A Synthetic Yeast Model for Differentiation and Division of Labor

A dissertation presented

by

Mary Elizabeth Wahl

to

the Department of Molecular and Cellular Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biochemistry

Harvard University

Cambridge, Massachusetts

April 2014

Dissertation Advisor: Professor Andrew Wood Murray Mary Elizabeth Wahl

A Synthetic Yeast Model for Differentiation and Division of Labor

Abstract

To maintain high average fitness, populations must effect selection against the deleterious mutations that continuously arise *de novo*. Theoretical models of mutation-selection balance predict that the maximum tolerable mutation rate is much lower for organisms growing in colonies than for those in well-mixed liquid media due to drift imposed by competition for position along the growing colony front. Simplifying assumptions made in these models, including the irreversibility and fixed fitness cost of mutations, do not strictly hold in extant species. To explore the applicability of these models in natural contexts, we have constructed a yeast strain which undergoes recombinase-mediated irreversible gene excision at a single locus with tunable fitness cost, but also possesses the random genomic mutation profile characteristic of yeast. We find that several theoretical predictions hold for our strain, including the dependence of maximum tolerable mutation rate on growth condition and selective coefficient. These results constitute the first direct biological test of mutation-selection balance theory.

The gene excision system in our yeast strain also mimics the irreversible conversion seen during terminal cellular differentiation, a key feature of development in many multicellular organisms. Terminal differentiation and multicellularity have distinct biological underpinnings, and each has evolved independently more than one dozen times in the eukaryotes alone, yet the two traits are very often found together in extant species. In the few lineages where evolutionary intermediates appear to have persisted (e.g., the volvocine algae and cyanobacteria), it appears that multicellularity evolved first. To investigate the importance of evolutionary order of these two traits, we created yeast strains that are multicellular and/or differentiate in such a way that the two cell types produced can cooperate to support

culture growth. We find that unicellular, differentiating yeast strains are highly susceptible to invasion by non-differentiating revertants. This suggests that such populations are inherently unstable, and thus evolution of multicellularity is likely to precede evolution of differentiation.

Contents

List of Tables

List of Figures

Acknowledgements

I have had the great fortune to be supported, both scientifically and emotionally, by many mentors and friends throughout the years. In addition to mounting debts of gratitude, most of these people are owed an apology.

We begin of course with Andrew Murray, of whom too much good cannot be said. We met through his graduate seminar in genetics, but I had some advance warning from a Caltech colleague, who summed Andrew up nicely as "a real hoot." Andrew's talent for instruction – and seamless incorporation of colorful metaphors – were immediately evident in the classroom and still serve as inspiration for my own teaching. Over the semester, I came to suspect that a guy like Andrew couldn't possibly be doing research as boring as his outdated faculty profile suggested.

Suspicion became conviction during our discussion of potential rotation projects. Andrew had prepared for this meeting by filling a whiteboard with sketches of fanciful ideas such as the experimental evolution of circadian rhythms, multicellularity, and meiosis. After his spiel, I choked down my gut reaction – that none of this would ever work[1] – and asked instead a naïve question: which hypotheses were these experiments designed to test?

Andrew closed his eyes, cradled his temples, sighed in exasperation, and after a dramatic pause, informed me that this was exactly the sort of question that a benighted, NIH-funded developmental biologist could be expected to ask. A true pioneer explores the frontiers of knowledge where, for want of facts to grapple with, hypotheses defy generation, let alone testing. Young scientists, he said, are at a unique advantage because only at this stage in their careers can they probe the world around them without accountability to funding agencies. Remarkable as it sounds, no one else has proselytized for discovery science in my presence before or since. Andrew's commitment to audacious, innovative research has been truly inspiring and one of my favorite aspects of life in his lab.

[1] As of this writing, I have been proven wrong in at least two cases.

Andrew demonstrated boundless patience and delivered profound advice in moments of soul-searching and personal crisis. He tolerated mercurial tempers, irrationality, "side projects," direct criticism, and impudent ribbing with exemplary grace and poise. Andrew whole-heartedly supported my teaching career while simultaneously nurturing my development as a researcher[2]. It has been an honor and a pleasure to work with this brilliant scientist and close friend.

Eternal thanks also to my freshman adviser, Eric H. Davidson, for refusing to sign my study card electronically and thus instigating my first visit to a biology lab. Eric ushered me into a research group immediately; years later, he would muster enough courage to accept me into his own. His brutal but refreshing honesty on all things – study abroad, stem cells, potential mentors – has saved me from myself a dozen times over. I regret only that I did not take more of his suggestions.

Other notable research advisers included David J. Anderson, who sparked my interest in Monty Python and developmental biology (unfortunately, I was in his behavioral neuroscience subgroup at the time); Rich Losick, a dedicated instructor deeply committed to the intellectual development of his charges; and the piercingly-insightful, encyclopedically-knowledgeable, inexhaustible Cassandra Extavour.

I am deeply grateful to my many committee members, particularly those who were forced to read this thesis. Sharad Ramanathan stuck with it through the best and worst of times, offering support not merely for my research, but with my teaching career as well. Michael Desai provided support and helpful discussions on current and future research, injected much-needed doses of realism into teaching plans, and shared many memorable lunchtime conversations with humor ranging from the morose to the scatological. I've been fortunate to benefit from both insightful instruction and scientific collaborations with David Nelson as well as his students Jayson Paulose and Max Lavrentovich. Rich Losick's comments and suggestions during the qualifying exam helped me to think abstractly about the

[2]The reader will appreciate how extraordinary it is, in our time, to be encouraged in the former pursuit without being written off entirely for the latter.

realized and potential contributions of experimental evolution to biology at large: Rich will be pleased to learn that I never performed any myself. Anne Pringle helped me draw clear, restrained comparisons between my synthetic system and the varied life strategies of natural organisms, while Chris Marx encouraged more liberal analogies: it is a travesty that Harvard has lost them both. The unfortunate timing of Alex Schier's latest sabbatical prevented him from staying on as my committee chair, but I greatly appreciated his thoughtful experimental suggestions and the developmental biology perspective he brought to this work.

I have had the privilege of learning the teaching trade from many wise and diligent instructors. From my time at Caltech, I thank Henry Lester for hiring me at eighteen for my first teaching position, despite severe underqualification, as well as Rob Phillips and Pamela Björkman, who tolerated problem set questions on subjects like pregnancy tests, HPV, and HIV. Rich Losick, Briana Burton, Tom Torello, and Mary Ellen Wiltrout baptized me into Harvard's tradition of discovery-based labs and taught me to handle academic dishonesty cases with measured indignation. With Andrew Murray and Cassandra Extavour I practiced leading paper discussions by the Socratic method, learned to bake, and discovered an excellent Mormon chalk company. Michael Brenner, Pia Sörensen, and Dave Weitz deserve an "A+ with garlands" for their consummate devotion to teaching and the science of cooking: who else can claim the distinction of staying up past midnight to discuss the viscoelastic differences between Easy Mac and Kraft Dinner? Thanks once more to Sharad Ramanathan and Michael Desai, for invigorating the training grant's math requirement with exciting applications of Bayesian inference and for allowing me to learn the subject as we went along, (almost always) staying one lesson ahead of the kids.

It is plain that I would not have survived this whole affair without my labmates, whose good humor was in evidence from my first step into the Murray lab. On my way to set down my bag before safety training, I was nearly concussed by a flying, oinking stuffed pig launched from a slingshot by the unassuming Derek Lau. At lunch I entered into what would be the first of many discussions led by Erik Hom on the relative intelligences of ethnic groups.

The wholesome warmth of Natalie Nannas (née Funk) and Lori Huberman, then my baymates, served as foil to the crass badinage I enjoyed with Beverly Neugeboren, often over coffee in the morning. The dynamic duo of Gregg Wildenberg and John Koschwanez contributed good espresso and much-needed virility in a lab overrun by women. Quincey Justman and Joana de Sá taught, through their example, a philosophical devotion to scientific ideals; together with Erik Hom and Andrew, they helped me to craft a credible thesis proposal (though ultimately none of that proposed work would appear herein). Nilay Karahan and Dario Cabrera taught me the truth of the old mantra that research is not for everyone, regardless of competence and drive. Scott Schuyler and I overlapped by about a week: he let me borrow his lucky hat once, and (unwittingly) shoulders the blame for the singed shelf at my bench. I profoundly thank Linda Kefalas for stocking snacks, remembering my birthday, booking rooms, organizing meetings, and commiserating in hatred of people (never individual persons). To Bodo Stern, a former lab member and ongoing colleague, I am indebted for free trips to Yale and the French Riviera as well as critiquing my practice talks and bestowing honorary membership in the Physics of Living Systems fraternity.

Later additions brought fresh flavor as old friends departed. Edel demonstrated excellence in mentorship and that an embrace of femininity and good taste is not fundamentally in conflict with science.I will never forget Max Lavrentovich's stoic calm as we fishtailed through the 2013 blizzard, nor his patient tolerance of whoppers such as "Ising model simulation with yeast." Board games, applied math seminars, and teasing Phoebe were among the favorite pastimes shared with the Germanic and biophysical trio of Wolfram Möbius, Melanie Müller, and Liedewij Laan. Liedewij in particular deserves a round for combining classic Dutch honesty with organic optimism to deliver very effective pep talks and constructive criticism. Nichole Collins, who remarkably joined the lab despite her rotation mentor's throes of depression, has also been an exceptional baymate, contributing the refreshing perspective of someone who – at least at one point – had a life. Our disco divas Phoebe Hsieh and Miguel Coelho introduced much-needed hardy-partying. Laura

Bagamery, Patrick Stoddard, and Sri Srikant: you were half-students, but you had my whole heart.

Friends outside of lab made life tolerable. Chris "Gonzo" Gonzales and Sarah Stidham taught the art of the conversation starter through taxidermy auctions, terraformed islands, beautiful but stupid pets, and high-stakes bets. Will Steinhardt deserves better than to have all his hair fall out: perhaps he will take some solace in having baptized a new Eagles fan. Russell McClellan, Eric Johlin, and Tamas Szalay threw epic parties with home-distilled liquors and fine punches. Arturo Pizano, Guarav "Gino" Giri, Marie Giron, Daniel Tofan, and Phil Muñoz kept the memory of Ricketts House alive. Ben Ranish, Jon Sadowski, and Becky Russell introduced me to good beer. No one can cuss better than Jamie Webster. Fellow MCOers Marina, Scott, Fred, Katherine, Vu, McGee, John, Naked John, Andrew, André, Chewie, Julio, Nick, Eddie, Amanda, Ania, Sarah, Jiang, . . . every time I see them, I smile.

Graduate fellowships from the National Science Foundation and the Department of Defense granted me some additional freedom to pursue teaching beyond the department's expectation. I'm certain this defied their intent but the money was no less appreciated because of it. Dan Gottschling and Derek Lindstrom shared a construct without which this project would have been nigh on impossible. Mike Lawrence put up with a tremendous hassle and always offered a hand. Susan Foster, in addition to many other forms of support, scored me an unbelievably sweet office.

During grad school I developed an interest in genomics and genealogy: I now see that the traits I inherited have greatly shaped my approach to science. Colleagues will recognize the influence of my grandparents: a horse trader, chemistry professor, no-nonsense homemaker, and clairvoyant-*cum*-Wiccan. I share my mother's healthy (?) pessimism and my father's compulsion to stick a finger in every pot: I hope also that I will one day exhibit their kindness, wisdom, and nurturing ability . . . but these things sometimes skip a generation.

Known widely for his keen sense of humor, love, compassion, and good company, Nick Hutzler is also a true gentleman and a scholar. I look forward to being a better partner to him once this is over.

To Nick.

Chapter 1

Evolution of multicellularity and differentiation

"Theirs not to reason why; theirs but to do and die."

- Alfred, Lord Tennyson

1.1 Introduction

Most authorities would agree to define cellular differentiation as a prolonged change in a cell's gene expression state. The appropriate timescale for distinguishing differentiation from transient responses like environmental sensing remains open for debate, with some authors in frustration paraphrasing Justice Potter Stewart's famous quote:

> I shall not today attempt further to define what I understand to be embraced within that description, and perhaps I could never succeed in intelligibly doing so. But I know it when I see it [97].

To avoid ambiguity, we will take the hardline definition that, barring scientific intervention, cellular differentiation is a permanent change and therefore that a differentiated cell cannot propagate the species. While harsh, this definition nonetheless admits dozens of independently-evolved forms of differentiation; it also highlights the lost reproductive potential associated with the process, and suggests that differentiated cells must perform functions beneficial to an organism's germline in order to compensate for the costs of their production. In this chapter, we will explore several cases in which differentiation has evolved in nature to better understand the circumstances under which this trait can appear.

The most obvious pattern we will encounter is that differentiating species are unfailingly multicellular. Here again, formal definitons for multicellularity are as varied as the organismal lifestyles under consideration. Some authors would include any prolonged association between cells, including purely accidental cases such as the formation of colonies on agar plates. Others require that multicellularity reflect "complexity," which in turn is defined so as to blackball the author's choice of life forms: colonies, biofilms, filaments, multinucleate syncytia, non-clonal cell aggregates, or even all species that lack cellular differentiation, regardless of their morphology. We will tend toward the more inclusive definition, though we will require some form of permanent physical attachment. As a counterbalance to this liberalism, we will note levels of complexity that are more exceptional for their rarity.

Our choice is motivated by considering the unifying traits of multicellular organisms as we have defined them. In all cases, a transition of the level of "individuality" has occurred from the single cell or nucleus to the level of the organism [19, 88]: multiple copies of the genome (which may not be perfectly identical) cooperate for the benefit of the group. The pressures and mechanisms that maintain this state, despite the potential for each genome to act in its own best interests, are fundamentally similar regardless of whether an organism is multinucleate or truly multi-celled, or whether its genomes are brought together through aggregation or simply remain stuck together clonally.

Multicellularity, too, has evolved many times, and the most prolific multicellular clades are those that also display differentiation. It is therefore natural to ask why and how these two traits co-evolve. Is there a small number of fundamentally-analogous mechanisms that can effect each trait? Does each trait tend to evolve under particular ecological circumstances? Why are these traits so often found together if their biological bases are separate? What advantages can multicellularity offer in the absence of differentiation, and vice versa?

These questions and many others have historically been addressed with a comparative approach. The process begins with untangling the timing of a trait's appearance on a phylogenetic tree: it is often difficult to discern similarity caused by convergent evolution from true homology, particularly when timescales are long and traits evolve frequently (both of which are true for multicellularity and differentiation). The phylogeny itself may be erroneous, causing misidentification of a trait's origin by maximum parsimony approaches. Once the trait's origin is guessed, speculations are made about the ancestor in which the trait appeared based on all living descendants and sister groups. Inferences about a common ancestor which lived hundreds of millions or billions of years ago – including the traits it possessed, the ecological pressures it experienced, and its potential means of adaptation – are, by their nature, speculative at best. The examples that follow are chosen to showcase hypotheses about the costs and benefits of multicellularity and differentiation: though improved phylogenies and comparative analyses may negate the conclusions drawn for specific

groups, the underlying reasoning may remain applicable in other cases. We begin with the eukaryotes, where acquisitions of multicellularity and differentiation have been relatively recent (reducing the required level of speculation) and better studied, then proceed to the other domains of life.

1.2 Bikonta

As of this publication, prevailing phylogenetic wisdom divides the eukaryotes between the unikonts, whose common ancestor likely had a single flagellum, and the bikonts – including multicellular groups like the green plants, red algae, volvocine algae, brown algae, and water molds – whose common ancestor had two flagella [136][1]. We highlight four examples from this group with very different multicellular lifestyles: the volvocine algae, the red algae, tubular forms of diatoms and labrynthulids, and branched, acellular stalked ciliates.

Volvocine algae

The volvocine algae (members of the division Chlorophyta) include multicellular species which form small or hollow spherical "colonies", as well as their closest unicellular relatives [56, 70] (Figure 1.2). When displayed in the proper order, it is easy to see a progression in complexity from the unicellular state through four-celled, undifferentiated groups to colonies with thousands of cells and a differentiated soma: phylogenetic analysis confirms that this suggestive arrangment reflects the order of evolution of these traits [103]. The divergence time of the multicellular volvocine algae from their nearest unicellular relatives, previously thought to be within the last 75 million years [113], has been recently revised to 234 ± 25 million years ago on the basis of multigene phylogenies [56]. Nevertheless, the evo-

[1] But see He et al. [54], which suggests that the former bikont clade Excavata (including species such as *Euglena* and the acrasid slime molds) are basal to all other eukaryotes. This placement remains controversial because the excavates are parasites and symbionts with highly-divergent sequences that confuse phylogenetic rooting through long-branch attraction [38].

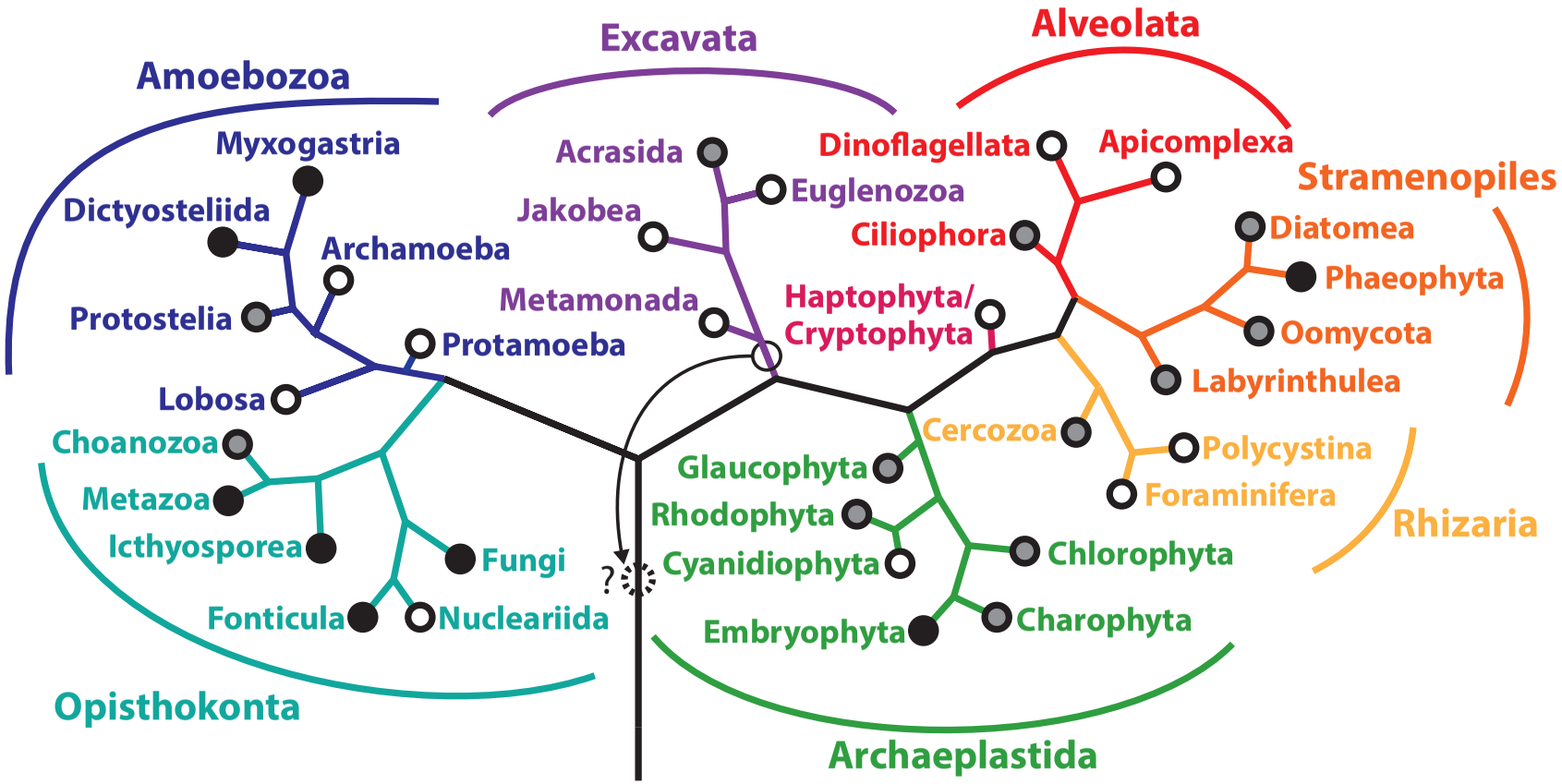

Figure 1.1: Phylogram of the eukaryotes

White circles indicate unicellular taxa; gray circles indicate the presence of derived multicellular/multinucleate species in the clade; black circles indicate that the common ancestor of the clade was multicellular/multinucleate. The position of the excavates remains controversial. Major clades of interest include the animals (Metazoa), cellular (Dictyosteliida) and plasmodial (Myxogastria) amoeboid slime molds, red algae (Rhodophyta), volvocine algae (subset of Chlorophyta), land plants (Embryophyta), and brown algae (Phaeophyta). Phylogram adapted from Baldauf [5], incorporating more recent findings [24, 47, 54, 68].

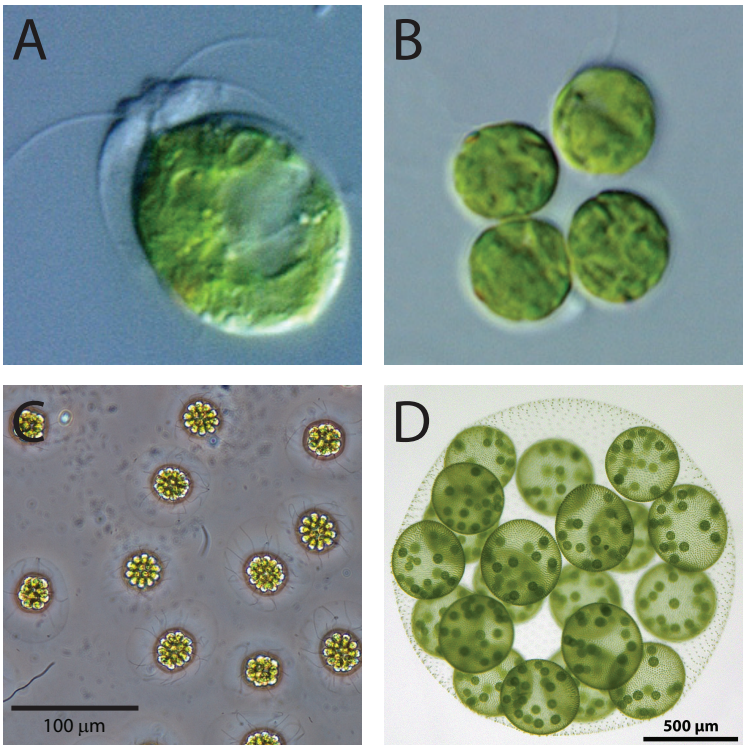

Figure 1.2: Morphology of volvocine algae

(A) The unicellular green alga *Haematococcus* sp., close relative of the model green alga *Chlamydomonas reinhardtii*. Photo courtesy of Peter Siver [130]. (B) Species of the primitively-multicellular *Gonium* phylum have 4-16 cells. Photo courtesy of Peter Siver [130]. (C) A colony of the 32-celled, undifferentiated species *Eudorina elegans* Ehrenberg. Photo courtesy of Karl Bruun, via Algaebase [48]. (D) The multicellular, differentiated species *Volvox carteri*: juvenile colonies developing inside the parent contain large gonidial (germline) cells interspersed between smaller somatic cells. Photo courtesy of Ichiro Nishii.

lution of multicellularity and differentiation in the volvocine alge remains one of the most recent cases known.

Evolution of multicellularity in this group may have beenfacilitated by a pre-existing feature of green algal division called palintomy: newborn cells grow manifold in volume, then undergo a rapid series of nuclear replications and cell divisions (reviewed in [10]). The daughter cells thus produced are initially confined to the cell wall of their largest parent, forming a group called a coenobium from which they emerge by secreting enzymes that degrade their container. It has been hypothesized that primitive multicellularity in the volvocine algae emerged through the failure of daughter cells to separate following palintomy [70]. Indeed, colonies of species in the most basal clades (such as *Tetrabaena socialis*, not pictured) are held together by the coenobial sac, sometimes with additional structural support from cytoplasmic bridges; each cell then produces an independent cell wall [4, 101]. More morphologically "advanced" forms, such as the 4 to 16-celled *Gonium pectorale*, se-

crete a unified extracellular matrix around the colony [102, 104], but palintomy persists as a developmental feature even in the most advanced forms.

The volvocine algae are not the only group believed to have evolved multicellularity through failure to separate following palintomy. The Chlorococcales, a distinct clade of chlorophyte algae, also form groups of 2^n cells through multiple divisions within a parent cell's wall: their shapes are even more varied, ranging from tiled rows of lozenge-like cells of *Scenedesmus dimorphis* to reticulated webs in *Hydrodictyon reticulatum* to stellate spheres in *Pediastrum duplex* [12].

Multicellular, undifferentiated species have persisted in this clade for over 200 million years, suggesting inherent benefits to this lifestyle. One demonstrated advantage of mulicellularity among the volvocine algae is the avoidance of predators only capable of consuming smaller prey [7]. Constraints on surface-to-volume ratio for efficient nutrient uptake prevent the algae from simply increasing their cell size indefinitely [52]: multicellularity is therefore an attractive option for avoiding predators. This selection pressure can be so acute, and the mutations needed to acquire multicellularity so accessible, that green algae can be experimentally evolved to form multicellular clumps in a few dozen generations simply by co-culturing them with a predator [13].

Another potential advantage of undifferentiated multicellularity in the volvocine algae is improved phototaxis. During the day, many chlorophyte algae must swim against gravity for improved access to the light they require for photosynthesis [69]; sinking (or active vertical migration) of up to 20 meters [134] at night can also be beneficial, since it permits access to nutrients like phosphate which are limiting near the surface [64]. The spherical topology of volvocine algae colonies and the orientation of eyespots and flagella permit positive phototaxis by allowing each cell in the colony to sense light from a particular direction and, when light is detected, to decrease flagellar beating frequency [33, 57] and/or reorieint flagella laterally [86, 133]; since cells on the "shady" side of the colony continue swimming normally, the colony tends to move toward the light. By contrast, single-celled algae like

Chlamydomonas tumble to reorient relative to the light [122], a less effective strategy in open water. Surface currents created by flagellar beating around colonies may also improve uptake of nutrients by diffusion [129].

Chlorophyte algae typically have two flagella and, in interphase, two microtubule organizing complexes (MTOCs) that orchestrate their motion [74]. In species that lack cell walls, the MTOCs can remain attached to the flagella during mitosis as they each migrate to opposite ends of the dividing cell[40]. The common ancestor of the volvocine algae, however, had thick cell walls that prevented the movement of flagella along the cell membrane: in these species, the MTOCs must detach from the flagella (which are then typically shed or resorbed) before migrating to each daughter cell[74]. During this time, the flagella are not beating and the algae therefore fall vertically through the water column [114], away from the light they require for photosynthesis [74].

In *Volvox carteri* (pictured), cellular differentiation produces flagellated somatic cells and large, rapidly-dividing germ cells [69]. Since both cell types are present in the same colony, there is no need to sacrifice motion for continued cell division[2]. The importance of this role for the soma is highlighted by the somatic cells' position: either at the anterior end of the colony (despite the superficial radial symmetry, some species have an anterior-posterior axis) or completely surrounding it with the germline dividing in the interior[3].

Cellular differentiation in *Volvox* begins with the asymmetric divisions of an enlarged germline cell into daughters which, due to incomplete cytokinesis, form a syncytium connected by cytoplasmic bridges. These bridges provide structural support to the colony until it can secrete an extracellular matrix of its own, at which point the cytoplasmic bridges disappear and cells begin to express genes differentially. For reasons yet unknown, but ap-

[2]In other clades, similar evolutionary constraints have been resolved by increasing the number of basal bodies per cell or by permitting the parental flagella to continue beating after MTOCs detach[58].

[3]This is not the only rationaliztion proposed for the position of germline cells within the spherical cavity of *Volvox*. Phosphate, which is limiting for growth in some natural environmental conditions[106, 115], is actively transported into the cavity, suggesting that internalization of the germline may ensure its access to essential nutrients[7].

parently dependent primarily on cell volume [71], smaller daughter cells begin to produce a transcriptional repressor that inhibits expression of proteins required for full chloroplast function: these cells are therefore incapable of further growth and division, and therefore become somatic cells [93]. Since this system relies on cell size to specify cell type, it may not be readily scalable, which may explain why additional cell types have not yet evolved in *Volvox*. We will later argue by contrast that transcription factor networks and intercellular signalling methods for cellular differentiation are more easily extensible due to their modularity, which may partly explain why the animals and plants that employ them have each evolved many cell types while the volvocine algae have only two.

Rhodophyta

In contrast to the volvocine algae, red algae (Rhodophyta) lack both flagella and centrioles [29]. Multicellular forms of red algae include coralline algae and seaweeds which can grow to be over a meter long. Differentiated, multicellular forms have existed in this clade for at least 1.2 billion years [20], yet a range of morphologies representing likely evolutionarily-intermediate forms (including unicellular species and undifferentiated filaments) persists today [29]. While unicellular forms are predominantly planktonic, adhesion to a substrate is necessary for spore germination for many multicellular red algae [29].

This finding has led to the speculation that primitive multicellularity may have evolved in a sessile red alga, which would then have received preferential access to light (and additional room for expansion) by growing vertically through the water column [18]. In support of this hypothesis, apical growth to produce long, narrow filaments is one of the earliest adaptations among the multicellular red algae [44], and "holdfasts" which secure one end of the filament to a substrate (while other cells reproduce and undergo photosynthesis) are the most primitive forms of differentiation found in the red algae [29]. As in volvocine algae, intercellular cohesion is mediated by cytoplasmic bridges (called "pit connections") as

well as shared cell walls. In some species, exchange of nutrients through these channels is blocked by "pit plugs" to effectively restore full cellularization [29].

A notable feature of red algal filaments is that, because they are attached at one end to a substrate, they have a trivial polarity along their axis. Some filamentous species in other groups, including the green alga *Ulothrix*, also have a simple holdfast that differentiates one end of the filament from the other [12]. Not all filamentous forms have this distinction: for example, in the cyanobacteria we will discuss later in this chapter, the two growing tips of a filament are indistinguishable.

Diatoms and ciliates

Not all multicellular bikonts use cytoplasmic bridges and a shared cell wall for multicellular adhesion. Diatoms produce elaborate, siliceous cell walls with two interlocking halves that must separate with each division to make room for new growth: this life strategy would presumably interfere with intercellular adhesion through the cell wall. According to Bonner [12], a single multicellular species of diatoms is known: *Navicula grevillei*, a colonial form that secretes a tubular network through which individual cells can migrate back and forth. A similar strategy of cellular migration within secreted tubules is taken by the amoeboid bikont genus *Labyrinthula* which lacks cell walls [12].

