulatory module may have been necessary to differentiate ciliated from non-ciliated cell types during early animal evolution.

Results and Discussion

Choanoflagellates express orthologs of animal cilia-associated transcription factors

Key features of the progenitors of animals can be inferred by comparing animals with their closest living relatives, the choanoflagellates (King 2004; Leadbeater 2015; Brunet and King 2017). Choanoflagellate cells feature a distinctive “collar complex” composed of a single apical cilium surrounded by a collar of actin-filled microvilli (King 2004; Leadbeater 2015). Structural conservation of cilia across eukaryotic diversity suggests that the last common ancestor of eukaryotes had a cilium (Fritz-Laylin 2020; Carvalho-Santos et al. 2011) and that the cilia of choanoflagellates and animals are homologous (Pinskey et al. 2022).

RFX and FoxJ1 are two transcription factors (TFs) that regulate animal ciliogenesis. Loss of either RFX or FoxJ1 function in animals reduces the transcription of many ciliary genes (Efimenko et al. 2005; Quigley and Kintner 2017; Lemeille et al. 2020) and results in ciliogenesis defects (Chung et al. 2012; Swoboda, Adler, and Thomas 2000; Yu et al. 2008; Bonnafe et al. 2004; Dubruille et al. 2002; Jianchun Chen, Heather J. Knowles, Jennifer L. Herbert, and Brian P. Hackett 1998; Stubbs et al. 2008). Despite their essentiality for proper ciliogenesis in animals, RFX and FoxJ1 are either missing (e.g., in *Chlamydomonas, Naegleria*, and ciliates), of unknown function (e.g., in choanoflagellates and chytrids), or of non-ciliary function (e.g., in ascomycete fungi (Wu and McLeod 1995; Bugeja, Hynes, and Andrianopoulos 2010; Huang, Zhou, and Elledge 1998; Hao et al. 2009) in non-animals (Figure 2.1A). To better understand the phylogenetic distribution of *RFX* and *foxJ1* genes, we used DNA-binding domain (DBD) sequences from diverse FoxJ1 and RFX predicted protein sequences to query EukProt (Richter et al. 2022) (Figure 2.1A; Materials and Methods; Files S1, S2). Confirming previous reports (Nakagawa et al. 2013; Brunet and King 2017), we found an ortholog of animal *foxJ1* genes in *S. rosetta*. Choanoflagellate RFX genes fall into three paralogous sub-families, provisionally named *cRFXa*, *cRFXb*, and *cRFXc* (Figure 2.1B; Figure 2.2A). *cRFXa* homologs were detected in nearly all choanoflagellate species analyzed, while *cRFXb* and *cRFXc* homologs have more restricted phylogenetic distributions (Figure 2.1B).

The life history of *S. rosetta* includes transitions between diverse ciliated cell types – including slow swimmers, fast swimmers, thecate cells, and multicellular rosettes (Dayel et al. 2011). We found that *cRFXa* was transcribed in each life history stage, while *cRFXb* and *cRFXc* expression was restricted to thecate cells (Figure 2.1C; File S3). *foxJ1* was down-regulated in thecate cells and up-regulated in fast swimmers, a starvation-induced cell type with longer cilia and a faster swimming velocity (H. Nguyen et al. 2019) (Figure 2.1C; File S3).

Phylogenetic analysis of RFX protein sequences from diverse opisthokonts and amoebozoans recovered the three choanoflagellate sub-families (cRFXa, cRFXb, and cRFXc), three RFX sub-families previously reported in animals (RFX1/2/3, RFX4/6/8 and RFX5/7) (Chu, Baillie, and Chen 2010), and distinct clades of amoebozoan and fungal RFX proteins (Figure 2.1D; Figure 2.2B). The cRFXa sub-family branched with the animal RFX1/2/3 sub-family, which regulates ciliogenesis in many tissues across diverse animals (Choksi et al. 2014; Chung et al. 2012;

Dubruille et al. 2002; Swoboda, Adler, and Thomas 2000) (Figure 2.1D; Figure 2.2B, 2.2C). The cRFXb and cRFXc sub-families grouped with the animal RFX5/7 and RFX4/6/8 sub-families, respectively, both of which serve diverse functions in animals and regulate ciliogenesis only in specific contexts (Sedykh et al. 2018; Ashique et al. 2009; Castro et al. 2018; Manojlovic et al. 2014). We thus infer that the last common ancestor of choanoflagellates and animals encoded at least three *RFX* paralogs, one related to modern-day *RFX1/2/3/cRFXa* genes, one related to *RFX5/7/cRFXb* genes, and one related to *RFX4/6/8/cRFXc* genes (Figure 2.1D).

Disruption of S. rosetta cRFXa *delays cell proliferation and ciliogenesis*

To investigate the function of the *cRFXa*, *cRFXb*, *cRFXc*, and *foxJ1* genes in *S. rosetta*, we used CRISPR-mediated gene editing (Booth and King 2020) to introduce an early stop codon near the 5’ ends of each gene (Figure 2.3A; Figure 2.4A; File S4). The resulting strains were cultured under conditions that favor the proliferation of slow swimmers, the cell type used for all experiments here (Materials and Methods). Mutants for *foxJ1*, *cRFXb*, or *cRFXc* showed normal growth and displayed no obvious phenotypic defects (Figure 2.4B, C). In contrast, two independently isolated *cRFXa* mutant lines, each encoding a truncated allele of *cRFXa* (*cRFXa^{PTS-1}* and *cRFXa^{PTS-2}*), proliferated more slowly than a wild-type control (*cRFXa^{WT}*; Figure 2.3B). A strain in which the *cRFXa^{PTS-1}* allele was reverted to the wild-type amino acid sequence (*cRFX^{REV}*) had comparable growth to that of *cRFXa^{WT}* cells, confirming that the growth defect in the *cRFXa^{PTS-1}* strain was a direct result of the cRFXa truncation (Figure 2.3B).

Cilia lengths were indistinguishable between *cRFXa^{WT}* and *cRFXa^{PTS-1}* cells (Figure 2C), but this did not reveal the dynamics of ciliogenesis itself. Therefore, we performed live imaging of ciliary regeneration (Figure 2.3D; Materials and Methods). In *cRFXa^{WT}* cells, the nascent cilium emerged rapidly and proceeded to lengthen (Figure 2.3E; Video S1, S2). In comparison, the nascent cilia of *cRFXa^{PTS-1}* mutant cells collapsed and were resorbed into the cell frequently (6.24 ciliary collapse events/cell/60 minutes compared to 1.00 for *cRFXa^{WT}* cells; p-value = 0.0012, unpaired t-test; Figures 2.3F, G; Videos S3, S4).

To quantify the rate of ciliogenesis, we established a metric by which cells were scored as having a regenerated cilium once the apical tip of the cilium grew past the apical boundary of the microvillar collar (Figure 2.3D). Within 60 minutes after ciliary removal, only 55% of *cRFXa^{PTS-1}* mutant cells and 50% of *cRFXa^{PTS-2}* cells had successfully regenerated their cilium, whereas 90% of *cRFXa^{WT}* cells and 97% of *cRFXa^{REV}* cells completed ciliary regeneration (Figure 2.3H; Figure 2.4D). In contrast, the *cRFXb^{PTS}*, *cRFXc^{PTS}*, and *foxJ1^{PTS}* mutants did not display any detectable ciliogenesis defect (Figure 2.4E, F, G). Moreover, a *cRFXa^{PTS-1};foxJ1^{PTS}* double mutant, generated by CRISPR editing of *foxJ1* in the *cRFXa^{PTS-1}* background, showed no additional defect in ciliary regeneration beyond that observed in *cRFXa^{PTS-1}* cells (Figure 2.4H). In summary, *cRFXa* is required for proper cilia regeneration in *S. rosetta* slow swimmers, while *cRFXb*, *cRFXc, and foxJ1* are not.

cRFXa *promotes transcription of conserved ciliogenesis genes and foxJ1*

To investigate how disruption of *cRFXa* in *S. rosetta* leads to aberrant ciliogenesis, we next investigated the transcriptional profiles of *cRFXa^{WT}* and *cRFXa^{PTS-1}* cells. In animals, the ciliary phenotypes of RFX loss-of-function mutants are associated with reduced expression of many ciliary genes (Kistler et al. 2015; Lemeille et al. 2020; Chung et al. 2014; Quigley and Kintner 2017) and we hypothesized that the same might be true for choanoflagellates. To identify candidate ciliary genes in *S. rosetta*, we curated the "HsaSro conserved ciliome," a list of 201 genes that (1) are required for proper assembly of cilia in humans, (2) have a well-characterized molecular function, and (3) are conserved between humans and *S. rosetta* (Materials and Methods; File S6). The HsaSro conserved ciliome includes axonemal dyneins, genes involved in intraflagellar transport (IFT), radial spokes, the BBSome, tubulin modifiers, the ciliary transition zone, ciliary vesicle formation, and more (Figure 2.5A).

Of the 201 genes in the HsaSro conserved ciliome, 93 were significantly down-regulated in *cRFXa^{PTS-1}* cells compared to *cRFXa^{WT}* cells (edgeR FDR < 0.001; Figure 2.5B; Files S5, S6). The down-regulated ciliary genes had slightly more than a 2-fold reduction in expression (Figure 2.5B, C), while genes not in the HsaSro conserved ciliome had, on average, no change in expression (Figure 2.5B). Among the most down-regulated ciliary genes in *cRFXa^{PTS-1}* cells were the ciliary GTPase *arl13B* (Larkins et al. 2011), the ciliary tip component *cep104* (Frikstad et al. 2019), and the tubulin glutamylation enzyme *ttll6* (Pathak, Austin, and Drummond 2011) (Figure 3B). Moreover, genes previously detected in the *S. rosetta* ciliome by mass spectrometry (Sigg et al. 2017) were preferentially down-regulated in *cRFXa^{PTS-1}* cells (Figure 2.6). Manual annotation of the most down-regulated genes in the *cRFXa^{PTS-1}* mutant uncovered a preponderance of genes of putative ciliary function (Figure 2.5D; File S6). These data indicate that cRFXa exerts widespread influence on ciliary gene transcription.

Previous work has shown that animal RFX and FoxJ1 cross-regulate each other's expression (Didon et al. 2013; Yu et al. 2008). For example, in mouse ependymal cells, RFX3 is required for full *foxJ1* expression (El Zein et al. 2009), while mouse *foxJ1$^{-/-}$* embryos fail to transcribe *rfx3* (Alten et al. 2012). Intriguingly, the most differentially expressed gene in the *S. rosetta cRFXa^{PTS-1}* mutant was *foxJ1*, which was 29-fold down-regulated (Figure 3B). This raised the question of whether cRFXa regulates ciliary genes partially through the action of FoxJ1. We found that no single HsaSro conserved ciliary gene was significantly down-regulated in *foxJ1^{PTS}* cells (Figure 2.5B; Files S5, S6). In fact, the only gene significantly differentially expressed in *foxJ1^{PTS}* was *trpm3,* which was up-regulated 30-fold in *foxJ1^{PTS}* cells (Figure 2.5B). Together with the observation that ciliogenesis proceeds normally in *foxJ1^{PTS}* cells, these data suggest that under standard growth conditions, *foxJ1* is a downstream target of *cRFXa*, but itself has no detectable effect on ciliary gene expression.

Finally, in contrast with the cell cycle regulatory function of RFX in some fungi (Bugeja, Hynes, and Andrianopoulos 2010; Wu and McLeod 1995), none of the strongly down-regulated genes in *cRFXa^{PTS-1}* mutants had clear connections to cell cycle regulation.

Predicted RFX binding sites are enriched in promoters of choanoflagellate ciliary genes

The DNA-contacting residues of RFX DBDs are largely invariant (Gajiwala et al. 2000; Chu, Baillie, and Chen 2010; Piasecki, Burghoorn, and Swoboda 2010) (Figure 2.8A), and the RFX monomeric recognition sequence – GTTRCY – is conserved across fungi and animals (Reith et al. 1990, 1994; Badis et al. 2008; Jolma et al. 2013) (Figure 2.7A). RFX binding sites often occur as tandem inverted repeats, forming a palindromic sequence referred to as an "X-box" (GTNRCC N_{0-3} RGYAAC; Figure 2.8B) (Reith et al. 1994; Emery et al. 1996; Gajiwala et al. 2000; Efimenko et al. 2005), which is bound by a dimer of RFX TFs (Reith et al. 1990; Gajiwala et al. 2000). To examine whether RFX might directly regulate ciliary genes in *S. rosetta*, we investigated motif enrichment in the promoters of *S. rosetta* ciliary genes and the DNA binding preferences of cRFXa.

Using the HOMER algorithm (Heinz et al. 2010), we detected a single motif in *S. rosetta* that was significantly enriched in the promoters of HsaSro conserved ciliome genes (Figure 2.7B). The motif closely resembles monomeric RFX-bound sequences from humans (Figure 2.7A) and was detected in 21.9% of promoters from conserved ciliome genes (44 total) as opposed to just 2.0% of all promoters (239 total; Figure 2.7C). The detected enrichment of the RFX motif in HsaSro conserved ciliome promoters was robust to variable definitions of promoter length (Figure 2.8C). Out of the 44 HsaSro conserved ciliome genes with RFX motifs, 33 (75%) were significantly down-regulated in *cRFXa*$^{PTS\text{-}1}$ cells (File S6). In *M. brevicollis,* the HOMER algorithm also detected an RFX-like motif as the most enriched motif among HsaMbrev conserved ciliome promoters (Figure 2.7B, C; File S6). In contrast, analysis of conserved ciliome promoters in *Spizellomyces punctatus,* a ciliated chytrid fungus that expresses RFX (Medina and Buchler 2020), did not identify any significantly enriched motifs, RFX or otherwise.

Because the predicted choanoflagellate ciliome motifs matched functionally validated RFX motifs from animals and fungi, we sought to investigate whether cRFXa shares this binding preference. To this end, we used an *in vitro* protein-binding microarray (PBM) (Lam et al. 2011; Weirauch et al. 2013, 2014) in which full-length cRFXa from *S. rosetta* was screened against multiple panels of short DNA oligonucleotides. The consensus motif recovered (Figure 2.7D) showed clear similarity to both the enriched choanoflagellate ciliome motifs and the binding sites of animal RFX monomers, including those derived from PBM approaches (Reith et al. 1990, 1994; Weirauch et al. 2014) (Figure 2.7A). No similarity to animal FoxJ1 PBM motifs was detected (Figure 2.8D).

In animals, RFX binding motifs are enriched near transcription start sites (Piasecki, Burghoorn, and Swoboda 2010; Sugiaman-Trapman et al. 2018). We found the same to be true in choanoflagellates, with 60.4% of RFX-like motifs located within 50 bp of the transcription start sites (TSS) of HsaSro conserved ciliary genes (Figure 2.7E; Figure 2.8E; File S7). Because we do not know whether choanoflagellates engage in distal regulation of gene transcription, we do not know whether RFX binding motifs detected further from the TSS may still be functional. Interestingly, the *foxJ1* promoter proximal region does not have an RFX binding site meeting our strict criteria, but does have a closely matched sequence (GTTGCGA, compared to the RFX GTTGCCA consensus) 701 base pairs upstream of its transcription start site.

If predicted RFX binding sites are essential for activating transcription of RFX-responsive ciliary genes, disruption of a predicted RFX binding site might be expected to reduce gene transcription. To test this, we focused on the *S. rosetta spag6* ciliary gene, which shows reduced expression in *cRFXa*$^{PTS\text{-}1}$ cells ($\log_2$FC = -1.50) and has a predicted RFX binding sequence (GTTGCCAA) in its promoter (Figure 2.7F). We built two reporter constructs: one with the *nanoluc* luciferase gene fused downstream of the wild-type *spag6* promoter ($P_{spag6\text{-}wt}$) and a second construct with key nucleotides in the RFX-binding motif mutated from GTTG to ACTG ($P_{spag6\text{-}\Delta TFBS}$). These constructs were transfected into wild-type and *cRFXa*$^{PTS\text{-}1}$ cells. Compared to *cRFXa*^{WT} cells transfected with the $P_{spag6\text{-}wt}$ reporter, *cRFXa*$^{PTS\text{-}1}$ cells transfected with the $P_{spag6\text{-}wt}$ reporter showed reduced NanoLuc activity (36%; Figure 2.7G), further implicating cRFXa in the regulation of *spag6*. Furthermore, *cRFXa*^{WT} cells transfected with the $P_{spag6\text{-}\Delta TFBS}$ reporter showed reduced NanoLuc activity compared to *cRFXa*^{WT} cells transfected with the $P_{spag6\text{-}wt}$ reporter (61%; Figure 2.7G). These results are consistent with the RFX consensus motif being required to mediate full transcription of *spag6*, which can be affected by either mutating the RFX motif or mutating the *cRFXa* gene (Figure 2.7G).

The pre-animal ancestry of the RFX ciliogenesis regulatory module

It has previously been unclear whether RFX or FoxJ1 transcription factors regulate ciliogenesis in any non-animal (Piasecki, Burghoorn, and Swoboda 2010; Brunet and King 2017; Chu, Baillie, and Chen 2010). One prior study looked for X-box sequences in the promoters of 12 ciliary genes in *M. brevicollis* and suggested that RFX gained control of ciliary genes in animals only after their divergence from choanoflagellates (Piasecki, Burghoorn, and Swoboda 2010), a conclusion we here revisit in light of increased genomic data and the establishment of transgenics in *S. rosetta* (Booth, Szmidt-Middleton, and King 2018; Booth and King 2020; Richter et al. 2018, 2022).

We have uncovered four lines of evidence indicating that cRFXa regulates ciliogenesis in *S. rosetta*: (1) targeted disruption of *cRFXa* results in aberrant ciliogenesis; (2) *cRFXa* mutants show significant down-regulation of 93 ciliary genes that are conserved between *S. rosetta* and humans; (3) an unbiased *in silico* approach identified an RFX motif enriched in ciliary gene promoters; (4) an RFX motif is necessary for wild-type levels of gene expression from a ciliary gene promoter.

Disruption of *cRFXa* also results in delayed cell proliferation, which is interesting because RFX homologs regulate the cell cycle in fungi (Hao et al. 2009; Wu and McLeod 1995). While we did not observe known cell cycle regulators among the most differentially expressed genes in the *cRFXa*$^{PTS\text{-}1}$ mutant strain, these experiments were not done in synchronized cells, which would allow more sensitive detection of differences in oscillatory gene expression. The defect in cell proliferation may also be due to the ciliogenesis defect, as ciliary function is essential for bacterial prey capture in *S. rosetta* (Dayel and King 2014). A defect in prey capture can be seen in our ciliogenesis assay, in which bacteria do not accumulate on the collar until the cilium is fully grown and begins to beat (e.g., time stamp 21:00 in Video S1 and time stamp 47:00 in Video S2 for wild-type cells). In *cRFXa*^{PTS} cells that do not assemble cilia in the ciliogenesis assay, bacteria never accumulate on the collar (Videos S3, S4). Therefore, post-mitotic *cRFXa*^{PTS} mutant cells may experience nutrient limitation as a secondary consequence of aberrant and delayed ciliogenesis.

Intriguingly, *cRFXa*$^{PTS-1}$ cells have steady state ciliary lengths comparable to that of *cRFXa*^{WT} cells. This fact, combined with the down-regulation but not total loss of ciliary gene expression (Figure 2.7B), suggests the presence of other transcriptional regulators of ciliogenesis. These are likely to be factors other than *cRFXb* and *cRFXc*, which were not appreciably transcribed in either *cRFXa*^{WT} or *cRFXa*$^{PTS-1}$ slow swimmer cells (Figure 2.1C; File S5).

The comparable roles of *S. rosetta cRFXa* and animal *RFX1/2/3* paralogs in regulating ciliogenesis (Choksi et al. 2014; Swoboda, Adler, and Thomas 2000; Dubruille et al. 2002; Chung et al. 2012), coupled with the predicted orthology between these two gene sub-families (Figure 2.1D; also see (Chu, Baillie, and Chen 2010)), suggests that the last common ancestor of animals and choanoflagellates expressed an RFX transcription factor that regulated ciliogenesis. Might the RFX regulatory module be more ancient than the choanoflagellate-animal clade (Choanozoa)? Functional data on ciliated opisthokonts outside the Choanozoa are missing, but our bioinformatic analysis of ciliome promoters in the chytrid *S. punctatus* did not suggest RFX involvement. RFX may have been co-opted to regulate ciliary genes in the Choanozoan stem lineage, perhaps potentiated by RFX family expansion. Alternatively, RFX might have regulated ciliogenesis in stem opisthokonts, but was then recruited for other functions in fungi, including in chytrids. In either scenario, the divergence of RFX functions between choanozoans and fungi required many changes in the *cis*-regulatory sequences of ciliary genes.

The RFX ciliogenesis regulatory module in the evolution of animal development

One question raised by this work is how the RFX-ciliogenesis regulatory module, likely already present in the protozoan progenitors of choanoflagellates and animals, was integrated into animal developmental programs. Was RFX activity sufficient for specifying ciliated cells, or did it require accessory regulators? If the founders of the modern-day *cRFXa/RFX1/2/3* sub-family had non-ciliogenesis roles, how was pleiotropy resolved when utilizing this network in novel cell type contexts? Finally, the function of FoxJ1 appears to differ in animals and *S. rosetta*. In animals, FoxJ1 regulates many ciliogenesis genes (Yu et al. 2008) and shows cross-regulation with RFX. The cross-regulation of these families is also seen in *S. rosetta*, as *foxJ1* is one of the most down-regulated genes upon *cRFXa* disruption. However, disruption of *foxJ1* in *S. rosetta* had no detectable effect on ciliogenesis efficiency and negligible impact on the expression of HsaSro conserved ciliary genes in the slow swimmer cell type. This raises the question of whether FoxJ1 was a sub-module of RFX ancestrally and was later “promoted” to a higher level of the gene regulatory hierarchy or whether the role of FoxJ1 in *S. rosetta* reflects a diminished role from that of its ancestral counterpart.

Finally, our data may add something useful to a growing discussion on the origins of animal cell types. Proposed modes and drivers of cell type evolution include division of labor (Mackie 1970; Arendt 2008), integration of life cycles (Mikhailov et al. 2009; Zakhvatkin 1949), stress responses (Wagner, Erkenbrack, and Love 2019), and gene or genome duplication (Kin et al. 2022; Ohno 1970). A common theme in many of these models is the re-purposing of ancestral regulatory connections in novel cell types, in which a single transcription factor can coordinate the activity of a suite of genes sharing complementary functions. The work reported here provides a concrete example of a pre-animal regulatory module, the regulation of which evolved alongside animal development to help differentiate ciliated from non-ciliated cells.

Acknowledgements

We thank Thibaut Brunet, Josean Reyes-Rivera, Michael Carver, Alain Garcia de Las Bayonas, Jacob Steenwyk, Arnau Sebé-Pedrós, Lillian Fritz-Laylin, Mike Eisen, Fyodor Urnov, Iswar Hariharan, Elçin Ünal, Monika Sigg, Flora Rutaganira, and Erika López-Alfonzo for critical reading of the manuscript and valuable feedback, Lily Helfrich for standardizing cell type growth conditions, Alexandra Mulligan for assistance with molecular cloning, and Jacob Steenwyk for advice on phylogenetics. MCC was supported by an NIH Molecular Basis of Cell Function T32 Training Grant (3T32GM007232-40S1) and an NSF Graduate Research Fellowship Program Award (#1752814). An Investigator Award from the Howard Hughes Medical Institute supports research in the King Lab. Work in the Hughes Lab was funded by a CIHR grant to TRH (FDN-148403).

Materials and Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the main contact, Nicole King (nking@berkeley.edu)

Materials availability

Plasmids generated in this study have been deposited to Addgene (#196406, #196407, #196408).

Choanoflagellate cell lines used in this study are available from the American Type Culture Collection (PRA-390 for wild-type *Salpingoeca rosetta*) or available upon request for mutant cell lines.

Data and code availability

RNA sequencing data generated in this study have been deposited to the NCBI Short Read Archive (Project PRJNA91984).

This paper does not report original code. For the use of existing bioinformatic packages, the Method Details specify the options used and the Key Resources Table lists software version numbers.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Choanoflagellate culture

All experiments used *Salpingoeca rosetta* co-cultured with a single prey bacterial species, *Echinicola pacifica* (ATCC PRA-390, strain designation: SrEpac). Cells were grown in artificial known sea water (AKSW) supplemented with 4% cereal grass media (CGM3) and 4% sea water complete(Booth, Szmidt-Middleton, and King 2018). Cells were grown at 22°C and 60% humidity. For consistency, experiments were done with cells in the mid-log phase of growth, which in this media formulation occurs between 5 x 10^5 and 3 x 10^6 cells/ml.

Mutant strains generated by CRISPR/Cas9 genome editing were maintained under the same conditions as wild-type SrEpac, and liquid nitrogen stocks of all generated strains were created.

The following mutant lines were generated (see File S4 for editing information and Figure 2.4A for genotyping traces):

cRFXa$^{PTS\text{-}1}$
cRFXa$^{PTS\text{-}2}$
cRFXa^{REV}
cRFXb^{PTS}
cRFXc^{PTS}
foxJ1^{PTS}
cRFXa$^{PTS\text{-}1}$*;foxJ1*^{PTS}

Since the *cRFXb*^{PTS}, *cRFXc*^{PTS}, and *foxJ1*^{PTS} mutants were generated using a co-editing strategy that confers cycloheximide resistance, the reference wild-type strain for these was SrEpac bearing the P56Q mutation in *rpl36a* (Booth and King 2020).

METHOD DETAILS

BLAST searches for RFX and FoxJ1 genes

To determine the presence of RFX genes throughout eukaryotic diversity, we used a variety of functionally validated RFX DBDs as BLAST queries, searching against the EukProt database, which includes 993 species (Richter et al. 2022). First, to define the broad phylogenetic distribution of RFX genes, we queried the DBDs of *Xenopus laevis* RFX2 and *Saccharomyces cerevisiae* RFX1 against the EukProt Comparative Set of 196 species, chosen for taxonomic diversity and genome/transcriptome completeness. EukProt implements the BLASTP 2.13.0 algorithm. We defined bone fide RFX hits as those with at least 75% query coverage and at least 30% sequence identity (see File S1 for DBD probe sequences and EukProt BLAST results).

To develop a comprehensive set of amorphean RFX hits, we used six RFX DBD sequences (*X. laevis* RFX2, *S. cerevisiae* RFX1, *M. musculus* RFX4, *M. musculus* RFX5, *C. elegans* Daf-19, and *S. rosetta* cRFXa) as BLAST probes against a set of 95 amorphean taxa. RFX hits within these taxa were used for the data shown in Figure 2.1A and to construct the phylogenetic trees in Figure 2.1D, 2.2A, 2.2B, and 2.2C. All sequences used for phylogenetic tree construction are detailed in File S1. For *S. mediterranea*, which is of interest due to it having demonstrated FoxJ1 function in ciliogenesis(Vij et al. 2012), but is not hosted on EukProt, we used the BLASTP server hosted on https://planosphere.stowers.org/, which implements BLASTP 2.3.0.

We used a similar procedure to identify Fox genes, first within the EukProt Comparative Set using the DBDs from *X. laevis* FoxJ1 and *S. mediterranea* FoxJ1 as probes (see File S2 for probe sequences and BLAST results) and a 75% query coverage / 30% query identity threshold criteria. To identify candidate FoxJ1 orthologs for the taxa represented in Figure 1A , reciprocal best BLAST searches were performed, using FoxJ1 DBDs from *M. musculus, X. laevis*, *S. mediterranea*, and *S. rosetta*. For these BLAST searches, we used EukProt for all except two taxa (which are not hosted on EukProt): *S. mediterranea*, hosted at https://planosphere.stowers.org/, and *X. laevis*, for which we used the NCBI BLAST server with the Uniprot reference database. In Figure 2.1A we report taxa containing reciprocal best BLAST for either *X. laevis* or *S. mediterranea*, which are phylogenetically disparate (within animals) and both have functionally validated FoxJ1 genes with known roles in regulating motile ciliogenesis.

When surveying the distribution of RFX and Fox genes across eukaryotic diversity, our results largely confirmed that RFX genes are widespread among opisthokonts and amoebozoans, while Fox genes are widespread among opisthokonts. However, we did observe rare exceptions to this pattern. Among 539 taxa in EukProt that are not opisthokonts or amoebozoans, three had RFX hits: *Madagascaria erythrocladioides* (a rhodophyte alga), *Gloeochaete wittrockiana* (a glaucophyte alga), and *Siedleckia nematoides* (an alveolate) (File S1). Among 824 non-opisthokonts in EukProt, 14 had Fox hits (File S2). For both the few RFX and Fox hits, the taxa in which they were observed were distributed across eukaryotic diversity. The only obvious pattern was that four out of the eight heterolobosean taxa hosted on EukProt contained Fox hits. Given the rare and dispersed nature of RFX and Fox hits outside of the amoebozoans/opisthokonts and opisthokonts, respectively, we interpret these hits as being more likely due to some combination of horizontal gene transfer, convergent evolution, and possibly sequencing contamination, than due to the presence of RFX or Fox genes in the last common ancestor of eukaryotes.

Phylogenetic trees

To build maximum-likelihood trees for RFX family genes, we aligned the protein sequences with MAFFT (Katoh et al. 2002; Katoh and Standley 2013) (v. 7.312) using default options, trimmed with ClipKIT (Steenwyk et al. 2020) (v 1.3.0) using the default smart-gap trimming mode, and built trees with IQ-TREE (L.-T. Nguyen et al. 2015) (v. 2.2.0-beta COVID-edition) using ModelFinder (Kalyaanamoorthy et al. 2017) and 1000 Ultrafast Bootstraps (UF-boot) (Minh, Nguyen, and von Haeseler 2013) or 1000 iterations of SH-aLRT (Guindon et al. 2010). Trees were visualized with iTOL (Letunic and Bork 2021). To test the robustness of our phylogenetic inferences, alignments were also trimmed with trimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) (v1.4.rev22) using the gappyout setting and trees were inferred with RAxML (Stamatakis 2014) (8.2.11) using the "-f a", "-m PROTGAMMAAUTO", and "-N 100" flags to find the best model and perform 100 bootstraps. For IQ-TREE analyses, the best substitution model (as determined by ModelFinder) for the choanoflagellate RFX tree was Q.pfam+F+R5 and for the amorphean RFX tree was Q.pfam+F+R6. For the amorphean RFX tree trimmed with trimAl, the best substitution model was Q.yeast+F+R5.

