Title
Extracellular matrix and the origin of animal multicellularity

Author
Wetzel, Laura Anne

Publication Date
2018

Peer reviewed|Thesis/dissertation

Extracellular matrix and the origin of animal multicellularity

By

Laura Anne Wetzel

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 Iswar Hariharan
Professor Douglas Koshland
Professor John Taylor

Fall 2018

Abstract

Extracellular matrix and the origin of animal multicellularity

by

Laura Anne Wetzel

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Nicole King, Chair

The evolution of animals from their unicellular ancestors was a major transition in evolutionary history that enabled the diversity and complexity of extant animals. Due to the lack of a clear fossil record of the progenitors of animals, we know relatively little about the first animals and the developmental events that predicated their evolution. My doctoral research utilized *Salpingoeca rosetta*, a model choanoflagellate and one of the closest living relatives of animals, to investigate the molecular changes that might have led to the origin of multicellularity in animals.

S. rosetta facultatively forms multicellular "rosettes" through serial cell division in a process reminiscent of early animal embryogenesis. To determine the genetic underpinnings of rosette development in *S. rosetta*, I performed a forward genetic screen for rosette defect mutants (Chapter 2). I identified a new class of mutants that aggregate promiscuously into amorphous clumps of cells, but that do not develop into orderly rosettes. Two clumpy mutants, named Jumble and Couscous, mapped to lesions in genes encoding predicted glycosyltransferases. The mutations in the *jumble* and *couscous* genes were shown to disrupt glycosylation patterns at the basal pole of the extracellular matrix (ECM). The only previously identified gene required for rosette formation, *rosetteless*, was found to encode a protein that localizes at the basal pole of cells in the ECM-filled center of rosettes. Thus, all three genes known to be required for rosette development in *S. rosetta* play a role in establishing the ECM of rosettes and implicate the ECM in the regulation of multicellular development.

Beyond the specific genes required for cell adhesion and cell signaling, animal multicellular development relies on transcriptional regulation of specific genes for cell differentiation. To examine the role of transcriptional regulation, I generated an improved *S. rosetta* genome assembly that allows for better annotation of regulatory regions and analyzed chromatin accessibility using an assay for transposase-accessible chromatin (ATAC-seq) in distinct *S. rosetta* cell types, including unicellular slow swimmers and rosettes (Appendix). Slow swimmers and rosettes were found to have nearly identical chromatin accessibility profiles—consistent with previous transcriptome sequencing that showed remarkably similar expression profiles between slow swimmers and rosettes.

Taken together, *S. rosetta* may rely on translational and/or post-translational regulation, including modification of the ECM, to build multicellular rosettes. Continuation

of the forward genetic screen for rosette defects to saturation and utilization of recently developed methods for reverse genetics will allow future scientists to more fully elucidate the genetic basis of rosette development in *S. rosetta* and its possible homology to animal development.

Table of Contents

List of Figures and Tables

Chapter 1:

Chapter 2:

Appendix:

Acknowledgements

"I immediately loved working with flies. They fascinated me and followed me around in my dreams." -Christiane Nüsslein-Volhard

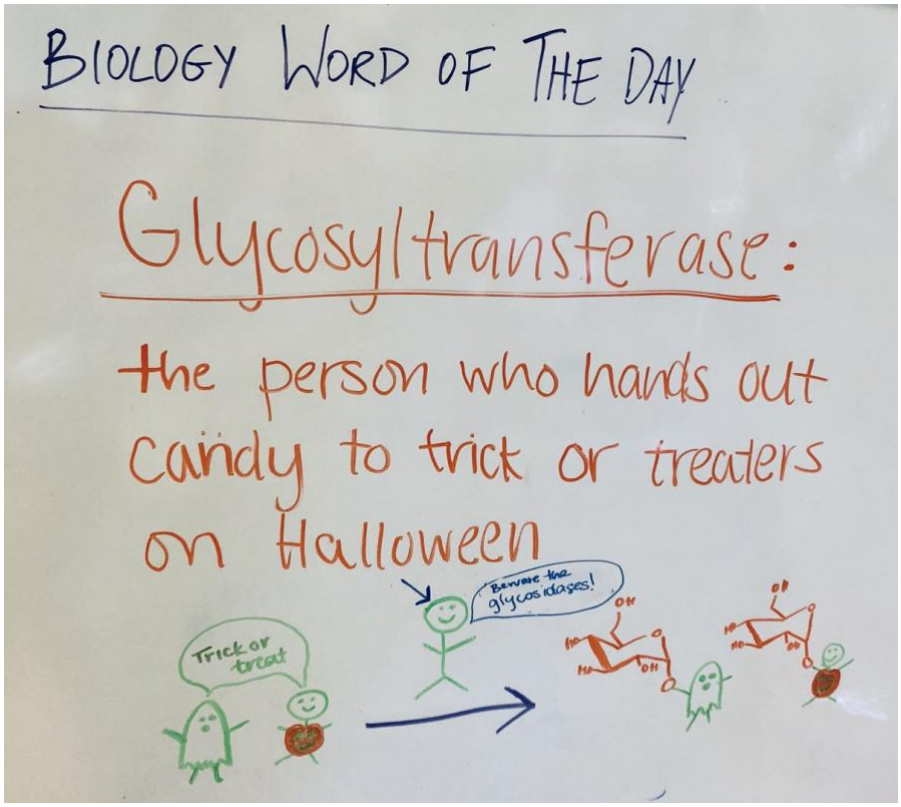
Christiane Nüsslein-Volhard's pioneering forward genetic screen in fruit flies has been a constant source of inspiration throughout my graduate career. She demonstrated how persistence and simple genetics can reveal a wealth of information about the genes required for conserved developmental processes. Much like my science hero, I was immediately taken with my model organism, the gregarious choanoflagellate. However, over the years, I would often face what felt like insurmountable challenges in studying a system lacking genetic tools. I cannot imagine overcoming these hurdles without the incredible scientific and emotional support of so many mentors, colleagues, friends, and family.

First and foremost, I would like to thank my advisor Nicole King for providing the freedom to pursue a project that took time and patience to yield results, for showing me how to give clear talks and write beautiful papers, for giving me so many opportunities to interact with the larger scientific community, and for helping me to grow into an independent scientist.

I would also like to acknowledge the useful suggestions and encouragement from my thesis committee members: Iswar Hariharan, Doug Koshland, and John Taylor.

I want to thank my previous scientific advisors and mentors, Claire Walczak, Rania Rizk, and Ke Hu, for instilling a love of looking through a microscope to generate scientific questions and for having the patience to teach me. Thank you to the Integrated Freshman Learning Experience (IFLE) at Indiana University for exposing me to the excitement of biological research—which eventually led me to graduate school.

Thanks to all the King lab members, past and present, for making the lab a collegial and friendly environment to work and do science. In particular, thanks to Pawel Burkhardt to introducing me to the wonder of choanoflagellates (and sponges). Thanks to Tera Levin, the founder of choanoflagellate genetics and a truly amazing mentor without whom I never would have passed my qualifying exams. Thanks to David Booth and his persistence in developing choanoflagellate transgenics without which I may never have been able to graduate. Thanks to Monika Sigg, Thibaut Brunet, and Ben Larson for their helpful scientific discussions and input. Thanks to the rogue members of the King Lab that created the joy of the twitter account, @KingLabWOTD (see figure below).



Thanks to Kayley Hake who joined the King lab at the same time I did. From biking a keg across Berkeley to planning a scientific conference in France together, we have had so many adventures throughout the years. I cannot imagine graduate school without your friendship.

I want to acknowledge the MBL Physiology course at Woods Hole for reminding me of the pure joy of observation and discovery. To all the people that I've met through the course and that keep the course running, thank you for keeping the Woods Hole spirit alive.

Graduate school would not have been possible without the following amazing friends and housemates who have been there to complain, grab coffee, dance at MoLo, or adventure outdoors with me: Shawn Beckman, Ellie Bondra, Gina Caldas, Justin De Leon, James Hart, Rachel Kjolby, Rosalie Lawerence, Adam Levine, Katie Lien, Shion Lim, Ethan McSpadden, Ambika Nadkarni, Alli Quan, Alex Seletsky, Amy Strom, Akshay Tambe, Lan Vu, and Lindsey Young.

Finally, thanks to the unwavering support of my family—especially my mom who always knew how to say the right thing when I would call her upset about my latest lab failure.

Chapter 1

Cell adhesion and the glycocalyx in the evolution of multicellular development in animals

INTRODUCTION

Animals develop from a single founding cell, the zygote, that undergoes serial rounds of cell division along with cell movement, cell differentiation, tissue morphogenesis, and apoptosis to produce adult animals in a wide array of body plans (Gilbert, 2000). All modern animals are composed of hundreds to trillions of cells (Savage, 1977; Wood, 1988) that function cooperatively and differentiate into at least five distinguishable cell types (Valentine, 2006). Yet, little is known about the evolutionary events that lead to coordinated multicellular development in animals.

Beyond animals, multicellularity evolved independently in at least 15 other lineages throughout the eukaryotic tree, notably including land plants, fungi (possibly including separate origins in ascomycetes and in basidiomycetes) (Nagy et al., 2018), cellular slime molds and different types of algae (Bonner, 1998; King, 2004; Rokas, 2008) (Figure 1.1). Each of these transitions occurred at different times in the history of life (Sebé-Pedrós et al., 2017) and from a distinct unicellular ancestor (Baldauf, 2003; King, 2004). Comparing multicellular eukaryotes may allow us to extrapolate whether there are common molecular mechanisms for the evolution of multicellularity and may help to inform our understanding of animal origins.

TWO PATHWAYS TO MULTICELLULARITY: AGGREGATION AND DIVISION

Each origin of multicellularity occurred through either aggregation or by clonal cell division. Aggregative multicellularity happens through the adhesion of individual cells to each other whereas clonal multicellularity results from successive rounds of division without separation of sister cells. Interestingly, there are no reported descriptions of protists that are able to form both stable aggregative and clonal multicellular forms, indicating that the two types of multicellularity may not be easily interconvertible (Bonner, 1998; Brunet and King, 2017).

Overall, aggregative multicellularity is a less common strategy among eukaryotes than clonal multicellularity. Aggregation is seen as a roadblock to the evolution of division of labor between cell types (Buss, 1988) because aggregated cells are not required to be genetically related and therefore are vulnerable to “cheaters” that can benefit from aggregation without sharing resources or contributing labor (Santorelli et al., 2008; Strassmann et al., 2000). Thus, aggregation is predicted to be evolutionarily stable only if restricted to close relatives (Gilbert et al., 2007; Kuzdzal-Fick et al., 2011). High relatedness ensures that cheaters and cooperators will tend to be in different groups, which limits the opportunity for cheaters to exploit cooperators and exposes any group-level defects for cheaters to selection (Gilbert et al., 2007).

One of the best-studied models of aggregative multicellularity, the slime mold *Dictyostelium discoideum*, relies on specific kin recognition mechanisms to form fruiting bodies with viable spores and dead stalk cells upon starvation (Benabentos et al., 2009; Hirose et al., 2011). A matching pair of the highly polymorphic *tgrB1* and *tgrC1* alleles is required for self-recognition through the predicted direct binding of the extracellular protein products of the two alleles (Gruenheit et al., 2017; Hirose et al., 2011). TgrB1 and TgrC1 contain immunoglobulin-like folds—a fold that is tolerant of sequence variation and convergently evolved in many organisms to mediate self-recognition (Hirose et al., 2011).

Despite the safeguard of kin selection, the multicellular form of *D. discoideum* remains a transient stage of in its overall life history. In fact, in all known cases, aggregative multicellularity is only found as a temporary life stage (Sebé-Pedrós et al., 2017).

On the other hand, clonal multicellularity is more phylogenetically widespread. Clonal multicellularity might evolve relatively easily since it avoids many of the genetic conflicts associated with aggregation since related cells remain stuck together. Moreover, "complex multicellularity" marked by cells that undergo spatial cell differentiation and have intercellular communication (Knoll, 2011) is restricted to lineages with clonal development. Complex multicellularity has evolved only six times among eukaryotes suggesting that the evolutionary hurdle is relatively high—possibly due to the strong demand for cooperation among cells that are specialized and dependent on each other.

In fact, clonal multicellularity can arise with simple mutations that lead to a loss of complete cell division (Brunet and King, 2017). Within 100 generations after the introduction of a predator, the green alga*, Chlorella vulgaris*, was able to form multicellular clusters of tens to hundreds of cells, apparently through incomplete cell division (Boraas et al., 1998). During selection for rapid settling in liquid media or through predation, *Chlamydomonas reinhardtii* evolved multicellular clusters that developed clonally by daughter cells remaining together after mitotic reproduction (Herron et al., 2018; Ratcliff et al., 2013). In a similar selection scheme for settling, disruption the transcription factor *ACE2* of *Saccharomyces cerevisiae* was found to prevent mother-daughter cell separation and generate multicellular "snowflake" yeast (Ratcliff et al., 2015). In each of these studies, laboratory-evolved *de novo* multicellularity resulted from mutations that led to incomplete cytokinesis that were acquired in between 100-315 generations (Boraas et al., 1998; Ratcliff et al., 2015, 2013).

Since most transitions to multicellularity happened hundreds of millions of years ago, how can we unravel their mechanisms? There are two possible approaches: (1) First, comparative genomics might reveal gene families whose origins, elaboration, or expansion correlates with the transition to multicellularity in the diverse groups of interest. (2) Second, in forms with facultative multicellularity, loss-of-function approaches (i.e, mutant screens or knock-outs) can reveal genes that are necessary for multicellularity. Strikingly, these two complementary approaches have converged toward revealing a central role for cell adhesion genes.

MULTICELLULAR CELL ADHESION IS MEDIATED BY GLYCOSYLATION

Comparative genomics has revealed a dramatic expansion of genes encoding adhesion molecules in multicellular organisms (Gagneux et al., 2015), which is consistent with data from developmental and cell biology that emphasizes the importance of adhesion molecules for the establishment and maintenance of the complex multicellular forms of animals, plants, and fungi. The specific adhesion molecules themselves vary between lineages, but in many multicellular lineages an extracellular matrix (ECM) that is heavily modified with sugars mediates adhesion (Colley et al., 2015) (Figure 1.2). Evolution has repeatedly and consistently selected polysaccharides, often referred to as glycans when on the exterior surface of cells, as being the most diverse and flexible molecules at the interface between cells and the extracellular environment (Colley et al., 2015). Potential reasons that glycans cover all cell types include: their relative

hydrophilicity, flexibility, and mobility in aqueous environments and their extreme diversity that allows facile short-term and long-term adaptations to changing environments and pathogen regimes (Colley et al., 2015). Alternatively, glycans could have evolved early in life and remained covering cells as an evolutionary contingency.

Glycans vary immensely in structure and expression both within and between evolutionary lineages (Gagneux et al., 2015) and our knowledge about this diversity remains limited largely due the inherent difficulties in elucidating their structures. Most major glycan classes in animal cells are represented in some related form among other eukaryotes or even archaea (Gagneux et al., 2015). One trend, however, is that there are far fewer N-glycoproteins in unicellular than multicellular organisms (Gagneux et al., 2015). Beyond covering the cells, glycans can play either a structural role in holding cells together or can modify the activity of other molecules which hold cells together.

For example, cell adhesion in the aggregative slime mold *D. discoideum* requires carbohydrate binding lectins, CBP-26 (also called discoidin I) and CBP-24 (also called discoidin II) (Ray et al., 1979; Shinnick and Lerner, 1980; Springer et al., 1984) (Figure 1.2). CBP-26 binds cell surface galactose molecules on other cells, allowing cell aggregation (Ray et al., 1979; Shinnick and Lerner, 1980). Inhibition of all protein glycosylation does not alter protein synthesis, but blocks cell aggregation (McDonald and Sampson, 1983).

Land plants and fungi are surrounded by cell walls composed of carbohydrates that mediate cell attachment (Figure 1.2). Land plants are covered by a cell wall composed of the polysaccharides cellulose, hemicellulose, and pectin (Daher and Braybrook, 2015; Iwai et al., 2002). Pectin encompasses a family of plant cell wall polysaccharides that contain galacturonic acid and are generally grouped into three majors types: homogalacturonan, rhamogalacturonan I, and the substituted galacturonan rhamnogalacturonan II (Atmodjo et al., 2013). Chemical modification of pectin, including methylation and acetylation, affect its ability to gel and act as glue between cells (Daher and Braybrook, 2015; Iwai et al., 2002). The fungal cell wall is composed mainly of the polysaccharides glucan and chitin (Bowman and Free, 2006; Gow et al., 2017). Fungal cell adhesion is mediated by a combination of adhesive cell wall proteins and secreted carbohydrates, although the precise composition of fungal adhesives is highly heterogenous between species and understudied in fruiting bodies and other complex multicellular forms (Nagy et al., 2018).

In animals, epithelial adhesion is mediated through cell adhesion molecules (CAMs) on the cell surface including: cadherins, integrins, selectins (a class of C-type lectin), and immunoglobin cell adhesion molecules (IgCAMs) (Edelman, 1986; Takeichi, 1988) (Figure 1.2). Cadherins, integrins, and IgCAMs are regulated by their glycosylation patterns (Carvalho et al., 2016; Guo et al., 2009; Kadmon et al., 1990; Zhao et al., 2008); while selectins bind sugar modifications to hold cells together. CAMs, such as cadherins and C-type lectins, are encoded in the genomes of choanoflagellates, the closest living relatives of animals (Abedin and King, 2008; Fairclough et al., 2010; King et al., 2008; Nichols et al., 2012; Richter and King, 2013). Illuminating the ancestral roles of CAMs and how glycans regulate or interact with CAMs in choanoflagellates and other close living relatives of animals can help to reveal which cell adhesion molecules were available to the last common ancestor of animals.

ESTABLISHING THE GLYCOCALYX

There is no single gene controlling glycan biosynthesis—they are synthesized and modified by a network of enzymes. The enormous complexity of the glycans found on the outside of all cells, also known as the glycocalyx, is derived from the orchestration of the enzymatic formation and breakdown of glycosidic linkages achieved by glycosyltransferases (GTs), glycoside hydrolase, glycoside phosphorylases, and polysaccharide lyases (Lairson et al., 2008; Rademacher et al., 1988) (Figure 1.3).

A key class of enzymes, the GTs, establish glycosidic linkages by transferring an activated donor sugar substrate that contains a phosphate leaving group to an acceptor substrate (Lairson et al., 2008). Most often, acceptor substrates are other sugars, but they can also be a lipid, protein, nucleic acid, or another small molecules (Lairson et al., 2008). GTs have been classified into 90 families based on amino acid sequence similarities (Campbell et al., 1997). However, making precise functional predictions based on sequence alone is often unreliable or inaccurate because many GT families exhibit polyspecificity, whereby enzymes within the same family have different donors and/or acceptors (Campbell et al., 1997). Based on more than 500 sequenced organisms, GTs account for about 1% to 2% of the gene products of an organism whether archaeal, bacterial, or eukaryotic—and even doubled-stranded DNA viruses (Lairson et al., 2008); thus, the enzyme family is very ancient. As of January 2008, there were 33,000 open reading frames encoding GTs, yet the donor and acceptor specificity for the vast majority (>95%) is not known (Lairson et al., 2008). Given their importance in establishing the glycocalyx, identifying the molecular and cellular functions of additional GTs in multicellular organisms and their relatives may reveal important modifications for proper cell adhesion in multicellular organisms.

GLYCOSYLATION IN ANIMAL DEVELOPMENT AND CANCER

Glycans are essential for developmental and differentiation events required for the assembly of complex multicellularity body plans in animals (Haltiwanger and Lowe, 2004). A strong link has been established between developmental phenotypes and genetic deficiencies in glycan expression and structures in humans, mice, *Drosophila melanogaster*, and *Caenorhabditis elegans* (Haltiwanger and Lowe, 2004). One of the best studied examples is the regulation of Notch signaling, essential for cell-fate specification in metazoans, by glycosylation of its extracellular domain (Takeuchi and Haltiwanger, 2014). The Notch extracellular domain is modified with different types of carbohydrates including asparagine-linked N-glycans and several serine- or threonine-linked O-glycans (Takeuchi and Haltiwanger, 2014). In mammals alone, there are nine different GTs that preferentially modify the epidermal growth factor (EGF) repeats in the Notch extracellular domain with O-glycans (Takeuchi and Haltiwanger, 2014). The nine GTs create a complex code of regulation as glycosylation by each enzyme is either essential for viability, activates Notch, or inhibits Notch (Takeuchi and Haltiwanger, 2014).

Cancer has been proposed by many researchers as a disease of multicellularity—whereby cancer cells revert to acting like unicellular life and can be viewed as cheating within a cooperative multicellular system (Aktipis et al., 2015; Chen et al., 2015; Trigos et al., 2018). Given the role glycosylation plays in ensuring multicellular cooperation in

animals, it should not be surprising that alterations in glycoproteins, glycosphingolipids, and proteoglycans are common features of cancer cells (Pinho and Reis, 2015). Irregular glycan expression in cancer has been attributed to under- or overexpression of glycosyltransferases (Kannagi et al., 2008), dysregulation of chaperone function (Schietinger et al., 2006), and/or altered glycosidase activity (Kakugawa et al., 2002).

Increased sialylation, branched-glycan structures and expression of 'core' fucosylation, are the most-widely occurring cancer-associated changes in protein glycosylation (Pinho and Reis, 2015). Several of these changes interfere directly with CAMs. Overexpression of branched-N-glycan structures disrupts epithelial cadherin-mediated cell-cell adhesion, which promotes tumor cell dissociation and invasion (Pinho et al., 2013, 2011, 2009). Modifications of integrins with branched N-glycans (Bassagañas et al., 2014), truncated O-glycans (Zhao et al., 2008), and/or sialylated structures (Asada et al., 1997) modulate tumor cell-matrix interaction and stimulate the process of tumor cell migration. Increased glycosylation with the sugars, sialyl Lewis^{X} and SLe^{a}, are associated with tumors and serve as ligands for adhesion receptors expressed in activated endothelial cells (E-selectin), platelets (P-selectin), and leukocytes (L-selectin), promoting cancer cell adhesion and metastasis (Rosen and Bertozzi, 1994). These examples underscore the importance of proper glycan modifications in ensuring proper cell adhesion mechanisms, yet the molecular mechanisms dictating glycosylation of CAMs remain difficult to study in the context of complex multicellular animals. All animals have complex multicellularity, making it difficult to know by looking at animals alone which changes paved the way for the initial evolution of multicellularity. On the other hand, the facultative multicellular forms of their closest relatives can serve as a proxy of the multicellular ancestors of animals.

HOLOZOANS CAN HELP REVEAL THE ORIGIN OF ANIMAL MULTICELLULARITY

Many of the genes required for cell-cell adhesion, cell-ECM adhesion, and developmental patterning are shared among almost all animals (Adamska et al., 2011; Nichols et al., 2006; Pires-daSilva and Sommer, 2003; Richter et al., 2018; Srivastava et al., 2010). The near ubiquity and conserved function of their gene products among extant animals implies the last common ancestor of all animals was multicellular and had a complex developmental toolkit. Yet, the first animals lived over 600 million years ago and left little mark on the fossil record for scientists to directly interrogate their origins (Conway Morris, 1993). Therefore, we must look to phylogenetically relevant relatives of animals to understand the early steps in the evolution of animal multicellularity. Three unicellular lineages, choanoflagellates, filastereans, and ichthyosporeans, form a clade with animals called Holozoa (Lang et al., 2002) which forms the reference point for studies of the origins of animals (Figure 1.1). Studying all unicellular holozoans can provide the most complete possible view of the genetic and regulatory toolkit available in the last common ancestor of animals.

Ichthyosporeans

The ichthyosporeans are the earliest branching holozoan lineage (Torruella et al., 2015). The clade, comprised of around 40 described species, includes parasites and commensals of a wide diversity of animals (Glockling et al., 2013; Mendoza et al., 2002).

The model species, *Creolimax fragrantissima*, has a life cycle that includes coenocytic development where the nucleus replicates multiple times without the cell itself dividing before undergoing cellularization and release of motile amoeboid zoospores (Marshall et al., 2008; Marshall and Berbee, 2011). Coenocytic development, despite being quite distinct from canonical animal development, is found in some animal lineages, such as in *Drosophila* syncytial blastoderm (Suga and Ruiz-Trillo, 2013); therefore, studying coenocytic development might inform our understanding of yet another developmental program available to the last common ancestor of animals.

The genome of *C. fragrantissima* has been sequenced and has shown distinct transcriptomic profiles for its two life stages that involve hundreds of differentially expressed genes (de Mendoza et al., 2015). Interestingly, the unicellular amoeboid cell type and not the multinucleate cell type is defined by multicellularity-related activities (de Mendoza et al., 2015). For example, the integrin adhesome, which is a major cell-ECM adhesion system in animals, and the transcription factor Brachyury, which is essential for animal development, are significantly upregulated in the unicellular amoeba stage (de Mendoza et al., 2015). Thus, while these genes may have an ancestral role in cell type specification, their roles in animal development may have been co-opted over time for different functions.

Filastereans

Filastereans are the sister group of the clade composed of choanoflagellates and animals (Torruella et al., 2015). To date, there are only four described filasterean species: the predatory flagellates *Pigoraptor vietnamica* and *Pigoraptor chileana* (Hehenberger et al., 2017), the marine free-living *Ministeria vibrans* (Schalchian-Tabrizi et al., 2008) and the snail symbiont *Capsaspora owczarzaki* (Hertel et al., 2002). The life cycle of the model species *C. owczarzaki* includes three different cell stages: an amoeboid stage, a cystic stage, and an aggregative multicellular stage (Sebé-Pedrós et al., 2016). *C. owczarzaki* has the only known example of aggregative multicellularity in Holozoa (Sebé-Pedrós et al., 2017). Transcriptome analysis has revealed that aggregate-stage *C. owczarzaki* cells highly express integrin adhesome genes, as well as ECM proteins, including fibronectin domain-containing and laminin domain-containing proteins (Sebé-Pedrós et al., 2016). This work has helped to reveal that major cell-cell and cell-matrix adhesion mechanisms in animals have a deeper evolutionary origin than previously thought. Studying further the molecular mechanisms of aggregation in *C. owczarzaki* may inform our understanding of adhesion and cell signaling mechanisms available to the last common ancestor of animals and whether aggregative multicellularity might have been an intermediate on the way to clonal multicellularity in animals.

Choanoflagellates

Choanoflagellates are a globally distributed group of marine and freshwater protozoans (Leadbeater, 2015). In 1867, Henry James-Clark first hypothesized a close relationship between choanoflagellates and the early branching animals, sponges, based on the shared morphology of choanoflagellates and choanocytes of sponges (James-Clark, 1867). With the advent of molecular phylogenetics, choanoflagellates unambiguously have been shown to be the closest living relatives of animals (King, 2004; Ruiz-Trillo et al., 2008; Schalchian-Tabrizi et al., 2008). Genomes and transcriptomes

from several choanoflagellate species have revealed that many gene families once thought to be animal-specific are present in choanoflagellates (Fairclough et al., 2013; King et al., 2008; Richter et al., 2018).

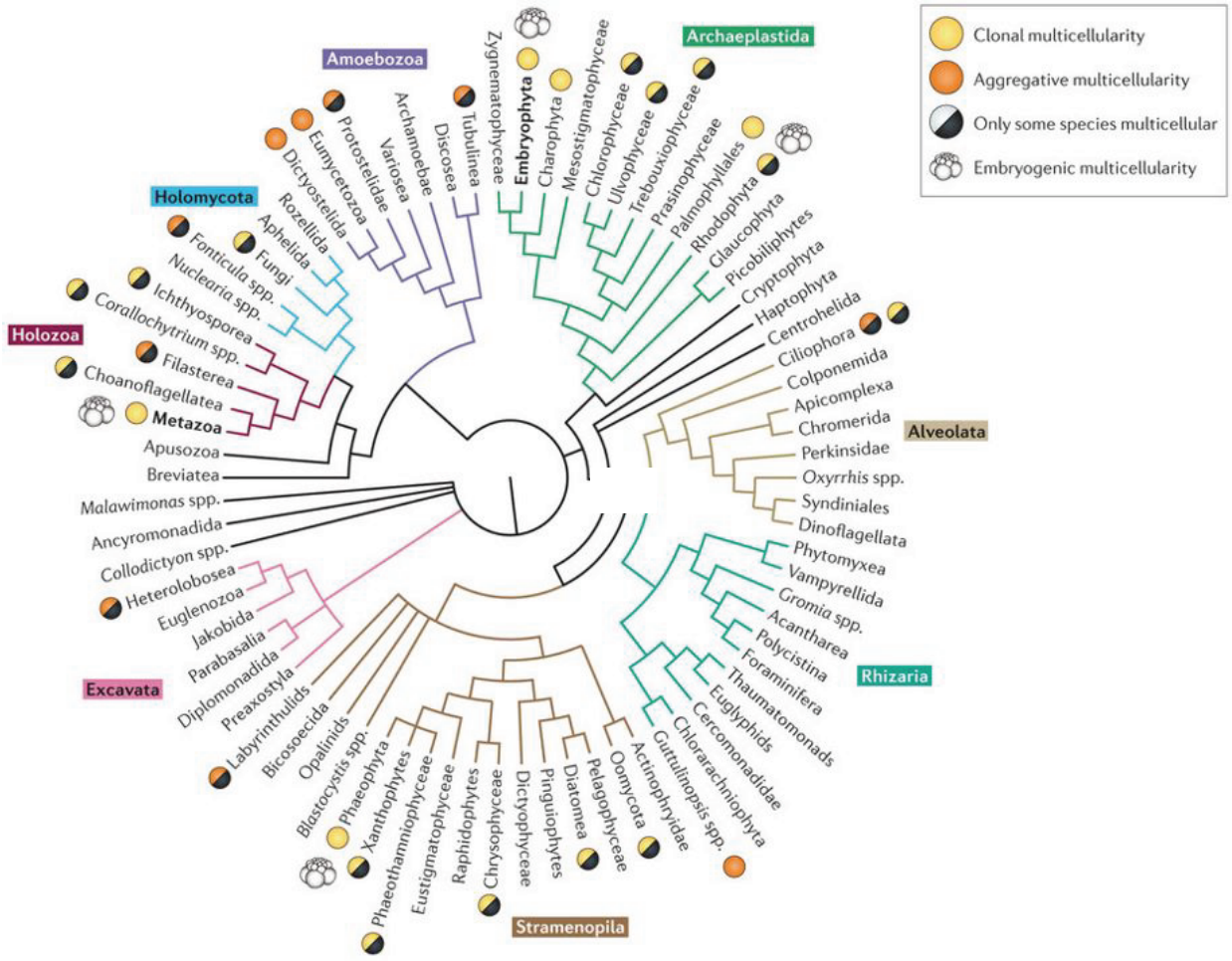
Enticingly, many choanoflagellates have facultatively clonal multicellular forms, including swimming spherical (rosettes), linear, or flat colonies, and sessile branching colonies (Dayel et al., 2011; Leadbeater, 2015, 1983). It remains unknown whether choanoflagellate and animal multicellularity are homologous or have independent origins. Studying choanoflagellate colony formation may help to clarify the relationship and reveal a plausible pathway to multicellularity along the animal lineage.

Salpingoeca rosetta is an emerging model choanoflagellate that has a complex life history that includes multicellular rosettes. Under standard laboratory conditions *S. rosetta* proliferates as solitary cells or linear chains that easily break apart into solitary cells (Dayel et al., 2011). However, upon exposure to the prey bacterium, *Algoriphagus machipongonensis, S. rosetta* develops into highly organized and structurally stable rosettes through a process of serial cell division reminiscent of early animal embryogenesis (Alegado et al., 2012; Dayel et al., 2011; Fairclough et al., 2010; Woznica et al., 2016). Recent advances – including a fully-sequenced genome (Fairclough et al., 2013), the discovery of a sexual phase of its life cycle that enables controlled mating (Levin et al., 2014; Levin and King, 2013; Woznica et al., 2017), and techniques that allow for transfection and expression of transgenes (Booth et al., 2018) — have enabled the detailed study of the molecular mechanisms underlying rosette development in *S. rosetta.*

A groundbreaking genetic screen for rosette defective mutants in *S. rosetta* revealed the first gene required for rosette formation, a C-type lectin named *rosetteless* (Levin et al., 2014). C-type lectins, until then thought to be animal specific (Fairclough et al., 2013; King et al., 2008; Richter et al., 2018), play many roles in animals, including cell-cell adhesion, cell-ECM adhesion, cell signaling, and innate immune recognition of pathogens (Cambi et al., 2005; Geijtenbeek and Gringhuis, 2009; Ruoslahti, 1996; Švajger et al., 2010; Zelensky and Gready, 2005). Although the molecular mechanisms by which *rosetteless* regulates rosette formation remain unknown, the wild type Rosetteless protein localizes to the basal pole of solitary swimmers and becomes heavily enriched in the ECM-filled center of rosettes upon rosette induction (Levin et al., 2014). This localization led to the hypothesis that the Rosetteless-filled ECM plays an essential role in holding together cells of a rosette and highlights the role of the ECM in multicellular development (Levin et al., 2014).

Several other rosette defective mutants were recovered in the screen and all lacked mutations at the *rosetteless* locus; thus, multiple genes are required for rosette formation (Levin et al., 2014). Intriguingly, transcriptome sequencing of *S. rosetta* cell types revealed that chain colonies and rosette colonies have remarkably similar genome-wide transcriptional profiles (Fairclough et al., 2013). Thus, it remains an open question what regulatory mechanisms—transcriptional, translational, or post-translational—direct the switch to rosette development. Uncovering additional required genes may help to reveal whether this pathway is conserved in the regulation of animal multicellularity.