Ciliates have multiple nuclei (and thus trivially fall under our definition of multicellularity) but lack cell walls, and thus cannot evolve multicellularity in the manner described for green and red algae. However, some ciliates (including those of genus *Zoothmnium*) form visually-stunning, snowflake-shaped clonal colonies through the secretion of a shared, branching stalk. This type of adhesion has appeared in other lineages that lack cell walls (including some thecate choanoflagellates) and even in clades that do possess walls, such as green algae of genus *Dictyosphaerium*: it is thus evidently a common strategy for the evolution of multicellularity.

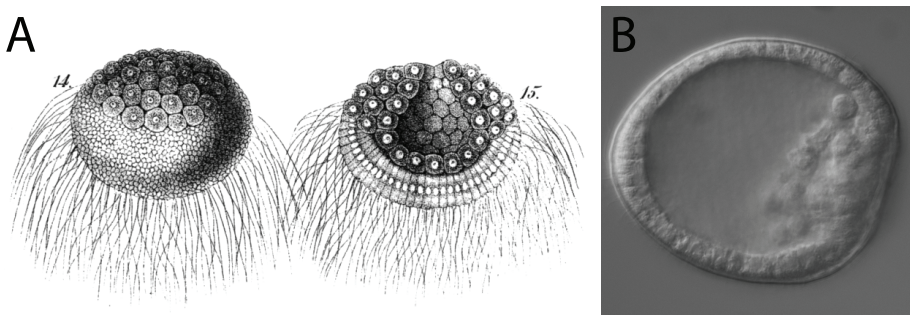

Figure 1.3: Early differentiation defines flagellated and dividing cell types in metazoan embryos

(A) Haeckel's [49] depictions of the *Sycyssa huxleyi* sponge gastrula (reproduced from Leys et al. [80]). Flagellated cells at the anterior supports motion of the embryo as unflagellated cells continue to divide. (B) Longitudinal section of *Strongylocentrotus purpuratus* sea urchin embryo in late blastula stage. Ectodermal cells are flagellated; mesodermal cells have begun to ingress into the central cavity.

1.3 Opisthokonta

The unikonts consist of the Amoebozoa and the supergroup Opisthokonta [23, 136], which includes the familiar fungi and metazoans [22] (Figure 1.3). Phylogenetic analyses suggest that multicellularity evolved independently in the fungi and metazoans, as well as in less well-known groups including the nucleariids, ichthyosporea, and choanoflagellates[61, 91]. These phylogenetic conclusions are supported morphologically: for example, all major fungal clades display simple hyphal growth [91], while the only known multicellular species in the closest outgroup is a slime mold that associates by aggregation [17]. Similarly, all metazoans develop through division of a fertilized zygote, while some of the closely-related choanoflagellates aggregate facultatively [2, 126], and their mutual sister clade, the icthyosporea, are multinucleate. Irreversible cellular differentiation has also evolved in most lineages. We will review the best-studied cases, the fungi and metazoans.

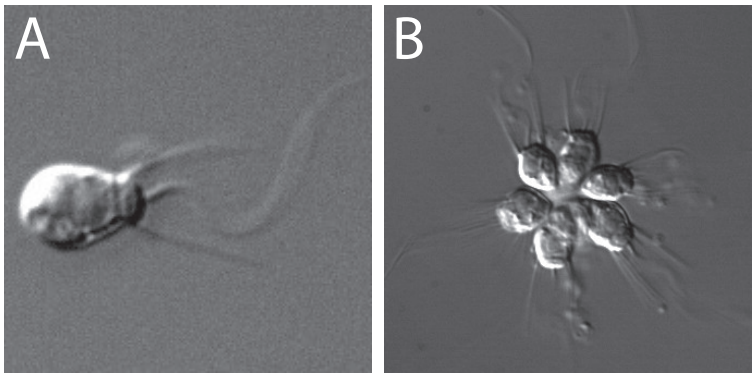

Figure 1.4: Prey-mediated adhesion induces multicellularity in the choanoflagellate *Salpingoeca rosetta*

(A) *S. rosetta* in unicellular form, with collar and flagellum visible. (B) A multicellular colony or "rosette" of *S. rosetta*. Photos courtesy of Mark Dayel.

Metazoa

Early metazoan embryogenesis parallels chlorophyte algae reproduction in three major ways. First, the fertilized zygote is massive in volume and thus able to undergo a series of rapid divisions without intervening growth that parallels algal palintomy. Second, in basal metazoan groups including the sponges and cnidarians, the somatic cells of embryos are flagellated and positioned at the anterior end or covering the exterior surface, suggesting the importance of their role in locomotion [80, 89, 100]. Third, division and flagellation appear to be mutually-exclusive processes in both lineages: flagella are shed or resorbed before mitosis [19, 74][4]. Buss [19] and others have therefore argued that, as is posited for the volvocine algae, metazoans likely evolved a soma due to a flagellation-division constraint: non-dividing somatic cells at the surface ensure motility by remaining flagellated while germline lineages actively divide.

Given the examples we have seen thus far, it is remarkable that metazoans have developed multicellularity, since they cannot or have not used any of the most common methods of cell adhesion. The common ancestor of metazoans lacked a rigid cell wall, didn't adhere through secreted stalks, and was not multinucleate. The key to multicellularity in the meta-

[4]To my knowledge there is no known explanation for why this should be necessary, given that metazoa do not possess cell walls that would prevent the migration and segregation of the flagella while still attached to the MTOCs, and the existence of multiple MTOCs per cell (e.g., supporting multiple cilia) in higher metazoans. Let us assume that the universality of the observation reflects a limitation which is not easily circumvented in the most recent common ancestor of metazoans.

zoans (besides a secreted extracellular matrix) was the presence of intercellular adhesion protein families such as the cadherins [1] and integrins [125]. These genes were present in the last common ancestor of metazoans and their nearest sister group, the choanoflagellates [1, 125]: while the metazoans became obligately multicellular, choanoflagellates lost integrin family members [125] and now are predominantly unicellular. However, some choanoflagellates form multicellular colonies through the failure of cell separation after cytokinesis, creating colonies termed "rosettes" in which the flagella point outward: in at least one case, this association appears to be a response to an inducer released by a bacterial prey species [2], though the potential benefit of this arrangement – to either species – remains unknown. This suggests that some mechanism for protein-mediated adhesion remains despite the loss of integrins. Some choanoflagellates also form multicellular groups through secretion of a theca (rigid encasement) attached to a shared stalk, analogous to the multicellular ciliates described above.

Several common developmental mechanisms have been revealed through extensive study of metazoan model species; some hold well for other complex differentiated multicellular life, particularly plants. Many eggs contain a determinant (usually protein or mRNA) which is asymmetrically distributed in the cytoplasm such that it is inherited by only a subset of cells during early embryonic cleavage. These cytoplasmic determinants initiate a distinct developmental program in the cells that inherit them: usually this will include activating a secondary, more permanent system to maintain cell fate even after the determinant disappears with time [35]. This secondary system is usually a gene regulatory network: a system of transcription factors which regulate one another's expression [30]. For example, a cytoplasmic determinant may drive the initial expression a transcription factor, which then maintains its expression by binding and activating its own promoter (directly or indirectly), forming a positive feedback loop. The transcription factor may also induce downstream genes required for cell type-specific functions, prevent differentiation to alternative cell fates by functioning as a repressor (perhaps indirectly), and affect the differentiation of nearby

cells by upregulating the expression of intercellular signalling molecules [30]. Ultimately the cell's developmental fate may be locked down yet further through epigenetic controls such as DNA methylation [131] and post-translational modification of histones [108]; autocrine or paracrine signalling [30]; or by excision of unnecessary DNA [3, 66, 76][5].

Each cell type must be present in the correct number and position in order for the embryo to survive: gene regulatory networks therefore usually contain redundant features which "canalize" cells along the correct differentiation path despite noisy conditions and changing environments [144]. It may not be surprising, then, that the transcription factors and their binding sites which specify germ layers and pattern body axes are conserved within and often between metazoan phyla [30, 82]. Still, portions of these networks work modularly and can be co-opted to define additional cell types. Expansion of signalling protein and transcription factor gene families, which extends the repertoire of possible developmental regulation systems, is thought to be pre-requisite for the evolution of many multicellular, differentiating organisms: this conjecture is corroborated by the significant expansion of these gene families in most such taxa, including the metazoans.

Fungi

Like many algae, fungi have evolved multicellularity not as a means of enhancing motility, but rather of remaining stuck in place. The common ancestor of the fungi was flagellated [22]: some groups still reproduce through flagellated single cells, but most species have lost

[5]This mechanism of differentiation was first speculated by August Weismann in his treatise *The Germ-Plasm* [147], where he postulated that the "idioplasm" is subdivided between daughter cells during somatic divisions in the embryo, such that each division produces cells with an increasingly smaller fraction of the hereditary material. Although we now know that gene excision is not the primary means of differentiation in most organisms, it remains the most definitive, for reversion to an undifferentiated state is impossible once the necessary genomic material is lost. Cases of differentiation through gene excision can be found in a variety of taxa, where many appear to result from independent co-options of tranposable elements and their recombinases [3, 66, 76]. For example, gene excision is required for the generation of functional T-cell receptors and immunoglobulins (an essential step of lymphocyte development) and is a characteristic event in mammalian erythropoiesis [14, 78, 107].

motility entirely. The phylogeny of fungi has recently been overhauled to include a new, basal clade: the Cryptomycota, which include the microsporidia [62, 63]. Like other opisthokonts, and unlike all other fungi, cryptomycetes lack a chitinous cell wall in most life cycle phases and do not form multinucleate (or septated) hyphae, placing an upper bound on the acquisition times for these traits [62, 63]. Two other major clades of fungi, the chytrids and zygomycetes, are unfortunately polyphyletic. More complex morphologies, including septated hyphae, many cell types, and macroscopic fruiting bodies, appear to have evolved once in the common ancestor of the Ascomycota (which include truffles and morels) and the Basidiomycota (which include chanterelles and more typical mushrooms).

Two explanations for evolution of multicellularity in the fungi have been advanced, with each positing a different lifestyle for the most recent common ancestor [22, 90], reflecting two strategies taken by modern-day chytrid fungi[6]: parasitic and saprophytic growth. Saprophytic chytrids extract nutrients from decaying organic material while clinging to it with a primitively-multicellular structure called a rhizoid, which consists of a branching network of hyphae; not all parasitic chytrids have rhizoids, but in those that do, their purpose is likewise to anchor the fungus in the host.

We will consider first the possibility that the common ancestor of fungi was a saprophytes. All opisthokonts share the capacity to form slender "filose" cellular projections, in contrast to the thick pseudopods found in amoebas [127]. Whereas these projections are thick and rigidly-supported by actin in the animals and choanoflagellates, analogous structures in the nucleariids and fungi are flexible, tapering, and sometimes branched. It has been proposed that this morphological innovation might have served to root saprobes to a substrate, eventually being elaborated into hyphae to perform a function similar to the rhizoids of chytrid fungi[7]. This innovation would have interfered with locomotion, perhaps

[6]These explanations were proposed long before the cryptomycetes were known to be the most basal clade of fungi; each remains valid in light of the new phylogenetic information.

[7]Whereas algal holdfasts are merely physical attachments, rhizoids are more like land plant roots in that they permit the uptake of nutrients.

explaining the loss of flagella in some lineages (despite their clear advantage for dispersal in aquatic species). A disadvantage of this theory is that it presumes the fungal ancestor had already specialized in a saprotrophic lifestyle before acquiring structural adaptations to hold it in place: a chicken-and-egg problem. This concern has since been addressed with the further conjecture that attachment may have been initially beneficial for another reason. For example, if the ancestor of fungi maintained a lifestyle similar to the common ancestor of metazoans and choanoflagellates (using a flagellum to trap prey) then anchoring to a substrate might improve its efficacy; transition to saprotrophy could then have been secondary to a sessile lifestyle [90].

An alternative theory holds that the common ancestor of fungi was parasitic, and the initial benefit of rhizoid formation was to prolong the parasites' stay and/or to gain access to nutrients within the host [22]. Support for this interpretation comes from the observation of parasitic cryptomycetes, which (lacking hyphae) attach to their diatom hosts by other means [63]; as well as of parasitic higher fungi which parasitize other single-celled hosts by attaching through their branched hyphae [6].

In either event, an analogous evolutionary trajectory is proposed for the brown algae (Phaeophyta) [28] and in sea lettuces (green algae of genus *Ulva*), which form rhizoid-like holdfasts to affix to the seafloor [12]. These algae do not form filose projections, tapered downward growth to form the holdfast begins directly in the zygote in some species of each group [11, 12]. Individual cells in the sea lettuces and brown algae remain together through failure to degrade the rigid cell walls that connect them: in this sense they are similar to the red algae discussed above. The common ancestor of fungi lacked a cell wall: perhaps this explains why the fungi formed multinucleate hyphae rather than cellularized filaments.

Multicellularity in the form of hyphal growth is believed to have been present in the common ancestor of fungi. Reversions to unicellularity have occurred frequently to generate yeast species scattered throughout the fungal phylogenetic tree [37, 77]. The unicellular state, too, can be reverted through mutations that prevent cell separation by chitinases following

cytokinesis: recently, this result has been achieved by experimental evolution [73, 105, 112]. It is thus quite possible that hyphal growth has evolved multiple times in the fungal lineage. A more complex form of multicellularity is believed to have evolved once in the Dikarya (consisting of the crown groups Ascomycota and Basidiomycota), which often have septated hyphae and macroscopic fruiting bodies. Evidence for a single acquisition event is mainly morphological, though buttressed by the observation of related gene family expansions in the common ancestor of both groups [135].

Evolution of cellular differentiation in the Dikarya is naturally thought to have been facilitated by the appearance of fully septated hyphae [75]. This is perhaps an oversimplification, however, as gene expression states can differ dramatically between nuclei even in synctial hyphae [79]. The sessile, terrestrial lifestyle of the Dikarya [55] also likely contributed to the advent of differentiation by engendering a strong selective pressure for the effective dispersal of spores by air: structures mediating spore ejection into the airstream are often morphologically complex [42]. Differentiation during fruiting body formation appears to be mediated by the usual suspects, gene regulatory networks of transcription factors, hormones [25], and WD40 domain-containing proteins involved in intercellular signal transduction [85, 109]. True terminal differentiation is rare among the fungi: morphologically-distinct hyphae, once experimentally removed from the fruiting body, can regenerate all forms in many of the cases examined carefully [95]. However, since differentiation states are maintained stably when hyphae remain within in the fruiting body (i.e. under natural circumstances), these fungi effectively possess terminal differentiation.

1.4 Amoebozoa

The sister group to the opisthokonts is Amoebozoa: it consists of non-photosynthetic, phagotrophic protozoa that move primarily by cytoplasmic flow into pseudopods, though some groups retain flagella. The slime molds, which include multinucleate acellular as well

as aggregative cellular forms, have evolved at least three times in this clade[8] and often show evidence of nuclear or cellular differentiation. In this section we will highlight *Dictyostelium*, a cellular slime mold which has been the focus of biochemical, genetic, and mechanical studies for over four decades and is representative of an aggregative multicellular lifestyle which has evolved independently in many other taxa, including the Rhizaria (*Guttulinopsis vulgaris*), Excavata (the acrasid slime molds), Ciliophora (*Sorogena stoianovitchae*), and Nucleariida (*Fonticula alba*) [16, 17, 141]. We note, however, that the acellular slime mold *Physarum polycephalum* is also fast becoming a model organism for the study of morphogenesis [39, 142] and is representative of a multinucleate hyphal lifestyle which has also evolved independently in the fungi and oomycetes.

Dictyostelium

Aggregative multicellularity has arisen independently in a diverse collection of prokaryotic and eukaryotic clades, but the best studied example remains the cellular slide mold *Dictyostelium discoideum*. *Dictyostelium* live as unicellular amoebae when their bacterial prey are plentiful, but under starvation conditions, as many as 100,000 cells may aggregate to form a macroscopic "slug." Cells in the slug secrete an extracellular matrix along which the slug moves using a repetitive series of contractions and extensions [15]: this coordinated motion between cells allows the slug to migrate up to 6 cm – much further than the individual cells could achieve alone [145]. The slug travels toward the light (which in their natural environment typically corresponds to the surface of the leaf layer) for several days, then halts and rears up to form a fruiting body. The cells at the apex sporulate, while the remainder form a thin, rigid stalk of ≈2 mm [8] that elevates the spores above the boundary layer to facilitate their dispersal[9].

[8] As the name implies, prior to the advent of molecular phylogeny, slime molds were thought to be primitive forms of fungi.

[9] Though at this height the spores have surpassed the boundary layer, dispersal by wind is unlikely because of a thick, sticky shell that joins the spores to the stalk. Dispersal by animal vectors, however, has been noted in the wild [60].

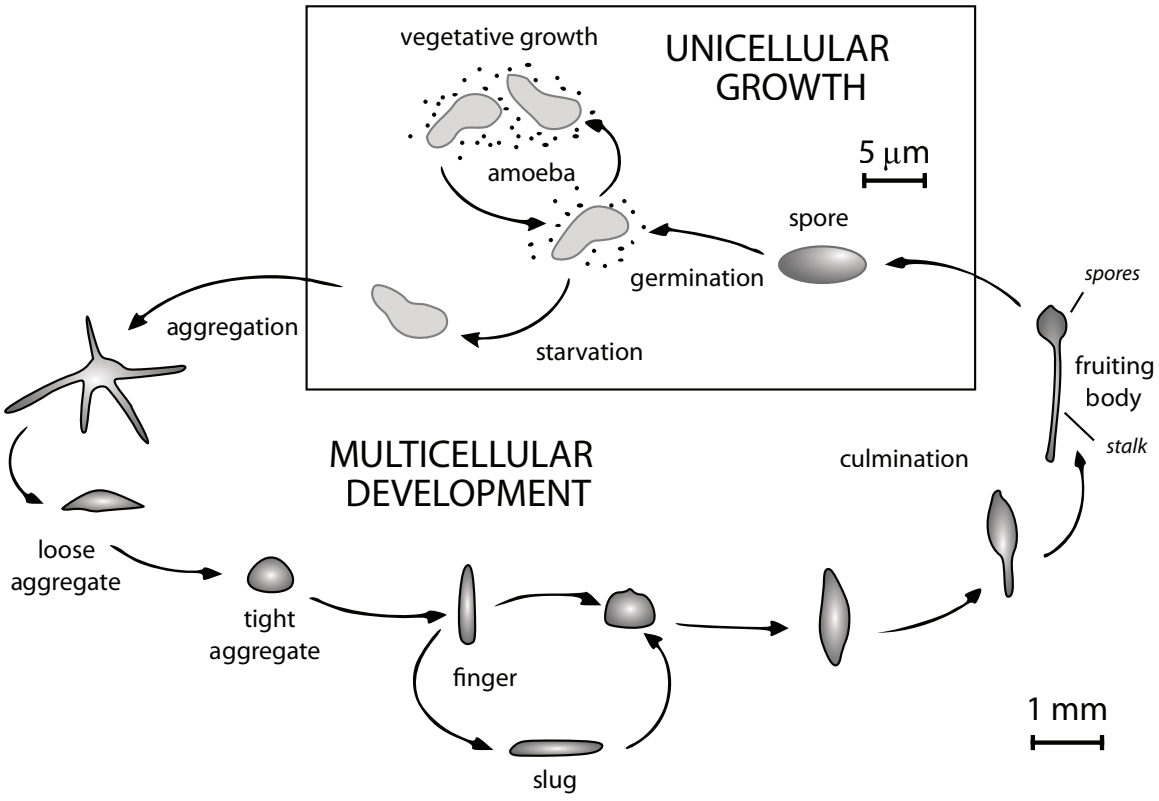

Figure 1.5: Life cycle of the cellular slime mold *Dictyostelium discoideum*.

Image provided under the GNU Free Documentation License by Tijmen Stam.

To understand the evolutionary adaptations required to adopt its multicellular lifestyle, we must understand the basis of *Dictyostelium* cells' cohesion and coordinated motion. Much of the slug's motion can be attributed to mechanical properties of the slimy "shell" it continuously secretes. The shell is weak at the anterior end but elastic enough to exert a surface tension of ≈50 mN/m (comparable to water) at the middle and rear of the slug [116]: the resulting pressure pushes cells toward the anterior end[10] [81, 116]. The shell remains stationary while the slug advances through it: the hollow tube eventually collapses at the rear end, and the resulting force is also expected to contribute to the slug's forward

[10]The same force is predicted to drive the expansion of plant meristems [36], which have an extracellular matrix of similar biochemical composition to the *Dictyostelium* slug shell [149]. The outward pressure exerted by slug's cells is analogous to the outward pressure exerted by cytoplasm of hyphal tips of fungi [137] and of pseudopods in amoebae.

motion [150]. The slug's forward motion does not appear to be entirely passive, however, as the molecular motor myosin II is heavily expressed on the slug's ventral surface [34], where its activity prevents "slippage"; the importance of this contribution is shown by failure of slug (but not individual cell) motility in myosin heavy chain mutants [148]. Thus, at a minimum, evolution of slug motility likely required a mechanism for aggregation, secretion of a shell, and coordinated motion within the slug. A possible evolutionarily-intermediate form, suggested by the "tipped mound" appearance of *Dictyostelium* aggregates before movement begins, is the secretion of a shell around an aggregate to force cells vertically upward and thus directly into a non-motile stalk.

The cells which comprise the stalk of the fruiting body are effectively somatic: unlike the spores, they will not directly contribute reproductively to future generations of *Dictyostelium*, yet are clearly necessary for forming a tall rigid structure for spore dispersal. The differentiation system which specifies stalk cells is already in motion long before cells first aggregate to form the slug: cells which were the least well-fed prior to aggregation are more responsive to Differentiation Inducing Facgtor (DIF) family signalling molecules that bias cells toward a stalk fate [26, 143]. This arrangement, christened "survival of the fattest" [96], might be beneficial because it ensures that the cells which will form spores are the most likely to survive, while only relatively moribund cells sacrifice their reproductive potential[11].

Since slugs are formed by aggregation, there is no guarantee of clonality: mutants less prone to differentiate and wildtype cells form slugs together: when this occurs, production of the stalk is "exploited" by the mutants, which make up a larger-than-representative fraction of the spores. This effect is seen in wild isolates and in the laboratory [45, 140]: it is presumed that spore dispersal eventually forces a unicellular bottleneck, limiting the spread of mutants incapable of differentiation.

[11] As with the fungi, there is evidence that differentiation in the slime mold can be reversed by experimentally disrupting a slug or stalk to separate the cell types. Within the natural context, however, differentiation state is effectively permanent.

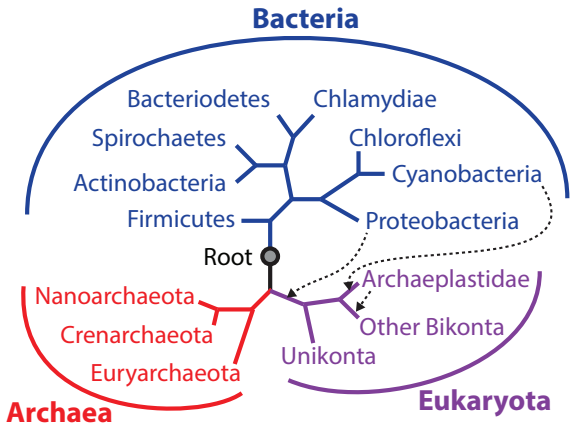

Figure 1.6: Phylogenetic relationships between the kingdoms of life

Major clades of interest for this section include the cyanobacteria, proteobacteria (including the myxobacteria), and crenarchaea (including *Pyrodictium*). Dotted lines represent major endosymbiont acquisitions which formed the eukaryotic mitochondrion (from a proteobacterium [84]), chloroplast (from a cyanobacterium [87]), and secondary chloroplasts (from green and red algae [67]). This cladogram reflects the phylogeny of Ciccarelli et al. [27].

1.5 Prokaryota

Bacteria and archaea have evolved multicellular forms akin to many of the examples described above. The most common morphologies are filaments, as found in the cyanobacteria highlighted below, and biofilms, cohesive colonies initated through pilial or flagellar adhesion and maintained by production of an extracellular matrix [43, 53, 83, 117]. Some actinobacteria, including the well-studied *Streptomyces*, have multinucleate hyphal forms similar to those of fungi. The methane-producing euryarcheons *Methanosarcina* form macroscopic, amorphous, clonal clumps, while the myxobacteria aggregate to form fruiting bodies in a manner similar to *Dictyostelium* [31]. Despite much overlap in morphology with eukaryotic multicellular clades, these domains of life also contain unparalleled forms like the woven network of cytoplasmic bridges formed by the *Pyrodictium* crenarcheon hyperthermophiles, discussed below.

Cyanobacteria

Cyanobacteria are believed to have caused the Great Oxygenation Event ≈2.8 billion years ago, dramatically increasing Earth's oxygen levels through their photosynthetic activity and triggering the global glaciation event known as "snowball Earth" [72]. The world's oldest

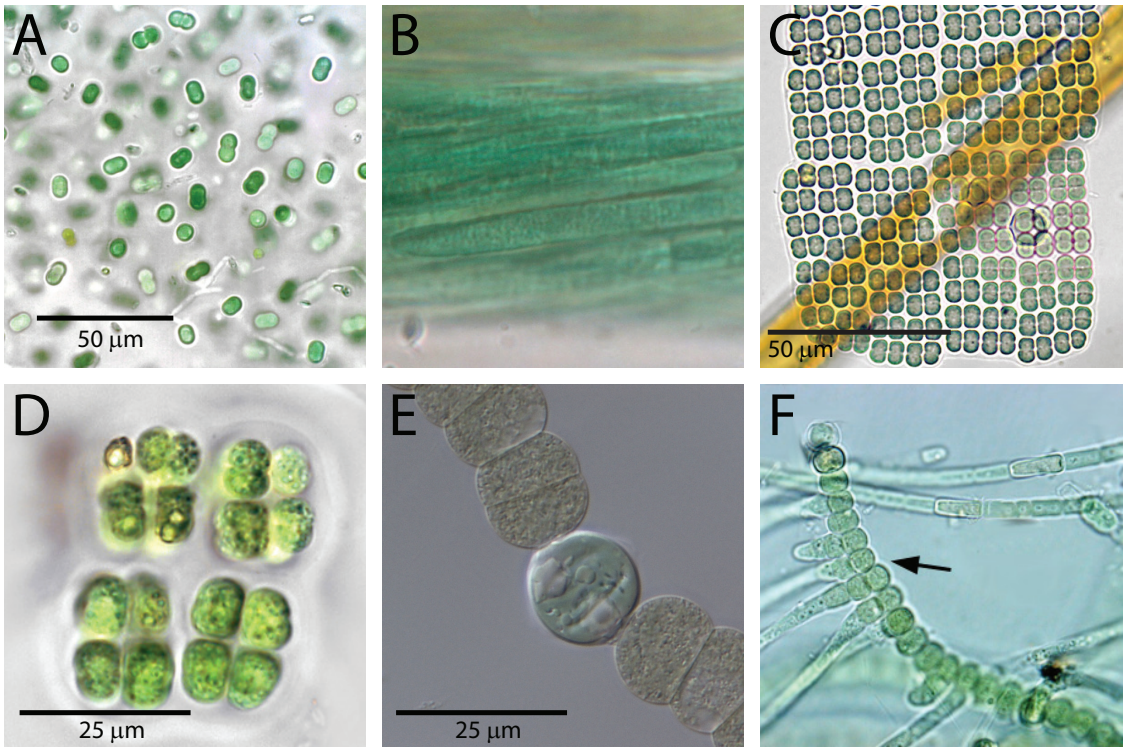

Figure 1.7: Diversity of forms in the cyanobacteria

(A) The unicellular cyanobacterium *Synechococcus elongatus*. Photo courtesy of Chris Carter. (B) The filamentous, non-differentiating species *Coleofasciculus chthonoplastes*. Photo courtesy of Ignacio Bárbara. (C) Planar colonies of *Merismopedia elegans*. Photo courtesy of Chris Carter. (D) Small, cubic colonies of *Eucapsis alpina*. Photo courtesy of Chris Carter. (E) A heterocyst within an *Anabaena crassa* filament. Photo courtesy of Karl Bruun. (F) Branched filaments of differentiating *Fischerella* sp.: note elongated morphology of cells at ramifications (arrow). Photo courtesy of Peter Siver [130].

multicellular fossils, dating back over 2 billion years, are believed to represent filamentous cyanobacteria[12] [111]. Molecular clock estimates support these fossil interpretations by dating the most recent common ancestor of extant multicellular cyanobacteria to approximately 2 billion years ago [32, 124]. The common ancestor of cyanobacteria was most likely unicellular, and many unicellular species persist, though phylogenetic analyses suggest that at least seven independent reversions to unicellularity have occurred [123, 128]. These studies also reveal that contrary to "Dollo's Law," which states that complex traits are exceedingly unlikely to be regained once they are lost [46], multicellularity has been gained, lost, and reacquired multiple times in the cyanobacteria. Loss of multicellularity has apparently not occurred in cyanobacteria, however, in clades that have also evolved cellular differentiation [123, 128].

Evolution of a filamentous lifestyle requires axial growth and adhesion at the poles following division: the former is common to many prokaryotes including the unicellular cyanobacteria, suggesting that the only necessary adaptation for filamentous growth was a mechanism for maintaining attachment between cells. This seemingly straightforward innovation is complicated slightly by the fact that the cyanobacteria are diderm (gram-negative) bacteria with two plasma membranes: in multicellular species, the outer membrane is continuous across the entire filament, holding the cells together. While average filament length depends partly on external factors like mechanical stress and bacteriophage-induced lysis, cyanobacteria also directly regulate this property through programmed cell death [9] as well as other genetically-encoded mechanisms that remain to be elucidated [94]. The inherent advantages of filamentous growth, as in many other forms of multicellularity, have been hypothesized to include improved efficiency of nutrient uptake requiring secreted enzymes and predator avoidance[13] [11, 50, 51]; both passive and active changes in filament size un-

[12] Filamentous fossils up to 3 billion years old have also been claimed for the cyanobacteria on the basis of morphological similarity to extant species, but these claims are somewhat more controversial.

[13] Predator avoidance was likely not the initial selective pressure for multicellularity, since the first fossils of phagotrophic species do not appear until bout 750 million years ago [110], but it may have been advantageous since their advent.

der appropriate conditions facilitate dispersal and avoid excessive local depletion of limiting nutrients [120].

Multicellular cyanobacteria can terminally differentiate to form heterocysts: these sterile cells are specialized for the fixation of nitrogen, which they share with neighboring cells primarily through the secretion of glutamine. Cyanobacteria are also commonly considered to have two other types of specialized cells: hormogonia and akinetes (resting cells akin to asexual spores). Hormogonia are transiently-motile single cells [118] induced by growth conditions and/or secretions of their plant symbionts: they return to a regular sedentary lifestyle within a few days (reviewed in [21, 92]). Since cyanobacterial cells can and do often transition between these forms, by the definition given above, they would be considered transient gene expression states rather than differentiated cells. Heterocysts, hormogonia, and akinetes are clearly visible in fossils over 2 billion years old, when cyanobacterial morphological complexity evidently plateaued [123]. The question of whether cyanobacteria were incapable of evolving more complicated forms (e.g., due to the absence of genomic pre-adaptations) or simply did not experience evolutionary pressure to do so remains open.