The protein sequences used for phylogenetic reconstruction are shown in File S1. Note that we do not necessarily use all of the RFX genes within a given taxon, for the purposes of both clarity of presentation and the efficiency of computational bandwidth. This is especially true for vertebrates, with their abundance of RFX duplications within well-established sub-families (e.g. RFX1/2/3 genes), and for some ichthyosporeans (e.g. *C. fragrantissima*), which contain extra RFX genes with long branches that lack consistent placement in phylogenetic re-constructions. These are likely more recent lineage-restricted duplications with extensive divergence.

The only surveyed choanoflagellates without a detectable RFX homolog were uncultured species whose genomes have been sequenced using single-cell technologies (López-Escardó et al. 2019; Needham et al. 2019). These species show relatively lower genome completeness as measured by BUSCO (Richter et al. 2022; Manni et al. 2021). Therefore, the apparent absence of RFX from these species may well be artefactual.

RFX DNA-binding domain alignment

For the presentation of RFX DBD alignments in Figure 2.8A, selected RFX DBD sequences were aligned using MUSCLE (Edgar 2004) (v. 3.8.425) with a maximum of 8 iterations and all other options as default, implemented in Geneious. However, alignments of full RFX protein sequences were used for the phylogenetic analysis (see previous section on "Phylogenetic Trees" and data in File S1).

Choanoflagellate culturing

Unless otherwise specified, all experiments were performed using *Salpingoeca rosetta* co-cultured with a single prey bacterial species: *Echinicola pacifica* (ATCC PRA-390, strain designation: SrEpac). Cells were grown in artificial known sea water (AKSW) supplemented with 4% cereal grass media (CGM3) and 4% sea water complete (Booth, Szmidt-Middleton, and King 2018). Cells are grown at 22°C and 60% humidity. For consistency, experiments were done with cells in the mid-log phase of growth, which in this media formulation occurs between 5 x 10^5 and 3 x 10^6 cells/ml.

S. rosetta cell type RNA sequencing and analysis

Cultures were grown in triplicate for each of four *S. rosetta* cell types. Samples of slow swimmers and rosettes were prepared from cultures of 5% SWC media inoculated with 10^4 cells/ml of *S. rosetta* feeding on *Echinicola pacifica* bacteria, and rosettes were induced with the addition of outer membrane vesicles (OMVs) from *Algoriphagus machipongonensis* (Alegado et al. 2012) . Both of those cultures were grown for 48 h at 22°C to mid-log phase. Cultures of fast swimmers were inoculated the same as slow swimmers and then grown to starvation for 3 d at 22°C, at which point we transitioned the culture to 30°C for 2.75 h to increase the population of fast swimmers. Thecate cells were prepared by inoculating the HD1 strain of *S. rosetta* – a strain that maintains a higher proportion of thecate cells while also feeding on *E. pacifica* – to 10^4 cells/ml 10% (v/v) CGM3 and then growing for 48 h at 22°C in square plates.

For each replicate of each cell type, 5 x 10^6 cells were processed for lysis and RNA extraction. Cells were centrifuged and washed with AKSW. Thecate cells were scraped off the plate first. Cells were resuspended in AKSW, counted, and aliquoted to 10 x 10^6 per aliquot, then resuspended in 100 µl of lysis buffer (Booth, Szmidt-Middleton, and King 2018): 20 mM Tris-HCl, pH 8.0; 150 mM KCl; 5 mM $MgCl_2$; 250 mM sucrose; 1 mM DTT; 10 mM digitonin; 1 mg/mL sodium heparin; 1 mM Pefabloc SC; 100 µg/mL cycloheximide; 0.5 U/µl Turbo DNase; 1 U/µl

SUPERaseIN. This was incubated on ice for 10 minutes, passed ten times through a 30G needle and centrifuged at 6,000 x g for 10 minutes at 4°C. The supernatant was collected, brought to 100 µl with RNAse-free water, and RNA was purified using the RNAeasy kit from Qiagen (Cat. No. 74104)., eluting in 30 µl of water.

500 ng were of RNA were used for library prep, first purified with two rounds of polyA mRNA selection with oligo-dT magnetic beads and then converted to sequencing-compatible cDNA using the KAPA mRNA HyperPrep kit (KAPA biosystems, Cat. No. KK8580), using the KAPA single-indexed adapter kit for multiplexing (KAPA biosystems, Cat. No. KK8701). RNA integrity was assessed by Agilent Bioanalyzer 2100 before library prep using an Agilent RNA 6000 Nano Kit (Cat. No. 5067-1511). Sequencing libraries were also confirmed by Bioanalyzer 2100 for the correct size distribution using the Agilent High Sensitivity DNA Kit (Cat. No. 5067-4626). Library concentration was quantified by Qubit and libraries were pooled at equal concentrations before sequencing.

Library sequencing was performed by the QB3-Berkeley Genomics core labs (QB3 Genomics, UC Berkeley, Berkeley, CA, RRID:SCR_022170). Sequencing was performed in one lane on the Illumina HiSeq 4000, collecting between 12.4 million and 61.3 million reads for each sample. Reads were de-multiplexed, checked for quality with fastqc (v 0.11.9), and aligned to predicted transcripts from the *S. rosetta* genome (Fairclough et al. 2013) using Salmon (Patro et al. 2017) (v 1.5.2.) and called for differential expression using edgeR (M. D. Robinson, McCarthy, and Smyth 2010), both implemented within the Trinity software package (Grabherr et al. 2011) (v 2.14.0). TPM values for RFX gene expression amongst the different cell stages, as well as differential expression tests comparing slow swimmers with thecate cells, are available in File S3.

CRISPR guide RNA and repair template design

Candidate guide RNA sequences were obtained for each gene of interest using the EuPaGDT tool (http://grna.ctegd.uga.edu/) and the *S. rosetta* genome(Fairclough et al. 2013). Guide RNA length was set at 15 and an expanded PAM consensus sequence, HNNRRVGGH, was used. Coding sequences for genes of interest are easily obtained from the Ensembl Protists hosting of the *S. rosetta* genome. Guide RNA candidates were filtered for guides with one on-target hit (including making sure the guides do not span exon-exon boundaries), zero off-target hits (including against the genome of the co-cultured bacterium *E. pacifica*), lowest strength of the predicted secondary structure (assessed using the RNAfold web server: http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi), and annealing near the 5' end of the targeted gene, particularly before the region encoding the DNA-binding domain. crRNAs with the guide sequence of interest, as well as universal tracrRNAs, were ordered from IDT (Integrated DNA Technologies, Coralville, IA).

Repair templates were designed as single-stranded DNA oligos, in the same sense strand as the guide RNA, with 50 base pairs of genomic sequence on either side of the DSB cut site. Between the homology arms is the TTTATTTAATTAAATAAA insertion cassette. Repair oligos were ordered from IDT as Ultramers.

Genome editing

48 h prior to the transfection, *S. rosetta* cells (see File S4 for background genotype of each editing experiment) were inoculated in 120 ml of media at 8,000 cells/ml. This seeding density brings the culture to mid-log phase at the time of transfection. Prior to the day of transfection, lyophilized crRNA and tracrRNA from IDT were each resuspended in duplex buffer (30 mM HEPES-KOH pH 7.5; 100 mM potassium acetate, IDT Cat. No. 11-0103-01) to a concentration of 200 µM. Equal volumes of crRNA and tracrRNA were mixed, incubated for 5 minutes at 95°C in an aluminum heating block, and then cooled to 25°C slowly by removing the heat block from the heating source (with the tube still in it) and cooling to RT. The annealed crRNA/tracrRNA is referred to as the gRNA and can be stored at -20°C for weeks before use. Also prior to the day of transfection, the lyophilized repair oligo was resuspended to 250 µM in 10 mM HEPES-KOH, pH 7.5 and incubated at 55°C for 1 hour, then stored at -20°C.

On the day of transfection, to wash away bacteria from the choanoflagellates, the culture was split into three 50 ml conical tubes and centrifuged for 5 minutes at 2000 x g. The cell pellets were resuspended and combined in 50 ml of AKSW, followed by a 5 min spin at 2200 x g. The cells were washed once more with 50 ml AKSW and spun at 2400 x g. The pellet is resuspended in 100 µl AKSW and diluted 1:100 in AKSW for counting. Cells are diluted to 5 x 10^7 / mL in AKSW, then 100 µl aliquots (with 5 x 10^6 cells each) are prepared.

Priming buffer is prepared by diluting 10 µl of 1 mM papain (Sigma-Aldrich Cat. No. P3125-100MG) in 90 µl of dilution buffer (50 mM HEPES-KOH, pH 7.5, 200 mM NaCl, 20% glycerol, 10 mM cysteine, filter-sterilized and stored in aliquots at -80°C). This is then diluted 1:100 in the rest of the priming buffer (40 mM HEPES-KOH, pH 7.5, 34 mM lithium citrate, 50 mM L-cysteine, 15% PEG-8000, filter-sterilized and stored in aliquots at -80°C) for a final concentration of 1 µM papain. The priming buffer can be prepared while washing the cells.

Also while washing the cells, equal volumes of pre-annealed gRNA and *Sp*Cas9 (20 µM, NEB Cat. No. M0646M) are mixed and incubated for 1 h at RT to form the RNP. 4 µl of RNP is used per transfection reaction. The resuspended repair oligo is incubated for 1 hour at 55°C to completely solubilize the material.

Each aliquot of cells is spun at 800 x g for 5 minutes and resuspended in 100 µl priming buffer and incubated for 35 minutes at RT. The priming reaction is quenched by adding 10 µl of 50 mg/ml bovine serum albumin fraction V (Thermo Fisher Scientific Cat. No. BP1600-1000). Cells are spun at 1250 x g for 5 minutes and resuspended in 25 µl Lonza SF buffer (Lonza Cat. No. V4SC-2960) if cycloheximide selection was not used or 200 µl of SF buffer if cycloheximide selection was used.

For each transfection, 16 µl of Lonza SF buffer is mixed with 4 µl of RNP targeting the gene of interest, 2 µl of resuspended repair oligo, and 1 µl of washed and primed cells. If cycloheximide selection is being used, 1 µl of CHX-R RNP is added as well as 0.5 µl of CHX-R repair oligo. These engineer a P56Q mutation in *rpl36a* that confers resistance to cycloheximide(Booth and King 2020). The nucleofection reactions are added to a 96-well nucleofection plate (Lonza Cat. No. V4SC-2960) and pulsed with a CM156 pulse in the Lonza 4D-Nucleofector (Cat. No. AAF-1003B for the core unit and AAF-1003S for the 96-well unit).

After the pulse, 100 µl of ice-cold recovery buffer (10 mM HEPES-KOH, pH 7.5; 0.9 M sorbitol; 8% [wt/vol] PEG 8000) is immediately added to each well of the nucleofection plate and incubated for 5 minutes. Then the entire contents of the well are added to 1 mL of 1.5% SWC + 1.5% CGM3 in AKSW in a 12-well plate and cultured at 22C. After one hour of culture, 10 µl of re-suspended *E. pacifica* bacteria (10 mg/ml in 1 ml AKSW) are added to each culture not undergoing cycloheximide selection, and 50 µl are added for each culture that is undergoing cycloheximide selection.

The following day, 10 µl of 1 ug/ml cycloheximide is added to wells undergoing cycloheximide selection. Selection was done for 4 days.

Clonal dilutions were done 24 hours after transfection for cells not undergoing cycloheximide selection, and 5 days after transfection (with 4 days of selection) for cells undergoing cycloheximide selection. Cells were counted and diluted to 2 cells/ml in 1.5% SWC + 1.5% CGM3 in AKSW. To this was added a 1:1000 dilution of re-suspended *E. pacifica* (10 mg/ml in 1 ml AKSW). 200 µl of diluted culture was added per well for 96-well plates. For each editing experiment, between 5 and 20 96-well plates were prepared.

To genotype, 96-well plates were screened by microscopy and wells containing choanoflagellates were marked. These were re-arrayed into fresh 96-well plates with each well containing a separate clone. To extract genomic DNA, 50 µl of cell culture was mixed with 50 µl of DNAzol direct (Molecular Research Center, Inc [MRC, Inc.], Cincinnati, OH; Cat. No. DN131), incubated at RT for 10 minutes and stored at -20°C. Genotyping PCRs were performed in 96-well plates (Brooks Life Sciences Cat. No. 4ti-0770/c) using Q5 polymerase (NEB Cat. No. M0491L), and 40 cycles of amplification. 5 µl of genomic DNA template were used in a 50 µl PCR reaction. PCR products were purified by magnetic bead clean-up and were analyzed by Sanger sequencing (UC Berkeley DNA Sequencing Facility).

Measuring ciliary lengths

To measure cilium length, cells grown to mid-log phase were fixed and stained using 1 part Lugol's solution (EMD Millipore Cat. No. 1.09261.1000) with 3 parts culture (usually 25 µl and 75 µl). 4 µl were loaded onto a slide, spread by placing a No. 1.5 coverslip on thee sample, and imaged coverslip slide down with a Zeiss Axio Observer.Z1/7 Widefield microscope with a Hamamatsu Orca-Flash 4.0 LT CMOS Digital Camera (Hamamatsu Photonics, Hamamatsu City, Japan) and 40×/NA 1.1 LD C-Apochromatic water immersion objective. Images were acquired with 10 ms exposure and 8.0 V of light intensity, using the PH3 phase contrast ring. Ciliary lengths were traced and measured in Fiji (Schindelin et al. 2012).

Genome editing for *cRFXa* revertant

To revert the *cRFXa*$^{PTS\text{-}1}$ strain to a wild-type amino acid sequence, we transfected Cas9 with guide RNAs that cut on either side of the PTS allele and included a repair template that introduces a GTC > GTG (Valine) synonymous mutation in the wild-type gene sequence, allowing us to distinguish revertants from wild-type cells by genotyping. We first transfected various single and dual gRNA combinations into the *cRFXa*$^{PTS\text{-}1}$ strain and assessed editing frequency by next-generation amplicon sequencing 24 hours post-transfection. To do this we extracted DNA as in the “Genome Editing” section, PCR amplified around the PTS insertion using primers TGTCATGTTCTTTGCTGGCG and GTCGAAGGCGTTGAAGTTGC, and submitted purified PCR products for Genewiz Amplicon-EZ services (Azenta Life Sciences, Chelmsford, MA). Editing efficiency was very low for all gRNAs tested, with a maximum of 0.04% for the combination listed in File S4. This may be due to using an NGG PAM instead of the stricter HNNRRVGGH PAM(Booth and King 2020), which had no consensus sites near the PTS insertion.

Despite the low efficiency, we reasoned that due to the growth defect of the *cRFXa*^{PTS} mutant, a revertant might out-compete non-reverted cells in a mixed population. To test this, we cultured the transfected cultures for 4 weeks, isolated clones, and genotyped the locus. All genotyped clones were had the reverted allele, showing the success of this competition strategy.

Ciliogenesis assay

For step-by-step protocol, see protocols.io:
dx.doi.org/10.17504/protocols.io.q26g7y9n3gwz/v2

To monitor ciliogenesis, cells were grown to mid-log phase, counted, and 6 x 10^6 cells were centrifuged in a 15 ml falcon tube for 10 minutes at 2000 x g. The cell pellet was resuspended in 1 ml of 90% AKSW / 10% glycerol, added to a FluoroDish (World Precision Instruments Cat. No. FD35-100) and incubated for 7 minutes at -20°C. This method of ciliary removal was inspired by a ciliary removal protocol from *Chlamydomonas* (Brokaw 1960). For *S. rosetta*, we observed that on average 85% of cells lost their cilium, with a range of 68%-98%. A second FluoroDish was treated with 10 seconds of corona discharge (Electro-Technic Products BD-20AC), then rinsed with 1 ml of 0.1 mg/ml poly-D-Lysine (Millipore Sigma Cat. No. P6407-5MG). The dish was rinsed 3x with water and air dried.

After incubation at -20°C, the cells were transferred to a 1.5 ml Eppendorf tube and spun for 10 minutes at 4200 x g. The cell pellet was resuspended in 25 µl AKSW and transferred to the lysine-coated FluoroDish. A 22 mm circular diameter #1.5 coverslip (Electron Microscopy Sciences Cat No. 72224-01) was gently laid on top. The dish was positioned on the microscope stage and after the cells were brought into focus, the dish was flooded with 1 mL of AKSW to dislodge the coverslip while leaving the cells stuck to the surface. Cells were imaged with a Zeiss

Axio Observer.Z1/7 Widefield microscope with a Hamamatsu Orca-Flash 4.0 LT CMOS Digital Camera (Hamamatsu Photonics, Hamamatsu City, Japan) and 100 × NA 1.40 Plan-Apochromatic oil immersion objective (Zeiss) using a differential interference contrast (DIC) filter. Images were acquired at 10 z-slices spanning 10 µm, with one stack acquired every 30 seconds for one hour. We used 12.2 V bulb intensity and a short exposure (5 ms) to best capture the position of the flagellum as it regrew.

Image analysis was done in Fiji, marking the time point at which ciliogenesis was complete. This was defined as the point at which the growing cilium crossed the outer edge of the microvillar collar. In cases where the microvillar collar was significantly shortened by the glycerol treatment, the collar was able to re-lengthen quickly, almost always faster than the pace of ciliary re-generation. The time point at which the cilium crossed the microvillar collar could be assessed by DIC microscopy, while exact ciliary lengths were hard to extrapolate from live cells, due to ciliary motion and the various angles at which cells were oriented relative to the imaging plane. Cells were excluded from analysis for the following reasons: if it was impossible to determine when or whether the cilium crossed the outer edge of the microvillar collar; if the cell still maintained a cilium at time 0 (occasionally a nub of a cilium had already started to regenerate by the time the cells were put on the microscope, so a pre-existing cilium was defined as a cilium greater than 2 µm in length); if the cell divided or fused with a nearby cell during the time-course; if a cell contained multiple cilia (due to fusion or incomplete cytokinesis); if the cell was obviously dead (this could be diagnosed by the cell having irreversibly lost its microvillar collar and not making any attempts to regenerate the cilium or collar).

Growth curves

Cells in mid-log phase were diluted to 5,000 / ml and supplemented with 10 µg/ml *E. pacifica* bacteria (diluted 1:1000 from a stock of 10 mg/ml in AKSW). 500 µl of culture was aliquoted into each well of a 24-well plate (Fisher Scientific Cat. No. 09-761-146) and cultured at 22°C. Plates were kept in a Tupperware box with dampened paper towels and the lid loosely affixed to prevent cultures from drying out but to allow gas exchange.

Every 12 hours for 96 hours, 3 wells from each strain were fixed with 10 µl of 16% paraformaldehyde (Fisher Scientific Cat. No. 50-980-487) and stored at 4°C. After all time points were collected, each sample was counted by vortexing the sample at high speed for 10 seconds to fully mix the sample, then aliquoting 10 µl into a counting slide (Logos Biosystems Cat. No. L12001 [disposable] or L12011 [reusable]) and counting using a Luna-FL automated cell counter (Logos Biosystems, Anyang, KOR; Cat. No. L20001).

RNA sequencing and differential expression analysis for cRFXa and FoxJ1 mutants

30 ml of cells were grown to mid-log phase. For *cRFXa*$^{PTS\text{-}1}$, wild-type *S. rosetta* was used as the wild-type comparison strain. For *foxJ1*^{PTS}, which was isolated using cycloheximide resistance selection and contains the co-edited *rpl36a*^{P56Q} allele, the wild-type comparison strain was a

clone with only the *rpl36a*^{P56Q} mutation(Booth and King 2020). Three biological replicates were prepared, each on a separate day, processing one wild-type and one mutant culture at a time for cell lysis and RNA extraction.

For each replicate of each strain, 5 x 10^6 cells were processed for lysis and RNA extraction. Cells were centrifuged and washed with AKSW. Cells were resuspended in AKSW, counted, and aliquoted to 10 x 10^6 per aliquot, then resuspended in 100 µl of lysis buffer. This was incubated on ice for 10 minutes, passed ten times through a 30G needle and centrifuged at 6,000 x g for 10 minutes at 4°C. The supernatant was collected, brought to 100 µl with RNAse-free water, and RNA was purified using the RNAeasy kit from Qiagen, eluting in 30 µl of water (Cat. No. 74104).

Library preparation and sequencing was performed by the QB3-Berkeley Genomics core labs (QB3 Genomics, UC Berkeley, Berkeley, CA, RRID:SCR_022170). 500 ng were of RNA were used for library prep using the KAPA mRNA capture kit (Cat. No. 07962240001) for poly-A selection and the KAPA RNA HyperPrep kit (Cat. No. 08105952001). Truncated universal stub adapters were ligated to cDNA fragments, which were then extended via PCR using unique dual indexing primers into full length Illumina adapters. RNA integrity was assessed by Agilent Bioanalyzer 2100 before library prep using an Agilent RNA 6000 Nano Kit (Cat. No. 5067-1511). Sequencing libraries were also confirmed by Bioanalyzer 2100 for the correct size distribution using the Agilent High Sensitivity DNA Kit (Cat. No. 5067-4626). Library concentration was quantified by qPCR using the KAPA Library Quantification Kit (Cat. No. 079601400001) and libraries were pooled at equal concentrations before sequencing.

Sequencing was performed in one lane of an SP flow cell on the Illumina NovaSeq 6000 with an S4 flowcell, collecting between 45.4 million and 73.3 million 50 bp paired-end reads for each sample. Reads were de-multiplexed using Illumina bcl2fastq2 (v 2.20) and default settings, on a server running CentOS Linux 7. Reads checked for quality with fastqc (v 0.11.9), and aligned to predicted transcripts from the *S. rosetta* genome(Fairclough et al. 2013) using Salmon(Patro et al. 2017) (v 1.5.2.) and called for differential expression using edgeR(M. D. Robinson, McCarthy, and Smyth 2010), both implemented within the Trinity software package(Grabherr et al. 2011) (v 2.14.0). Transcripts with an average TPM value less than 1 for both wild-type and mutant cells were excluded from analysis. Further analysis and comparisons were done using Python scripts in Jupyter Notebook with plotting in Prism 9. TPM values for all replicates and differential expression tests are shared in File S5.

Conserved ciliome genes

Lists of evolutionarily conserved ciliary genes have been assembled by comparing datasets across eukaryotic diversity using approaches such as comparative genomics and mass spectrometry. Previous compilations of ciliary genes have been published as the Ciliary proteome database (Adrian Gherman, Erica E. Davis, and Nicholas Katsanis 2006), Cildb (Arnaiz et al. 2009) and SYSCILIA (van Dam et al. 2013).

Building on these databases, we curated our own set of human ciliary genes, focusing on components with a described functional role in ciliogenesis (File S6). Our list contained 269 genes. We identified likely orthologs of these genes in *S. rosetta, M. brevicollis,* or *S. punctatus* using the criteria of reciprocal best BLAST hits or a BLAST e-value < $1e^{-20}$. Finally, we removed duplicate hits to finalize a list of conserved ciliary genes, which was used for downstream analysis of RNA sequencing data and promoter motif content. 201 human ciliary genes were conserved in *S. rosetta*, 176 in *M. brevicollis*, and 182 in *S. punctatus*.

Protein binding microarray

RNA was prepared from wild-type *S. rosetta* cells grown to mid-log phase using the methods for lysis and RNA extraction described previously (see: *S. rosetta* cell type RNA sequencing and analysis). cDNA was prepared form this RNA using the SuperScript IV reverse transcriptase kit (Thermo Fisher Scientific, Cat. No. 18091050), with 150 ng of RNA input and dT(20) primers. The cRFXa CDS was amplified from cDNA using primers ATGTCACAGCAACAGGGGGT and CACGTCCGGTGGCCG using Q5 DNA polymerase (NEB Cat. No. M0491L)**,** with 2 µl of cDNA template in a 50 µl PCR reaction and 35x cycles. The PCR product was gel purified (Qiagen, Venlo, NLD, Cat. No. 28706) and cloned into TOPO pCR2.1 (Thermo Fisher Scientific Cat. No. K450001) after A-tailing with Taq polymerase (NEB Cat. No. M0273S) for 15 minutes at 72°C. The TOPO reaction was transformed into TOPO OneShot cells, cultured over-night, mini-prepped (Qiagen, Cat. No. 27106) and confirmed for correct insertion with Sanger sequencing (UC Berkeley DNA Sequencing Facility) using M13R primer.

The cRFXa CDS was amplified from the TOPO vector using primers TGCAGAGCTCAGGCGCGCCATGTCACAGCAACAGGGGGT and GCCGGATCCTCACCTGCAGGTCACGTCCGGTGGCCG using Q5 DNA polymerase in a 50 µl PCR reaction. The primers contain homology arms for Gibson assembly into the pTH6838 vector, which was linearized with restriction enzyme XhoI (NEB Cat. No. R016S). The pTH6838 vector is a T7-driven expression vector with a N-terminal GST tag. The amplified CDS and XhoI-digested vector were gel purified. Gibson assemblies were performed using the NEB HiFi Assembly Kit (New England Biolabs, Cat. No. E2621L) with 100 ng of insert and a 2:1 molar ratio of insert:vector. The Gibson reaction was transformed into chemically competent XL10 Gold *E. coli* (Agilent, Santa Clara, CA, Cat. No. 200315), cultured over-night, mini-prepped and confirmed for correct insertion with Sanger sequencing.

The TF samples were expressed by using a PURExpress In Vitro Protein Synthesis Kit (New England BioLabs) and analyzed in duplicate on two different PBM arrays (HK and ME) with differing probe sequences. PBM laboratory methods including data analysis followed the procedure described previously (Lam et al. 2011; Weirauch et al. 2013). PBM data were generated with motifs derived using Top10AlignZ (Weirauch et al. 2014).

Promoter transcription factor motif analysis

From the conserved ciliary genes in *S. rosetta*, *M. brevicollis,* or *S. punctatus* (File S6), we extracted the promoter regions, defined as 1000 base pairs upstream and 200 base pairs downstream of annotated transcription start sites, although other promoter definitions were tested to ascertain the robustness of the results (Figure 2.8C). Using the ciliary promoters and a background set of all promoters (-1000 to 200 bp from all protein-coding genes), we looked for ciliome-enriched motifs using HOMER(Heinz et al. 2010), specifically the findMotifs.pl script with default options. To create a list of motif instances from a HOMER-identified motif, we also called findMotifs.pl with the -find option.

For *S. rosetta*, we used gene models from assembly Proterospongia_sp_ATCC50818, hosted on Ensembl Protist. For *M. brevicollis*, we used gene models from assembly GCA_000002865.1, hosted on Ensembl Protist. For *S. punctatus*, we used gene models from assembly DAOM BR117, hosted on Ensembl Fungi.

Luciferase Reporter Assays

To compare luciferase activity between promoters, we built plasmids expressing both nanoluc and firefly luciferases codon-optimized for *S. rosetta*. This allows one promoter to be variable between plasmids while keeping the other promoter constant as a control for efficiency of transfection and plasmid retention. A codon-optimized *nanoluc* was previously published (Booth, Szmidt-Middleton, and King 2018); therefore we ordered a codon-optimized *firefly* as a gBlock (Integrated DNA Technologies) and ligated this in between 5' and 3' regulatory regions of *S. rosetta actin* (XM_004993513.1) in the NK587 backbone (Addgene), creating a new plasmid called NK621 (Addgene).

To construct the dual-luciferase plasmid, a fragment containing the *S. rosetta efl* (XM_004996684.1) 5' and 3' regulatory regions flanking the *nanoluc* ORF was digested from plasmid NK606 (Addgene) using MfeI-HF (Cat. No. R3589S) and KpnI-HF (Cat. No. R3142S) restriction enzymes from New England Biolabs. NK809, containing the *S. rosetta actin* (XM_004993513.1) 5' and 3' regulatory regions flanking the *firefly* ORF, was linearized using KpnI-HF (Cat. No. R3142S), EcoRI-HF (Cat. No. R3101S), and CIP (M0290S) from New England Biolabs. The fragments were purified on a 1% agarose gel and extracted with QIAquick Gel Extraction Kit (Qiagen Cat. No. 28706). The purified fragments were ligated using the Roche Rapid DNA Ligation Kit (Roche Diagnostics Cat. No. 11635379001) using 90 ng of total DNA and 5:1 ratio of insert:vector, then transformed into chemically competent XL10 Gold *E. coli* (Agilent, Santa Clara, CA, Cat. No. 200315), cultured over-night, mini-prepped and confirmed for correct assembly with Sanger sequencing. The resulting plasmid is identified as NK809 (Addgene #196406).

To test different promoters with this reporter plasmid, the 5' *efl* region next to *nanoluc* was replaced with a 5'UTR/promoter of interest. From *S. rosetta* genomic DNA, the 5' upstream region of the *spag6* gene (XM_004991453.1) including the 133 bp annotated 5' UTR plus an additional 852 bp upstream of that were amplified using forward primer ACTCACTCATTCTCTGCTGC and reverse primer CTTGTCTGTTTCGTGTGTGTG using Q5 DNA polymerase (NEB Cat. No. M0491L) in a 50 µl PCR reaction with 35x cycles. This was gel purified (Qiagen Cat. No. 28706) and cloned into TOPO pCR2.1 (Thermo Fisher Scientific Cat. No. K450001) after A-tailing with Taq polymerase (NEB Cat. No. M0273S) for 15 minutes at 72°C. The TOPO reaction was transformed into TOPO OneShot cells, cultured over-night, mini-prepped (Qiagen, Cat. No. 27106) and confirmed for correct insertion with Sanger sequencing (UC Berkeley DNA Sequencing Facility) using M13R primer. The NK809 backbone was amplified to include everything except for the *pEFL* sequence using primers TGCAAATTGTACAGAAGTCACTGT and ATGTCTGTCTTCACCCTCG using Q5 DNA polymerase. A minimal *spag6* promoter containing the 133 bp 5' UTR and 138 bp of additional 5' sequence was amplified to include homology arms for pMC001 without *pEFL* using primers ACTTCTGTACAATTTGCAAGACAACGCGCTGAAGAAGA and GAGGGTGAAGACAGACATCTTGTCTGTTTCGTGTGTGTGT. These two PCR products were ligated in a Gibson assembly using the NEB HiFi Assembly Kit (New England Biolabs, Cat. No. E2621L) with 100 ng of insert and a 2:1 molar ratio of insert:vector. The Gibson reaction was transformed into chemically competent XL10 Gold *E. coli* (Agilent, Santa Clara, CA, Cat. No. 200315), cultured over-night, mini-prepped and confirmed for correct insertion with Sanger sequencing. The resulting plasmid is called NK810 (Addgene #196407).