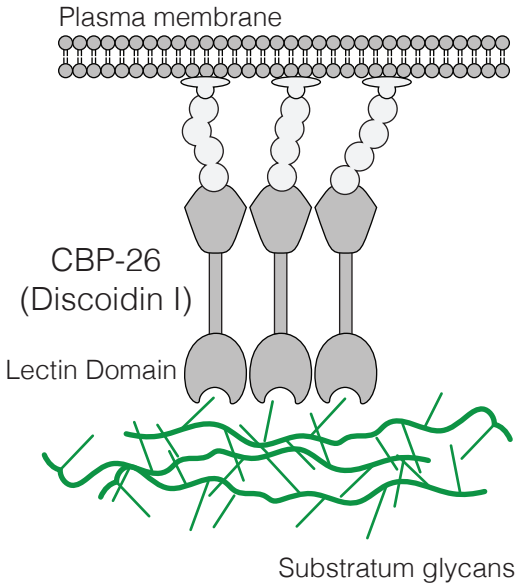
Figure 1.1. The multiple origins of multicellularity. Multicellular forms are present in many eukaryotic lineages. Several of these lineages have only a few multicellular species, but animals (Metazoa; highlighted in bold) and plants (Embryophyta; highlighted in bold) are entirely multicellular. Figure adapted from Sebé-Pedrós et al., 2017.



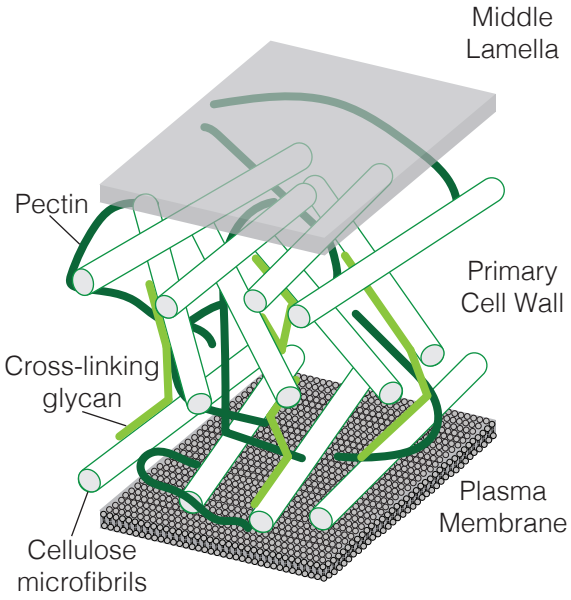
Adapted from Sebé-Pedrós, et al., 2017

Figure 1.2. Multicellular cell adhesion in diverse organisms is mediated by glycosylation. In both aggregative and clonal multicellularity, glycans (green) play key roles in cell-cell adhesion. The slime mold, *Dictyostelium discoideum,* forms an aggregative fruit body through the action of lectin, CBP-26, which binds to specific glycans to hold cells together. Figure adapted from Vasta et al., 2017. Land plants are covered by a cell wall composed of the polysaccharides: cellulose and pectin. Figure adapted from Loqué et al., 2015. Fungal cell adhesion mediated by a combination of adhesive cell wall proteins and secreted carbohydrates, although the precise composition of fungal adhesives is highly heterogenous between species and understudied in fruiting bodies and other complex multicellular forms. Figure adapted from Gow et al., 2017. In animals, epithelial adhesion is mediated through cell adhesion molecules (CAMs) on the cell surface including: cadherins, integrins, selectins (a class of C-type lectin), and immunoglobin cell adhesion molecules (IgCAMs). Cadherins, integrins, and IgCAMs are all regulated by their glycosylation patterns. Selectins are lectins that bind specific glycans to hold cells together. Figure adapted from Lodish et al., 2000.

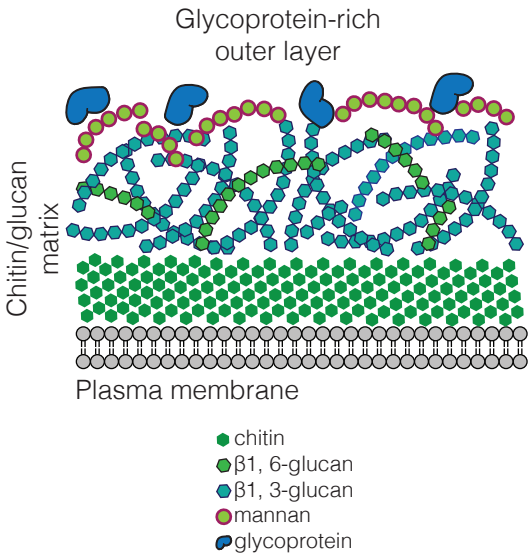
Dictyostelium discoideum



Plants



Fungi



Animals

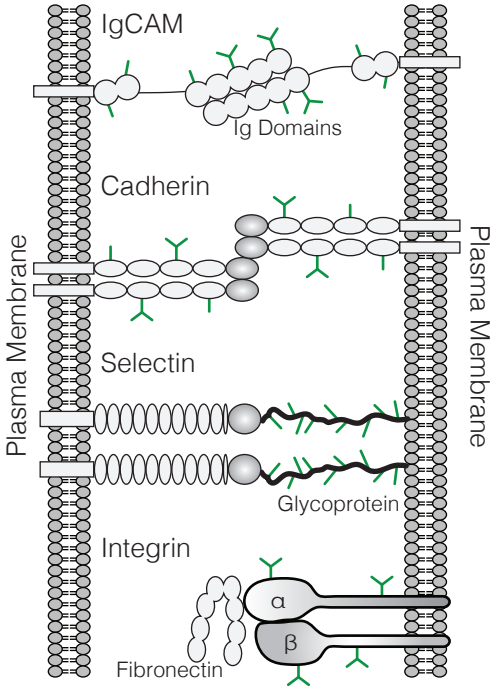
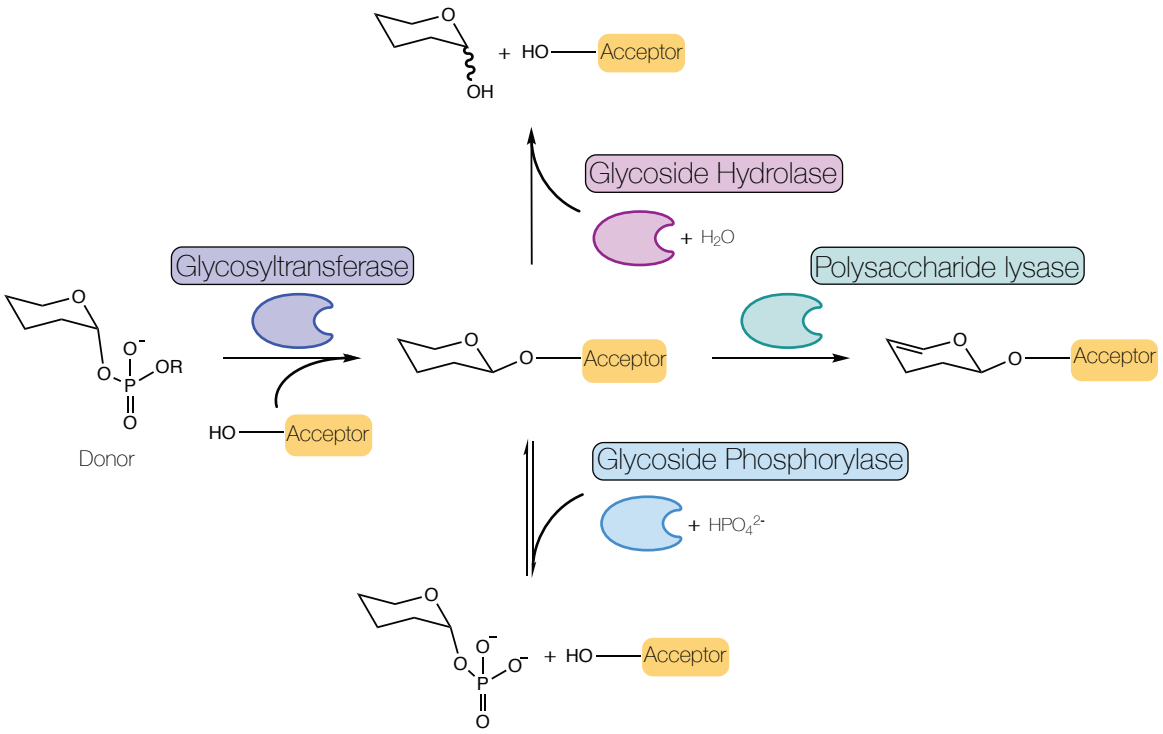


Figure 1.3. Multiple enzymes shape the glycocalyx. Glycans, found on the outside of all cells, also known as the glycocalyx, are derived from the action of four classes of enzymes: glycosyltransferases, glycoside hydrolases, glycoside phosphorylases, and polysaccharide lyases. Glycosyltransferases establish glycosidic linkages by transferring an activated donor sugar substrate that contains a phosphate leaving group to nucleophilic group, normally an alcohol, on an acceptor substrate. Most commonly acceptor substrates are other sugars, but they can also be a lipid, protein, nucleic acid, or other small molecules. Glycoside hydrolases catalyze the hydrolysis of glycosidic bonds. Glycoside phosphorylases catalyze the cleavage of a glycosidic bond through substitution with phosphate. Polysaccharide lyases cleave uronic acid-containing polysaccharides to generate an unsaturated hexuronic acid residue and a new reducing end at the point of cleavage.


+ HO Acceptor
OH
Glycoside Hydrolase
+ H_2O
Glycosyltransferase
Polysaccharide lysase
OR
Donor
HO Acceptor
Acceptor
Acceptor
Glycoside Phosphorylase
+ HPO_4^{2-}
+ HO Acceptor

Chapter 2

Predicted glycosyltransferases promote development and prevent promiscuous cell aggregation in the choanoflagellate *S. rosetta*

The results presented here were published as part of the following paper:
Wetzel L., Levin T., Hulett R.E., Chan D., King G., Aldayafleh R., Booth D., Sigg M.A., & King N. (2018). Glycosyltransferases promote development and prevent promiscuous cell aggregation in the choanoflagellate *S. rosetta*. *bioRxiv. doi: 10.1101/384453* (also in review at *eLife).*

IMPACT STATEMENT

A genetic screen reveals that two predicted glycosyltransferases promote proper rosette development and prevent of cell clumping in one of the closest living relatives of animals, the choanoflagellate *S. rosetta*.

ABSTRACT

In a previous study (Levin *et al.* 2014), we established forward genetics in the choanoflagellate *Salpingoeca rosetta* and found that a C-type lectin gene is required for rosette development. Here we report on critical improvements to genetic screens in *S. rosetta* while also investigating the genetic basis for rosette defect mutants in which single cells fail to develop into orderly rosettes but instead aggregate promiscuously into amorphous clumps of cells. Two of the mutants, Jumble and Couscous, mapped to lesions in genes encoding two different predicted glycosyltransferases and displayed aberrant glycosylation patterns in the basal extracellular matrix (ECM). In animals, glycosyltransferases sculpt the polysaccharide-rich ECM, regulate integrin and cadherin activity, and, when disrupted, contribute to tumorigenesis. The finding that predicted glycosyltransferases promote proper rosette development and prevent cell aggregation in *S. rosetta* suggests a pre-metazoan role for glycosyltransferases in regulating development and preventing abnormal tumor-like multicellularity.

INTRODUCTION

The transition to multicellularity was essential for the evolution of animals from their single celled ancestors (Szathmary and Smith, 1995). However, despite the centrality of multicellularity to the origin of animals, little is known about the genetic and developmental mechanisms that precipitated the evolution of multicellularity on the animal stem lineage. All modern animals develop clonally through serial cell division, suggesting that the same was true for their last common ancestor. While the closest living relatives of animals, choanoflagellates, develop clonally into multicellular rosettes, more distant relatives such as *Capsaspora owczarzaki* (Sebé-Pedrós et al., 2013) and *Dictyostelium discoideum* (Bonner, 1967; Brefeld, 1869) become multicellular through cell aggregation (which is vulnerable to cheating) (Santorelli et al., 2008; Strassmann et al., 2000). This raises a general question of how stem animals might have suppressed cell aggregation in favor of clonal multicellular development.

Although the first animals evolved over 600 million years ago, studying their closest living relatives, choanoflagellates, allows the reconstruction of important aspects of animal origins (Brunet and King, 2017; King et al., 2008; Ruiz-Trillo et al., 2008; Schalchian-Tabrizi et al., 2008; Sebé-Pedrós et al., 2017). *Salpingoeca rosetta* is an emerging model choanoflagellate that was isolated from nature as a spherical colony of cells called a rosette. Under standard laboratory conditions, *S. rosetta* proliferates as solitary cells or as linear chain colonies that easily break apart into solitary cells (Dayel et al., 2011). When exposed to rosette inducing factors (RIFs) produced by the co-isolated prey bacterium *Algoriphagus machipongonensis, S. rosetta* instead develops into highly organized and structurally stable rosettes through a process of serial cell division (Alegado et al., 2012; Dayel et al., 2011; Fairclough et al., 2010; Woznica et al., 2016). Recent advances, including a sequenced genome (Fairclough et al., 2010), the discovery of a sexual phase to the *S. rosetta* life cycle that enables controlled mating (Levin et al., 2014; Levin and King, 2013; Woznica et al., 2017), and techniques that allow for transfection and expression of transgenes (Booth et al., 2018) have enabled increasingly detailed studies of molecular mechanisms underlying rosette development in *S. rosetta*.

In the first genetic screen to identify genes required for rosette formation in *S. rosetta*, multiple rosette defect mutants were recovered that displayed a range of phenotypes (Levin et al., 2014). The first mutant to be characterized in detail was named Rosetteless; while Rosetteless cells did not develop into rosettes in the presence of RIFs, they were otherwise indistinguishable from wild type cells (Levin et al., 2014). The mutation underlying the Rosetteless phenotype was mapped to a C-type lectin, encoded by the gene *rosetteless,* the first gene shown to be required for rosette formation (Levin et al., 2014). In animals, C-type lectins function in signaling and adhesion to promote development and innate immunity (Cambi et al., 2005; Geijtenbeek and Gringhuis, 2009; Ruoslahti, 1996; Švajger et al., 2010; Zelensky and Gready, 2005). Although the molecular mechanisms by which *rosetteless* regulates rosette development remain unknown, the localization of Rosetteless protein to the rosette interior suggests that it functions as part of the extracellular matrix (ECM) (Levin et al., 2014).

Here we report on the largest class of mutants from the original rosette defect screen (Levin et al., 2014), all of which fail to develop into organized rosettes and instead form large, amorphous clumps of cells in both the absence and presence of RIFs. By

mapping the mutations underlying the clumpy, rosette defect phenotypes of two mutants in this class, we identified two predicted glycosyltransferase genes that are each essential for proper rosette development. The causative mutations led to similar perturbations in the glycosylation pattern of the basal ECM. The essentiality of the predicted glycosyltransferases for rosette development combined with prior findings of the requirement of a C-type lectin highlight the importance of the ECM for regulating multicellular rosette development and preventing spurious cell adhesion in a close relative of animals.

RESULTS

Rosette defect mutants form amorphous clumps of cells through promiscuous cell adhesion

The original rosette defect screen performed by Levin et al., 2014 yielded nine mutants that were sorted into seven provisional phenotypic classes. For this study, we screened 21,925 additional clones and identified an additional seven mutants that failed to form proper rosettes in the presence of *Algoriphagus* RIFs. (For this study, we used *Algoriphagus* outer membrane vesicles as a source of RIFs, as described in Woznica et al., 2016). Comparing the phenotypes of the 16 total rosette defect mutants in the presence and absence of RIFs allowed us to classify four broad phenotypic classes: (1) Class A mutants that have wild type morphologies in the absence of RIFs and entirely lack rosettes in the presence of RIFs, (2) Class B mutants that have wild type morphologies in the absence of RIFs and develop reduced levels of rosettes with aberrant structures in the presence of RIFs, (3) Class C mutants that produce large clumps of cells in both the presence and absence of RIFs while forming little to no rosettes in the presence of RIFs, and (4) a Class D mutant that exist primarily as solitary cells, with no linear chains of cells detected in the absence of RIFs and no rosettes detected in the presence of RIFs (Table 2.2).

Of the 16 rosette defect mutants isolated, seven mutants fell into Class C. For this study, we focused on four Class C mutants — Seafoam, Soapsuds, Jumble, and Couscous (previously named Branched in Levin et al., 2014) — that form amorphous, tightly packed clumps of cells, both in the presence and absence of RIFs, but never develop into rosettes (Table 2.1; Figure 2.1A,B). We found that the clumps contain a few to hundreds of mutant cells that pack together haphazardly, unlike wild type rosettes in which all cells are oriented with their basal poles toward the rosette center and their apical flagella extending out from the rosette surface (Alegado et al., 2012; Levin et al., 2014; Woznica et al., 2016). Moreover, in contrast with the structural stability and shear resistance of wild type rosettes (Figure 2.1A) (Levin et al., 2014), the cell clumps formed by Class C mutants were sensitive to shear and separated into solitary cells upon pipetting or vortexing the culture (Figure 2.1A).

Following exposure to shear, we observed that mutant cells re-aggregated into new clumps within minutes, while wild type cells never formed clumps (Figure 2.1C, D; rare cell doublets were likely due to recent cell divisions). Within 30 minutes after disruption by shear force, cell clumps as large as 75, 55, 32, and 23 cells formed in Couscous, Soapsuds, Seafoam, and Jumble mutant cultures, respectively. Moreover, blocking cell division with the cell cycle inhibitor aphidicolin did not prevent clump

formation (Figure 2.2). Both the speed of clump reformation (less than the ~6 hours required for a single cell division (Levin et al., 2014; Figure 2.3) and the observation of cell clumping in the absence of cell division (Figure 2.2) demonstrate that cell aggregation alone is sufficient to drive clump formation. Indeed, each of the mutants tested also displayed a mild defect in cell proliferation (Figure 2.3).

Therefore, the cell clumps are not aberrant rosettes, which never form through aggregation and instead require at least 15 – 24 hours to develop clonally through serial rounds of cell division (Dayel et al., 2011; Fairclough et al., 2010). Nonetheless, we tested whether Jumble and Couscous clump formation might be influenced by the presence or absence of RIFs. Clumps formed in both the presence and absence of RIFs were comparable in size (K-S test; Figure 2.4). Cell aggregation was not strain-specific, as unlabeled Jumble and Couscous mutant cells adhered to wild type cells identified by their expression of cytoplasmic mWasabi (Figure 2.5).

The fact that the seven clumping/aggregating Class C mutants isolated in this screen were also defective in rosette development suggests a direct link between promiscuous cell adhesion and failed rosette development.

Improving genetic mapping in *S. rosetta* through bulk segregant analysis

We next set out to identify the causative mutation(s) underlying the clumping and rosette defect phenotypes in each of these mutants. In the Levin *et al.* 2014 study, the Rosetteless mutant was crossed to a phenotypically wild type Mapping Strain (previously called Isolate B in Levin et al., 2014) and relied on genotyping of haploid F1s at 60 PCR-verified genetic markers that differed between the Rosetteless mutant and the Mapping Strain (Levin et al., 2014). The 60 markers were distributed unevenly across the 55 Mb genome and proved to be insufficient for mapping the Class C mutants for this study. Compounding the problem, the low level of sequence polymorphism among *S. rosetta* laboratory strains and abundance of repetitive sequences in the draft genome assembly (Fairclough et al., 2013; Levin et al., 2014) made it difficult to identify and validate additional genetic markers, while genotyping at individual markers proved labor intensive and costly.

To overcome these barriers, we modified bulk segregation methods developed in other systems (Doitsidou et al., 2010; Leshchiner et al., 2012; Lister et al., 2009; Pomraning et al., 2011; Schneeberger et al., 2009; Voz et al., 2012; Wenger et al., 2010) for use in *S. rosetta.* Our strategy involved: (1) crossing mutants to the Mapping Strain (which contains previously identified sequence variants); (2) isolating heterozygous diploids identified through genotyping at a microsatellite on supercontig 1; (3) inducing meiosis; (4) growing clonal cultures of haploid F1 offspring; (5) phenotyping the F1 offspring; (6) pooling F1 offspring based on their clumping phenotype; and (7) deeply sequencing pooled genomic DNA from the F1 mutants to find mutations that segregated with the clumping phenotype (Figure 2.6).

To test whether a bulk segregant approach would work in *S. rosetta*, we first analyzed a cross between the previously mapped Rosetteless mutant and the Mapping Strain (Levin et al., 2014). We isolated 38 F1s with the rosette defect phenotype from a Mapping Strain×Rosetteless cross (Levin et al., 2014), grew clonal cultures from each, pooled the resulting cultures, extracted their genomic DNA, and sequenced the pooled mutant genomes to an average coverage of 187X. Against a background of sequence

variants that did not segregate with the Rosetteless phenotype, five unlinked single nucleotide variants (SNVs) and insertions/deletions (INDELs) were found to segregate with the phenotype (Table 2.3). Four of these detected sequence variants likely had spurious correlations with the phenotype resulting from relatively low sequencing coverage at those variants (>0.25X coverage of the entire genome) (Table 2.3). In contrast, the remaining SNV was detected in a well-assembled portion of the genome at a sequencing depth approaching the average coverage of the entire genome. The segregating SNV, at position 427,804 on supercontig 8, was identical to the causative mutation identified in Levin et al., 2014 (Table 2.3). Thus, a method based on pooling F1 haploid mutants, identifying sequence variants that segregated with the phenotype, and masking those SNVs/INDELs that were detected with >0.25X coverage of the total genome was effective for correctly pinpointing the causal mutation for Rosetteless (Figure 2.6). Therefore, we used this validated bulk segregant method to map the clumping mutants.

Mapping crosses were carried out for the four clumping/rosette defect mutants characterized in this study (Seafoam, Soapsuds, Jumble, and Couscous) and all four crosses yielded heterozygous diploids, demonstrating that they were competent to mate. As observed in prior studies of *S. rosetta* mating (Levin et al., 2014; Woznica et al., 2017), the diploid cells each secreted a flask-shaped attachment structure called a theca and were obligately unicellular. Therefore, the heterozygous diploids were not informative about whether the mutations were dominant or recessive as the phenotypes could only be detected in haploid cells. For Seafoam and Soapsuds, we isolated heterozygous diploids, but never recovered F1 offspring with the mutant phenotype (Table 2.1). The inability to recover haploids with either clumping or rosette defect phenotypes from the Seafoam×Mapping Strain and Soapsuds×Mapping Strain crosses might be explained by any of the following: (1) the clumping/rosette defect phenotypes are polygenic, (2) meiosis defects are associated with the causative mutations, and/or (3) mutant fitness defects allowed wild type progeny to outcompete the mutant progeny. In contrast, heterozygous diploids from crosses of Jumble and Couscous to the Mapping Strain produced F1 haploid progeny with both wild type and mutant phenotypes and thus allowed for the successful mapping of the causative genetic lesions, as detailed below.

Jumble maps to a putative glycosyltransferase

Following the bulk segregant approach, we identified 5 sequence variants in Jumble that segregated with both the clumping and rosette defects. Only one of these – at position 1,919,681 on supercontig 1 – had sequencing coverage of at least 0.25X of the average sequence coverage of the rest of the genome (Figure 2.7A; Table 2.4). In a backcross of mutant F1 progeny to the Mapping Strain, we confirmed the tight linkage of the SNV to the rosette defect phenotype (Figure 2.7B). Moreover, all F2 progeny that displayed a rosette defect also had a clumping phenotype. Given the tight linkage of both traits with the SNV and the absence of any detectable neighboring sequence variants, we infer that the single point mutation at genome position 1:1,919,681 causes both the clumping and rosette defect phenotypes in Jumble mutants.

The mutation causes a T to C transition in a gene hereafter called *jumble* (GenBank accession EGD72416/NCBI accession XM_004998928; Figure 2.7A). The *jumble* gene contains a single exon and is predicted to encode a 467 amino acid protein

containing a single transmembrane domain. Following the convention established in Levin et al. 2014, the mutant allele, which is predicted to confer a leucine to proline substitution at amino acid position 305, is called *jumble*^{lw1}.

We used recently developed methods for transgene expression in *S. rosetta* (Booth et al., 2018) to test whether expression of a *jumble* with an N- or C-terminal *monomeric teal fluorescent protein* (*mTFP*) gene fusion under the *S. rosetta elongation factor L* (*efl*) promoter could complement the mutation and rescue rosette development in the Jumble mutant (Figure 2.7C,D). We were able to enrich for rare transformed cells by using a plasmid in which the puromycin resistance gene (*pac*) was expressed under the same promoter as the *jumble* fusion gene, with the two coding sequences separated by a sequence encoding a self-cleaving peptide (Kim et al., 2011). Transfection of Jumble mutant cells with wild type *jumble-mTFP* followed by puromycin selection and the addition of RIFs yielded cultures in which 9.33%±5.07% of cells were in rosettes (Figure 2.7C). Similarly, transfection of Jumble with *mTFP-jumble* followed by puromycin selection and rosette induction resulted in cultures with 7.00%±4.91% of cells in rosettes (Figure 2.7C). Importantly, we did not detect any rosettes when we transfected Jumble cells with *mTFP* alone*, jumble*^{lw1}*-mTFP*, or *mTFP*-jumble^{lw1}. Complementation of the Jumble mutant by the wild type *jumble* allele, albeit in a subset of the population, provided further confirmation that the *jumble*^{lw1} mutation causes the cell clumping and rosette defect phenotypes. The fact that the transfection experiment did not allow all cells to develop into rosettes may be due to any number of reasons, including incomplete selection against untransformed cells, differences in transgene expression levels in different transformed cells, and the possibility that the mTFP tag reduces or otherwise changes the activity of the Jumble protein.

We next sought to determine the function and phylogenetic distribution of the *jumble* gene. BLAST searches uncovered unannotated *jumble* homologs in nine other choanoflagellates (Figure 2.8A) and in fungi (Figure 2.9), but none in animals. The choanoflagellate homologs of *jumble* were detected in the transcriptomes of species representing each of the three major choanoflagellate clades (Richter et al., 2018), suggesting that *jumble* evolved before the origin and diversification of choanoflagellates. Although Interpro (Finn et al., 2017) and Pfam (Finn et al., 2016) did not reveal any known protein domains in Jumble, the NCBI Conserved Domain Search (Marchler-Bauer et al., 2017) predicted a glycosyltransferase domain with low confidence (E-value 3.87^{-03}). Moreover, two different algorithms that use predicted secondary and tertiary structures to identify potential homologs, HHphred (Zimmermann et al., 2017) and Phyre2 (Kelly et al., 2015), predict that Jumble is related to well-annotated glycosyltransferases (HHphred: E-value 7.5^{-19} to polypeptide N-acetylgalactosaminyltransferase 4; Phyre2: Confidence 94.5% to human polypeptide n-acetylgalactosaminyltransferase 2) (Figure 2.8B). The Leu305Pro substitution in Jumble^{lw1} disrupts a predicted alpha helix, which we hypothesize would prevent proper folding of the Jumble protein (Figure 2.7A).

Glycosyltransferases play essential roles in animal development (Sawaguchi et al., 2017; L. Zhang et al., 2008) and cell adhesion (Müller et al., 1979; Stratford, 1992). Their biochemical functions include transferring an activated nucleotide sugar, also called a glycosyl donor, to lipid, protein, or carbohydrate acceptors (Lairson et al., 2008). Target acceptors in animals include key signaling and adhesion proteins such as integrins and cadherins, whose activities are regulated by N- and O-linked polysaccharide

modifications, also referred to as N- and O-linked glycans (Larsen et al., 2017; Zhao et al., 2008). Notably, many well-characterized glycosyltransferases act in the Golgi apparatus, where they glycosylate molecules that are trafficked through the secretory system (El-Battari, 2006; Tu and Banfield, 2010). To investigate the localization of Jumble, we transfected wild type cells with a *jumble-mWasabi* gene fusion transcribed under the control of the *S. rosetta efl* promoter. Jumble-mWasabi protein localized to the apical pole of the cell body near the base of the flagellum. Based on comparisons with transmission electron micrographs of *S. rosetta* and other choanoflagellates, Jumble-mWasabi localization corresponds to the location of the Golgi apparatus, for which there is not yet a fluorescent marker in *S. rosetta* (Figure 2.7E,G; Figure 2.10A) (Leadbeater, 2015). In contrast, Jumble^{lw1}-mWasabi, was distributed in a tubular pattern throughout the cell and co-localized with an endoplasmic reticulum (ER) marker (Figure 2.7F,H; Figure 2.10B) (Booth et al., 2018). The ER localization of Jumble^{lw1} is consistent with the hypothesis that the missense mutation disrupts proper protein folding as often misfolded proteins are retained in the ER and targeted for degradation (Kopito, 1997). The failure of the Jumble^{lw1} protein to localize properly at the Golgi apparatus strongly suggests a loss of function.

Couscous maps to a lesion in a predicted mannosyltransferase

We followed a similar strategy to map the genetic lesion(s) underlying the Couscous mutant phenotype. Using the bulk segregant approach on F1 mutant offspring from a Couscous × Mapping Strain cross, we identified eight sequence variants that segregated with the clumping and rosette defect phenotypes, of which only one – a single nucleotide deletion at position 462,534 on supercontig 22 – had sequencing coverage at least 0.25X of the average sequence coverage of the rest of the genome (Figure 2.11A; Table 2.5). The tight linkage of the deletion to both the clumping and rosette defect phenotypes was further confirmed by genotyping the sequence variant in F2 mutants resulting from backcrosses of F1 mutants to the Mapping Strain (Figure 2.11B). Given the tight linkage, we infer that the deletion at position 462,534 on supercontig 22 causes both clumping and the disruption of rosette development in Couscous mutant cells.

The single nucleotide deletion at position 462,534 on supercontig 22 lies in a four-exon gene, hereafter called *couscous* (GenBank accession EGD77026/ NCBI accession XM_004990809). The mutation causes a predicted frameshift leading to an early stop codon in the mutant protein, Couscous^{lw1} (Figure 2.11A). As with the Jumble mutant, we were able to rescue rosette formation in a portion of the population by transfecting cells with either a *couscous-mTFP* or *mTFP-couscous* gene fusion under the *efl* promoter (Figure 2.11C, D), thereby increasing confidence in the mapping results.

The predicted Couscous amino acid sequence contains a specific type of glycosyltransferase domain, an alpha-mannosyltransferase domain, that transfers activated mannose onto the outer chain of core N-linked polysaccharides and O-linked mannotriose (Strahl-Bolsinger et al., 1999). The predicted mannosyltransferase domain shares 28% and 35% amino acid sequence identity to alpha 1-2 mannosyltransferase (MNN2) proteins in *Saccharomyces cerevisiae* and *Candida albicans*, respectively, including the conserved DXD motif found in many families of glycosyltransferases (Wiggins and Munro, 1998) (Figure 2.12A). MNN2 proteins catalyze the addition of the first branch of mannose-containing oligosaccharides found on cell wall proteins (Rayner

and Munro, 1998) and proper MNN2 activity is required for flocculation, or non-mating aggregation, in *S. cerevisiae* (Stratford, 1992). In addition to the mannosyltransferase domain, Couscous is predicted to have a PAN/Apple domain composed of a conserved core of three disulfide bridges (Ho et al., 1998; Tordai et al., 1999). PAN/Apple domains are broadly distributed among eukaryotes, including animals, where they mediate protein-protein and protein-carbohydrate interactions, often on the extracellular surface of the cell (Ho et al., 1998; Tordai et al., 1999).

In wild type cells transfected with a *couscous-mWasabi* transgene under the *efl* promoter, Couscous was found in puncta scattered throughout the cytosol, collar and cell membrane (Figure 2.12B, C). While Couscous-mWasabi was clearly not localized to the Golgi, the puncta may co-localize with the ER, where glycosyltransferases are also known to function (El-Battari, 2006; Tu and Banfield, 2010). However, despite attempting to co-transfect cells with *couscous-mWasabi* and a marker of the ER (mCherry fused to a C-terminal HDEL ER retention signal sequence (Booth et al., 2018)), we were unable to detect any cells expressing both gene fusions. In addition, it is possible that the fusion of Couscous to a fluorescent protein or its overexpression interfered with its proper localization in *S. rosetta.* Therefore, we are currently uncertain about the subcellular localization of Couscous protein.

Jumble and Couscous mutants lack proper sugar modifications at the basal pole

Because both Jumble and Couscous have mutations in predicted glycosyltransferases, we hypothesized that the abundance or distribution of cell surface sugars, called glycans, on Jumble and Couscous mutant cells might be altered. To investigate the distribution of cell surface glycans, we stained live *S. rosetta* with diverse fluorescently labelled sugar-binding lectins. Of the 22 lectins tested, 21 either did not recognize *S. rosetta* or had the same staining pattern in wild type, Jumble and Couscous cells (Table 2.6).

The remaining lectin, jacalin, bound to the apical and basal poles of wild type cells (Figure 2.13A, B, B'). Jacalin also brightly stained the ECM filling the center of rosettes in a pattern reminiscent of the Rosetteless C-type lectin (Levin et al., 2014) (Figure 2.13A, B'), although the two were not imaged simultaneously because jacalin does not bind after cell fixation and labelled Rosetteless antibodies accumulate strongly in the food vacuoles of live cells. In contrast with wild type cells, the basal patch of jacalin staining was absent or significantly diminished in Couscous and Jumble mutants, both in the presence and absence of RIFs (Figure 2.13C-F). Interestingly, the apical patch of jacalin staining in mutant cells appeared similar to wild type cells. This may explain the lack of a clear difference in bands detected with jacalin by western blot between wild type and mutants whole cell lysates (Figure 2.14). Transformation of Jumble cells with *mTFP-jumble* not only rescued rosette development (Figure 2.7C, D), but also restored the wild type glycosylation pattern, as revealed by jacalin staining in the center of complemented rosettes (Figure 2.15). The same was true for Couscous cells, in which transformation with *couscous-mTFP* rescued both rosette development and the wild type glycosylation pattern (Figure 2.11C, D; Figure 2.15). Thus, the glycosylation defects in Jumble and Couscous mutant cells were directly linked to the genetic lesions in *jumble* and *couscous,* respectively.