The benefit of heterocyst differentiation in the cyanobacteria derives from the incompatibility of two critical processes, nitrogen fixation and photosynthesis, due to the inactivation of nitrogenase by oxygen. While some cyanobacteria have responded to this pressure by simply relegating nitrogen fixation to the evening hours, the differentiation of heterocysts permits nitrogen fixation throughout the day. It is not surprising that heterocyst differentiation is terminal rather than transient because to avoid exposure to oxygen, heterocysts must forsake photosynthesis through the O_2-evolving photosystem II, purge gas vacuoles, and form an extracellular envelope to reduce gas diffusion into the cell below the rate at which it can be respired [146]. Furthermore, since heterocysts provide nitrogen to neighboring cells by diffusion, the expected returns of continued heterocyst division would be marginal relative to differentiation of new heterocysts further away on the filament. A complex patterning mechanism involving multiple ligand-receptor systems ensures the differentiation

of an appropriate number of new heterocysts when cells begin to experience nitrogen deprivation [41, 98, 151]. The expansion of multiple signal transduction pathways in heterocystic cyanobacteria suggest that intercellular signaling was a key pre-adaptation for the evolution of differentiation in this clade [65].

Methanosarcina

Like the cyanobacteria, the *Methanosarcina* are so prolific that their evolution and subsequent methane production is thought to have caused a major geological transition [121]. Members of this genus are anaerobes found in such diverse locations as hydrothermal vents, deep freshwater, and the digestive tracks of ruminants, where they transition between a unicellular dispersal stage and the formation of clonal clumps large enough to be visible to the naked eye [119]. Bonner and Stetter have proposed that multicellularity evolved in this group to protect cells in the interior of the clump from exposure to oxygen [12]. If accurate, the benefits derived from clumping would be similar to those demonstrated for flocculating yeast. Flocculation is a form of aggregation through the expression of cell surface lectins that has evolved in some yeast (themselves unicellular revertants of multicellular fungi) which is initiated under stressful conditions: cells in the center of the macroscopic clumps that form are protected from environmental stressors by the physical shield of dead cells around them [132].

Pyrodictia

Members of the genus *Pyrodictium* were the first extremophiles demonstrated to grow at temperatures above 100°C [138]. Some species form long, thin cytoplasmic bridges called cannulae that seem to both connect cells and, through their elasticity, to enforce their regular spacing. Cannulae are formed initially as looping periplasmic projections doubly-attached to the surface of the same cell: if the cell division plane happens to fall between the two ends of a cannula, it will connect them after cytokinesis [59]. These cytoplasmic bridges are fairly

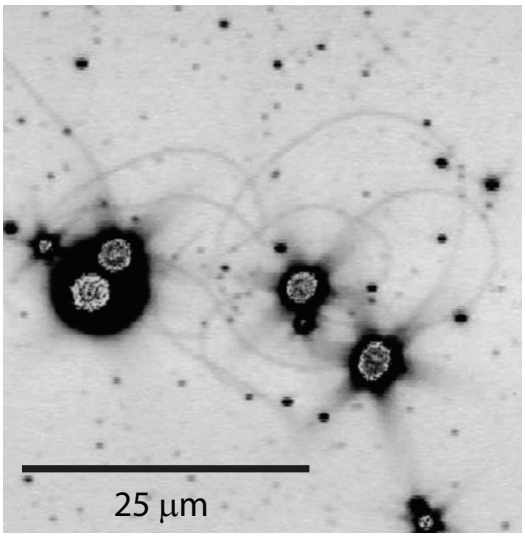

Figure 1.8: Cell adhesion through cannulae in *Pyrodictium abyssi*.

Cells of *Pyrodictium abyssi* produce cannulae that often form looping projections onto the cell's own surface. When a looped cannula's two ends are separated by the mitotic division plane, it becomes a linkage between the daughter cells, as shown here. Photograph courtesy of Harald Huber [59].

stable and elongate slowly during cell growth, allowing dense clusters of clonally-related cells to remain interconnected [59]: as they extend only into the periplasmic space [99], the outer membrane presumably further stabilizes the multicellular groups as it does in the cyanobacteria. Cannulae comprise a large fraction of the biomass of *Pyrodictium* cultures, and are thus expected to provide substantial benefits to offset their costs of production. The advantages of multicellularity *per se* for *Pyrodictium*, which grows in hydrothermal vents, are unknown. Noting that the cannulae of *Pyrodictium* are able to withstand much hotter temperatures than the cells themselves, Stetter [139] speculates that they serve to force portions of the colony into much hotter regions of the vent, where perhaps different bioelectrochemistry can be accomplished to the benefit of the surviving, connected cells.

1.6 Conclusion

The examples above illustrate the diverse circumstances and means by which differentiation and multicellularity have evolved previously. We have seen that multicellularity can join cells that are related to one another by descent from a common ancestor (clonal multicellularity) or free-living, potentially-unrelated single cells which join together facultatively (aggregative multicellularity). Organisms with aggregative multicellularity are small in size

and have few cell types, suggesting that this strategy does not extend effectively to greater levels of complexity, which may be due in part to inherent intergenomic competition within the aggregate. Mechanistically, multicellularity can be implemented through protein-based cell-to-cell adhesion, enclosure within a common extracellular matrix, incomplete cytokinesis, cell fusion, and even elaborate cytoplasmic bridges. Evolution of multicellularity may be driven by size increase alone (for example, to avoid predators or gain preferential access to sunlight), rooting to a substrate, generating an internal environment (e.g. for phosphate storage or germline cell division), facilitating access to a public good, or improving locomotion and dispersal.

Differentiation, too, comes in many forms. It can be strictly irreversible, as for enucleated red blood cells and entombed heterocysts; effectively irreversible, as when position effects govern gene expression state in Dikarya fruiting bodies; or transient, as for motile propagule life stages like cyanobacterial hormogonia – even cytoplasmic separation is not a requirement, as we have seen for syncytial nuclei in fungi. Differentiation often separates two or more processes that are inherently antagonistic, including photosynthesis and nitrogen fixation in cyanobacteria; flagellar motion and cell division in the volvocine algae and metazoans; or spore dispersal and germline propagation in fungal and slime mold fruiting bodies. Differentiation can be mediated by differential segregation of determinants, intercellular signalling, transcription factor-based gene regulatory networks, epigenetic modifications, heritable changes in protein activity, and, most drastically, gene excision.

In all cases where speculation is possible or where "intermediate forms" remain, it appears that multicellularity evolved prior to differentiation. Many of the functions of differentiated cells described above would not be achievable in a unicellular species, with a notable exception: nutrient exchange between cell types, e.g. in the heterocystous cyanobacteria. In chapter three, we will describe the production of a unicellular budding yeast strain which differentiates to produce cells that release monosaccharides into the culture media, just as heterocysts secrete glutamine for uptake by neighboring cells. First, however, we de-

vote chapter two to the problem of engineering cellular differentiation *de novo*, including how to create a complete dichotomy between cell types, impose irreversible conversion, and arbitrarily assign functions to daughter cells.

1.7 Bibliography

[1] M. Abedin and N. King. The premetazoan ancestry of cadherins. *Science*, 319(5865):946–948, 2008.

[2] R. A. Alegado, L. W. Brown, S. Cao, R. K. Dermenjian, R. Zuzow, S. R. Fairclough, J. Clardy, and N. King. A bacterial sulfonolipid triggers multicellular development in the closest living relatives of animals. *eLife*, 1, 2012.

[3] S. Apte and N. Prabhavathi. Rearrangements of nitrogen fixation (*nif*) genes in the heterocystous cyanobacteria. *Journal of Biosciences*, 19:579–602, 1994.

[4] Y. Arakaki, H. Kawai-Toyooka, Y. Hamamura, T. Higashiyama, A. Noga, M. Hirono, B. J. S. C. Olson, and H. Nozaki. The simplest integrated multicellular organism unveiled. *PLoS ONE*, 8(12):e81641, 12 2013.

[5] S. L. Baldauf. The deep roots of eukaryotes. *Science*, 300(5626):1703–1706, 2003.

[6] G. L. Barron. A new *Amoebophilus* (zygomycetes) ectoparasitic on amoebae. *Canadian Journal of Botany*, 61(12):3091–3094, 1983.

[7] G. Bell. The origin and early evolution of germ cells as illustrated in the Volvocales. In H. O. Halvorson and A. Monroy, editors, *The origin and evolution of sex*. A. R. Liss., 1985.

[8] M. A. Benjamin, J. A. Gilchriest, and S. Lehrer. Development in altered gravity influences height in *Dictyostelium*. *Gravitational and Space Research*, 1(1):51 – 58, 2013.

[9] I. Berman-Frank, K. D. Bidle, L. Haramaty, and P. G. Falkowski. The demise of the marine cyanobacterium, *Trichodesmium* spp., via an autocatalyzed cell death pathway. *Limnology and Oceanography*, 49(4):997 – 1005, 2004.

[10] K. Bišová and V. Zachleder. Cell-cycle regulation in green algae dividing by multiple fission. *Journal of Experimental Botany*, 2014.

[11] J. Bonner. *On Development: the Biology of Form*. Harvard University Press, 1974.

[12] J. Bonner. *First Signals*. Princeton University Press, 2000.

[13] M. Boraas, D. Seale, and J. Boxhorn. Phagotrophy by a flagellate selects for colonial prey: A possible origin of multicellularity. *Evolutionary Ecology*, 12(2):153–164, 1998.

[14] C. Brack, M. Hirama, R. Lenhard-Schuller, and S. Tonegawa. A complete immunoglobulin gene is created by somatic recombination. *Cell*, 15:1–14, 1978.

[15] E. Breen, P. Vardy, and K. Williams. Movement of the multicellular slug stage of *Dictyostelium discoideum*: an analytical approach. *Development*, 101(2):313–321, 1987.

[16] M. W. Brown, M. Kolisko, J. D. Silberman, and A. J. Roger. Aggregative multicellularity evolved independently in the eukaryotic supergroup Rhizaria. *Current Biology*, 22(12):1123–1127, 2012.

[17] M. W. Brown, F. W. Spiegel, and J. D. Silberman. Phylogeny of the "forgotten" cellular slime mold, *Fonticula alba*, reveals a key evolutionary branch within Opisthokonta. *Molecular Biology and Evolution*, 26(12):2699–2709, 2009.

[18] L. W. Buss. Competition within and between encrusting clonal invertebrates. *Trends in Ecology & Evolution*, 5(11):352 – 356, 1990.

[19] L. Buss. *The evolution of individuality*. Princeton University Press, 1987.

[20] N. J. Butterfield. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the mesoproterozoic/neoproterozoic radiation of eukaryotes. *Paleobiology*, 26(3):386–404, 2000.

[21] E. L. Campbell, H. Christman, and J. C. Meeks. DNA microarray comparisons of plant factor- and nitrogen deprivation-induced hormogonia reveal decision-making transcriptional regulation patterns in *Nostoc punctiforme*. *Journal of Bacteriology*, 190(22):7382–7391, 2008.

[22] T. Cavalier-Smith. The origin of fungi and pseudofungi. In A. D. Rayner, editor, *Evolutionary Biology of Fungi*. Cold Spring Harbor Laboratory Press, 1987.

[23] T. Cavalier-Smith. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *International Journal of Systematic and Evolutionary Microbiology*, 52(2):297–354, 2002.

[24] T. Cavalier-Smith and E.-Y. Chao. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom chromista). *Journal of Molecular Evolution*, 62(4):388–420, 2006.

[25] Y. Champavier, M.-T. Pommier, N. Arpin, A. Voiland, and G. Pellon. 10-oxo-trans-8-decenoic acid (ODA): production, biological activities, and comparison with other hormone-like substances in *Agaricus bisporus*. *Enzyme and Microbial Technology*, 26(2fb4):243 – 251, 2000.

[26] A. Chattwood, K. Nagayama, P. Bolourani, L. Harkin, M. Kamjoo, G. Weeks, and C. R. Thompson. Developmental lineage priming in *Dictyostelium* by heterogeneous Ras activation. *eLife*, 2, 2013.

[27] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–1287, 2006.

[28] J. M. Cock, L. Sterck, P. Rouze, D. Scornet, A. E. Allen, G. Amoutzias, V. Anthouard, F. Artiguenave, J.-M. Aury, J. H. Badger, B. Beszteri, K. Billiau, E. Bonnet, J. H. Bothwell, C. Bowler, C. Boyen, C. Brownlee, C. J. Carrano, B. Charrier, G. Y. Cho, S. M. Coelho, J. Collen, E. Corre, C. Da Silva, L. Delage, N. Delaroque, S. M. Dittami, S. Doulbeau, M. Elias, G. Farnham, C. M. M. Gachon, B. Gschloessl, S. Heesch, K. Jabbari, C. Jubin, H. Kawai, K. Kimura, B. Kloareg, F. C. Kupper, D. Lang, A. Le Bail, C. Leblanc, P. Lerouge, M. Lohr, P. J. Lopez, C. Martens, F. Maumus, G. Michel, D. Miranda-Saavedra, J. Morales, H. Moreau, T. Motomura, C. Nagasato, C. A. Napoli, D. R. Nelson, P. Nyvall-Collen, A. F. Peters, C. Pommier, P. Potin, J. Poulain, H. Quesneville, B. Read, S. A. Rensing, A. Ritter, S. Rousvoal, M. Samanta, G. Samson, D. C. Schroeder, B. Segurens, M. Strittmatter, T. Tonon, J. W. Tregear, K. Valentin, P. von Dassow, T. Yamagishi, Y. Van de Peer, and P. Wincker. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature*, 465(7298):617–621, Jun 2010.

[29] K. Cole and R. Sheath. *Biology of the Red Algae*. Cambridge University Press, 1990.

[30] E. Davidson. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Academic Press, 2006.

[31] W. Dawid. Biology and global distribution of myxobacteria in soils. *FEMS Microbiology Reviews*, 24(4):403–427, 2000.

[32] J. Domínguez-Escobar, Y. Beltrán, B. Bergman, B. Dez, K. Ininbergs, V. Souza, and L. I. Falcón. Phylogenetic and molecular clock inferences of cyanobacterial strains within Rivulariaceae from distant environments. *FEMS Microbiology Letters*, 316(2):90–99, 2011.

[33] K. Drescher, R. E. Goldstein, and I. Tuval. Fidelity of adaptive phototaxis. *Proceedings of the National Academy of Sciences*, 107(25):11171–11176, 2010.

[34] S. Eliott, P. H. Vardy, and K. L. Williams. The distribution of myosin II in *Dictyostelium discoideum* slug cells. *The Journal of Cell Biology*, 115(5):1267–1274, 1991.

[35] B. Ewen-Campen, E. E. Schwager, and C. G. Extavour. The molecular machinery of germ line specification. *Molecular Reproduction and Development*, 77(1):3–18, 2010.

[36] P. Fayant, O. Girlanda, Y. Chebli, C.-E. Aubin, I. Villemure, and A. Geitmann. Finite element model of polar growth in pollen tubes. *The Plant Cell*, 22(8):2579–2593, 2010.

[37] J. Fell, T. Boekhout, A. Fonseca, and J. Sampaio. Basidiomycetous yeasts. In D. J. McLaughlin, E. G. McLaughlin, and P. A. Lemke, editors, *The Mycota*. Springer-Verlag, 2001.

[38] J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410, 1978.

[39] A. Fessel, C. Oettmeier, E. Bernitt, N. C. Gauthier, and H.-G. Döbereiner. *Physarum polycephalum* percolation as a paradigm for topological phase transitions in transportation networks. *Phys. Rev. Lett.*, 109:078103, Aug 2012.

[40] G. L. Floyd. Mitosis and cytokinesis in *Asteromonas gracilis*, a wall-less green monad. *Journal of Phycology*, 14(4):440–445, 1978.

[41] G. E. Fogg. Growth and heterocyst production in *Anabaena cylindrica* Lemm: II. in relation to carbon and nitrogen metabolism. *Annals of Botany*, 13(3):241–259, 1949.

[42] J. A. Fritz, A. Seminara, M. Roper, A. Pringle, and M. P. Brenner. A natural O-ring optimizes the dispersal of fungal spores. *Journal of The Royal Society Interface*, 10(85), 2013.

[43] S. Fröls. Archaeal biofilms: widespread and complex. *Biochemical Society Transactions*, 41:393 – 398, 2013.

[44] P. W. Gabrielson, D. J. Garbary, and R. F. Scagel. The nature of the ancestral red alga: Inferences from a cladistic analysis. *Biosystems*, 18(3fb4):335 – 346, 1985.

[45] O. M. Gilbert, K. R. Foster, N. J. Mehdiabadi, J. E. Strassmann, and D. C. Queller. High relatedness maintains multicellular cooperation in a social amoeba by controlling cheater mutants. *Proceedings of the National Academy of Sciences*, 104(21):8913–8917, 2007.

[46] S. J. Gould. Dollo on Dollo's law: Irreversibility and the status of evolutionary laws. *Journal of the History of Biology*, 3(2):189–212, 1970.

[47] R. K. Grosberg and R. R. Strathmann. The evolution of multicellularity: A minor major transition? *Annual Review of Ecology, Evolution, and Systematics*, 38(1):621–654, 2007.

[48] M. D. Guiry and G. M. Guiry. AlgaeBase, National University of Ireland, Galway. `http://www.algaebase.org`, 2014.

[49] E. Haeckel. *Die Kalkschwaemme. Eine monographie.* Verlag von Georg Reimer, 1872.

[50] M. W. Hahn and M. G. Höfle. Grazing pressure by a bacterivorous flagellate reverses the relative abundance of *Comamonas acidovorans* PX54 and *Vibrio* strain CB5 in chemostat cocultures. *Applied and Environmental Microbiology*, 64(5):1910–1918, 1998.

[51] M. W. Hahn, E. R. B. Moore, and M. G. Höfle. Bacterial filament formation, a defense mechanism against flagellate grazing, is growth rate controlled in bacteria of different phyla. *Applied and Environmental Microbiology*, 65(1):25–35, 1999.

[52] J. Haldane. On being the right size. In J. Maynard Smith, editor, *On Being the Right Size and Other Essays*, volume 10. Oxford University Press, 1985.

[53] L. Hall-Stoodley, J. W. Costerton, and P. Stoodley. Bacterial biofilms: from the natural environment to infectious diseases. *Nature Reviews Microbiology*, 2004.

[54] D. He, O. Fiz-Palacios, C.-J. Fu, J. Fehling, C.-C. Tsai, and S. Baldauf. An alternative root for the eukaryote tree of life. *Current Biology*, 24(4):465 – 470, 2014.

[55] D. S. Heckman, D. M. Geiser, B. R. Eidell, R. L. Stauffer, N. L. Kardos, and S. B. Hedges. Molecular evidence for the early colonization of land by fungi and plants. *Science*, 293(5532):1129–1133, 2001.

[56] M. D. Herron, J. D. Hackett, F. O. Aylward, and R. E. Michod. Triassic origin and early radiation of multicellular volvocine algae. *Proceedings of the National Academy of Sciences*, 106(9):3254–3258, 2009.

[57] S. J. Holmes. Phototaxis in *Volvox*. *Biological Bulletin*, 4(6):pp. 319–326, 1903.

[58] H. J. Hoops and G. B. Witman. Basal bodies and associated structures are not required for normal flagellar motion or phototaxis in the green alga *Chlorogonium elongatum*. *The Journal of Cell Biology*, 100(1):297–309, 1985.

[59] C. Horn, B. Paulmann, G. Kerlen, N. Junker, and H. Huber. In vivo observation of cell division of anaerobic hyperthermophiles by using a high-intensity dark-field microscope. *Journal of Bacteriology*, 181(16):5114–5118, 1999.

[60] M. J. Huss. Dispersal of cellular slime molds by two soil invertebrates. *Mycologia*, 81(5):677 – 682, 1989.

[61] T. Y. James, F. Kauff, C. L. Schoch, P. B. Matheny, V. Hofstetter, C. J. Cox, G. Celio, C. Gueidan, E. Fraker, J. Miadlikowska, H. T. Lumbsch, A. Rauhut, V. Reeb, A. E. Arnold, A. Amtoft, J. E. Stajich, K. Hosaka, G.-H. Sung, D. Johnson, B. OŔourke, M. Crockett, M. Binder, J. M. Curtis, J. C. Slot, Z. Wang, A. W. Wilson, A. Schuszler, J. E. Longcore, K. O/'Donnell, S. Mozley-Standridge, D. Porter, P. M. Letcher, M. J. Powell, J. W. Taylor, M. M. White, G. W. Griffith, D. R. Davies, R. A. Humber, J. B. Morton, J. Sugiyama, A. Y. Rossman, J. D. Rogers, D. H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R. A. Spotts, M. Serdani, P. W. Crous, K. W. Hughes, K. Matsuura, E. Langer, G. Langer, W. A. Untereiner, R. Lucking, B. Budel, D. M. Geiser, A. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D. S. Hibbett, F. Lutzoni, J. W. Spatafora, and R. Vilgalys. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*, 443(7113):818–822, Oct 2006.

[62] T. James, A. Pelin, L. Bonen, S. Ahrendt, D. Sain, N. Corradi, and J. Stajich. Shared signatures of parasitism and phylogenomics unite Cryptomycota and Microsporidia. *Current Biology*, 23(16):1548 – 1553, 2013.

[63] M. D. M. Jones, I. Forn, C. Gadelha, M. J. Egan, D. Bass, R. Massana, and T. A. Richards. Discovery of novel intermediate forms redefines the fungal tree of life. *Nature*, 474:2636–2643, 2011.

[64] S. K., R. I. Jones, and L. Arvola. Hypolimnetic phosphorus retrieval by diel vertical migrations of lake phytoplankton. *Freshwater Biology*, 14(4):431–438, 1984.

[65] D. Kaiser. Building a multicellular organism. *Annual Review of Genetics*, 35(1):103–123, 2001. PMID: 11700279.

[66] V. Kapitonov and J. Jurka. RAG1 core and V(D)J recombination signal sequences were derived from *transib* transposons. *PLoS Biology*, 3:e181, 2005.

[67] P. J. Keeling. Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany*, 91(10):1481–1493, 2004.

[68] N. King. The unicellular ancestry of animal development. *Developmental Cell*, 7(3):313 – 325, 2004.

[69] D. L. Kirk. *Volvox: molecular-genetic origins of multicellularity and cellular differentiation.* Cambridge University Press, 1998.

[70] D. L. Kirk. A twelve-step program for evolving multicellularity and a division of labor. *BioEssays*, 27(3):299–310, 2005.

[71] M. M. Kirk, A. Ransick, S. E. McRae, and D. L. Kirk. The relationship between cell size and cell fate in *Volvox carteri*. *The Journal of Cell Biology*, 123(1):191–208, 1993.

[72] R. E. Kopp, J. L. Kirschvink, I. A. Hilburn, and C. Z. Nash. The Paleoproterozoic snowball earth: A climate disaster triggered by the evolution of oxygenic photosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32):11131–11136, 2005.

[73] J. H. Koschwanez, K. R. Foster, and A. W. Murray. Improved use of a public good selects for the evolution of undifferentiated multicellularity. *eLife*, 2, 2013.

[74] V. Koufopanou. The evolution of soma in the Volvocales. *The American Naturalist*, 143(5):pp. 907–931, 1994.

[75] A. Kuhn. *Lectures on Developmental Physiology.* Springer-Verlag, 1971.

[76] B. Kunkel, R. Losick, and P. Straiger. The *Bacillus subtilis* gene for the developmental transcription factor σ_K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. *Genes & Development*, 4:525–535, 1990.

[77] C. P. Kurtzman and J. Sugiyama. Ascomycetous yeasts and yeastlike taxa. In D. J. McLaughlin, E. G. McLaughlin, and P. A. Lemke, editors, *The Mycota*. Springer-Verlag, 2001.

[78] J. Lafaille, A. DeCloux, M. Bonneville, Y. Takagaki, and S. Tonegawa. Junctional sequences of T cell receptor $\gamma\delta$ genes: implications for $\gamma\delta$ T cell lineages and for a novel intermediate of VDJ joining. *Cell*, 59:859–870, 1989.

[79] C. Lee, H. Zhang, A. E. Baker, P. Occhipinti, M. E. Borsuk, and A. S. Gladfelter. Protein aggregation behavior regulates cyclin transcript localization and cell-cycle control. *Developmental Cell*, 25:572 – 584, 2013.

[80] S. P. Leys and D. Eerkes-Medrano. Gastrulation in calcareous sponges: In search of Haeckel's Gastraea. *Integrative and Comparative Biology*, 45(2):342–351, 2005.

[81] W. F. Loomis. Role of the surface sheath in the control of morphogenesis in *Dictyostelium discoideum*. *Nature New Biology*, 240:6–9, 1972.

[82] I. Maeso, M. Irimia, J. J. Tena, F. Casares, and J. L. Gómez-Skarmeta. Deep conservation of *cis*-regulatory elements in metazoans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1632), 2013.

[83] H. Margot, C. Acebal, E. Toril, R. Amils, and J. Fernandez Puentes. Consistent association of crenarchaeal Archaea with sponges of the genus *Axinella*. *Marine Biology*, 140(4):739–745, 2002.

[84] L. Margulis. *Symbiosis in cell evolution.* W. H. Freeman, 1981.

[85] S. Masloff, S. Pggeler, and U. Kck. The *pro1*$^{+}$ gene from *Sordaria macrospora* encodes a c_6 zinc finger transcription factor required for fruiting body development. *Genetics*, 152(1):191–199, 1999.

[86] S. Mast. Reactions to light in *Volvox*, with special reference to the process of orientation. *Zeitschrift für vergleichende Physiologie*, 4(5):637–658, 1926.

[87] M. Matsuzaki, O. Misumi, T. Shin-i, S. Maruyama, M. Takahara, S.-y. Miyagishima, T. Mori, K. Nishida, F. Yagisawa, K. Nishida, Y. Yoshida, Y. Nishimura, S. Nakao, T. Kobayashi, Y. Momoyama, T. Higashiyama, A. Minoda, M. Sano, H. Nomoto, K. Oishi, H. Hayashi, F. Ohta, S. Nishizaka, S. Haga, S. Miura, T. Morishita, Y. Kabeya, K. Terasawa, Y. Suzuki, Y. Ishii, S. Asakawa, H. Takano, N. Ohta, H. Kuroiwa, K. Tanaka, N. Shimizu, S. Sugano, N. Sato, H. Nozaki, N. Ogasawara, Y. Kohara, and T. Kuroiwa. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10d. *Nature*, 428(6983):653–657, Apr 2004.

[88] J. Maynard Smith and E. Szathmáry. *The Major Transitions in Evolution*. W.H. Freeman, 1995.

[89] T. Mayorova and I. Kosevich. FMRF-amide immunoreactivity pattern in the planula and colony of the hydroid *Gonothyraea loveni*. *Zoology*, 116(1):9 – 19, 2013.

[90] D. McLaughlin, K. Esser, P. Lemke, M. Blackwell, E. McLaughlin, and J. Spatafora. *Systematics and Evolution*. Mycota : a comprehensive treatise on fungi as experimental systems for basic and applied research. K. Esser, ed. Springer, 2001.

[91] M. Medina, A. G. Collins, J. W. Taylor, J. W. Valentine, J. H. Lipps, L. Amaral-Zettler, and M. L. Sogin. Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. *International Journal of Astrobiology*, 2:203–211, 7 2003.

[92] J. C. Meeks. Symbiosis between nitrogen-fixing cyanobacteria and plants. *BioScience*, 48(4):pp. 266–276, 1998.

[93] M. Meissner, K. Stark, B. Cresnar, D. L. Kirk, and R. Schmitt. *Volvox* germline-specific genes that are putative targets of RegA repression encode chloroplast proteins. *Current Genetics*, 36(6):363–370, 1999.

[94] V. Merino-Puerto, A. Herrero, and E. Flores. Cluster of genes that encode positive and negative elements influencing filament length in a heterocyst-forming cyanobacterium. *Journal of Bacteriology*, 195(17):3957–3966, 2013.

[95] N. P. Money. Mushroom stem cells. *BioEssays*, 24(10):949–952, 2002.

[96] S. M. Morgani and J. M. Brickman. Survival of the fattest. *eLife*, 2, 2013.

[97] S. J. Morrison, N. M. Shah, and D. J. Anderson. Regulatory mechanisms in stem cell biology. *Cell*, 88(3):287 – 298, 1997.

[98] A. M. Muro-Pastor and W. R. Hess. Heterocyst differentiation: from single mutants to global approaches. *Trends in Microbiology*, 20(11):548 – 557, 2012.

[99] S. Nickell, R. Hegerl, W. Baumeister, and R. Rachel. *Pyrodictium* cannulae enter the periplasmic space but do not enter the cytoplasm, as revealed by cryo-electron tomography. *Journal of Structural Biology*, 141(1):34 – 42, 2003.

[100] C. Nielsen. Six major steps in animal evolution: are we derived sponge larvae? *Evolution & Development*, 10(2):241–257, 2008.

[101] H. Nozaki, M. Itoh, M. Watanabe, and T. Kuroiwa. Ultrastructure of the vegetative colonies and systematic position of *Basichlamys* (Volvocales, Chlorophyta). *European Journal of Phycology*, 31(1):67–72, 1996.

[102] H. Nozaki. Ultrastructure of the extracellular matrix and taxonomy of *Gonium* (Volvocales, Chlorophyta). *Phycologia*, 29:1 – 8, 1990.

[103] H. Nozaki, M. Itoh, R. Sano, H. Uchida, M. M. Watanabe, and T. Kuroiwa. Phylogenetic relationships within the colonial Volvocales (Chlorophyta) inferred from *rbcL* gene sequence data. *Journal of Phycology*, 31(6):970–979, 1995.

[104] H. Nozaki and T. Kuriowa. Ultrastructure of the extracellular matrix of *Eudorina*, *Pleodorina* and *Yamagishiella* gen. nov. (Volvocaceae, Chlorophyta). *Phycologia*, 29:1 – 8, 1990.

[105] B. Oud, V. Guadalupe-Medina, J. F. Nijkamp, D. de Ridder, J. T. Pronk, A. J. A. van Maris, and J.-M. Daran. Genome duplication and mutations in *ACE2* cause multicellular, fast-sedimenting phenotypes in evolved *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, 110(45):E4223–E4231, 2013.

[106] O. V. H. Owens and W. E. Esaias. Physiological responses of phytoplankton to major environmental factors. *Annual Review of Plant Physiology*, 27(1):461–483, 1976.

[107] J. Palis. Ontogeny of erythropoiesis. *Current Opinion in Hematology*, 15:155–161, 2008.

[108] A. R. Pengelly, O. Copur, H. Jaeckle, A. Herzig, and J. Mller. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. *Science*, 339(6120):698–699, 2013.

[109] S. Pöggeler and U. Kück. A WD40 repeat protein regulates fungal cell differentiation and can be replaced functionally by the mammalian homologue striatin. *Eukaryotic Cell*, 3(1):232–240, 2004.