To mutate the RFX binding site in the *spag6* regulatory region, from GTTGCCAA to ACGTCCAA, the SPAG6 plasmid was amplified using primers TGTTGGCGTTGGCGGTGGTTGGACGTCAAAACAACGAAAATTACCCCAAATC and GATTTGGGGTAATTTTCGTTGTTTTGACGTCCAACCACCGCCAACGCCAACA, then assembled using the Agilent QuikChange Lightning Side-Directed Mutagenesis Kit (Cat. No. 210518), using DpnI to degrade the methylated (and non-mutated) template backbone. The reaction was transformed into chemically competent XL10 Gold *E. coli* (Agilent, Santa Clara, CA, Cat. No. 200315), cultured over-night, mini-prepped and confirmed for correct insertion with Sanger sequencing. The resulting plasmid is called NK811 (Addgene #196408).

To prepare for plasmid transfection into *S. rosetta*, the NK809, NK810, and NK811 plasmids were transformed into dam-/dcm- *E. coli* (New England Biolabs Cat. No. C2925H), then sent for large-scale preps and concentration to a value of 10 µg/µl in 10 mM Tris pH 8.5 using the Genewiz service (Azenta Life Sciences, Chelmsford, MA).

The plasmids were transfected into *S. rosetta* using the following protocol, which is similar to the genome editing protocol with some important differences. 48 h prior to the transfection, *S. rosetta* cells were inoculated in 120 ml of media at 8,000 cells/mL. This seeding density brings the culture to mid-log phase at the time of transfection.

On the day of transfection, to wash away bacteria from the choanoflagellates, the culture was split into three 50 ml conical tubes and centrifuged for 5 minutes at 2000 x g. The cell pellets were resuspended and combined in 50 ml of AKSW, followed by a 5 min spin at 2200 x g. The cells were washed once more with 50 ml AKSW and spun at 2400 x g. The pellet is resuspended in 100 µl AKSW and diluted 1:100 in AKSW for counting. Cells are diluted to 5 x 10^7 / mL in AKSW, then 100 µl aliquots (with 5 x 10^6 cells each) are prepared.

Priming buffer is prepared by diluting 10 µl of 1 mM papain (Sigma-Aldrich Cat. No. P3125-100MG) in 90 µl of dilution buffer (50 mM HEPES-KOH, pH 7.5, 200 mM NaCl, 20% glycerol, 10 mM cysteine, filter-sterilized and stored in aliquots at -80°C). This is then diluted 1:67 in the rest of the priming buffer (40 mM HEPES-KOH, pH 7.5, 34 mM lithium citrate, 50 mM L-cysteine, 15% PEG-8000, filter-sterilized and stored in aliquots at -80°C) for a final concentration of 1.5 µM papain (compared to 1 µM papain for genome editing). The priming buffer can be prepared while washing the cells.

Each aliquot of cells is spun at 800 x g for 5 minutes and resuspended in 100 µl priming buffer and incubated for 35 minutes at RT. The priming reaction is quenched by adding 1 µl of 50 mg/ml bovine serum albumin fraction V (Thermo Fisher Scientific Cat. No. BP1600-1000). Cells are spun at 1250 x g for 5 minutes and resuspended in 25 µl Lonza SF buffer (Lonza Cat. No. V4SC-2960).

For each transfection, 16 µl of Lonza SF buffer is mixed 1 µl of 10 µg/µl plasmid, 1 µl of 10 mM Tris pH 8.5, 2 µl of re-suspended *S. rosetta* cells, and 4 µl of the plasmid nucleofection master mix (10 µg/µl pUC19 plasmid DNA, 62.5 mM ATP-NaOH pH 7.5, 25 mg/ml heparin).

The nucleofection reactions are added to a 96-well nucleofection plate (Lonza Cat. No. V4SC-2960) and pulsed with a CM156 pulse in the Lonza 4D-Nucleofector (Cat. No. AAF-1003B for the core unit and AAF-1003S for the 96-well unit).

After the pulse, 100 µl of ice-cold recovery buffer (10 mM HEPES-KOH, pH 7.5; 0.9 M sorbitol; 8% [wt/vol] PEG 8000) is immediately added to each well of the nucleofection plate and incubated for 5 minutes. Then the entire contents of the well are added to 1 mL of 1.5% SWC + 1.5% CGM3 in AKSW in a 12-well plate and cultured at 22C. After one hour of culture, 10 µl of re-suspended *E. pacifica* bacteria (10 mg/ml in 1 ml AKSW).

24 hours after transfection, cells were prepared for the reporter assay. For each sample, the 1 ml culture was centrifuged at 4200 x g for 15 mins at 4C. The cell pellet was resuspended in 50 µl of lysis buffer [50 mM HEPES, pH 7.6, 100 mM NaCl, 1% (v/v) Triton X-100, 2 mM Pefabloc, 1 Roche EDTA free complete mini/5 mL, 1 mM EDTA, 10% (v/v) glycerol, 2 mM DTT], transferred to a white flat-bottom 96-well plate (Greiner Bio-One Cat. No. 655083) and incubated at RT for 10 mins. The lysates were analyzed for luciferase activity using the Nano-Glo Dual-Luciferase Reporter Assay System (Promega Cat. No. N1610). Luminescence was read on the SpectraMax i3x plate reader (Molecular Devices), using photon counting with 1 second of integration.

QUANTIFICATION AND STATISTICAL ANALYSIS

Information about the quantification and statistical details of experiments can be found in the corresponding figure legends. Statistical tests and graphs were produced using Prism 9.0.0.

Figure 2.1. The evolutionary history of cilia-associated transcription factors and their expression in choanoflagellates.

(A) Cilia evolved before the emergence of RFX and Fox TFs. The presence (filled circle) or absence (open circle) of RFX and Fox domain proteins is indicated for diverse eukaryotes (Files S1, S2; Materials and Methods). Half shading in the Fox/J1 column indicates the presence of Fox family members, while full shading indicates the presence of a putative FoxJ1 homolog reciprocal best BLAST hit with either the *Xenopus laevis* or *Schmidtea mediterranea* FoxJ1 (File S2; Materials Methods). Cilia have been observed in most eukaryotic lineages, indicating a cilium was present in the last eukaryotic common ancestor. RFX TFs are more phylogenetically restricted, having been found across opisthokonts and amoebozoans, while Fox TFs are nearly entirely restricted to opisthokonts (see Materials and Methods for rare exceptions to these patterns.) All choanoflagellates express Fox TF homologs and FoxJ1 orthologs were detected in most choanoflagellate species. Species tree represents a consensus of recent well-supported eukaryotic and clade-specific phylogenies (Adl et al. 2012; Dunn et al. 2008; Philippe et al. 2011; King and Rokas 2017; Carr et al. 2017).

(B) The *cRFXa* sub-family is widespread in choanoflagellates. RFX family relationships were determined using maximum-likelihood phylogenetic trees built by IQ-TREE (L.-T. Nguyen et al. 2015) (Figure 2.2A; File S1). All RFX TFs in choanoflagellates grouped into one of three well-supported sub-families: *cRFXa*, *cRFXb*, and *cRFXc.* For representative choanoflagellates, the presence (filled circle) or absence (open circle) of each sub-family is indicated. While *cRFXa* was detected in all cultured choanoflagellates that have been sequenced, *cRFXb* and *cRFXc* were restricted to subsets of choanoflagellate diversity.

(C) *cRFXa* is expressed in all surveyed *S. rosetta* life history stages. *S. rosetta* can transition between multiple colonial and solitary cell types (Dayel et al. 2011), including slow swimmers, rosettes, fast swimmers, and thecate cells. Cells in all life history stages depicted here bear motile cilia. RNA-seq analysis showed that only *cRFXa* is expressed above background levels (average TPM [transcripts per million] ≥ 1) in all cell types. *cRFXb* and *cRFXc* are only expressed above background levels in thecate cells (File S3). *foxJ1* is expressed in all cell types and most highly in fast swimmers (File S3). Shading indicates average TPM value of the gene across three biological replicates. Note the separate scale bars for *RFX* and *foxJ1* expression levels due to the approximately ten-fold difference in maximum expression level between these genes.

(D) Choanoflagellate *cRFXa* genes form a clade with the animal *RFX1/2/3* family (File S1). Width of branches indicates scales with UFboot support for the ancestral node and all nodes with less than 75% bootstrap support are collapsed. Labels A, B, and C indicate ancestral nodes of homologous choanoflagellate/animal RFX sub-families. Node A has 81% bootstrap support, Node B has 81% bootstrap support, and Node C has 85% bootstrap support. Branch lengths do not scale with evolutionary distance in this rendering. See Figure 2.2B for full annotated version of this phylogeny, including branch lengths, bootstrap values, and all species names. See Figure 2.2C for phylogenetic trees built with different trimming and reconstruction algorithms.

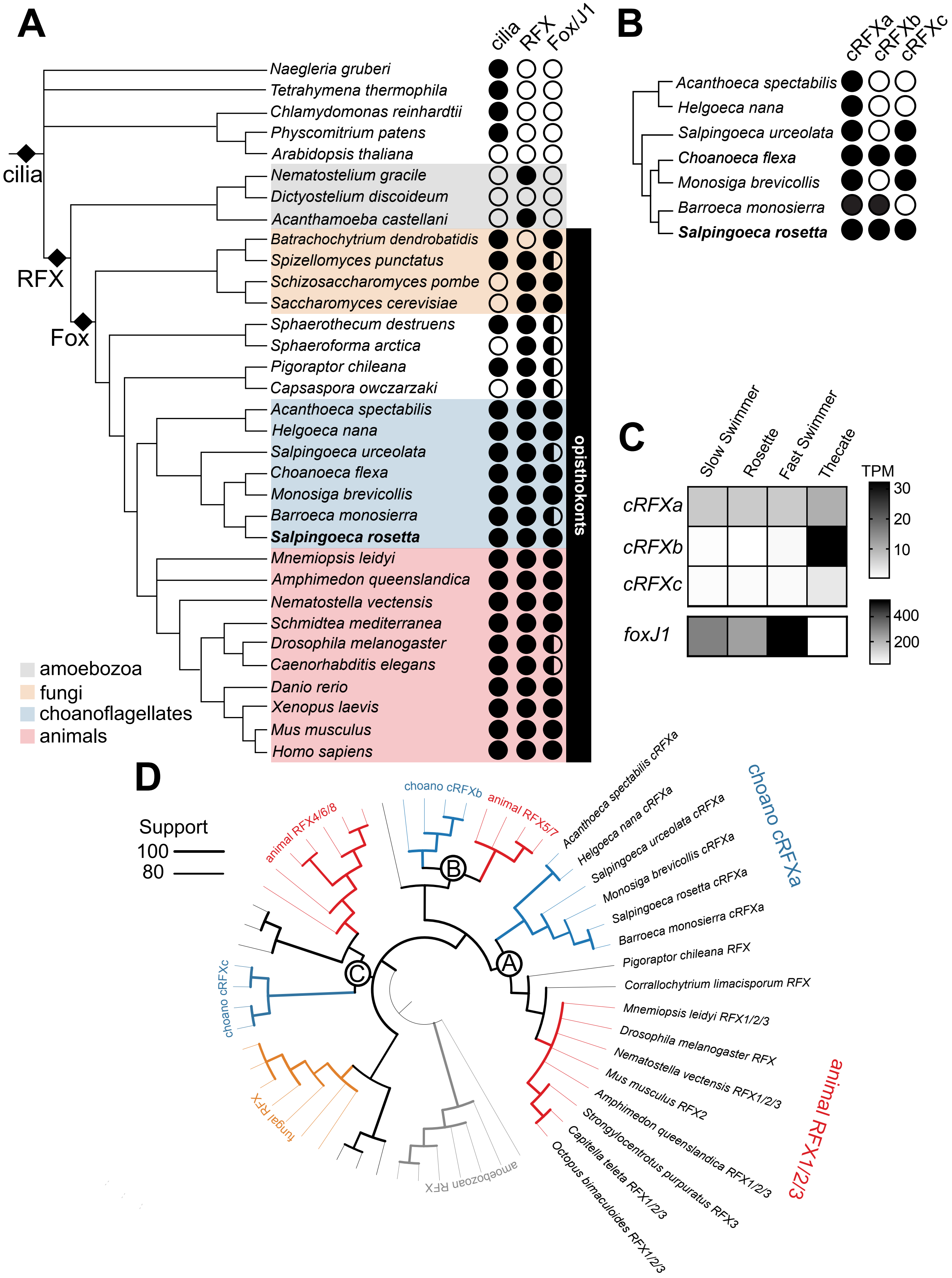

A
cilia
RFX
Fox/J1
cilia
RFX
Fox
Naegleria gruberi
Tetrahymena thermophila
Chlamydomonas reinhardtii
Physcomitrium patens
Arabidopsis thaliana
Nematostelium gracile
Dictyostelium discoideum
Acanthamoeba castellani
Batrachochytrium dendrobatidis
Spizellomyces punctatus
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Sphaerothecum destruens
Sphaeroforma arctica
Pigoraptor chileana
Capsaspora owczarzaki
Acanthoeca spectabilis
Helgoeca nana
Salpingoeca urceolata
Choanoeca flexa
Monosiga brevicollis
Barroeca monosierra
Salpingoeca rosetta
Mnemiopsis leidyi
Amphimedon queenslandica
Nematostella vectensis
Schmidtea mediterranea
Drosophila melanogaster
Caenorhabditis elegans
Danio rerio
Xenopus laevis
Mus musculus
Homo sapiens
opisthokonts
amoebozoa
fungi
choanoflagellates
animals
B
cRFXa
cRFXb
cRFXc
Acanthoeca spectabilis
Helgoeca nana
Salpingoeca urceolata
Choanoeca flexa
Monosiga brevicollis
Barroeca monosierra
Salpingoeca rosetta
C
Slow Swimmer
Rosette
Fast Swimmer
Thecate
TPM
30
20
10
400
200
cRFXa
cRFXb
cRFXc
foxJ1
D
Support
100
80
choano cRFXb
animal RFX5/7
animal RFX4/6/8
Acanthoeca spectabilis cRFXa
Helgoeca nana cRFXa
Salpingoeca urceolata cRFXa
Monosiga brevicollis cRFXa
Salpingoeca rosetta cRFXa
Barroeca monosierra cRFXa
choano cRFXa
Pigoraptor chileana RFX
Corrallochytrium limacisporum RFX
Mnemiopsis leidyi RFX1/2/3
Drosophila melanogaster RFX
Nematostella vectensis RFX1/2/3
Mus musculus RFX2
Amphimedon queenslandica RFX1/2/3
Strongylocentrotus purpuratus RFX3
Capitella teleta RFX1/2/3
Octopus bimaculoides RFX1/2/3
animal RFX1/2/3
A
B
C
choano cRFXc
fungal RFX
amoebozoan RFX

Figure 2.2. Phylogenetic trees of RFX genes.

(A) Choanoflagellate RFX genes form three sub-families. Choanoflagellate RFX protein sequences (File S1) were aligned with MAFFT, trimmed with ClipKIT, and assembled into a maximum-likelihood phylogenetic tree with IQ-TREE. Width of branches indicate UF-boot support. All nodes with less than 75% bootstrap support are collapsed. Every choanoflagellate with RFX genes contains a copy of *cRFXa* (green, 100% UF-boot support), while *cRFXb* (pink, 100% UF-boot support) and *cRFXc* (blue, 82% UF-boot support) are found in subsets of choanoflagellate taxa. Tree scale indicates length of branch corresponding to one substitution per site in amino acid alignment.

(B) Choanoflagellate *cRFXa* genes are orthologous to the animal *RFX1/2/3* sub-family, *cRFXb* genes are orthologous to animal *RFX5/7*, and *cRFXc* genes are orthologous to animal *RFX4/6/8*. Selected RFX protein sequences from across diverse opisthokonts and amoebozoans (File S1) were aligned with MAFFT, trimmed with ClipKIT, and assembled into a maximum-likelihood phylogenetic tree with IQ-TREE. The three previously discovered choanoflagellate RFX families were well-resolved, as were the three animal RFX families, fungal RFX genes, and amoebozoan RFX genes. Red letters and arrows indicate UF-boot support for nodes that connect animal and choanoflagellate RFX gene families. Note that ichthyosporeans (*A. parasiticum*, *C. fragrantissima*, *I. hoferi*) contain at least two RFX genes, one of which groups with *cRFXc* and *aRFX4/6/8*. Width of branches indicates bootstrap support and all nodes with less than 75% bootstrap support are collapsed. Tree scale indicates length of branch corresponding to one substitution per site in amino acid alignment.

(C) Different phylogenetic software packages recover similar phylogenetic relationships between animal and choanoflagellate RFX sub-families. A set of opisthokont and amoebozoan full-length RFX protein sequences (File S1, the same sequences used for Figure 2.1D and Figure 2.2B) were aligned with MAFFT, followed by alignment trimming with either ClipKIT or trimAl, and then maximum-likelihood tree construction with either IQ-TREE or RAxML. For both ML algorithms, automatic best model finding was used. For the combination of ClipKIT and IQ-TREE, we show SH-aLRT statistics (1000 iterations), for comparison with UF-boot statistics (the tree shown in Figure 2.1D). For Trimal/IQ-TREE, 1000 Ultrafast bootstraps were used. For RAxML trees, 100 bootstraps were used. Width of branches indicates bootstrap support. Animal, choanoflagellate, fungal, and amoebozoan RFX sub-families are indicated. Labels A,B,C indicate ancestral nodes of homologous choanoflagellate/animal RFX sub-families. Bootstrap supports are: ClipKIT/IQ-TREE/SH-aLRT (A: 66%, B: 89%, C: 58%); Trimal/IQ-TREE (A: 72%, B: 78%, C: 96%); ClipKIT/RAxML/Bootstrap (A: 19%, B: 39%, C: 30%); Trimal/RAxML/Bootstrap (A: 26%, B: 29%, C: 71%).

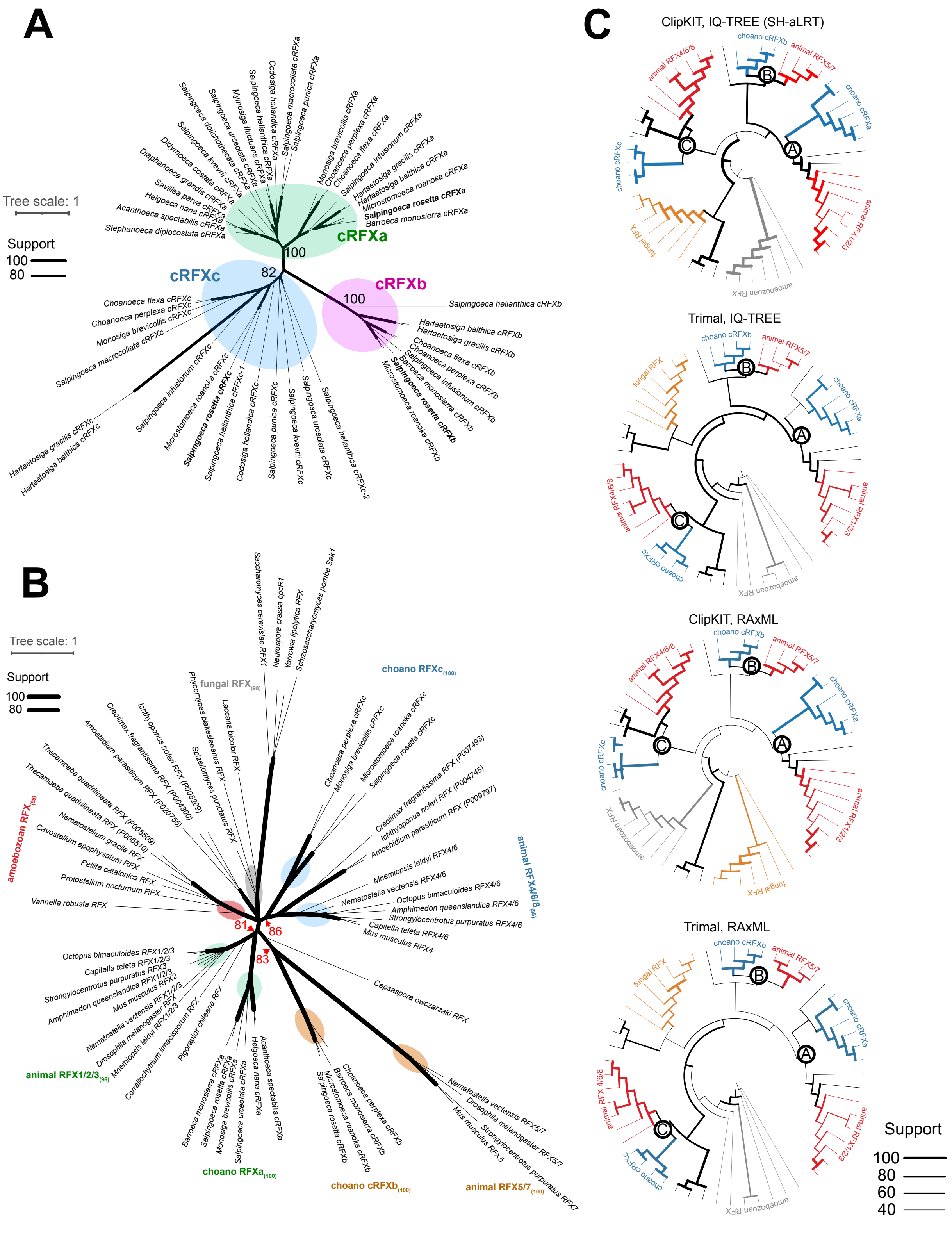

A
Tree scale: 1
Support
100
80
cRFXa
cRFXb
cRFXc
100
82
100
Stephanoeca diplocostata cRFXa
Acanthoeca spectabilis cRFXa
Helgoeca nana cRFXa
Savillea parva cRFXa
Diaphanoeca grandis cRFXa
Didymoeca costata cRFXa
Salpingoeca kvevrii cRFXa
Salpingoeca dolichothecata cRFXa
Salpingoeca urceolata cRFXa
Mylnosiga fluctuans cRFXa
Salpingoeca helianthica cRFXa
Codosiga hollandica cRFXa
Salpingoeca macrocollata cRFXa
Salpingoeca punica cRFXa
Monosiga brevicollis cRFXa
Choanoeca perplexa cRFXa
Choanoeca flexa cRFXa
Salpingoeca infusionum cRFXa
Hartaetosiga gracilis cRFXa
Hartaetosiga balthica cRFXa
Microstomoeca roanoka cRFXa
Salpingoeca rosetta cRFXa
Barroeca monosierra cRFXa
Salpingoeca helianthica cRFXb
Hartaetosiga balthica cRFXb
Hartaetosiga gracilis cRFXb
Choanoeca flexa cRFXb
Choanoeca perplexa cRFXb
Salpingoeca infusionum cRFXb
Barroeca monosierra cRFXb
Salpingoeca rosetta cRFXb
Microstomoeca roanoka cRFXb
Choanoeca flexa cRFXc
Choanoeca perplexa cRFXc
Monosiga brevicollis cRFXc
Salpingoeca macrocollata cRFXc
Hartaetosiga gracilis cRFXc
Hartaetosiga balthica cRFXc
Salpingoeca infusionum cRFXc
Microstomoeca roanoka cRFXc
Salpingoeca rosetta cRFXc
Salpingoeca helianthica cRFXc-1
Codosiga hollandica cRFXc
Salpingoeca punica cRFXc
Salpingoeca kvevrii cRFXc
Salpingoeca urceolata cRFXc
Salpingoeca helianthica cRFXc-2
B
Tree scale: 1
Support
100
80
fungal RFX(98)
Saccharomyces cerevisiae RFX1
Neurospora crassa cpcR1
Yarrowia lipolytica RFX
Schizosaccharomyces pombe Sak1
Phycomyces blakesleeanus RFX
Laccaria bicolor RFX
Spizellomyces punctatus RFX
choano RFXc(100)
Choanoeca perplexa cRFXc
Monosiga brevicollis cRFXc
Microstomoeca roanoka cRFXc
Salpingoeca rosetta cRFXc
Creolimax fragrantissima RFX (P007493)
Ichthyophonus hoferi RFX (P004745)
Amoebidium parasiticum RFX (P009797)
animal RFX4/6/8(98)
Mnemiopsis leidyi RFX4/6
Nematostella vectensis RFX4/6
Octopus bimaculoides RFX4/6
Amphimedon queenslandica RFX4/6
Strongylocentrotus purpuratus RFX4/6
Capitella teleta RFX4/6
Mus musculus RFX4
Capsaspora owczarzaki RFX
Nematostella vectensis RFX5/7
Drosophila melanogaster RFX5/7
Strongylocentrotus purpuratus RFX7
Mus musculus RFX5
animal RFX5/7(100)
Choanoeca perplexa cRFXb
Barroeca monosierra cRFXb
Microstomoeca roanoka cRFXb
Salpingoeca rosetta cRFXb
choano cRFXb(100)
Acanthoeca spectabilis cRFXa
Helgoeca nana cRFXa
Salpingoeca urceolata cRFXa
Monosiga brevicollis cRFXa
Salpingoeca rosetta cRFXa
Barroeca monosierra cRFXa
choano RFXa(100)
Pigoraptor chileana RFX
Corallochytrium limacisporum RFX
Mnemiopsis leidyi RFX1/2/3
Drosophila melanogaster RFX
Nematostella vectensis RFX1/2/3
Mus musculus RFX2
Amphimedon queenslandica RFX1/2/3
Strongylocentrotus purpuratus RFX3
Capitella teleta RFX1/2/3
Octopus bimaculoides RFX1/2/3
animal RFX1/2/3(96)
81
86
83
amoebozoan RFX(98)
Vannella robusta RFX
Protostelium nocturnum RFX
Pellita catalonica RFX
Cavostelium apophysatum RFX
Nematostelium gracile RFX
Thecamoeba quadrilineata RFX (P005510)
Thecamoeba quadrilineata RFX (P005509)
Amoebidium parasiticum RFX (P020755)
Creolimax fragrantissima RFX (P004300)
Ichthyophonus hoferi RFX (P005309)
C
ClipKIT, IQ-TREE (SH-aLRT)
choano cRFXb
animal RFX5/7
animal RFX4/6/8
choano cRFXa
choano cRFXc
animal RFX1/2/3
fungal RFX
amoebozoan RFX
A
B
C
Trimal, IQ-TREE
choano cRFXb
animal RFX5/7
fungal RFX
choano cRFXa
animal RFX4/6/8
animal RFX1/2/3
choano cRFXc
amoebozoan RFX
A
B
C
ClipKIT, RAxML
choano cRFXb
animal RFX5/7
animal RFX4/6/8
choano cRFXa
choano cRFXc
animal RFX1/2/3
amoebozoan RFX
fungal RFX
A
B
C
Trimal, RAxML
choano cRFXb
animal RFX5/7
fungal RFX
choano cRFXa
animal RFX 4/6/8
animal RFX1/2/3
choano cRFXc
amoebozoan RFX
A
B
C
Support
100
80
60
40

Figure 2.3. Truncation of cRFXa results in cell proliferation and ciliogenesis defects.

(A) The *S. rosetta cRFXa* locus encodes a protein that contains an N-terminal DNA-binding domain (DBD) followed by two conserved domains of unknown function (B, C) and a dimerization domain (DIM) (Choksi et al. 2014). The *cRFXa* locus was targeted by a guide RNA (gRNA) that anneals to an exon near the 5' end of the gene coupled with a homology-directed repair template that inserts a cassette (TTTATTAATTAAATAAA) that encodes an early stop codon (* in translation product, grey shaded letters). The edited allele is called *cRFXa^{PTS}* (for Premature Termination Signal(Booth and King 2020)) and codes for a truncated polypeptide of 24 amino acids. Two independent *cRFXa^{PTS}* mutants, *cRFXa^{PTS-1}* and *cRFXa^{PTS-2}*, were recovered. The *cRFXa^{PTS-1}* strain was reverted to a wild-type polypeptide sequence to create the *cRFXa^{REV}* strain, which harbors a synonymous GTC→GTG (Valine) that allows its genotype to be distinguished from that of *cRFXa^{WT}* cells. DSB = double-strand break, PAM = protospacer adjacent motif. Numbers indicate amino acid positions in coding DNA sequence. See Figure 2.4A for genotyping confirmation.

(B) Truncation of cRFXa in the *cRFXa^{PTS-1}* and *cRFXa^{PTS-2}* strains resulted in delayed cell proliferation compared to *cRFXa^{WT}* and *cRFXa^{REV}* cells. Cells were diluted to 1,000 cells/ml and triplicate samples were collected and counted every 12 hours for 96 hours. The mean values were plotted with the standard error of the mean shown as dotted lines. See Figures 2.4B and 2.4C for growth curves of other TF mutant strains.

(C) Cilia lengths were comparable in *cRFXa^{WT}* (19.73 µm) and *cRFXa^{PTS-1}* (19.63 µm) cells. Cilia lengths in randomly selected cells from three biological replicates were analyzed (see Materials and Methods), measuring 20 cells/genotype/replicate, for 60 cells total/replicate. Colored dots show replicate mean values and grey dots show the lengths of individual cilia. Unpaired t-test compares mean values of biological replicates (n = 3), p-value = 0.959. ns = not significant.