The loss of basal jacalin staining in Jumble and Couscous mutants indicated that *jumble*^{lw1} and *couscous*^{lw1} either disrupt proper trafficking of sugar-modified molecules to the basal pole of cells or alter the glycosylation events themselves. Thus, we examined whether the basal secretion of Rosetteless protein was disrupted in the mutant strains. In both Jumble and Couscous cells, Rosetteless protein properly localized to the basal pole, but its expression did not increase nor was it secreted upon treatment with RIFs, as normally occurs in wild type cells (Figure 2.16). Because Rosetteless is required for rosette development, this failure to properly upregulate and secrete Rosetteless might contribute to the rosette defect phenotype in Jumble and Couscous cells.

DISCUSSION

Of the 16 rosette defect mutants isolated in Levin *et al.* 2014 and in this study, almost half (7) also display a mild to severe clumping phenotype. This suggests that mechanisms for preventing promiscuous adhesion among wild type cells can be easily disrupted. We found that the clumping phenotype results from promiscuous adhesion of mutant cells to other mutant or wild type cells rather than from incomplete cytokinesis. A recent study revealed that the bacterium *Vibrio fischeri* induces *S. rosetta* to form swarms of cells, visually similar to the mutant clumps, as part of their mating behavior (Woznica et al., 2017). However, it seems unlikely that the clumping Class C mutants is related to swarming; the cell fusion and subsequent settling of diploid cells characteristic of *V. fischeri*-induced mating have not been observed in the class C mutants cultured without *V. fischeri*.

For both Jumble and Couscous, the causative mutations mapped to predicted glycosyltransferase genes. Consistent with its role as a glycosyltransferase, Jumble localized to the Golgi apparatus, but Couscous appeared to localize in cytoplasmic puncta and to the cell membrane. We predict that the mutations in predicted glycosyltransferases are loss of function alleles, given that transfection of mutant *S. rosetta* with the wild type alleles was sufficient to complement each of the mutations. While we have not uncovered the target(s) of the predicted glycotransferases or the exact nature of the interplay between the two phenotypes, disruption of the glycocalyx at the basal pole of both Jumble and Couscous mutant cells (Figure 2.13) hints that the regulation of ECM could play a role in preventing clumping and in promoting proper rosette development.

One possible explanation for the clumping phenotype is that *jumble* and *couscous* are required to regulate the activity of cell surface adhesion molecules and receptors. Glycosylation regulates the activities of two key adhesion proteins in animals: integrins that regulate ECM adhesion, and cadherins that, among their various roles in cell signaling and animal development, bind other cadherins to form cell-cell adhesions called adherens junctions (Larsen et al., 2017; Zhao et al., 2008). Cadherin activity can be either positively or negatively regulated by glycosyltransferases. For example, epithelial cadherin (E-cadherin) is modified by N-acetylglucosaminyltransferase III (GnT-III) whose activity leads to increased cell adhesion and N-acetylglucosaminyltransferase V (GnT-V) whose activity leads to decreased cell adhesion (Carvalho et al., 2016; Granovsky et al., 2000). GnT-V knockdown enhances cell-cell adhesion mediated by E-cadherin and the related N-cadherin (Carvalho et al., 2016; Guo et al., 2009). The inactivation of E-cadherin, including through over- or under- expression of GnT-V or GnT-III, is considered

to be a hallmark of epithelial cancers (Hirohashi and Kanai, 2003). *S. rosetta* expresses 29 different cadherins (Nichols et al., 2012) and it is possible that mutations to *jumble* and *couscous* disrupt regulatory glycosylation of a cell adhesion molecules like cadherins.

Another possibility is that *jumble* and *couscous* add a protective sugar layer to the cell surface and loss of glycosyltransferase activity reveals underlying sticky surfaces. If *jumble* and *couscous* add branches to existed sugar modifications, their loss of function could expose new sugar moieties at the cell surface that act as ligands for lectins that aggregate cells. Lectins mediate cell aggregation in diverse organisms (Colin Hughes, 1992). For example, sponges such as *Geodia cydonium* can be disaggregated into single cells and then reaggregated through lectin binding of a post-translational sugar modification (Müller et al., 1979). In *S. cerevisiae*, the mannosyltransferase MNN2 adds mannose structures to the cell wall that are recognized by aggregating lectins and MNN2 is required for proper flocculation (Rayner and Munro, 1998; Stratford, 1992). Exposing new sugars on the cell surface in Jumble and Couscous could lead to spurious aggregation, potentially by lectins or other sugar binding proteins.

It is somewhat more difficult to infer how increased clumping in single cells might interfere with rosette development. One possibility is that the disruption of ECM glycosylation that we hypothesize might promote clumping may also prevent the proper maturation of the ECM needed for rosette development (Figure 2.17). A prior study showed that only *S. rosetta* cells recognized with the lectin wheat germ agglutinin (WGA) are competent to form rosettes, which suggests that glycosylation might be necessary for rosette development (Dayel et al., 2011). While WGA staining does not appear to be perturbed in Jumble and Couscous (Table 2.6), jacalin staining at the basal pole appears severely reduced or abolished compared to wild type. Jacalin staining was enriched in the center of wild type rosettes in a pattern reminiscent of Rosetteless, which is required for rosette development (Levin et al., 2014). Intriguingly, in Jumble and Couscous, Rosetteless localized to the correct pole, but did not become enriched upon rosette induction, indicating that the ECM did not properly mature. Rosetteless has mucin-like Ser/Thr repeats that are predicted sites of heavy glycosylation and two C-type lectin domains that would be expected to bind to sugar moieties (Levin et al., 2014). Therefore, it is possible that Rosetteless might be regulated either through direct glycosylation or through the glycosylation of potential binding partners by Jumble and Couscous.

The clumping, rosette defect mutants underscore the differences between cell aggregation and a regulated clonal developmental program, such as animal embryogenesis or choanoflagellate rosette development. Importantly, the multicellularity of all extant animals arises through cell division rather than cell aggregation, suggesting that the suppression of cell aggregation in favor of clonal development was a prerequisite for animal origins. Our results raise the possibility that glycosylation and the regulation of the ECM suppressed cell aggregation while stabilizing obligate clonal multicellularity on the animal stem lineage. Glycosylation remains an important regulator of tissue organization in modern animals (L. Zhang et al., 2008). Interestingly, cancer suppression is thought to have been important for ensuring organismal integrity in multicellular animals (Aktipis et al., 2015) and disruption of glycosylation is often implicated in metastatic cancers (Pinho and Reis, 2015). Understanding the molecular mechanisms that prevent spurious aggregation in *S. rosetta* may provide new insights into the mechanisms that

ensured cell cooperativity in stem animals while also revealing cancer vulnerabilities in modern animals.

MATERIALS AND METHODS

Media preparation, strains, and cell culture

Unenriched artificial seawater (ASW), AK artificial seawater (AK), cereal grass media (CG), and high nutrient (HN) media were prepared as described previously (Booth et al., 2018; Levin et al., 2014; Levin and King, 2013). The wild type strain, from which each mutant was generated, was the described strain SrEpac (ATCC PRA-390; accession number SRX365844) in which *S. rosetta* is co-cultured monoxenically with the prey bacterium *Echinicola pacifica* (DSM 19836, Levin et al., 2014; Levin and King, 2013; Nedashkovskaya et al., 2006). Seafoam, Soapsuds, and Couscous (previously named Branched) were generated through X-ray mutagenesis and Jumble was generated by EMS mutagenesis as documented in Levin et al., 2014. In Levin et al., 2014, Branched/Couscous was not thoroughly characterized and was named based on the hypothesis that the clumps formed through cell divisions that resulted in unusually branched chain colonies. (Wild type chain colonies are primarily linear, with rare branches.) Our thorough characterization of the mutant in this study revealed that the clumps form through aggregation and not through branching cell divisions. In order for the mutant name to better reflect the phenotype, we renamed it Couscous. For routine culturing, wild type and mutant cultures were diluted 1:10 every 2-3 days in HN media. The Mapping Strain, (previously called Isolate B in Levin et al., 2014) used for mapping crosses (accession number SRX363839) was grown in the presence of rosette-inducing *A. machipongonensis* bacteria (ATCC BAA-2233). The Mapping Strain was maintained in 25% CG media diluted in ASW and passaged 1:10 every 2-3 days. For transfection of *S. rosetta*, cells were maintained in 5% (vol/vol) HN media in AK seawater (Booth et al., 2018). Rosette formation initially was assayed using both live *A. machipongonensis* and *A. machipongonensis* outer membrane vesicles (OMVs) prepared as in Woznica et al., 2016. For each strain tested, both methods of rosette induction resulted in similarly low/non-existent percentages of cells in rosettes and visually similar clumps for Class C mutants (Table 2.2). Therefore, unless stated otherwise, rosette induction was performed with *A. machipongonensis* OMVs and referred to here as rosette inducing factors (RIFs).

Imaging and quantifying rosette phenotypes

To image rosette phenotypes (Figure 2.1A), cells were plated at a density of $1x10^4$ cells/ml in 3 ml HN media either with or without *Algoriphagus* RIFs. Cultures were imaged after 48 hr of rosette induction in 8-well glass bottom dishes (Ibidi 15 µ-Slide 8 well Cat. No. 80826) that were coated with 0.1 mg/mL poly-D-lysine (Sigma) for 15 min and washed 3 times with water to remove excess poly-D-lysine. For imaging wild type and mutant cultures in the presence and absence of RIFs (Figure 2.1A top two panels), 200 µl of cells were plated with a wide bore pipette tip for minimal disruption and allowed to settle for 5 min. For images of vortexed cells (Figure 2.1A bottom panel), 200 µl of cells were vortexed for 15 s before plating and imaged within 10 min of plating to prevent re-clumping. Cells were imaged live by differential interference contrast microscopy using a Zeiss Axio Observer.Z1/7 Widefield microscope with a Hammatsu Orca-Flash 4.0 LT

CMOS Digital Camera and a 63x/NA1.40 Plan-Apochromatic oil immersion lens with 1.6X optivar setting.

To quantify rosette induction (Figure 2.1B), cells were plated at a density of $1x10^4$ cells/ml in 3 ml HN media with RIFs. After 48 hr, an aliquot of cells was vortexed vigorously for 15 secs and fixed with formaldehyde. To determine the percentage of cells in rosettes, the relative number of single cells and cells within rosettes were scored using a hemocytometer. Rosettes were counted as a group of 3 or more cells with organized polarity relative to a central focus after exposure to vortexing.

Imaging and quantification of cell clumping

Clumps were quantified using a modified protocol from Woznica et al., 2017 (Figure 2.1C; Figure 2.4). Mutant cells, and to some extent wild type cells, will adhere to glass. Therefore, to prevent cells from simply sticking to the bottom of the 8-well glass bottom dishes (Ibidi 15 µ-Slide 8 well Cat. No. 80826), the dishes were coated with 1% BSA for 1 hr and washed 3 times with water to remove any residual BSA. Importantly, the addition of BSA to the imaging dishes did not cause wild type cells to stick to the bottom of the dishes or to each other. Cells were diluted to $5x10^5$ cells/mL, vortexed for 15 s to break apart any pre-formed clumps and plated in the BSA pre-treated dishes. For quantification, DIC images were taken using a Zeiss Axio Observer.Z1/7 Widefield microscope with a Hammatsu Orca-Flash 4.0 LT CMOS Digital Camera and a 20x objective. Images were collected for each strain from 10 distinct locations throughout the well.

Images were batch processed in ImageJ for consistency. To accurately segment the phase bright cells and limit signal from the phase dark bacteria the following commands were applied with default settings: ‘Smooth’ (to reduce background bacterial signal), ‘Find Edges’ (to highlight the phase-bright choanoflagellate cells), ‘Despeckle’ (to remove noise), ‘Make Binary’ (to convert to black and white), ‘Dilate’ (to expand to smooth jagged edges from segmentation), ‘Erode’ (to return to the same size as before dilate), and ‘Fill Holes’ (to fill any remaining small holes). Finally, images were analyzed with the ‘Analyze Particles’ command to calculate the area of the clump and only particles larger than 20 μm^2 were kept to filter out any remaining bacterial signal. Cell equivalents/clump (Figure 2.1C and Figure 2.4, right y axis) were calculated by dividing the area of the clump by the area of a representative individual cell (as approximated by averaging the area of the wild type cells). Data are presented as violin boxplots, showing the median cell number (middle line), interquartile range (white box), and range excluding outliers (thin line). A minimum of 630 clumps from two biological replicates were measured for each condition.

To examine whether cell division was required for clump formation, we used aphidicolin to block cell division (Figure 2.2). Cells in vortexed wild type, Jumble, and Couscous cultures were counted and diluted to $1x10^5$ cells/mL in 5%(vol/vol) HN media in AK seawater. For each strain, either 250 µM aphicidolin, an equal volume of a DMSO control, or no additional control were added to each condition. After 24 h, DIC images were taken using a Zeiss Axio Observer.Z1/7 Widefield microscope with a Hammatsu Orca-Flash 4.0 LT CMOS Digital Camera and a 40x air objective.

Performing mapping crosses

Mapping crosses for each mutant strain (Seafoam, Soapsuds, Jumble, and Couscous) with Mapping Strain (previously described as Isolate B) were attempted using both methods previously shown to induce mating in *S. rosetta*: nutrient limitation for 11 days and addition of 2.5-5% *Vibrio fischeri* (ATCC 700601) conditioned media (Levin and King, 2013; Woznica et al., 2017). Both methods were effective at inducing mating for all attempted crosses; here, we report which method was used to generate data for each individual cross. Cells induced to mate were plated by limiting dilution to isolate diploid clones. Clonal isolates were allowed to grow for 5-7 days and screened for populations of thecate cells, as these are the only documented diploid cell type (Levin et al., 2014; Woznica et al., 2017). From each population of thecate cells, we extracted DNA from 75 µl of cells by scraping cells from the plate, harvesting and pelleting the cells, resuspending in 10 µl of base solution (25 mM NaOH, 2 mM EDTA), transferring samples into PCR plates, boiling at 100°C for 20 min, followed by cooling at 4°C for 5 min, and then adding 10 µl Tris solution (40 mM Tris-HCl, pH 7.5). We used 2 µl of this sample as the DNA template for each genotyping reaction. We identified heterozygous strains through genotyping by PCR at a single microsatellite genotyping marker at position 577,135 on supercontig 1 (Forward primer: GACAGGGCAAAACAGACAGA and Reverse primer: CCATCCACGTTCATTCTCCT) that distinguishes a 25 bp deletion in the Mapping Strain (199 bp) from the strain used to generate the mutants (217 bp). Isolates containing PCR products of both sizes were inferred to be diploid. Meiosis was induced by rapid passaging every day in CG medium.

For both Seafoam and Soapsuds, we were able to generate putative outcrossed diploids by crossing to the Mapping Strain based on the genotyping marker on supercontig 1, but we only could only clonally isolate populations of F1 haploids with rosettes and never isolated any F1 haploids with the clumpy, rosetteless phenotype. Whole genome resequencing of Seafoam and Soapsuds revealed no mutations at the *rosetteless*, *jumble,* or *couscous* loci. Seafoam and Soapsuds have 17 and 34 predicted nonsense or missense mutations, respectively, in coding sequences and additional mutations in non-coding portions of the genome. Of the lesions in Seafoam and Soapsuds, none were detected in genes encoding a predicted glycosyltransferase, lectin, or related gene family. Without being able to do mapping crosses, it was not possible to identify the causative mutations from Seafoam or Soapsuds.

For the successful cross of Jumble to the Mapping Strain, we induced mating by starvation using the approach of Levin and King 2013. First, we started with rapidly growing, regularly passaged strains, pelleted $2x10^6$ cells/mL of each strain together and resuspended in 10mL of ASW lacking any added nutrients. After 11 days of starvation in ASW, we pelleted all cells (presumably including diploid cells resulting from mating) and resuspended in 100% CG media to recover any diploids. After 3 days of recovery, we isolated clones by limiting dilution in 10% CG media in ASW (vol/vol). The probability of clonal isolation in this step was 0.91-0.93 (calculated using the Poisson distribution and the number of choanoflagellate-free wells per plate; Levin and King, 2013). Three clonally isolated heterozygous populations, each containing almost exclusively thecate cells, were identified through genotyping by PCR at a supercontig 1 microsatellite as described above. To induce meiosis, heterozygotes were diluted 1:2 in 25% CG media in ASW (vol/vol) every 1-2 days for 8 days. As soon as rosettes and swimming cells were

observed, we repeated the serial dilution to isolate clones (probability of clonal isolation 0.85-0.98). We collected any clonally isolated populations that formed rosettes or clumps and ignored any wells containing thecate cells assuming that these represented diploid cells that had not undergone meiosis. 56% of all non-thecate isolates displayed the cell clumping phenotype and 44% of all non-thecate isolates were capable of forming rosettes, consistent with Mendelian segregation of a single locus (X^2=1.162, df=1, p=0.28). Isolates were genotyped with the marker on supercontig 1 to ensure that independent assortment of the genotype and the phenotype indeed occurred. In total, 30 clumpy F1s were collected for bulk segregation analysis.

For the successful Couscous cross, we induced mating using *V. fisheri* conditioned media using the approach of Woznica et al., 2017. A mixture of $1x10^6$ Couscous and Mapping Strain cells at stationary growth were pelleted and resuspended in 5% *V. fischeri* conditioned media in ASW (vol/vol). After 24 hr, the cells were pelleted, resuspended in 5% HN media in ASW (vol/vol), and allowed to recover for 24 hr. We then isolated clones by limiting dilution in 10% CG media in ASW (vol/vol). The probability of clonal isolation in this step was between 0.97-0.98. We extracted DNA as described above and identified heterozygous clones through genotyping by PCR at a single microsatellite genotyping marker on supercontig 1. Four clonally isolated heterozygous populations, containing almost exclusively thecate cells, were identified. To induce meiosis, heterozygotes were passaged 1:2 in 25% CG media in ASW (vol/vol) every 1-2 days for 8 days. As soon as rosettes and swimming cells were observed, we repeated clonal isolation (probability of clonal isolation 0.78-0.97). We collected any clonally isolated populations that formed rosettes or clumps and ignored any wells containing thecate cells assuming that these represented diploid cells that had not undergone meiosis. Only 14.6% of non-thecate isolates were clumps; this deviation from a Mendelian ratio (X^2=225.63, df=1, $p<5.34^{-51}$) may indicate a potential fitness defect of the mutant phenotype. Isolates were genotyped with the marker on supercontig 1 to ensure that independent assortment indeed occurred. In total, 22 clumpy F1s were collected for bulk segregant analysis.

Whole genome sequencing

Jumble, Couscous, Seafoam, and Soapsuds were whole genome sequenced individually to identify the mutation(s) carried in each strain. To do this, Jumble, Couscous, Seafoam, and Soapsuds cells were grown to stationary phase in 500 mL of 5% HN media in ASW (vol/vol). To generate pooled genomic DNA for bulk segregant analysis, we grew up $5x10^6$ cells of each of the 38 F1s with the rosetteless phenotype from the Rosetteless×Mapping Strain cross (Levin et al., 2014), $5x10^6$ cells of each of the 30 F1s with the clumpy phenotype from the Jumble×Mapping Strain cross, and $5x10^6$ cells of each of the 22 F1s with the clumpy phenotype from the Couscous×Mapping Strain cross. For each cross, the F1 cells were pelleted, frozen, and combined during lysis for DNA extraction. For all samples, we performed a phenol-chloroform DNA extraction and used a CsCl gradient to separate *S. rosetta* DNA from contaminating *E. pacifica* DNA by GC content (King et al., 2008).

Multiplexed, 100 bp paired-end libraries were prepared and sequenced on an Illumina HiSeq 2000 for the Jumble, Couscous, Seafoam, and Soapsuds mutant DNA alone. Multiplexed, 150 bp paired-end libraries were prepared and sequenced on an Illumina HiSeq 2500 for the Rosetteless×Mapping Strain cross and the Jumble x Mapping

Strain cross pooled DNA. For the Couscous×Mapping Strain cross DNA, a multiplexed, 300 bp paired-end library was prepared and sequenced on an Illumina MiSeq. Raw reads are available at the NCBI Short Read Archive with the BioProject identifier PRJNA490902. BioSample and SRA accession numbers are as follows: Jumble mutant-SAMN10061445 and SRR7866767, Couscous mutant-SAMN10061446 and SRR7866768, Seafoam mutant-SAMN10501893 and SRR8263910, Soapsuds mutant-SAMN10501894 and SRR8263909, Rosetteless×Mapping Strain cross-SAMN10061447 and SRR7866769, Jumble×Mapping Strain cross-SAMN10061448 and SRR7866770, and Couscous×Mapping Strain cross- SAMN10061449 and SRR7866771. Raw reads were trimmed with TrimmomaticPE (Bolger et al., 2014) to remove low quality base calls. Trimmed reads were mapped to the *S. rosetta* reference genome (Fairclough et al., 2013) using Burrows-Wheeler Aligner (Li and Durbin, 2009), and we removed PCR duplicates with Picard (http://broadinstitute.github.io/picard/). We realigned reads surrounding indel calls using GATK (Depristo et al., 2011) and called variants using SAMtools and bcftools (Li et al., 2009).

Bulk segregant sequencing analysis

No large region of the genome (i.e. haplotype block) was found to co-segregate with the mutant phenotype in any of the crosses, likely because of the sparse, uneven distribution of genetic markers and/or high recombination rates. Sequence variants from the pooled samples were culled using vcftools vcf-isec (Danecek et al., 2011): (1) to keep only any sequence variants in the pooled samples that were shared with the parental mutant strain since any causative mutations should be present in both the pooled sample and the parental mutant strain, and (2) to remove any sequence variants in the pooled samples that were shared with the Mapping Strain (Isolate B), wild type (previously Isolate C), or the unmutagenized control from the Rosetteless mutagenesis (C2E5) since any of these sequence variants should not be causative for rosette defects (Levin et al., 2014; Levin and King, 2013). The remaining variants were filtered by quality: depth >2, quality score >10, and reference allele not N. The remaining list represents high quality variants in the pooled population that are shared with the mutant to the exclusion 3 different strains competent to form rosettes. Segregating variants were determined by dividing the number of reads that map to the alternative allele by the total number of high quality reads determined by SAMtools and bcftools (Li et al., 2009); any variants with >99% of reads that map to the alternative allele were considered variants that segregated with the mutant phenotype.

Backcrosses

To test the linkage of clumpy phenotype and the predicted causative mutation from the bulk segregant analysis, F1s with the clumpy phenotype from the Jumble×Mapping Strain and Couscous×Mapping Strain were backcrossed to the Mapping Strain. For the Jumble F1 backcross, $1x10^6$ cells grown up from a clonally isolated F1 with the clumpy phenotype from Jumble×Mapping Strain and $1x10^6$ Mapping Strain cells were mixed, pelleted, and resuspended in 10 mL of 5% *V. fischeri* conditioned media in ASW (vol/vol). After 24 hr, the *V. fischeri* conditioned media was replaced with 25% CG media in ASW (vol/vol) and cells were plated to limiting dilution. Clonally isolated thecate populations were genotyped by PCR of the microsatellite on supercontig 1 as described above and 4

heterozygous diploids populations were identified (probability of clonal isolation 0.79-0.95). The heterozygotes were rapidly passaged for 2 weeks to induce meiosis before being plated for clonal isolation (probability of clonal isolation 0.95-0.98). 12 F2s with the clumpy phenotype and 9 F2s with the rosette phenotype were identified (Figure 2.7B). Their DNA was extracted using Base-Tris method described above and the region around the causal mutation was amplified. The resultant PCR product was digested for 4 hr with BfaI, which cleaves the mutant allele but not the wild type allele, and products of the digest were distinguished by agarose gel electrophoresis.

For the two Couscous F1 backcrosses, $2.5x10^5$ cells from either one of two F1s with the clumpy phenotype from Couscous×Mapping Strain cross and $2.5x10^5$ Mapping Strain cells were mixed, pelleted, resuspending in 0.5 mL of 2.5% *V. fischeri* conditioned media in ASW (vol/vol). After 24 hr, *V. fischeri* conditioned media was replaced with 25% CG media in ASW (vol/vol) and cells were plated to limiting dilution (probability of clonal isolation 0.85-0.97). Clonally isolated thecate populations were genotyped by PCR of the microsatellite on supercontig 1 as described above and 3 heterozygous diploids (6 total) were identified in each cross. Isolates were rapidly passaged for 2 weeks to induce meiosis before being plated for clonal isolation (probability of clonal isolation 0.88-0.97). 51 F2s with the clumpy phenotype and 38 F2s with the rosette phenotype were identified (Figure 2.11B); their DNA was extracted using Base-Tris method described above, the region around the causal mutation was amplified, and the resultant PCR product was Sanger sequenced.

Jumble and Couscous domain and structure prediction and alignment

Protein domains encoded by *jumble* (Figure 2.7A) and *couscous* (Figure 2.11A) were predicted using Interpro (Finn et al., 2017), PFAM (Finn et al., 2016), and the NCBI Conserved Domain Search (Marchler-Bauer et al., 2017). Structural homology analysis of Jumble was performed with Phyre2 (Kelly et al., 2015) and HHphred (Zimmermann et al., 2017). The structure of the human N-acetylgalactosaminyltransferase 4 (GlcNAc T4) catalytic domain (HHphred: E-value 7.5^{-19}) was aligned to the predicted Jumble structure generated by HHphred using the PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC (Figure 2.8B). Other choanoflagellate homologs of *jumble* were determined by reciprocal BLAST of the 20 sequenced choanoflagellate transcriptomes (Richter et al., 2018) and alignment was performed with ClustalX (Larkin et al., 2007) (Figure 2.8A). Four fungal homologs [*Saitoella complicata* (NCBI accession XP_019021578.1), *Dactylellina haptotyla* (NCBI accession EPS43829.1), *Naematelia encephala* (NCBI accession ORY22834.1), and *Tuber magnatum* (NCBI accession PWW71609.1)] were identified by best reciprocal BLAST using the *S. rosetta* Jumble protein sequence and aligned with ClustalX (Larkin et al., 2007) (Figure 2.9).The alignment of Couscous to yeast MNN2 glycosyltransferase domains was performed with ClustalX (Larkin et al., 2007) (Figure 2.12A).

Generating transgenic constructs

Jumble (GenBank accession EGD72416/NCBI accession XM_004998928) and Couscous (GenBank accession EGD77026/ NCBI accession XM_004990809) were cloned from wild type cDNA prepared as described in Booth et al., 2018. Jumble^{lw1} was cloned from cDNA prepared from the Jumble mutant. Couscous^{lw1} could not be cloned

from cDNA directly (possibly because of low mRNA levels due to nonsense mediate decay or simply because of high GC content of the gene). However, the 1 bp deletion in *Couscous^{lw1}* was confirmed by Sanger sequencing of genomic Couscous DNA. Site directed mutagenesis of the wild type gene was used to generate the mutant allele.

For complementation (Figure 2.7C,D and 3C,D), constructs were generated from a plasmid with a pUC19 backbone with a 5' *S. rosetta* elongation factor L (*efl)* promoter, monomeric teal fluorescent protein (*mTFP)*, and the 3' UTR from actin (Addgene ID NK633) (Booth et al., 2018). A puromycin resistance gene was synthesized as a gene block and codon optimized for *S. rosetta*. The puromycin resistance gene *(pac)* was inserted after the *efl* promoter and separated from fluorescent reporters by self-cleaving 2A peptide from the porcine virus (P2A) (Kim et al., 2011). Copies of *jumble, jumble^{lw1}, couscous,* and *couscous^{lw1}*were inserted either 5' or 3' of the mTFP and separated from mTFP by a flexible linker sequence (SGGSGGS) through Gibson cloning.

For fluorescent localization (Figure 2.7E-H, Figure 2.10B, Figure 2.12B,C), constructs were generated from a pUC19 backbone with a 5' *S. rosetta* elongation factor L (*efl)* promoter, mWasabi, and 3' UTR from actin. Copies of *jumble* (Addgene ID NK690)*, jumble^{lw1}* (Addgene ID NK691)*,* and *couscous* (Addegene ID NK692) were inserted either 5' of the mWasabi separated by a flexible linker sequence (SGGSGGS) through Gibson cloning. Plasma membrane and ER markers from Booth et al., 2018 were used as previously described (Addgene ID NK624 and NK644).

S. rosetta transfection and transgene expression

Transfection protocol was followed as described in Booth et al., 2018 (http://www.protocols.io/groups/king-lab). Two days prior to transfection, a culture flask (Corning, Cat. No. 353144) was seeded with Jumble, Couscous, or wild type cells at a density of 5,000 cells/ml in 200 ml of 1x HN Media. After 36-48 hr of growth, bacteria were washed away from the cells in three consecutive rounds of centrifugation and resuspension in sterile AK seawater. After the final wash, the cells were resuspended in a total volume of 100 µl AK and counted on a Luna-FL automated cell counter (Logos Biosystems). The remaining cells were diluted to a final concentration of $5x10^7$ cells/ml and divided into 100 µl aliquots. Each aliquot of cells pelleted at 2750 x g, resuspend in priming buffer (40 mM HEPES-KOH, pH 7.5; 34 mM Lithium Citrate; 50 mM L-Cysteine; 15% (w/v) PEG 8000; and 1 µM papain), and incubated at room temperature for 30 mins to remove extracellular material coating the cells. Priming buffer was quenched with 50 mg/ml bovine serum albumin-fraction V (Sigma). Cells were pelleted at 1250 x g and resuspend in 25 µl of SF buffer (Lonza). Each transfection reaction was prepared by adding 2 µl of "primed" cells to a mixture of 16 µl of SF buffer, 2 µl of 20 µg/ µl pUC19; 1 µl of 250 mM ATP, pH 7.5; 1 µl of 100 mg/ml Sodium Heparin; and 1 µl of each reporter DNA construct at 5 µg/µl. Transfections were carried out in 96-well nucleofection plate (Lonza) in a Nucleofector 4d 96-well Nucleofection unit (Lonza) with the CM-156 pulse. Immediately after nucleofection, 100 µl of ice-cold recovery buffer (10 mM HEPES-KOH, pH 7.5; 0.9 M Sorbitol; 8% (w/v) PEG 8000) was added to the cells and incubated for 5 min. The whole volume of the transfection reaction plus the recovery buffer was transferred to 1 ml of 1x HN media in a 12-well plate. After cells recovered for 1 hr, 5 µl of a 10 mg frozen *E. pacifica* pellet resuspend in 1 ml of AK seawater was added to each well and RIFs were added if looking at rosette induction.

Transgenic Complementation

For complementation, Jumble mutants were transfected with the following plasmids: (1) *pEFl5'-Actin3'::pac-P2A-Jumble-mTFP* (Addgene ID NK694), (2) *pEFl5'-Actin3'::pac-P2A-Jumble^{lw1}-mTFP* (Addgene ID NK695), (3) *pEFl5'-Actin3'::pac-P2A-mTFP-Jumble* (Addgene ID NK696), (4) *pEFl5'-Actin3'::pac-P2A-mTFP-Jumble^{lw1}* (Addgene ID NK697), and (5) *pEFl5'-Actin3'::pac-P2A-mTFP* (Addgene ID NK676); and Couscous with the following plasmids: (1) pEFl5'-Actin3'::pac-P2A-Couscous-mTFP (Addgene ID NK698), (2) *pEFl5'-Actin3'::pac-P2A-Couscous^{lw1}-mTFP* (Addgene ID NK699), (3) *pEFl5'-Actin3'::pac-P2A-mTFP-Couscous* (Addgene ID NK700), (4) *pEFl5'-Actin3'::pac-P2A-mTFP-Couscous^{lw1}* (Addgene ID NK701), and (5) *pEFl5'-Actin3'::pac-P2A-mTFP* (Addgene ID NK676). Transformed cells were grown an additional 24 hr after transfection to allow for transgene expression, and then 40 µg/ml puromycin was added for selection. Selection occurred for 48 hr before rosette induction was counted by hemocytometer. After vortexing for 15 sec and fixing with formaldehyde, 200 cells of each transfection well were counted on a hemocytometer to determine percentage of cells in rosettes (Figure 2.7C, Figure 2.11C). Complementation was repeated on 2 biological replicates with 3 technical transfection replicates each. Representative rosette images (Figure 2.7D, Figure 2.11D) were taken on by confocal microscopy using Zeiss Axio Observer LSM 880 a C-Apochromat 40x/NA1.20 W Korr UV-Vis-IR water immersion objective.

Live cell imaging

Glass-bottom dishes for live cell imaging were prepared by corona-treating and poly-D-lysine coating as described in Booth et al., 2018. Transfected cells were prepared for microscopy by pelleting 1-2 ml of cells and resuspend in 200 µl of 4/5 ASW with 100 mM LiCl to slow flagellar beating. Cells were plated on glass-bottom dishes and covered by 200 µl of 20% (w/v) Ficoll 400 dissolved in 4/5 ASW with 100 mM LiCl. Confocal microscopy was performed on a Zeiss Axio Observer LSM 880 with an Airyscan detector and a 63x/NA1.40 Plan-Apochromatic oil immersion objective.
Confocal stacks were acquired in super-resolution mode using ILEX
Line scanning and two-fold averaging and the following settings: 35 nm x 35 nm pixel size, 100 nm z-step, 0.9-1.0 µsec/pixel dwell time, 850 gain, 458 nm laser operating at 1-6% laser power, 561 nm laser operating at 1-2% laser power, 458/561 nm multiple beam splitter, and 495-550 nm band-pass/570 nm long-pass filter. Images were processed using the automated Airyscan algorithm (Zeiss).