[110] S. Porter. The rise of predators. *Geology*, 39(6):607–608, 2011.

[111] B. Rasmussen, I. R. Fletcher, J. J. Brocks, and M. R. Kilburn. Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature*, 455:1101–1104, 2008.

[112] W. C. Ratcliff, R. F. Denison, M. Borrello, and M. Travisano. Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences*, 2012.

[113] H. Rausch, N. Larsen, and R. Schmitt. Phylogenetic relationships of the green algae *Volvox carteri* deduced from small-subunit ribosomal RNA comparisons. *Journal of Molecular Evolution*, 29(3):255–265, 1989.

[114] C. Reynolds. *The ecology of freshwater phytoplankton.* Cambridge University Press, 1984.

[115] G.-Y. Rhee. A continuous culture study of phosphate uptake, growth rate and polyphosphate in *Scenedesmus* sp. *Journal of Phycology*, 9(4):495–506, 1973.

[116] J.-P. Rieu and H. Delanoë-Ayari. Shell tension forces propel *Dictyostelium* slugs forward. *Physical Biology*, 9(6):066001, 2012.

[117] K. D. Rinker and R. M. Kelly. Growth physiology of the hyperthermophilic archaeon *Thermococcus litoralis*: Development of a sulfur-free defined medium, characterization of an exopolysaccharide, and evidence of biofilm formation. *Applied and Environmental Microbiology*, 62(12):4478–85, 1996.

[118] R. Rippka, J. Deruelles, J. B. Waterbury, M. Herdman, and R. Y. Stanier. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *Journal of General Microbiology*, 111(1):1–61, 1979.

[119] R. W. Robinson. Life cycles in the methanogenic archaebacterium *Methanosarcina mazei*. *Applied and Environmental Microbiology*, 52(1):17–27, 1986.

[120] V. Rossetti, M. Filippini, M. Svercel, A. D. Barbour, and H. C. Bagheri. Emergent multicellular life cycles in filamentous bacteria owing to density-dependent population dynamics. *Journal of The Royal Society Interface*, 8(65):1772–1784, 2011.

[121] D. H. Rothman, G. P. Fournier, K. L. French, E. J. Alm, and C. Cao. Methanogenic burst in the end-Permian carbon cycle. *Proceedings of the National Academy of Sciences*, in press.

[122] U. Rüffer and W. Nultsch. Flagellar photoresponses of *Chlamydomonas* cells held on micropipettes: I. Change in flagellar beat frequency. *Cell Motility and the Cytoskeleton*, 15(3):162–167, 1990.

[123] B. Schirrmeister, A. Antonelli, and H. Bagheri. The origin of multicellularity in cyanobacteria. *BMC Evolutionary Biology*, 11(1):45, 2011.

[124] B. E. Schirrmeister, J. M. de Vos, A. Antonelli, and H. C. Bagheri. Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proceedings of the National Academy of Sciences*, 110(5):1791–1796, 2013.

[125] A. Sebé-Pedrés, A. J. Roger, F. B. Lang, N. King, and I. n. Ruiz-Trillo. Ancient origin of the integrin-mediated adhesion and signaling machinery. *Proceedings of the National Academy of Sciences*, 107(22):10142–10147, 2010.

[126] A. Sebé-Pedrós, M. Irimia, J. del Campo, H. Parra-Acero, C. Russ, C. Nusbaum, B. J. Blencowe, and I. naki Ruiz-Trillo. Regulated aggregative multicellularity in a close unicellular relative of Metazoa. *eLife*, 2, 2013.

[127] K. Shalchian-Tabrizi, M. A. Minge, M. Espelund, R. Orr, T. Ruden, K. S. Jakobsen, and T. Cavalier-Smith. Multigene phylogeny of Choanozoa and the origin of animals. *PLoS ONE*, 3(5):e2098, 05 2008.

[128] P. M. Shih, D. Wu, A. Latifi, S. D. Axen, D. P. Fewer, E. Talla, A. Calteau, F. Cai, N. Tandeau de Marsac, R. Rippka, M. Herdman, K. Sivonen, T. Coursin, T. Laurent, L. Goodwin, M. Nolan, K. W. Davenport, C. S. Han, E. M. Rubin, J. A. Eisen, T. Woyke, M. Gugger, and C. A. Kerfeld. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences*, 110(3):1053–1058, 2013.

[129] M. B. Short, C. A. Solari, S. Ganguly, T. R. Powers, J. O. Kessler, and R. E. Goldstein. Flows driven by flagella of multicellular organisms enhance long-range molecular transport. *Proceedings of the National Academy of Sciences*, 103(22):8315–8319, 2006.

[130] P. A. Siver and H. Shayler. Algal Ed project. `http://fmp.conncoll.edu/Silicasecchidisk/Algal-ED_finished.html`.

[131] Z. D. Smith and A. Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14:204 – 220, 2013.

[132] S. Smukalla, M. Caldara, N. Pochet, A. Beauvais, S. Guadagnini, C. Yan, M. D. Vinces, A. Jansen, M. C. Prevost, J.-P. Latgé, G. R. Fink, K. R. Foster, and K. J. Verstrepen. *FLO1* is a variable green beard gene that drives biofilm-like cooperation in budding yeast. *Cell*, 135:726–737, 2008.

[133] C. A. Solari, K. Drescher, and R. E. Goldstein. The flagellar photoresponse in *Volvox* species (Volvocaceae, Chlorophyceae)1. *Journal of Phycology*, 47(3):580–583, 2011.

[134] U. Sommer and Z. M. Gliwicz. Long range vertical migration of *Volvox* in tropical Lake Cahora Bassa (Mozambique). *Limnology and Oceanography*, 31(3):650–653, 1986.

[135] J. E. Stajich, S. K. Wilke, D. Ahrn, C. H. Au, B. W. Birren, M. Borodovsky, C. Burns, B. Canbck, L. A. Casselton, C. Cheng, J. Deng, F. S. Dietrich, D. C. Fargo, M. L. Farman, A. C. Gathman, J. Goldberg, R. Guig, P. J. Hoegger, J. B. Hooker, A. Huggins, T. Y. James, T. Kamada, S. Kilaru, C. Kodira, U. Kes, D. Kupfer, H. Kwan, A. Lomsadze, W. Li, W. W. Lilly, L.-J. Ma, A. J. Mackey, G. Manning, F. Martin, H. Muraguchi, D. O. Natvig, H. Palmerini, M. A. Ramesh, C. J. Rehmeyer, B. A. Roe, N. Shenoy, M. Stanke, V. Ter-Hovhannisyan, A. Tunlid, R. Velagapudi, T. J. Vision, Q. Zeng, M. E. Zolan, and P. J. Pukkila. Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proceedings of the National Academy of Sciences*, 107(26):11889–11894, 2010.

[136] A. Stechmann and T. Cavalier-Smith. Rooting the eukaryote tree by using a derived gene fusion. *Science*, 297(5578):89–91, 2002.

[137] G. Steinberg. Hyphal growth: a tale of motors, lipids, and the Spitzenkörper. *Eukaryotic Cell*, 6(3):351–360, 2007.

[138] K. O. Stetter. Ultrathin mycelia-forming organisms from submarine volcanic areas having an optimum growth temperature of 105°C. *Nature*, 300:258–260, 1982.

[139] K. O. Stetter. Smallest cell sizes within hyperthermophilic Archaea ("archaebacteria"). In A. Knoll and M. J. Osborn, editors, *Size limits of very small microorganisms*. National Academies Press, 1999.

[140] J. E. Strassmann, Y. Zhu, and D. C. Queller. Altruism and social cheating in the social amoeba *Dictyostelium discoideum*. *Nature*, 408:965–967, 2000.

[141] H. Sugimoto and H. Endoh. Differentially expressed genes during fruiting body development in the aggregative ciliate *Sorogena stoianovitchae* (ciliophora: Colpodea). *Journal of Eukaryotic Microbiology*, 55(2):110–116, 2008.

[142] A. Tero, S. Takagi, T. Saigusa, K. Ito, D. P. Bebber, M. D. Fricker, K. Yumiki, R. Kobayashi, and T. Nakagaki. Rules for biologically inspired adaptive network design. *Science*, 327(5964):439–442, 2010.

[143] C. R. Thompson and R. R. Kay. Cell-fate choice in *Dictyostelium*: Intrinsic biases modulate sensitivity to DIF signaling. *Developmental Biology*, 227(1):56 – 64, 2000.

[144] C. H. Waddington. Canalization of development and the inheritance of acquired characters. *Nature*, 150:563–565, 1942.

[145] E. Wallraff and H. G. Wallraff. Migration and bidirectional phototaxis in *Dictyostelium discoideum* slugs lacking the actin cross-linking 120 kDa gelation factor. *Journal of Experimental Biology*, 200(24):3213–20, 1997.

[146] A. E. Walsby. The permeability of heterocysts to the gases nitrogen and oxygen. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 226(1244):pp. 345–366, 1985.

[147] A. Weismann. *The Germ-Plasm: A Theory of Heredity*. Charles Scribner's Sons, 1893.

[148] D. Wessels, D. R. Soll, D. Knecht, W. F. Loomis, A. D. Lozanne, and J. Spudich. Cell motility and chemotaxis in *Dictyostelium* amoebae lacking myosin heavy chain. *Developmental Biology*, 128(1):164 – 177, 1988.

[149] M. Wilkins and K. Williams. The extracellular matrix of the *Dictyostelium discoideum* slug. *Experientia*, 51(12):1189–1196, 1995.

[150] K. L. Williams, P. H. Vardy, and L. A. Segel. Cell migrations during morphogenesis: Some clues from the slug of *Dictyostelium discoideum*. *BioEssays*, 5(4):148–152, 1986.

[151] C.-C. Zhang, S. Laurent, S. Sakr, L. Peng, and S. Bdu. Heterocyst differentiation and pattern formation in cyanobacteria: a chorus of signals. *Molecular Microbiology*, 59(2):367–375, 2006.

Chapter 2

Design and characterization of a yeast model system for cellular differentiation

"What I cannot create, I do not understand."

- Richard Feynman (attrib.)

2.1 Abstract

Cellular differentiation is a complex trait whose molecular underpinnings have been the subject of decades of developmental biology research. Such investigations have resulted in a clear philosophical definition of differentiation as well as a catalogue of the required features and their implementations in extant organisms. We herein describe the design and characterization of a synthetic gene construct in yeast which achieves irreversible differentiation through recombinase-mediated gene excision. We show that our system allows fundamental properties like conversion rate, cell-specific gene expression, and relative growth rate to be readily modified. To demonstrate the predictability of our system's behavior, we recapitulate phenomena predicted by common mutation-selection balance models.

2.2 Introduction

Cellular differentiation – a long-term change in a cell's gene expression state – has evolved independently over a dozen times in nature. Despite this convergence in function, the mechanisms employed to implement irreversible cell fate decisions are varied and complex. Attempts to recreate differentiation *ab initio* have thus far been fraught with noise, instability, and intransigence to future modifications. We sought to construct a model system where the conditions and consequences of differentiation are, to the greatest extent possible, under the experimenter's control. Our system will find direct application in chapter three: here, we explore its design and characterization as matters of intrinsic interest from a synthetic biology perspective.

Most eukarytoic cell types are established and maintained by groups of interregulated transcription factors and promoters called gene regulatory networks [18]. As a consequence of inheriting cytoplasmic factors, receiving intercellular signals, or even experiencing random fluctuations in gene expression, a differentiating cell will begin to express a combination of transcription factors characteristic of the cell type it will become. These transcription factors then maintain their own expression indefinitely through direct or indirect regulatory connections, suppress alternative cell fates, and orchestrate the expression of effector genes specific to the differentiated cell type. Cell fate may then be further reinforced by epigenetic changes in chromatin state and/or autocrine and paracrine signaling.

Gene regulatory networks are often composed of dozens of nodes: while smaller motifs, including positive feedback and antagonism between states, can produce the required bistability [82], the incorporation of redundancy ensures the maintenance of the differentiation state in the face of inherent noise in transcriptional and translational regulation [51, 86]. In synthetic systems that use only bare-bones regulation systems (e.g., a single feedback loop), noisy gene expression and transcription factor binding, insufficient cooperativity in gene regulation, or weak induction/repression can permit undesired interconversion between states or produce an effectively monostable system [26, 35].

Distinct gene expression states can also be produced by heritable cytoplasmic determinants. Yeast that possess the *[PSI+]* prion, a self-templating protein aggregate of the translation termination factor Sup35 [19, 72, 76], read through stop codons at a higher rate (0.2 - 35% of translations, depending on the strain and codon context) and therefore have a dramatically different proteome from wildtype yeast [8, 22, 58, 79, 89]. Under normal conditions, the number of aggregates per cell is large enough to ensure that most daughters will inherit at least one prion and therefore maintain the *[PSI+]* phenotype [16, 50, 80, 81]. Prion loss occurs only when chaperone activity is altered significantly: overexpression of heat shock proteins can cause the breakdown of aggregates, whereas a decrease in chaperone activity results in fewer, larger aggregates per cell and therefore decreases the probability of inheritance by daughter cells [59, 60].

A third mechanism of differentiation was first hypothesized by August Weismann in his treatise *The Germ-Plasm* [88], where he postulated that the hereditary material ("idioplasm") is partitioned between daughter cells during somatic divisions in the embryo: according to his theory, differentiation occurred when the remaining genetic material was only sufficient to encode one cell type. Although we now know that genome reduction is not the primary means of differentiation in most organisms, it remains the most definitive, for reversion to an undifferentiated state is impossible once the necessary genomic material is lost. Examples of differentiation through gene excision can be found in a variety of taxa, where they have apparently resulted from many independent co-options of tranposable elements and their recombinases [3, 37, 42]. Gene excision is required for the generation of functional T-cell receptors and immunoglobulins (an essential step of lymphocyte development), while complete enucleation is a characteristic event in mammalian erythropoiesis [7, 43, 63]. Production of nitrogen-fixing cells in heterocystous cyanobacteria requires the excision of two transposons in the *nif* gene cluster [3]. During sporulation, *B. subtilis* mother cells undergo gene excision to produce functional σ_K, their cell type-specific RNA polymerase II sigma fac-

tor [42]. Thus there is substantial biological precedent for differentiation through genome reduction.

We chose gene excision via the commonly-used bacteriophage P1 Cre recombinase [74] as the molecular basis for differentiation within our model system (Figure 2.1A). Cre recombinase binds as a dimer to a 34-bp directional sequence loxP [1]: once two Cre dimers have each bound a loxP site and to one another, they induce recombination between the two loxP sites [27, 83]. When the loxP sites were originally aligned end-to-end, this results in excision of the intervening region as a small circular fragment that will be lost by dilution if it lacks an origin of replication. In our system, "unconverted" cells express a gene encoded between two loxP sites. Conversion occurs when the recombinase excises the intervening region, permanently halting expression of the removed open reading frame and allowing expression of another transcript to begin. The cell type-specific genes are chosen so that unconverted cells have the higher growth rate (due to improved tolerance of the antibiotic cycloheximide) and each cell type can be readily visualized by fluorescence (Figure 2.1B).

The field of synthetic biology is guided by the principle that the elemental components of life act modularly and can therefore be recombined in appropriate ways to achieve new functionality. While great strides towards this ideal have been made [11, 25, 73], rational design remains a lofty ideal given the current depth of understanding of the underlying molecular biology [1]. Close scrutiny is thus required to ensure the proper function of a synthetic biological system, regardless of how straightforward its implementation may appear. The most direct test of a new synthetic system is its ability to recapitulate behaviors predicted by theory, simulation, or orthogonal experimental approaches: comparisons between the expected and observed activity of our system will therefore form the bulk of this chapter.

[1] For example, a common bioengineering problem is to express a gene in a non-native condition, and the logical approach is to replace its transcriptional promoter with that of a gene expressed at the correct place and time. But even this simple strategy is not guaranteed, for the promoter's function may depend heavily on its genomic context, such as *cis*-regulatory elements or nucleosome positioning sites in the adjacent sequence.

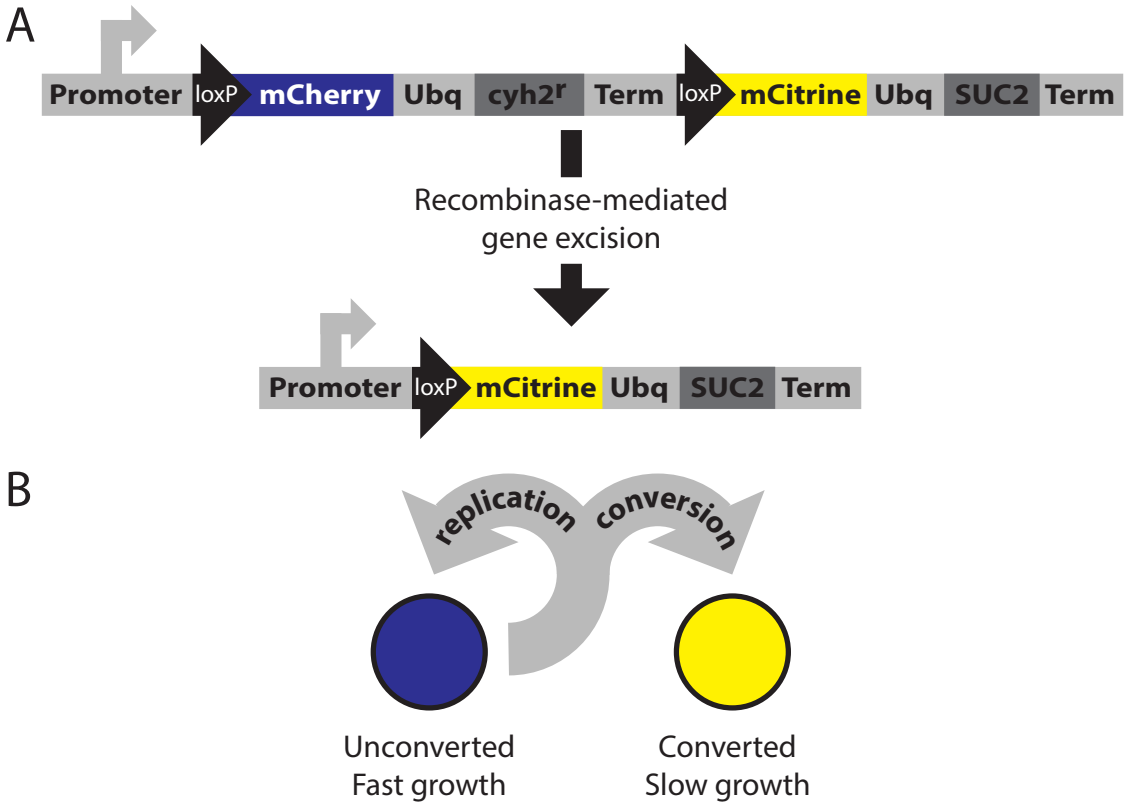

Figure 2.1: Schematic of the yeast model system.

(A) Diagram of the locus conferring cell type-specific gene expression. In unconverted cells, a strong promoter (P_{ENO2}) drives expression of the polypeptide mCherry-Ubq-cyh2^r; a transcriptional terminator prevents expression of downstream genes. After translation, the polypeptide is cleaved at its ubiquitin moiety by cellular proteases to release a stable fluorescent protein-ubiquitin fusion and a mutant version of the ribosomal protein L28, Cyh2^r, in its native sequence. The Cyh2^r protein is not able to bind cycloheximide, and thus confers a growth advantage in media containing this drug. Conversion occurs via gene excision: mCherry-Ubq-cyh2^r expression halts and mCitrine-Ubq-Suc2 expression begins: this polypeptide encodes a second fluorescent marker and the yeast invertase Suc2 (used in this context to balance the fitness effects of cyh2^r expression).

(B) Diagram of functionality conferred to each cell type. Unconverted cells express the fluorescent marker mCherry (pseudo-colored blue throughout this document) and have relatively fast growth in cycloheximide due to cyh2^r expression. Cre recombinase activity in newborn daughters of unconverted cells may induce their conversion, which results in expression of the yellow fluorescent marker mCitrine and slower growth in cycloheximide.

Mutation-selection balance theory, which has been applied to diverse problems ranging from the estimation of mutation rates [17, 30] to eugenics and medical policy [17, 29, 33, 55, 56, 65], provides several testable predictions of the behavior of our strains. For the haploid, single-locus case most relevant to our system, the simplest mutation-selection balance model proposes that wildtype cells undergo one type of irreversible deleterious mutation at a rate μ to produce mutant clones with selection coefficient s, which is taken to be negative. The fraction f of wildtype cells at equilibrium reflects the ability of the population to effectively eliminate the continually-spawning mutants through selection [17]. The process of conversion in our system is akin to mutation, in that it produces a heritable change in gene expression state; differences in gene expression decrease the growth rate of converted cells under appropriate media conditions, producing an effective selection coefficient. Under this model, J. B. S. Haldane showed that the reduction in population fitness at mutation-selection equilibrium is a function solely of the conversion rate:

> It is at once clear that in equilibrium abnormal genes are wiped out by natural selection at exactly the same rate as they are produced by mutation. It does not matter whether the gene is lethal or almost harmless. [...] The loss of fitness of the species depends entirely on the mutation rate and not at all on the effect of the gene upon the fitness of the individual carrying it.

This decrease in population fitness was later deemed the mutation "load" by Muller [56], who arrived at the result independently. Later in this chapter, we show that, as predicted, the mutation load in our strains varies with the conversion rate but not significantly with the selection coefficient.

Due to the irreversibility of mutations assumed in this model, it is possible to lose the wildtype phenotype indefinitely; i.e., $f = 0$ is an absorbing state. The average survival time of mutant clones diverges ($f \to 0$) when the mutation rate μ is too high or the selection coefficient s is too low for selection to be effective. The sudden divergence in average survival time is analogous to a phase transition. These characteristics make mutation-selection balance a

member of a universality class (a group of fundamentally-analogous models) of nonequilibrium critical phenomena called directed percolation models ([9], reviewed in [31]), which include models of fluid percolation through porous materials, traffic jams, certain catalytic chemical reactions, and epidemic spread. Each model in this class specifies units that can exist in two states, one of which is absorptive. Results obtained for this universality class in general give insight into mutation-selection balance as a special case.

For example, the work of Domany and Kinzel on directed percolation can be used to predict the shape of the phase transition line for mutation-selection balance during growth on solid media [46]. The production of new cells at the linear front of a colony can be modeled by the iterative propagation of cellular automata. By simulating directed percolation in cellular automata with varying model parameters, Domany and Kinzel characterized the phase transition at which the absorbing state prevails [20, 40]. Lavrentovich, Korolev, and Nelson have adapted this model to mutation-selection balance, accommodating important features of biological systems such as the expanding front of radially-growing colonies and explicitly predicting the phase transition diagram in terms of μ and s: as they note, the predicted threshold for wildtype cell loss occurs at much lower mutation rates/higher selection coefficients for populations growing on solid media than in well-mixed liquid media [46]. Later in this chapter, we will explore the fit between their predictions and a phase diagram obtained experimentally using our converting yeast strain.

2.3 Methods

Strain and plasmid construction

Synthesis of the cell type specification locus

loxP, the binding site of Cre recombinase, is 34 bp long and contains inverted repeats that can interfere with polymerase chain reactions when they occur at the boundaries of amplicons (and thus within primer sequences). We reasoned that inclusion of this inverted repeat within a transcript could also form a hairpin or otherwise interfere with translation of the transcript. To facilitate cloning and improve the odds of successful protein expression, we modified a yeast artificial intron [91] to contain a *loxP* site between the 5' splice donor site and the branch point. This artificial intron was introduced at the 5' end of the open reading frame of yeast-optimized version of mCherry [described in [68]; generously provided with *ADH1* terminator by Nick Ingolia] via overhang PCR, then cloned into pFA6a-HIS3MX6 [49] by sequence ligation-independent cloning (SLIC, [47]) to generate pMEW56. The constitutive *ACT1* promoter was then fused to the 5' end of the artificial intron by ligation-independent cloning, producing pMEW61. A ubiquitin moiety amplified from the *UBI4* locus was then introduced between mCherry and the *ADH1* terminator by ligation-independent cloning to generate pMEW72.

The artificial intron described above (excluding the 5' splice donor site) was fused to the 5' end of the *SUC2* open reading frame and terminator by overhang PCR and introduced into pRS402 [71] by NotI/SacI restriction digest and ligation to generate pMEW54. The P_{ACT1}-*AI-mCherry*-T_{ADH1} construct was amplified from pMEW61 by PCR to introduce XmaI and NotI restriction sites, with which the construct was introduced 5' to the *AI-SUC2*-T_{SUC2} insert on pMEW54 to produce pMEW63.

Our application requires single-copy genomic integration of a DNA fragment too large for efficient and accurate amplification by PCR: we therefore designed a plasmid whose insert, once released by restriction digest, could integrate at the *SUC2* locus via homologous re-

combination [62]. A 300-bp region 5' to the *SUC2* open reading frame was amplified via PCR and introduced 5' to the *ADE2* cassette of pRS402 by sequence ligation-independent cloning to generate pMEW71. The P_{ACT1}-*AI-mCherry-UBQ-AI-SUC2-*T_{SUC2} cassette of pMEW63 was extracted by XmaI/SacI digest and inserted into pMEW71: the resulting plasmid pMEW73, once digested with Bsu36I and SacI, has insert ends homologous to the 5' and 3' ends of the *SUC2* locus and thus can be used for efficient integration.

An intron-free version of the ribosomal protein-encoding gene *CYH2* containing a previously-described N38K mutation conferring cycloheximide resistance [23, 39, 75] was amplified by fusion PCR and integrated at the C terminus of the P_{ACT1}-*AI-mCherry-UBQ* construct in pMEW61 by ligation-independent cloning to produce pMEW67. To increase the expression level and improve consistency in expression during transition from log to stationary phase, the *ACT1* promoter of pMEW73 was replaced by P_{ENO2} through PCR amplification and restriction digest, creating plasmid pMEW79. Replacement of P_{ACT1}-*AI-mCherry-*T_{ADH1} in pMEW79 with P_{ENO2}-*AI-mCherry-UBQ-cyh2*^{r}-T_{ADH1} by restriction digest and ligation was used to create pMEW82.

To allow fluorescent labeling of converted cells, the yeast-optimized version of mCitrine [69] was then introduced 5' to the *SUC2* open reading frame with a ubiquitin linker as follows. The mCherry open reading frame of pMEW67 was replaced with the mCitrine open reading frame by ligation-independent cloning to generate pMEW83. The SUC2-T_{SUC2} fragment was then introduced at the C terminus of the *AI-mCitrine-UBQ* insert to produce pMEW84, and the resulting *AI-mCitrine-UBQ-SUC2-*T_{SUC2} cloned into pMEW82 by restriction digest to create pMEW90.

Yeast strains and culture media

All yeast strains were constructed in the W303 background (*ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100*) containing the S288c (functional) *BUD4* allele [85], which was introduced by a *URA3* loopout strategy [28]. Integration of plasmid fragments and PCR amplicons by

homologous recombination was confirmed by diagnostic PCR across the insertion boundaries and (when insert length permitted) across the complete insertion cassette.

All media components used in this research were purchased from Sigma-Aldrich (http://www.sigmaaldrich.com/). Yeast extract peptone dextrose (YPD) media were prepared using a standard recipe with 2% w/v dextrose [10]. Minimal media and dropout media were produced using a Yeast Nitrogen Base recipe modified to minimize autofluorescence and eliminate undesired carbon sources [28]. Cycloheximide and β-estradiol were resuspended at 1 mM in EtOH, aliquoted, and stored less than one year at -80°C.

Liquid cultures were grown at 30°C on roller drum. Colony assays on plates were performed with 1% agar medium to facilitate lateral colony spread. Four inocula of 0.5 μL of saturated culture were pipetted onto agar plates with equal spacing, allowed to dry, then placed in a box containing an open beaker of water to minimize desiccation during a five-day growth period at 30°C [57].

Fitness and steady-state ratio assays

Fitness assays were made as described previously [45, 77]. Two strains pre-grown in log phase in like media were combined at a known ratio and passaged, maintaining log phase growth by frequent dilution and collecting samples for flow cytometry at regular timepoints. Linear regression was used to estimate the (negative) selection coefficient s according to the relation:

$$\frac{P_1}{P_2}(t) = \frac{P_1(0)\,2^{-\gamma t}}{P_2(0)\,2^{-(\gamma+s)t}} = \frac{P_1}{P_2}(0)\,2^{st} \implies \log\left[\frac{P_1}{P_2}(t)\right] = c + st\log 2$$

where the elapsed time t is measured in units of generations of the reference strain, determined by maintaining a pure culture in parallel and measuring cell density on a Beckman Coulter counter before and after each dilution and flow cytometry timepoint. Steady-state

Table 2.1: Strains used in Chapter 2

Strain name	Genotype	Plasmids used
yMEW17	*ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100 bar1 BUD4*	-
yMEW106	*ade2-1 HIS3::P_{ACT1}-GAL3 leu2-3 trp1-1 ura3-1 can1-100 bar1 BUD4*	-
yMEW109	*ade2-1 HIS3::P_{ACT1}-GAL3 leu2-3 trp1-1 ura3-1 can1-100 bar1 BUD4 HPHMX4@gal1-10*	-
yMEW118	*ade2-1 HIS3::P_{ACT1}-GAL3 leu2-3 trp1-1 ura3-1 can1-100 bar1 BUD4 URA3::KANMX6@suc2 HPHMX4@gal1-10*	-
yMEW131	*ade2-1 HIS3::P_{ACT1}-GAL3 leu2-3 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@ho URA3::KANMX6@suc2 HPHMX4@gal1-10*	pDL12
yMEW139	*ade2-1 HIS3::P_{ACT1}-GAL3 leu2-3 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@ho ADE2::P_{ENO2}-AI-ymCherry-T_{ADH1}-AI-SUC2@suc2 HphMX4@gal1-10*	pMEW79, pDL12
yMEW141	*ade2-1 HIS3::P_{ACT1}-GAL3 leu2-3 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@ho ADE2::P_{ENO2}-AI-ymCherry-UBQ-T_{ADH1}-AI-SUC2@suc2 HphMX4@gal1-10*	pDL12, pMEW82
yMEW145	*ade2-1 his3-11 leu2-3 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO*	pDL12
yMEW151	*ade2-1 his3-11 HPHMX4::P_{ACT1}-yCerulean@LEU2 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO*	pDL12
yMEW163	*ade2-1 his3-11 HPHMX4::P_{ACT1}-yCerulean@LEU2 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2*	pMEW90, pDL12
yMEW187	*ade2-1 HIS3 HPHMX4::P_{ACT1}-yCerulean@LEU2 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2@SUC2*	pMEW90, pDL12
yMEW192	*ade2-1 HIS3 HPHMX4::P_{ACT1}-yCerulean@LEU2 TRP1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2@SUC2*	pMEW90, pDL12

ratio assays were performed analogously, concluding when cell type ratio did not change significantly between at least two consecutive timepoints.