(D) Choanoflagellate ciliogenesis can be synchronized and quantified following ciliary removal. To this end, *S. rosetta* cells were treated with 10% glycerol and cold-shocked (STAR Methods), which results in the severing of cilia. We observed that nascent cilia sometimes collapse and resorb before a new round of ciliary growth begins (grey arrows). The point at which the growing cilium passed the edge of the microvillar collar was used as a marker of successful ciliogenesis (asterisk).

(E) A representative time series shows a *cRFXa^{WT}* cell in the process of ciliogenesis, from cilia removal (00:00 mm:ss) to growth (15:30-17:00 mm:ss). The nascent cilium (arrowhead) extended as a thin, straight protrusion; ciliary beating had not begun yet. The cell shifted slightly in position under the coverslip between 00:00 and 15:30. Scale bar = 5 µm. See Video S1 and S2 for complete examples of *cRFXa^{WT}* regeneration.

(F) A representative time series shows a *cRFXa^{PTS-1}* cell in the process of ciliogenesis. Arrowhead marks a nascent cilium that collapsed (20:00 time point) and resorbed back into the cell. Resorption here was complete in one minute, which was typical. The cell shifted slightly in position under the coverslip between 00:00 and 19:30. Scale bar = 5 µm. See Video S3 and S4 for complete examples of *cRFXa^{WT}* regeneration.

(G) Nascent cilia in *cRFXa*$^{PTS\text{-}1}$ cells collapse more frequently than *cRFXa*^{WT} cells during ciliogenesis. For each of two biological replicates, 20+ randomly selected cells were scored for the number of ciliary collapses during a 60-minute ciliary regeneration period. Colored dots show mean values of each biological replicate and grey dots show values for individual cells. The mean number of collapses (across biological replicates) was 1.00 collapses/cell/60 minutes for *cRFXa*^{WT} and 6.24 for *cRFXa*$^{PTS\text{-}1}$. Unpaired t-test compares mean values of biological replicates (n = 2), p-value = 0.0012.

(H) *cRFXa*$^{PTS\text{-}1}$ cells are delayed in ciliary regeneration relative to *cRFXa*^{WT} and *cRFXa*^{REV} cells. Graph shows the percent of cells that have completed ciliary regeneration as a function of time (three biological replicates, 20 cells each). Regeneration was defined as the point at which the cilium grows past the collar (see panel D). Dotted lines show standard error of the mean across three replicates. See Figure 2.4D-F for ciliary regeneration curves for other TF mutant strains.

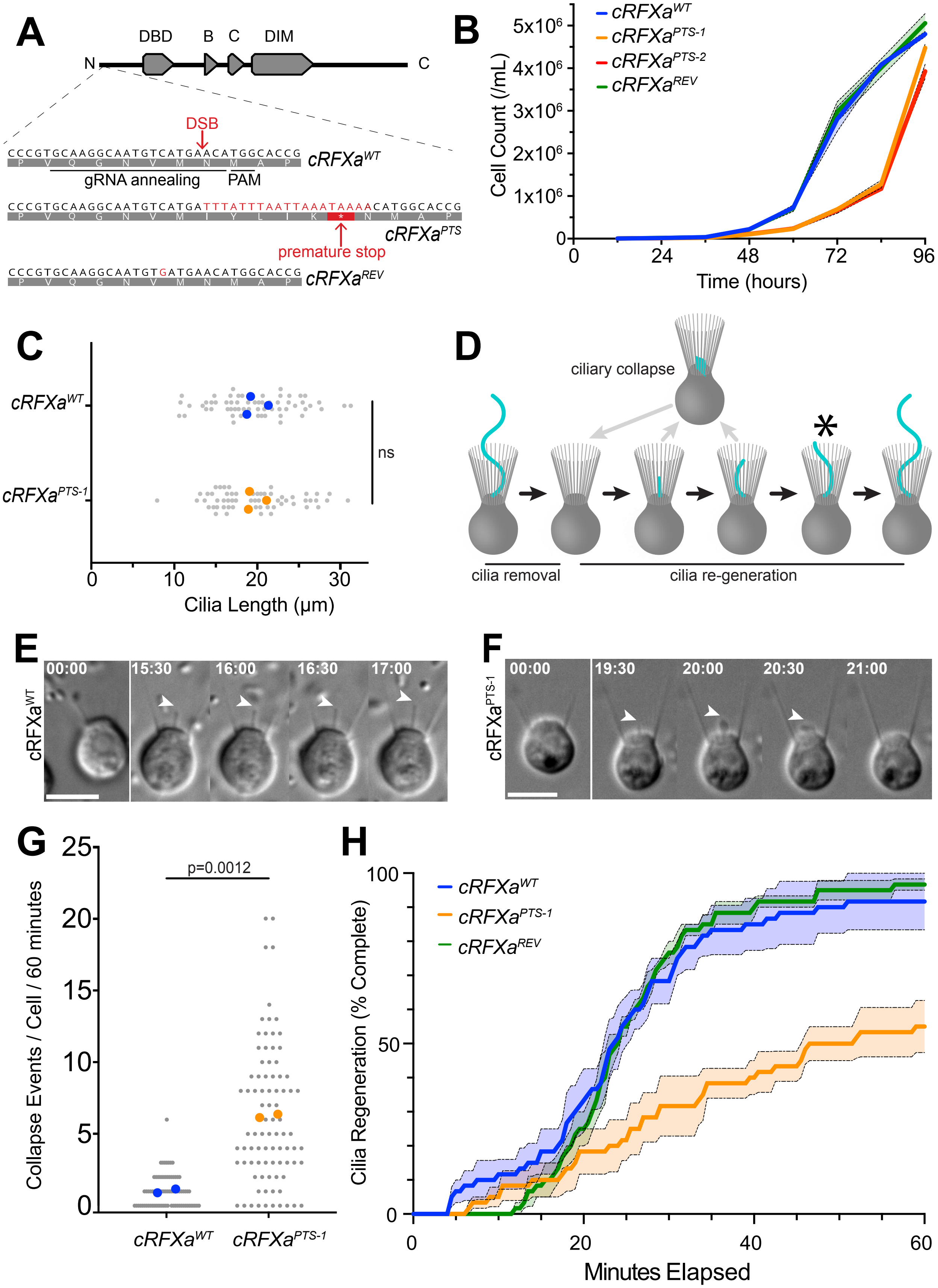

A
DBD
B
C
DIM
N
C
DSB
CCCGTGCAAGGCAATGTCATGAACATGGCACCG
P V Q G N V M N M A P
cRFXa^WT
gRNA annealing
PAM
CCCGTGCAAGGCAATGTCATGATTTATTTAATTAAATAAAACATGGCACCG
P V Q G N V M I Y L I K * N M A P
cRFXa^PTS
premature stop
CCCGTGCAAGGCAATGTGATGAACATGGCACCG
P V Q G N V M N M A P
cRFXa^REV
B
cRFXa^WT
cRFXa^PTS-1
cRFXa^PTS-2
cRFXa^REV
Cell Count (/mL)
5x10^6
4x10^6
3x10^6
2x10^6
1x10^6
0
0
24
48
72
96
Time (hours)
C
cRFXa^WT
cRFXa^PTS-1
ns
0
10
20
30
Cilia Length (µm)
D
ciliary collapse
*
cilia removal
cilia re-generation
E
cRFXa^WT
00:00
15:30
16:00
16:30
17:00
F
cRFXa^PTS-1
00:00
19:30
20:00
20:30
21:00
G
p=0.0012
Collapse Events / Cell / 60 minutes
25
20
15
10
5
0
cRFXa^WT
cRFXa^PTS-1
H
cRFXa^WT
cRFXa^PTS-1
cRFXa^REV
Cilia Regeneration (% Complete)
100
50
0
0
20
40
60
Minutes Elapsed

Figure 2.4. Genotyping, growth rates, and ciliogenesis of transcription factor mutant strains.

(A) Clonally isolated cells from CRISPR genome editing experiments were genotyped by PCR and Sanger sequencing (File S3). Numbers show relative position in the coding DNA sequence of the target gene. The TTTATTTAATTAAATAAA cassette is introduced in an exon and creates a stop codon in every possible reading frame. All genes are truncated before the DNA-binding domain.

(B) *cRFXb*^{PTS} and *cRFXc*^{PTS} strains show equivalent proliferation rates compared to an isogenic strain (Materials and Methods). As in Figure 2.3B, cells were diluted to 1,000 cells / ml and triplicate samples were collected and counted every 12 hours for 96 hours. The mean values are plotted with the standard error of the mean shown as dotted lines.

(C) *foxJ1*^{PTS} shows an equivalent proliferation rate compared to an isogenic strain. Growth rates were assayed and quantified as in Figure 2.3B.

(D) Ciliogenesis for *cRFXa*^{WT}, *cRFXa*$^{PTS\text{-}1}$, and *cRFXa*$^{PTS\text{-}2}$ was compared under standard growth conditions as described in Figure 2G. For each strain, triplicate experiments were done, quantifying the time point of completed regeneration for each of 20 cells, and plotting the percent that have completed ciliary regeneration as a function of time. Dotted lines show standard error of the mean across the three replicates.

(E) The *cRFXb*^{PTS} strain shows no defect in ciliogenesis. The data represents the average of three triplicate experiments (n=20 cells each) plotting the percent that have completed ciliary regeneration as a function of time. Dotted lines represent standard error of the mean.

(F) The *cRFXc*^{PTS} strain shows no defect in ciliogenesis. The data represents the average of three triplicate experiments (n=20 cells each) plotting the percent that have completed ciliary regeneration as a function of time. Dotted lines represent standard error of the mean.

(G) The *foxJ1*^{PTS} strain shows no defect in ciliogenesis. The data represents the average of three triplicate experiments (n=20+ cells each) plotting the percent that have completed ciliary regeneration as a function of time. Dotted lines represent standard error of the mean.

(H) The *cRFXa*^{PTS}*foxJ1*^{PTS} double mutant strain shows a ciliogenesis defect comparable to that observed in *cRFXa*$^{PTS\text{-}1}$. The data represents the average of three triplicate experiments (n=20 cells each) plotting the percent that have completed ciliary regeneration as a function of time. Dotted lines represent standard error of the mean.

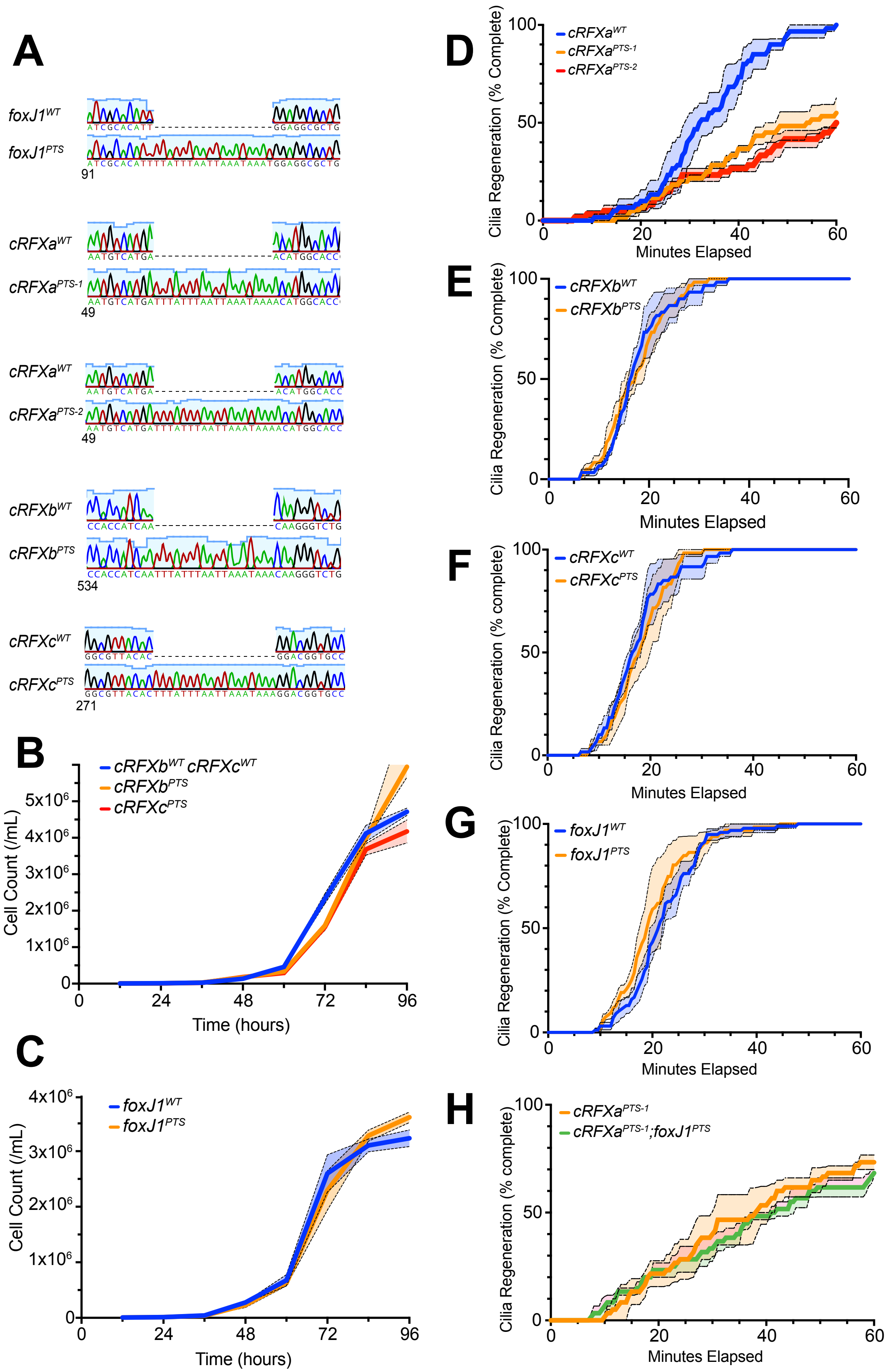

A
foxJ1^WT
ATCGCACATT----GGAGGCGCTG
foxJ1^PTS
ATCGCACATTTTATTTAATTAAATAAATGGAGGCGCTG
91
cRFXa^WT
AATGTCATGA----ACATGGCACC
cRFXa^PTS-1
AATGTCATGATTTATTTAATTAAATAAAACATGGCACC
49
cRFXa^WT
AATGTCATGA----ACATGGCACC
cRFXa^PTS-2
AATGTCATGATTTATTTAATTAAATAAAACATGGCACC
49
cRFXb^WT
CCACCATCAA----CAAGGGTCTG
cRFXb^PTS
CCACCATCAATTTATTTAATTAAATAAACAAGGGTCTG
534
cRFXc^WT
GGCGTTACAC----GGACGGTGCC
cRFXc^PTS
GGCGTTACACTTTATTTAATTAAATAAAGGACGGTGCC
271
B
cRFXb^WT cRFXc^WT
cRFXb^PTS
cRFXc^PTS
Cell Count (/mL)
5x10^6
4x10^6
3x10^6
2x10^6
1x10^6
0
0
24
48
72
96
Time (hours)
C
foxJ1^WT
foxJ1^PTS
Cell Count (/mL)
4x10^6
3x10^6
2x10^6
1x10^6
0
0
24
48
72
96
Time (hours)
D
cRFXa^WT
cRFXa^PTS-1
cRFXa^PTS-2
Cilia Regeneration (% Complete)
100
50
0
0
20
40
60
Minutes Elapsed
E
cRFXb^WT
cRFXb^PTS
Cilia Regeneration (% Complete)
100
50
0
0
20
40
60
Minutes Elapsed
F
cRFXc^WT
cRFXc^PTS
Cilia Regeneration (% complete)
100
50
0
0
20
40
60
Minutes Elapsed
G
foxJ1^WT
foxJ1^PTS
Cilia Regeneration (% Complete)
100
50
0
0
20
40
60
Minutes Elapsed
H
cRFXa^PTS-1
cRFXa^PTS-1;foxJ1^PTS
Cilia Regeneration (% complete)
100
50
0
0
20
40
60
Minutes Elapsed

Figure 2.5. *cRFXa^{PTS}* cells down-regulate conserved ciliary genes.

(A) Eukaryotic motile cilia are constructed from conserved macromolecular complexes encoded by dozens of genes (File S6). The side view of a cilium shows how the basal body, which nucleates the microtubules of the cilium, docks to the cell membrane. Intraflagellar transport (IFT) trains traverse in both anterograde and retrograde directions to shuttle ciliary components to the growing tip. Axoneme cross-section shows the organization of microtubule doublets in the cilium as well as the inter-doublet links and dynein arms that power ciliary motility.

(B) Ciliary genes, including *foxJ1*, were significantly down-regulated in the *cRFXa^{PTS-1}* mutant compared to *cRFXa^{WT}* cells. Shown are log_2FC values for HsaSro conserved ciliary genes (n = 201), compared to all other predicted genes in the *S. rosetta* genome, for both *cRFXa^{PTS-1}* and *foxJ1^{PTS}* strains, relative to wild-type cells. All strains were sequenced while cells were in mid-log growth phase as slow swimmers. Red dots indicate genes whose differential expression was called as significant by edgeR using a false discovery rate (FDR) cut-off of < 0.001. For *cRFXa^{PTS-1}*, the average log_2FC of all ciliary genes was -0.68 compared to -0.017 for non-ciliary genes (Mann-Whitney p-value < 0.0001). For *foxj1^{PTS}*, the average log_2FC of all ciliary genes was -0.04 compared to -0.0086 for non-ciliary genes (Mann-Whitney p-value = 0.0187). See Figure 2.6 for RNA-seq expression of *S. rosetta* ciliary genes identified by mass spectrometry(Sigg et al. 2017).

(C) Many categories of ciliary genes were down-regulated in *cRFXa^{PTS-1}* cells. For each category, the horizontal bar shows the average log_2FC value for genes in that category, while dots indicate the expression changes of individual genes. Red dots indicate a gene with an edgeR false discovery rate (FDR) < 0.001.

(D) Predicted functions for all 65 genes down-regulated more than four-fold (log_2FC < -2) in the *cRFXa^{PTS-1}* mutant. Categories were called based on protein domain annotation by InterProScan and the closest human BLAST hit (File S6).

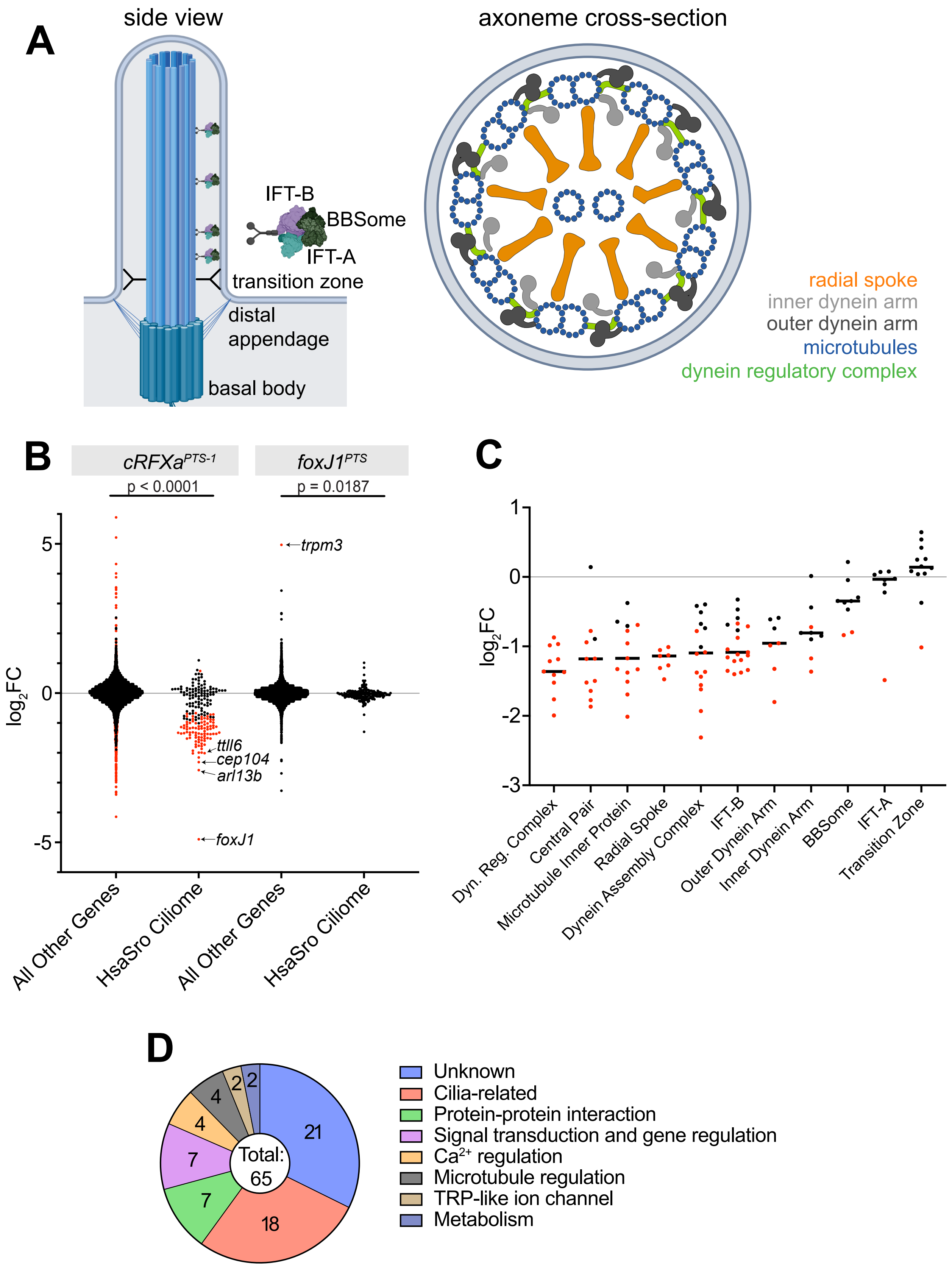

A
side view
IFT-B
BBSome
IFT-A
transition zone
distal
appendage
basal body
axoneme cross-section
radial spoke
inner dynein arm
outer dynein arm
microtubules
dynein regulatory complex
B
cRFXa^{PTS-1}
foxJ1^{PTS}
p < 0.0001
p = 0.0187
log$_2$FC
5
0
-5
trpm3
ttll6
cep104
arl13b
foxJ1
All Other Genes
HsaSro Ciliome
All Other Genes
HsaSro Ciliome
C
log$_2$FC
1
0
-1
-2
-3
Dyn. Reg. Complex
Central Pair
Microtubule Inner Protein
Radial Spoke
Dynein Assembly Complex
IFT-B
Outer Dynein Arm
Inner Dynein Arm
BBSome
IFT-A
Transition Zone
D
21
18
7
7
4
4
2
2
Total:
65
Unknown
Cilia-related
Protein-protein interaction
Signal transduction and gene regulation
Ca^{2+} regulation
Microtubule regulation
TRP-like ion channel
Metabolism

Figure 2.6. Genes with protein products identified in *S. rosetta* cilia by mass spectrometry (MS) are down-regulated in *cRFXa*$^{PTS\text{-}1}$ cells.

Dots show $\log_2$FC values for proteins not identified by MS in the cilia, genes that were identified by MS in cilia, and a subset of the MS hits: proteins whose sea urchin and sea anemone orthologs were also identified in the ciliary proteome of those respective taxa (MS hit conserved). MS data from (Sigg et al. 2017)*,* with 464 proteins identified in the *S. rosetta* ciliome. 131 of these are likely to have conserved ciliary function across Choanozoa, due to the presence of orthologs detected in the ciliary proteomes of sea urchins and sea anemones. Transcripts whose products were detected in the ciliary proteome were on average down-regulated in *cRFXa*$^{PTS\text{-}1}$ mutant cells (avg $\log_2$FC = -0.50), and the subset of choanozoan-conserved ciliary genes showed more extensive down-regulation ($\log_2$FC = -1.00), suggesting that ciliary genes with evolutionarily conserved function have greater dependence on RFX-mediated transcriptional regulation in *S. rosetta*.

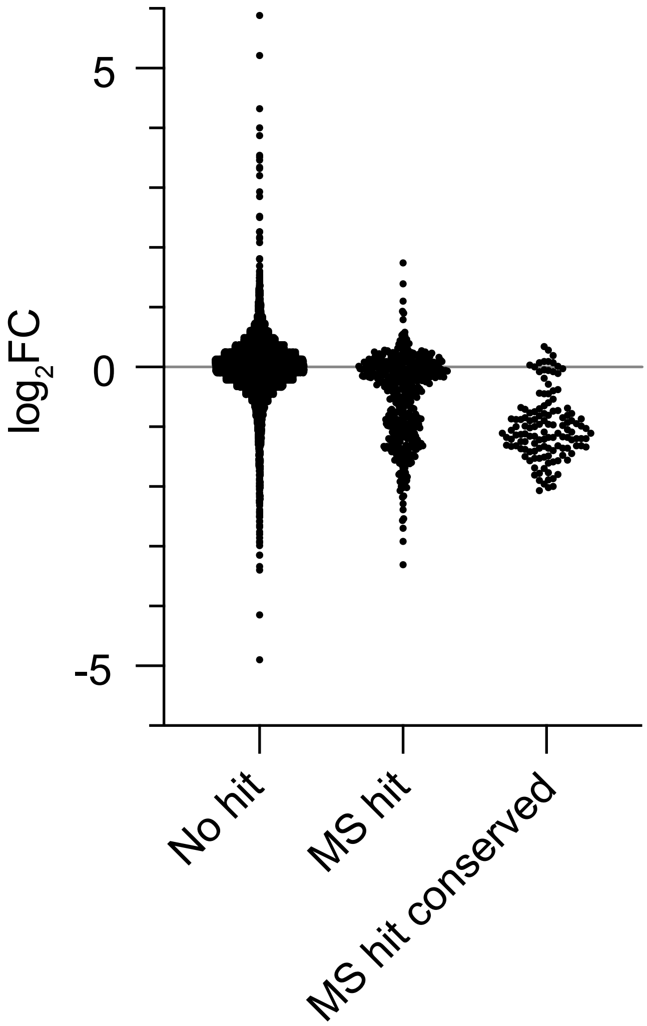

5
0
-5
log$_2$FC
No hit
MS hit
MS hit conserved

Figure 2.7. Functional RFX motifs are enriched in choanoflagellate ciliary gene promoters.

(A) The *H. sapiens* RFX2 consensus motif as determined by PBM (Cis-BP ID #M02449_2.00) (Weirauch et al. 2014). This motif represents the binding preferences of a single monomer, although RFX binding sites can occur as a tandem inverted repeat of two monomeric sites (the X-box) that bind to an RFX dimer. The DNA binding specificity for RFX TFs is conserved across animal and fungal RFX proteins(Piasecki, Burghoorn, and Swoboda 2010; Efimenko et al. 2005; Quigley and Kintner 2017). See Figure 2.8A for an RFX DBD alignment, Figure 2.8B for the *H. sapiens* RFX2 motif as determined by ChIP-Seq, and Figure 2.8D for the *M. musculus* FoxJ1 PBM motif.

(B) The only enriched sequence motif in the promoters of choanoflagellate ciliary genes matched RFX binding sites from animal studies. Shown are the most enriched HsaSro ciliome promoter motifs for *S. rosetta* and *M. brevicollis*, as determined by the HOMER *de novo* motif finding algorithm. Note the GTTGYCA consensus shared between the two choanoflagellate HsaSro ciliome-enriched motifs and the *H. sapiens* RFX2 motif. This represents the binding specificity of a single RFX DBD. For HOMER, ciliome promoters were defined as 1000 bp upstream and 200 bp downstream of annotated transcription start sites of HsaSro conserved ciliome genes (File S6), although the same RFX motif was recovered using variable definitions of promoter length (Figure 2.8C). Asterisk indicates a position not shared by animal or fungal RFX motifs.

(C) Percentage of HsaSro ciliome promoters with RFX-like motif compared to all mRNA promoters for both *S. rosetta* and *M. brevicollis*. RFX motifs are significantly enriched in HsaSro ciliome promoters compared to all promoters, with enrichment p-values reported by HOMER.

(D) The DNA binding specificity of *S. rosetta* cRFXa *in vitro*, as determined by protein binding microarray. The *in vitro* motif was built from the top ten scoring 8-mer hits (E-score range: 0.481-0.486). Asterisk indicates a position not shared by animal RFX motifs.

(E) In HsaSro conserved ciliary genes, RFX motifs are preferentially located near transcription start sites. The motif density within promoters is shown for HsaSro conserved ciliome promoters and for all other promoters. The RFX motif identified by HOMER (Figure 2.7B) in *S. rosetta* was used. Normalized motif density (y-axis) describes the proportion of all motifs that fall into a 100 bp sliding window centered on any given position on the x-axis. The x-axis gives promoter position relative to the predicted transcription start sites of conserved ciliary genes (black line) or all other genes (grey line). See Figure 2.8E for the same analysis applied to HsaMbrev ciliary promoters using the *M. brevicollis* RFX motif shown in Figure 2.7B.

(F) To functionally test the necessity of predicted an RFX binding site for gene activation, the 5’ UTR and proximal promoter of the *spag6* gene from *S. rosetta* was cloned in front of a *nanoluc* open reading frame which was codon-optimized for *S. rosetta*. A second reporter construct was made in which the predicted RFX binding site was mutated. As an internal normalization step, the plasmid also encodes the *firefly* luciferase under strong expression from the *S. rosetta actin* promoter.

(G) Mutation of the predicted RFX binding site in the *spag6* promoter/5’UTR decreased expression of the nanoluc luciferase to an average of 61% of wild-type activity. Three biological replicates were assayed, with 3-6 transfections per construct in each replicate. To normalize for transfection efficiency, the reporter plasmid coded for a second luciferase (*firefly*) under the *actin* promoter. This allowed for the normalization of transfection efficiency by taking the ratio of nanoluc signal to firefly signal. This ratio was then normalized to the expression from the strong *EFL* promoter, included as a positive control in all experiments. Individual values are plotted in gray and averages for each biological replicate plotted in orange. The average across biological replicates is represented by a horizontal bar. P-values are shown for a paired t-test between $P_{spag6\text{-}wt}$ and $P_{spag6\text{-}\Delta TFBS}$ reporters transfected into each genotype, using the mean value for each biological replicate (n = 3), as well as un unpaired t-test for $P_{spag6\text{-}wt}$ transfected into either *cRFXa*^{WT} or *cRFXa*^{PTS} cells, again using the mean value for each biological replicate (n = 3).