Lectin staining and jacalin quantification

All fluorescein lectins from kits I, II, and III from Vector Lab (FLK-2100, FLK-3100, and FLK-4100) were tested for recognition in wild type, Jumbled, and Couscous (Table 2.6). Cells were plated on poly-D-Lysine coated wells of a 96-well glass bottom plate, lectins were added at a concentration of 1:200 and imaged immediately using Zeiss Axio Observer.Z1/7 Widefield microscope with a Hammatsu Orca-Flash 4.0 LT CMOS Digital Camera and a 20x objective. For further jacalin image analysis (Figure 2.13), cells were plated on a poly-D-Lysine coated glass bottom dish, 1:400 fluorescein labelled-jacalin and 1:200 lysotracker Red DN-99 (overloaded to visualize the cell body) and were imaged

immediately by confocal microscopy using Zeiss Axio Observer LSM 880 a 63x/NA1.40 Plan-Apochromatic oil immersion objective. Images were taken with the following settings: 66 nm x 66 nm pixel size, 64 nm z-step, 0.34 µsec/pixel dwell time, 488 nm laser operating at 0.2% laser power with 700 master gain, and 561 nm laser operating at 0.0175% laser power with 750 master gain. Fifteen unique fields of view chosen based on lysotracker staining. Induced cells were treated with OMVs 24 hr before imaging.

To process images, Z-stack images were max projected using ImageJ. Individual cells were chosen based on the ability to clearly see a horizontally oriented collar by lysotracker and cropped to only include a single cell. The maximum fluorescence intensity pixel of the jacalin channel was determined for the cropped image and was used to normalize the fluorescence intensity. To measure jacalin staining around the cell body, a line was drawn using only the lysotracker staining from the point where the collar and the cell body meet on one side of the cell around the cell to the other and the fluorescence intensity was measured along the line. To compare between cells, the lines drawn around the cell body were one-dimensional interpolated in R to include 150 points and normalized to the length of the line. The average fluorescence intensity was plotted over the length of the line drawn around the cell body for Jumble, Couscous, and wild type induced and uninduced with a 95% confidence interval (Figure 2.13F). Measurements were taken from two biological replicates with at least 59 cells in total from each condition.

To examine jacalin localization for the Jumble and Couscous rescue experiments (Figure 2.15), fluorescein-conjugated jacalin could not be used due to its overlapping emission spectrum with the mTFP fusion protein used for complementation. Therefore, cells were incubated with 1 mg/mL biotinylated jacalin (Vector Labs, Cat. No. B-1155) for 5 min at room temperature and pelleted at 3000xg for 5 min. Once the supernatant was removed, the cells were incubated with 1:1000 Streptavidin Alexa Fluor 647 conjugate (Thermo Fisher Scientific, Cat. No. 32357) for 5 min at room temperature to fluorescently label the jacalin. The cells were then pelleted at 3000xg for 5 min, the supernatant was removed, and the cells were resuspended in ASW and plated for imaging. Jacalin localization was imaged by confocal microscopy using a Zeiss Axio Observer LSM 880 with a 63x/NA1.40 Plan-Apochromatic oil immersion objective.

Wild type and mutant clumping assays

Wild type cells transfected with the puromycin resistance gene and mWasabi separated by the P2A self-cleaving peptide under the *efl* promoter and maintained in 40 µg/mL puromycin to enrich for positive transformants. For clumping assays, equal numbers of mWasabi-wt cells either uninduced or induced to form rosettes were mixed with either Jumble or Couscous, vortexed, and plated on BSA treated 8-well glass bottom dishes. DIC and fluorescent images were obtained after 30 mins using Zeiss Axio Observer.Z1/7 Widefield microscope with a Hammatsu Orca-Flash 4.0 LT CMOS Digital Camera and a 40x/NA1.40 Plan-Apochromatic lens (Figure 2.5).

Wild type and mutant growth curves

All cells strains were plated at a density of $1x10^4$ cells/ml in 3 ml HN media. Every 12 hr an aliquot of cells was vortexed vigorously for 15 sec, fixed with formaldehyde, and counted by hemacytometer. Curves were generated from the average ± SD from 2 biological replicates with 3 technical replicates each (Figure 2.3).

Jacalin Western blot

Whole cell lysates were made from pelleting $1x10^7$ cells at 4C at 3,000 x g and resuspending in lysis buffer (20 mM Tris-HCl, pH 8.0; 150 mM KCl; 5 mM MgCl2; 250 mM Sucrose; 1 mM DTT; 10 mM Digitonin; 1 mg/ml Sodium Heparin; 1 mM Pefabloc SC; 0.5 U/µl DNaseI; 1 U/µl SUPERaseIN). Cells were incubated in lysis buffer for 10 min on ice and passed through 30G needle 5x. Insoluble material was pelleted at 6,000 x g for 10 min at 4C. Lysate ($1x10^6$ cells/sample) was run on 4-20% TGX mini-gel (Bio-Rad) for 45 min at 200 V and transferred onto 0.2 µm nitrocellulose membrane using Trans-Blot Turbo Transfer System (Bio-Rad) with semi-dry settings 25V for min. The blot was blocked for 30 min with Odyssey PBS Block (Li-cor). The blot was probed with biotinylated jacalin (1:4,000; Vector Labs) and E7 anti-tubulin antibody (1:10,000; Developmental Studies Hybridoma Bank) diluted in block for 1 hr, and then with IRDye 800 streptavidin (1:1,000; Li-cor) and IRDye 700 mouse (1:1,000; Li-cor) in PBST [PBS with %1 Tween 20 (v/v)]. Blot was imaged on Licor Odyssey (Figure 2.14).

Rosetteless immunofluorescence staining and imaging

Immunofluorescence (Figure 2.16) was performed previously described in Levin et al., 2014 with the modifications for better cytoskeleton preservation described in Booth et al., 2018. Two mL of dense wild type, Jumble, and Couscous cells, that were either uninduced or induced with RIFs for 24 hr, were allowed to settle on poly-L-lysine coated coverslips (BD Biosciences) for 30 min. Cells were fixed in two steps: 6% acetone in cytoskeleton buffer (10 mM MES, pH 6.1; 138 KCl, 3 mM $MgCl_2$; 2 mM EGTA; 675 mM Sucrose) for 5 and 4% formaldehyde with diluted in cytoskeleton buffer for 20 min. The coverslips were gently washed three times with cytoskeleton buffer. Cells were permeabilized with permeabilization buffer [100 mM PIPES, pH 6.95; 2 mM EGTA; 1 mM $MgCl_2$; 1% (w/v) bovine serum albumin-fraction V; 0.3% (v/v Triton X-100)] for 30 min. Cells were stained with the anti-Rosetteless genomic antibody at 3.125 ng/µl (1:400), E7 anti-tubulin antibody (1:1000; Developmental Studies Hybridoma Bank), Alexa fluor 488 anti-mouse and Alexa fluor 647 anti-rabbit secondary antibodies (1:1000 each; Molecular Probes), and 6 U/ml rhodamine phalloidin (Molecular Probes) before mounting in Prolong Gold antifade reagent with DAPI (Molecular Probes).

Images were acquired on a Zeiss LSM 880 Airyscan confocal microscope with a 63x objective (as described for live cell imaging) by frame scanning in the super-resolution mode with the following settings: 30 nm x 30 nm pixel size; 100 nm z-step; 561 nm laser operating at 1.5% power with 700 master gain, and 488 nm laser operating at 2.0% power with 800 master gain. Wild type rosettes were imaged with 633 nm laser operating at 0.3% laser power and 650 master gain to prevent overexposure of Rosetteless, but all other conditions were operating at 2% laser power and 650 master gain in the 633 nm channel.

ACKNOWLEDGEMENTS

Hannah Elzinga, Lily Helfrich, and Max Coyle helped with experiments and reagent preparation. We thank Iswar Hariharan and members of the King lab for helpful discussions, research support, mutant naming suggestions, and comments on the

manuscript, especially Kayley Hake, Ben Larson, Tess Linden, and Thibaut Brunet. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

Figure 2.1. Mutant cells aggregate and fail to form rosettes. (A) Wild type cells are unicellular or form linear chains in the absence of rosette inducing factors (RIFs) and develop into organized spherical rosettes. Rosettes are resistant to shear force and survive vortexing. Four class C mutants — Seafoam, Soapsuds, Couscous, and Jumble — form disordered clumps of cells in the presence and absence of RIFs. The clumps are not resistant to vortexing and fall apart into single cells. **(B)** Class C mutants do not form any detectable rosettes. Rosette development was measured as the % of cells in rosettes after 48 hr in the presence of RIFs and is shown as mean ± SEM. n.d. = no detected rosettes. **(C)** Class C mutants quickly aggregated into large clumps after disruption by vortexing. After vortexing, wild type and mutant cells were incubated for 30 minutes in the absence of RIFs and clump sizes were quantified by automated image analysis. Data are presented as violin boxplots, showing the median cell number (horizontal line), interquartile range (white box), and range excluding outliers (vertical line). All mutants had significantly larger masses of cells (K-S test, ****p<0.0001) than found in cultures of wild type cells. **(D)** Clumping occurred within minutes after vortexing in the Class C mutants without RIFs, revealing that the clumps form by aggregation and not through cell division. DIC images obtained at 0, 15, and 30 minutes post-vortexing. Scale bar = 20 μm.

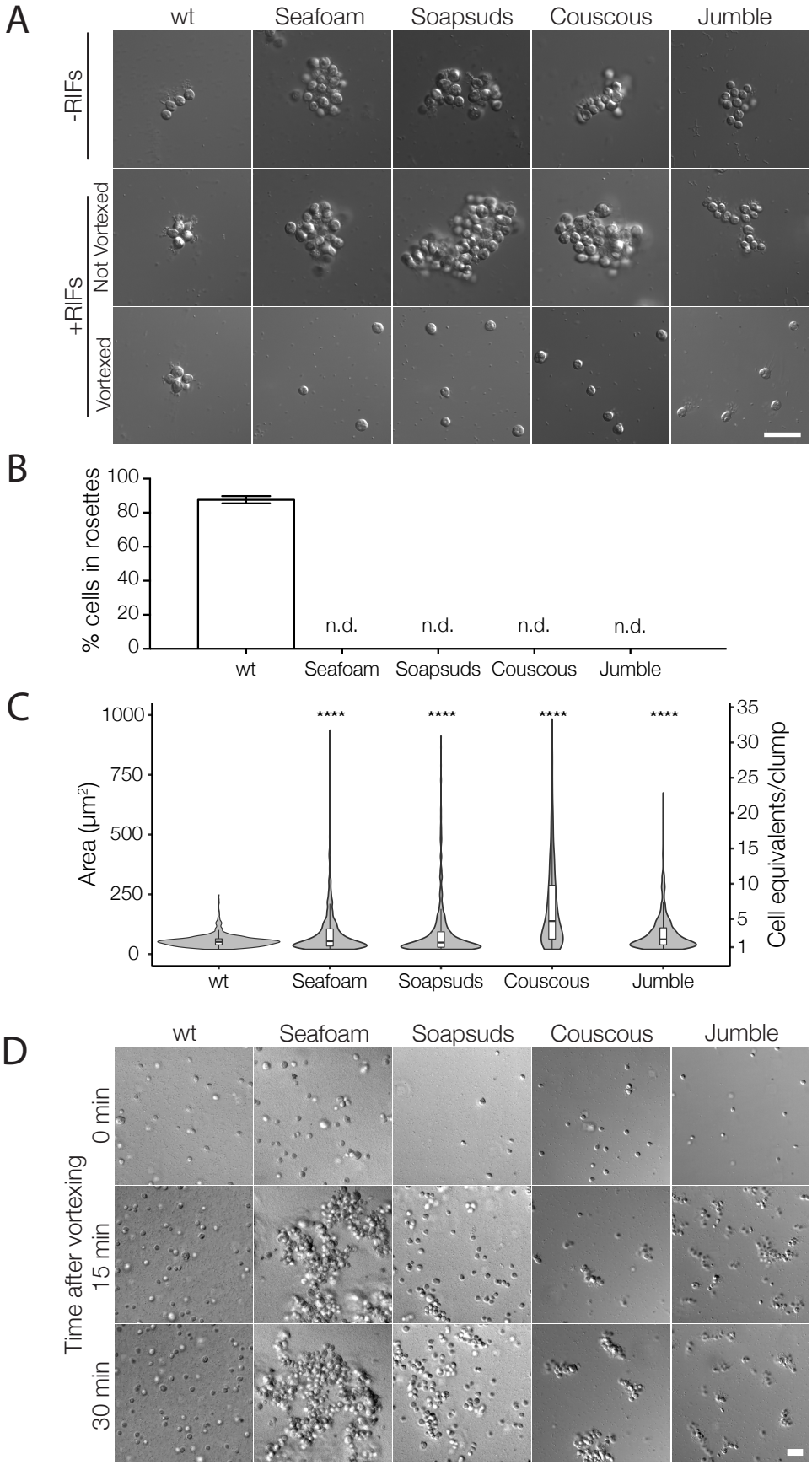
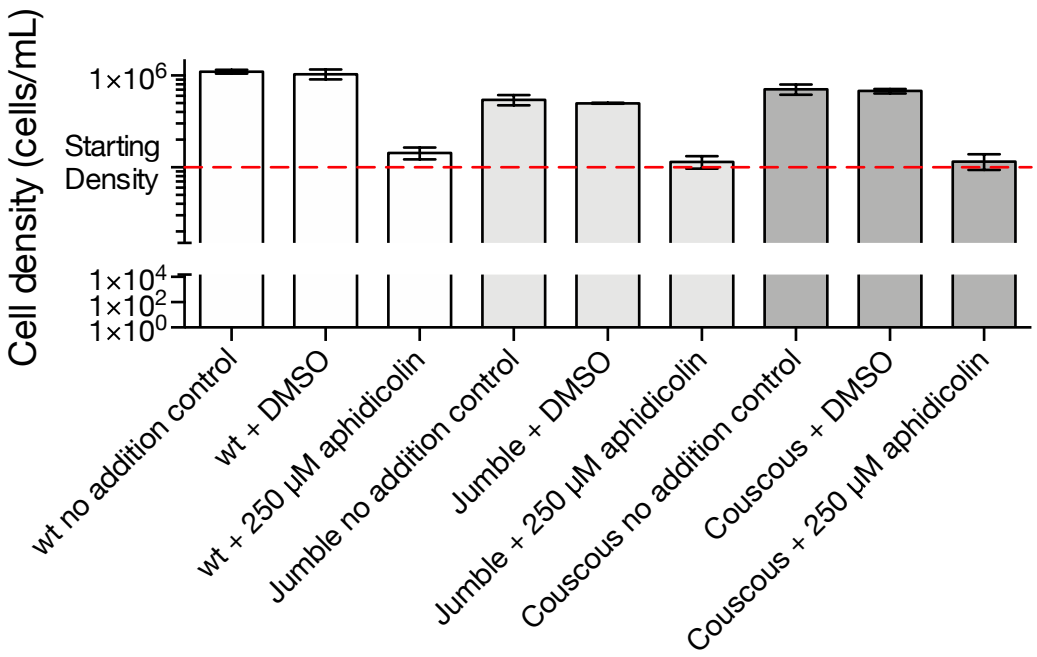

A
wt
Seafoam
Soapsuds
Couscous
Jumble
-RIFs
+RIFs
Not Vortexed
Vortexed
B
% cells in rosettes
100
80
60
40
20
0
n.d.
n.d.
n.d.
n.d.
wt
Seafoam
Soapsuds
Couscous
Jumble
C
Area (μm²)
1000
750
500
250
0
****
****
****
****
35
30
25
20
15
10
5
1
Cell equivalents/clump
wt
Seafoam
Soapsuds
Couscous
Jumble
D
wt
Seafoam
Soapsuds
Couscous
Jumble
Time after vortexing
0 min
15 min
30 min

Figure 2.2. Cell division is not required for clump formation in mutants. Wild type, Jumble, and Couscous cells were vortexed and diluted to $1x10^5$ cells/mL.Either no addition, DMSO, or 250 µM aphidicolin were added to wild type, Jumble, and Couscous. After 24 hours, cells were counted and imaged. **(A)** Aphidicolin successfully blocked cell division, while the no addition control and DMSO control grew for all conditions. Mean density plotted ± S.D for 3 technical replicates on the same day as imaged. **(B)** In wild type cells, no chains were observed in the aphidicolin treated cells, but they were able to successfully grow chains. For Jumble and Couscous, clumps formed in all conditions. Clumps formed in the presence of aphidicolin appear smaller, perhaps due to the lower cell density of the cultures or the lack of cell division, both of which may contribute to clump size. Scale bar = 20 µm.

A


Cell density (cells/mL)
1×10^6
Starting Density
1×10^4
1×10^2
1×10^0
wt no addition control
wt + DMSO
wt + 250 µM aphidicolin
Jumble no addition control
Jumble + DMSO
Jumble + 250 µM aphidicolin
Couscous no addition control
Couscous + DMSO
Couscous + 250 µM aphidicolin

B

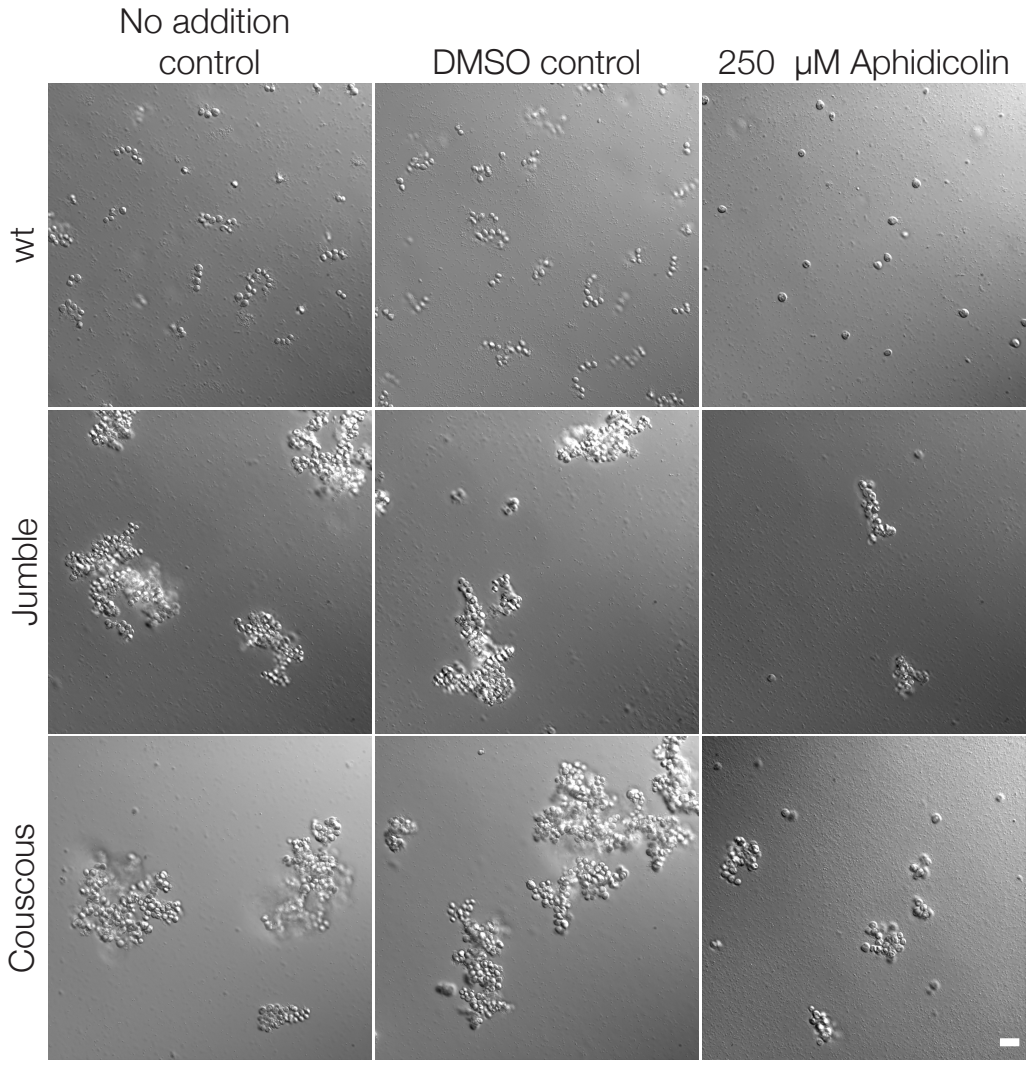

No addition control
DMSO control
250 µM Aphidicolin
wt
Jumble
Couscous

Figure 2.3. Class C mutant growth curve. Mutant and wild type cells were plated at a density of $1x10^4$ cells/mL and counted every 12 hours to assess growth. Mean density plotted $\pm$ SD (n=2 biological replicates with 3 technical replicates).

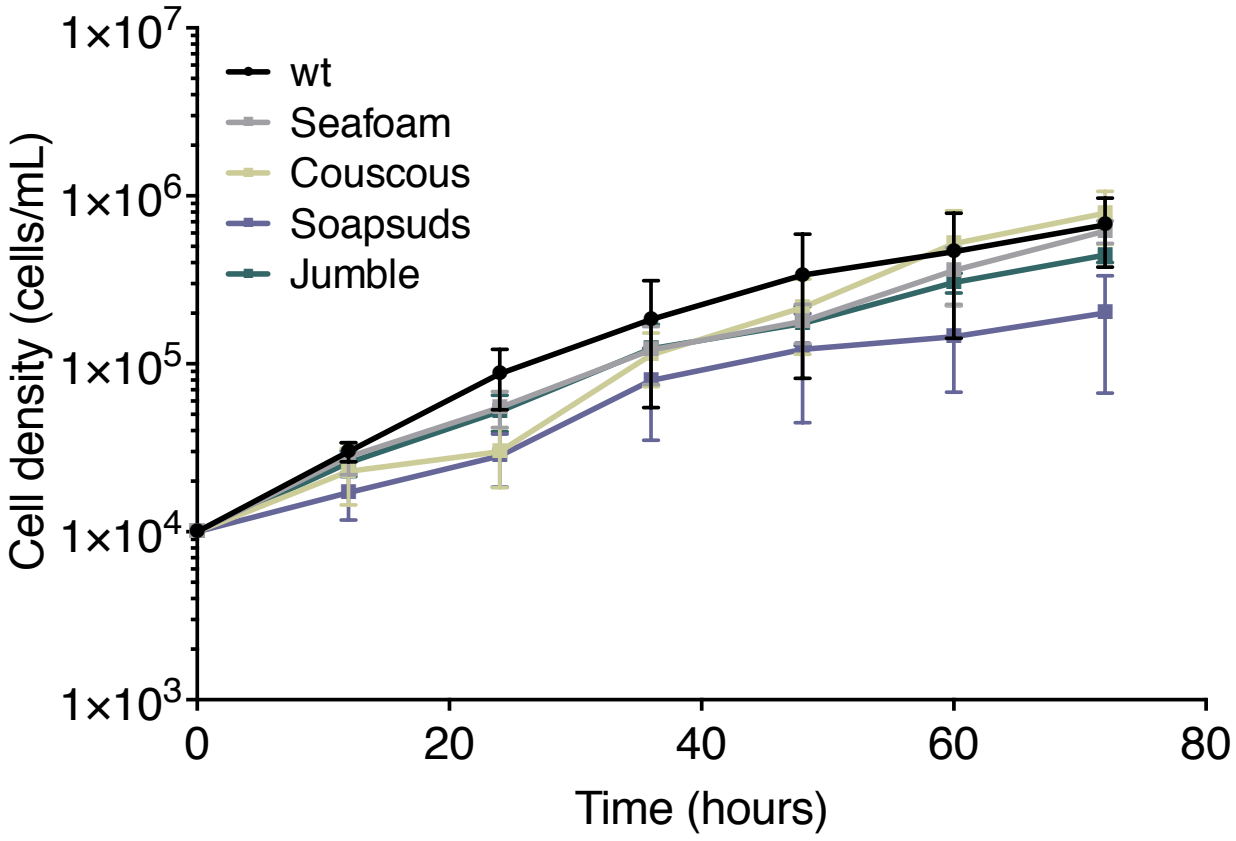

1×10^7
1×10^6
1×10^5
1×10^4
1×10^3
Cell density (cells/mL)
wt
Seafoam
Couscous
Soapsuds
Jumble
0
20
40
60
80
Time (hours)

Figure 2.4. Jumble and Couscous clumps formed in the absence or presence of RIFs are comparable in size. Jumble and Couscous were cultured for 24 hours, either without RIFs or with RIFs. To perform the clumping assay, cells cultured either with or without RIFs were vortexed and then incubated for 30 minutes. Wild type cells without RIFs were included as a negative control. Clump sizes were quantified by automated image analysis. Data are presented as violin boxplots, showing the median cell number (horizontal line), interquartile range (white box), and range excluding outliers (vertical line). There were no significant differences in clump size in mutants treated with RIFs or without RIFs (K-S test, n.s.=not significant, $p>0.05$).

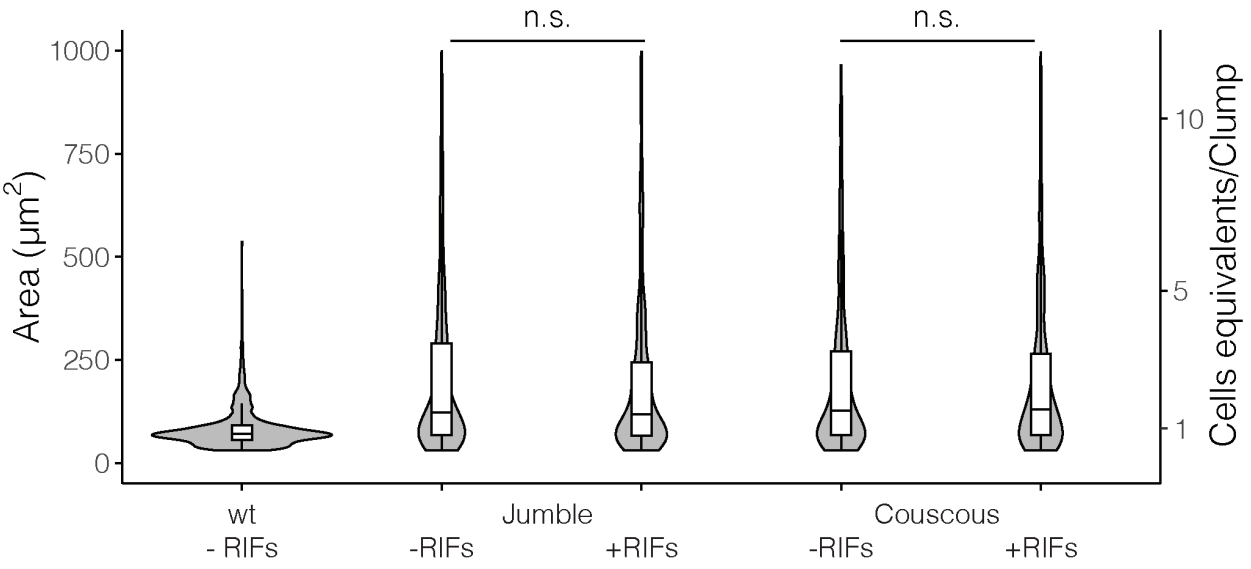

n.s.
n.s.
1000
750
500
250
0
Area (µm2)
10
5
1
Cells equivalents/Clump
wt
- RIFs
Jumble
-RIFs
+RIFs
Couscous
-RIFs
+RIFs

Figure 2.5. Jumble and Couscous cells adhere to wild type cells. Fluorescent mWasabi expressing wild type cells uninduced or induced to form rosettes by the addition of RIFs were mixed with either Jumble or Couscous cells and imaged after 30 minutes. Mutant cells adhered non-specifically to each other and wild type cells. Scale bar = 20 μm.

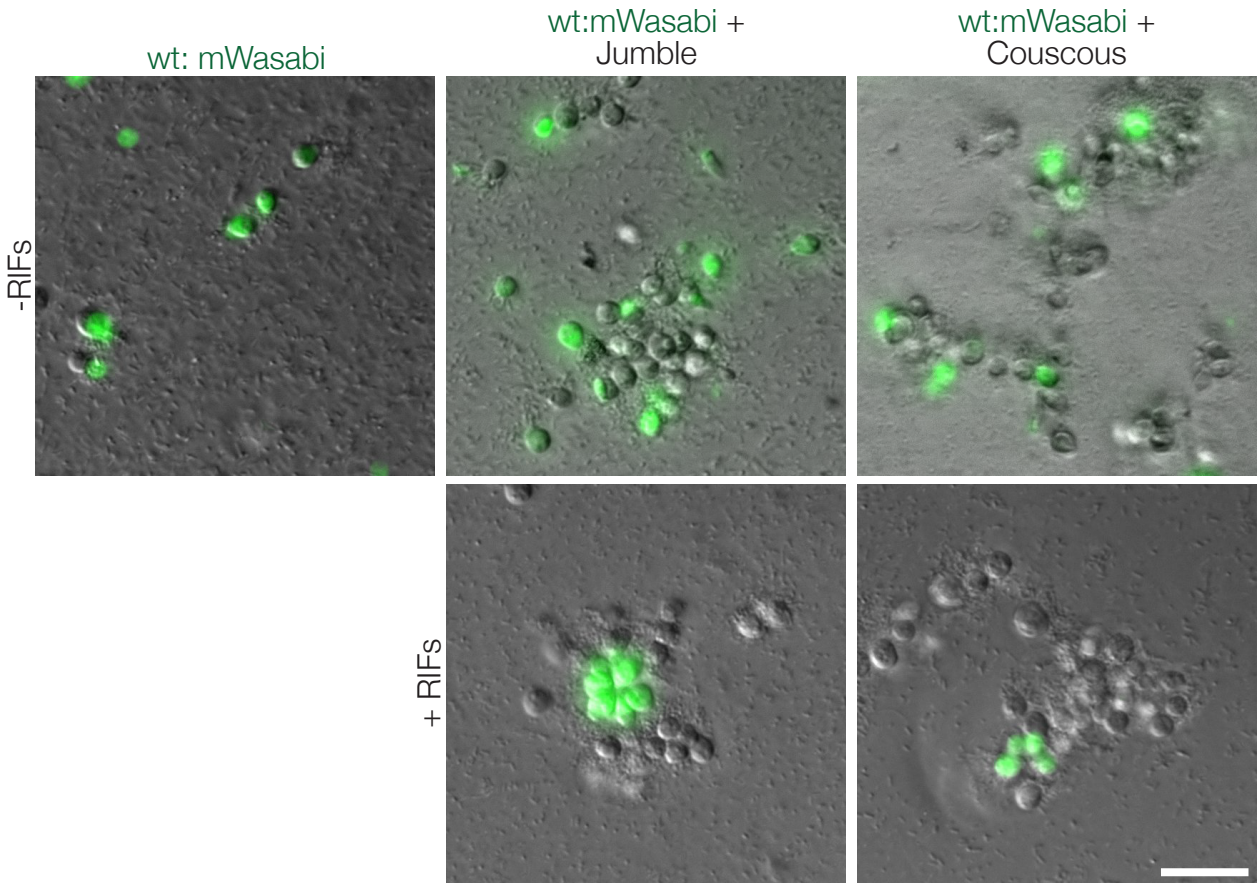

wt: mWasabi
wt:mWasabi +
Jumble
wt:mWasabi +
Couscous
-RIFs
+ RIFs

Figure 2.6. Mapping cross scheme. Flow chart of the steps used in mapping cross and bulk segregant analysis.