Flow cytometry

Flow cytometry samples were prepared by diluting YPD cultures in phosphate-buffered saline (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4) to minimize media fluorescence and halt growth without inducing osmotic stress. Measurements were collected on a Becton, Dickinson, and Company LSRFortessa flow cytometer. Flow cytometry data were analyzed in Matlab using custom-written scripts (see Appendix A) to identify the single cell population by its scattering profile and to compare subpopulations with fluorescence above and below defined thresholds.

Imaging and flow chamber recording

Still images and movies were collected at room temperature using a Nikon Eclipse Ti-E inverted microscope with a 20x objective lens and a Photometrics CoolSNAP HQ camera (Roper Scientific). The image processing program Fiji [66] was used to produce pseudocolored fluorescence images and movies.

Timelapse movies of monolayer growth were collected using a CellASICs Y04C ONIX Live Cell Imaging microfluids flow chamber pre-treated by perfusion of concanavalin A solution (1 mg/mL concanavalin A, 1 mM $CaCl_2$, 10 mM Na_2HPO_4 at pH 6) for 5 minutes at 2 psi, followed by media washout for 5 minutes at 2 psi. Cells were loaded into the chamber at 5 psi for 10 seconds. Metamorph 7.7 (Molecular Devices) with Nikon Perfect Focus System was used to acquire images at multiple stage positions at regular intervals and fixed exposure times. For movies, fluorescence and differential interference contrast images were collected every fifteen minutes for 24 hours. Colony images were collected on a Zeiss Lumar stereoscope.

2.4 Results

Our design uses a single locus to specify cell type, but requires two functions, growth stimulation (cycloheximide tolerance) and a cellular reporter (mCherry fluorescence) for unconverted cells: it was thus necessary to encode the fluorescent marker mCherry and the cycloheximide resistance-conferring protein Cyh2^{r} within a single transcript. Translation initiation in eukaryotes typically occurs via ATG scanning by the 43S ribosomal complex starting from the 5' cap of the transcript [38]: in *S. cerevisiae*, initiation of translation at multiple sites on a single mRNA is relatively rare [67]. Internal ribosome binding site (IRES) elements that recruit ribosomes to secondary initiation sites within the transcript, first recognized in polycistronic viral RNA [36, 64], have been employed modularly with great success in some eukaryotes [15, 53]; unfortunately, the IRES elements available in yeast are more context-dependent and produce significantly-reduced expression from the second open reading frame relative to the first [54, 78]. We therefore chose to encode all cell type-specific proteins as a single fusion peptide rather than polycistronically.

Fusion proteins are normally constructed by introducing a linker sequence of amino acids between the open reading frames of the two component proteins. While the linker is usually chosen to be unstructured and flexible (e.g. poly-glycine), it still may not be possible for each domain to fold, localize, or function in its usual way: fusion peptide design can therefore require multiple rounds of trial and error [90]. We desired a system where the cell type-specific proteins could be readily swapped out as needed without the requirement to troubleshoot new fusion proteins with each modification. We therefore introduced a linker between domains that would be proteolytically cleaved, releasing the individual proteins for natural folding, transport, and function. We chose to use ubiquitin (Ubq) as the linker because its C terminus is recognized by the endogenous UBP family of proteases [5] and is fully removed from the C-terminal peptide fragment during cleavage, restoring the C-terminal protein's native sequence and thus minimizing the potential for functional disruption [4].

The fusion peptide mCherry-Ubq-cyh2^r produced a fluorescence level similar to mCherry alone, suggesting that inclusion of a ubiquitin moiety in the fusion peptide did not affect its function or degradation rate significantly (Figure 2.2). The cycloheximide tolerance conferred by the fusion peptide was also comparable to that of cyh2^r expressed alone (data not shown). Similarly, cells expressing the fusion peptide mCitrine-Ubq-Suc2 displayed mCitrine fluorescence and were able to grow on minimal media containing sucrose as the sole carbon source (Figure 2.3), suggesting that the resulting Suc2 protein entered the secretory pathway, which would likely not be functional if cleavage of the ubiquitin moiety had failed and the N-terminal signal sequence had remained unexposed [12, 87].

Alternative splicing in *S. cerevisiae* is rare: as of this writing, only six genes are known to contain two introns [6] and no alternative splicing with exon exclusion has been previously described. However, to further limit the potential for alternative splicing in our construct, we have included the *CYH2* transcriptional terminator upstream of the second 3' splice site. With this preventative measure, expression of a fluorescent reporter (ymCitrine) from the second open reading frame was undetectable relative to a YFP$^-$ control prior to conversion (data not shown).

In our system, conversion is effected by Cre recombinase-mediated gene excision. Our application requires the ability to tune conversion rate over a wide range of values, ranging from undetectable basal activity levels (to allow culture propagation without conversion) to high inducible rates $\mu \sim 0.3$ conversions per division. To permit changes in conversion rate without requiring genomic modifications, we utilized a Cre recombinase-estrogen binding domain fusion construct (Figure 2.4) previously described by Lindstrom and Gottschling [48]. The nuclear localization sequence on the estrogen binding domain is obscured by bound heat shock proteins in the absence of hormone, and therefore this fusion protein is normally retained in the cytoplasm, away from Cre's genomic targets. Binding of the estrogen analog β-estradiol reduces heat shock protein binding, thus permitting entry into the

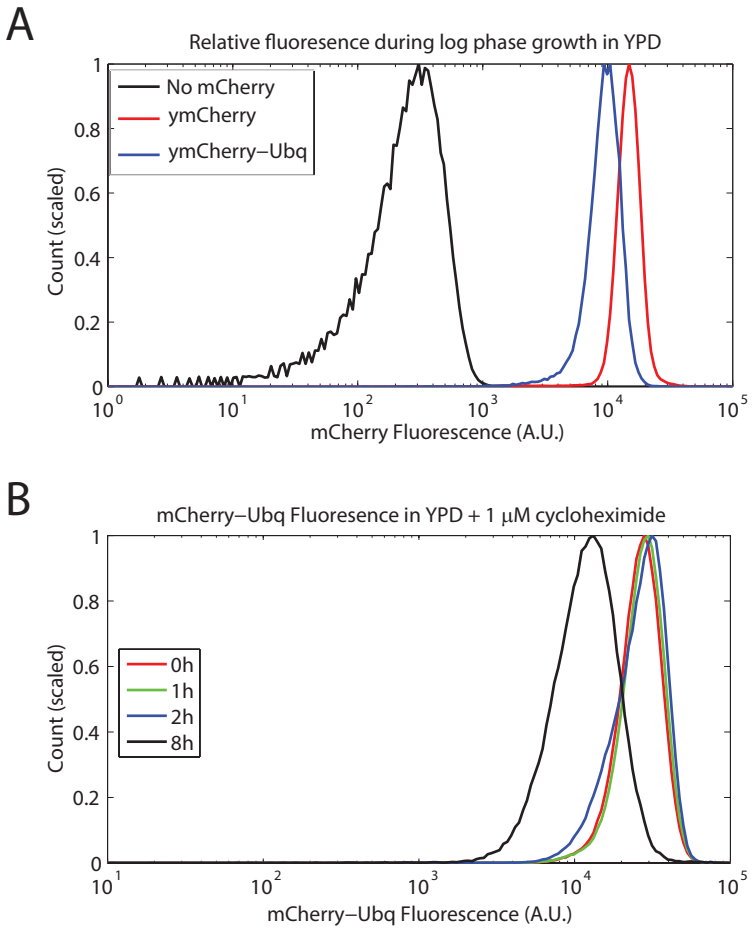

Figure 2.2: C-terminal ubiquitin moiety does not reduce stability of mCherry

(A) Strains containing no fluorescent protein (yMEW17), mCherry (yMEW139), and mCherry-Ubq (yMEW141) were grown in log phase for >24 h in YPD prior to mCherry fluorescence measurement by flow cytometry. The steady-state fluorescence level of mCherry-Ubq was found to be modestly lower than that of mCherry, suggesting sufficient accumulation and activity.

(B) A strain expressing mCherry-Ubq (yMEW141) was grown in log phase for >24 h in YPD prior to addition of 1 μM cycloheximide to halt translation. mCherry fluorescence was then assayed at several timepoints to estimate degradation rate. An approximately four-fold decrease in median fluorescence level was observed over an eight-hour interval, suggesting a low degradation rate relative to dilution during log phase growth.

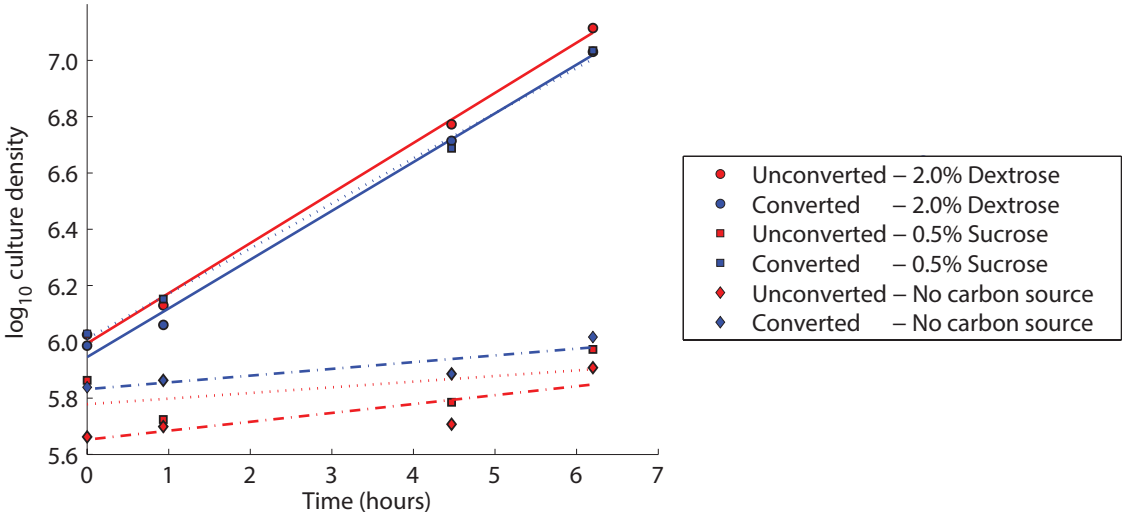

Figure 2.3: Converted cells can use sucrose as a carbon source.

Representative result of growth curve comparing unconverted and converted cultures grown on media with differing carbon sources. An unconverted culture (yMEW163) and a culture derived from a single converted isolate (yMEW163c) were pre-grown >24 h in log phase in 2% dextrose minimial media, washed twice with phosphate-buffered saline, and resuspended in the indicated media. Following a four hour acclimation period, culture densities were measured by Coulter counting at several timepoints and the growth rates determined by linear regression of log cell density vs. time. Best estimate doubling times were as follows: in 2% glucose, 1.62 hours for unconverted cells and 1.63 hours for converted cells; in 0.5% sucrose, 15.02 hours for unconverted cells and 1.82 hours for converted cells; with no known carbon source present, 9.52 hours for converted cells and 12.55 hours for unconverted cells.

nucleus and the onset of recombinase activity [13]. This fusion protein is expressed from the promoter of *SCW11*, a cell wall degrading-enzyme produced transiently in newborn daughter cells under mitotic exit network regulation [14, 21]. Lindstrom et al. also subjected the P_{SCW11}-*Cre-EBD* construct to several rounds of PCR mutagenesis to further lower its basal activity level.

We found that in the absence of inducer, the conversion rate was sufficiently low to maintain pure populations of unconverted cells (Figure 2.5A). After induction of conversion by β-estradiol addition, converted cells expressing ymCitrine began to appear in the population (Figure 2.5B). Loss of mCherry fluorescence in converted cells was gradual and

Figure 2.4: The Lindstrom-Gottschling Cre recombinase construct.

Lindstrom and Gottschling fused the open reading frame of Cre recombinase to the estrogen binding domain of the estrogen receptor to prevent its entry into the nucleus in the absence of inducer (β-estradiol). A promoter with activity limited to a brief interval after daughter cell formation (P_{SCW11}) ensures low basal activity and prevents continued conversion of cultures at rest in stationary phase. PCR mutagenesis of the construct further limited basal activity, presumably by reducing protein stability [48].

likely due to dilution by cell growth and division (Figure 2.5C-E), consistent with the high stability of mCherry-Ubq measured previously (Figure 2.2B). Following β-estradiol washout and continued growth, converted cells formed an easily-distinguished mCherry$^-$ mCitrine$^+$ population (Figure 2.5F).

The process of conversion could also be visualized during microcolony growth. Unconverted cells were fixed in a flow chamber and maintained in media containing a high concentration of β-estradiol (1 μM) for 24 hours. These movies showed that expression of mCitrine began shortly after conversion, with mCherry expression (primarily limited to the vacuoule) fading over time (Figure 2.6A-D).

Since Cre activity in our system requires inducer-dependent nuclear localization, we expected that conversion rate should be a function of β-estradiol concentration. To determine the attainable range of conversion rates, we transferred pure cultures of our converting strain to media containing different concentrations of β-estradiol and performed flow cytometry on samples collected at several timepoints during culture growth. The fraction of unconverted (mCherry$^+$ mCitrine$^-$) cells remaining decreased exponentially with time as predicted for cells converting at a fixed rate per generation (Figure 2.7A). The attainable conversion rates were found to span several orders of magnitude, ranging from unmeasurable on experimental timescales in the absence of inducer to $\mu \approx 0.3$ at the highest concentration

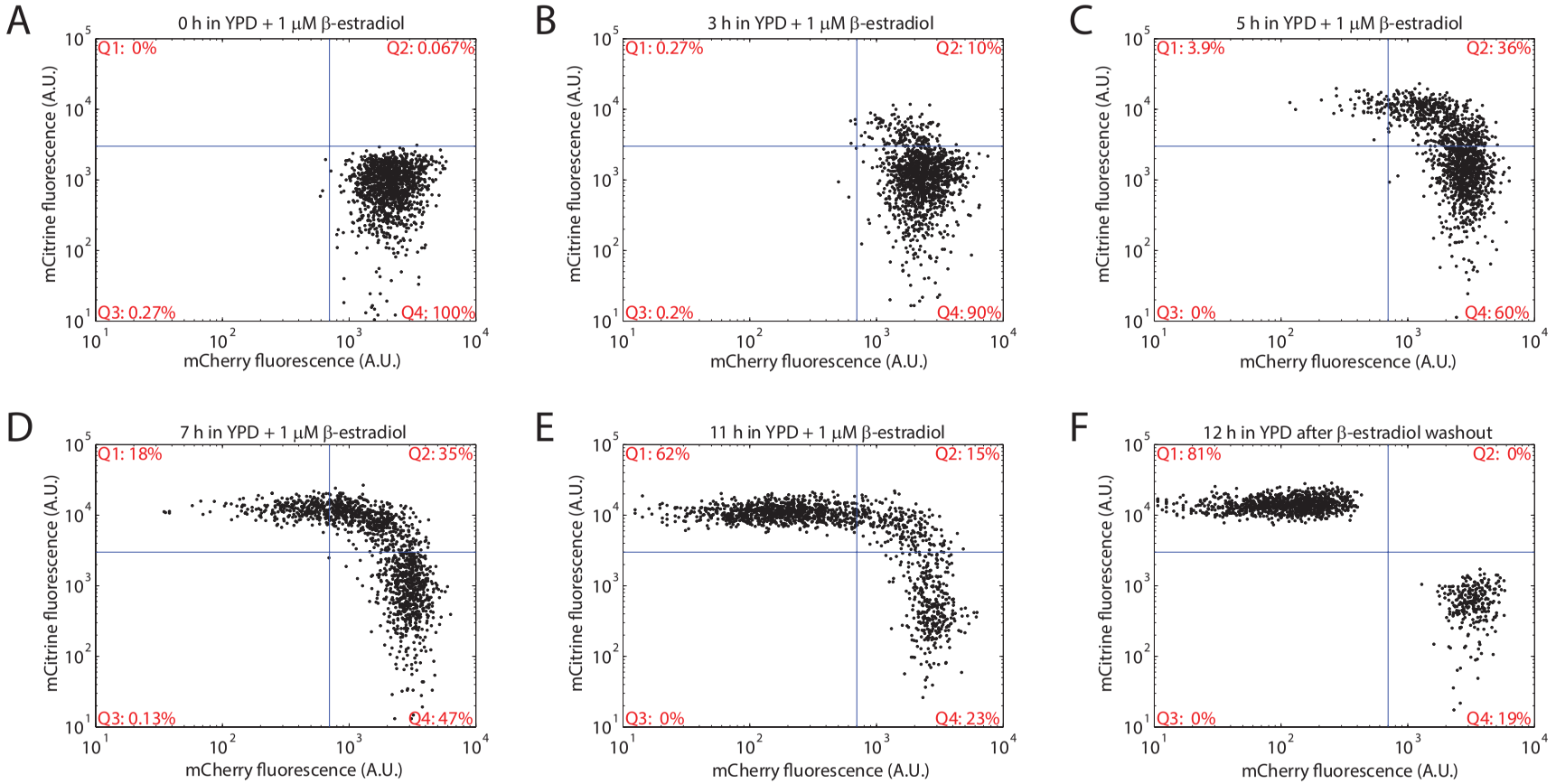

Figure 2.5: Timecourse of conversion following β-estradiol addition.

A converting strain (yMEW192) was grown in log phase in YPD for > 24 h prior to β-estradiol addition. Log phase growth was maintained during culturing in 1 μM β-estradiol medium while samples were collected at multiple timepoints for flow cytometry (A-E). After 11 hours, cells were washed twice with phosphate-buffered saline (PBS) and resuspended in YPD, preventing further conversion and allowing the dilution of mCherry/degradation of mCherry for a further 12 hours (F).

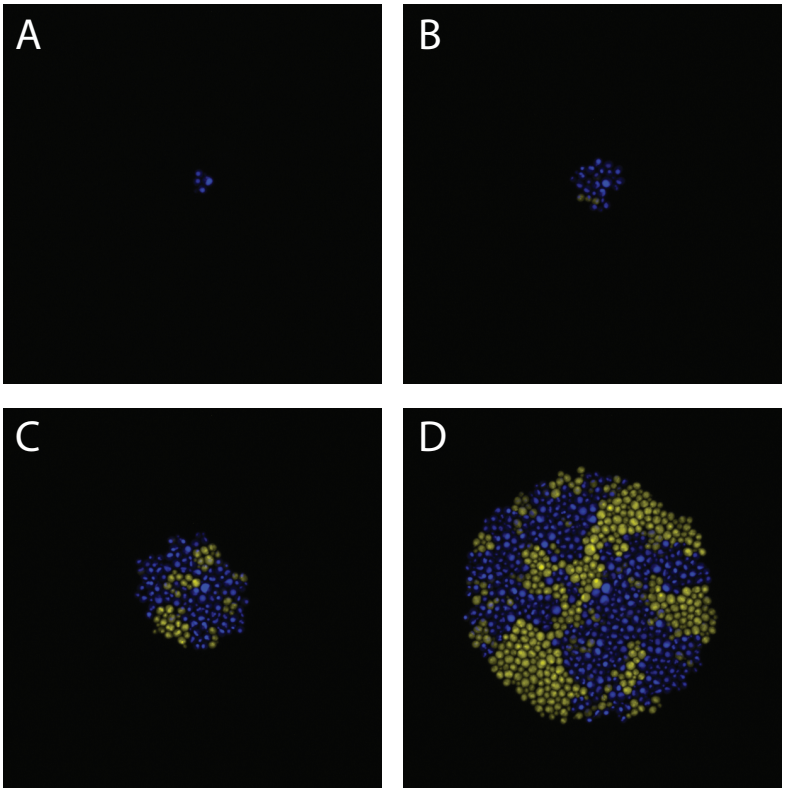

Figure 2.6: Still images of microcolony formation during conversion.

Still images of a monolayer microcolony originating from a single colony (A) after 2 hours, (B) 7 hours, (C) 12 hours, and (D) 17 hours after addition of 1 μM β-estradiol. mCherry fluorescence (present in unconverted cells and recently-converted cells) is pseudocolored blue; mCitrine fluorescence is shown in yellow.

permissible for normal culture growth (Figure 2.7B). This maximum conversion rate was somewhat lower than previously reported [48], but we note that the distance between our loxP sites and their genomic context may account for this difference [32, 92].

These conversion rate measurements assume that conversion is irreversible. We justify this assumption on the basis that the circular excised fragment does not contain an origin of replication and thus is expected to be inherited by only one descendant of each conversion event. Furthermore, while re-integration of the circular fragment through further recombinase activity is possible [24], the rate of integration will be significantly lower due to the much larger average distance between loxP sites after excision. We therefore conclude,

given the low conversion rates used throughout this work, that conversion is effectively irreversible. No reverse conversion was observed during timelapse imaging of converting cultures or as sectoring following converted culture plating (data not shown).

A consistent and tunable selection coefficient for conversion was implemented using the antibiotic cycloheximide, which inhibits growth by binding the eukaryotic ribosomal protein Cyh2, impeding the translocation step and thus stalling elongating ribosomes [61]. The growth defect imposed on haploid sensitive cells by cycloheximide increases linearly with concentration [41]. The *cyh2*^{r} allele of Cyh2 is known to convey partial resistance to cycloheximide in heterozygous diploids [23]: this incomplete dominance can be rationalized as an inhibition of functional ribosome progression by one or more stalled ribosomes on the same transcript. Our unconverted cells are effectively heterozygous for this locus, possessing both a cycloheximide-sensitive allele at the native locus and the resistant allele (under the much stronger *ENO2* promoter) at the cell type-defining locus. We therefore anticipated that increasing cycloheximide concentration would slow growth of both converted and unconverted cells, but in a differential manner conferring a selective advantage on the unconverted merodiploids.

To determine the range of attainable selection coefficients, we developed isogenic cultures from independent mCitrine$^+$ convertants, mixed these with pure cultures of unconverted mCherry$^+$ cells, and observed the change in cell type ratio by flow cytometry as the cultures were propagated in media containing different concentrations of cycloheximide. As expected, the ratio of converted to unconverted cells decreased exponentially with time (Figure 2.8A). Selection coefficients estimated from these timecourse measurements were found to vary over an order of magnitude; no fitness disadvantage was measurable on our experimental timescale in the absence of cycloheximide. At higher cycloheximide concentrations, where growth of both cell types was significantly impeded by cycloheximide, the selection coefficient of converted cells (relative to unconverted cells) appeared to plateau around $s \approx -0.3$ (i.e., a 30% fitness defect).

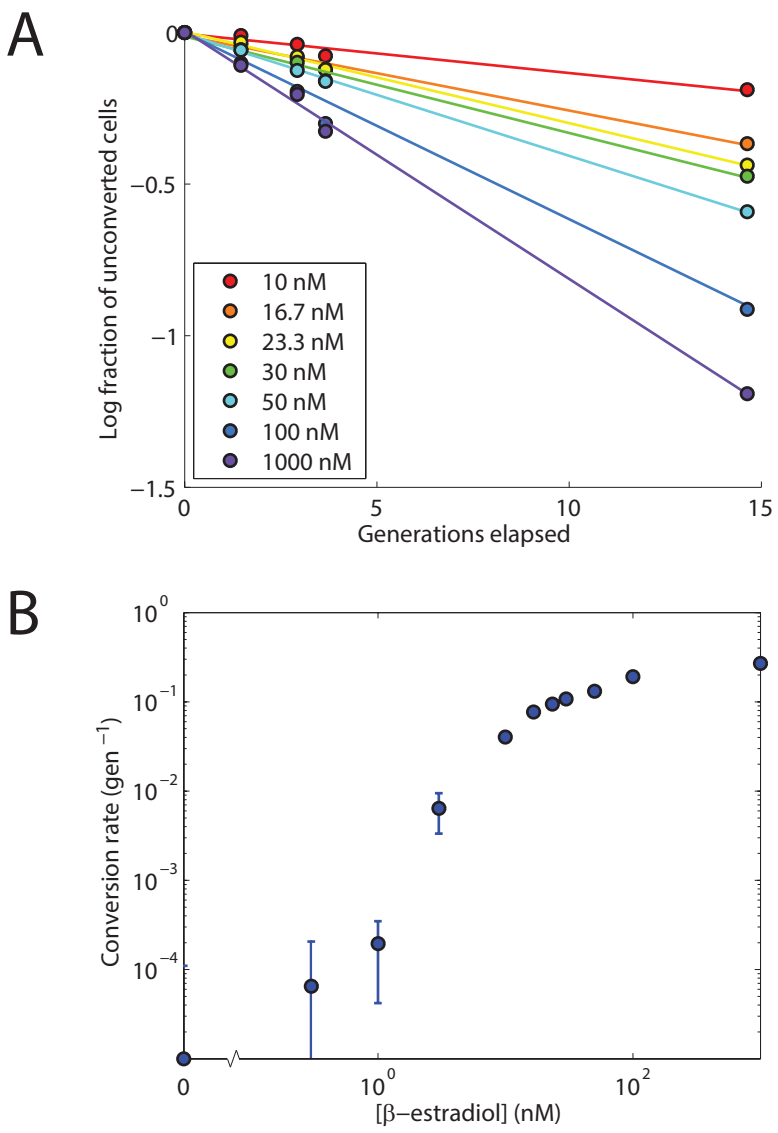

Figure 2.7: Measurement of the conversion rate by flow cytometry.

(A) Representative results of conversion rate assay. Cultures of unconverted cells (192) were grown in log phase in YPD medium for >24 h prior to addition of β-estradiol. Log phase growth was maintained during culture in β-estradiol medium while samples were collected at multiple timepoints for flow cytometry. The fraction of unconverted cells is expected to decrease exponentially under these conditions at a rate proportional to the conversion rate: regression lines calculated to estimate conversion rate are shown.

(B) Plot of conversion rate vs. β-estradiol concentration. Conversion rates were found to span three orders of magnitude, plateauing at a maximum conversion rate of ≈ 0.3 conversions per division.

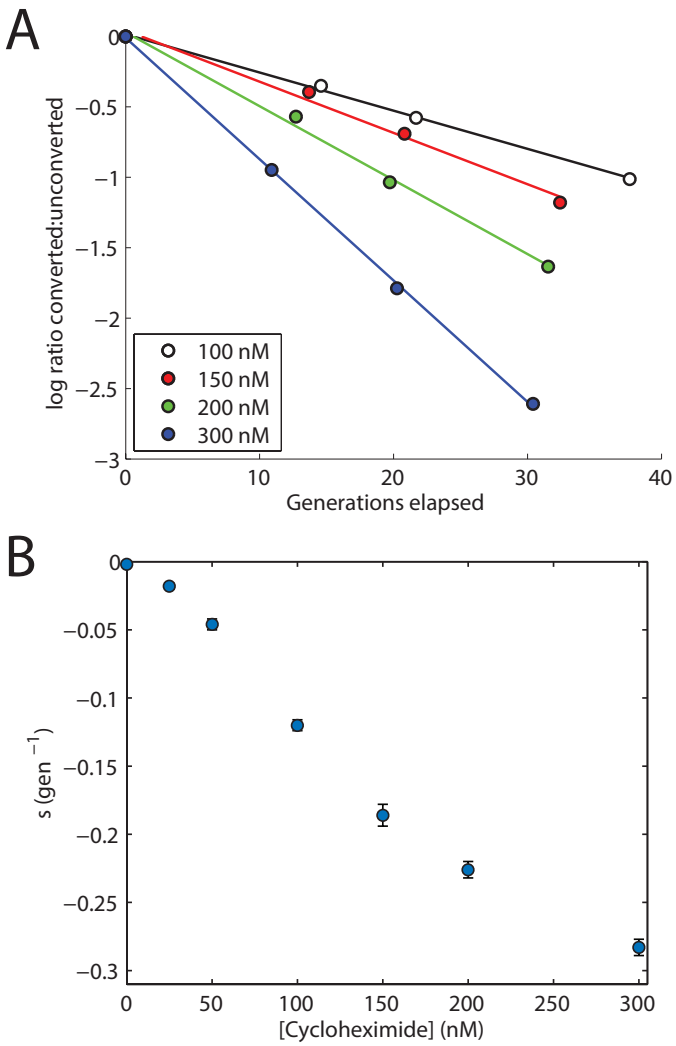

Figure 2.8: Selection coefficient of converted cells scales with cycloheximide concentration.

(A) Representative results of fitness assay. Cultures of unconverted (yMEW192) and converted (yMEW192c) cells were grown separately in log phase in YPD, combined, and resuspended in YPD containing cycloheximide. Cultures were maintained in log phase while samples were collected periodically for analysis by flow cytometry to determine cell type ratio. Generation values reflect the number of divisions experienced by an unconverted strain maintained in log phase in parallel, assayed by Coulter counting. Linear regression lines used to estimate selection coefficients are illustrated.
(B) Plot of selection coefficient vs. cycloheximide concentration. Selection coefficients were found to plateau at larger cycloheximide concentrations as both converted and unconverted cells experienced severe growth limitations (data not shown).

The uniformity of the cycloheximide-induced fitness defect across independant convertants appears evident from design, but in practice may be modified by undesired mutations. Of particular concern is the potential for gain of cycloheximide resistance through point mutation or gene conversion between the copy of *CYH2* at its normal chromosomal locus and the introduced copy of $cyh2^r$ prior to gene excision. All mutations known to confer cycloheximide resistance are substitution mutations at residue N37 of Cyh2 [39, 75], suggesting a small target size: estimating a point mutation rate of 5×10^{-10} per basepair per generation [44], such mutants would be unlikely to arise even during longterm culture in typical laboratory population sizes of $< 10^8$ cells. The rate of gene conversion is highly dependent on the length of the region of homology and the proximity to the homologous region's edge [2]. In our cell type-defining locus, the codon for N37 occurs 65 bp from the edge of an 870 bp region of contiguous homology (the native intron of *CYH2* was removed from our construct): the gene conversion rate would therefore be expected to be much less than the rate of 10^{-4} per generation observed for regions with 1.3 kb of homology and centrally-located variable sites [2]. Culture takeover by high-fitness mutants from both $mCherry^+$ and $mCitrine^+$ populations was occasionally observed, but generally did not interfere with the long-term (~5 day) experiments described below.