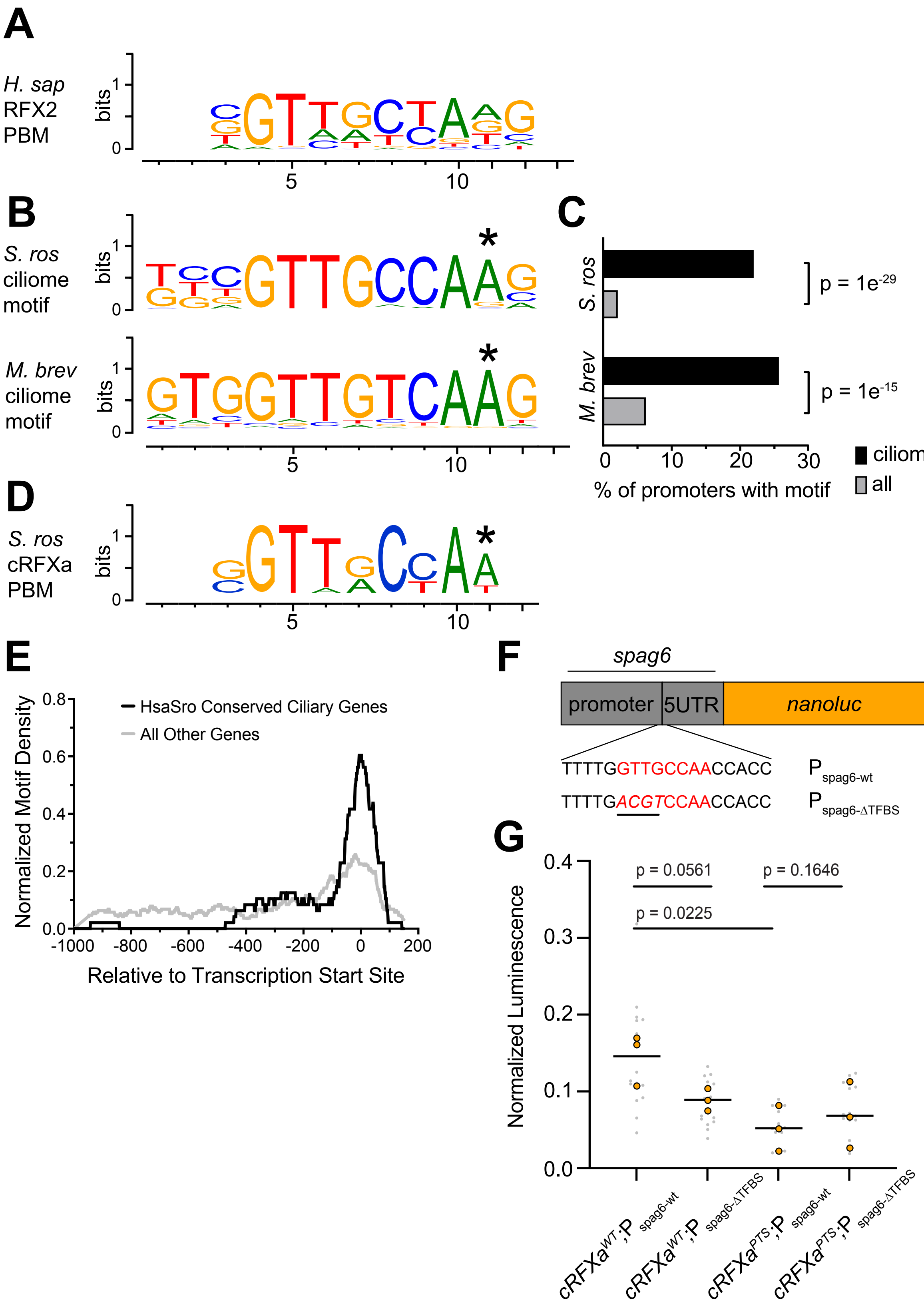

A
H. sap
RFX2
PBM
bits
0
1
5
10
B
S. ros
ciliome
motif
bits
*
M. brev
ciliome
motif
C
S. ros
M. brev
p = 1e-29
p = 1e-15
0
10
20
30
% of promoters with motif
ciliome
all
D
S. ros
cRFXa
PBM
*
E
Normalized Motif Density
0.0
0.2
0.4
0.6
0.8
HsaSro Conserved Ciliary Genes
All Other Genes
-1000
-800
-600
-400
-200
0
200
Relative to Transcription Start Site
F
spag6
efl
promoter
5UTR
nanoluc
3UTR
TTTTGGTTGCCAACCACC
TTTTGACGTCCAACCACC
P spag6-wt
P spag6-ΔTFBS
G
Normalized Luminescence
0.0
0.1
0.2
0.3
0.4
p = 0.0561
p = 0.1646
p = 0.0225
cRFXaWT;P spag6-wt
cRFXaWT;P spag6-ΔTFBS
cRFXaPTS;P spag6-wt
cRFXaPTS;P spag6-ΔTFBS

Figure 2.8. DNA-binding motifs for RFX and FoxJ1 transcription factors.

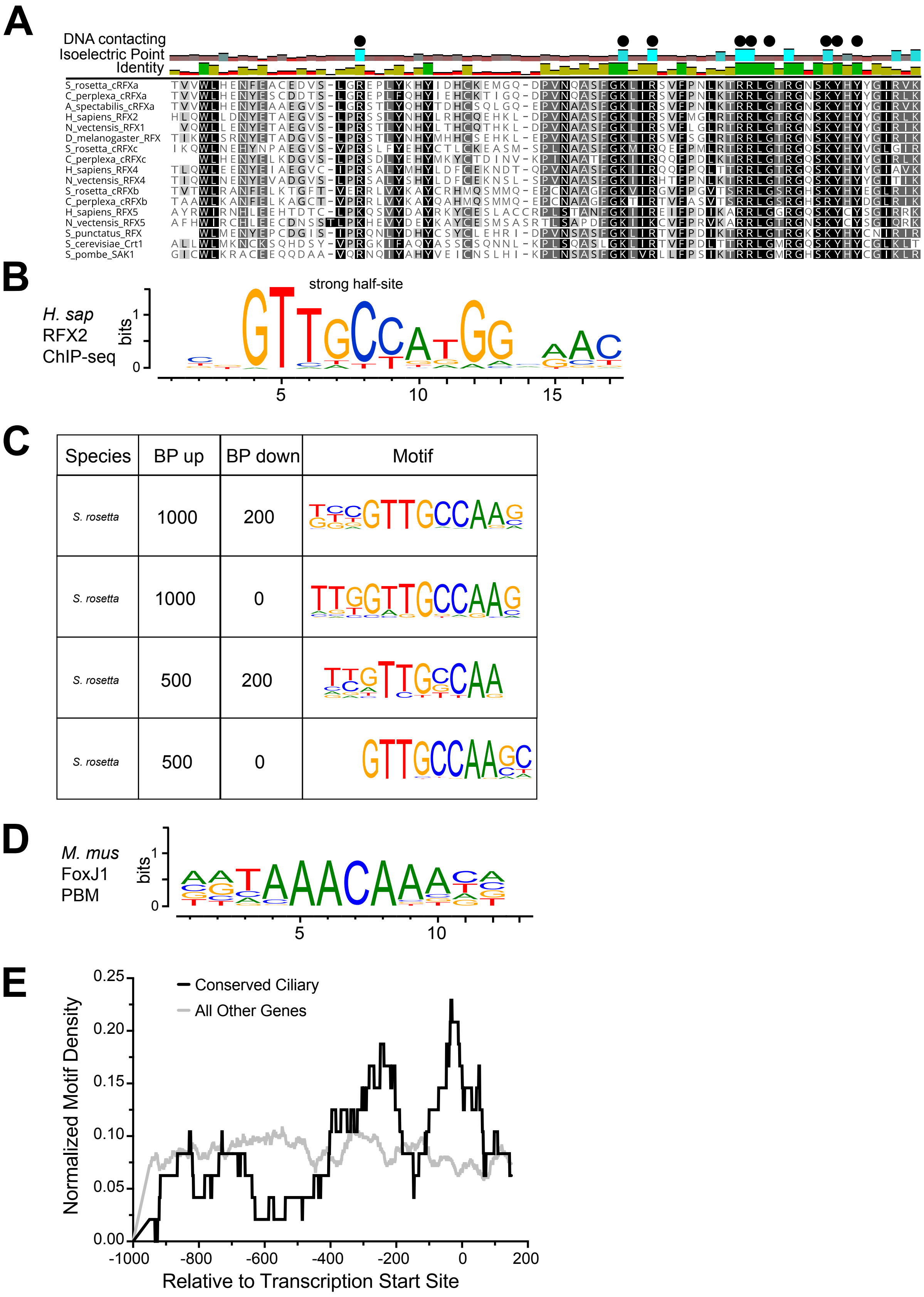
(A) RFX DNA-binding domain sequences have highly conserved DNA contacting residues. Selected RFX DNA-binding domains were aligned with MUSCLE and individual residues shaded according to identity. DNA contacting residues as determined by a crystal structure of *H. sapiens* RFX1 are labeled with black circles. These largely basic residues (note their correspondence with the average isoelectric point of each residue in the alignment) are almost perfectly conserved across all RFX sequences.

(B) The *H. sapiens* RFX2 consensus motif derived from ChIP-seq (JASPAR MA0600.1). This motif consists of two inverted, palindromic half-sites, one of which (here shown on the left) has stricter specificity requirements. The DNA binding specificity for RFX TFs is conserved across animal RFX proteins(Piasecki, Burghoorn, and Swoboda 2010; Efimenko et al. 2005; Quigley and Kintner 2017).

(C) Identification of a ciliome-enriched RFX motif is robust to definitions of promoter length. RFX-like motifs are identified as the most enriched in *S. rosetta* ciliome promoters across different definitions of promoter length, relative to annotated transcription start sites. Promoters were extracted using the criteria displayed and analyzed for motif enrichment using HOMER and our set of HsaSro conserved ciliary genes (File S6).

(D) The DNA binding preferences of *Mus musculus* FoxJ1 as determined by PBM (Cis-BP ID #M00161_2.00). The consensus motif is distinct from RFX binding motifs and does not show similarity to the motif identified as enriched in HsaSro ciliary promoter regions.

(E) *M. brevicollis* RFX motifs are preferentially located near transcription start sites of ciliome genes. The motif density within promoters is shown for motif instances in conserved ciliome genes, as well as motif instances in all other promoters. The RFX motif identified by HOMER (Figure 2.7B) in *M. brevicollis* was used. Normalized motif density (y-axis) describes the proportion of all motifs that fall into a 100 bp sliding window centered on any given position on the x-axis. The x-axis gives promoter position relative to the predicted transcription start sites of conserved ciliary genes (black line) or all other genes (grey line).



Species	BP up	BP down	Motif
S. rosetta	1000	200	TCCGTTGCCAAG
S. rosetta	1000	0	TTGGTTGCCAAG
S. rosetta	500	200	TTGTTGCCAA
S. rosetta	500	0	GTTGCCAAGC

Appendix

Entomophthovirus: An insect-derived iflavirus that infects a behavior manipulating fungal pathogen of dipterans

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

Coyle, M. C., Elya, C. N., Bronski, M. & Eisen, M. B. Entomophthovirus: An insect-derived iflavirus that infects a behavior manipulating fungal pathogen of dipterans. *bioRxiv* 371526 (2018) doi:10.1101/371526.

Abstract

We discovered a virus infecting *Entomophthora muscae*, a behavior-manipulating fungal pathogen of dipterans. The virus, which we name Entomophthovirus, is a capsid-forming, positive-strand RNA virus in the viral family iflaviridae, whose known members almost exclusively infect insects. We show that the virus RNA is expressed at high levels in fungal cells *in vitro* and during *in vivo* infections of *Drosophila melanogaster*, and that virus particles are present in *E. muscae*. Two close relatives of the virus had been previously described as insect viruses based on the presence of viral genomes in transcriptomes assembled from RNA extracted from wild dipterans. By analyzing sequencing data from these earlier reports, we show that both dipteran samples were co-infected with *E. muscae*. We also find the virus in RNA sequencing data from samples of two other species of dipterans, *Musca domestica* and *Delia radicum*, known to be infected with *E. muscae*. These data establish that Entomophthovirus is widely, and seemingly obligately, associated with *E. muscae*. As other members of the iflaviridae cause behavioral changes in insects, we speculate on the possibility that Entomophthovirus plays a role in *E. muscae* involved host manipulation.

Introduction

A wide variety of microbes have evolved the ability to manipulate animal behavior in ways that appear to advance microbial fitness (Eisthen and Theis 2016; Forsythe, Kunze, and Bienenstock 2012; Hoover et al. 2011; Hughes et al. 2016; Libersat 2003; Rohrscheib and Brownlie 2013; Roy et al. 2006; Sampson and Mazmanian 2015; Wang et al. 2015). Among them is the fungal pathogen of dipterans *Entomophthora muscae*. Originally identified in the 19th century (Cohn 1855), *E. muscae* has been observed infecting a wide variety of fly species (Kramer and Steinkraus 1981; Steinkraus and Kramer 1987), which exhibit a distinct series of behaviors prior to death: they climb to a high location (summiting), extend their proboscides which become attached to the surface via fungal growths that protrude from the tip (Bałzy 1984), and extend their wings in a characteristic "death pose."

Shortly after flies killed by *E. muscae* die, conidiophores emerge through the weakest points of the cuticle and forcibly eject spores at high velocity, ideally (from the fungal point of view) landing on a new host and propagating the infection. The wing position removes a major obstacle to spores escaping the immediate vicinity of the fly, and the elevation benefits the fungus by increasing the target range covered by traveling spores.

We recently reported the isolation of a strain of *E. muscae* from wild *Drosophila* and its propagation in lab-reared *D. melanogaster* and as an *in vitro* culture (Elya et al. 2018). *Drosophila* infected by this strain of *E. muscae* manifest the same set of behavioral changes as have been described in other flies.

In order to build a reference *E. muscae* transcriptome free of *Drosophila* RNA to study the behavior of the fungus in infected flies, we sequenced mRNA from our *in vitro* culture. As part of our initial quality checks of the *in vitro* mRNA sequence data, we used BLAST to search a small random subset of reads against GenBank for related sequences. We were surprised to find that a large number of reads from the *E. muscae* transcriptome aligned with near 100% identity to a virus identified by mRNA sequencing of wild *Drosophila* (Twyford virus; GenBank: KP714075.1) (Webster et al. 2015).

We initially suspected that this virus was misannotated - that the sample of flies from which the virus was isolated was infected with *E. muscae* and that the sequence was actually a transposable element from the repeat rich *E. muscae* genome. However when we subsequently began examining reads and initial assemblies of the *E. muscae* genome generated from our *in vitro* culture, we were astonished to find that the virus was not in our assembly or in any of the genomic reads from either Illumina or Pacific Biosciences sequencing.

This led us to more closely examine the original discovery of Twyford virus by ((Webster et al. 2015). Twyford is one of roughly two dozen new viruses identified by assembling mRNAs isolated from multiple large collections of wild caught *Drosophila melanogaster* and screening for virus-like sequences not found in the *D. melanogaster* genome. It is an Iflavirus, a family of positive-strand RNA viruses, related to picornaviruses, that infect a wide variety of insect hosts. A nearly complete 8 kb genome of Twyford was assembled from flies caught in the southern England village that gave it its name.

Of the viruses isolated in this study two things stand out about Twyford. First, it is rare, only appearing once in a panel of 16 independent wild populations screened for known and newly identified *Drosophila* viruses. Second, and more notably, small RNAs derived from Twyford virus have unusual characteristics compared to those derived from other viruses: they show a strong negative strand bias and have an almost complete bias for a 5' U base. Since this pattern was unique for *Drosophila* viruses, it suggested that small RNAs aligning to Twyford virus had not been generated by the canonical *Drosophila* Ago2-Dcr system. Webster et al. explored the hypothesis that the virus was infecting a eukaryotic commensal of *Drosophila*, but rejected candidate mites, nematodes and fungi for various reasons, and suggested the small RNAs they observed may have come from a previously unknown *Drosophila* pathway.

Here we present direct evidence that the virus they identified is infecting *E. muscae* and appears to be obligately associated with *E. muscae* in the wild. The viral RNA is found at high concentration in our *E. muscae* liquid culture, it cannot be washed away from fungal cells, and its expression tracks with *E. muscae* levels during infection of *Drosophila* in the lab. Small RNAs from our *in vitro* culture have precisely those characteristics described by (Webster et al. 2015).

We purified the virus from our *in vitro* culture and have used electron microscopy (EM) to show that it forms structures of approximately the same size as other picornaviruses. We also used EM to show that there are intact virus particles in *E. muscae* cells, confirming that the virus is infecting the fungus.

We obtained reads from the original Twyford sample from public databases, and using our genomic data as a reference against which to screen reads, we show that at least one fly in that sample was infected with *E. muscae*. We also find close relatives of the virus present in three other transcriptomes of wild dipterans, two from individual flies known to be infected with *E. muscae*, and a third from a mixed collection of flies that also contains mRNAs from *E. muscae*. Below we present details of our discovery, isolation and characterization of this virus, which we propose renaming Entomophthovirus to reflect its host and its novelty as the first picornavirus known to infect a fungus and the first virus of the insect-infecting iflaviridae known to infect a fungal pathogen of insects. We also discuss the possibility that the virus may be involved in behavior manipulation.

Results

Discovery of an iflavirus in in vitro culture of an isolate of *Entomophthora muscae*

We recently described the isolation of a strain of *E. muscae* from wild *Drosophila* species caught in Berkeley, CA in the summer of 2016. Following standard protocols (Hajek, Papierok, and

Eilenberg 2012), we captured spores ejected from an individual *D. melanogaster* recently killed by *E. muscae* on a liquid medium (Elya et al. 2018). We verified that the culture contained *E. muscae* by genotyping both the ITS and 28S loci. We sequenced mRNA from the liquid culture approximately five months after it was established, obtaining 38.9 million paired-end reads of 150 bp.

As discussed above we ran individual reads from mRNAs from the *in vitro* culture through NCBI's blastn, which we did to confirm that the sample was from *E. muscae*, and noticed that many reads aligned with ~95% identity to Twyford virus (GenBank: KP714075.1). After assembling an *E. muscae in vitro* transcriptome using TRINITY (Grabherr et al. 2011), we compared all of the assembled transcripts against Twyford, and found several highly similar versions of a slightly longer sequence (it appears the original Twyford virus sequence was truncated).

The genome of the virus in our liquid culture, which we refer to here as *D. melanogaster* Entomophthovirus (DmEV), is 8,832 basepairs and encodes a single 2,901 amino acid open reading frame. This single viral pro-protein contains the six proteins characteristic of iflaviruses: three coat proteins, an RNA helicase, a protease and a RNA-dependent RNA polymerase (Figure A.1A). We built a phylogenetic tree to compare DmEV to other iflaviruses and picornaviruses, which shows its placement in a subclade with other insect viruses as well as a plant iflavirus (Tomato matilda virus) (Figure A.1B). Reads aligning to this DmEV genome account for ~15% of the reads from the *in vitro* culture, suggesting that the virus is abundant and actively replicating within *E. musace* cells *in vitro*.

Entomophthovirus expression during *in vivo* infection of *D. melanogaster* with *E. muscae*

We previously described an experiment in which we sequenced RNAs from either whole animals or extracted brains from *D. melanogaster* infected with *E. muscae* over the course of an infection, as well as time-matched uninfected controls (Elya et al. 2018). We examined reads from this experiment for DmEV and found significant levels of DmEV RNA in some flies infected with *E. muscae* (but not unexposed controls) beginning 72 hours post exposure (Figure A.2), which is when we begin to see a significant fraction of reads aligning to the *E. muscae* transcriptome. In six of the twelve flies sampled 72 to 96 hours after infection, DmEV represents over ten percent of all reads (with a maximum of an astonishing 38 percent). There is, however, considerable inter-animal variation: at 72 hours between two and 21 percent of all reads align to DmEV, and at 96 hours the range is one to 38 percent. At 120 hours the fraction of reads aligning to DmEV drops while those aligning to *E. muscae* continues to rise.

Evidence of intracellular and extracellular virus in *E. muscae in vitro* culture

At this point our only evidence for the existence of DmEV in the fungus was the presence of iflaviral RNA in a *Drosophila*-free *in vitro* culture, and we therefore sought to demonstrate that there are viral particles in the *in vitro* culture. We started with the supernatant, assuming that if viral particles are being made they would either actively exit cells or be released as cells are lysed.

We spun down fungal cells and filtered the supernatant through a 0.22 um filter to retain potential viral particles. Using primers to specifically amplify a 831 bp segment of the DmEV genome, we found DmEV enriched in the retentate by RT-PCR (Figure A.3A). We then re-suspended the cell pellet and collected cells on a filter by vacuum filtration. After thoroughly washing the cells by vacuum filtration, we found a strong signal for DmEV in the eluted cell fraction (Figure A.3A). No viral signal was detected in the media used to culture *E. muscae* or in stocks of the flies we use for our *in vivo E. muscae* infections.

We used TEM to directly confirm the presence of virus particles, first in a sample purified from the *in vitro* supernatant (Figure A.3B). Iflavirus capsids have been reported to have a diameter of around 30 nanometers (Silva et al. 2015), and the regular size and shape of viral particles help them stand out by TEM. Indeed, a uranyl acetate negative stain of viral particles collected by ultracentrifugation from the extracellular fraction showed an abundance of symmetric ~30 nm objects, consistent with our expectations for the iflaviridae (Figure A.3B).

We next carried out double contrast uranyl acetate/lead citrate staining of fixed *E. muscae* cells to look for virus particles. A large fraction of sections had intracellular particles consistent with viral capsids by virtue of their ~30 nm diameter and their strong electron density (Figure A.3C). Notably, these particles were never found inside the nucleus, mitochondria, or other clearly demarcated organelles (Figure A.3C). Nor was their concentration noticeably higher near the plasma membrane or endomembranes, although the fixation process might have obscured fine spatial information. Only a fraction of sections seemed to possess any viral-like particles at all, while the infected cells showed a high viral titer.

E. muscae is present in the Twyford samples

Having established that DmEV is present in *E. muscae* cells and is replicating at high levels in *D. melanogaster* infected with *E. muscae*, we were curious if we could find evidence of *E. muscae* in samples in which closely related viruses were identified. We began with data from (Webster et al. 2015). We obtained reads for the original samples from the NCBI's Sequence Read Archive (SRA): SRR1914527 which contains flies from UK (including Twyford) and SRR1914484 which contains flies from non-UK sources.

Using an set of 17,826 genes from a preliminary annotation of the *E. muscae* genome filtered to remove regions that cross-align with the *D. melanogaster* genome, we found 823 read pairs that align to *E. muscae* in the Twyford sample while there are 0 in the non-Twyford sample. An additional 575 read pairs from the Twyford sample align discordantly to the *E. muscae* annotation, while 1,220 have a single read from the pair that aligns, reflecting the incomplete and fragmented nature of the current *E. muscae* annotation. In total reads aligning to 1,500 distinct *E. muscae* genes were identified with an average mismatch frequency of 0.012, consistent with the sample containing a strain of *E. muscae* closely related to but divergent from the Berkeley sample from which the genome was derived.

As described above, (Webster et al. 2015) had noted an unusual profile of small RNAs isolated from the Twyford sample that aligned to the Twyford virus genome. We therefore sequenced small RNAs from our *in vitro* culture. The 924,861 small RNA reads that align to DmEV from our *in vitro E. muscae* culture have very similar properties to those described for Twyford (Figure A.4). They show a ~70/30 negative strand bias with a strong preference for a 5' U base. Furthermore, using Augustus software to predict open reading frames from our de novo-assembled *E. muscae* genome, we see at least one clear Dicer homolog (g4150, e-value < e^-134, % identity > 35%), suggesting that small RNAs may be processed by a Dicer pathway in *E. muscae*. Collectively this evidence demonstrates that the Twyford virus described by (Webster et al. 2015) as a *D. melanogaster* virus was in fact DmEV present in their sample because of a concurrent infection of at least one of their flies with *E. muscae*.

***Entomophthovirus* in other samples**

GenBank contains a second virus closely related to DmEV and Twyford, Hubei picorna-like virus 39 (H39; KX883974.1). H39 was identified using similar methods to those of (Webster et al. 2015) as part of a large survey of viruses from different collections of arthropod taxa (Shi et al. 2016). We obtained the raw sequencing reads for the 67 different samples used in this experiments and aligned them to both DmEV and the *E. muscae* transcriptome.

The sample from which H39 was identified, a collection of diverse dipterans including one species of *Drosophila*, contains 13,258 reads that align to DmEV as well as 762 reads (out of 96,396,434) that align to 342 different *E. muscae* transcripts. This established that a second wild sample of flies from which a close relative of DmEV was isolated also contained *E. muscae*. None of the remaining 66 samples from (Shi et al. 2016) contain reads that align to either *E. musae* or to any version of DmEV.

Having initially discovered Entomophthovirus (EV) in a shotgun transcriptome assembly, we searched NCBI's Transcriptome Sequence Assembly (TSA) database for closely related transcripts and identified a series of hits from an individual of *Delia radicum*, a dipteran known as the cabbage fly, infected with *E. muscae* (De Fine Licht, Jensen, and Eilenberg 2017). The hits include an essentially full-length EV annotated as an *E. muscae* transcript (GenBank locus GENB01034640), which we henceforth refer to as DrEV. All three wild-caught *E. muscae* infected *D. radicum* had large numbers of reads aligning to DrEV. This dataset also included mRNA sequencing data from six individuals of the housefly *Musca domestica* infected in the laboratory from two wild *M. domestica* infected with *E. muscae*. We identified a different variant of EV in all six of these samples. We refer to this variant as MdEV.

Interestingly, we found no reads aligning to any version of EV in sequencing data from the *in vitro* cultures of *E. muscae* derived from one of the wild caught *M. domestica* described by (De

Fine Licht, Jensen, and Eilenberg 2017), demonstrating that it is possible to clear the viral infection of *E. muscae*.

Finally, we searched for EV in transcriptome data from flies at various stages of infection with a variety of different pathogenic bacteria to explore whether EV might be an opportunistic infection in flies undergoing immune collapse (Troha et al. 2018) and did not find any, consistent with the observation from (Webster et al. 2015) that Twyford virus is rare in *Drosophila*.

In summary, our survey of currently published sequencing data suggests that EV is obligately associated with *E. muscae* in *in vivo* infections. We cannot find a single hit for EV and its very close relatives (including Twyford and H39), where *E. muscae* infection was not confirmed phenotypically or suggested by a large number of reads aligning specifically to numerous *E. muscae* transcripts (Shi et al. 2016; Webster et al. 2015). This co-occurrence of fungus and virus appears in a variety of dipteran hosts, including *D. melanogaster*, *M. domestica*, and *D. radicum*.

Diversity of Entomophthovirus

Detailed analysis of the reads from our *D. melanogaster* samples revealed the presence of three substantially different versions of EV, two (DmEV1 and DmEV2) dominant in the *in vitro* culture, the other (DmEV3) dominant in the *in vivo* samples. DmEV1 and DmEV2 have a pairwise nucleotide divergence of .05, meaning they differ at one base in 20, while these two have an average pairwise divergence of .19 to DmEV3.

Both nucleotide and protein phylogenies (Figure A.5) of the seven sequences - DmEV1, DmEV2, DmEV3, Twyford, H39, DrEV and MdEV - place DmEV1 and DmEV2 together with Twyford (which we also assume was derived from *D. melanogaster*) with the three in a clade with DrEV, while DmEV3 is the sister taxa of MdEV in a separate clade. The placement of H39 is inconsistent: the nucleotide tree places H39 as an outgroup to the other six, the protein tree as a deeply branching member of the clade with DmEV1, DmEV2, Twyford and DrEV.

Unsurprisingly, we also see evidence for ongoing evolution of the virus. We see a small number of polymorphisms within DmEV3 in our *in vivo* time course. They are too distant from each other to phase, but there are a set of four polymorphisms that are always at the same frequency in the reads from an individual fly but at different frequencies between individuals, suggesting there is some form of bottlenecking of a mixed population, likely during spore transmission or the early phases of infection, although we have not demonstrated this experimentally.

Discussion

We believe these data unambiguously establish that the virus we originally identified in our *E. muscae in vitro* culture, along with two closely related viruses previously described as dipteran viruses, are variants of an iflavirus that infects *E. muscae* and is transmitted along with *E. muscae* to and from infected dipteran hosts. We have demonstrated that the virus is actively replicating in *E. muscae* in culture, that *E. muscae* generates anti-viral RNAs from the virus, that the virus forms capsids in the fungus, that it is transmitted along with the fungus from fly to fly, and that where the virus is found in nature it is always associated with *E. muscae* infected dipterans. For these reasons we formally propose naming this virus Entomophthovirus.

This is, to our knowledge, the first known member of the viral order Picornavirales to infect a fungus. As virtually all other known iflaviruses are insect pathogens, it seems likely that the virus moved from an insect host to an ancestor of *E. muscae*. When this happened and how broad the association is remains to be seen, but there is precedent for host switching in iflaviruses (Saqib, Wylie, and Jones 2015), and many fungal viruses are members of families that also infect animals (Son, Yu, and Kim 2015).

The consistent association suggests either that, in spite of an active anti-viral response from the fungus, the virus is never cleared from fungal cells and is present in fungal spores. The extent of virus-associated fungal mortality is unclear, but the high levels of viral RNA and the large number of capsids in the *in vitro* culture suggest that the virus has relatively low pathogenicity.