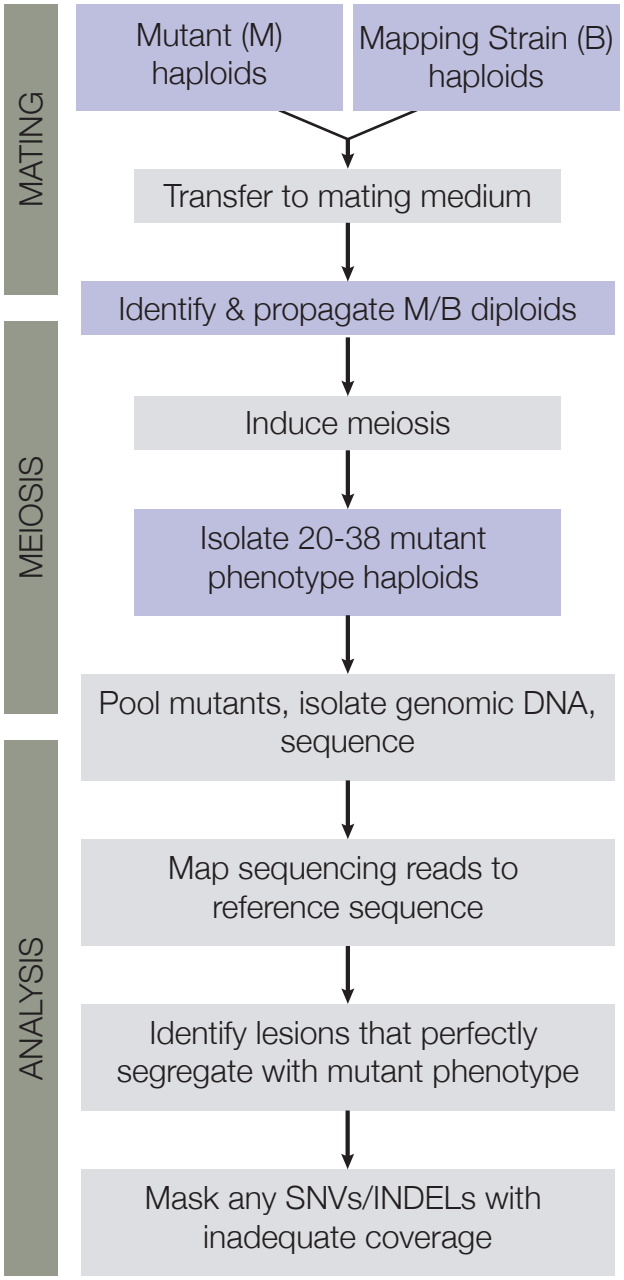
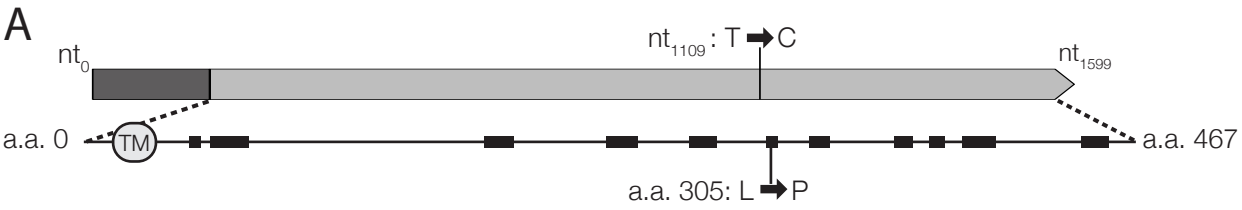

MATING
MEIOSIS
ANALYSIS
Mutant (M) haploids
Mapping Strain (B) haploids
Transfer to mating medium
Identify & propagate M/B diploids
Induce meiosis
Isolate 20-38 mutant phenotype haploids
Pool mutants, isolate genomic DNA, sequence
Map sequencing reads to reference sequence
Identify lesions that perfectly segregate with mutant phenotype
Mask any SNVs/INDELs with inadequate coverage

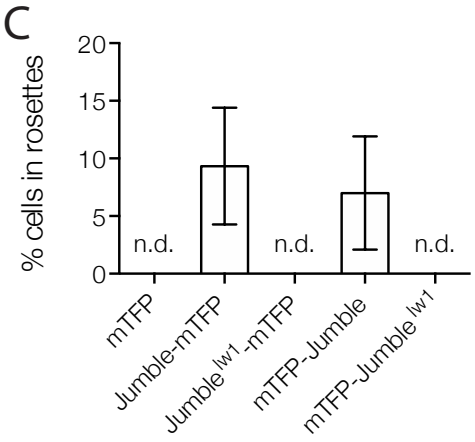
Figure 2.7. Jumble maps to a predicted glycosyltransferase that localizes to the Golgi apparatus. (A) Jumble has a predicted transmembrane domain (marked TM) and secondary structure (alpha helices marked by black rectangles). Structural homology algorithms predict that Jumble is structurally related to well-characterized glycosyltransferases (Figure 2.8B). The mutant gene has a T to C mutation at nucleotide 1109 that causes an amino acid substitution of proline to leucine at amino acid position 305. **(B)** A backcross of a mutant F1 progeny to the Mapping Strain yielded nine rosette-forming F2 isolates with the wild type T allele and twelve clumpy F2 isolates with the *jumble*^{lw1} C allele. The inheritance significantly deviated from expected Mendelian inheritance of unlinked traits and confirmed the tight linkage between the *jumble*^{lw1} allele to the clumpy, rosetteless phenotype. X^2 = Chi-squared value, d.f. = degrees of freedom. **(C,D)** Transgenic expression of *jumble-mTFP* and *mTFP-jumble* rescued rosette development in the Jumble mutant, but *jumble*^{lw1}*-mTFP*, *mTFP-jumble*^{lw1}, or *mTFP* did not. RIFs were added immediately after transfection and 40 µg/mL puromycin was added 24 hours post-transfection to select for transformants. **(C)** Rosette development was measured as the % of cells in rosettes 72 hr post-transfection and shown as mean ± SD. n.d. = no detected rosettes. (n=200 cells counted from each of 3 technical replicates; 2 biological replicates). **(D)** Rosettes transgenically complemented with *jumble-mTFP* in the Jumble mutant appeared phenotypically wild type and most cells in rosettes had detectable fluorescent expression at the apical base of the cell. Representative rosette shown. **(E-H)** To examine localization, Jumble-mWasabi or Jumble^{lw1}-mWasabi (cyan) under the *efl* promoter were co-expressed with membrane marker-mCherry (magenta) in wild type *S. rosetta.* Jumble-mWasabi localizes to the apical pole of cells grown **(E)** without RIFs or **(G)** with RIFs, consistent with the localization of the Golgi apparatus. When expressed in otherwise wild type cells grown **(F)** without RIFs or **(H)** with RIFs, the mutant Jumble^{lw1}-mWasabi incorrectly localizes to the ER and food vacuole. Boxes indicate the inferred location of the Golgi apparatus at the apical pole of the cell. The food vacuole (asterisk) was often visualized due to autofluoresence from ingested bacteria or through accumulation of the fluorescent markers in the food vacuole, perhaps through autophagy. For reference, arrows indicate the base of the flagellum although the flagellum may not be visible in the plane of focus shown. Scale bar = 5 µm.

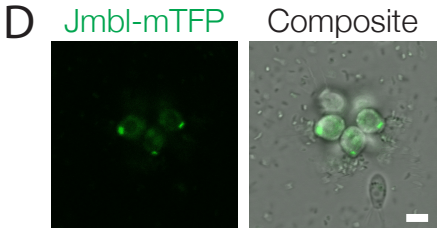


B

Jumble		
	Rosettes	Clumps
T	9	0
C	0	12

X^2=21.9 , df=3, $p < 0.001$





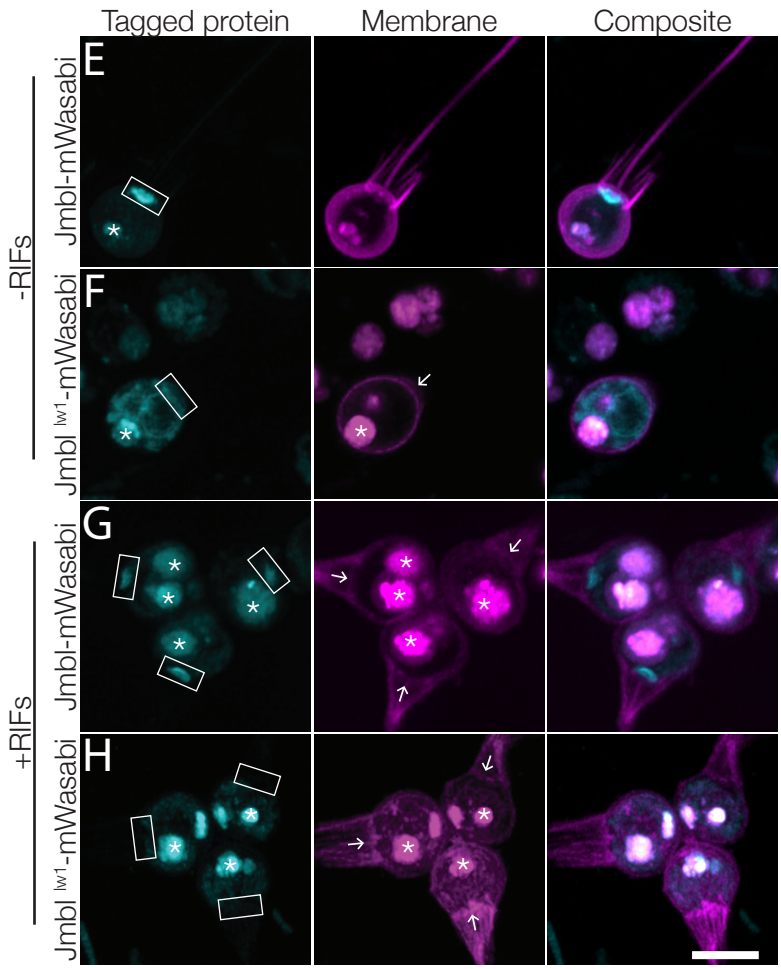
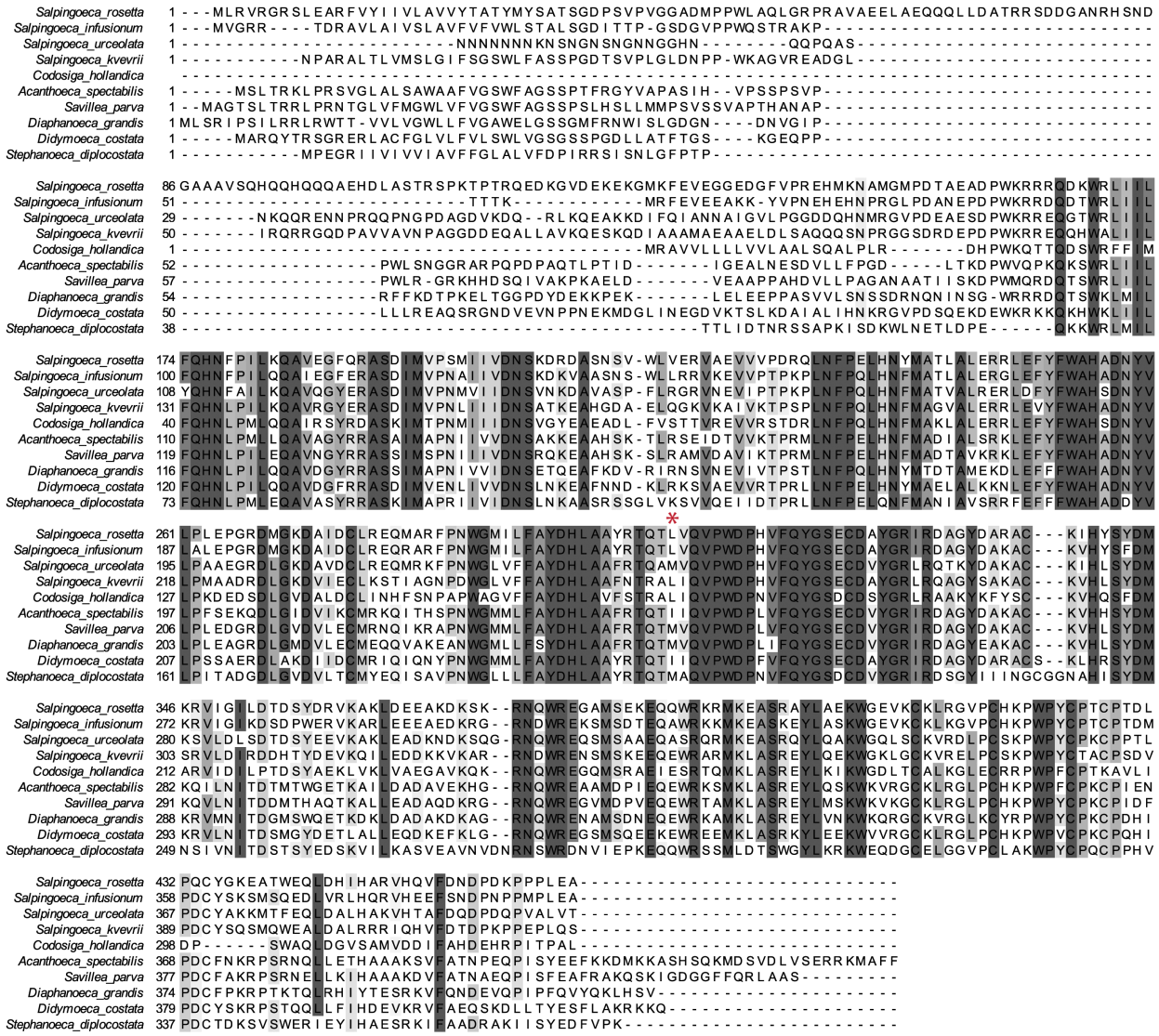


Figure 2.8. Alignment of Jumble homologs and predicted structure. (A) *S. rosetta* Jumble amino acid sequence was aligned to the predicted sequences encoded by homologs from nine other choanoflagellate species, first identified by best reciprocal BLAST using the transcriptomes reported in Richter et al., 2018. Red asterisk indicates the location of the causative mutation in the *S. rosetta jumble* gene. **(B)** The structure of Jumble protein predicted by HHphred (teal) was aligned to the catalytic domain of human polypeptide N-acetylgalactosaminyltransferase 4 (GalNAc-T4; purple). The mutated leucine at 305 is found in a predicted alpha helix.

A


Salpingoeca_rosetta
Salpingoeca_infusionum
Salpingoeca_urceolata
Salpingoeca_kvevrii
Codosiga_hollandica
Acanthoeca_spectabilis
Savillea_parva
Diaphanoeca_grandis
Didymoeca_costata
Stephanoeca_diplocostata

B

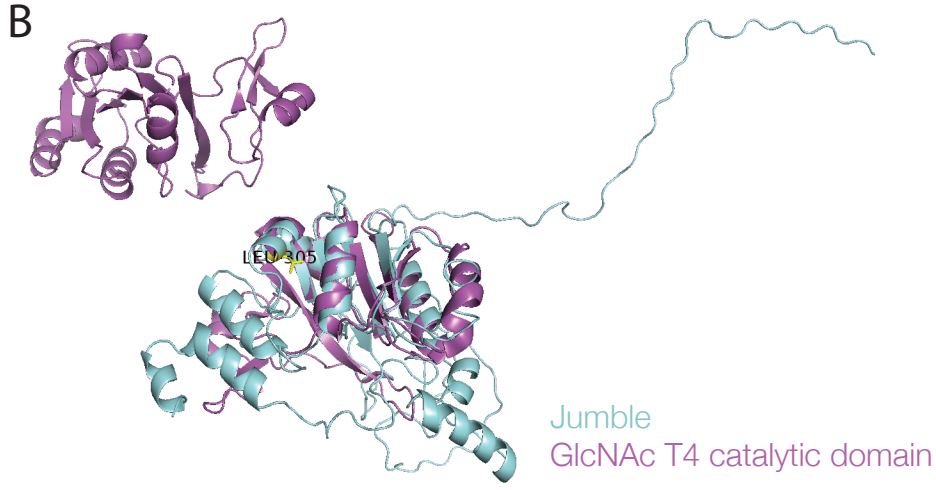

LEU305
Jumble
GlcNAc T4 catalytic domain

Figure 2.9. Alignment of Jumble to fungal homologs. *S. rosetta* Jumble protein sequence was aligned to predicted/unannotated protein sequences from four fungal species identified by best reciprocal BLAST: *Saitoella complicata* (NCBI accession XP_019021578.1), *Dactylellina haptotyla* (NCBI accession EPS43829.1), *Naematelia encephala* (NCBI accession ORY22834.1), and *Tuber magnatum* (NCBI accession PWW71609.1).

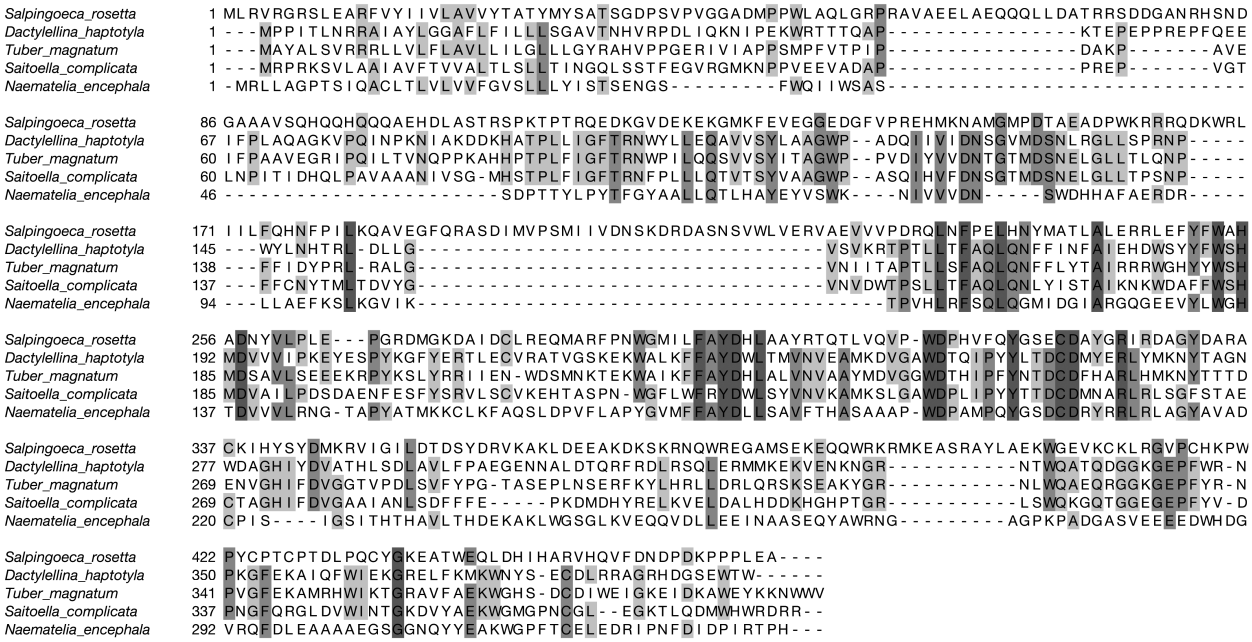
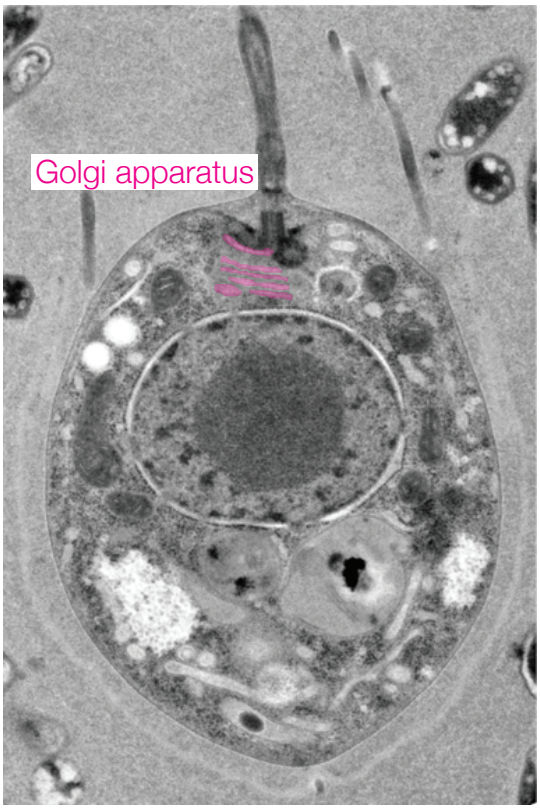

Salpingoeca_rosetta 1 MLRVRGRSLEARFVYIIVLAVVYTATYMYSATSGDPSVPVGGADMPPWLAQLGRPRAVAEELAEQQQLLDATRRSDDGANRHSND
Dactylellina_haptotyla 1 ---MPPITLNRRAIAYLGGAFLFILLLSGAVTNHVRPDLIQKNIPEKWRTTTQAP----------KTEPEPPREPFQEE
Tuber_magnatum 1 ---MAYALSVRRRLLVLFLAVLLILGLLLGYRAHVPPGERIVIAPPSMPFVTPIP----------DAKP-------AVE
Saitoella_complicata 1 ---MRPRKSVLAAIAVFTVVALTLSLLTINGQLSSTFEGVRGMKNPPVEEVADAP----------PREP-------VGT
Naematelia_encephala 1 -MRLLAGPTSIQACLTLVLVVFGVSLLLYISTSENGS---------FWQIIWSAS----------
Salpingoeca_rosetta 86 GAAAVSQHQQHQQQAEHDLASTRSPKTPTRQEDKGVDEKEKGMKFEVEGGEDGFVPREHMKNAMGMPDTAEADPWKRRRQDKWRL
Dactylellina_haptotyla 67 IFPLAQAGKVPQINPKNIAKDDKHATPLLIGFTRNWYLLEQAVVSYLAAGWP--ADQTIVIDNSGVMDSNLRGLLSPRNP-----
Tuber_magnatum 60 IFPAAVEGRIPQILTVNQPPKAHHPTPLFIGFTRNWPILQQSVVSYITAGWP--PVDIYVVDNTGTMDSNELGLLTLQNP-----
Saitoella_complicata 60 LNPITIDHQLPAVAAANIVSG-MHSTPLFIGFTRNFPLLLQTVTSYVAAGWP--ASQIHVFDNSGTMDSNELGLLTPSNP-----
Naematelia_encephala 46 ----------SDPTTYLPYTFGYAALLQTLHAYEYVSWK----NIVVVDN------SWDHHAFAERDR-----
Salpingoeca_rosetta 171 IILFQHNFPILKQAVEGFQRASDIMVPSMIIVDNSKDRDASNSVWLVERVAEVVVPDRQLNFPELHNYMATLALERRLEFYFWAH
Dactylellina_haptotyla 145 ---WYLNHTRL-DLLG----------VSVKRTPTLLTFAQLQNFFINFAIEHDWSYYFWSH
Tuber_magnatum 138 ---FFIDYPRL-RALG----------VNIITAPTLLSFAQLQNFFLYTAIRRRWGHYYWSH
Saitoella_complicata 137 ---FFCNYTML TDVYG----------VNVDWTPSLLTFAQLQNLYISTAIKNKWDAFFWSH
Naematelia_encephala 94 ---LLAEFKSLKGVIK----------TPVHLRFSQLQGMIDGIARGQGEEVYLWGH
Salpingoeca_rosetta 256 ADNYVLPLE---PGRDMGKDAIDCLREQMARFPNWGMILFAYDHLAAYRTQTLVQVP-WDPHVFQYGSECDAYGRIRDAGYDARA
Dactylellina_haptotyla 192 MDVVVIPKEYESPYKGFYERTLECVRATVGSKEKWALKFFAYDWLTMVNVEAMKDVGAWDTQIPYYLTDCDMYERLYMKNYTAGN
Tuber_magnatum 185 MDSAVLSEEEKRPYKSLYRRIIEN-WDSMNKTEKWAIKFFAYDHLALVNVAAYMDVGGWDTHIPFYNTDCDFHARLHMKNYTTTD
Saitoella_complicata 185 MDVAILPDSDAENFESFYSRVLSCVKEHTASPN-WGFLWFRYDWLSYVNVKAMKSLGAWDPLIPYYTTDCDMNARLRLSGFSTAE
Naematelia_encephala 137 TDVVVLRNG-TAPYATMKKCLKFAQSLDPVFLAPYGVMFFAYDLLSAVFTHASAAAP-WDPAMPQYGSDCDRYRRLRLAGYAVAD
Salpingoeca_rosetta 337 CKIHYSYDMKRVIGILDTDSYDRVKAKLDEEAKDKSKRNQWREGAMSEKEQQWRKRMKEASRAYLAEKWGEVKCKLRGVPCHKPW
Dactylellina_haptotyla 277 WDAGHIYDVATHLSDLAVLFPAEGENNALDTQRFRDLRSQLERMMKEKVENKNGR----------NTWQATQDGGKGEPFWR-N
Tuber_magnatum 269 ENVGHIFDVGGTVPDLSVFYPG-TASEPLNSERFKYLHRLLDRLQRSKSEAKYGR----------NLWQAEQRGGKGEPFYR-N
Saitoella_complicata 269 CTAGHIFDVGAAIANLSDFFFE-----PKDMDHYRELKVELDALHDDKHGHPTGR----------LSWQKGQTGGEGEPFYV-D
Naematelia_encephala 220 CPIS----IGSITHTHAVLTHDEKAKLWGSGLKVEQQVDLLEEINAASEQYAWRNG---------AGPKPADGASVEEEEDWHDG
Salpingoeca_rosetta 422 PYCPTCPTDLPQCYGKEATWEQLDHIHARVHQVFDNDPDKPPPLEA----
Dactylellina_haptotyla 350 PKGFEKAIQFWIEKGRELFKMKWNYS-ECDLRRAGRHDGSEWTW------
Tuber_magnatum 341 PVGFEKAMRHWIKTGRAVFAEKWGHS-DCDIWEIGKEIDKAWEYKKNWWV
Saitoella_complicata 337 PNGFQRGLDVWINTGKDVYAEKWGMGPNCGL--EGKTLQDMWHWRDRR--
Naematelia_encephala 292 VRQFDLEAAAAEGSGGNQYYEAKWGPFTCELEDRIPNFDIDPIRTPH---

Figure 2.10. Ultrastructure of *S. rosetta* and ER co-localization of Jumble^{lw1}. (A) A transmission electron micrograph shows the ultrastructure of *S. rosetta*. The Golgi apparatus has been pseudo-colored pink and labelled. Image provided courtesy of Kent McDonald and adapted from Booth et al., 2018. **(B)** Jumble^{lw1}-mWasabi fusion protein shows partial co-localization with the mCherry-ER marker when expressed in wild type *S. rosetta.* Dashed line marks the inferred location of the nucleus. Scale bar = 1 μm.

A



B

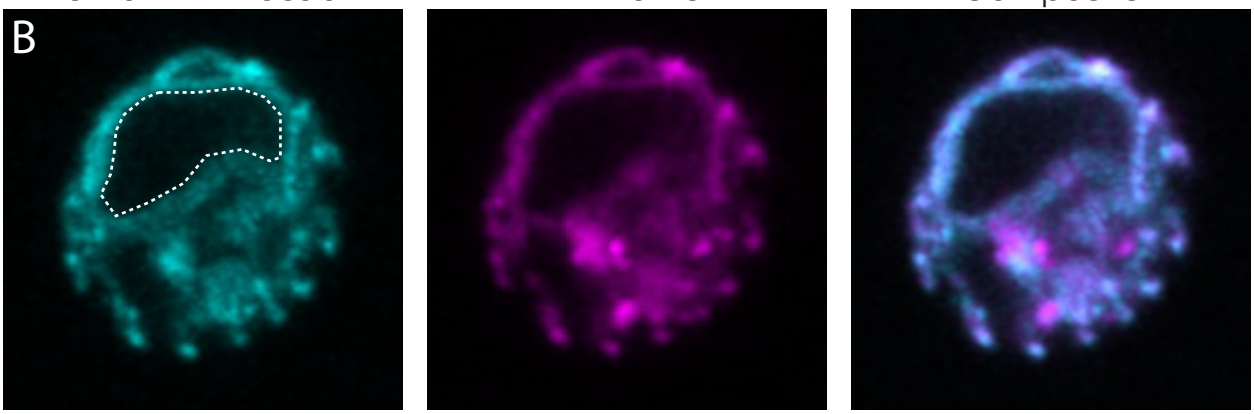
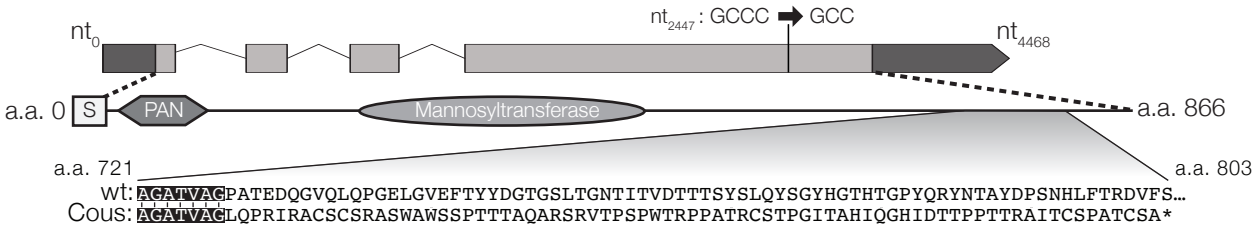


Figure 2.11. Couscous maps to a predicted mannosyltransferase with a PAN/Apple domain. (A) Couscous has a predicted signal sequence (S), a PAN/Apple domain (PAN), and a mannosyltransferase domain. The causative lesion is a 1-base pair deletion at nucleotide position 2447 that causes a frameshift at amino acid 728, resulting in 75 amino acids that do not align between the wild type and mutant (Cous) sequences, and an early stop codon (*) at amino acid 803. **(B)** Independent backcrosses of two individual mutant F1 progeny to the Mapping Strain yielded 38 rosette-forming F2 isolates with the wild type GCCC allele and 51 clumpy F2 isolates with the *couscous*^{lw1} GCC allele. The inheritance significantly deviated from expected Mendelian inheritance of unlinked traits and confirmed the tight linkage between the *couscous*^{lw1} allele to the clumpy, rosetteless phenotype. X^2 = Chi-squared value, d.f. = degrees of freedom. **(C, D)** Rosette formation in Couscous mutant cells can be rescued by transgenic expression of *couscous-mTFP* or *mTFP-couscous*, but not *couscous*^{lw1}*-mTFP*, *mTFP-couscous*^{lw1}, or *mTFP* alone. RIFs were added immediately after transfection and 40 µg/mL puromycin was added 24 hours post-transfection to select for positive transformants. **(C)** Rosette development (mean ± SD) was measured as the % of cells in rosettes 72 hr after transfection and treatment with RIFs. n.d. = no detected rosettes. (n=200 cells counted from each of 3 technical replicates; 2 biological replicates). **(D)** Rosettes transgenically complemented with *couscous-mTFP* in the Couscous mutant appeared phenotypically wild type. Representative rosette shown. Scale bar = 5 µm.

A

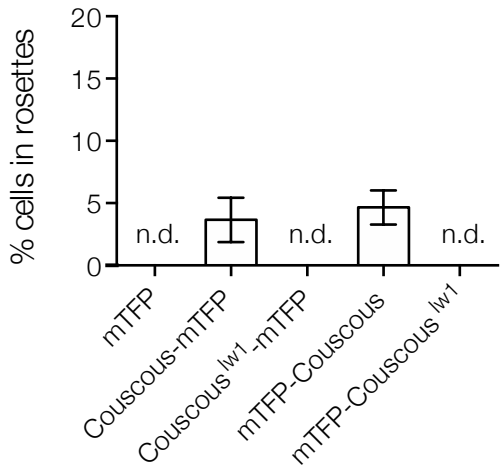


B

Couscous		
	Rosettes	Clumps
GCCC	38	0
GCC	0	51

X^2=92.8 , df=3, $p < 0.001$

C



D

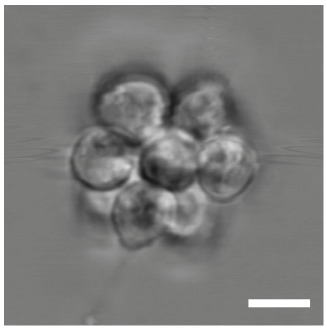


Figure 2.12. Couscous homology and localization. (A) The predicted mannosyltransferase domain from *S. rosetta* was aligned to the alpha-mannosyltransferase domain of MNN2 genes from *S. cerevisiae* (NCBI accession NP_009571.1) and *C. albicans* (NCBI accession XP_710276.1). Red asterisks highlight the conserved DXD motif of many glycosyltransferases. **(B)** The transgenes mCherry-membrane marker and Couscous-mWasabi fusion protein were expressed in wild type *S. rosetta.* The Couscous fusion localized to puncta distributed throughout the cell (to the exclusion of an unidentified organelle; circle) and faintly at the cell collar. Scale bar = 5 µm.

A

```
S_rosetta_Cous      218 KQYLENVP--TYPEGKYRGRGIIIVAGGRYLASALVSVKMLRDLGCTLRIQIWHLGPEEVDAAQR-EVLAAYDVEPRDFTAVV
S_cerevisiae_MNN2   152 KDYVEHIATLVSPKGTYKGSGIATVGGGKFSLMAFLIIKTLRNMGTTLPVEVLIPPGDEGETEFCNKILPKYNSKCIYVSDIL
C_albicans_MNN2     229 -KYVVENLPEDAPDGLYKKNGIVYVAGGSFNWLTLLSIKSLRAVGCHLPIEVFIPKIEEYESDLCNRILPELDARCIYMKNQL
                                                                  * *
S_rosetta_Cous      298 PPELLKP---IESNVGMRLFQLKPLALLYSDLQEVLLIDSDNTPLMDPTYLFSEQGFQDTGTVFWPDYWKTSFDNPIWATLGM
S_cerevisiae_MNN2   235 PRET--------IEKFVFKGYQFKSLALIASSFENLLLLDADNFPIKPLDNIFNEEPYVSTGLVMWPDFWRRTTHPLYYDIAGI
C_albicans_MNN2     311 MNPNKDNSDSFANKFEFKGYQYKALAILLSSFENVLLLDSDNIPAHSPEELFENDPFKSYGLVVWPDYWKRATSPYYYNIADI

S_rosetta_Cous      378 EPKAMWE----------------------------------------QESGQLLINKRAAWRGLNLCVHFNS---AFYMGLIN-
S_cerevisiae_MNN2   311 AVDKKKR--VRNSRDDITPPAVYTKDLKDLSDVPLSDLDGTIPDVSTESGQLMINKTKHLATALLSLFYNVNGPTWYYPIFSQ
C_albicans_MNN2     394 DVSEKYLGSKYNEVEGQYTDLSVEKGSVELDKIPLHQRLGSIPDPTSESGQLLISKKTHLKPLLLALYYNLYGPSHYYPLFSQ

S_rosetta_Cous      418 -GDKDTFRFAWLAAGVPFVMNSWMPSAVGTVKERHSDTDLGFCSHTMLQHDL
S_cerevisiae_MNN2   392 KAAGEGDKETFIAAANFYGLSFYQVRTRTGVEGYHDEDG--FHGVAMLQHDF
C_albicans_MNN2     477 GSDGEGDKETFLAATVTLGKRYYQVAKFLVSLGHFKVPGGDFEGCGMGQFD-
```

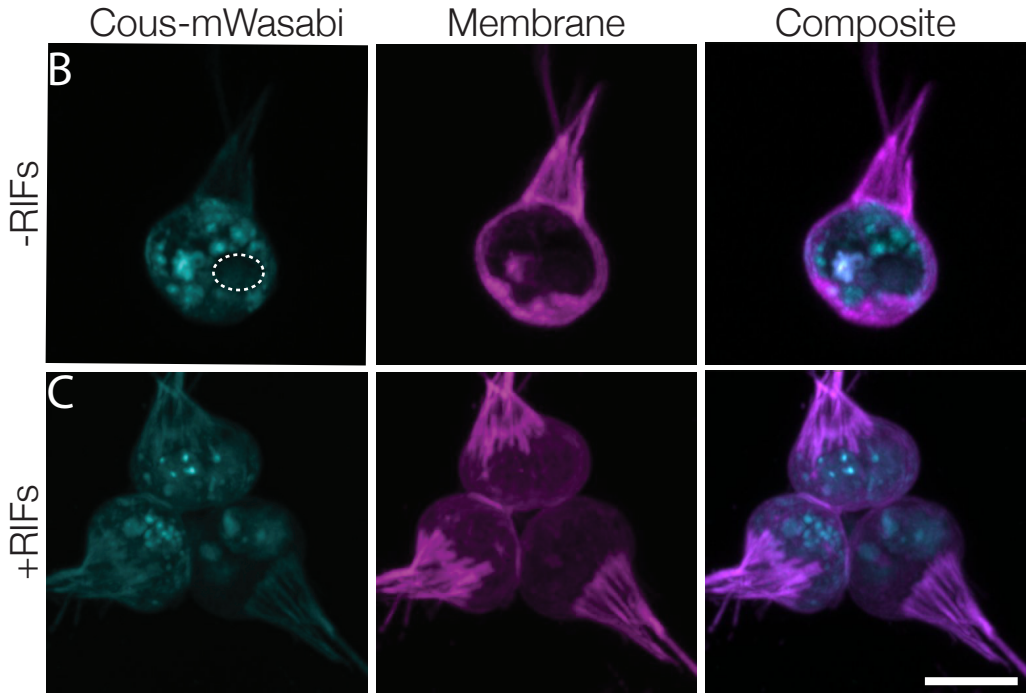


Figure 2.13. Disruption of basal glycosylation patterns in Jumble and Couscous mutants. FITC-labelled jacalin binds the apical and basal poles of wild type single cells **(B)** and becomes enriched in the ECM in the center of rosettes **(A, B'** boxed region from **A)**. Although FITC-jacalin staining appeared normal at the apical poles of Jumble (**C)** and Couscous **(D)** mutant cells, FITC-jacalin staining at the basal poles of cells was undetectable in cells grown either in the absence (-RIFs; **C, D**) or presence (+RIFs; **C', D'**) RIFs. Arrows mark the apical pole and arrowheads mark the basal pole. **(E)** Cartoon depicts how jacalin fluorescence was measured. Starting with micrographs of FITC-jacalin stained cells, a line was drawn tracing from one edge of the collar around the cell body to the other edge of the collar, and the underlying fluorescent signal was normalized for cell size and background intensity. **(F)** The average normalized fluorescence intensity of jacalin measured in at least 59 cells for each condition was graphed against the normalized length of the cell body (n=2 biological replicates). Jumble and Couscous -/+RIFs have reduced jacalin binding at the basal pole compared to wild type -/+RIFs. Gray shadows indicate 95% confidence intervals. Scale bar = 5 μm.