Mutation-selection balance predicts that a non-zero steady-state fraction f of unconverted cells is achievable during culture on both solid and liquid media. In the case of well-mixed liquid culture, the change in population of unconverted cells (u) and converted cells (c) with time will be:

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ c \end{pmatrix}(t) = \begin{pmatrix} 1-\mu & 0 \\ \mu & 1+s \end{pmatrix}\begin{pmatrix} u \\ c \end{pmatrix}(t)$$

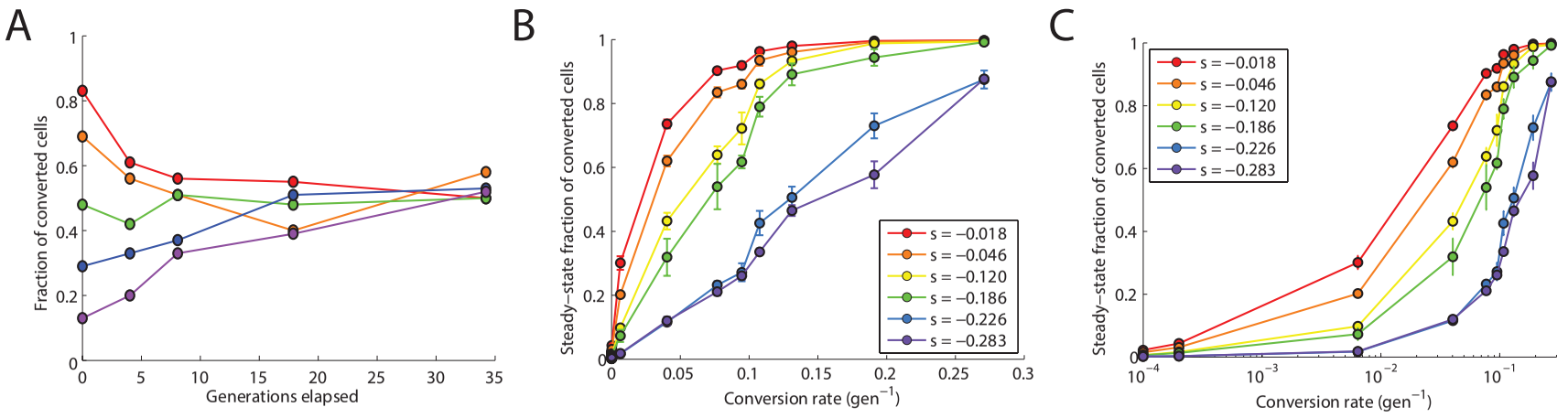

Figure 2.9: Converting cultures in liquid media attain steady-state cell type ratios dependent on cycloheximide and β-estradiol concentrations.

(A) Cultures of unconverted (yMEW192) and converted (yMEW192c) cells were grown separately in log phase in YPD, combined at several ratios, and resuspended in YPD containing 300 nM cycloheximide and 10 nM β-estradiol. Samples were collected at multiple timepoints during culture passaging and the fraction of converted cells was determined by flow cytometry.

(B) Cultures of a converting strain (yMEW192) were passaged in YPD media containing β-estradiol and cycloheximide. The steady-state fraction of converted cells was determined by flow cytometry and plotted according to the conversion rate and selection coefficient of the β-estradiol and cycloheximide concentrations used.

(C) Same data as (B), visualized on a semilog plot to highlight differences at low conversion rates.

where we have again taken the selection coefficient s to be negative. When $\mu \neq |s|$, this linear system has a general solution of the form:

$$\begin{pmatrix} u \\ c \end{pmatrix}(t) = a_1 \begin{pmatrix} -s-\mu \\ \mu \end{pmatrix} e^{(1-\mu)t} + a_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{(1+s)t}$$

It can be seen that if $\mu < |s|$, then the fraction f of unconverted cells will approach $1 - \mu/|s|$ as $t \to \infty$; however, if $\mu > |s|$, then unconverted cells will be lost with time. This solution is equivalent to the more common representation of the system:

$$\frac{\partial f}{\partial t} = |s| f (1-f) - \mu f$$

where the first term describes selection following Verhulst's logistic equation [84] and the second term describes conversion. This simple model does not take into account the finite population size, the population bottlenecking during back-dilution, or the mutation at other loci experienced by cultures of our converting strain. Furthermore, it does not take into account the loss of existing Cyh2^r protein by dilution following conversion and the intermediate selection coefficients thus conferred on newly-converted cells.

We therefore asked whether the predicted stability and steady-state cell type ratios could be achieved in long-term cultures of our converting strain. Cultures initiated at a range of cell type ratios converged to the same steady-state cell type ratio after passaging in media with $\mu < s$ set by a fixed concentration of cycloheximide and β-estradiol (Figure 2.9A). The steady-state fraction of converted cells increased with β-estradiol concentration (conversion rate) and decreased with cycloheximide concentration (selection coefficient) as expected (Figure 2.9BC). Complete or near loss of the unconverted cell population was observed when the conversion rate exceeded the selection coefficient, consistent with the prediction that the second fixed point should disappear in this regime [2]. The accordance

[2]Once the unconverted cell fraction becomes small relative to s, the timing of final loss is highly variable and often exceeded feasible experimental timescales. (Extended culturing could allow novel

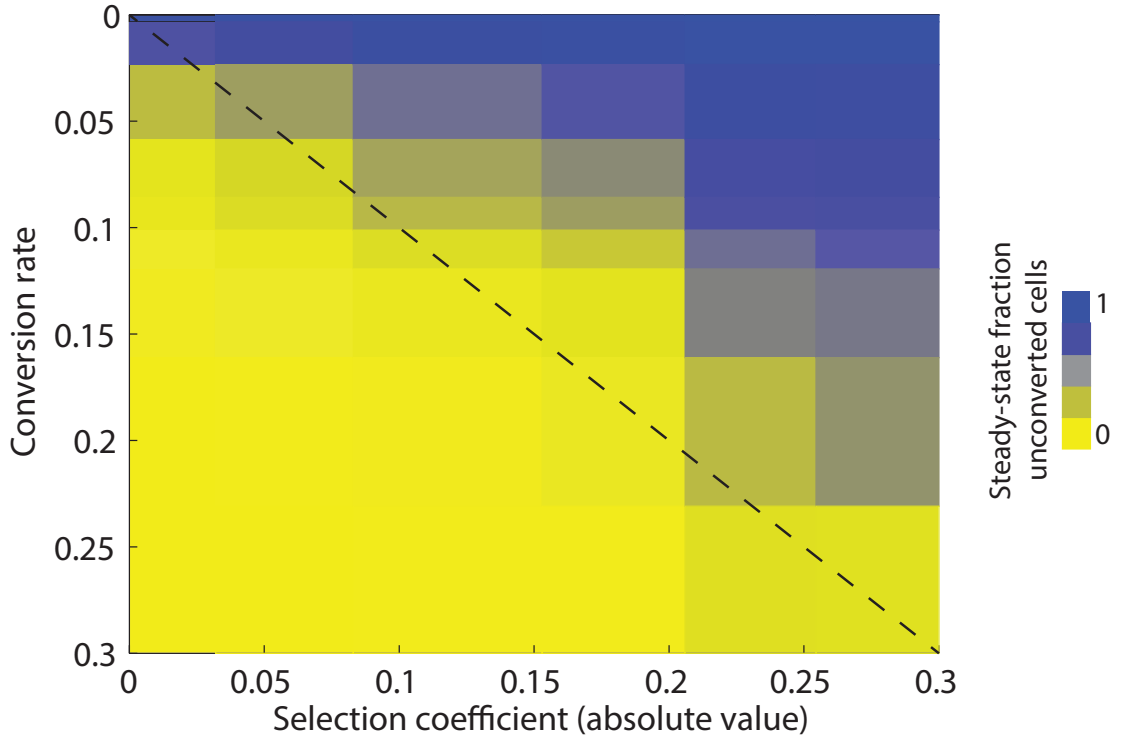

Figure 2.10: Comparison of theoretical and experimental phase diagram of unconverted cell type stability in liquid media.

Fraction of converted cells at steady state (data from Figure 2.9BC) are displayed as a heat map, with yellow representing 100% converted cells and blue representing 100% unconverted cells. The theoretical prediction of the threshold for unconverted cell type loss is shown as a dotted line.

between the predicted and observed loss of unconverted cell type stability is visualized in the heat map of Figure 2.10 and has quantitiative support ($\chi^2(2, 45) = 1.5778$, $p \approx 1$).

The corresponding prediction for steady-state unconverted cell fraction on solid media must account for the diffusive variation in boundary position at the colony frontier [46]:

$$\frac{\partial f(\phi)}{\partial t} = |s| f(\phi)\left[1 - f(\phi)\right] - \mu f(\phi) + \frac{D}{R(t)^2}\frac{\partial^2 f}{\partial \phi^2} + \eta(\phi)$$

beneficial mutations to arise and sweep, biasing the apparent timing of loss.) We therefore assume for simplicity that conditions in which the final fraction of unconverted cells was <1% would ultimately result in unconverted cell type loss.

where $\eta(\phi)$ is a noise term with correlation function $\sim f(\phi)[1 - f(\phi)]/R(t)$, which has been shown to give a fixed point at $f \approx 1 - k\mu/s^2$ for $\mu < \sqrt{|s|}$ [46]. We tested this prediction for colonies by inoculating a small droplet of saturated unconverted cell culture onto agar plates and visualizing cell type ratio at the colony boundaries following five days' growth (approx. 30 generations). As expected, loss of unconverted cells at the colony frontier occurred at much lower conversion rates than had been required in liquid media (Figure 2.11). To quantify cell type ratio, samples were collected from the colony frontier, resuspended in phosphate buffered saline, and analyzed by flow cytometry. The measured unconverted cell fractions were consistent with visual inspection and with the theoretical prediction (Figure 2.12; $\chi^2(2, 25) = 16.5872$, $p = 0.7857$). A data collapse representation of the data is shown in Figure 2.13.

The Haldane-Muller principle states that at mutation-selection equilibrium, the mutational load (decrease in average population fitness caused by the presence of mutations) will be a function only of the mutation rate. In our model system, the mutational load is easily calculated by comparing the culture growth rate in the presence or absence of conversion (i.e., β-estradiol). We tested this principle in our model system by comparing the mutational loads of steady-state cultures experiencing different selection coefficients (cycloheximide concentrations). While the data were consistent with this hypothesis within error (Figure 2.14), the pattern of residuals suggests an increase in mutational load with the selection coefficient of converted cells (viz., with cycloheximide concentration). This could result from a minor increase in probability of conversion: the growth rate of unconverted cells is also decreased in the presence of cycloheximide, potentially extending the active time of the mitotic exit network pathway and thus of Cre activity.

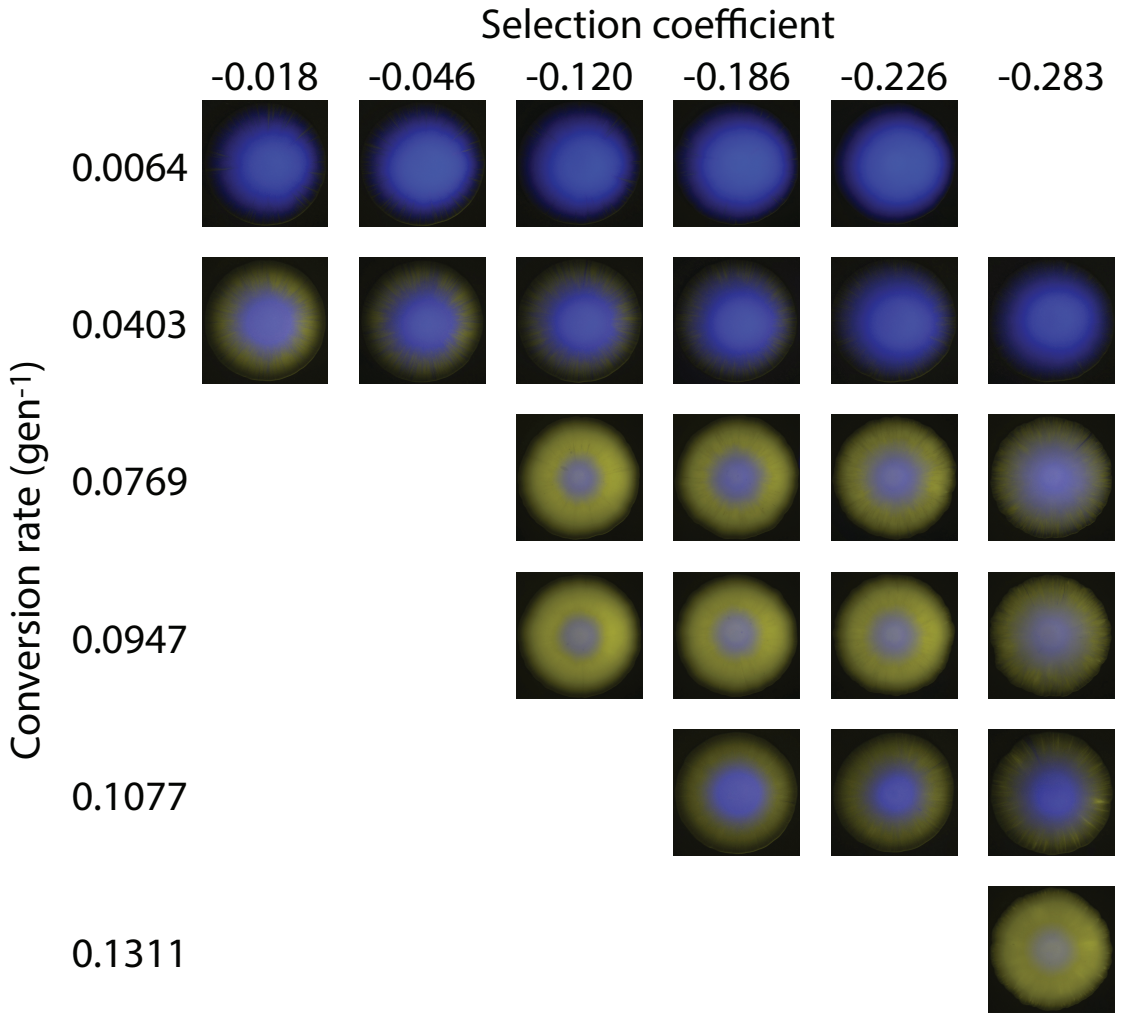

Figure 2.11: Change in cell type ratio during colony growth.

Colonies were grown for five days on 1% agar YPD plates containing β-estradiol and cycloheximide, then imaged for mCherry (unconverted, pseudocolored blue) and mCitrine (converted) fluorescence. Representative images are shown on axes corresponding to the conversion rate and selective coefficient imposed by the respective β-estradiol and cycloheximide concentrations.

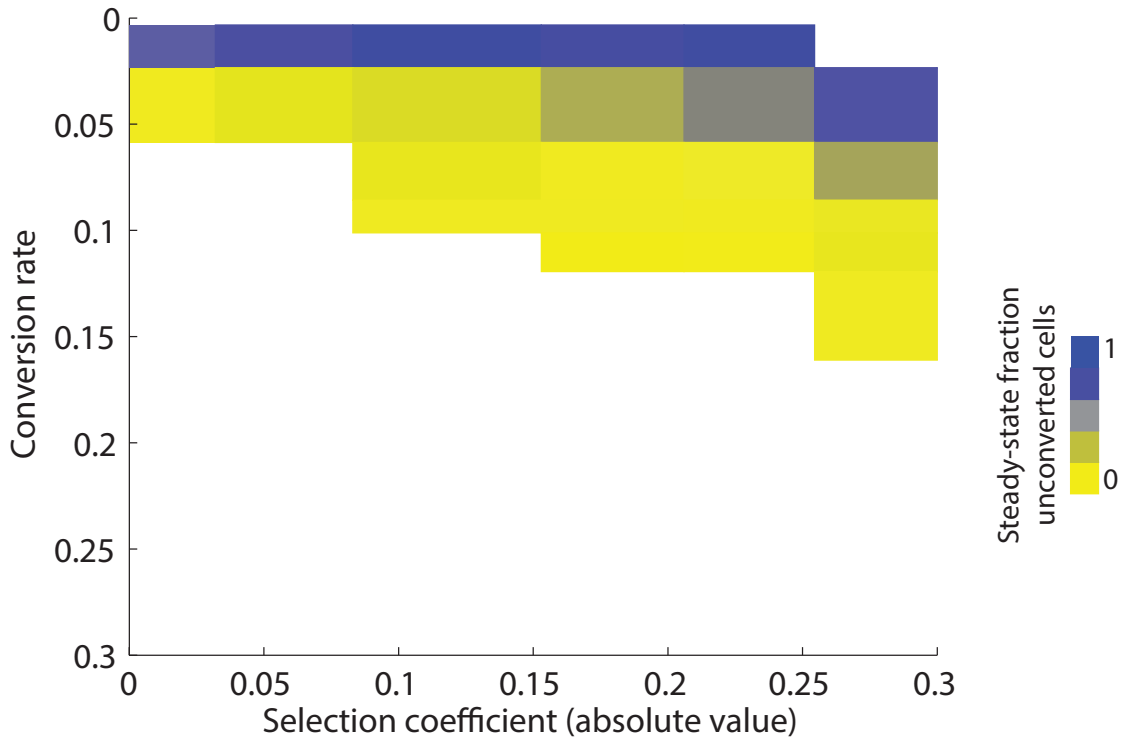

Figure 2.12: Comparison of theoretical and experimental phase diagram of unconverted cell type stability on solid media.

Colonies were grown for five days on 1% agar YPD plates containing β-estradiol and cycloheximide. Plugs of cells at the colony frontier were collected with a pipette tip and resuspended in PBS to determine the fraction of converted cells by flow cytometry. Results are displayed as a heat map, with yellow representing 100% converted cells and blue representing 100% unconverted cells.

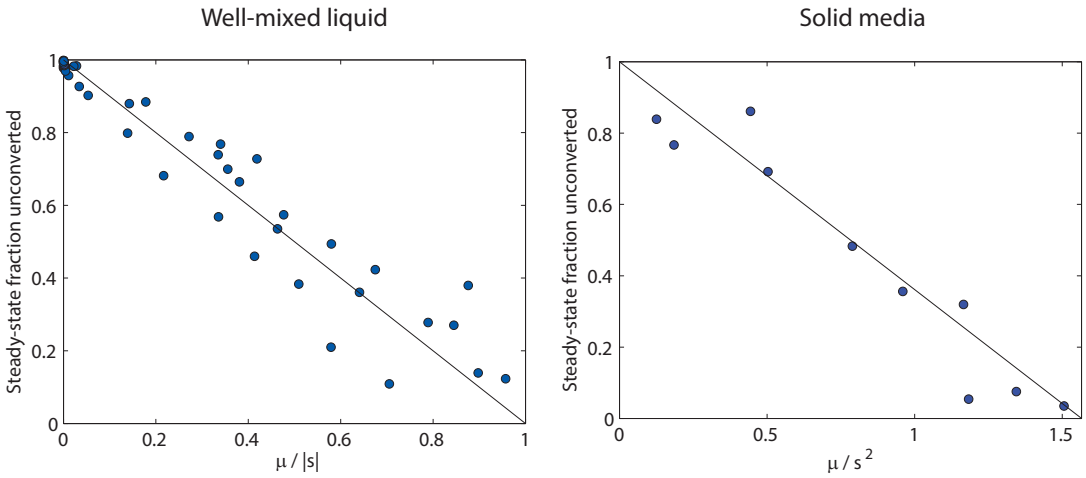

Figure 2.13: Data collapse representation of steady-state fractions of unconverted cells in well-mixed liquid media and on solid media.

Steady-state fractions of unconverted cells are plotted against the predicted parameters of interest ($\mu/|s|$ for well-mixed liquid media and μ/s^2 for solid media).

2.5 Discussion

Two major strategies for synthesizing gene expression switches have been employed previously. The first entails engineering cross-regulation between two strong repressors such that only one can be actively transcribed at a time. The original Collins toggle switch uses LacI and TetR: the repressor LacI is expressed from a promoter containing the TetR binding site, and TetR is expressed from a promoter containing the LacI binding site [26]. When activity levels are asymmetic due to differences in repressor efficacy, such systems can fail to exhibit bistability: however, Gardner et al. chose repressors that bind cooperatively to alleviate this concern. Switching in their system is promoted by interfering with repressor activity (e.g., adding IPTG to prevent LacI binding). Spontaneous switching is also possible, however, if the protein level of the active repressor drops spontaneously due to fluctuations in gene expression; the authors do not describe the change in relative populations during free running,

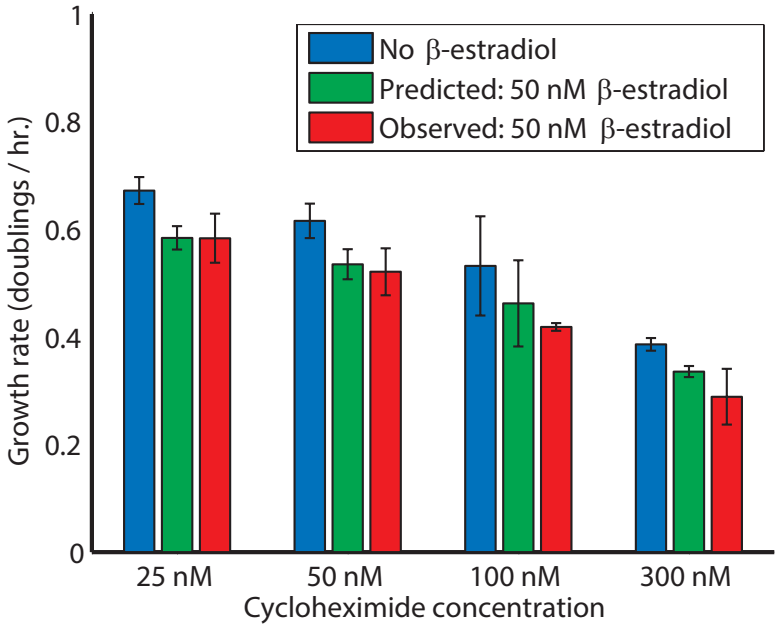

Figure 2.14: Apparent dependence of mutational load on selection coefficient.

The growth rate of strain yMEW192 was determined at its steady-state ratio during growth in YPD ± 50 nM β-estradiol at the indicated concentrations of cycloheximide. The mutational load was determined from the relative growth rate difference according to the formula:

$$L = \frac{w_{\text{no } \beta\text{-estradiol}} - w_{50 \text{ nM } \beta\text{-estradiol}}}{w_{\text{no } \beta\text{-estradiol}}}$$

and is expected to equal the conversion rate, $\mu \approx 0.1311$ (red line), with no dependence on selection coefficient. Error bars reflect a $\pm 2\sigma$ confidence interval after accounting for propagation of error in growth rate determinations from Coulter counter growth curves.

so it is difficult to assess the frequency of such reversion events. A related method using one repressor and one protease showed state persistence times of >40 generations [34].

A second common approach involves autoregulation of a transcriptional activator to generate a positive feedback loop: once expression occurs spontaneously or is induced, the transcription factor binds to its own promoter and ensures its continued expression [35, 70]: here, too, noise in gene expression can allow reversion to the "off" state, sometimes requiring excessive elaborations (multiple integrations of constructs, mutually activating transcription factors, etc.) to achieve stability.

Our model system offers several advantages over these previous approaches: recombinase-mediated gene excision ensures conversion events remain irreversible, and conversion rate can be tuned without affecting the gene expression state of each cell type. We have shown that our system permits tuning of selection coefficient and conversion rate over orders of magnitude, allowing a range of steady-state ratios between cell types to be maintained. Although loss of unconverted cell proteins (mCherry and cyh2^r) after conversion is not immediate, it is sufficiently rapid that culture growth rates and conversion-selection phase diagrams are in excellent agreement with predictions from mutation-selection balance theory.

In the following chapter, we will apply this model system to investigate selection pressures underlying the coevolution of multicellularity and differentiation. Another potential application of our work is to overexpress costly proteins for commercial purposes. In industry, yeast strains engineered to secrete desirable compounds are sometimes maintained at large populations in chemostats ("bioreactors") to allow continual harvesting of the byproducts over the course of days or weeks. During this culturing period, mutants that do not produce the desired compound, and therefore have more energy to invest in cell division, arise and sweep the population; eventually the bioreactor must be emptied and the culture reinitiated. Our system could prolong the useful culturing period by limiting expression of

the costly genes to differentiated cells: under these circumstances, no selective advantage exists for mutants that do not produce the desired but costly compounds.

Another possible extension of our work is to the development of reversible switches. In the model presented in this chapter, recombination between loxP sites in sense orientation causes gene excision. Anti-parallel loxP sites can be engineered to flank the promoter and allow its orientation to be reversed by Cre recombinase, resulting in expression of a different complement of proteins. Preliminary steps have been taken for engineering such a construct: Bryan Weinstein of the Nelson lab plans to use this work to explore theoretical predictions of cell type ratio and culture dynamics for reversible switches.

Acknowledgements

P_{SCW11}-Cre-EBD78 was a generous gift of Derek Lindstrom and Dan Gottschling [48]. Thanks are also extended to R. Scott McIsaac and David Botstein for providing *GAL4DBD-ER-VP16* [52], a construct considered for an earlier version of our differentiation system, and to Max Lavrentovich, Bryan Weinstein, and David Nelson for helpful discussions and experimental suggestions.

2.6 Bibliography

[1] K. Abremski and R. Hoess. Bacteriophage P1 site-specific recombination. Purification and properties of the Cre recombinase protein. *Journal of Biological Chemistry*, 259(3):1509–1514, 1984.

[2] B. Y. Ahn, K. J. Dornfeld, T. J. Fagrelius, and D. M. Livingston. Effect of limited homology on gene conversion in a *Saccharomyces cerevisiae* plasmid recombination system. *Molecular and Cellular Biology*, 8(6):2442–2448, 1988.

[3] S. Apte and N. Prabhavathi. Rearrangements of nitrogen fixation (*nif*) genes in the heterocystous cyanobacteria. *Journal of Biosciences*, 19:579–602, 1994.

[4] A. Bachmair, D. Finley, and A. Varshavsky. *In vivo* half-life of a protein is a function of its amino-terminal residue. *Science*, 234(4773):179–186, 1986.

[5] R. T. Baker, J. W. Tobias, and A. Varshavsky. Ubiquitin-specific proteases of *Saccharomyces cerevisiae*: Cloning of *UBP2* and *UBP3*, and functional analysis of the *UBP* gene family. *Journal of Biological Chemistry*, 267(32):23364–75, 1992.

[6] J. D. Barrass and J. D. Beggs. Splicing goes global. *Trends in Genetics*, 19(6):295 – 298, 2003.

[7] C. Brack, M. Hirama, R. Lenhard-Schuller, and S. Tonegawa. A complete immunoglobulin gene is created by somatic recombination. *Cell*, 15:1–14, 1978.

[8] M. E. Bradley, S. Bagriantsev, N. Vishveshwara, and S. W. Liebman. Guanidine reduces stop codon read-through caused by missense mutations in *SUP35* or *SUP45*. *Yeast*, 20(7):625–632, 2003.

[9] S. R. Broadbent and J. M. Hammersley. Percolation processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, 53:629–641, 1957.

[10] D. J. Burke, D. C. Amberg, and J. N. Strathern. *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*. Cold Spring Harbor Laboratory Press, 2005.

[11] B. Canton, L. Anna, and D. Endy. Refinement and standardization of synthetic biological parts and devices. *Nature Biotechnology*, 26(7):787–793, 2008.

[12] M. Carlson and D. Botstein. Two differentially regulated mRNAs with different 5' ends encode secreted and intracellular forms of yeast invertase. *Cell*, 28(1):145–154, 1982.

[13] T.-H. Cheng, C.-R. Chang, P. Joy, S. Yablok, and M. R. Gartenberg. Controlling gene expression in yeast by inducible site-specific recombination. *Nucleic Acids Research*, 28(24):e108, 2000.

[14] A. Colman-Lerner, T. E. Chin, and R. Brent. Yeast Cbk1 and Mob2 activate daughter-specific genetic programs to induce asymmetric cell fates. *Cell*, 107(6):739 – 750, 2001.

[15] M. Constante, R. Grunberg, and M. Isalan. A Biobrick library for cloning custom eukaryotic plasmids. *PLoS ONE*, 6(8):e23685, 08 2011.

[16] B. S. Cox. Psi, a cytoplasmic suppressor of super-suppressor in yeast. *Nature*, 20:505–521, 1965.

[17] H. C. Danforth. The frequency of mutatton and the incidence of hereditary traits in man. In C. B. Davenport, editor, *Eugenics, genetics and the family : Scientific papers of the second international congress of eugenics*. Williams & Wilkins Company, 1923.

[18] E. Davidson. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Academic Press, 2006.

[19] S. M. Doel, S. J. McCready, C. R. Nierras, and B. S. Cox. The dominant PNM2- mutation which eliminates the psi factor of *Saccharomyces cerevisiae* is the result of a missense mutation in the *SUP35* gene. *Genetics*, 137(3):659–70, 1994.

[20] E. Domany and W. Kinzel. Equivalence of cellular automata to Ising models and directed percolation. *Phys. Rev. Lett.*, 53:311–314, Jul 1984.

[21] M.-T. Doolin, A. L. Johnson, L. H. Johnston, and G. Butler. Overlapping and distinct roles of the duplicated yeast transcription factors Ace2p and Swi5p. *Molecular Microbiology*, 40(2):422–432, 2001.

[22] N. S. Enattah, T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, and I. Jarvela. Quantitation of readthrough of termination codons in yeast using a novel gene fusion assay. *Yeast*, 7(2):173–183, 1991.

[23] H. M. Fried and J. R. Warner. Molecular cloning and analysis of yeast gene for cycloheximide resistance and ribosomal protein L29. *Nucleic Acids Research*, 10(10):3133–3148, 1982.

[24] S. Fukushige and B. Sauer. Genomic targeting with a positive-selection lox integration vector allows highly reproducible gene expression in mammalian cells. *Proceedings of the National Academy of Sciences*, 89(17):7905–7909, 1992.

[25] K. E. Galloway, E. Franco, and C. D. Smolke. Dynamically reshaping signaling networks to program cell fate via genetic controllers. *Science*, 341(6152), 2013.

[26] T. Gardner, C. Cantor, and J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403:339–342, 2000.

[27] F. Guo, D. Gopaul, and G. Van Duyne. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature*, 389:40–46, 1997.

[28] J. H. Koschwanez, K. R. Foster, and A. W. Murray. Sucrose utilization in budding yeast as a model for the origin of undifferentiated multicellularity. *PLoS Biol*, 9(8):e1001122, 2011.

[29] J. B. S. Haldane. The effect of variation of fitness. *The American Naturalist*, 71(735):pp. 337–349, 1937.

[30] J. Haldane. The rate of spontaneous mutation of a human gene. *Journal of Genetics*, 31(3):317–326, 1935.

[31] H. Hinrichsen. Non-equilibrium critical phenomena and phase transitions into absorbing states. *Advances in Physics*, 49(7):815–958, 2000.

[32] R. Hoess, A. Wierzbicki, and K. Abremski. Formation of small circular DNA molecules via an *in vitro* site-specific recombination system. *Gene*, 40(2fb3):325 – 329, 1985.

[33] O. W. Holmes. Buck v. Bell. 274 U.S. 200. `http://supreme.justia.com/cases/federal/us/274/200/case.html#207`, 1927.

[34] D. Huang, W. Holtz, and M. Maharbiz. A genetic bistable switch utilizing nonlinear protein degradation. *Journal of Biological Engineering*, 6(1):9, 2012.

[35] N. Ingolia and A. Murray. Positive-feedback loops as a flexible biological module. *Current Biology*, 17:668–677, 2007.

[36] S. K. Jang, H. G. Krausslich, M. J. Nicklin, G. M. Duke, A. C. Palmenberg, and E. Wimmer. A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during *in vitro* translation. *Journal of Virology*, 62(8):2636–2643, 1988.