An alternative hypothesis is that EV provides some fitness advantage to *E. muscae* and its presence is essential to successful transmission of the infection, serving as a form of positive selection to maintain the association. One tantalizing possibility is that the virus is involved in behavior manipulation. Many animal viruses have behavioral effects on their hosts, including a baculovirus that induces summiting behavior in caterpillars (Katsuma et al. 2012), and, more specifically several other iflaviruses that induce a range of behaviors in their insect hosts (Dheilly et al. 2015; Fujiyuki et al. 2005).

For example, Kakugo Virus, which is a subtype of Deformed-Wing Virus, is an iflavirus that infects honeybees, and has been shown to be associated with aggressive colony behavior (Fujiyuki et al. 2005). Another iflavirus (*D. coccinellae* paralysis virus; DcPV) was recently shown to be actively involved in the parasitoid induced paralytic behavioral manipulation of ladybugs (Dheilly et al. 2015). Finally, an iflavirus was recently found associated with *Bombyx mori* infected with the behavioral manipulating ascomycete fungus *Cordyceps militaris* (Suzuki et al. 2015), although the nature of the fungal-viral association and its significant are unknown.

Many important questions regarding the relationship between EV and *E. muscae* remain open. Can *E. muscae* cleared of virus infect flies, induce behaviors and transmit the infection between flies? Does the virus replicate in fly cells during infection, or is it restricted to the fungus? Can the virus infect flies in the absence of fungus, and, if so, does it induce behavioral changes? How widespread is the association between entomopathogenic fungi and viruses?

On this later point, another distantly related picorna-like virus, RiPV-1, was recently identified in mRNA sequencing reads from bean bugs, *Riptortus pedestris*, infected with a distantly related ascomycete gentomopathogenic fungus, *Beauveria bassiana* (Yang et al. 2016). Like EV, RiPV-1 replicates at high rate in infected insects. In this case the authors concluded that the virus is not restricted to fungal infected animals, rather that there is a persistent low-level infection in their laboratory stocks. Nonetheless, the observation is intriguing and may suggest a broader relationship between fungal insect pathogens and positive-strand RNA viruses.

Whether it turns out the EV is involved in behavioral manipulation or not, it is a fascinating example of viral adaptation that represents the discovery of the expansion of a major viral lineage to a new kingdom, and understanding how this relationship evolved and is maintained will illuminate new aspects of virus biology.

Methods

Confirming virus in samples by RT-PCR

A liquid culture of *E. muscae* was propagated in Grace's Insect Media supplemented with 5% FBS. One milliliter of cultured cells was spun at 10,000 x g for 5 min to pellet fungal cells. The supernatant was filtered through a 0.22 um syringe filter, and RNA was extracted with Trizol. The pellet was resuspended in 10 mL PBS and vacuum filtered through a Whatman 0.8 um cellulose ester filter to collect cells. This filter was washed four times with 10 mL of PBS. RNA was extracted from all washes with Trizol. Finally the filter paper was equilibrated in 10 mL PBS for 30 minutes and the eluted cells were pelleted and RNA-extracted by Trizol. Additionally, Trizol extraction was performed on our media stocks and 25 CantonS flies from the fly stocks we have used to propagate *in vivo E. muscae* infections.

All RNA samples were reverse transcribed with SuperScript III reverse transcriptase using 150 ng of random hexamer primers per reaction. The RT reaction was heat inactivated at 70C for 15 minutes and 1/10 of the cDNA was used to amplify a 831-bp sequence specific to EV using Taq polymerase. Amplification primers were "GGGTTAGAAGTGTGCGAGAAT" and "GCGACAAGGACTACACGATAAG". Amplicon presence was assayed with a 1% agarose gel with ethidium bromide.

Analysis of *E. muscae* infected *D. melanogaster* RNA-seq

We used RNA-seq data from (Elya et al. 2018) available in the NCBI GEO database under ID GSE111046. We used published read counts for each sample from (Elya et al. 2018) for *E. muscae* and *D. melanogaster*, and determined read counts for EV by aligning reads to all variants of the EM genome using bowtie2 (Langmead and Salzberg 2012).

Transmission Electron Microscopy

To prepare a crude sample of extracellular EV, we pelleted 10 mL of *E. muscae* liquid culture, filtered the supernatant through a 0.2 um syringe filter, and ultra-centrifuged the sample for 2 hours at 25,000 RPM and 4C, using the SW28 swinging bucket rotor. The pelleted material was fixed in 2.5% glutaraldehyde (in 0.1M sodium cacodylate, pH 7.4) and a 1:100 dilution was negative stained with 1% uranyl acetate and imaged on a Tecnai 12 TEM.

To image sections of *E. muscae* cells, pelleted cells were fixed in 2.5% glutaraldehyde in 0.1M sodium cacodylate, pH 7.4 and then embedded in 2% agarose. After washing away fixation buffer with 0.1M sodium cacodylate, samples were treated with 1% osmium tetroxide and 1.6% potassium ferricyanide for 30 minutes, then washed again with 0.1M sodium cacodylate. Fixed and embedded samples were then dehydrated with increasing concentrations of EtOH, followed by pure EtOH and then pure acetone. Increasing concentrations of Eponate resin in acetone (25%, 50%, 75%, then 100%) were infiltrated into the samples for one hour each, followed by pure resin infiltration overnight. Then eponate resin with BDMA accelerant was infiltrated into samples for 5 hours. Samples were embedded into a mold and incubated ta 60C for one week. 70 nm sections of samples were cut and stained with 2% uranyl acetate, followed by Reynolds lead citrate, before imaging on the Tecnai 12 TEM.

Small RNA sequencing

4 mL of *E. muscae* liquid culture was pelleted and RNA was extracted with Trizol. The sample was treated with Turbo DNase and Trizol-extracted again. RNA integrity was confirmed with an RNA 6000 Pico chip on the Agilent 2100 Bioanalyzer. RNA was diluted to 200 ng/ul, and 1 ug (5 ul) was used as input for the Illumina TruSeq small RNA kit (RS-200-0012). The size range was confirmed on the 2100 Bioanalyzer with a HS DNA chip. 204.5 million 50 SR reads were obtained with a HiSeq 4000.
Cutadapt software was used to trim 3' bases with a Phred score <= 10, and to remove the 3' Illumina small RNA adapter. Next, cutadapt was used to select sequences between 17 and 29 bp and these reads were aligned to the EV genome with Hisat2.

Analysis of samples with previously identified Entomophthovirus

Twyford

We obtained reads for the original samples from the NCBI's Sequence Read Archive (SRA): SRR1914527 which contains flies from UK (including Twyford) and SRR1914484 which contains flies from non-UK sources. We aligned reads using bowtie2 with default parameters to a set of 17,826 genes from a preliminary annotation of the *E. muscae* genome filtered to remove regions that cross-align with the *D. melanogaster* genome (identified using blastn with an e-value cutoff of .0000001) and highly conserved fungal genes which have regions that align cross species (e.g. beta-tubulin, histones, ribosomal proteins), we found 823 read pairs that align to *E. muscae* in the Twyford sample while there are 0 in the non-Twyford sample. An additional 575 read pairs from the Twyford sample align discordantly to the *E. muscae* annotation, while 1,220 have a single read from the pair that aligns, reflecting the incomplete and fragmented nature of the current *E. muscae* annotation. In total reads aligning to 1,511 distinct *E. muscae* genes were identified with an average mismatch frequency of .012. Six reads align from SRR1914484 align to the filtered *E. muscae* transcriptome, but have a high number of mismatches demonstrating they are not from *E. muscae*. 807 reads aligned to EV from SRR1914527 with an average mismatch frequency of .0036. Four reads from SRR1914484 aligned to the highly conserved RDRP portion of EV with mismatch patterns suggesting they are from a different virus.

H39

We obtained reads from all 67 samples from the experiment in which H39 was identified (Shi et al. 2016) (NCBI SRA BioProject ID SRP073469) and aligned with the same procedure as for Twyford above. H39 was identified in sample SRR3400838 labeled “Diptera mix”. It contains 13,258 reads that align to EV and 762 reads that align to 342 different *E. muscae* genes. None of the remaining 66 samples appear to contain *E. muscae*. 15 have a handful of reads (between 1 and 18) that align to the filtered *E. muscae* transcript set, but in all cases they contain many mismatches (from 10 to 25 per 100bp) demonstrating that the reads are from another fungal species. Four samples have a small number of reads (1-25) aligning to the conserved RDRP, but with multiple mismatches showing that they are no EV. To confirm that these samples do not contain EV we assembled the reads from these samples using TRINITY and did not find any even fragmentary versions of EV in the assemblies.

Identification of *Delia radicum* and *Musca domestica* Entomophthovirus

We searched NCBI’s Transcriptome Sequence Assembly (TSA) database for additional and identified a series of hits from an individual of *Delia radicum*, a dipteran known as the cabbage fly, infected with *E. muscae* (De Fine Licht, Jensen, and Eilenberg 2017). The hits include an essentially full-length EV annotated as an *E. muscae* transcript (GenBank locus GENB01034640), which we henceforth refer to as DrEV.

We downloaded reads for this paper (NCBI SRA BioProject ID PRJEB10825), which involved the sampling of two infected individuals of the house fly *Musca domestica* and three of *D. radicum*. Spores from the first *M. domestica* sample were used to infect *M. domestica* in the laboratory, and mRNA from three laboratory-infected *M. domestica* were sequenced (A_Md1, A_Md2, and A_Md3) and a fourth was used to inoculate three *in vitro* cultures, from which RNA was also sequenced (yielding mRNA samples A_Gl1, A_Gl2, and A_Gl3). The first part of the process was repeated for a second wild infected *M. domestica* (yielding mRNA samples B_Md1, B_Md2 and B_Md3). Finally mRNA from three wild caught individuals of *D. radicum* were sequenced (yielding mRNA samples C_Dr, D_Dr, E_Dr).

All three wild *E. muscae* infected *D. radicum* samples contain EV. One wild *D. radicum* had 50,000 EV reads, while the other two had only a few hundred. Preliminary alignments of the *M. domestica* samples to *D. melanogaster* EV showed a wide range of EV titres: laboratory *M.*

domestica infected from the first individual had only approximately 100 EV reads each, while laboratory *M. domestica* infected from the second wild individual had several hundred thousand EV reads.

We used reads from *M. domestica* infected by the second wild *M. domestica* sample to assemble to *de novo* transcriptome using Trinity (Grabherr et al. 2011). This transcriptome contains a nearly full length copy of EV which we refer to as MdEV. We aligned the *M. domestica* samples against MdEV confirming the presence of this virus in all six infected *M. domestica*, and its absence from the *in vitro* culture.

Acknowledgements

We are grateful to Joseph DeRisi for helpful discussions about the identification, isolation and characterization of viruses. We thank Danielle Jorgens and Reena Zalpuri at the UC Berkeley Electron Microscopy Lab. This work was funded by an HHMI Investigator award to MBE. CNE was supported by an NSF predoctoral fellowship.

Figure A.1. The genome and phylogenetic placement of Entomophthovirus.

(A) The genome, assembled by Trinity from RNA-sequencing of our in vitro *Entomophthora muscae* culture, encodes a polyprotein with the characteristic open reading frames of a picornavirus, including three coat proteins, a helicase, a protease, and an RNA-dependent RNA polymerase.

(B) A neighbor-joining phylogenetic re-construction of RdRp protein sequences clusters Entomophthorvirus with Twyford virus and Hubei picorna-like virus 39 as a sub-clade of the iflaviruses. Other characterized iflaviruses, including some linked to behavioral manipulation, are shown. The tree was calculated using MEGA7 (Kumar, Stecher, and Tamura 2016) using the Neighbor-Joining method (Zhang and Sun 2008). The evolutionary distances were computed using the Poisson correction method and are in the units of the number of substitutions per site. The optimal tree is shown, drawn to scale using iTOL (Letunic and Bork 2016).

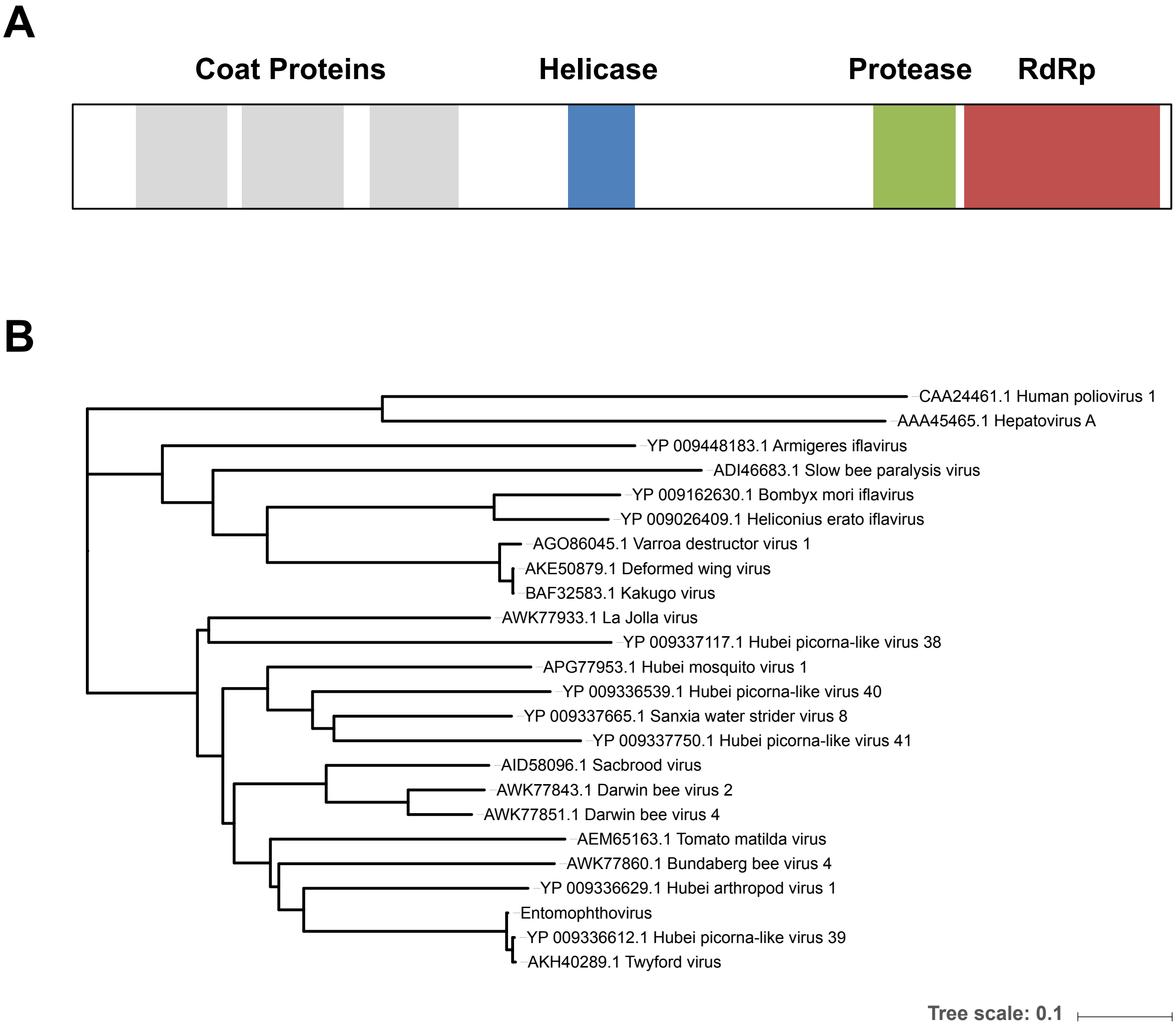

A
Coat Proteins
Helicase
Protease
RdRp
B
CAA24461.1 Human poliovirus 1
AAA45465.1 Hepatovirus A
YP 009448183.1 Armigeres iflavirus
ADI46683.1 Slow bee paralysis virus
YP 009162630.1 Bombyx mori iflavirus
YP 009026409.1 Heliconius erato iflavirus
AGO86045.1 Varroa destructor virus 1
AKE50879.1 Deformed wing virus
BAF32583.1 Kakugo virus
AWK77933.1 La Jolla virus
YP 009337117.1 Hubei picorna-like virus 38
APG77953.1 Hubei mosquito virus 1
YP 009336539.1 Hubei picorna-like virus 40
YP 009337665.1 Sanxia water strider virus 8
YP 009337750.1 Hubei picorna-like virus 41
AID58096.1 Sacbrood virus
AWK77843.1 Darwin bee virus 2
AWK77851.1 Darwin bee virus 4
AEM65163.1 Tomato matilda virus
AWK77860.1 Bundaberg bee virus 4
YP 009336629.1 Hubei arthropod virus 1
Entomophthovirus
YP 009336612.1 Hubei picorna-like virus 39
AKH40289.1 Twyford virus
Tree scale: 0.1

Figure A.2. Entomophthovirus in *in vivo Entomophthora muscae* infections of *Drosophila melanogaster*.

To characterize the growth of virus in animals infected with *E. muscae* aligned reads from mRNA-seq data of (Elya et al. 2018) to Entomophthovirus. Samples were from individual whole flies exposed to *E. muscae* at 24, 28, 72, 96 and 120 hours after infection as well as controls. Plotted are the fraction of total reads that aligned to *D. melanogaster* mRNAs, to E. muscae transcripts or annotated genes, and to Entomophthovirus.

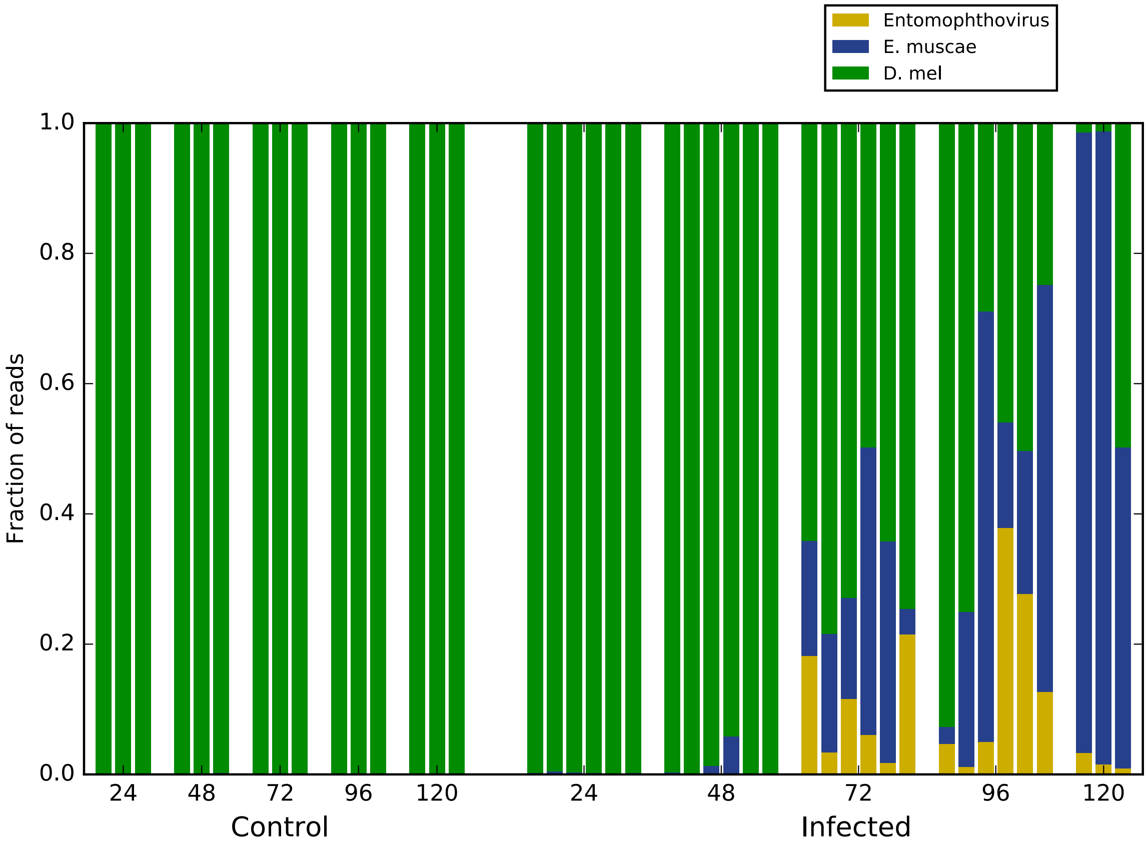

Entomophthovirus
E. muscae
D. mel
1.0
0.8
0.6
0.4
0.2
0.0
Fraction of reads
24
48
72
96
120
24
48
72
96
120
Control
Infected

Figure A.3. The presence of Entomophthovirus in *in vitro Entomophthora muscae* culture.

(A) RT-PCR with primers specific for the Entomophthovirus genome shows the virus in the supernatant and cellular fraction of in vitro culture of E. muscae. Viral signal is lac,king from the media we use for in vitro culture, the *Drosophila* stocks used to propagate the infection in lab, and washes of the cellular fraction.

(B) Transmission electron microscopy of virus collected from the supernatant of in vitro cultures by ultra-centrifugation and negative stained by uranyl acetate. The viral particles (red arrows) have a tight size distribution and the expected diameter (~30 nm) of an iflavirus capsid.

(C) Transmission electron microscopy of cellular sections of *E. muscae*, with a double-contrast staining of uranyl acetate and lead citrate. Electron-dense particles (red arrows) consistent with the size distribution of an iflavirus capsid are abundant in a fraction of cellular sections and are localized cytoplasmically. Examples of E. muscae nucleus, mitochondria (M), and chromatin (Ch) are also marked.

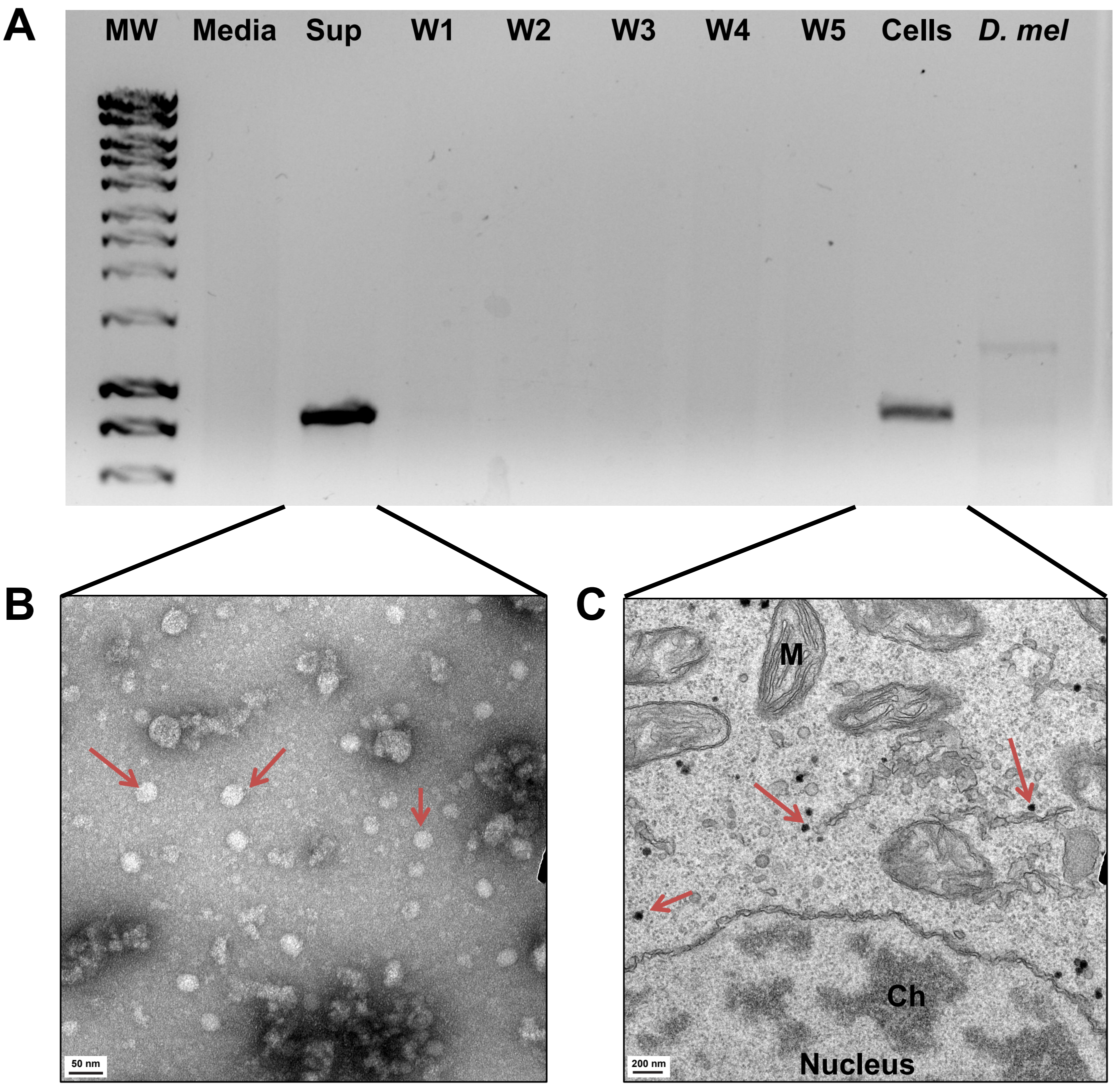
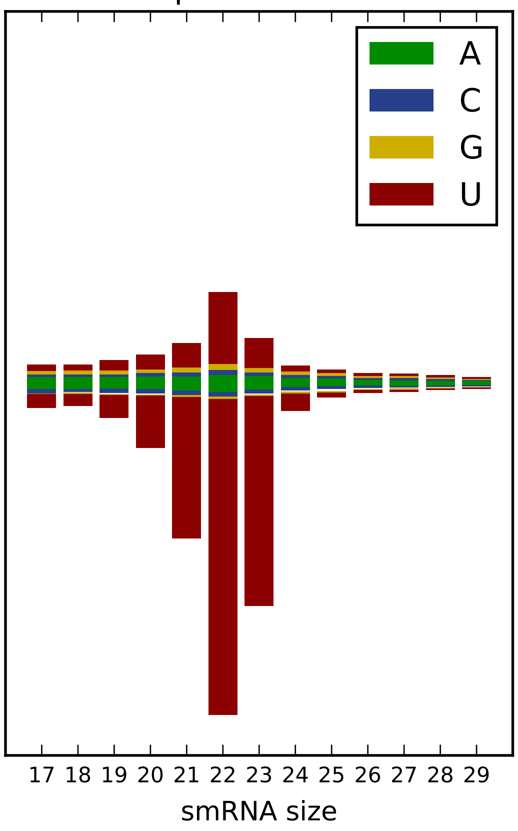

A
MW
Media
Sup
W1
W2
W3
W4
W5
Cells
D. mel
B
50 nm
C
M
Ch
Nucleus
200 nm

Figure A.4. Small RNAs from Entomophthovirus have a characteristic 5' U bias

We sequenced 204.5 million small RNAs from our *E. muscae in vitro* culture. After quality trimming and adapter removal (using cutadapt (Martin 2011)), the 17-29 bp fraction was aligned to the Entomophthovirus genome (Hisat2 (Kim, Langmead, and Salzberg 2015)). Following the presentation of data from (Webster et al. 2015), the length and 5' nucleotide bias of aligned reads is shown, with the data from (Webster et al. 2015) for Twyford virus in wild *Drosophila* shown alongside. In both, the aligned reads are mostly 21-23 nt, with a negative strand bias and a strong 5' U bias.


Entomophthovirus in vitro

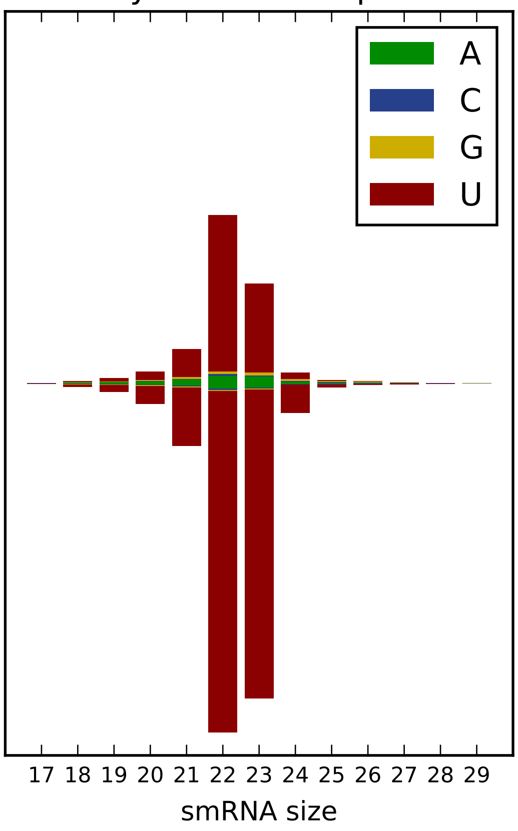
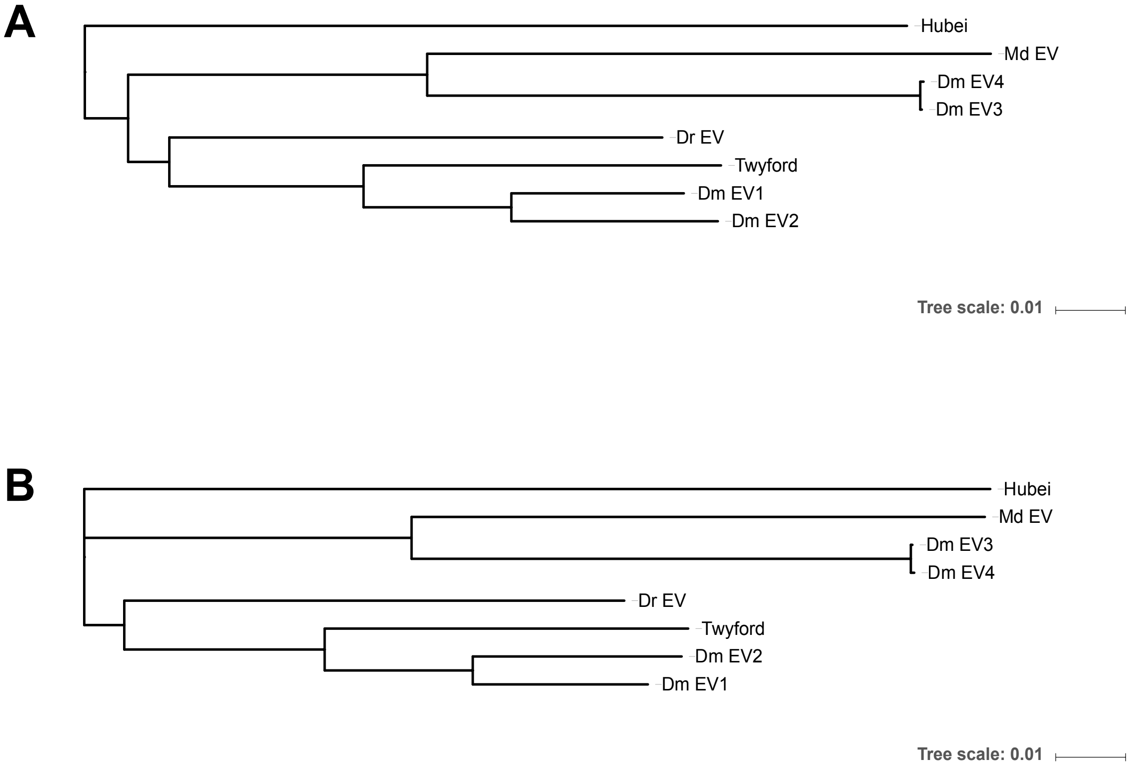

Twyford in Drosophila

Figure A.5. Relationship of *Entomophthora muscae* associated Entomophthovirus variants.