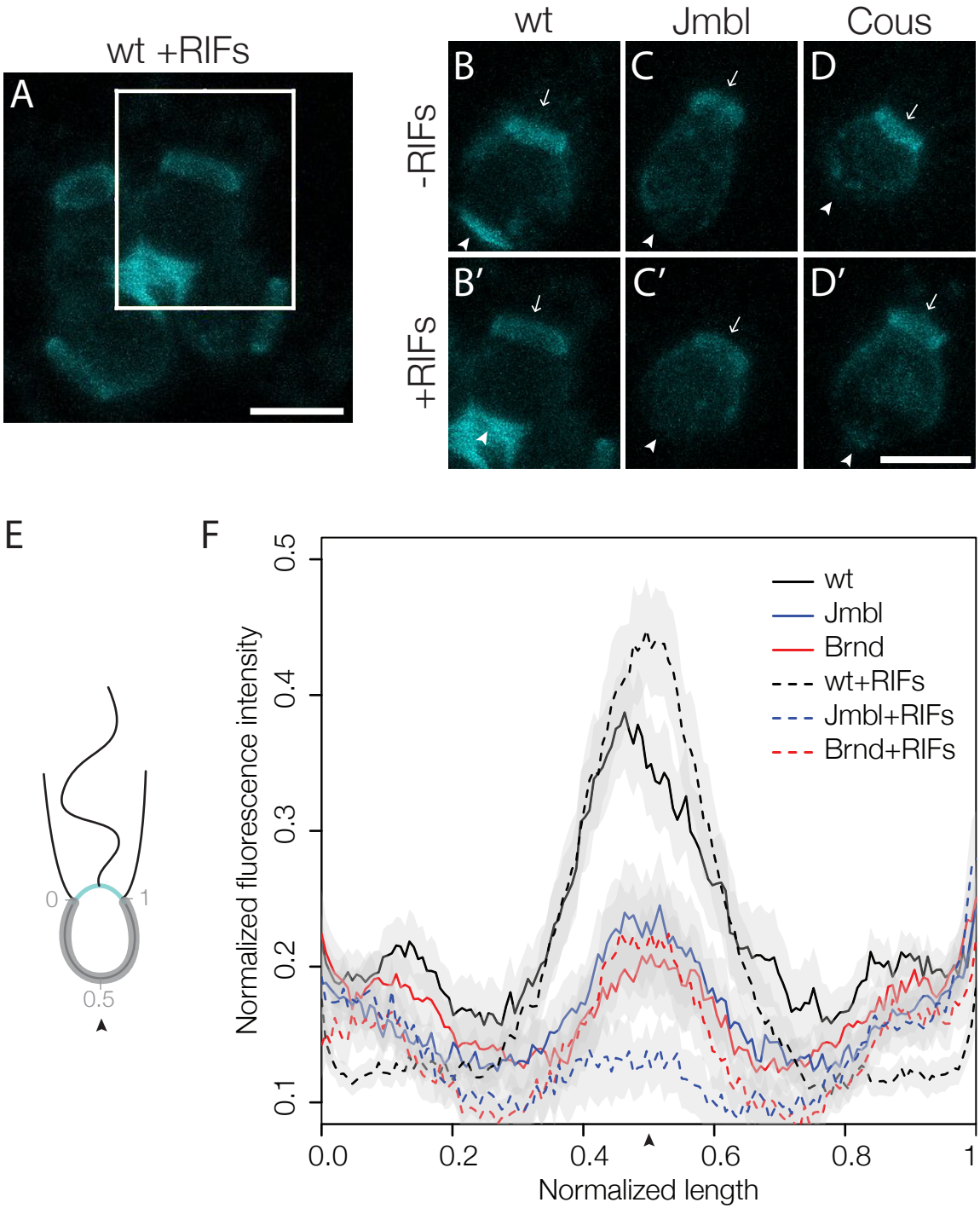

wt +RIFs
A
wt
Jmbl
Cous
B
C
D
-RIFs
B′
C′
D′
+RIFs
E
0
1
0.5
F
Normalized fluorescence intensity
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Normalized length
wt
Jmbl
Brnd
wt+RIFs
Jmbl+RIFs
Brnd+RIFs

Figure 2.14. Jacalin Western blot in cell lysates. Whole cell lysates from *E. pacifica* (co-cultured prey bacteria), wild type *S. rosetta*, Couscous, and Jumble were probed with jacalin. No clear differences in banding pattern were observed among the *S. rosetta* strains, except for a small band ~25 kD in the Couscous lysate that is likely from *E. pacifica* contamination. Tubulin was probed as a loading control.

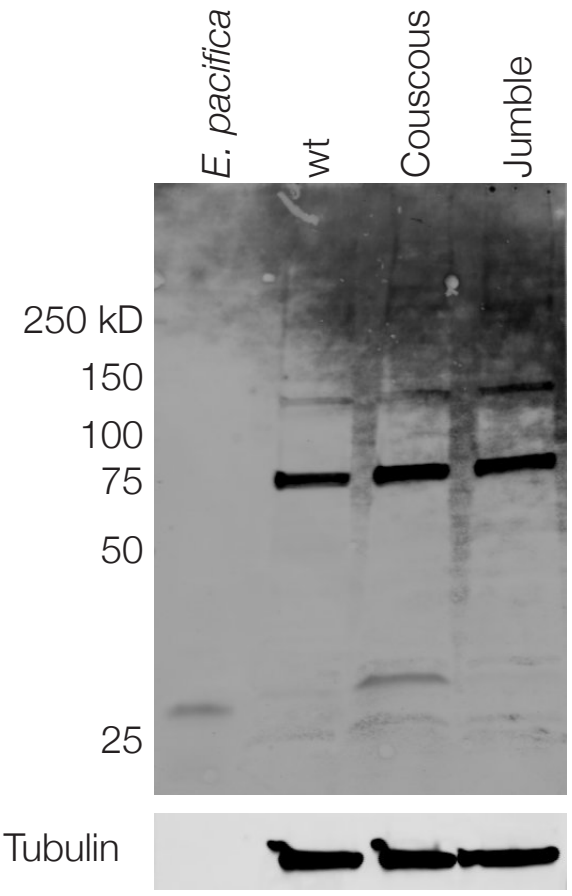

E. pacifica
wt
Couscous
Jumble
250 kD
150
100
75
50
25
Tubulin

Figure 2.15. Transgenic rescue restores jacalin staining at the center of complemented rosettes. (A-F) Biotinylated-jacalin labelled with streptavidin Alexa Fluor 647 has the same localization pattern in the absence **(A, C, E)** and in the presence **(B, D, F)** of RIFs as that observed with FITC-labelled jacalin (Figure 2.13A-D). In wild type cells, jacalin binds the apical and basal poles of single cells **(A)** and becomes enriched at the center of wild type rosettes **(B)**. In the mutants Jumble **(C, D)** and Couscous **(E, F)**, jacalin staining was severely reduced at the basal poles both in the absence **(C, E)** and in the presence **(D, F)** of RIFs, while the apical pole staining appeared similar to wild type single cells. **(G, I)** Transfection of Jumble **(G)** and Couscous **(I)** with *mTFP* alone did not restore jacalin localization to the basal pole. Shown here in the presence of RIFs. **(H, J)** However, Jumble **(H)** and Couscous **(J)** complemented with *mTFP-jumble* or *couscous-mTFP*, respectively, form rosettes with jacalin localized in the center as observed in wild type rosettes. Arrows mark the apical pole and arrowheads mark the basal pole. Scale bar = 5 µm.

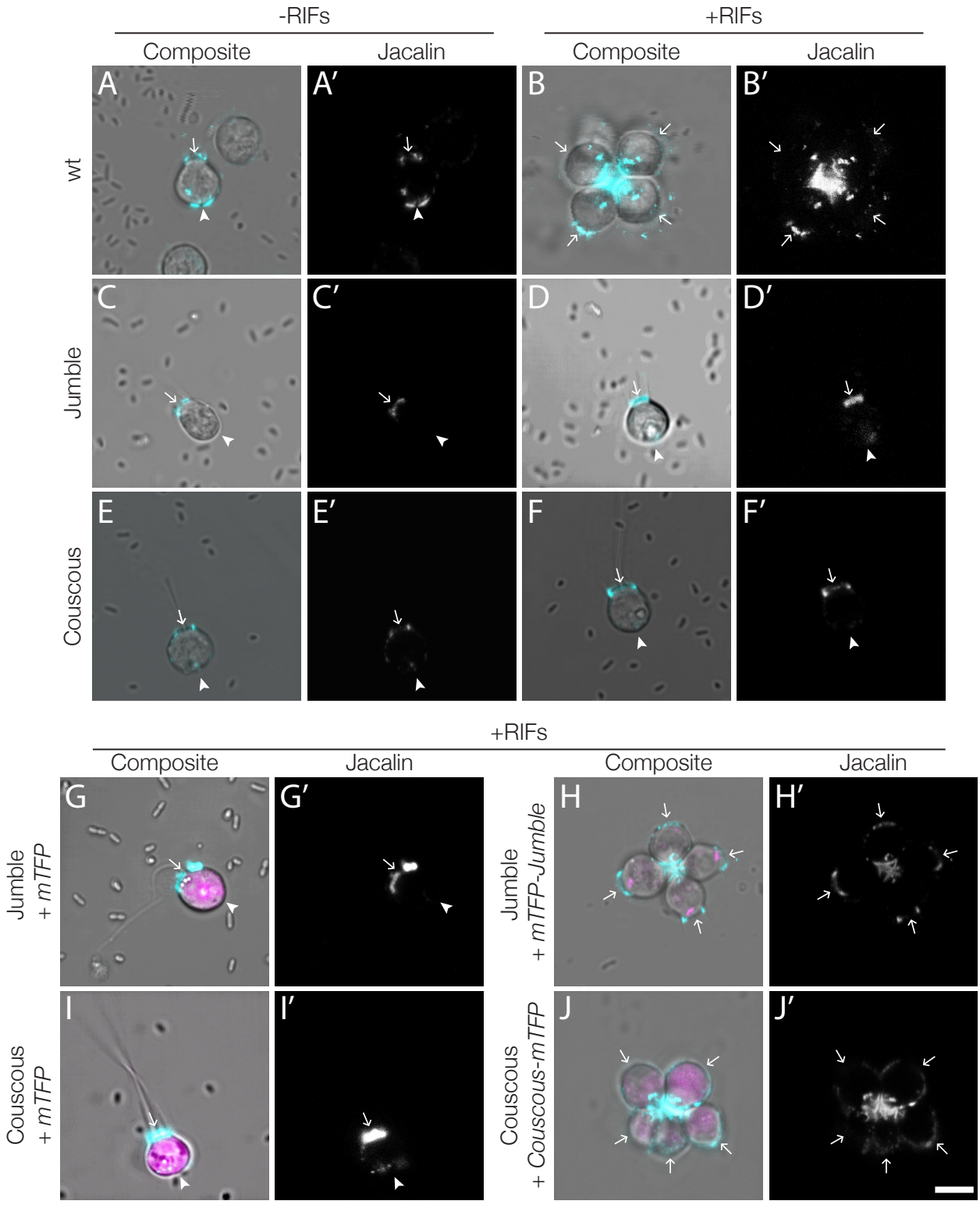

-RIFs
+RIFs
Composite
Jacalin
Composite
Jacalin
wt
Jumble
Couscous
A
A′
B
B′
C
C′
D
D′
E
E′
F
F′
+RIFs
Composite
Jacalin
Composite
Jacalin
Jumble + mTFP
Jumble + mTFP-Jumble
Couscous + mTFP
Couscous + Couscous-mTFP
G
G′
H
H′
I
I′
J
J′

Figure 2.16. Rosetteless staining in wild type and mutant cells. Jumble and Couscous cells grown without and with RIFs were stained for Rosetteless (magenta), tubulin (gray), and actin (cyan). In uninduced Jumble and Couscous cells Rosetteless staining localizes to the basal pole, similar to wild type cells. Following treatment of wild type cells with RIFs, Rosetteless staining becomes highly enriched in the center of rosettes and must be imaged with less gain for clarity (0.3% laser power, gain=650 indicated by green boxes), while Rosetteless is not enriched or apparently secreted from the basal poles of Jumble or Couscous cells (2% laser power, gain=800). Scale bar = 20 μm.

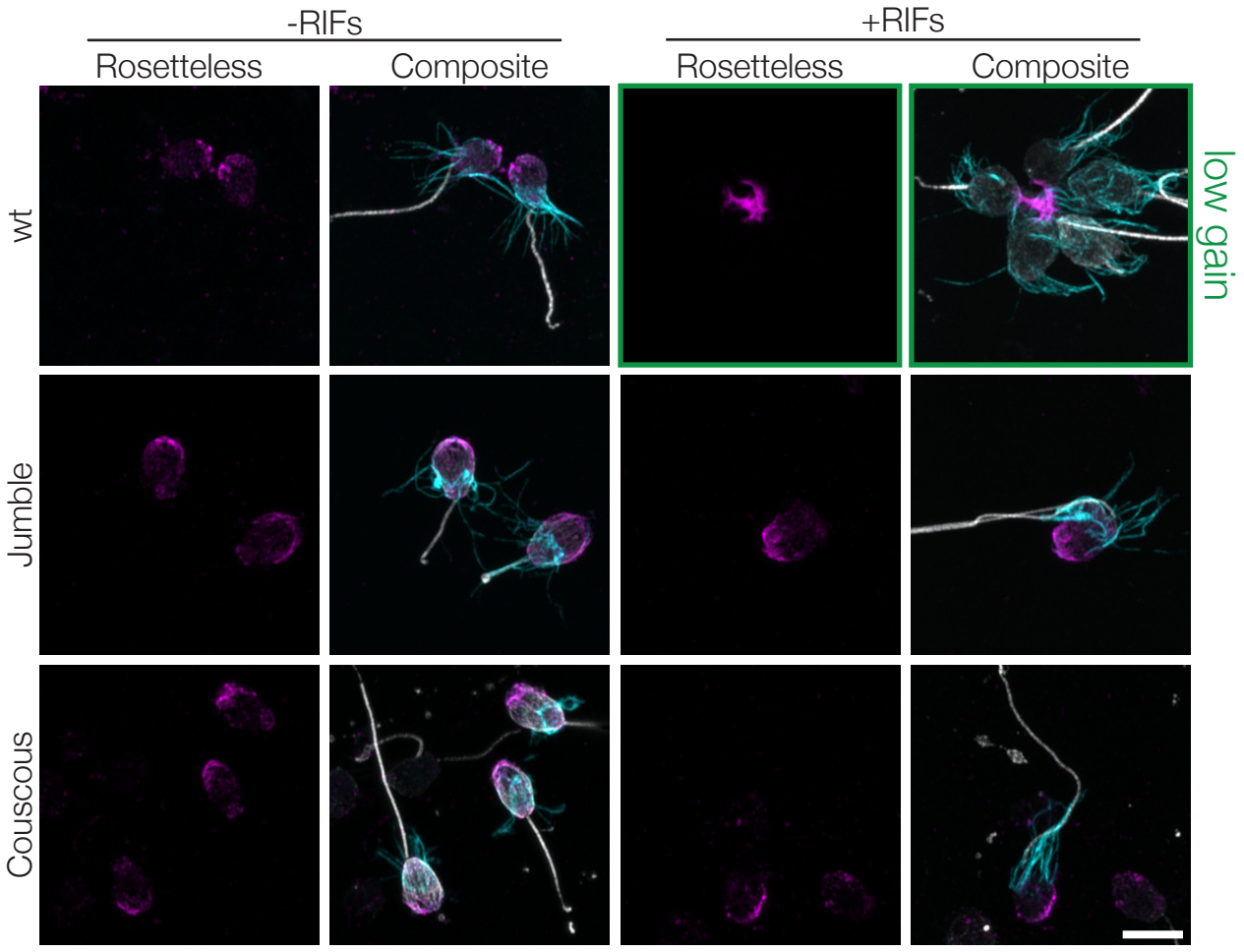

-RIFs
+RIFs
Rosetteless
Composite
Rosetteless
Composite
wt
Jumble
Couscous
low gain

Figure 2.17. Model for promiscuous clumping in rosette defective Class C mutants. Wild type *S. rosetta* has a glycosylated basal patch of ECM (red) as marked by the lectin jacalin that becomes enriched during the course of rosette formation. The Rosetteless protein, required for rosette formation and speculated to play a structural role in holding rosettes together, localizes to the same location on the basal pole of cells and becomes similarly enriched as rosette form. Mutants lack the glycosylated basal patch of jacalin staining. The altered cell surface could lead to clumping, either through mis-regulation of cell adhesion molecules or exposure of a normally masked adhesive cell surface. The same alteration that allows clumping of Class C mutants also prevents rosette development, perhaps by disrupting glycan modification on the Rosetteless protein or one of its interaction partners.

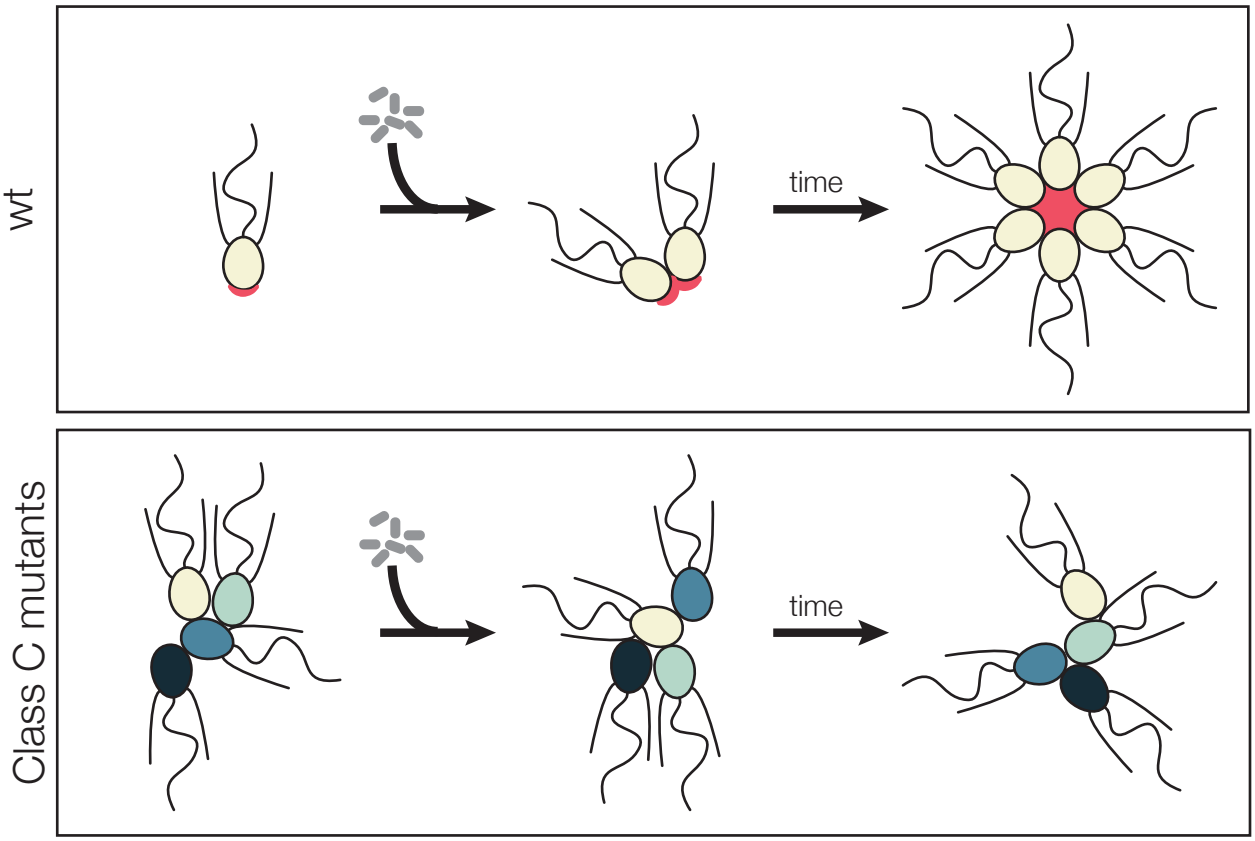

wt
time
Class C mutants
time

Table 2.1. Phenotypes of wild type and Class C mutants.

Strain	% cells in rosettes	Cell interactions	Successful outcross?
wild type	87.7	Non-clumping	Yes
Seafoam	0	Clumping	No
Soapsuds	0	Clumping	No
Couscous	0	Clumping	Yes
Jumble	0	Clumping	Yes

Table 2.2. Phenotypic classes of mutants isolated in this study and in the Levin *et al.* 2014 screen.

Mutant Class	Strain	Mutagen	Single cell interactions	Rosette morphology	% cells in rosettes	
					Live *Algori-phagus*	OMVs
	wt	-	Non-clumping	wt	87	95.2
	Rosetteless†	EMS	Non-clumping	n.d.	0	0
Class A	M7G9	X-rays	Non-clumping	n.d.	0	0
	M1A1.F3	Spontaneous	Non-clumping	n.d.	0	0
	Insensate†	X-rays	Non-clumping	Irregular	3.1	5.6
	Slacker†	X-rays	Non-clumping	Irregular	42.4	nt
Class B	Uptight†	X-rays	Non-clumping	Irregular	53.1	nt
	M14A9.D5	X-rays	Non-clumping	Irregular	56.9*	76.3*
	M17C12	X-rays	Non-clumping	Irregular	nt	24*
	Jumbled†	EMS	Clumping	n.d.	0	0
	Couscous†	X-rays	Clumping	n.d.	0	0
	Seafoam†	X-rays	Clumping	n.d.	0	0
Class C	Soapsuds†	X-rays	Clumping	n.d.	0	0
	M5G11.E8	X-rays	Mild clumping	n.d.	0	0
	M13H12.G2	X-rays	Mild clumping	wt	5.2	4
	M1C5.D2	X-rays	Mild clumping	wt	0	37*
Class D‡	Solo	X-rays	Non-clumping	No rosettes	0	nt

†Originally reported in Levin *et al.,* 2014; n.d.= not detected; nt= not tested; * ≤ 2 biological replicates; ‡Class D mutant fails to form chains in the absence of RIFs and is therefore distinct from Class A mutants.

Table 2.3. Segregating variants in Rosetteless mapping cross.

Supercontig	Location	Position relative to genes	Type	Coverage†
4	516,051	intron	INDEL*	8
6	1,139,589	5' UTR	INDEL*	40
8	**427,804**	**splice donor**	**SNV****	**253**
11	524,974	intron	INDEL*	12
11	1,660,350	intron	INDEL*	6
Average genome-wide coverage†: 187				

†Number of high quality reads determined by SAMtools (Li et al., 2009) at nucleotide position; *Insertion or deletion; **Single nucleotide variant; Highlighted sequence variant indicates known causative lesion (Levin et al., 2014).

Table 2.4 Segregating variants in Jumble mapping cross.

Supercontig	Location	Position relative to genes	Type	Coverage†
1	**1,919,681**	**coding sequence**	**SNV****	**165**
20	530,561	intron	INDEL*	3
22	65,983	intron	INDEL*	37
32	134,832	intron	INDEL*	5
49	3,863	intron	SNV**	2
Average genome-wide coverage†: 187				

†Number of high quality reads determined by SAMtools (Li et al., 2009) at nucleotide position; *Insertion or deletion; **Single nucleotide variant; Highlighted sequence variant indicates predicted causative lesion.

Table 2.5. Segregating variants in Couscous mapping cross.

Supercontig	Location	Position relative to genes	Type	Coverage†
3	1,812,030	splice acceptor	INDEL*	2
4	475,982	intron	INDEL*	10
4	518,253	intron	INDEL*	12
5	533	intron	INDEL*	3
9	141,246	intron	INDEL*	3
13	698,752	intron	INDEL*	6
22	110,265	intron	INDEL*	5
22	**462,534**	**coding sequence**	**INDEL***	**128**
Average genome-wide coverage†: 72				

†Number of high quality reads determined by SAMtools (Li et al., 2009) at nucleotide position; *Insertion or deletion; Highlighted sequence variant indicates predicted causative lesion.

Table 2.6. Fluorescent lectins tested.

Lectin	Preferred Sugar Specificity*	wild type Localization	Jumble Localization	Couscous Localization
Con A (Concanavalin A)	αMan, αGlc	Bacteria	Bacteria	Bacteria
SBA (*Glycine max* (soybean) agglutinin)	α>βGalNAc	Faint cell body	Faint cell body	Faint cell body
DBA (*Dolichos biflorus* agglutinin)	αGalNAc	Faint cell body	Faint cell body	Faint cell body
DSL (*Datura Stramonium* lectin)	$(GlcNAc)_{2-4}$	n.d.	n.d.	n.d.
ECL (*Erythrina cristagalli* lectin)	Galβ4GlcNAc	n.d.	n.d.	n.d.
GSL I (*Griffonia (Bandeiraea) simplicifolia* lectin I)	αGal, αGalNAc	Patchy cytoplasmic	Patchy cytoplasmic	Patchy cytoplasmic
GSL II (*Griffonia (Bandeiraea) simplicifolia* lectin II)	α or βGlcNAc	n.d.	n.d.	n.d.
Jacalin	**Galβ3GalNAc**	**Basal pole and collar base**	**Collar base**	**Collar base**
LCA (*Lens culinaris* agglutinin)	αMan, αGlc	n.d.	n.d.	n.d.
LEL (*Lycopersicon esculentum* (tomato) lectin)	$(GlcNAc)_{2-4}$	Collar base	Collar base	Collar base
MAL I (*Maackia Amurensis* lectin I)	Galβ4GalNAc	Cell membrane	Cell membrane	Cell membrane
PHA-E (*Phaseolus vulgaris* Erythroagglutinin)	Galβ4GlcNAcβ 2Manα6 (GlcNAcβ4) (GlcNAcβ4Man α3) Manβ4	Faint cell body	Faint cell body	Faint cell body
PHA-L (*Phaseolus vulgaris* Leucoagglutinin)	Galβ4GlcNAcβ 6(GlcNAc β2Manα3)Man α3	n.d.	n.d.	n.d.
PNA (*Arachis hypogaea* (peanut) agglutinin)	Galβ3GalNAc	n.d.	n.d.	n.d.
PSA (*Pisum sativum* Agglutinin)	αMan, αGlc	Faint cell body	Faint cell body	Faint cell body
RCA_{120} (*Ricinus communis* agglutinin)	Gal	Cell body	Cell body	Cell body
SNA (*Sambucus Nigra* Lectin)	Neu5Acα6Gal/ GalNAc	n.d.	n.d.	n.d.
STL (*Solanum tuberosum* (potato) lectin)	$(GlcNAc)_{2-4}$	Collar base	Collar base	Collar base
Succinylated WGA (Wheat germ agglutinin, succinylated)	GlcNAc	Cell membrane	Cell membrane	Cell membrane
UEA I (*Ulex europaeus* agglutinin I)	αFuc	n.d.	n.d.	n.d.
VVL (*Vicia villosa* agglutinin)	GalNAc	Faint cell body	Faint cell body	Faint cell body
WGA (*Triticum vulgaris* (wheat germ) agglutinin)	GlcNAc	Cell membrane	Cell membrane	Cell membrane

*From Vector Laboratories Product Information

Symbols and abbreviations:			
n.d.	not detected	Glc	D-Glucose
>	preference for first over	GlcNAc	*N*-Acetyleglucosamine
Fuc	second sugar	Man	Mannose
Gal	L-Fucose	Neu5Ac	N-Acetylneuraminic acid (sialic acid)
GalNAc	D-Galactose *N*-Acetylgalactoamine		

Appendix

Improved genome assembly and the regulatory genome of *S. rosetta*

Laura Wetzel conceived and designed genome assembly experiments, analyzed data, and wrote the manuscript. ATAC-seq nuclei and library preparation were performed by David Booth and Lily Helfrich. Stefan Prost aided in the genome assembly analysis.

INTRODUCTION

Complex multicellularity, including numerous specialized cell types arising from temporally and spatially regulated developmental programs, is a hallmark of animals. Recent genomic analyses of holozoans have revealed that the last common unicellular ancestor of animals already had a complex gene repertoire involved in multicellular functions, including transcription factors, extracellular matrix components, and intricate signaling pathways that were previously considered to be animal specific (de Mendoza et al., 2015; Fairclough et al., 2013; Grau-Bové et al., 2017; King et al., 2008; Richter et al., 2018; Sebé-Pedrós et al., 2016; Srivastava et al., 2010; Suga and Ruiz-Trillo, 2013). Since many genes required for animal development were present before animals arose, an emerging hypothesis proposes that animal evolution may have relied on new genome regulatory capabilities to generate cell types during animal development (Sebé-Pedrós et al., 2016).

Transcriptional regulation underlies much of cell differentiation during animal development. Distinct transcriptional profiles are established and maintained by a complex combination of chromatin regulatory dynamics, distal *cis*-regulatory elements and transcription factor networks (Bernstein et al., 2007; Buecker and Wysocka, 2012; de Laat and Duboule, 2013; Ho et al., 2014; Levine, 2010; Levine and Tjian, 2003). However, the evolutionary origins of these regulatory mechanisms remain unclear.

The closest unicellular relatives of animals, including ichthyosporeans, filastereans, and choanoflagellates, have complex life histories complete with morphologically distinct cell types (Dayel et al., 2011; Marshall et al., 2008; Marshall and Berbee, 2011; Sebé-Pedrós et al., 2016). Recent studies of the filasterean, *Capsaspora owczarzaki*, and the ichthyosporean, *Creolimax fragrantissima*, have revealed that different cell types have distinct transcriptional profiles (de Mendoza et al., 2015; Sebé-Pedrós et al., 2016). Moreover, *C. owczarzaki* appears to lack animal promoter types and its regulatory sites are small, proximal, and lack signatures of animal enhancers (Sebé-Pedrós et al., 2016).

Choanoflagellates, as the closest living relatives of animals, may provide additional insights in the regulatory toolkit available to progenitor of animals. Transcriptional analysis of facultatively multicellular *Salpingoeca rosetta* revealed distinct transcriptomes between attached thecate cells, swimming cells (including slow and fast swimmers), chain colonies, and rosettes colonies (Fairclough et al., 2013). However, little is known about mechanisms, such as cell-type specific promoters or enhancers, that regulate these transcriptional differences.

The recently developed technique assay for transposase-accessible chromatin using sequencing (ATAC-seq), which enables rapid detection of open chromatin and the regions of nucleosome-bound and nucleosome-free positions in regulatory regions, can provide a basis for uncovering epigenetic regulation (Buenrostro et al., 2013). ATAC-seq relies on the transposase, Tn5, to insert sequencing adapters into only accessible regions of chromatin; this is followed by a PCR reaction that amplifies DNA fragments between inserted adapters of preferentially open chromatin (Buenrostro et al., 2013). Here, we use ATAC-seq to assay the open chromatin of four distinct *S. rosetta* cell types: thecate cells, fast swimmers, slow swimmers, and rosettes (Figure A.1). Additionally, to improve ATAC-seq analysis and to increase sequence information about regulatory elements, we

generated an improved genome assembly using newly developed long-read sequencing technologies and using Hi-C chromatin capture to increase scaffold lengths.

RESULTS

Improving the *S. rosetta* genome assembly

To improve the assembly of the *S. rosetta* genome, we developed a new method of reducing contaminating DNA from prey bacteria through 48 h of rifampicin treatment and by pelleting choanoflagellates at low-speed centrifugation and washing with seawater. This method avoided the need to use a CsCl gradient, which separated bacteria and choanoflagellate DNA by GC-content and therefore prevented recovery of any low GC-content choanoflagellate DNA. For sequencing, we utilized two new long-read sequencing techniques: single-molecule, real-time (SMRT) sequencing technology using the PacBio RSII and nanopore sequencing on the MinION from Oxford Nanopore Technologies. Draft genomes were assembled from each of these technologies alone; then, we used Dovetail Hi-C and HiRise assembly to improve scaffold lengths for each assembly and the original 454 assembly in Fairclough et al., 2013. Here we report on 3 improved *de novo* assemblies of the *S. rosetta* genome and will use the Pac+Hi-Rise assembly as the new *S. rosetta* reference genome for future studies (Table A.1).