[37] V. Kapitonov and J. Jurka. RAG1 core and V(D)J recombination signal sequences were derived from *transib* transposons. *PLoS Biology*, 3:e181, 2005.

[38] L. D. Kapp and J. R. Lorsch. The molecular mechanics of eukaryotic translation. *Annual Review of Biochemistry*, 73(1):657–704, 2004.

[39] N. F. Kaufer, H. M. Fried, W. F. Schwindinger, M. Jasin, and J. R. Warner. Cycloheximide resistance in yeast: the gene and its protein. *Nucleic Acids Research*, 11(10):3123–3135, 1983.

[40] W. Kinzel. Phase transitions of cellular automata. *Zeitschrift für Physik B Condensed Matter*, 58(3):229–244, 1985.

[41] K. S. Korolev, M. J. I. Mller, N. Karahan, A. W. Murray, O. Hallatschek, and D. R. Nelson. Selective sweeps in growing microbial colonies. *Physical Biology*, 9(2):026008, 2012.

[42] B. Kunkel, R. Losick, and P. Straiger. The *Bacillus subtilis* gene for the developmental transcription factor σ_K is generated by excision of a dispensable DNA element containing a sporulation recombinase gene. *Genes & Development*, 4:525–535, 1990.

[43] J. Lafaille, A. DeCloux, M. Bonneville, Y. Takagaki, and S. Tonegawa. Junctional sequences of T cell receptor $\gamma\delta$ genes: implications for $\gamma\delta$ T cell lineages and for a novel intermediate of VDJ joining. *Cell*, 59:859–870, 1989.

[44] G. I. Lang and A. W. Murray. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. *Genetics*, 178(1):67–82, 2008.

[45] G. I. Lang, A. W. Murray, and D. Botstein. The cost of gene expression underlies a fitness trade-off in yeast. *Proceedings of the National Academy of Sciences*, 106(14):5755–5760, 2009.

[46] M. O. Lavrentovich, K. S. Korolev, and D. R. Nelson. Radial Domany-Kinzel models with mutation and selection. *Phys. Rev. E*, 87:012103, Jan 2013.

[47] M. Li and S. Elledge. SLIC: A method for sequence- and ligation-independent cloning. In J. Peccoud, editor, *Gene Synthesis*, volume 852 of *Methods in Molecular Biology*, pages 51–59. Humana Press, 2012.

[48] D. L. Lindstrom and D. E. Gottschling. The mother enrichment program: A genetic system for facile replicative life span analysis in *Saccharomyces cerevisiae*. *Genetics*, 183(2):413–422, 2009.

[49] M. S. Longtine, A. Mckenzie III, D. J. Demarini, N. G. Shah, A. Wach, A. Brachat, P. Philippsen, and J. R. Pringle. Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast*, 14(10):953–961, 1998.

[50] P. M. Lund and B. S. Cox. Reversion analysis of [psi-] mutations in *Saccharomyces cerevisiae*. *Genetics Research*, 37:173–182, 4 1981.

[51] L. T. MacNeil and A. J. Walhout. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, 21(5):645–657, 2011.

[52] R. S. McIsaac, S. J. Silverman, M. N. McClean, P. A. Gibney, J. Macinskas, M. J. Hickman, A. A. Petti, and D. Botstein. Fast-acting and nearly gratuitous induction of gene expression and protein depletion in *Saccharomyces cerevisiae*. *Molecular biology of the cell*, 22(22):4447–4459, 2011.

[53] A. D. Miller. Development and applications of retroviral vectors. In J. M. Coffin, S. H. Hughes, and H. E. Varmus, editors, *Retroviruses*. Cold Spring Harbor Laboratory Press, 2002.

[54] M. Mokrejs, V. Vopalensky, O. Kolenaty, T. Masek, Z. Feketova, P. Sekyrova, B. Skaloudova, V. Kriz, and M. Pospisek. IRESite: the database of experimentally verified IRES structures (www.iresite.org). *Nucleic Acids Research*, 34(suppl 1):D125–D130, 2006.

[55] H. Muller. *Out of the Night: A Biologist's View of the Future.* Vanguard Press, 1935.

[56] H. Muller. Our load of mutations. *The American Journal of Human Genetics*, 2(2):111 – 174, 1950.

[57] M. J. I. Müller, B. I. Neugeboren, D. R. Nelson, and A. W. Murray. Genetic drift opposes mutualism during spatial population expansion. *Proceedings of the National Academy of Sciences*, 111(3):1037–1042, 2014.

[58] O. Namy, G. Duchateau-Nguyen, and J.-P. Rousset. Translational readthrough of the *PDE2* stop codon modulates cAMP levels in *Saccharomyces cerevisiae*. *Molecular Microbiology*, 43(3):641–652, 2002.

[59] F. Ness, P. Ferreira, B. S. Cox, and M. F. Tuite. Guanidine hydrochloride inhibits the generation of prion "seeds" but not prion protein aggregation in yeast. *Molecular and Cellular Biology*, 22(15):5593–5605, 2002.

[60] G. P. Newnam, R. D. Wegrzyn, S. L. Lindquist, and Y. O. Chernoff. Antagonistic interactions between yeast chaperones Hsp104 and Hsp70 in prion curing. *Molecular and Cellular Biology*, 19(2):1325–1333, 1999.

[61] T. G. Obrig, W. J. Culp, W. L. McKeehan, and B. Hardesty. The mechanism by which cycloheximide and related glutarimide antibiotics inhibit peptide synthesis on reticulocyte ribosomes. *Journal of Biological Chemistry*, 246(1):174–181, 1971.

[62] T. L. Orr-Weaver, J. W. Szostak, and R. J. Rothstein. Yeast transformation: a model system for the study of recombination. *Proceedings of the National Academy of Sciences*, 78(10):6354–6358, 1981.

[63] J. Palis. Ontogeny of erythropoiesis. *Current Opinion in Hematology*, 15:155–161, 2008.

[64] J. Pelletier and N. Sonenberg. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature*, 334(6180):320–325, 1988.

[65] D. Plotz. *The Genius Factory: The Curious History of the Nobel Prize Sperm Bank.* Random House Trade Paperbacks, 2006.

[66] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 2012.

[67] Selpi, C. Bryant, G. Kemp, J. Sarv, E. Kristiansson, and P. Sunnerhagen. Predicting functional upstream open reading frames in *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 10(1):451, 2009.

[68] N. C. Shaner, R. E. Campbell, P. A. Steinbach, B. N. G. Giepmans, A. E. Palmer, and R. Y. Tsien. Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nature*, 22(12):1567–1572, 2004.

[69] M. A. Sheff and K. S. Thorn. Optimized cassettes for fluorescent protein tagging in *Saccharomyces cerevisiae*. *Yeast*, 21(8):661–670, 2004.

[70] V. Siciliano, F. Menolascina, L. Marucci, C. Fracassi, I. Garzilli, M. N. Moretti, and D. di Bernardo. Construction and modelling of an inducible positive feedback loop stably integrated in a mammalian cell-line. *PLoS Comput Biol*, 7(6):e1002074, 06 2011.

[71] R. S. Sikorski and P. Hieter. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics*, 122(1):19–27, 1989.

[72] I. Stansfield, K. M. Jones, V. V. Kushnirov, A. R. Dagkesamanskaya, A. I. Poznyakovski, S. V. Paushkin, B. S. Nierras, C R amd Cox, M. D. Ter-Avanesyan, and M. F. Tuite. The products of the *SUP45* (eRF1) and *SUP35* genes interact to mediate translation termination in *Saccharomyces cerevisiae*. *EMBO Journal*, 14:4365–4373, 1995.

[73] B. C. Stanton, A. A. K. Nielsen, A. Tamsir, K. Clancy, T. Peterson, and C. A. Voigt. Genomic mining of prokaryotic repressors for orthogonal logic gates. *Nature Chemical Biology*, 10:99–105, 2014.

[74] N. Sternberg and D. Hamilton. Bacteriophage P1 site-specific recombination: I. Recombination between loxP sites. *Journal of Molecular Biology*, 150(4):467 – 486, 1981.

[75] W. Stoecklein, W. Piepersberg, and A. Boeck. Amino acid replacements in ribosomal protein YL24 of *Saccharomyces cerevisiae* causing resistance to cycloheximide. *FEBS Letters*, 136(2):265 – 268, 1981.

[76] M. D. Ter-Avanesyan, A. R. Dagkesamanskaya, V. V. Kushnirov, and V. N. Smirnov. The SUP35 omnipotent suppressor gene is involved in the maintenance of the non-Mendelian determinant [psi+] in the yeast *Saccharomyces cerevisiae*. *Genetics*, 137(3):671–6, 1994.

[77] D. A. Thompson, M. M. Desai, and A. W. Murray. Ploidy controls the success of mutators and nature of mutations during budding yeast evolution. *Current Biology*, 16(16):1581–1590, 2006.

[78] S. R. Thompson, K. D. Gulyas, and P. Sarnow. Internal initiation in *Saccharomyces cerevisiae* mediated by an initiator tRNA/eIF2-independent internal ribosome entry site element. *Proceedings of the National Academy of Sciences*, 98(23):12972–12977, 2001.

[79] H. L. True and S. L. Lindquist. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*, 407:477–483, 2000.

[80] M. F. Tuite, C. R. Mundy, and B. S. Cox. Agents that cause a high frequency of genetic change from [psi+] to [psi-] in *Saccharomyces cerevisiae*. *Genetics*, 98(4):691–711, 1981.

[81] J. Tyedmers, M. L. Madariaga, and S. Lindquist. Prion switching in response to environmental stress. *PLoS Biol*, 6(11):e294, 11 2008.

[82] A. Uri. *An Introduction to Systems Biology: Design Principles for Biological Circuits*. Chapman & Hall/CRC, 2007.

[83] G. D. Van Duyne. A structural view of Cre-loxP site-specific recombination. *Annual Review of Biophysics and Biomolecular Structure*, 30(1):87–104, 2001. PMID: 11340053.

[84] P. F. Verhulst. Notice sur la loi que la population poursuit dans son accroissement. In A. Quetelet, editor, *Correspondance Mathématique et Physique*, volume 10. Hauman and Co., 1838.

[85] W. P. Voth, A. E. Olsen, M. Sbia, K. H. Freedman, and D. J. Stillman. Ace2, Cbk1, and Bud4 in budding and cell separation. *Eukaryotic Cell*, 4(6):1018–1028, 2005.

[86] C. H. Waddington. Canalization of development and the inheritance of acquired characters. *Nature*, 150:563–565, 1942.

[87] P. Walter, I. Ibrahimi, and G. Blobel. Translocation of proteins across the endoplasmic reticulum. I. Signal recognition protein (SRP) binds to *in vitro*-assembled polysomes synthesizing secretory protein. *The Journal of Cell Biology*, 91(2):545–550, 1981.

[88] A. Weismann. *The Germ-Plasm: A Theory of Heredity*. Charles Scribner's Sons, 1893.

[89] I. Williams, J. Richardson, A. Starkey, and I. Stansfield. Genome-wide prediction of stop codon readthrough during translation in the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 32(22):6605–6616, 2004.

[90] W. Wriggers, S. Chakravarty, and P. A. Jennings. Control of protein functional dynamics by peptide linkers. *Peptide Science*, 80(6):736–746, 2005.

[91] T. Yoshimatsu and F. Nagawa. Control of gene expression by artificial introns in *Saccharomyces cerevisiae*. *Science*, 244(4910):1346–1348, 1989.

[92] B. Zheng, M. Sage, E. A. Sheppeard, V. Jurecic, and A. Bradley. Engineering mouse chromosomes with Cre-loxP: Range, efficiency, and somatic applications. *Molecular and Cellular Biology*, 20(2):648–655, 2000.

Chapter 3

Evolutionary pressures underlying the coevolution of multicellularity and cellular differentiation

"Non-cooperation with evil is as much a duty as is cooperation with good."

- Mahatma Gandhi

3.1 Abstract

Many multicellular organisms produce terminally-differentiated cells with finite division potential. Producing cells that cannot contribute to reproduction dramatically reduces an organism's maximum fitness: this cost must be offset by benefits gained from the specialized functions that differentiated cells can perform. Differentiated multicellularity is an effective life strategy that has evolved independently in dozens of clades, yet little is known of the evolutionary trajectory required to attain this phenotype. Multicellularity and differentiation have separate biological underpinnings and therefore most likely appear sequentially: the order of evolution of these two traits can seldom be inferred, but when evolutionary intermediates persist, they are invariably multicellular species that lack cellular differentiation.

We propose that unicellular differentiation is an unstable phenotype due to the potential for population invasion by non-differentiating (revertant) mutants. We test this claim by engineering yeast strains which differentiate and/or form multicellular clumps, showing that multicellular strains can resist invasion by such mutants while unicellular strains cannot. This result explains the paucity of extant unicellular, differentiating species and suggests that their limited duration makes evolution of differentiated multicellularity through such intermediates unlikely.

3.2 Introduction

Terminal differentiation is an irreversible change in a cell's gene expression accompanied by loss of the ability to grow indefinitely. Producing daughter cells which cannot propagate the genome dramatically impacts an organism's fitness: the theoretical minimum effect is to decrease the exponential growth rate by an amount equal to the differentiation rate, but the fitness costs may be higher if differentiated cells consume limiting resources. The prevalence of species with soma shows that this life strategy succeeds when differentiated cells perform a useful function whose benefits more than make up for the costs. For example, differentiated cells may provide structure, motility, nutrient uptake and metabolism, sensation and response to the external environment, maintenance of homeostasis, and protection from predators: by relegating such resource-intensive tasks to the soma, undifferentiated cells can maintain high viability while increasing their own division rate (and thus the speed of the organism's reproduction).

It can be seen from the list above that intercellular adhesion is required for many potential somatic cell functions. The exception is the secretion of products which can be shared between cells through the extracellular milieu. Exchange of nutrients in this manner is effective in mutualisms[1] and would be straightforward for processes that must begin outside the cell, such as phosphate scavenging and the breakdown of nutrients too large to import. Despite this, no examples of terminally-differentiating, unicellular species are known to us (Figure 3.1).

Multicellularity and differentiation have each evolved independently over a dozen times and have separate biological bases, yet these two traits are most often seen together in extant species. The order of evolution of these two traits can in some cases be inferred by identifying persistent species that appear to represent evolutionary intermediates (Figure 3.1). For example, the existence of multicellular, undifferentiated cyanobacteria such

[1]Enforced physical associations in mutualisms are traditionally thought to be acquired after the initial productive association between species, though counterexamples are known (E.H. Hom, p.c.).

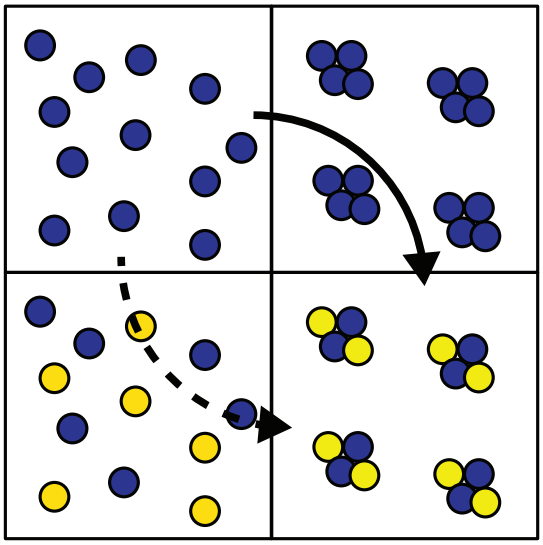

Figure 3.1: Alternative pathways to the evolution of differentiated multicellularity.

The molecular bases of cellular differentiation and multicellularity are independent, and therefore are unlikely to evolve concurrently. Evolution of differentiated multicellularity therefore likely proceeds through an intermediate that is either multicellular or produces differentiated cells (yellow), but not both. The persistence of multicellular species which do not differentiate in some taxa strongly suggests that multicellularity evolved first in those groups (solid arrow; see Chapter 1 for a review). No examples of evolution through a unicellular, differentiating intermediate have been demonstrated (dashed arrow).

as *Crinalium magnum* suggests that multicellularity evolved prior to differentiation in that clade [27, 28]. Unfortunately, for many multicellular taxa, remaining intermediates are not known and this method cannot be applied. Inferences made from extant species are also inherently probabilistic, as they may represent revertant rather than intermediate forms. Must multicellularity evolve first in all cases? What evolutionary pressures might disfavor the alternative route, and what can be understood by the lack of unicellular, differentiating species alive today?

We propose that unicellular differentiation is an inherently unstable life strategy because of the potential for invasion by reversion mutants (Figure 3.2A). The problem arises from the requirement that differentiated cell byproducts be shared through the growth medium in order for undifferentiated cells to profit from them. In this case, all cells have equal access to the byproducts, including any mutants that may arise which can no longer differentiate. Such mutants are not likely to be rare: even if just one basepair were essential for differentiation, this position would mutate in >1 in 10^8 cells (given a typical microbial genome length), which is far less than a typical species's population size. Such a mutant

would have a fitness advantage over the rest of the population because it would have equal access to nutrients without ever producing differentiated daughters itself: barring loss by drift, the mutant would be expected to spread through the population and revert the species to a unicellular, non-differentiating phenotype. A multicellular species, however, could potentially resist population invasion by such mutants because physical association provides cells with preferential access to the products secreted by their own differentiated daughter cells (Figure 3.2B).

We chose to test this hypothesis directly by comparing yeast strains bioengineered to be multicellular, to terminally differentiate, or to do both. A major advantage of this approach over analytical modeling is the lack of dependence on inference of biologically-important parameters[2]: by employing true biological systems, we demonstrate that the conversion rates, fitness benefits, growth rates, etc. in our system are realistic. Moreover, the production of the necessary yeast strains poses questions of inherent interest from the perspective of synthetic biology. Must differentiation pathways rely on gene regulatory networks, as is the norm in natural systems? What alternatives exist, and can they be reliably implemented given our current understanding of the underlying molecular biology?

In this chapter, we describe the production and analysis of a yeast strain which terminally differentiates through the recombinase-mediated excision of *CDC28*, a gene required for progression through the cell cycle (Figure 3.3). Our differentiated cells secrete invertase, an enzyme which hydrolyzes sucrose into its component monosaccharides, glucose and fructose. These simple sugars have a high probability of diffusing away from the cell which originally produced them, and can therefore be consumed by undifferentiated cells in the same medium. Invertase secretion is a canonical system for the study of cooperation between cells [11, 12, 13, 15]. Budding yeast invertase has two isoforms, which result from transcription initiation at distinct start sites [6]. The first produces a full-length peptide containing an N-terminal signal sequence that targets the protein to the secretory pathway,

[2]It would be a philosophical challenge, for example, to accurately estimate the range of benefits that could be derived from the theoretical byproducts of a differentiated cell.

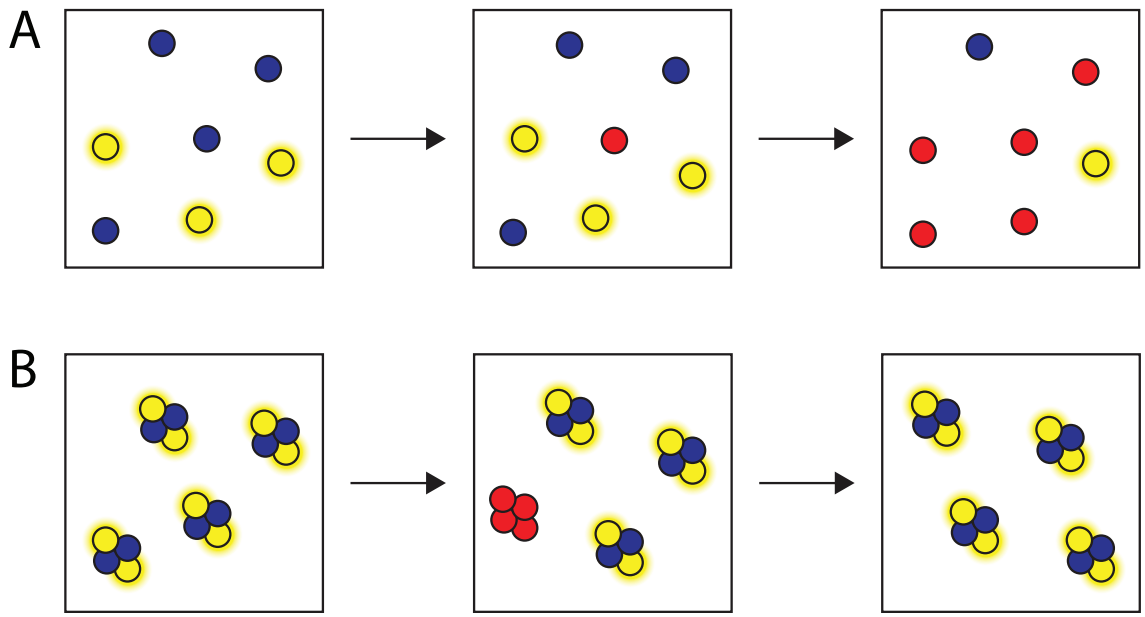

Figure 3.2: Illustration of the hypothesis that unicellular differentiation is inherently unstable.

(A) In a hypothetical unicellular, terminally-differentiating species, differentiated cells (yellow) perform some useful function for undifferentiated cells (blue) that justifies the costs of their production. Since both cell types interact only through the medium, the differentiated cell function must consist of secreting a product into the medium. If an undifferentiated cell mutates so that it cannot differentiate (red), it will still have equal access to the differentiated cell products present in the medium, but will not invest any cell divisions in the production of differentiated cells. The mutant cell thus has a fitness advantage and spreads through the culture, overtaking it.
(B) In a clonally-multicellular, terminally-differentiating species, non-differentiating mutants may be close to differentiated cells when they first appear, but fragmentation of the cell clumps during continued growth will eventually cause the non-differentiating cells to be farther on average from the differentiated cells. Normal undifferentiated cells, meanwhile, will be close enough to differentiated cells to have preferential access to differentiated cell products by diffusion.

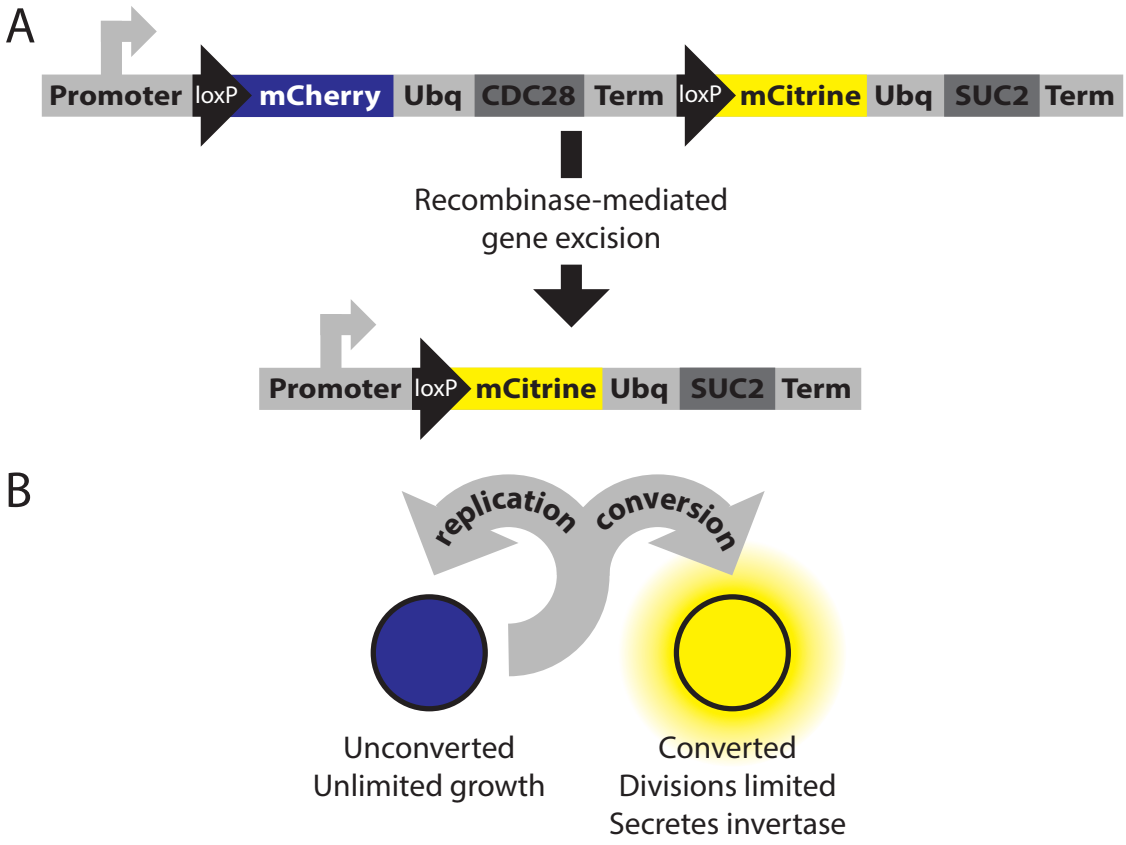

Figure 3.3: Schematic of the Cdc28-based terminal differentiation system.

(A) Diagram of the locus conferring cell type-specific gene expression. In unconverted cells, a strong promoter (P_{ENO2}) drives expression of the polypeptide mCherry-Ubq-Cdc28; a transcriptional terminator prevents expression of downstream genes. After translation, the polypeptide is cleaved at the ubiquitin C terminus by cellular proteases to release a stable fluorescent reporter-ubiquitin fusion protein and the cyclin-dependent protein kinase Cdc28. Conversion ("differentiation") occurs via gene excision: mCherry-Ubq-Cdc28 expression halts and mCitrine-Ubq-Suc2 (encoding a second fluorescent marker and the secreted invertase Suc2) production begins.

(B) Diagram of functionality conferred to each cell type. Unconverted cells express the fluorescent marker mCherry (pseudo-colored blue throughout this document) and progress normally through the cell cycle. Cre recombinase activity in newborn daughters of unconverted cells may induce their conversion, which results in expression of the yellow fluorescent marker mCitrine, a limited number of further divisions, and secretion of invertase.

where it can hydrolyze extracellular sucrose; the second, shorter variant lacks the signal sequence and thus remains in the cytoplasm. Many wild isolates of *S. cerevisiae* express a maltose transporter, Mal11, that facilitates uptake of sucrose into the cytoplasm [4]. The commonly-used lab strains S288c and W303 do not express this transporter due to loss of its transcriptional activator, *MAL63* [4]: in these strains, the cytoplasmic variant of invertase therefore has vacuous function, as demonstrated by the failure of strains expressing only the cytoplasmic variant to grow in sucrose media [13]. The difficulty with which Suc^+ strains grow at low densities attests to the high fraction of monosaccharides which are not captured by the cell that produced them: by comparing a culture's monosaccharide production rate to its growth rate, the percentage of monosaccharides captured has been estimated at less than 1% [11]. In dense cultures, colonies, or cell aggregates, the combined local monosaccharide concentration resulting from diffusion away from all productive cells can become high enough to support growth [13, 15]. It is not necessary for all cells to contribute to invertase secretion in order for a culture to grow in sucrose: indeed, Suc^- mutants have a slight fitness advantage that allows them to invade Suc^+ populations [11, 12].

We demonstrate that in our system, invertase secretion has little direct benefit to the differentiated cells which produce it, but improves the culture growth rate by supplying monosaccharides to all cells in the media. We then show that our terminally-differentiating strain is susceptible to invasion by non-differentiating mutants, which take advantage of monosaccharides shared through the media. We further demonstrate that multicellular, differentiating strains can resist this type of invasion, suggesting that multicellular differentiation is a stable strategy under circumstances where unicellular differentiation is not. Finally, we conclude that unicellular, differentiating species are unlikely to serve as intermediates for the evolution of differentiated multicellularity due to their short duration.

3.3 Methods

Strain and plasmid construction

The construct P_{ENO2}-*AI-mCherry-UBQ* was amplified from pMEW90 (see Section 2.3) by PCR and integrated into pFA6a-HIS3MX6 [19] by ligation-independent cloning [17]. The *CDC28* open reading frame and 400 bp of 3' sequence including transcriptional terminator were then amplified from genomic DNA by PCR and integrated at the 3' end of the ubiquitin moiety by ligation-independent cloning. The complete P_{ENO2}-*AI-mCherry-UBQ-CDC28-*T_{CDC28} construct was amplified by PCR, adding flanking NotI and XmaI restriction sites with which the amplicon was integrated in place of P_{ENO2}-*AI-mCherry-UBQ-cyh2*^{r}-T_{CYH2} in pMEW90 to generate pMEW94.

Bsu36I/SacI-digested pMEW94 was integrated at the *SUC2* locus of yMEW151 (see Table 2.3) by homologous recombination to generate yMEW169. To delete the native *CDC28* locus, the *LEU2* marker was amplified from pRS406 [30], adding homology to the 5' and 3' regions of *CDC28* by extension PCR; this amplicon was then integrated into yMEW151 at the *CDC28* locus of yMEW169 to create yMEW170. Tryptophan and histidine auxotrophies were corrected by homologous recombination at the native loci to produce yMEW194.

To produce a corresponding strain which lacked Cre recombinase for use in competition assays, the Bsu36I/SacI fragment of pMEW90 was integrated into yMEW17 (see Table 2.3) to create yMEW178. The *ace2-4-472* mutation was introduced into yMEW163 and yMEW178 by a *URA3* loop-in/loop-out strategy using pJHK167 [15] to produce yMEW182 and yMEW184, respectively. An *AMN1* gain-of-function allele from strain RM11 [36] was introduced into yMEW163 and yMEW178 via a loop-in/loop-out strategy using pEF607 to generate yMEW179 and yMEW181, respectively. Deletion of *CTS1* was performed by integration of the *HIS3* marker amplified from pRS403 [30] through homologous recombination: yMEW208 and yMEW209 were created from yMEW163 and yMEW178 by *CTS1* deletion and *TRP1* correction.

The construction of others strains mentioned in this chapter is described in Section 2.3 and Table 2.3.

Yeast strains and culture media

Yeast media and culturing conditions were as described in Section 2.3.

Measurement of cell division number following *CDC28* excision

To induce a brief pulse of conversion events, 1 μM β-estradiol was added to log-phase cultures of unconverted cells for one hour prior to cell division (less than one cell cycle). Cells were then washed twice with phosphate-buffered saline and streaked onto a YPD plate. A micromanipulation needle was used to position individual cells with sufficient spacing to allow unrestricted growth. Converted and unconverted cells were not separated prior to plating: the growth of unconverted cells into normal colonies provided an internal control for growth and handling. The number of cells per microcolony was determined by separating them into a monolayer with the dissection needle.

Flow chamber assays were performed as described in Section 2.3. Log-phase cultures of unconverted cells were loaded into the flow chamber and grown in YPD + 1 μM β-estradiol to induce conversion throughout the experiment: while only cells that converted within the first division were used for analysis, continued β-estradiol application prevented unconverted cells from quickly spreading throughout the fields of view. Timelapse videos were recorded and the number of descendants per converted cell determined through frame-by-frame analysis. The average number of divisions per converted cell was inferred from the final number of descendants.