Evolutionary relationships of eight Entomophthovirus samples identified from *Entomophthora muscae* infected *Drosophila melanogaster* (Dm and Twyford), *Musca domestica* (Md), *Delia radicum* (Dr) and an unknown dipteran (Hubei) based on (A) nucleotide sequences and (B) inferred amino acid sequences. Trees were calculated using MEGA7 (Kumar, Stecher, and Tamura 2016) using the Neighbor-Joining method (Zhang and Sun 2008). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura, Nei, and Kumar 2004) and are in the units of the number of substitutions per site. The optimal tree is shown, drawn to scale using iTOL (Letunic and Bork 2016).


A
Hubei
Md EV
Dm EV4
Dm EV3
Dr EV
Twyford
Dm EV1
Dm EV2
Tree scale: 0.01
B
Hubei
Md EV
Dm EV3
Dm EV4
Dr EV
Twyford
Dm EV2
Dm EV1
Tree scale: 0.01

References

Abegglen, Lisa M., Aleah F. Caulin, Ashley Chan, Kristy Lee, Rosann Robinson, Michael S. Campbell, Wendy K. Kiso, et al. 2015. "Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans." *JAMA: The Journal of the American Medical Association* 314 (17): 1850–60.

Adl, Sina M., Alastair G. B. Simpson, Christopher E. Lane, Julius Lukeš, David Bass, Samuel S. Bowser, Matthew W. Brown, et al. 2012. "The Revised Classification of Eukaryotes." *The Journal of Eukaryotic Microbiology* 59 (5): 429–93.

Adrian Gherman, Erica E. Davis, and Nicholas Katsanis. 2006. "The Ciliary Proteome Database: An Integrated Community Resource for the Genetic and Functional Dissection of Cilia." *Nature Genetics*.

Albertin, Caroline B., Oleg Simakov, Therese Mitros, Z. Yan Wang, Judit R. Pungor, Eric Edsinger-Gonzales, Sydney Brenner, Clifton W. Ragsdale, and Daniel S. Rokhsar. 2015. "The Octopus Genome and the Evolution of Cephalopod Neural and Morphological Novelties." *Nature* 524 (7564): 220–24.

Alegado, Rosanna A., Laura W. Brown, Shugeng Cao, Renee K. Dermenjian, Richard Zuzow, Stephen R. Fairclough, Jon Clardy, and Nicole King. 2012. "A Bacterial Sulfonolipid Triggers Multicellular Development in the Closest Living Relatives of Animals." *ELife* 1 (October): e00013.

Alten, Leonie, Karin Schuster-Gossler, Anja Beckers, Stephanie Groos, Bärbel Ulmer, Jan Hegermann, Matthias Ochs, and Achim Gossler. 2012. "Differential Regulation of Node Formation, Nodal Ciliogenesis and Cilia Positioning by Noto and Foxj1." *Development* 139 (7): 1276–84.

Aravind, L., Vivek Anantharaman, Santhanam Balaji, M. Mohan Babu, and Lakshminarayan M. Iyer. 2005. "The Many Faces of the Helix-Turn-Helix Domain: Transcription Regulation and Beyond." *FEMS Microbiology Reviews* 29 (2): 231–62.

Aravind, L., and D. Landsman. 1998. "AT-Hook Motifs Identified in a Wide Variety of DNA-Binding Proteins." *Nucleic Acids Research* 26 (19): 4413–21.

Arendt, Detlev. 2008. "The Evolution of Cell Types in Animals: Emerging Principles from Molecular Studies." *Nature Reviews. Genetics* 9 (11): 868–82.

Arnaiz, Olivier, Agata Malinowska, Catherine Klotz, Linda Sperling, Michal Dadlez, France Koll, and Jean Cohen. 2009. "Cildb: A Knowledgebase for Centrosomes and Cilia." *Database: The Journal of Biological Databases and Curation* 2009 (December): bap022.

Ashique, Amir M., Youngshik Choe, Mattias Karlen, Scott R. May, Khanhky Phamluong, Mark J. Solloway, Johan Ericson, and Andrew S. Peterson. 2009. "The Rfx4 Transcription Factor Modulates Shh Signaling by Regional Control of Ciliogenesis." *Science Signaling* 2 (95): ra70.

Babu, M. M., L. M. Iyer, S. Balaji, and L. Aravind. 2006. "The Natural History of the WRKY–GCM1 Zinc Fingers and the Relationship between Transcription Factors and Transposons." *Nucleic Acids Research*. https://academic.oup.com/nar/article-abstract/34/22/6505/3112387.

Badis, Gwenael, Esther T. Chan, Harm van Bakel, Lourdes Pena-Castillo, Desiree Tillo, Kyle Tsui, Clayton D. Carlson, et al. 2008. "A Library of Yeast Transcription Factor Motifs Reveals a Widespread Function for Rsc3 in Targeting Nucleosome Exclusion at Promoters." *Molecular Cell* 32 (6): 878–87.

Baker, Christopher R., Brian B. Tuch, and Alexander D. Johnson. 2011. "Extensive DNA-Binding Specificity Divergence of a Conserved Transcription Regulator." *Proceedings of the National Academy of Sciences of the United States of America* 108 (18): 7493–98.

Balaji, S., M. Madan Babu, Lakshminarayan M. Iyer, and L. Aravind. 2005. "Discovery of the Principal Specific Transcription Factors of Apicomplexa and Their Implication for the Evolution of the AP2-Integrase DNA Binding Domains." *Nucleic Acids Research* 33 (13): 3994–4006.

Bałzy, S. 1984. "On Rhizoids of Entomophthora Muscae (Cohn) Fresenius (Entomophthorales: Entomophthoraceae)." *Mycotaxon* 19: 397–407.

Banerji, J., S. Rusconi, and W. Schaffner. 1981. "Expression of a Beta-Globin Gene Is Enhanced by Remote SV40 DNA Sequences." *Cell* 27 (2 Pt 1): 299–308.

Bell, Graham, and Arne O. Mooers. 1997. "Size and Complexity among Multicellular Organisms." *Biological Journal of the Linnean Society. Linnean Society of London* 60 (3): 345–63.

Bloodgood, Robert A. 2010. "Sensory Reception Is an Attribute of Both Primary Cilia and Motile Cilia." *Journal of Cell Science* 123 (Pt 4): 505–9.

Bonnafe, E., M. Touka, A. AitLounis, D. Baas, E. Barras, C. Ucla, A. Moreau, et al. 2004. "The Transcription Factor RFX3 Directs Nodal Cilium Development and Left-Right Asymmetry Specification." *Molecular and Cellular Biology* 24 (10): 4417–27.

Booth, David S., and Nicole King. 2020. "Genome Editing Enables Reverse Genetics of Multicellular Development in the Choanoflagellate Salpingoeca Rosetta." *ELife* 9 (June): e56193.

Booth, David S., Heather Szmidt-Middleton, and Nicole King. 2018. "Transfection of Choanoflagellates Illuminates Their Cell Biology and the Ancestry of Animal Septins." *Molecular Biology of the Cell* 29 (25): 3026–38.

Brokaw, C. J. 1960. "Decreased Adenosine Triphosphatase Acivity of Flagella from a Paralyzed Mutant of Chlamydomonas Moewusii." *Experimental Cell Research* 19 (March): 430–32.

Brown, Seth J., Michael D. Cole, and Albert J. Erives. 2008. "Evolution of the Holozoan Ribosome Biogenesis Regulon." *BMC Genomics* 9 (September): 442.

Brunet, Thibaut, Marvin Albert, William Roman, Maxwell C. Coyle, Danielle C. Spitzer, and Nicole King. 2021. "A Flagellate-to-Amoeboid Switch in the Closest Living Relatives of Animals." *ELife* 10 (January). https://doi.org/10.7554/eLife.61037.

Brunet, Thibaut, and Nicole King. 2017. "The Origin of Animal Multicellularity and Cell Differentiation." *Developmental Cell* 43 (2): 124–40.

Bugeja, Hayley E., Michael J. Hynes, and Alex Andrianopoulos. 2010. "The RFX Protein RfxA Is an Essential Regulator of Growth and Morphogenesis in Penicillium Marneffei." *Eukaryotic Cell* 9 (4): 578–91.

Buss, Leo W. 1988. *The Evolution of Individuality*. Princeton University Press.

Cao, Chen, Laurence A. Lemaire, Wei Wang, Peter H. Yoon, Yoolim A. Choi, Lance R. Parsons, John C. Matese, Wei Wang, Michael Levine, and Kai Chen. 2019. “Comprehensive Single-Cell Transcriptome Lineages of a Proto-Vertebrate.” *Nature* 571 (7765): 349–54.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73.

Carr, Martin, Daniel J. Richter, Parinaz Fozouni, Timothy J. Smith, Alexandra Jeuck, Barry S. C. Leadbeater, and Frank Nitsche. 2017. “A Six-Gene Phylogeny Provides New Insights into Choanoflagellate Evolution.” *Molecular Phylogenetics and Evolution* 107 (February): 166–78.

Carvalho-Santos, Zita, Juliette Azimzadeh, José B. Pereira-Leal, and Mónica Bettencourt-Dias. 2011. “Evolution: Tracing the Origins of Centrioles, Cilia, and Flagella.” *The Journal of Cell Biology* 194 (2): 165–75.

Castro, Wilson, Sonia T. Chelbi, Charlène Niogret, Cristina Ramon-Barros, Suzanne P. M. Welten, Kevin Osterheld, Haiping Wang, et al. 2018. “The Transcription Factor Rfx7 Limits Metabolism of NK Cells and Promotes Their Maintenance and Immunity.” *Nature Immunology* 19 (8): 809–20.

Chan, Yingguang Frank, Melissa E. Marks, Felicity C. Jones, Guadalupe Villarreal Jr, Michael D. Shapiro, Shannon D. Brady, Audrey M. Southwick, et al. 2010. “Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer.” *Science* 327 (5963): 302–5.

Chapman, D. M. 1974. “Cnidarian Histology. Coelenterate Biology. Reviews and New Perspectives.” New York: Academic.

Choksi, Semil P., Gilbert Lauter, Peter Swoboda, and Sudipto Roy. 2014. “Switching on Cilia: Transcriptional Networks Regulating Ciliogenesis.” *Development* 141 (7): 1427–41.

Chu, Jeffrey S. C., David L. Baillie, and Nansheng Chen. 2010. “Convergent Evolution of RFX Transcription Factors and Ciliary Genes Predated the Origin of Metazoans.” *BMC Evolutionary Biology* 10 (1): 130.

Chung, Mei-I, Taejoon Kwon, Fan Tu, Eric R. Brooks, Rakhi Gupta, Matthew Meyer, Julie C. Baker, Edward M. Marcotte, and John B. Wallingford. 2014. “Coordinated Genomic Control of Ciliogenesis and Cell Movement by RFX2.” *ELife* 3 (January): e01439.

Chung, Mei-I, Sara M. Peyrot, Sarah LeBoeuf, Tae Joo Park, Kriston L. McGary, Edward M. Marcotte, and John B. Wallingford. 2012. “RFX2 Is Broadly Required for Ciliogenesis during Vertebrate Development.” *Developmental Biology* 363 (1): 155–65.

Cohn, F. 1855. “Empusa Muscae Und Die Krankheit Der Stubenfliegen. Ein Beitrag Zur Lehre von Den Durch Parasitische Pilze Charakterisierten Epidemien.” *Nova Acta Academiae Caesareae Leopoldino-Carolinae Germanicae Naturae Curiosorum* 25: 299–360.

Cowling, Victoria H., Sanjay Chandriani, Michael L. Whitfield, and Michael D. Cole. 2006. “A Conserved Myc Protein Domain, MBIV, Regulates DNA Binding, Apoptosis, Transformation, and G2 Arrest.” *Molecular and Cellular Biology* 26 (11): 4226–39.

Coyle, Maxwell C., Adia M. Tajima, Fredrick Leon, Semil P. Choksi, Ally Yang, Sarah Espinoza, Timothy R. Hughes, Jeremy F. Reiter, David S. Booth, and Nicole King. 2023. “An RFX Transcription Factor Regulates Ciliogenesis in the Closest Living Relatives of Animals.” *Current Biology: CB*, August. https://doi.org/10.1016/j.cub.2023.07.022.

Dam, Teunis Jp van, Gabrielle Wheway, Gisela G. Slaats, SYSCILIA Study Group, Martijn A. Huynen, and Rachel H. Giles. 2013. “The SYSCILIA Gold Standard (SCGSv1) of Known Ciliary Components and Its Applications within a Systems Biology Consortium.” *Cilia* 2 (1): 7.

Dayel, Mark J., Rosanna A. Alegado, Stephen R. Fairclough, Tera C. Levin, Scott A. Nichols, Kent McDonald, and Nicole King. 2011. “Cell Differentiation and Morphogenesis in the Colony-Forming Choanoflagellate Salpingoeca Rosetta.” *Developmental Biology* 357 (1): 73–82.

Dayel, Mark J., and Nicole King. 2014. “Prey Capture and Phagocytosis in the Choanoflagellate Salpingoeca Rosetta.” *PloS One* 9 (5): e95577.

De Fine Licht, Henrik H., Annette B. Jensen, and Jørgen Eilenberg. 2017. “Comparative Transcriptomics Reveal Host-Specific Nucleotide Variation in Entomophthoralean Fungi.” *Molecular Ecology* 26 (7): 2092–2110.

Degnan, Bernard M., Michel Vervoort, Claire Larroux, and Gemma S. Richards. 2009. “Early Evolution of Metazoan Transcription Factors.” *Current Opinion in Genetics & Development* 19 (6): 591–99.

Dheilly, Nolwenn M., Fanny Maure, Marc Ravallec, Richard Galinier, Josée Doyon, David Duval, Lucas Leger, et al. 2015. “Who Is the Puppet Master? Replication of a Parasitic Wasp-Associated Virus Correlates with Host Behaviour Manipulation.” *Proceedings. Biological Sciences / The Royal Society* 282 (1803): 20142773.

Didon, Lukas, Rachel K. Zwick, Ion Wa Chao, Matthew S. Walters, Rui Wang, Neil R. Hackett, and Ronald G. Crystal. 2013. “RFX3 Modulation of FOXJ1 Regulation of Cilia Genes in the Human Airway Epithelium.” *Respiratory Research* 14 (July): 70.

Dobi, Krista C., and Fred Winston. 2007. “Analysis of Transcriptional Activation at a Distance in Saccharomyces Cerevisiae.” *Molecular and Cellular Biology* 27 (15): 5575–86.

Dorighi, Kristel M., Tomek Swigut, Telmo Henriques, Natarajan V. Bhanu, Benjamin S. Scruggs, Nataliya Nady, Christopher D. Still 2nd, Benjamin A. Garcia, Karen Adelman, and Joanna Wysocka. 2017. “Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation.” *Molecular Cell* 66 (4): 568-576.e4.

Dubruille, Raphaelle, Anne Laurençon, Camille Vandaele, Emiko Shishido, Madeleine Coulon-Bublex, Peter Swoboda, Pierre Couble, Maurice Kernan, and Bénédicte Durand. 2002. “Drosophila Regulatory Factor X Is Necessary for Ciliated Sensory Neuron Differentiation.” *Development* 129 (23): 5487–98.

Dunn, Casey W., Andreas Hejnol, David Q. Matus, Kevin Pang, William E. Browne, Stephen A. Smith, Elaine Seaver, et al. 2008. “Broad Phylogenomic Sampling Improves Resolution of the Animal Tree of Life.” *Nature* 452 (7188): 745–49.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97.

Efimenko, Evgeni, Kerry Bubb, Ho Yi Mak, Ted Holzman, Michel R. Leroux, Gary Ruvkun, James H. Thomas, and Peter Swoboda. 2005. “Analysis of Xbx Genes in C. Elegans.” *Development* 132 (8): 1923–34.

Eilers, Martin, and Robert N. Eisenman. 2008. “Myc’s Broad Reach.” *Genes & Development* 22 (20): 2755–66.

Eisthen, Heather L., and Kevin R. Theis. 2016. “Animal-Microbe Interactions and the Evolution of Nervous Systems.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371 (1685): 20150052.

El Zein, Loubna, Aouatef Ait-Lounis, Laurette Morlé, Joëlle Thomas, Brigitte Chhin, Nathalie Spassky, Walter Reith, and Bénédicte Durand. 2009. “RFX3 Governs Growth and Beating Efficiency of Motile Cilia in Mouse and Controls the Expression of Genes Involved in Human Ciliopathies.” *Journal of Cell Science* 122 (Pt 17): 3180–89.

Elya, Carolyn, Tin Ching Lok, Quinn E. Spencer, Hayley McCausland, Ciera C. Martinez, and Michael Eisen. 2018. “Robust Manipulation of the Behavior of Drosophila Melanogaster by a Fungal Pathogen in the Laboratory.” *ELife* 7 (July). https://doi.org/10.7554/eLife.34414.

Emery, P., M. Strubin, K. Hofmann, P. Bucher, B. Mach, and W. Reith. 1996. “A Consensus Motif in the RFX DNA Binding Domain and Binding Domain Mutants with Altered Specificity.” *Molecular and Cellular Biology* 16 (8): 4486–94.

Erwin, Douglas H. 2020. “Evolutionary Dynamics of Gene Regulation.” *Current Topics in Developmental Biology* 139 (March): 407–31.

Fairclough, Stephen R., Zehua Chen, Eric Kramer, Qiandong Zeng, Sarah Young, Hugh M. Robertson, Emina Begovic, et al. 2013. “Premetazoan Genome Evolution and the Regulation of Cell Differentiation in the Choanoflagellate Salpingoeca Rosetta.” *Genome Biology* 14 (2): R15.

Feng, Chenzhuo, Wenhao Xu, and Zhiyi Zuo. 2009. “Knockout of the Regulatory Factor X1 Gene Leads to Early Embryonic Lethality.” *Biochemical and Biophysical Research Communications* 386 (4): 715–17.

Finn, Robert D., Teresa K. Attwood, Patricia C. Babbitt, Alex Bateman, Peer Bork, Alan J. Bridge, Hsin-Yu Chang, et al. 2016. “InterPro in 2017—beyond Protein Family and Domain Annotations.” *Nucleic Acids Research* 45 (D1): D190–99.

Forsythe, Paul, Wolfgang A. Kunze, and John Bienenstock. 2012. “On Communication between Gut Microbes and the Brain.” *Current Opinion in Gastroenterology* 28 (6): 557–62.

Frikstad, Kari-Anne M., Elisa Molinari, Marianne Thoresen, Simon A. Ramsbottom, Frances Hughes, Stef J. F. Letteboer, Sania Gilani, et al. 2019. “A CEP104-CSPP1 Complex Is Required for Formation of Primary Cilia Competent in Hedgehog Signaling.” *Cell Reports* 28 (7): 1907-1922.e6.

Fritz-Laylin, Lillian K. 2020. “The Evolution of Animal Cell Motility.” *Current Biology: CB* 30 (10): R477–82.

Fujiyuki, Tomoko, Hideaki Takeuchi, Masato Ono, Seii Ohka, Tetsuhiko Sasaki, Akio Nomoto, and Takeo Kubo. 2005. “Kakugo Virus from Brains of Aggressive Worker Honeybees.” *Advances in Virus Research* 65: 1–27.

Gaiti, Federico, Katia Jindrich, Selene L. Fernandez-Valverde, Kathrein E. Roper, Bernard M. Degnan, and Miloš Tanurdžić. 2017. “Landscape of Histone Modifications in a Sponge Reveals the Origin of Animal Cis-Regulatory Complexity.” *ELife* 6 (April). https://doi.org/10.7554/eLife.22194.

Gajiwala, K. S., H. Chen, F. Cornille, B. P. Roques, W. Reith, B. Mach, and S. K. Burley. 2000. “Structure of the Winged-Helix Protein HRFX1 Reveals a New Mode of DNA Binding.” *Nature* 403 (6772): 916–21.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52.

Grandori, C., S. M. Cowley, L. P. James, and R. N. Eisenman. 2000. "The Myc/Max/Mad Network and the Transcriptional Control of Cell Behavior." *Annual Review of Cell and Developmental Biology* 16: 653–99.

Grau-Bové, Xavier, Guifré Torruella, Stuart Donachie, Hiroshi Suga, Guy Leonard, Thomas A. Richards, and Iñaki Ruiz-Trillo. 2017. "Dynamics of Genomic Innovation in the Unicellular Ancestry of Animals." *ELife* 6 (July). https://doi.org/10.7554/eLife.26036.

Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." *Systematic Biology* 59 (3): 307–21.

Hajek, Ann E., Bernard Papierok, and Jørgen Eilenberg. 2012. "Methods for Study of the Entomophthorales." *Manual of Techniques in Invertebrate Pathology*, 285–316.

Hao, Binghua, Cornelius J. Clancy, Shaoji Cheng, Suresh B. Raman, Kenneth A. Iczkowski, and M. Hong Nguyen. 2009. "Candida Albicans RFX2 Encodes a DNA Binding Protein Involved in DNA Damage Responses, Morphogenesis, and Virulence." *Eukaryotic Cell* 8 (4): 627–39.

Hare, Emily E., Brant K. Peterson, Venky N. Iyer, Rudolf Meier, and Michael B. Eisen. 2008. "Sepsid Even-Skipped Enhancers Are Functionally Conserved in Drosophila despite Lack of Sequence Conservation." *PLoS Genetics* 4 (6): e1000106.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4): 576–89.

Hoover, Kelli, Michael Grove, Matthew Gardner, David P. Hughes, James McNeil, and James Slavicek. 2011. "A Gene for an Extended Phenotype." *Science* 333 (6048): 1401.

Huang, M., Z. Zhou, and S. J. Elledge. 1998. "The DNA Replication and Damage Checkpoint Pathways Induce Transcription by Inhibition of the Crt1 Repressor." *Cell* 94 (5): 595–605.

Hughes, D. P., J. P. M. Araújo, R. G. Loreto, L. Quevillon, C. de Bekker, and H. C. Evans. 2016. "From So Simple a Beginning: The Evolution of Behavioral Manipulation by Fungi." *Advances in Genetics* 94 (February): 437–69.

Hulett, Ryan E., Julian O. Kimura, D. Marcela Bolaños, Yi-Jyun Luo, Carlos Rivera-López, Lorenzo Ricci, and Mansi Srivastava. 2023. "Acoel Single-Cell Atlas Reveals Expression Dynamics and Heterogeneity of Adult Pluripotent Stem Cells." *Nature Communications* 14 (1): 2612.

Imbeault, Michaël, Pierre-Yves Helleboid, and Didier Trono. 2017. "KRAB Zinc-Finger Proteins Contribute to the Evolution of Gene Regulatory Networks." *Nature* 543 (7646): 550–54.

Jianchun Chen, Heather J. Knowles, Jennifer L. Herbert, and Brian P. Hackett. 1998. "Mutation of the Mouse Hepatocyte Nuclear Factor/Forkhead Homologue 4 Gene Results in an Absence of Cilia and Random Left-Right Asymmetry." *The Journal of Clinical Investigation* 233 (5321): 575–575.

Johnston, J. R., and R. K. Mortimer. 1959. "Use of Snail Digestive Juice in Isolation of Yeast Spore Tetrads." *Journal of Bacteriology* 78 (2): 292.

Joho, K. E., M. K. Darby, E. T. Crawford, and D. D. Brown. 1990. “A Finger Protein Structurally Similar to TFIIIA That Binds Exclusively to 5S RNA in Xenopus.” *Cell* 61 (2): 293–300.

Jolma, Arttu, and Jussi Taipale. 2011. “Methods for Analysis of Transcription Factor DNA-Binding Specificity In Vitro.” *Sub-Cellular Biochemistry* 52: 155–73.

Jolma, Arttu, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, et al. 2013. “DNA-Binding Specificities of Human Transcription Factors.” *Cell* 152 (1–2): 327–39.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. “ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates.” *Nature Methods* 14 (6): 587–89.

Katoh, Kazutaka, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research* 30 (14): 3059–66.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Katsuma, Susumu, Yasue Koyano, Wonkyung Kang, Ryuhei Kokusho, Shizuo George Kamita, and Toru Shimada. 2012. “The Baculovirus Uses a Captured Host Phosphatase to Induce Enhanced Locomotory Activity in Host Caterpillars.” *PLoS Pathogens* 8 (4): e1002644.

Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. “HISAT: A Fast Spliced Aligner with Low Memory Requirements.” *Nature Methods* 12 (4): 357–60.

Kin, Koryu, Zhi-Hui Chen, Gillian Forbes, and Pauline Schaap. 2022. “Evolution of a Novel Cell Type in Dictyostelia Required Gene Duplication of a CudA-like Transcription Factor.” *Current Biology: CB* 32 (2): 428-437.e4.

King, Nicole. 2004. “The Unicellular Ancestry of Animal Development.” *Developmental Cell* 7 (3): 313–25.

King, Nicole, and Antonis Rokas. 2017. “Embracing Uncertainty in Reconstructing Early Animal Evolution.” *Current Biology: CB* 27 (19): R1081–88.

King, Nicole, M. Jody Westbrook, Susan L. Young, Alan Kuo, Monika Abedin, Jarrod Chapman, Stephen Fairclough, et al. 2008. “The Genome of the Choanoflagellate Monosiga Brevicollis and the Origin of Metazoans.” *Nature* 451 (7180): 783–88.

Kistler, W. Stephen, Dominique Baas, Sylvain Lemeille, Marie Paschaki, Queralt Seguin-Estevez, Emmanuèle Barras, Wenli Ma, et al. 2015. “RFX2 Is a Major Transcriptional Regulator of Spermiogenesis.” *PLoS Genetics* 11 (7): e1005368.

Kożyczkowska, Aleksandra, Sebastián R. Najle, Eduard Ocaña-Pallarès, Cristina Aresté, Victoria Shabardina, Patricia S. Ara, Iñaki Ruiz-Trillo, and Elena Casacuberta. 2021. “Stable Transfection in Protist Corallochytrium Limacisporum Identifies Novel Cellular Features among Unicellular Animals Relatives.” *Current Biology: CB* 31 (18): 4104-4110.e5.

Kramer, J. P., and D. C. Steinkraus. 1981. “Culture of Entomophthora Muscae in Vivo and Its Infectivity for Six Species of Muscoid Flies.” *Mycopathologia* 76 (3): 139–43.

Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. “MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.” *Molecular Biology and Evolution* 33 (7): 1870–74.

Kvon, Evgeny Z., Olga K. Kamneva, Uirá S. Melo, Iros Barozzi, Marco Osterwalder, Brandon J. Mannion, Virginie Tissières, et al. 2016. “Progressive Loss of Function in a Limb Enhancer during Snake Evolution.” *Cell* 167 (3): 633-642.e11.

Lam, Kathy N., Harm van Bakel, Atina G. Cote, Anton van der Ven, and Timothy R. Hughes. 2011. “Sequence Specificity Is Obtained from the Majority of Modular C2H2 Zinc-Finger Arrays.” *Nucleic Acids Research* 39 (11): 4680–90.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. “The Human Transcription Factors.” *Cell* 175 (2): 598–99.

Lambert, Samuel A., Ally W. H. Yang, Alexander Sasse, Gwendolyn Cowley, Mihai Albu, Mark X. Caddick, Quaid D. Morris, Matthew T. Weirauch, and Timothy R. Hughes. 2019. “Similarity Regression Predicts Evolution of Transcription Factor Sequence Specificity.” *Nature Genetics* 51 (6): 981–89.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Larkins, Christine E., Gladys D. Gonzalez Aviles, Michael P. East, Richard A. Kahn, and Tamara Caspary. 2011. “Arl13b Regulates Ciliogenesis and the Dynamic Localization of Shh Signaling Proteins.” *Molecular Biology of the Cell* 22 (23): 4694–4703.

Larroux, Claire, Graham N. Luke, Peter Koopman, Daniel S. Rokhsar, Sebastian M. Shimeld, and Bernard M. Degnan. 2008. “Genesis and Expansion of Metazoan Transcription Factor Gene Classes.” *Molecular Biology and Evolution* 25 (5): 980–96.

Leadbeater, Barry S. C. 2015. *The Choanoflagellates*. Cambridge University Press.

Lemeille, Sylvain, Marie Paschaki, Dominique Baas, Laurette Morlé, Jean-Luc Duteyrat, Aouatef Ait-Lounis, Emmanuèle Barras, et al. 2020. “Interplay of RFX Transcription Factors 1, 2 and 3 in Motile Ciliogenesis.” *Nucleic Acids Research* 48 (16): 9019–36.