Due to the ability to recover low GC content DNA, we were able to assemble a contig of putative mitochondrial DNA (mtDNA) of *S. rosetta*. The majority of animal mtDNAs are relatively small (~13-19 kbp in size) due to gene loss from the mitochondrial genome, but the only choanoflagellate mtDNA genome assembly to date, *Monosiga brevicollis,* was 76.6 kbp and conserved genes that were thought to be present in the alphaproteobacterial ancestor (Burger et al., 2003). The size of the *M. brevicollis'* unreduced mtDNA was a key piece of supporting evidence that choanoflagellates are the outgroup to animals (Burger et al., 2003). Consistent with the expanded size of the *M. brevicollis*, *S. rosetta* mtDNA was found to be 81.7 kbp (Table A.2). This provides additional evidence that last common ancestor of animals and choanoflagellates retained an expanded mitochondrial genome which was specifically reduced along the animal stem lineage. A similar process is thought to have occurred within the fungi (Bullerwell and Lang, 2005).

Comparing chromatin accessibility between *S. rosetta* life histories

Eukaryotic genomes are hierarchically packaged into chromatin (Kornberg, 1974) and the nature of this packaging plays a central role in gene regulation (Kornberg and Lorch, 1992; Mellor, 2005). To interrogate accessible chromatin as a proxy for active *cis*-regulatory elements, we carried out ATAC-seq on ten million unfixed nuclei from four distinct *S. rosetta* cell types: thecate cells, fast swimmers, slow swimmers, and rosettes (Figure A.1). Mapped reads were reproducible between technical replicates (Spearman Correlation Coefficient R=0.91-0.98; Figure A.2D); thus, the reads from technical replicates were combined for some comparative analyses between cell types.

Previous studies found that ATAC-seq reads produced detailed information about nucleosome packing and positioning (Buenrostro et al., 2013). The insert size distribution of sequenced fragments from different *S. rosetta* cell types showed a hint of periodicity of approximately 200 bp, primarily in the fast swimmers, suggesting fragments might be

protected by integer multiples of nucleosome (Figure A.2E). Based on this periodicity and previously observed periodicity (Buenrostro et al., 2013), reads were separated into nucleosome-free and mono-nucleosome by length. Others have found nucleosome-free fragments are enriched at canonical nucleosome-free promoter regions overlapping with transcription start sites (TSSs), whereas nucleosome signal is enriched both upstream and downstream of the active TSS (Buenrostro et al., 2013). Examining the highly expressed, *elongation factor L* (*efl*) in *S. rosetta* slow swimmer populations upholds this expected pattern (Figure A.3A).

To identify potential regulatory differences between cell types, we identified genomic regions enriched with ATAC-seq reads using the MAC2s peak caller for each cell type (Y. Zhang et al., 2008). We compared the called ATAC-seq peaks between cell types. PCA analysis showed that slow swimmers and rosettes clustered together while fast swimmers and thecate cells formed individual clusters (Figure A.3B). Moreover, all ATAC-seq peaks were compared between each cell type (Figure A.3C). Interestingly, there were no differential peaks between slow swimmers and rosettes; but there were >1000 differential peaks between every other cell type pairwise (Wald test $p<0.01$; Figure A.3C). Thus, while all other *S. rosetta* cell types have distinct chromatin accessibility profiles, there is no detectable difference between slow swimmers and rosettes in open chromatin and thereby, presumably no differential transcriptional regulation.

DISCUSSION AND FUTURE DIRECTIONS

The improved genome assembly and ATAC-seq of *S. rosetta* cell types provides a new set of tools to untangle the role of transcription regulation in life history transitions and infer the regulatory mechanisms present in the last common ancestor of animals. We found distinct chromatin accessibility profiles for each life history examined, although rosettes and slow swimmers appear indistinguishable. Consistent with transcriptome sequencing that revealed that slow swimmers and rosettes have remarkably similar genome-wide transcription profiles (Fairclough et al., 2013), our results add further support to the hypothesis that the transition from slow swimmers to rosettes is not transcriptionally regulated and may be dictated by translational or post-translational modifications.

A key step to validating the observed ATAC-seq data is to corelate peaks to high quality transcriptomes from each examined life history. Highly expressed genes should have higher peak intensities at and around their TSSs if the assay is capturing open chromatin and chromatin accessibility is a true readout of transcription levels (Buenrostro et al., 2013). Correlating open chromatin to epigenetic marks, like histone modifications, and to RNA polymerase through ChIP-seq may provide further insight into the molecular mechanisms of regulation utilized in *S. rosetta*.

Examining the identified differentially open chromatin near TSSs can help to reveal life history specific promoters and enhancer motifs. If identified promoter motifs are conserved with other known motifs, prediction tools might be able to predict the transcription factors that control life history specific expression. It has also been proposed that distal enhancers are an animal innovation (Sebé-Pedrós et al., 2016). The ATAC-seq data can be used to search for the presence of distal enhancers in choanoflagellates

by identifying open chromatin at least 1 kb from the nearest gene and could be validated using newly developed transgenic techniques (Booth et al., 2018).

MATERIALS AND METHODS

Media and cell culture

AK artificial seawater (AK), high nutrient (HN) media, and cereal grass (CG) media were prepared as described previously (Booth et al., 2018; Levin et al., 2014; Levin and King, 2013). The clonally isolated strain SrEpac (ATCC PRA-390; accession number SRX365844) — *S. rosetta* co-cultured monoxenically with the prey bacterium *Echinicola pacifica* (Levin et al., 2014; Levin and King, 2013; Nedashkovskaya et al., 2006) — was used for all experiments. For routine maintenance, cultures were passaged every 2-3 days in 5% (vol/vol) HN media in AK.

Genomic DNA preparation and sequencing for improved assembly

To reduce bacterial contamination for genomic DNA preparation, 200 mL of stationary phase cultures were pelleted at 2400 x g, resuspended in AK, and treated with 5 µg/mL rifampicin for 48 h. Cells were harvested by pelleting at 2400 x g at 4°C. Phenol-chloroform extractions were performed as described previously with careful pipetting to avoid shearing DNA (King et al., 2008).

PacBio SMRT sequencing was performed on PacBio RSII instrument (Pacific Biosciences of California Inc., Menlo Park, CA, USA) at the UC Davis Genome Center. In total, 5 SMRT cells were sequenced.

For Nanopore sequencing, 8 µg DNA was sheared to 10 kb, and performed a 4 kb high pass size selection on the Blue Pippin (Sage Science). The sample was ligated with PCR adapters, split into four 100 ng PCR reactions, and amplified with NEB LongAMP (Cat. No. MO323). 5 µg of the PCR amplified sample was used in the standard ligation 1D protocol (Oxford Nanopore Technologies SQK-LSK108). Prepared libraries were sequenced on MinION (Oxford Nanopore Technologies).

Draft genome assembly

PacBio data were assembled with Sprai assembler using the default settings (http://zombie.cb.k.u-tokyo.ac.jp/sprai/index.html). Nanopore data were assembled with Miniasm assembler using the default settings (Li, 2016) and polished with RACON using the default settings three times (Vaser et al., 2017).

Chicago library preparation

To improve the assembly, we created a Chicago library (Putnam et al., 2016) at Dovetail Genomics. The Chicago library was sequenced on an MiSeq v3 2x300. The 3 draft genomes (the previous *S. rosetta* assembly (Fairclough et al., 2013), the PacBio assembly, and the Nanopore assembly) in FASTA format, and Chicago library sequence (3,159.77X- 5,018.49X based on input assembly; PE~300 bp) in FASTQ format were used as input data for HiRise (Putnam et al., 2016). HiRise is a software pipeline designed specifically for using Chicago library sequence data to assemble genomes. The number of BUSCOs (benchmarking Universal Single-Copy Orthologs) was determined after HiRise using the eukaryote odb9 dataset (Waterhouse et al., 2018). A mitochondrial

scaffold was determined by BLAST with the *M. brevicollis* mitochondrial genome; however, in the future, a better assembly may be generated using specific mitochondrial genome assembly programs.

ATAC-seq protocol

(1) Prepare nuclei

Nuclei were isolated from different cell types: (1) slow swimmers, (2) rosettes, (3) fast swimmers, and (4) thecate cells with two independent replicates. Slow swimmers were generated by maintaining cells in standard culture conditions. Rosettes were induced in SrEpac with outer membrane vesicles from *Algoriphagus machipongonensis* as previously described (Woznica et al., 2016). Fast swimmers were grown 3 days in standard media and then heat shocked at 30°C for 2.75 h. Slow swimmers, rosettes, and fast swimmers were harvested by pelleting at 2400 x g for 5 min, washed with 50 mL AK seawater, re-pelleted at 2400 x g for 5 min, counted with the Luna cell counter, diluted to $5x10^7$ cells/mL, and $1x10^7$ cells were pelleted at 2700 x g. Thecate cells were derived from an isolate of SrEpac, called HD1, and maintained in 10% CG in AK seawater (vol/vol) in petri dishes. To harvest thecate cells, plates were washed with 16.7 mL of AK seawater, cells lifted from the plate with a cell scraper, and filtered onto a 3 µm polycarbonate membrane filter to concentrate. Filtered cells were pelleted at 2700 x g for 5 min, washed with 50 mL AKSW two times, re-pelleted at 2700 x g for 5 min, counted with Luna cell counter, diluted to $5x10^7$ cells/mL, and $1x10^7$ cells pelleted at 2700 x g. All cell types were resuspended in 200 µL freshly prepared pretreatment buffer (10 mM citric acid, 100 mM LiOAC, 10% (w/v) PEG 800 pH 8.5 with Tris, 100 nM papain, and 10 mM thioglycolic acid) and incubated at room temperature for 22 min.

Nuclei were isolated in four steps wash, strip, lyse, and purify (Figure A.2A). To wash, cells were pelleted at 1200 x g for 5 min, the supernatant discarded, and resuspended in 200 µL of 0.7M sorbitol in 1x PBS and 1% (w/v) BSA, and pelleted at 1200 x g for 5 min. Pellets were resuspend in 250 µL cold buffer L (10 mM HEPES-KOH pH 7.9, 0.2 mM $MgCl_2$, 10 mM KCl, 0.1 mM EDTA-KOH pH 8.0, 0.5 mM EGTA-KOH pH 8.0, 0.5 mM DTT, 0.5 mM pefabloc, and 1x Roche protease inhibitor solution) and incubated for 10 min on ice. To lyse cells, 0.05% NP40 was added, cells were incubated on ice for 10 min, and then samples were passed through 30G needle ten times. Lysed cells were pelleted at 1000 x g for 5 min at 4°C, supernatant was removed. Pellets were resuspended in 250 µL buffer L with sucrose (Buffer L, 250 mM sucrose, 0.5 mM DTT, 0.5 mM pefabloc, and 1x Roche protease inhibitor solution), spun at 1000 x g for 5 min at 4°C, and both steps repeated.

To examine the purity of the nuclei, Western blots were run with a tubulin and H3 antibody (Figure A.2B).

(2) Transpose and purify

For transposition, nuclei were pelleted at 1000 x g for 5 min at 4°C, resuspended in 25 µL of 2x TD buffer and 2.5 µl of Tn5 transposase from the Nextera DNA Library Prep it (Illumina, San Diego, CA), and incubated at 37°C for 30 min. DNA was purified using the Qiagen MinElute kit (Cat. No. 28004) per PCR purification protocol provided by the manufacturer.

(3) PCR

Transposed DNA was originally amplified and barcoded in a PCR reaction using NEBnext PCR master mix (NEB Cat. No. #M0544) and 1.25 µM forward and reverse primers originally described in Buenrostro et al., 2013 (Table A.3), using the following PCR conditions: 72°C for 5 min; 98°C for 30 s; and thermocycling at 98°C for 10 s, 63°C for 30 s and 72°C for 1 min. To reduce GC and size bias in the PCR, we monitored the PCR reaction using qPCR in order to stop amplification before saturation. To do this, we amplified the full libraries for five cycles, after which we took an aliquot of the PCR reaction and added 10 µl of PCR cocktail with Sybr Green at a final concentration of 0.6x. We ran this reaction for 20 cycles to determine the additional number of cycles needed for the remaining 45 µL reaction. The libraries were purified using a Qiagen PCR cleanup kit. Libraries were amplified for a total of 10-12 cycles. An additional 0.9X SPRI bead clean-up was performed to eliminate a contaminating 50 bp peak. We quantified our libraries using qPCR-based methods.

(4) Primary data processing

Data were collected using 50 x 8 x 50 reads on an Illumina HiSeq 2500 run in rapid mode. *E. pacifica* contaminating reads varied between samples, but for all samples, we were able to collect at least 5,000,000 reads (Figure A.2C). Raw reads were trimmed with TrimmomaticPE (Bolger et al., 2014) to remove low quality base calls and barcodes/adaptors. Trimmed reads were aligned to the *S. rosetta* reference genome (Fairclough et al., 2013) using Bowtie (Langmead et al., 2009) with the parameters -X2000 and -m1. These parameters ensured that only fragments up to 2 kb were allowed to align (-X2000) and that only unique aligning reads were collected (-m1).) For all data files, we removed PCR duplicates with Picard tools (http://broadinstitute.github.io/picard/). Unmapped reads were aligned to the *E. pacifica* reference genome (Nedashkovskaya et al., 2006) to determine how much bacterial contamination was present (Figure A.2C).

To examine, the reproducibility between sample reads Deeptools (Ramírez et al., 2016) was used to: (1) normalize reads using reads per kilobase million with bamCoverage, (2) compare read counts using multiBigwigSummary, and (3) plot Spearman correlations between samples using plotCorrelation (Figure A.2D). Given high correlations between samples, replicates were combined for specified analyses.

The distribution of paired-end sequencing fragment sizes was determined with Picard tools (http://broadinstitute.github.io/picard/). The read density was plotted against insert size to look for evidence of insert sizes characteristic of mono-, di-, and tri-nucleosomes (Figure A.2E).

For peak-calling, we adjusted the read start sites to represent the center of the transposon binding event. Previous descriptions of the Tn5 transposase show that the transposon binds as a dimer and inserts two adaptors separated by a 9 bp (Adey et al., 2010; Buenrostro et al., 2013). Therefore, all reads aligning to the plus strand were offset by +4 bp, and all reads aligning to the minus strand were offset -5 bp as previously described (Buenrostro et al., 2013).

Finally, reads were classified into nucleosome free reads (paired-read distance <100 bp) and mononucleosome reads (pair-read distance between 180-247 bp), representing single nucleosomes.

(5) ATAC-seq peak-calling

Peak calling was performed using MACS2 (Y. Zhang et al., 2008) with the following parameters: -g 550000000 -q 0.1 --extsize 40 --call-summits --nomodel. Peaks from different samples were merged using Bedtools (Quinlan and Hall, 2010) to generate the final dataset of 45,176 peaks.

Peaks were visualized in the Integrative Genomics Viewer (Figure A.3A) (Thorvaldsdottir et al., 2013). Peaks were compared using DESeq2 (Figure A.3C) (Love et al., 2014).

ACKNOWLEDGMENTS

Winston Timp, Rachael Workman, Hitomi Asahara, and Aaron Pomerantz helped troubleshooting and performing Nanopore sequencing. We thank the members of the King lab for helpful discussions, research support, and comments on the manuscript. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

Figure A.1. *S. rosetta* life history. *S. rosetta* has several morphologically distinct life histories of solitary and colonial forms. For this study, we focus on studying the solitary fast swimmers, thecate cells and slow swimmers, and the colonial rosette colonies. Illustration credit: Janet Iwasa.

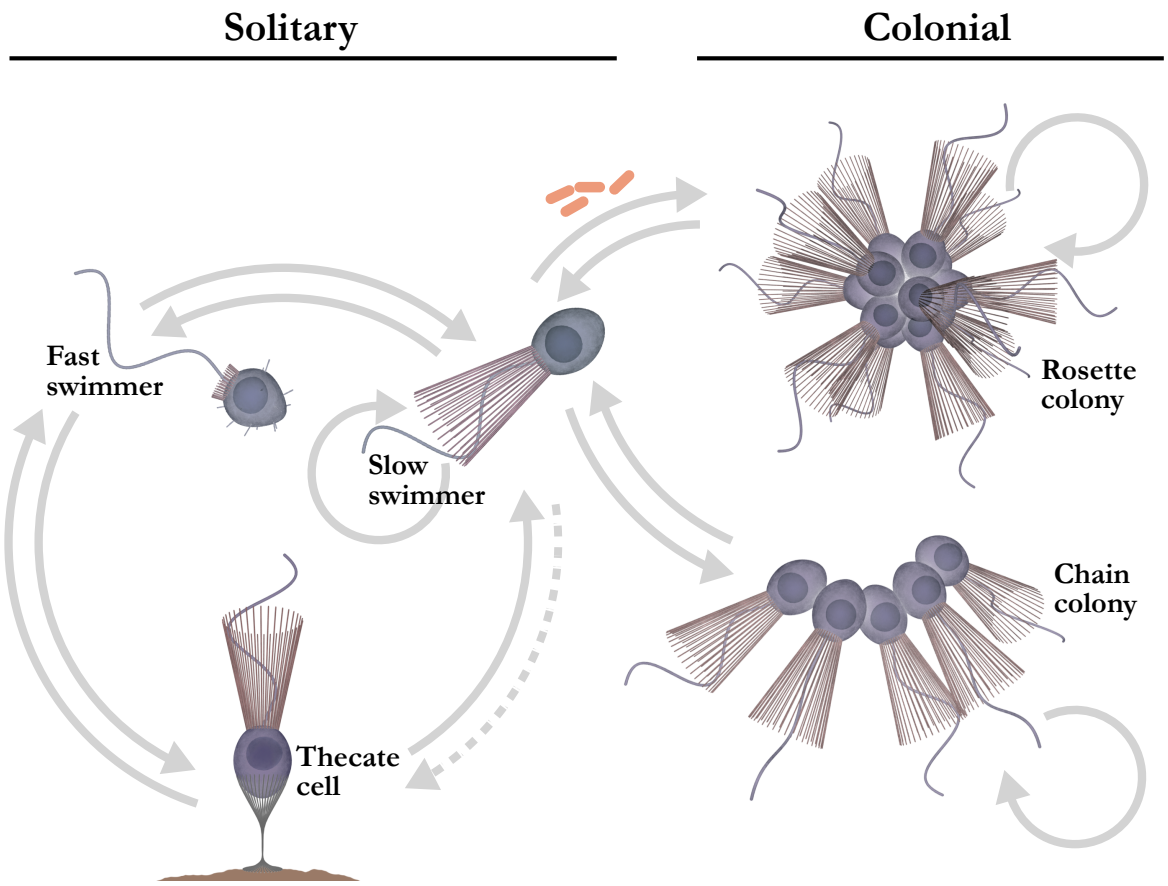


Illustration credit Janet Iwasa

Figure A.2. Development of ATAC-seq in *S. rosetta* cell types. (A) A novel protocol to purify nuclei from *S. rosetta* was established through four steps: (1) cells were washed to remove excess bacteria, (2) ECM was stripped from the remaining cells, (3) cells were lysed, and (4) nuclei were purified. **(B)** Whole cell lysate, intermediate supernatants, and final purified nuclei were probed for tubulin and histone H3 (H3). As expected for purified nuclei, we could not detect any tubulin, but retained the H3 signal. **(C)** The ATAC-seq reads contained variable amounts of contaminating *Echinicola pacifica* mapped reads from the prey bacteria. For each sample, we were able to collect at least 5,000,000 reads mapped to the *S. rosetta* genome. **(D)** Mapped reads between technical duplicates were compared using Spearman correlation and found to be reproducible between duplicates (Spearman correlation coefficient R=0.91-0.98). The similarity of samples based on the nearest point algorithm from Deeptools (Ramírez et al., 2016) is depicted in the red tree that is used to generate heatmap clustering. Heat map colors correspond to the calculated Spearman's correlation coefficient. **(E)** Insert size of the ATAC-seq *S. rosetta* mapped reads were determined and plotted as a function of read density. The insert size distribution of sequenced fragments from different *S. rosetta* life histories showed a hint of periodicity of approximately 200 bp in fast swimmers, suggesting fragments are likely protected by integer multiples of nucleosome, but is unclear in other cell types.

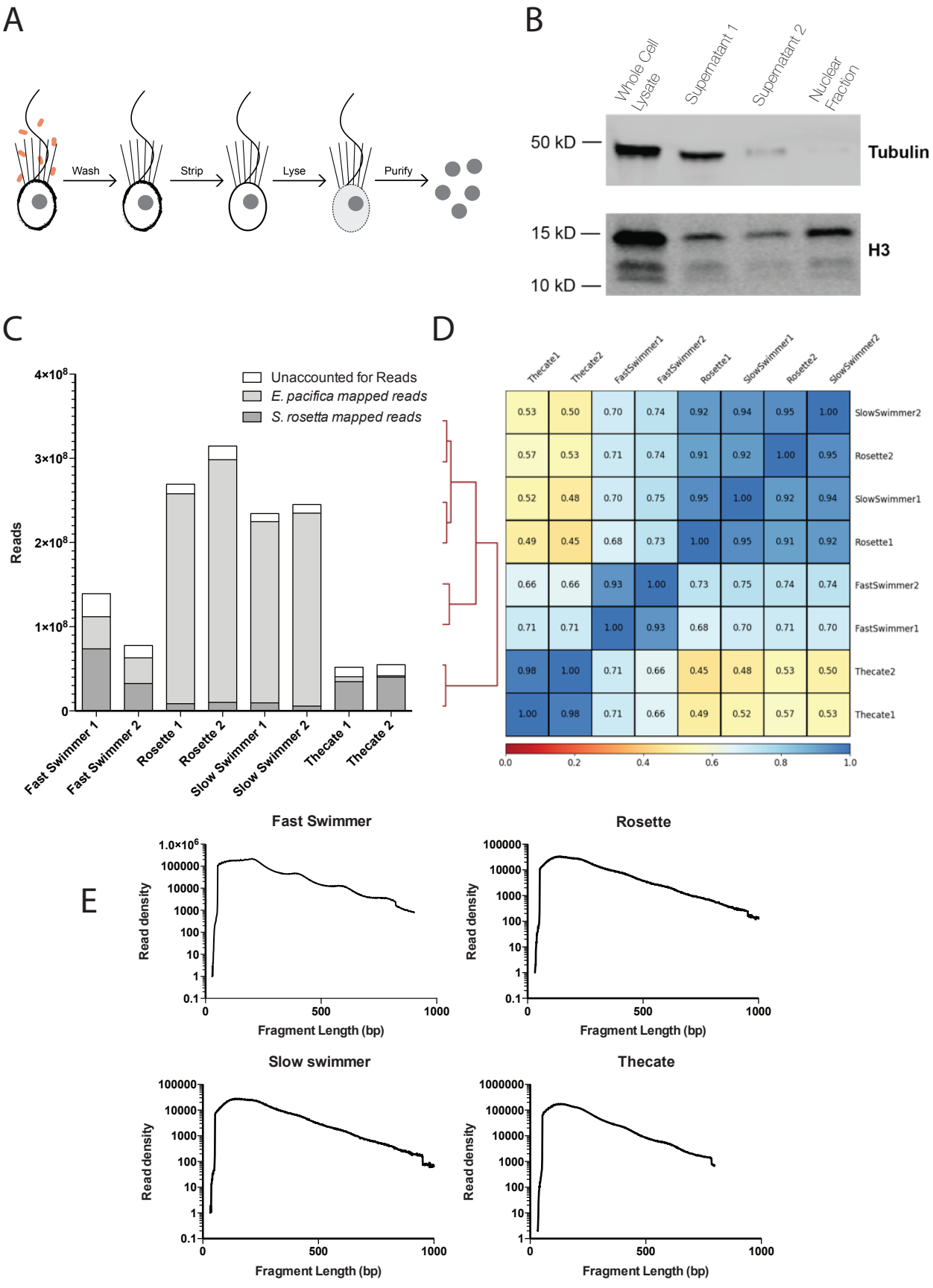
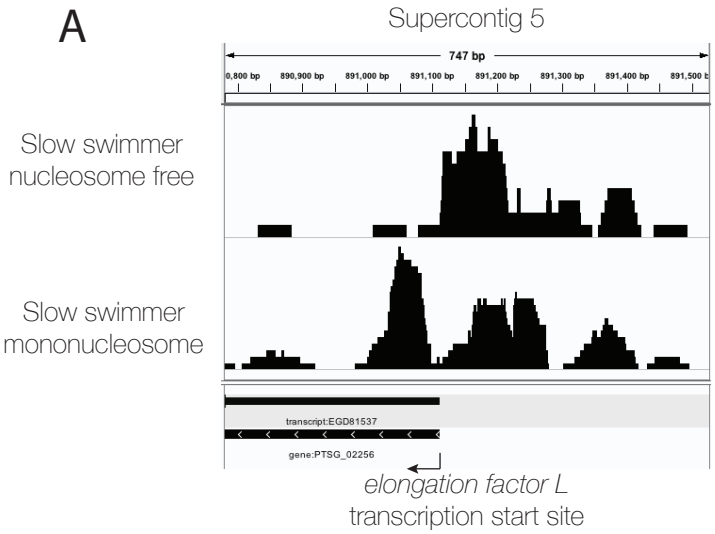
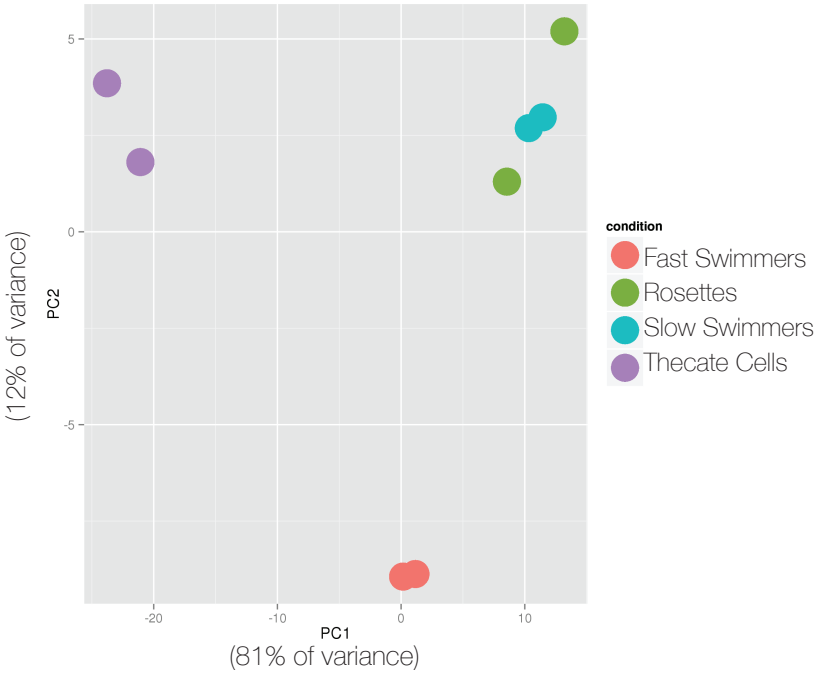

A
Wash
Strip
Lyse
Purify
B
Whole Cell Lysate
Supernatant 1
Supernatant 2
Nuclear Fraction
50 kD
Tubulin
15 kD
10 kD
H3
C
Unaccounted for Reads
E. pacifica mapped reads
S. rosetta mapped reads
Reads
4×10^8
3×10^8
2×10^8
1×10^8
0
Fast Swimmer 1
Fast Swimmer 2
Rosette 1
Rosette 2
Slow Swimmer 1
Slow Swimmer 2
Thecate 1
Thecate 2
D
Thecate1
Thecate2
FastSwimmer1
FastSwimmer2
Rosette1
SlowSwimmer1
Rosette2
SlowSwimmer2
0.0
0.2
0.4
0.6
0.8
1.0
E
Fast Swimmer
Rosette
Slow swimmer
Thecate
Read density
Fragment Length (bp)

Figure A.3. Enriched ATAC-seq reads at the *elongation factor L* (*efl*) transcription start site (TSS) and peak comparisons between *S. rosetta* cell types. (A) We identified genomic regions enriched with ATAC-seq reads using the MAC2s peak caller (Y. Zhang et al., 2008) for all cell types. An example of ATAC-seq peaks at the *efl* locus, a known highly expressed gene, are displayed for the slow swimmer population. Nucleosome-free reads are enriched at the TSS with mono-nucleosome reads flanking the *efl* TSS. **(B)** Principle component analysis (PCA) was performed on called peaks between cell types. PC1 accounted for 81% of the variance and PC2 accounted for 12% of the variance. Slow swimmers and rosettes cluster together, while thecate and fast swimmers form distinct clusters. **(C)** ATAC-seq peaks were compared between cell types for significant differences as determined by Wald test ($p < 0.01$). There were no significantly different peaks between slow swimmers and rosettes, but every other pairwise comparison produced >1000 differential peaks.

A



B



C

Number of peak differences between samples (Wald test p<0.01)

	Thecate	Fast Swimmer	Slow Swimmer	Rosette
Thecate				
Fast Swimmer	1,768			
Slow Swimmer	4,187	1,247		
Rosette	4,589	1,456	0	

Table A.1: Genome assembly statistics.

Input Assembly Basis	Sequencing Runs	Coverage	Assembly Size (Mb)	Scaffolds in Hi-Rise Assembly	L90/N90	Largest Scaffold	Single Copy BUSCOs
454 Original Assembly	-	33 x	55.54	154	28 scaffolds; 0.796 MB	2.6 Mb	218
454 Sequencing + Chicago Hi-C	-	33 x	55.55	95	29 scaffolds; 0.931 Mb	3.6 Mb	219
PacBio + Chicago Hi-C	3	70 x	60.18	358	35 scaffolds; 0.709 Mb	3.7 Mb	220
Nanopore + Chicago Hi-C	6	50x	52.79	266	31 scaffolds; 0.681 Mb	3.3 Mb	174

Table A.2: Mitochondrial genomes sizes.

Taxon	Size, kbp
S. rosetta	81.7
M. brevicollis	76.6*
Animals	13-22*
Fungi	19-94*
*(Burger et al., 2003)	

Table A.3: ATAC-seq primers.

Primer Name	Sequence	Samples
Ad1_noMX	AATGATACGGCGACCACCGAGATCTACACTCG TCGGCAGCGTCAGATGTG	All
Ad2.1_TAAGGCGA	CAAGCAGAAGACGGCATACGAGATtcgcctta GTCTCGTGGGCTCGGAGATGT	Slow swimmers 1
Ad2.2_CGTACTAG	CAAGCAGAAGACGGCATACGAGATctagtacg GTCTCGTGGGCTCGGAGATGT	Slow swimmers 2
Ad2.3_AGGCAGAA	CAAGCAGAAGACGGCATACGAGATttctgcct GTCTCGTGGGCTCGGAGATGT	Rosettes 1
Ad2.4_TCCTGAGC	CAAGCAGAAGACGGCATACGAGATgctcagga GTCTCGTGGGCTCGGAGATGT	Rosettes 2
Ad2.5_GGACTCCT	CAAGCAGAAGACGGCATACGAGATaggagtcc GTCTCGTGGGCTCGGAGATGT	Fast Swimmers 1
Ad2.6_TAGGCATG	CAAGCAGAAGACGGCATACGAGATcatgccta GTCTCGTGGGCTCGGAGATGT	Fast Swimmers 2
Ad2.7_CTCTCTAC	CAAGCAGAAGACGGCATACGAGATgtagagag GTCTCGTGGGCTCGGAGATGT	Thecate Cells 1
Ad2.8_CAGAGAGG	CAAGCAGAAGACGGCATACGAGATcctctctg GTCTCGTGGGCTCGGAGATGT	Thecate Cells 1

Lower case sequence indicates the unique barcode.

REFERENCES

Abedin M, King N. 2008. The premetazoan ancestry of cadherins. *Science* **319**:946–8. doi:10.1126/science.1151084

Adamska M, Degnan BM, Green K, Zwafink C. 2011. What sponges can tell us about the evolution of developmental processes. *Zoology* **114**:1–10. doi:10.1016/j.zool.2010.10.003

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology* **11**:R119. doi:10.1186/gb-2010-11-12-r119

Aktipis CA, Boddy AM, Jansen G, Hibner U, Hochberg ME, Maley CC, Wilkinson GS. 2015. Cancer across the tree of life: Cooperation and cheating in multicellularity. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**. doi:10.1098/rstb.2014.0219

Alegado RA, Brown LW, Cao S, Dermenjian RK, Zuzow R, Fairclough SR, Clardy J, King N. 2012. A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. *eLife* **1**. doi:10.7554/eLife.00013

Asada M, Furukawa K, Segawa K, Endo T, Kobata A. 1997. Increased expression of highly branched N-glycans at cell surface is correlated with the malignant phenotypes of mouse tumor cells. *Cancer research* **57**:1073–80.

Atmodjo MA, Hao Z, Mohnen D. 2013. Evolving Views of Pectin Biosynthesis. *Annual Review of Plant Biology* **64**:747–779. doi:10.1146/annurev-arplant-042811-105534

Baldauf SL. 2003. The deep roots of eukaryotes. *Science* **300**:1703–6. doi:10.1126/science.1085544

Bassagañas S, Carvalho S, Dias AM, Pérez-Garay M, Ortiz MR, Figueras J, Reis CA, Pinho SS, Peracaula R. 2014. Pancreatic Cancer Cell Glycosylation Regulates Cell Adhesion and Invasion through the Modulation of α2β1 Integrin and E-Cadherin Function. *PLoS ONE* **9**:e98595. doi:10.1371/journal.pone.0098595

Benabentos R, Hirose S, Sucgang R, Curk T, Katoh M, Ostrowski EA, Strassmann JE, Queller DC, Zupan B, Shaulsky G, Kuspa A. 2009. Polymorphic Members of the lag Gene Family Mediate Kin Discrimination in *Dictyostelium*. *Current Biology* **19**:567–572. doi:10.1016/j.cub.2009.02.037

Bernstein BE, Meissner A, Lander ES. 2007. Leading Edge The Mammalian Epigenome. *Cell* **128**:669–681. doi:10.1016/j.cell.2007.01.033

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120. doi:10.1093/bioinformatics/btu170

Bonner JT. 1998. The origins of multicellularity. *Integrative Biology* **1**:27–36. doi:10.1002/(SICI)1520-6602(1998)1:1<27::AID-INBI4>3.0.CO;2-6

Bonner JT. 1967. The Cellular Slime Molds, 2nd ed. Princeton University Press.