Letunic, Ivica, and Peer Bork. 2016. “Interactive Tree of Life (ITOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees.” *Nucleic Acids Research* 44 (W1): W242-5.

———. 2021. “Interactive Tree Of Life (ITOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation.” *Nucleic Acids Research* 49 (W1): W293–96.

Levin, Tera C., and Nicole King. 2013. “Evidence for Sex and Recombination in the Choanoflagellate Salpingoeca Rosetta.” *Current Biology: CB* 23 (21): 2176–80.

Levine, Michael. 2010. “Transcriptional Enhancers in Animal Development and Evolution.” *Current Biology: CB* 20 (17): R754-63.

Levine, Michael, Claudia Cattoglio, and Robert Tjian. 2014. “Looping Back to Leap Forward: Transcription Enters a New Era.” *Cell* 157 (1): 13–25.

Levine, Michael, and Robert Tjian. 2003. “Transcription Regulation and Animal Diversity.” *Nature* 424 (6945): 147–51.

Levy, Shani, Anamaria Elek, Xavier Grau-Bové, Simón Menéndez-Bravo, Marta Iglesias, Amos Tanay, Tali Mass, and Arnau Sebé-Pedrós. 2021. “A Stony Coral Cell Atlas Illuminates the

Molecular and Cellular Basis of Coral Symbiosis, Calcification, and Immunity.” *Cell* 184 (11): 2973-2987.e18.

Libersat, F. 2003. “Wasp Uses Venom Cocktail to Manipulate the Behavior of Its Cockroach Prey.” *Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology* 189 (7): 497–508.

López-Escardó, David, Xavier Grau-Bové, Amy Guillaumet-Adkins, Marta Gut, Michael E. Sieracki, and Iñaki Ruiz-Trillo. 2019. “Reconstruction of Protein Domain Evolution Using Single-Cell Amplified Genomes of Uncultured Choanoflagellates Sheds Light on the Origin of Animals.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 374 (1786): 20190088.

Mackie, G. O. 1970. “Neuroid Conduction and the Evolution of Conducting Tissues.” *The Quarterly Review of Biology* 45 (4): 319–32.

Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021. “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.” *Molecular Biology and Evolution* 38 (10): 4647–54.

Manojlovic, Zarko, Ryan Earwood, Akiko Kato, Branko Stefanovic, and Yoichi Kato. 2014. “RFX7 Is Required for the Formation of Cilia in the Neural Tube.” *Mechanisms of Development* 132 (May): 28–37.

Margulis, Lynn. 1992. *Symbiosis in Cell Evolution*. W. H. Freeman.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10–12.

Medina, Edgar M., and Nicolas E. Buchler. 2020. “Chytrid Fungi.” *Current Biology: CB* 30 (10): R516–20.

Mendoza, Alex de, and Arnau Sebé-Pedrós. 2019. “Origin and Evolution of Eukaryotic Transcription Factors.” *Current Opinion in Genetics & Development* 58–59 (October): 25–32.

Mendoza, Alex de, Arnau Sebé-Pedrós, Martin Sebastijan Šestak, Marija Matejcic, Guifré Torruella, Tomislav Domazet-Loso, and Iñaki Ruiz-Trillo. 2013. “Transcription Factor Evolution in Eukaryotes and the Assembly of the Regulatory Toolkit in Multicellular Lineages.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (50): E4858-66.

Mesika, Adi, Shifra Ben-Dor, Elad L. Laviad, and Anthony H. Futerman. 2007. “A New Functional Motif in Hox Domain-Containing Ceramide Synthases: Identification of a Novel Region Flanking the Hox and TLC Domains Essential for Activity.” *The Journal of Biological Chemistry* 282 (37): 27366–73.

Mikhailov, Kirill V., Anastasiya V. Konstantinova, Mikhail A. Nikitin, Peter V. Troshin, Leonid Yu Rusin, Vassily A. Lyubetsky, Yuri V. Panchin, et al. 2009. “The Origin of Metazoa: A Transition from Temporal to Spatial Cell Differentiation.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 31 (7): 758–68.

Minh, Bui Quang, Minh Anh Thi Nguyen, and Arndt von Haeseler. 2013. “Ultrafast Approximation for Phylogenetic Bootstrap.” *Molecular Biology and Evolution* 30 (5): 1188–95.

Muller, Patricia A. J., and Karen H. Vousden. 2013. “P53 Mutations in Cancer.” *Nature Cell Biology* 15 (1): 2–8.

Musser, Jacob M., Klaske J. Schippers, Michael Nickel, Giulia Mizzon, Andrea B. Kohn, Constantin Pape, Paolo Ronchi, et al. 2021. “Profiling Cellular Diversity in Sponges Informs Animal Cell Type and Nervous System Evolution.” *Science* 374 (6568): 717–23.

Nadimpalli, Shilpa, Anton V. Persikov, and Mona Singh. 2015. “Pervasive Variation of Transcription Factor Orthologs Contributes to Regulatory Network Evolution.” *PLoS Genetics* 11 (3): e1005011.

Najafabadi, Hamed S., Sanie Mnaimneh, Frank W. Schmitges, Michael Garton, Kathy N. Lam, Ally Yang, Mihai Albu, et al. 2015. “C2H2 Zinc Finger Proteins Greatly Expand the Human Regulatory Lexicon.” *Nature Biotechnology* 33 (5): 555–62.

Nakagawa, So, Stephen S. Gisselbrecht, Julia M. Rogers, Daniel L. Hartl, and Martha L. Bulyk. 2013. “DNA-Binding Specificity Changes in the Evolution of Forkhead Transcription Factors.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (30): 12349–54.

Needham, David M., Susumu Yoshizawa, Toshiaki Hosaka, Camille Poirier, Chang Jae Choi, Elisabeth Hehenberger, Nicholas A. T. Irwin, et al. 2019. “A Distinct Lineage of Giant Viruses Brings a Rhodopsin Photosystem to Unicellular Marine Predators.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (41): 20574–83.

Nguyen, Hoa, M. A. R. Koehl, Christian Oakes, Greg Bustamante, and Lisa Fauci. 2019. “Effects of Cell Morphology and Attachment to a Surface on the Hydrodynamic Performance of Unicellular Choanoflagellates.” *Journal of the Royal Society, Interface / the Royal Society* 16 (150): 20180736.

Nguyen, Lam-Tung, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2015. “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

Nielsen, Claus. 2008. “Six Major Steps in Animal Evolution: Are We Derived Sponge Larvae?” *Evolution & Development* 10 (2): 241–57.

Ohno, Susumu. 1970. *Evolution by Gene Duplication*. Springer Science & Business Media.

Pathak, Narendra, Christina A. Austin, and Iain A. Drummond. 2011. “Tubulin Tyrosine Ligase-like Genes Ttll3 and Ttll6 Maintain Zebrafish Cilia Structure and Motility.” *The Journal of Biological Chemistry* 286 (13): 11685–95.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Philippe, Hervé, Henner Brinkmann, Dennis V. Lavrov, D. Timothy J. Littlewood, Michael Manuel, Gert Wörheide, and Denis Baurain. 2011. “Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough.” *PLoS Biology* 9 (3): e1000602.

Phillips, Jonathan E., Maribel Santos, Mohammed Konchwala, Chao Xing, and Duojia Pan. 2022. “Genome Editing in the Unicellular Holozoan Capsaspora Owczarzaki Suggests a Premetazoan Role for the Hippo Pathway in Multicellular Morphogenesis.” *ELife* 11 (June). https://doi.org/10.7554/eLife.77598.

Piasecki, Brian P., Jan Burghoorn, and Peter Swoboda. 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 (29): 12969–74.

Pinskey, Justine M., Adhya Lagisetty, Long Gui, Nhan Phan, Evan Reetz, Amirrasoul Tavakoli, Gang Fu, and Daniela Nicastro. 2022. “Three-Dimensional Flagella Structures from Animals’ Closest Unicellular Relatives, the Choanoflagellates.” *ELife* 11 (November). https://doi.org/10.7554/eLife.78133.

Plass, Mireya, Jordi Solana, F. Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J. Theis, Christine Kocks, and Nikolaus Rajewsky. 2018. “Cell Type Atlas and Lineage Tree of a Whole Complex Animal by Single-Cell Transcriptomics.” *Science* 360 (6391). https://doi.org/10.1126/science.aaq1723.

Plevin, Michael J., Morgon M. Mills, and Mitsuhiko Ikura. 2005. “The LxxLL Motif: A Multifunctional Binding Sequence in Transcriptional Regulation.” *Trends in Biochemical Sciences* 30 (2): 66–69.

Quigley, Ian K., and Chris Kintner. 2017. “Rfx2 Stabilizes Foxj1 Binding at Chromatin Loops to Enable Multiciliated Cell Gene Expression.” *PLoS Genetics* 13 (1): e1006538.

Reece-Hoyes, John S., and A. J. Marian Walhout. 2012. “Yeast One-Hybrid Assays: A Historical and Technical Perspective.” *Methods* 57 (4): 441–47.

Reinke, Aaron W., Jiyeon Baek, Orr Ashenberg, and Amy E. Keating. 2013. “Networks of BZIP Protein-Protein Interactions Diversified over a Billion Years of Evolution.” *Science* 340 (6133): 730–34.

Reiter, Franziska, Sebastian Wienerroither, and Alexander Stark. 2017. “Combinatorial Function of Transcription Factors and Cofactors.” *Current Opinion in Genetics & Development* 43 (April): 73–81.

Reiter, Jeremy F., and Michel R. Leroux. 2017. “Genes and Molecular Pathways Underpinning Ciliopathies.” *Nature Reviews. Molecular Cell Biology* 18 (9): 533–47.

Reith, W., C. Herrero-Sanchez, M. Kobr, P. Silacci, C. Berte, E. Barras, S. Fey, and B. Mach. 1990. “MHC Class II Regulatory Factor RFX Has a Novel DNA-Binding Domain and a Functionally Independent Dimerization Domain.” *Genes & Development* 4 (9): 1528–40.

Reith, W., M. Kobr, P. Emery, B. Durand, C. A. Siegrist, and B. Mach. 1994. “Cooperative Binding between Factors RFX and X2bp to the X and X2 Boxes of MHC Class II Promoters.” *The Journal of Biological Chemistry* 269 (31): 20020–25.

Richter, Daniel J., Cédric Berney, Jürgen F. H. Strassert, Yu-Ping Poh, Emily K. Herman, Sergio A. Muñoz-Gómez, Jeremy G. Wideman, Fabien Burki, and Colomban de Vargas. 2022. “EukProt: A Database of Genome-Scale Predicted Proteins across the Diversity of Eukaryotes.” *Peer Community Journal*, September. https://doi.org/10.24072/pcjournal.173.

Richter, Daniel J., Parinaz Fozouni, Michael B. Eisen, and Nicole King. 2018. “Gene Family Innovation, Conservation and Loss on the Animal Stem Lineage.” *ELife* 7 (May). https://doi.org/10.7554/eLife.34226.

Rickels, Ryan, Hans-Martin Herz, Christie C. Sze, Kaixiang Cao, Marc A. Morgan, Clayton K. Collings, Maria Gause, et al. 2017. “Histone H3K4 Monomethylation Catalyzed by Trr and Mammalian COMPASS-like Proteins at Enhancers Is Dispensable for Development and Viability.” *Nature Genetics* 49 (11): 1647–53.

Riggelen, Jan van, Alper Yetil, and Dean W. Felsher. 2010. "MYC as a Regulator of Ribosome Biogenesis and Protein Synthesis." *Nature Reviews. Cancer* 10 (4): 301–9.

Robinson, K. A., and J. M. Lopes. 2000. "SURVEY AND SUMMARY: Saccharomyces Cerevisiae Basic Helix-Loop-Helix Proteins Regulate Diverse Biological Processes." *Nucleic Acids Research* 28 (7): 1499–1505.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.

Rodríguez-Martínez, José A., Aaron W. Reinke, Devesh Bhimsaria, Amy E. Keating, and Aseem Z. Ansari. 2017. "Combinatorial BZIP Dimers Display Complex DNA-Binding Specificity Landscapes." *ELife* 6 (February). https://doi.org/10.7554/eLife.19272.

Rohrscheib, Chelsie E., and Jeremy C. Brownlie. 2013. "Microorganisms That Manipulate Complex Animal Behaviours by Affecting the Host's Nervous System." *Springer Science Reviews* 1 (1): 133–40.

Ros-Rocher, Núria, Alberto Pérez-Posada, Michelle M. Leger, and Iñaki Ruiz-Trillo. 2021. "The Origin of Animals: An Ancestral Reconstruction of the Unicellular-to-Multicellular Transition." *Open Biology* 11 (2): 200359.

Roy, H. E., D. C. Steinkraus, J. Eilenberg, A. E. Hajek, and J. K. Pell. 2006. "Bizarre Interactions and Endgames: Entomopathogenic Fungi and Their Arthropod Hosts." *Annual Review of Entomology* 51: 331–57.

Ruiz-Trillo, Iñaki, Andrew J. Roger, Gertraud Burger, Michael W. Gray, and B. Franz Lang. 2008. "A Phylogenomic Investigation into the Origin of Metazoa." *Molecular Biology and Evolution* 25 (4): 664–72.

Ryan, Joseph F., Kevin Pang, Christine E. Schnitzler, Anh-Dao Nguyen, R. Travis Moreland, David K. Simmons, Bernard J. Koch, et al. 2013. "The Genome of the Ctenophore Mnemiopsis Leidyi and Its Implications for Cell Type Evolution." *Science* 342 (6164): 1242592.

Sampson, Timothy R., and Sarkis K. Mazmanian. 2015. "Control of Brain Development, Function, and Behavior by the Microbiome." *Cell Host & Microbe* 17 (5): 565–76.

Saqib, M., S. J. Wylie, and M. G. K. Jones. 2015. "Serendipitous Identification of a NewIflavirus-like Virus Infecting Tomato and Its Subsequent Characterization." *Plant Pathology* 64 (3): 519–27.

Schindelin, Johannes, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, et al. 2012. "Fiji: An Open-Source Platform for Biological-Image Analysis." *Nature Methods* 9 (7): 676–82.

Schmitges, Frank W., Ernest Radovani, Hamed S. Najafabadi, Marjan Barazandeh, Laura F. Campitelli, Yimeng Yin, Arttu Jolma, et al. 2016. "Multiparameter Functional Diversity of Human C2H2 Zinc Finger Proteins." *Genome Research* 26 (12): 1742–52.

Schmitz, Jonathan F., Fabian Zimmer, and Erich Bornberg-Bauer. 2016. "Mechanisms of Transcription Factor Evolution in Metazoa." *Nucleic Acids Research* 44 (13): 6287–97.

Schneuwly, S., R. Klemenz, and W. J. Gehring. 1987. "Redesigning the Body Plan of Drosophila by Ectopic Expression of the Homoeotic Gene Antennapedia." *Nature* 325 (6107): 816–18.

Schumacher, Maria A., Audrey O. T. Lau, and Patricia J. Johnson. 2003. "Structural Basis of Core Promoter Recognition in a Primitive Eukaryote." *Cell* 115 (4): 413–24.

Schwaiger, Michaela, Anna Schönauer, André F. Rendeiro, Carina Pribitzer, Alexandra Schauer, Anna F. Gilles, Johannes B. Schinko, Eduard Renfer, David Fredman, and Ulrich Technau. 2014. “Evolutionary Conservation of the Eumetazoan Gene Regulatory Landscape.” *Genome Research* 24 (4): 639–50.

Sebé-Pedrós, Arnau, Ana Ariza-Cosano, Matthew T. Weirauch, Sven Leininger, Ally Yang, Guifré Torruella, Marcin Adamski, et al. 2013. “Early Evolution of the T-Box Transcription Factor Family.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (40): 16050–55.

Sebé-Pedrós, Arnau, Cecilia Ballaré, Helena Parra-Acero, Cristina Chiva, Juan J. Tena, Eduard Sabidó, José Luis Gómez-Skarmeta, Luciano Di Croce, and Iñaki Ruiz-Trillo. 2016. “The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity.” *Cell* 165 (5): 1224–37.

Sebé-Pedrós, Arnau, Elad Chomsky, Kevin Pang, David Lara-Astiaso, Federico Gaiti, Zohar Mukamel, Ido Amit, Andreas Hejnol, Bernard M. Degnan, and Amos Tanay. 2018. “Early Metazoan Cell Type Diversity and the Evolution of Multicellular Gene Regulation.” *Nature Ecology & Evolution* 2 (7): 1176–88.

Sebé-Pedrós, Arnau, Bernard M. Degnan, and Iñaki Ruiz-Trillo. 2017. “The Origin of Metazoa: A Unicellular Perspective.” *Nature Reviews. Genetics* 18 (8): 498–512.

Sebé-Pedrós, Arnau, Manuel Irimia, Javier Del Campo, Helena Parra-Acero, Carsten Russ, Chad Nusbaum, Benjamin J. Blencowe, and Iñaki Ruiz-Trillo. 2013. “Regulated Aggregative Multicellularity in a Close Unicellular Relative of Metazoa.” *ELife* 2 (December): e01287.

Sebé-Pedrós, Arnau, Alex de Mendoza, B. Franz Lang, Bernard M. Degnan, and Iñaki Ruiz-Trillo. 2011. “Unexpected Repertoire of Metazoan Transcription Factors in the Unicellular Holozoan Capsaspora Owczarzaki.” *Molecular Biology and Evolution* 28 (3): 1241–54.

Sebé-Pedrós, Arnau, Baptiste Saudemont, Elad Chomsky, Flora Plessier, Marie-Pierre Mailhé, Justine Renno, Yann Loe-Mie, et al. 2018. “Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq.” *Cell* 173 (6): 1520-1534.e20.

Sedykh, Irina, Abigail N. Keller, Baul Yoon, Laura Roberson, Oleg V. Moskvin, and Yevgenya Grinblat. 2018. “Zebrafish Rfx4 Controls Dorsal and Ventral Midline Formation in the Neural Tube.” *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 247 (4): 650–59.

Shalchian-Tabrizi, Kamran, Marianne A. Minge, Mari Espelund, Russell Orr, Torgeir Ruden, Kjetill S. Jakobsen, and Thomas Cavalier-Smith. 2008. “Multigene Phylogeny of Choanozoa and the Origin of Animals.” *PloS One* 3 (5): e2098.

Shi, Mang, Xian-Dan Lin, Jun-Hua Tian, Liang-Jun Chen, Xiao Chen, Ci-Xiu Li, Xin-Cheng Qin, et al. 2016. “Redefining the Invertebrate RNA Virosphere.” *Nature* 540 (7634): 539–43.

Siebert, Stefan, Jeffrey A. Farrell, Jack F. Cazet, Yashodara Abeykoon, Abby S. Primack, Christine E. Schnitzler, and Celina E. Juliano. 2019. “Stem Cell Differentiation Trajectories in Hydra Resolved at Single-Cell Resolution.” *Science* 365 (6451). https://doi.org/10.1126/science.aav9314.

Sigg, Monika Abedin, Tabea Menchen, Chanjae Lee, Jeffery Johnson, Melissa K. Jungnickel, Semil P. Choksi, Galo Garcia 3rd, et al. 2017. “Evolutionary Proteomics Uncovers Ancient Associations of Cilia with Signaling Pathways.” *Developmental Cell* 43 (6): 744-762.e11.

Silva, Leonardo A., Daniel M. P. Ardisson-Araujo, Ricardo S. Tinoco, Odair A. Fernandes, Fernando L. Melo, and Bergmann M. Ribeiro. 2015. “Complete Genome Sequence and Structural Characterization of a Novel Iflavirus Isolated from Opsiphanes Invirae (Lepidoptera: Nymphalidae).” *Journal of Invertebrate Pathology* 130 (September): 136–40.

Simpson, T. L. 1984. *The Cell Biology of Sponges*. Springer Science & Business Media.

Smith, Carolyn L., Frédérique Varoqueaux, Maike Kittelmann, Rita N. Azzam, Benjamin Cooper, Christine A. Winters, Michael Eitel, Dirk Fasshauer, and Thomas S. Reese. 2014. “Novel Cell Types, Neurosecretory Cells, and Body Plan of the Early-Diverging Metazoan Trichoplax Adhaerens.” *Current Biology: CB* 24 (14): 1565–72.

Son, Moonil, Jisuk Yu, and Kook-Hyung Kim. 2015. “Five Questions about Mycoviruses.” *PLoS Pathogens* 11 (11): e1005172.

Srivastava, Mansi, Emina Begovic, Jarrod Chapman, Nicholas H. Putnam, Uffe Hellsten, Takeshi Kawashima, Alan Kuo, et al. 2008. “The Trichoplax Genome and the Nature of Placozoans.” *Nature* 454 (7207): 955–60.

Srivastava, Mansi, Oleg Simakov, Jarrod Chapman, Bryony Fahey, Marie E. A. Gauthier, Therese Mitros, Gemma S. Richards, et al. 2010. “The Amphimedon Queenslandica Genome and the Evolution of Animal Complexity.” *Nature* 466 (7307): 720–26.

Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.

Steenwyk, Jacob L., Thomas J. Buida 3rd, Yuanning Li, Xing-Xing Shen, and Antonis Rokas. 2020. “ClipKIT: A Multiple Sequence Alignment Trimming Software for Accurate Phylogenomic Inference.” *PLoS Biology* 18 (12): e3001007.

Steinkraus, Donald C., and John P. Kramer. 1987. “Susceptibility of Sixteen Species of Diptera to the Fungal PathogenEntomophthora Muscae (Zygomycetes: Entomophthoraceae).” *Mycopathologia* 100 (1): 55–63.

Stubbs, Jennifer L., Isao Oishi, Juan Carlos Izpisúa Belmonte, and Chris Kintner. 2008. “The Forkhead Protein Foxj1 Specifies Node-like Cilia in Xenopus and Zebrafish Embryos.” *Nature Genetics* 40 (12): 1454–60.

Suga, Hiroshi, and Iñaki Ruiz-Trillo. 2013. “Development of Ichthyosporeans Sheds Light on the Origin of Metazoan Multicellularity.” *Developmental Biology* 377 (1): 284–92.

Sugiaman-Trapman, Debora, Morana Vitezic, Eeva-Mari Jouhilahti, Anthony Mathelier, Gilbert Lauter, Sougat Misra, Carsten O. Daub, Juha Kere, and Peter Swoboda. 2018. “Characterization of the Human RFX Transcription Factor Family by Regulatory and Target Gene Analysis.” *BMC Genomics* 19 (1): 181.

Suzuki, Tomohiro, Yoshino Takeshima, Toshiyuki Mikamoto, Jun-David Saeki, Tatsuya Kato, Enoch Y. Park, Hirokazu Kawagishi, and Hideo Dohra. 2015. “Genome Sequence of a Novel Iflavirus from MRNA Sequencing of the Pupa of Bombyx Mori Inoculated with Cordyceps Militaris.” *Genome Announcements* 3 (5). https://doi.org/10.1128/genomeA.01039-15.

Swoboda, P., H. T. Adler, and J. H. Thomas. 2000. “The RFX-Type Transcription Factor DAF-19 Regulates Sensory Neuron Cilium Formation in C. Elegans.” *Molecular Cell* 5 (3): 411–21.

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, and Principal investigators. 2018. “Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris.” *Nature* 562 (7727): 367–72.

Tacheny, A., M. Dieu, T. Arnould, and P. Renard. 2013. “Mass Spectrometry-Based Identification of Proteins Interacting with Nucleic Acids.” *Journal of Proteomics* 94 (December): 89–109.

Takahashi, Kazutoshi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. 2007. “Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors.” *Cell* 131 (5): 861–72.

Tamura, Koichiro, Masatoshi Nei, and Sudhir Kumar. 2004. “Prospects for Inferring Very Large Phylogenies by Using the Neighbor-Joining Method.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (30): 11030–35.

Tikhonenkov, Denis V., Kirill V. Mikhailov, Elisabeth Hehenberger, Sergei A. Karpov, Kristina I. Prokina, Anton S. Esaulov, Olga I. Belyakova, et al. 2020. “New Lineage of Microbial Predators Adds Complexity to Reconstructing the Evolutionary Origin of Animals.” *Current Biology: CB*, September. https://doi.org/10.1016/j.cub.2020.08.061.

Troha, Katia, Joo Hyun Im, Jonathan Revah, Brian P. Lazzaro, and Nicolas Buchon. 2018. “Comparative Transcriptomics Reveals CrebA as a Novel Regulator of Infection Tolerance in D. Melanogaster.” *PLoS Pathogens* 14 (2): e1006847.

Valentine, James W., Allen G. Collins, and C. Porter Meyer. 1994. “Morphological Complexity Increase in Metazoans.” *Paleobiology* 20 (2): 131–42.

Vij, Shubha, Jochen C. Rink, Hao Kee Ho, Deepak Babu, Michael Eitel, Vijayashankaranarayanan Narasimhan, Varnesh Tiku, Jody Westbrook, Bernd Schierwater, and Sudipto Roy. 2012. “Evolutionarily Ancient Association of the FoxJ1 Transcription Factor with the Motile Ciliogenic Program.” *PLoS Genetics* 8 (11): e1003019.

Villar, Diego, Camille Berthelot, Sarah Aldridge, Tim F. Rayner, Margus Lukk, Miguel Pignatelli, Thomas J. Park, et al. 2015. “Enhancer Evolution across 20 Mammalian Species.” *Cell* 160 (3): 554–66.

Visel, Axel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2009. “ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers.” *Nature* 457 (7231): 854–58.

Wagner, Günter P., Eric M. Erkenbrack, and Alan C. Love. 2019. “Stress-Induced Evolutionary Innovation: A Mechanism for the Origin of Cell Types.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 41 (4): e1800188.

Wallbank, Richard W. R., Simon W. Baxter, Carolina Pardo-Diaz, Joseph J. Hanly, Simon H. Martin, James Mallet, Kanchon K. Dasmahapatra, et al. 2016. “Evolutionary Novelty in a Butterfly Wing Pattern through Enhancer Shuffling.” *PLoS Biology* 14 (1): e1002353.

Wang, Guobao, Jianjia Zhang, Yunwang Shen, Qin Zheng, Min Feng, Xingwei Xiang, and Xiaofeng Wu. 2015. “Transcriptome Analysis of the Brain of the Silkworm Bombyx Mori Infected with Bombyx Mori Nucleopolyhedrovirus: A New Insight into the Molecular Mechanism of Enhanced Locomotor Activity Induced by Viral Infection.” *Journal of Invertebrate Pathology* 128: 37–43.

Webster, Claire L., Fergal M. Waldron, Shaun Robertson, Daisy Crowson, Giada Ferrari, Juan F. Quintana, Jean-Michel Brouqui, et al. 2015. “The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila Melanogaster.” *PLoS Biology* 13 (7): e1002210.

Weirauch, Matthew T., Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, et al. 2013. “Evaluation of Methods for Modeling Transcription Factor Sequence Specificity.” *Nature Biotechnology* 31 (2): 126–34.

Weirauch, Matthew T., and T. R. Hughes. 2011. “A Catalogue of Eukaryotic Transcription Factor Types, Their Evolutionary Origin, and Species Distribution.” *Sub-Cellular Biochemistry* 52: 25–73.

Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. “Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity.” *Cell* 158 (6): 1431–43.

Wolfe, S. A., L. Nekludova, and C. O. Pabo. 2000. “DNA Recognition by Cys2His2 Zinc Finger Proteins.” *Annual Review of Biophysics and Biomolecular Structure* 29: 183–212.

Wong, Emily S., Dawei Zheng, Siew Z. Tan, Neil L. Bower, Victoria Garside, Gilles Vanwalleghem, Federico Gaiti, et al. 2020. “Deep Conservation of the Enhancer Regulatory Code in Animals.” *Science* 370 (6517). https://doi.org/10.1126/science.aax8137.

Woznica, Arielle, Joseph P. Gerdt, Ryan E. Hulett, Jon Clardy, and Nicole King. 2017. “Mating in the Closest Living Relatives of Animals Is Induced by a Bacterial Chondroitinase.” *Cell* 170 (6): 1175-1183.e11.

Wu, S. Y., and M. McLeod. 1995. “The Sak1 Gene of Schizosaccharomyces Pombe Encodes an RFX Family DNA-Binding Protein That Positively Regulates Cyclic AMP-Dependent Protein Kinase-Mediated Exit from the Mitotic Cell Cycle.” *Molecular and Cellular Biology*. https://doi.org/10.1128/mcb.15.3.1479.

Yang, Yi-Ting, Yu-Shin Nai, Se Jin Lee, Mi Rong Lee, Sihyeon Kim, and Jae Su Kim. 2016. “A Novel Picorna-like Virus, Riptortus Pedestris Virus-1 (RiPV-1), Found in the Bean Bug, R. Pedestris, after Fungal Infection.” *Journal of Invertebrate Pathology* 141 (November): 57–65.

Young, Susan L., Daniel Diolaiti, Maralice Conacci-Sorrell, Iñaki Ruiz-Trillo, Robert N. Eisenman, and Nicole King. 2011. “Premetazoan Ancestry of the Myc-Max Network.” *Molecular Biology and Evolution* 28 (10): 2961–71.

Yu, Xianwen, Chee Peng Ng, Hermann Habacher, and Sudipto Roy. 2008. “Foxj1 Transcription Factors Are Master Regulators of the Motile Ciliogenic Program.” *Nature Genetics* 40 (12): 1445–53.

Zabidi, Muhammad A., Cosmas D. Arnold, Katharina Schernhuber, Michaela Pagani, Martina Rath, Olga Frank, and Alexander Stark. 2015. “Enhancer-Core-Promoter Specificity Separates Developmental and Housekeeping Gene Regulation.” *Nature* 518 (7540): 556–59.

Zakhvatkin, A. A. 1949. “The Comparative Embryology of the Low Invertebrates. Sources and Method of the Origin of Metazoan Development.” *Soviet Science*.

Zhang, Wei, and Zhirong Sun. 2008. “Random Local Neighbor Joining: A New Method for Reconstructing Phylogenetic Trees.” *Molecular Phylogenetics and Evolution* 47 (1): 117–28.