Booth DS, Szmidt-Middleton H, King N. 2018. Choanoflagellate transfection illuminates their cell biology and the ancestry of animal septins. *Molecular Biology of the Cell* **9**:3026–3038. doi:10.1091/mbc.E18-08-0514

Boraas ME, Seale DB, Boxhorn JE. 1998. Phagotrophy by a flagellate selects for colonial prey: A possible origin of multicellularity. *Evolutionary Ecology* **12**:153–164.

doi:10.1023/A:1006527528063
Bowman SM, Free SJ. 2006. The structure and synthesis of the fungal cell wall. *BioEssays* **28**:799–808. doi:10.1002/bies.20441
Brefeld O. 1869. *Dictyostelium mucoroides*. Ein neuer Organismus aus der Verwandshaft der Myxomyceten. *Abhandl Senckenbergish Naturf Ges* **7**:85–107.
Brunet T, King N. 2017. The Origin of Animal Multicellularity and Cell Differentiation. *Developmental Cell* **43**:124–140. doi:10.1016/j.devcel.2017.09.016
Buecker C, Wysocka J. 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends in genetics : TIG* **28**:276–84. doi:10.1016/j.tig.2012.02.008
Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin , DNA-binding proteins and nucleosome position. *Nature Methods* **10**. doi:10.1038/nmeth.2688
Bullerwell CE, Lang BF. 2005. Fungal evolution: the case of the vanishing mitochondrion. *Current Opinion in Microbiology* **8**:362–369. doi:10.1016/J.MIB.2005.06.009
Burger G, Forget L, Zhu Y, Gray MW, Franz Lang B. 2003. Unique mitochondrial genome architecture in unicellular relatives of animals. *Proceedings of the National Academy of Sciences* **100**:892–897.
Buss LW. 1988. The Evolution of Individuality. Princeton University Press.
Cambi A, Koopman M, Figdor CG. 2005. How C-type lectins detect pathogens. *Cellular Microbiology* **7**:481–488. doi:10.1111/j.1462-5822.2005.00506.x
Campbell JA, Davies GJ, Bulone V, Henrissat B. 1997. A classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities. *The Biochemical journal* **326 (Pt 3)**:929–39.
Carvalho S, Catarino TA, Dias AM, Kato M, Almeida A, Hessling B, Figueiredo J, Gartner F, Sanches JM, Ruppert T, Miyoshi E, Pierce M, Carneiro F, Kolarich D, Seruca R, Yamaguchi Y, Taniguchi N, Reis CA, Pinho SS. 2016. Preventing E-cadherin aberrant N-glycosylation at Asn-554 improves its critical function in gastric cancer. *Oncogene* **35**:1619–1631. doi:10.1038/onc.2015.225
Chen H, Lin F, Xing K, He X. 2015. The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nature Communications* **6**. doi:10.1038/ncomms7367
Colin Hughes R. 1992. Lectins as cell adhesion molecules. *Current Opinion in Structural Biology* **2**:687–692. doi:10.1016/0959-440X(92)90202-I
Colley KJ, Varki A, Kinoshita T. 2015. Cellular Organization of Glycosylation, Essentials of Glycobiology. Cold Spring Harbor Laboratory Press. doi:10.1101/GLYCOBIOLOGY.3E.004
Conway Morris S. 1993. The fossil record and the early evolution of the Metazoa. *Nature* **361**:219–225. doi:10.1038/361219a0
Daher FB, Braybrook SA. 2015. How to let go: pectin and plant cell adhesion. *Frontiers in Plant Science* **6**:1–8. doi:10.3389/fpls.2015.00523
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158. doi:10.1093/bioinformatics/btr330

Dayel MJ, Alegado RA, Fairclough SR, Levin TC, Nichols SA, McDonald K, King N. 2011. Cell differentiation and morphogenesis in the colony-forming choanoflagellate *Salpingoeca rosetta*. *Developmental Biology* **357**:73–82. doi:10.1016/j.ydbio.2011.06.003

de Laat W, Duboule D. 2013. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**:499–506. doi:10.1038/nature12753

de Mendoza A, Suga H, Permanyer J, Irimia M, Ruiz-Trillo I. 2015. Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *eLife* **4**. doi:10.7554/eLife.08904.001

Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**:491–501. doi:10.1038/ng.806

Doitsidou M, Poole RJ, Sarin S, Bigelow H, Hobert O. 2010. *C. elegans* mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. *PLoS ONE* **5**:1–7. doi:10.1371/journal.pone.0015435

Edelman GM. 1986. Cell Adhesion Molecules in the Regulation of Animal Form and Tissue Pattern. *Annual Review of Cell Biology* **2**:81–116. doi:10.1146/annurev.cb.02.110186.000501

El-Battari A. 2006. Autofluorescent Proteins for Monitoring the Intracellular Distribution of Glycosyltransferases. *Methods in Enzymology* **416**:102–120. doi:10.1016/S0076-6879(06)16007-3

Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E, Richter DJ, Russ C, Westbrook MJ, Manning G, Lang BF, Haas B, Nusbaum C, King N. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate *Salpingoeca rosetta*. *Genome Biology* **14**:1–15. doi:10.1186/gb-2013-14-2-r15

Fairclough SR, Dayel MJ, King N. 2010. Multicellular development in a choanoflagellate. *Current Biology* **20**:875–876. doi:10.1016/j.cub.2010.09.014

Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL. 2017. InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research* **45**:D190–D199. doi:10.1093/nar/gkw1107

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research* **44**:D279–D285. doi:10.1093/nar/gkv1344

Gagneux P, Aebi M, Varki A. 2015. Evolution of Glycan Diversity, Essentials of Glycobiology. Cold Spring Harbor Laboratory Press. doi:10.1101/GLYCOBIOLOGY.3E.020

Geijtenbeek TBH, Gringhuis SI. 2009. Signalling through C-type lectin receptors: shaping immune responses. *Nature Reviews Immunology* **9**:465–479. doi:10.1038/nri2569

Gilbert OM, Foster KR, Mehdiabadi NJ, Strassmann JE, Queller DC. 2007. High relatedness maintains multicellular cooperation in a social amoeba by controlling cheater mutants. *Proceedings of the National Academy of Sciences* **104**:8913–8917. doi:10.1073/pnas.0702723104

Gilbert SF. 2000. Developmental biology. Sinauer Associates.

Glockling SL, Marshall WL, Gleason FH. 2013. Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). *Fungal Ecology* **6**:237–247. doi:10.1016/J.FUNECO.2013.03.005

Gow NAR, Latge J-P, Munro CA. 2017. The Fungal Cell Wall: Structure, Biosynthesis, and Function. *Microbiology Spectrum* **5**:267–292. doi:10.1128/microbiolspec.FUNK-0035-2016

Granovsky M, Fata J, Pawling J, Muller WJ, Khokha R, Dennis JW. 2000. Suppression of tumor growth and metastasis in Mgat5-deficient mice. *Nature Medicine* **6**:306–312. doi:10.1038/73163

Grau-Bové X, Torruella G, Donachie S, Suga H, Leonard G, Richards TA, Ruiz-Trillo I. 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife* **6**. doi:10.7554/eLife.26036

Gruenheit N, Parkinson K, Stewart B, Howie JA, Wolf JB, Thompson CRL. 2017. A polychromatic ‘greenbeard’ locus determines patterns of cooperation in a social amoeba. *Nature Communications* **8**. doi:10.1038/ncomms14171

Guo HB, Johnson H, Randolph M, Pierce M. 2009. Regulation of homotypic cell-cell adhesion by branched N-glycosylation of N-cadherin extracellular EC2 and EC3 domains. *Journal of Biological Chemistry* **284**:34986–34997. doi:10.1074/jbc.M109.060806

Haltiwanger RS, Lowe JB. 2004. Role of Glycosylation in Development. *Annual Review of Biochemistry* **73**:491–537. doi:10.1146/annurev.biochem.73.011303.074043

Hehenberger E, Tikhonenkov D V, Kolisko M, Del Campo J, Esaulov AS, Mylnikov AP, Keeling Correspondence PJ. 2017. Novel Predators Reshape Holozoan Phylogeny and Reveal the Presence of a Two-Component Signaling System in the Ancestor of Animals. *Current Biology* **27**. doi:10.1016/j.cub.2017.06.006

Herron MD, Borin JM, Boswell JC, Walker J, Chen I-CK, Knox CA, Boyd M, Rosenzweig F, Ratcliff WC. 2018. De novo origin of multicellularity in response to predation. *bioRxiv* 247361. doi:10.1101/247361

Hertel LA, Bayne CJ, Loker ES. 2002. The symbiont *Capsaspora owczarzaki*, nov. gen. nov. sp., isolated from three strains of the pulmonate snail Biomphalaria glabrata is related to members of the Mesomycetozoea. *International Journal for Parasitology* **32**:1183–1191. doi:10.1016/S0020-7519(02)00066-8

Hirohashi S, Kanai Y. 2003. Cell adhesion system and human cancer morphogenesis. *Cancer Science* **94**:575–581. doi:10.1111/j.1349-7006.2003.tb01485.x

Hirose S, Benabentos R, Ho H-I, Kuspa A, Shaulsky G. 2011. Self-recognition in social amoebae is mediated by allelic pairs of tiger genes. *Science* **333**:467–70. doi:10.1126/science.1203903

Ho DH, Badellino K, Baglia FA, Walsh PN. 1998. A Binding Site for Heparin in the Apple

3 Domain of Factor XI. *Journal of Biological Chemistry* **273**:16382–16390.
Ho JWK, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn K-A, Minoda A, Tolstorukov MY, Appert A, Parker SCJ, Gu T, Kundaje A, Riddle NC, Bishop E, Egelhofer TA, Hu SS, Alekseyenko AA, Rechtsteiner A, Asker D, Belsky JA, Bowman SK, Chen QB, Chen RA-J, Day DS, Dong Y, Dose AC, Duan X, Epstein CB, Ercan S, Feingold EA, Ferrari F, Garrigues JM, Gehlenborg N, Good PJ, Haseley P, He D, Herrmann M, Hoffman MM, Jeffers TE, Kharchenko P V., Kolasinska-Zwierz P, Kotwaliwale C V., Kumar N, Langley SA, Larschan EN, Latorre I, Libbrecht MW, Lin X, Park R, Pazin MJ, Pham HN, Plachetka A, Qin B, Schwartz YB, Shoresh N, Stempor P, Vielle A, Wang C, Whittle CM, Xue H, Kingston RE, Kim JH, Bernstein BE, Dernburg AF, Pirrotta V, Kuroda MI, Noble WS, Tullius TD, Kellis M, MacAlpine DM, Strome S, Elgin SCR, Liu XS, Lieb JD, Ahringer J, Karpen GH, Park PJ. 2014. Comparative analysis of metazoan chromatin organization. *Nature* **512**:449–452. doi:10.1038/nature13415
Iwai H, Masaoka N, Ishii T, Satoh S. 2002. A pectin glucuronyltransferase gene is essential for intercellular attachment in the plant meristem. *Proceedings of the National Academy of Sciences* **99**:16319–16324. doi:10.1073/pnas.252530499
James-Clark H. 1867. IV.—Conclusive proofs of the animality of the ciliate sponges, and of their affinities with the Infusoria flagellata. *The Annals and Magazine of Nautral History* 13–18.
Kadmon G, Kowitz A, Altevogt P, Schachner M. 1990. Functional cooperation between the neural adhesion molecules L1 and N-CAM is carbohydrate dependent. *The Journal of Cell Biology* **110**:209–18. doi:10.1083/JCB.110.1.209
Kakugawa Y, Wada T, Yamaguchi K, Yamanami H, Ouchi K, Sato I, Miyagi T. 2002. Up-regulation of plasma membrane-associated ganglioside sialidase (Neu3) in human colon cancer and its involvement in apoptosis suppression. *Proceedings of the National Academy of Sciences of the United States of America* **99**:10718–23. doi:10.1073/pnas.152597199
Kannagi R, Yin J, Miyazaki K, Izawa M. 2008. Current relevance of incomplete synthesis and neo-synthesis for cancer-associated alteration of carbohydrate determinants—Hakomori's concepts revisited. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1780**:525–531. doi:10.1016/J.BBAGEN.2007.10.007
Kelly LA, Mezulis S, Yates C, Wass M, Sternberg M. 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols* **10**:845–858. doi:10.1038/nprot.2015-053
Kim JH, Lee SR, Li LH, Park HJ, Park JH, Lee KY, Kim MK, Shin BA, Choi SY. 2011. High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice. *PLoS ONE* **6**:1–8. doi:10.1371/journal.pone.0018556
King N. 2004. The unicellular ancestry of animal development. *Developmental Cell* **7**:313–325. doi:10.1016/j.devcel.2004.08.010
King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzow R, Dirks W, Good M, Goodstein D, Lemons D, Li W, Lyons JB, Morris A, Nichols S, Richter DJ, Salamov A, Bork P, Lim WA, Manning G, Miller WT, McGinnis W, Shapiro H, Tjian R, Grigoriev I V., Rokhsar D, Sequencing JGI, Bork P, Lim WA,

Manning G, Miller WT, McGinnis W, Shapiro H, Tjian R, Grigoriev I V., Rokhsar D. 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* **451**:783–8. doi:10.1038/nature06617

Knoll AH. 2011. The Multiple Origins of Complex Multicellularity. *Annual Review of Earth and Planetary Sciences* **39**:217–239. doi:10.1146/annurev.earth.031208.100209

Kopito RR. 1997. ER quality control: the cytoplasmic connection. *Cell* **88**:427–30. doi:10.1016/S0092-8674(00)81881-4

Kornberg RD. 1974. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**:868–71.

Kornberg RD, Lorch Y. 1992. Chromatin Structure and Transcription. *Annual Review of Cell Biology* **8**:563–587. doi:10.1146/annurev.cb.08.110192.003023

Kuzdzal-Fick JJ, Fox SA, Strassman JE, Queller DC. 2011. High Relatedness Is Necessary and Sufficient to Maintain Multicellularity in *Dictyostelium*. *Science* **334**:1548–1551. doi:10.3334/ORNLDAAC/797

Lairson LL, Henrissat B, Davies GJ, Withers SG. 2008. Glycosyltransferases: Structures, Functions, and Mechanisms. *Annual Review of Biochemistry* **77**:521–555. doi:10.1146/annurev.biochem.76.061005.092322

Lang BF, O'Kelly C, Nerad T, Gray MW, Burger G. 2002. The Closest Unicellular Relatives of Animals. *Current Biology* **12**:1773–1778. doi:10.1016/S0960-9822(02)01187-9

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**:R25. doi:10.1186/gb-2009-10-3-r25

Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**:2947–2948. doi:10.1093/bioinformatics/btm404

Larsen ISB, Narimatsu Y, Joshi HJ, Siukstaite L, Harrison OJ, Brasch J, Goodman KM, Hansen L, Shapiro L, Honig B, Vakhrushev SY, Clausen H, Halim A. 2017. Discovery of an O-mannosylation pathway selectively serving cadherins and protocadherins. *Proceedings of the National Academy of Sciences* **114**:201708319. doi:10.1073/pnas.1708319114

Leadbeater BS. 2015. The choanoflagellates: evolution, biology, and ecology, 1st ed. Cambridge: Cambridge University Press.

Leadbeater BSC. 1983. Life-history and ultrastructure of a new marine species of *Proterospongia (Choanoflagellida)*. *Journal of the Marine Biological Association of the United Kingdom* **63**:135. doi:10.1017/S0025315400049857

Leshchiner I, Alexa K, Kelsey P, Adzhubei I, Austin-Tse CA, Cooney JD, Anderson H, King MJ, Stottmann RW, Garnaas MK, Ha S, Drummond IA, Paw BH, North TE, Beier DR, Goessling W, Sunyaev SR. 2012. Mutation mapping and identification by whole-genome sequencing. *Genome Research* **22**:1541–1548. doi:10.1101/gr.135541.111

Levin TC, Greaney AJ, Wetzel L, King N. 2014. The *rosetteless* gene controls development in the choanoflagellate *S. rosetta*. *eLife* **3**:e04070. doi:10.7554/eLife.04070

Levin TC, King N. 2013. Evidence for Sex and Recombination in the Choanoflagellate *Salpingoeca rosetta*. *Current Biology* **23**:2176–2180. doi:10.1016/j.cub.2013.08.061
Levine M. 2010. Transcriptional enhancers in animal development and evolution. *Current biology : CB* **20**:R754-63. doi:10.1016/j.cub.2010.06.070
Levine M, Tjian R. 2003. Transcription regulation and animal diversity. *Nature* **424**:147–51. doi:10.1038/nature01763
Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**:2103–2110. doi:10.1093/bioinformatics/btw152
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. doi:10.1093/bioinformatics/btp324
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. doi:10.1093/bioinformatics/btp352
Lister R, Gregory BD, Ecker JR. 2009. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current Opinion in Plant Biology* **12**:107–118. doi:10.1016/j.pbi.2008.11.004
Lodish H, Berk A, Zipurksy A. 2000. Molecular Cell Biology. New York: W. H. Freeman.
Loqué D, Scheller H V, Pauly M. 2015. Engineering of plant cell walls for enhanced biofuel production. *Current Opinion in Plant Biology* **25**:151–161. doi:10.1016/J.PBI.2015.05.018
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**:550. doi:10.1186/s13059-014-0550-8
Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Geer LY, Bryant SH. 2017. CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* **45**:D200–D203. doi:10.1093/nar/gkw1129
Marshall WL, Berbee ML. 2011. Facing Unknowns: Living Cultures (*Pirum gemmata* gen. nov., sp. nov., and *Abeoforma whisleri*, gen. nov., sp. nov.) from Invertebrate Digestive Tracts Represent an Undescribed Clade within the Unicellular Opisthokont Lineage Ichthyosporea (Mes. *Protist* **162**:33–57. doi:10.1016/J.PROTIS.2010.06.002
Marshall WL, Celio G, McLaughlin DJ, Berbee ML. 2008. Multiple Isolations of a Culturable, Motile Ichthyosporean (Mesomycetozoa, Opisthokonta), *Creolimax fragrantissima* n. gen., n. sp., from Marine Invertebrate Digestive Tracts. *Protist* **159**:415–433. doi:10.1016/J.PROTIS.2008.03.003
McDonald CJ, Sampson J. 1983. The effects of inhibition of protein glycosylation on the aggregation of *Dictyostelium discoideum*. *Development* **78**.
Mellor J. 2005. The Dynamics of Chromatin Remodeling at Promoters. *Molecular Cell* **19**:147–157. doi:10.1016/j.molcel.2005.06.023
Mendoza L, Taylor JW, Ajello L. 2002. The Class Mesomycetozoea: A Heterogeneous Group of Microorganisms at the Animal-Fungal Boundary. *Annual Review of Microbiology* **56**:315–344. doi:10.1146/annurev.micro.56.012302.160950
Müller WE, Zahn RK, Kurelec B, Müller I, Uhlenbruck G, Vaith P. 1979. Aggregation of sponge cells. A novel mechanism of controlled intercellular adhesion, basing on the

interrelation between glycosyltransferases and glycosidases. *Journal of Biological Chemistry* **254**:1280–1287.
Nagy LG, Kovács GM, Krizsán K. 2018. Complex multicellularity in fungi: Evolutionary convergence, single origin, or both? *Biological Reviews*. doi:10.1111/brv.12418
Nedashkovskaya OI, Kim SB, Vancanneyt M, Lysenko AM, Shin DS, Park MS, Lee KH, Jung WJ, Kalinovskaya NI, Mikhailov V V., Bae KS, Swings J. 2006. *Echinicola pacifica* gen. nov., sp. nov., a novel flexibacterium isolated from the sea urchin *Strongylocentrotus intermedius*. *International Journal of Systematic and Evolutionary Microbiology* **56**:953–958. doi:10.1099/ijs.0.64156-0
Nichols SA, Dirks W, Pearse JS, King N. 2006. Early evolution of animal cell signaling and adhesion genes. *Proceedings of the National Academy of Sciences* **103**:12451–12456. doi:10.1073_pnas.0604065103
Nichols SA, Roberts BW, Richter DJ, Fairclough SR, King N. 2012. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. *Proceedings of the National Academy of Sciences* **109**:13046–51. doi:10.1073/pnas.1120685109
Pinho SS, Figueiredo J, Cabral J, Carvalho S, Dourado J, Magalhães A, Gärtner F, Mendonça AM, Isaji T, Gu J, Carneiro F, Seruca R, Taniguchi N, Reis CA. 2013. E-cadherin and adherens-junctions stability in gastric carcinoma: Functional implications of glycosyltransferases involving N-glycan branching biosynthesis, N-acetylglucosaminyltransferases III and V. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1830**:2690–2700. doi:10.1016/J.BBAGEN.2012.10.021
Pinho SS, Reis CA. 2015. Glycosylation in cancer: mechanisms and clinical implications. *Nature Reviews Cancer* **15**. doi:10.1038/nrc3982
Pinho SS, Reis CA, Paredes J, Magalhaes AM, Ferreira AC, Figueiredo J, Xiaogang W, Carneiro F, Gartner F, Seruca R. 2009. The role of N-acetylglucosaminyltransferase III and V in the post-transcriptional modifications of E-cadherin. *Human Molecular Genetics* **18**:2599–2608. doi:10.1093/hmg/ddp194
Pinho SS, Seruca R, Gärtner F, Yamaguchi Y, Gu J, Taniguchi N, Reis CA. 2011. Modulation of E-cadherin function and dysfunction by N-glycosylation. *Cellular and Molecular Life Sciences* **68**:1011–1020. doi:10.1007/s00018-010-0595-0
Pires-daSilva A, Sommer RJ. 2003. The evolution of signalling pathways in animal development. *Nature Reviews Genetics* **4**:39–49. doi:10.1038/nrg977
Pomraning KR, Smith KM, Freitag M. 2011. Bulk segregant analysis followed by high-throughput sequencing reveals the *Neurospora* cell cycle gene, ndc-1, to be allelic with the gene for ornithine decarboxylase, spe-1. *Eukaryotic Cell* **10**:724–733. doi:10.1128/EC.00016-11
Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome research* **26**:342–50. doi:10.1101/gr.193474.115
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**:841–842. doi:10.1093/bioinformatics/btq033
Rademacher TW, Parekh RB, Dwek RA. 1988. Glycobiology.
Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing

data analysis. *Nucleic Acids Research* **44**:W160–W165. doi:10.1093/nar/gkw257
Ratcliff WC, Fankhauser JD, Rogers DW, Greig D, Travisano M. 2015. Origins of multicellular evolvability in snowflake yeast. *Nature Communications* **6**:1–9. doi:10.1038/ncomms7102
Ratcliff WC, Herron MD, Howell K, Pentz JT, Rosenzweig F, Travisano M. 2013. Experimental evolution of an alternating uni- and multicellular life cycle in *Chlamydomonas reinhardtii*. *Nature Communications* **4**:1–7. doi:10.1038/ncomms3742
Ray J, Shinnick T, Lerner R. 1979. A mutation altering the function of a carbohydrate binding protein blocks cell-cell cohesion in developing *Dictyostelium discoideum*. *Nature* **279**:215–221. doi:10.1038/279215a0
Rayner JC, Munro S. 1998. Identification of the MNN2 and MNN5 mannosyltransferases required for forming and extending the mannose branches of the outer chain mannans of *Saccharomyces cerevisiae*. *Journal of Biological Chemistry* **273**:26836–26843. doi:10.1074/jbc.273.41.26836
Richter DJ, Fozouni P, Eisen MB, King N. 2018. Gene family innovation, conservation and loss on the animal stem lineage. *eLife* **7**. doi:10.7554/eLife.34226
Richter DJ, King N. 2013. The Genomic and Cellular Foundations of Animal Origins. *Annual Review of Genetics* **47**:509–537. doi:10.1146/annurev-genet-111212-133456
Rokas A. 2008. The Origins of Multicellularity and the Early History of the Genetic Toolkit For Animal Development. *Annual Review of Genetics* **42**:235–251. doi:10.1146/annurev.genet.42.110807.091513
Rosen SD, Bertozzi CR. 1994. The selectins and their ligands. *Current Opinion in Cell Biology* **6**:663–673.
Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, Lang BF. 2008. A phylogenomic investigation into the origin of Metazoa. *Molecular Biology and Evolution* **25**:664–672. doi:10.1093/molbev/msn006
Ruoslahti E. 1996. Brain extracellular matrix. *Glycobiology* **6**:489–492. doi:https://doi.org/10.1093/glycob/6.5.489
Santorelli LA, Thompson CRL, Villegas E, Svetz J, Dinh C, Parikh A, Sucgang R, Kuspa A, Strassmann JE, Queller DC, Shaulsky G. 2008. Facultative cheater mutants reveal the genetic complexity of cooperation in social amoebae. *Nature* **451**:1107–1110. doi:10.1038/nature06558
Savage DC. 1977. Microbial Ecology of the Gastrointestinal Tract. *Annual Review of Microbiology* **31**:107–133. doi:10.1146/annurev.mi.31.100177.000543
Sawaguchi S, Varshney S, Ogawa M, Sakaidani Y, Yagi H, Takeshita K, Murohara T, Kato K, Sundaram S, Stanley P, Okajima T. 2017. O-GlcNAc on NOTCH1 EGF repeats regulates ligand-induced Notch signaling and vascular development in mammals. *eLife* **6**. doi:10.7554/eLife.24419
Schalchian-Tabrizi K, Minge MA, Espelund M, Orr R, Ruden T, Jakobsen KS, Cavalier-Smith T. 2008. Multigene phylogeny of Choanozoa and the origin of animals. *PLoS ONE* **3**. doi:10.1371/journal.pone.0002098
Schietinger A, Philip M, Yoshida BA, Azadi P, Liu H, Meredith SC, Schreiber H. 2006. A mutant chaperone converts a wild-type protein into a tumor-specific antigen. *Science* **314**:304–308. doi:10.1126/science.1130716

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jørgensen JE, Weigel D, Andersen SU. 2009. SHOREmap: Simultaneous mapping and mutation identification by deep sequencing. *Nature Methods* **6**:550–551. doi:10.1038/nmeth0809-550

Sebé-Pedrós A, Ballaré C, Parra-Acero H, Gó Mez-Skarmeta L, Croce L Di, Aki Ruiz-Trillo I, Chiva C, Tena JJ, Sabidó E, Gó Mez-Skarmeta JL. 2016. The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal Multicellularity. *Cell* **165**:1224–1237. doi:10.1016/j.cell.2016.03.034

Sebé-Pedrós A, Degnan BM, Ruiz-Trillo I. 2017. The origin of Metazoa: a unicellular perspective. *Nature Reviews Genetics* **18**:498–512. doi:10.1038/nrg.2017.21

Sebé-Pedrós A, Irimia M, del Campo J, Parra-Acero H, Russ C, Nusbaum C, Blencowe BJ, Ruiz-Trillo I. 2013. Regulated aggregative multicellularity in a close unicellular relative of metazoa. *eLife* **2**. doi:10.7554/eLife.01287

Shinnick TM, Lerner RA. 1980. The cbpA gene: role of the 26,000-dalton carbohydrate-binding protein in intercellular cohesion of developing *Dictyostelium discoideum* cells. *Proceedings of the National Academy of Sciences of the United States of America* **77**:4788–92. doi:10.1073/PNAS.77.8.4788

Springer WR, Cooper DNW, Barondes SH. 1984. Discoidin I is implicated in cell-substratum attachment and ordered cell migration of *Dictyostelium discoideum* and resembles fibronectin. *Cell* **39**:557–564. doi:10.1016/0092-8674(84)90462-8

Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, Mitros T, Richards GS, Conaco C, Dacre M, Hellsten U, Larroux C, Putnam NH, Stanke M, Adamska M, Darling A, Degnan SM, Oakley TH, Plachetzki DC, Zhai Y, Adamski M, Calcino A, Cummins SF, Goodstein DM, Harris C, Jackson DJ, Leys SP, Shu S, Woodcroft BJ, Vervoort M, Kosik KS, Manning G, Degnan BM, Rokhsar DS. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466**:720–726. doi:10.1038/nature09201

Strahl-Bolsinger S, Gentzsch M, Tanner W. 1999. Protein O-mannosylation. *Biochimica et Biophysica Acta* **1426**:297–307. doi:10.1016/S0304-4165(98)00131-7

Strassmann JE, Zhu Y, Queller DC. 2000. Altruism and social cheating in the social amoeba *Dictyostelium discoideum*. *Nature* **408**:965–967. doi:10.1038/35050087

Stratford M. 1992. Yeast Flocculation : Receptor Definition by *mnn* Mutants and Concanavalin A. *Yeast* **8**:635–645. doi:10.1002/yea.320080807

Suga H, Ruiz-Trillo I. 2013. Development of ichthyosporeans sheds light on the origin of metazoan multicellularity. *Developmental Biology* **377**:284–292. doi:10.1016/J.YDBIO.2013.01.009

Švajger U, Anderluh M, Jeras M, Obermajer N. 2010. C-type lectin DC-SIGN: An adhesion, signalling and antigen-uptake molecule that guides dendritic cells in immunity. *Cellular Signalling* **22**:1397–1405. doi:10.1016/J.CELLSIG.2010.03.018

Szathmary E, Smith JM. 1995. The major evolutionary transitions. *Nature* **374**:227–232. doi:10.2307/2410462

Takeichi M. 1988. The cadherins: cell-cell adhesion molecules controlling animal morphogenesis. *Development* **102**.

Takeuchi H, Haltiwanger RS. 2014. Significance of glycosylation in Notch signaling. *Biochemical and Biophysical Research Communications* **453**:235–242. doi:10.1016/J.BBRC.2014.05.115

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**:178–192. doi:10.1093/bib/bbs017

Tordai H, Bányai L, Patthy L. 1999. The PAN module: The N-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins. *FEBS Letters* **461**:63–67. doi:10.1016/S0014-5793(99)01416-7

Torruella G, de Mendoza A, Grau-Bové X, Antó M, Chaplin MA, del Campo J, Eme L, Pérez-Cordón G, Whipps CM, Nichols KM, Paley R, Roger AJ, Sitjà-Bobadilla A, Donachie S, Ruiz-Trillo I. 2015. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Current Biology* **25**:2404–2410. doi:10.1016/J.CUB.2015.07.053

Trigos AS, Pearson RB, Papenfuss AT, Goode DL. 2018. How the evolution of multicellularity set the stage for cancer. *British Journal of Cancer* **118**:145–152. doi:10.1038/bjc.2017.398

Tu L, Banfield DK. 2010. Localization of Golgi-resident glycosyltransferases. *Cellular and Molecular Life Sciences* **67**:29–41. doi:10.1007/s00018-009-0126-z

Valentine JW. 2006. On the Origin of Phyla. University of Chicago Press.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* **27**:737–746. doi:10.1101/gr.214270.116

Vasta GR, Amzel LM, Bianchet MA, Cammarata M, Feng C, Saito K. 2017. F-Type Lectins: A Highly Diversified Family of Fucose-Binding Proteins with a Unique Sequence Motif and Structural Fold, Involved in Self/Non-Self-Recognition. *Frontiers in Immunology* **8**:1648. doi:10.3389/fimmu.2017.01648

Voz ML, Coppieters W, Manfroid I, Baudhuin A, von Berg V, Charlier C, Meyer D, Driever W, Martial JA, Peers B. 2012. Fast homozygosity mapping and identification of a zebrafish enu-induced mutation by whole-genome sequencing. *PLoS ONE* **7**:1–10. doi:10.1371/journal.pone.0034671

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva E V, Zdobnov EM. 2018. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution* **35**:543–548. doi:10.1093/molbev/msx319

Wenger JW, Schwartz K, Sherlock G. 2010. Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from *Saccharomyces cerevisiae*. *PLoS Genetics* **6**:18. doi:10.1371/journal.pgen.1000942

Wiggins C a, Munro S. 1998. Activity of the yeast MNN1 alpha-1,3-mannosyltransferase requires a motif conserved in many other families of glycosyltransferases. *Proceedings of the National Academy of Sciences of the United States of America* **95**:7945–7950. doi:10.1073/pnas.95.14.7945

Wood WB. 1988. The Nematode Caenorhabditis elegens. CSH.

Woznica A, Cantley AM, Beemelmanns C, Freinkman E, Clardy J, King N. 2016. Bacterial lipids activate, synergize, and inhibit a developmental switch in choanoflagellates. *Proceedings of the National Academy of Sciences* **113**:7894–7899. doi:10.1073/pnas.1605015113

Woznica A, Gerdt JP, Hulett RE, Clardy J, King N. 2017. Mating in the Closest Living

Relatives of Animals Is Induced by a Bacterial Chondroitinase. *Cell* **170**:1175–1183.e11. doi:10.1016/j.cell.2017.08.005
Zelensky AN, Gready JE. 2005. The C-type lectin-like domain superfamily. *FEBS Journal* **272**:6179–6217. doi:10.1111/j.1742-4658.2005.05031.x
Zhang L, Zhang Y, Ten Hagen KG. 2008. A mucin-type O-glycosyltransferase modulates cell adhesion during *Drosophila* development. *Journal of Biological Chemistry* **283**:34076–34086. doi:10.1074/jbc.M804267200
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**:R137. doi:10.1186/gb-2008-9-9-r137
Zhao Y, Sato Y, Isaji T, Fukuda T, Matsumoto A, Miyoshi E, Gu J, Taniguchi N. 2008. Branched N-glycans regulate the biological functions of integrins and cadherins. *FEBS Journal* **275**:1939–1948. doi:10.1111/j.1742-4658.2008.06346.x
Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2017. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* **430**:2237–2243. doi:10.1016/j.jmb.2017.12.007