Table 3.1: Strains used in Chapter 3

Strain name	Genotype	Plasmids used
yMEW169	*ade2-1 his3-11 HPHMX4::P_{ACT1}-yCerulean@LEU2 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-CDC28-T_{CDC28}-AI-ymCitrine-UBQ-SUC2*	pMEW94, pDL12
yMEW170	*ade2-1 his3-11 HPHMX4::P_{ACT1}-yCerulean@LEU2 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-CDC28-T_{CDC28}-AI-ymCitrine-UBQ-SUC2 LEU2@CDC28*	pMEW94, pDL12
yMEW178	*ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100 bar1 BUD4 ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2@SUC2*	pMEW90
yMEW179	*ade2-1 his3-11 HPHMX4::P_{ACT1}-yCerulean@LEU2 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2 AMN1-RM11*	pMEW90, pDL12
yMEW181	*ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100 bar1 BUD4 ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2@SUC2 AMN1-RM11*	pMEW90
yMEW182	*ade2-1 his3-11 HPHMX4::P_{ACT1}-yCerulean@LEU2 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2 ace2-4-472*	pMEW90, pDL12
yMEW184	*ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100 bar1 BUD4 ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2@SUC2 ace2-4-472*	pMEW90
yMEW185	*ade2-1 his3-11 HPHMX4::P_{ACT1}-yCerulean@LEU2 trp1-1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-CDC28-T_{CDC28}-AI-ymCitrine-UBQ-SUC2 LEU2@CDC28 ace2-4-472*	pMEW94, pDL12
yMEW194	*ade2-1 HIS3 HPHMX4::P_{ACT1}-yCerulean@LEU2 TRP1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-CDC28-T_{CDC28}-AI-ymCitrine-UBQ-SUC2 LEU2@CDC28*	pMEW94, pDL12
yMEW208	*ade2-1 his3-11 HPHMX4::P_{ACT1}-yCerulean@LEU2 TRP1 ura3-1 can1-100 bar1 BUD4 P_{SCW11}-Cre-EBD78::NATMX4@HO ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2 HIS3@CTS1*	pMEW90, pDL12
yMEW209	*ade2-1 his3-11,15 leu2-3,112 TRP1 ura3-1 can1-100 bar1 BUD4 ADE2::P_{ENO2}-AI-ymCherry-UBQ-cyh2^r-T_{CYH2}-AI-ymCitrine-UBQ-SUC2@SUC2 HIS3@CTS1*	pMEW90

Competition assays

Strains used in competition assays were pre-grown in like media for 24h in log phase, mixed at a 1:1 ratio, washed with phosphate-buffered saline to remove existing monosaccharides, and resuspended in fresh media. Cultures were passaged at low densities (10^4 – 10^6 cells/mL) to limit accumulation of monosaccharides in the media. Samples were collected at each dilution for analysis by flow cytometry: the ratio between strain populations was determined on the basis of yCerulean fluorescence. In competition assays, the fitness differences between strains are typically expressed as selection coefficients with units of enrichment rate per generation of a reference strain: this approach could not be used in our case, as strain growth rates were likely dependent on strain frequency and accumulation of monosaccharides in the media with time. We therefore expressed fitness differences between strains as the effective percent change in strain frequency over the course of the experiment (5-6 days and three media changes) divided by the number of culture density doublings during this period.

3.4 Results

Cyclin-dependent protein kinases (CDKs) drive the eukaryotic cell cycle by phosphorylating target proteins required for cell cycle progression. CDKs consist of two subunits: a cyclin whose expression and degradation is regulated throughout the cell cycle and provides substrate specificity, and a catalytic domain which requires the cyclin for activation. Budding yeast possesses five cyclin-dependent protein kinases [22]: two of these, Cdc28 and Kin28, are essential for viability [9, 31]. Cdc28 is required for passage through the G1/S checkpoint (Start), spindle pole body duplication, and entry into mitosis (reviewed in [23]). Inactivation of Cdc28 in temperature-sensitive mutants leads to arrest at the next cell cycle checkpoint, but not to cell death: cells grow in volume throughout the arrest and resume division when returned to the permissive temperature [3, 10, 14].

We have integrated *CDC28* into our conversion system (see Chapter 2) and removed the native locus so that, following gene excision, converted cells will eventually cease division. The arrest is not immediate, however: *CDC28* is expressed under the strong promoter *ENO2*[3], so accumulated *CDC28* transcripts and Cdc28 protein must be diluted by cell division and/or actively degraded before cell division will halt (Figure 3.4).

To test the efficacy of our system in halting growth of converted cells, we determined the fraction of converted (mCitrine$^+$) cells by flow cytometry 7 hours after the Cre recombinase inducer β-estradiol was added to a log-phase culture. The cultures were then spun down, washed twice in phosphate-buffered saline, and resuspended in YPD without β-estradiol. After 17 hours of additional log-phase growth following washout, the fraction of converted cells was determined again. In strains where no essential genes are excised during conversion, the fraction of mCitrine$^+$ cells does not change appreciably during growth after β-estradiol washout (see Figure 2.5 and 3.5); by contrast, the fraction of mCitrine$^+$ cells had

[3] Fortuately, regulation of Cdc28 activity is primarily post-translational and its overexpression is well-tolerated [21]

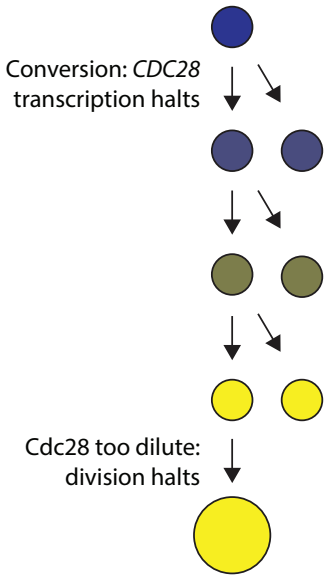

Figure 3.4: Expected cell growth and division pattern following *CDC28* excision.

Transcription of *CDC28* halts immediately after excision of Cre recombinase due to the absence of a promoter on the excised fragment. A large quantity of *CDC28* mRNA and Cdc28 protein will be present, however, due to the high activity level of the *ENO2* promoter. Loss by dilution (and potentially proteolytic degradation) are required before the Cdc28 concentration becomes too low to permit cell cycle progression. Failure to divide is not expected to interfere with the cell's ability to grow in mass/volume.

decreased more than one thousand-fold in our *CDC28* excision strain (Figure 3.5), suggesting that the converted cells had failed to keep pace with unconverted cells in the culture.

To demonstrate that division fully halted, unconverted cultures were exposed to β-estradiol for one hour, then arrayed on agar plates by micromanipulation so that the descendants of individual cells could be easily tracked (Figure 3.6A). Cells which did not convert during the brief induction grew into colonies over the course of two days (Figure 3.6B), while those which had converted formed microcolonies of approx. 30-80 cells (Figure 3.6C). Cells in microcolonies were much larger in size, suggesting that they remained metabolically active long after ceasing cell division.

The number of divisions the average cell undergoes following conversion were estimated from timelapse videos of microcolonies growing in monolayers. Unconverted cell cultures were loaded into a flow chamber and grown for twenty-four hours in YPD + 1 μM β-estradiol to induce conversion. Converted cells appeared to grow at a normal rate prior to halting suddenly after approximately 12 hours (Figure 3.7A-C). Unlike in the micromanipulation experiments above (Figure 3.6C), the cells' continued growth in volume eventually caused them to burst (Figure 3.7D-F), likely due to confinement in the flow chamber. The microcolonies' monolayer growth facilitated counting of descendants from single conver-

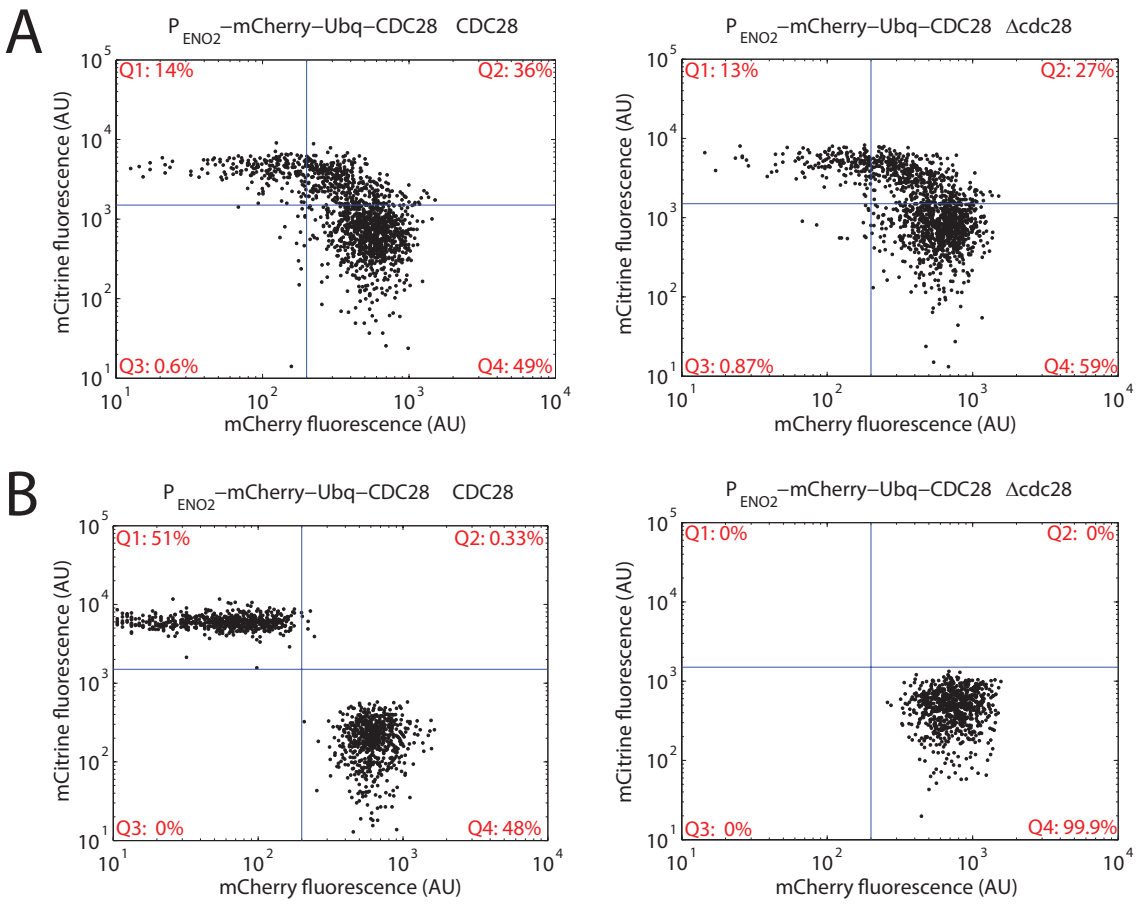

Figure 3.5: Loss of converted cells by dilution following *CDC28* excision.

(A) Strains containing the P_{ENO2}-*AI-mCherry-UBQ-CDC28-*T_{CDC28} construct with (yMEW169) and without (yMEW170) *CDC28* also expressed from its native locus were grown in 1 μM β-estradiol for seven hours. Both strains showed significant conversion to the mCitrine$^+$ state.
(B) Following the treatment in (A), cultures were washed and resuspended in YPD without β-estradiol and grown in log phase for seventeen hours (approx. 11 divisions for unconverted cells). Converted cells persisted in the yMEW169 culture due to expression of *CDC28* from the native locus.

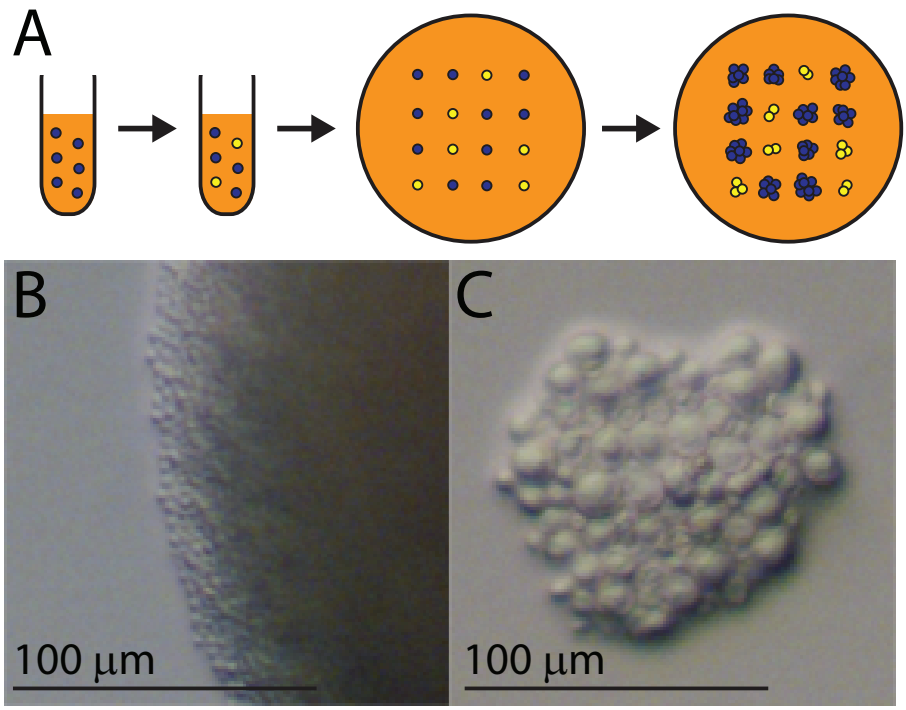

Figure 3.6: Estimation of average division number following *CDC28* loopout using micromanipulation

(A) Schematic of the experiment. To induce a pulse of conversion events, 1 μM β-estradiol was added to a pure log phase culture of unconverted cells (yMEW194) for one hour (less than one cell division length). After β-estradiol washout, individual cells were separated and positioned on a YPD plate using a dissection needle. After two days, unconverted cells had formed large colonies while converted cells had produced a small number of daughters, which could then be counted using the micromanipulation needle.
(B) Representative image of the frontier of an unconverted colony after 2 days' growth.
(C) Representative image of a microcolony formed from a converted cell after 2 days' growth.

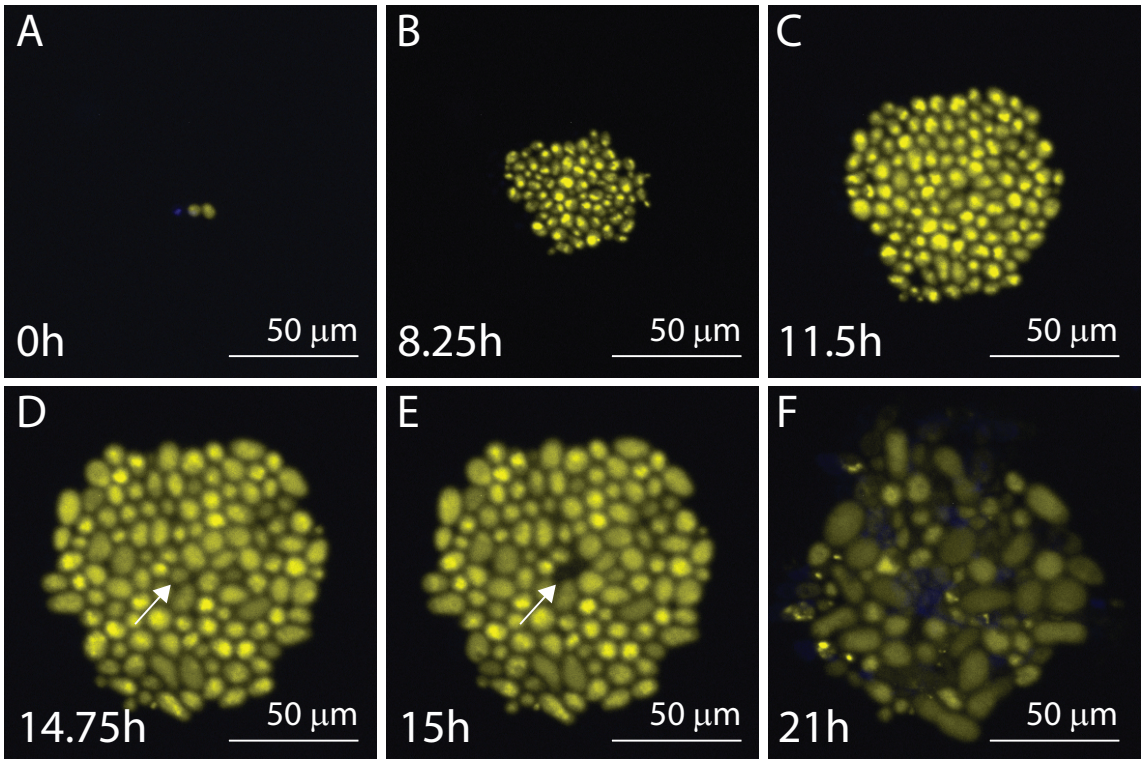

Figure 3.7: Still images from a timelapse video of microcolony formation following *CDC28* excision.

To induce a pulse of conversion events, 1 μM β-estradiol was added to a pure log phase culture of unconverted cells (yMEW194). The culture was loaded into a flow chamber after one hour (less than one cell division length) and individual converted cells were identified for timelapse imaging. [Identification of converted cells was time-consuming, and as a result the converted cell has already completed one cell division before the first frame (A) was captured.] Growth appeared normal for several divisions (B), but eventually slowed to a halt, with the last division occurring in frame (C). Existing cells continued to grow in size (D) and began to pop (E, see arrows), causing the accumulation of autofluorescent cellular debris (F).

sion events and thus the determination of the number of divisions required on average prior to arrest (Figure 3.8).

Growing cultures will eventually reach an equilibrium between the production of new converted cells and the loss by dilution of old ones: the fraction of converted cells at steady-state is determined by the conversion rate and number of divisions following conversion before arrest, both parameters which have been measured from our data. Our design re-

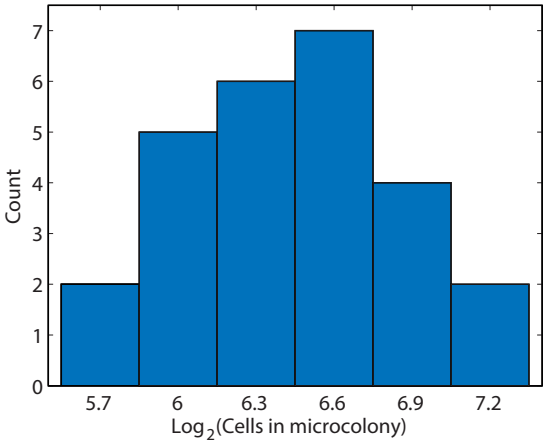

Figure 3.8: Number of divisions prior to cell cycle arrest following *CDC28* gene excision.

Timelapse videos of microcolony formation following *CDC28* excision were recorded as described in Figure 3.7. The maximum number of cells present in each microcolony was determined through frame-by-frame visual inspection; this cell count was converted to an estimate of the number of divisions experienced prior to cell cycle arrest. The mean number of divisions observed was 6.42 ± 0.41.

quired that converted cells secrete invertase to support growth of both cell types in sucrose medium: to perform this function adequately, converted cells must make up a non-negligible portion of the population. We were therefore interested in determining the expected fraction of converted cells at steady-state to ascertain whether it was likely to be sufficient to sustain a culture in sucrose media. Under the model that converted cells grow at a normal rate until Cdc28 is too dilute for progression through the cell cycle, we predict that the number of cells that are unconverted (n_u) and which have divided i times since converting ($n_{c,i}$) change with time according to a system of linear first-order differential equations. Let $\vec{n}(t) = (n_u(t), n_{c,1}(t), \ldots n_{c,k}(t))^T$ represent the number of cells of each type at time t. Then the rates of change in population should be given by:

$$\frac{\partial \vec{n}(t)}{\partial t} = \gamma \begin{pmatrix} 1-\mu & 0 & \cdots & 0 & \\ \mu & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} n_u \\ n_{c,0} \\ n_{c,1} \\ \vdots \\ n_{c,k} \end{pmatrix} (t) = \gamma \mathcal{A} \vec{n}(t)$$

where γ is the growth rate of all cells capable of division, μ is the conversion rate, and k is the number of divisions possible after conversion. This system has solutions of the form:

$$\vec{n}(t) = \sum_{i=1}^{k+1} a_i \vec{v}_i e^{\gamma \lambda_i t}$$

where λ_i and $\vec{v}_i$ are the eigenvalues and corresponding eigenvectors of $\mathcal{A}$, and the constants a_i may be chosen to satisfy an initial condition. At long times, the term with the largest eigenvalue will dominate this sum, and its corresponding eigenvector will therefore represent the steady-state ratio between cells in each class. The conversion rate μ has been experimentally determined for a range of β-estradiol concentrations (see Figure 2.7), allowing this system to be solved numerically. As the latter was estimated to be 6.42 ± 0.41, we compared the predictions of the model for $k = 6$ and $k = 7$ to steady-state ratio measurements obtained by passaging cultures in YPD containing different concentrations of β-estradiol. Our model predicted that the population fraction of converted cells could be tuned from <5% to > 70% by altering the β-estradiol concentration, in accord with experimental results (Figure 3.9).

Cultures grown in sucrose had higher steady-state fractions of converted cells than their counterparts grown in glucose at the same β-estradiol concentration (Figure 3.10A), likely reflecting a decrease in growth rate specific to unconverted cells[4]. The increase was particularly substantial for lower conversion rates, suggesting an inherent difficulty in culture growth when the fraction of Suc$^+$ cells is less than ≈50%. At the converted cell fractions observed observed for these β-estradiol concentrations, converted cells were able to support growth of the culture in sucrose media, as expected (Figure 3.10B).

One might imagine that if the conversion rate is too low, the growth rate of a culture in sucrose medium would be limited primarily by sucrose production rather than the limited growth of Δ*cdc28* convertants. In this case, there would be a marginal utility to increasing the conversion rate: the benefit of producing more Suc$^+$ cells would outweigh the costs of

[4] Even in shaken liquid cultures, converted cells have slight preferential access to monosaccharides because of the sugars' proximity immediately after hydrolysis by invertase in the converted cells' walls.

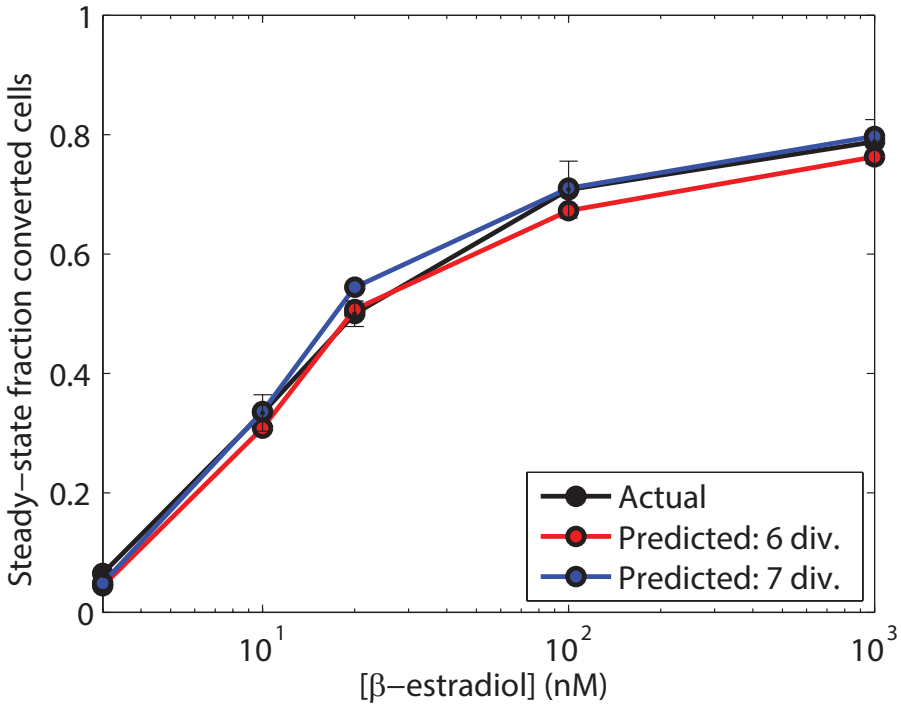

Figure 3.9: Predicted and observed steady-state fraction of converted cells vs. [β-estradiol]

Cultures of a converting strain (yMEW194) were passaged in YPD media containing β-estradiol at the indicated concentrations until the fraction of converted (mCitrine$^+$) cells did not change significantly between two timepoints and was thus considered to be at steady-state (black line). Predicted steady-state ratios assuming that converted cells divide at a normal rate for n divisions, then halt, are also shown (red line: 6 divisions; blue line: 7 divisions).

producing more differentiated cells: in other words, there would be an intermediate concentration of β-estradiol for which the culture growth rate would peak. (By contrast, growth rate should always decrease with conversion rate in glucose media.) With the exception of the fact that our cultures cannot grow in sucrose without some conversion, we observed a monotonic decrease in culture growth rate with β-estradiol concentration (Figure 3.10B). This suggests that the benefits of generating a converted daughter cell by *CDC28* excision do not justify the costs. While increasing the number of divisions per daughter cell is theoretically possible in this system – genomic modifications could increase the strength of the promoter or decrease the degradation rates of transcript and protein – no obvious avenues were identified for increasing division counts substantially.

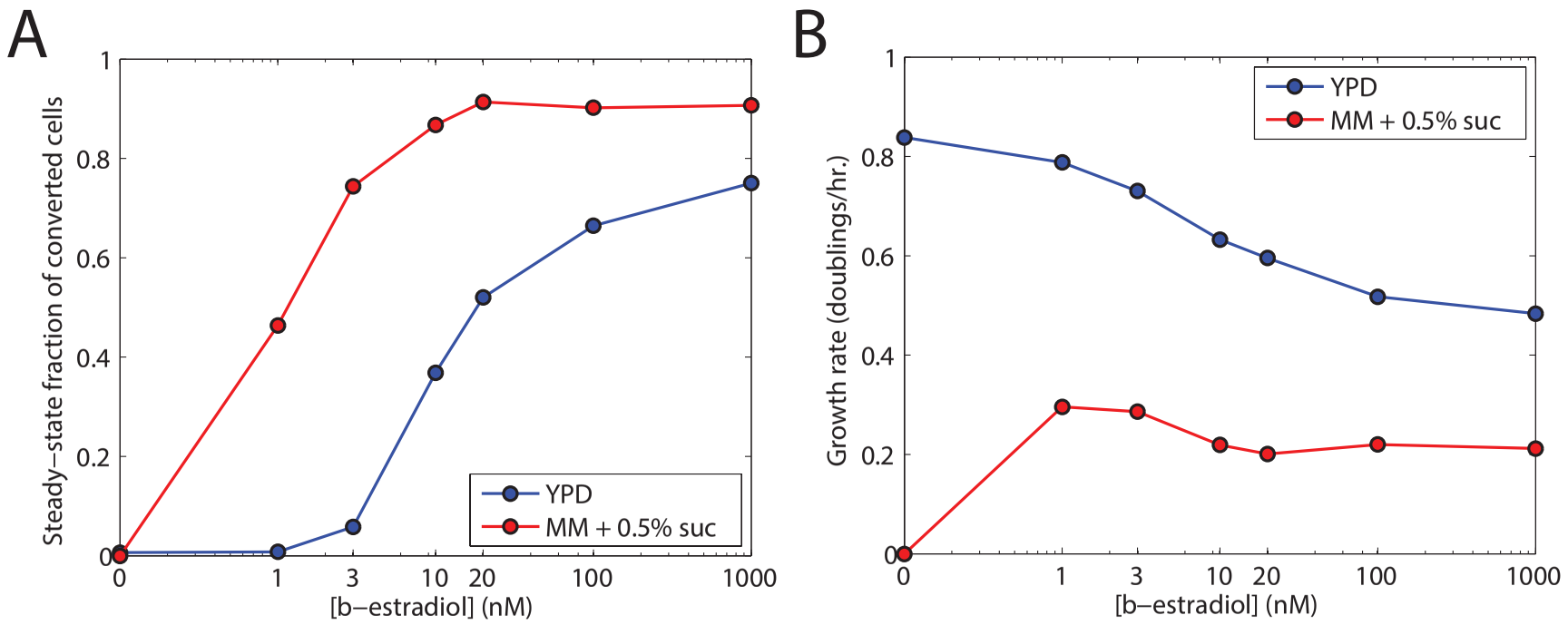

Figure 3.10: Δ*cdc28* convertants support culture growth in sucrose media

(A) The steady-state fraction of converted cells is higher for cultures utilizing sucrose as a carbon source. Cultures of a converting strain (yMEW194) were passaged in 2% dextrose YPD or 0.5% sucrose minimal media at the indicated β-estradiol concentrations until the fracton of converted (mCitrine$^+$) cells did not change between two successive timepoints.
(B) A growth curve was performed using the endpoint cultures from (A) in same media after a 4 hour adaptation period following back-dilution.

We therefore turned to an alternative system for limiting the growth of converted cells: instead of excising *CDC28* during conversion, converting cells excised an allele conferring tolerance to the antibiotic cycloheximide (Figure 2.1A). In this strain, unconverted cells are partially resistant to the growth-limiting effects of cycloheximide due to expression of the ribosomal protein L28 allele $cyh2^R$; after conversion, $Cyh2^r$ protein and transcript levels decrease through growth and division, ultimately leaving converted cells fully sensitive to cycloheximide. Unlike for cyclin-dependent protein kinase excision, isolated cycloheximide-sensitive convertants are capable of indefinite growth. However, in the context of cultures containing both cell types, converted cells have an effectively finite lifespan because of their slower growth rate. We thus feel that this strain also meets our definition of terminal differentiation.

Cultures of the floxed $cyh2^R$ strain described above were unable to grow substantially in sucrose media containing cycloheximide when only one cell type was present: unconverted cells alone did not produce the invertase necessary for growth, and converted cells were significantly impeded by the antibiotic (Figure 3.11). The growth rate of cultures that contained both cell types was significantly faster than cultures of either cell type alone, and did not appear to decrease monotonically with conversion rate (Figure 3.11). This suggested cooperation between cell types for the benefit of the genotype's propagation, akin to differentiation systems found in nature.

Clonal multicellularity was easily introduced by exploiting the biological properties of budding yeast. In *S. cerevisiae*, the spindle pole body-associated kinase Tem1 is activated in the bud, where it initiates a daughter cell-specific branch of the mitotic exit network following cytokinesis [1, 2, 29, 33]. This pathway ultimately leads to nuclearization of the transcription factor Ace2 [20], which in turn drives expression of effector genes required for the cell wall remodeling that permits full separation of mother and daughter cells, such as the chitinase *CTS1* [8, 24] and the glucanase *SCW11* [5, 34]. Ace2 also limits the duration of its own activity through a negative feedback loop by driving expression of *AMN1*, which com-

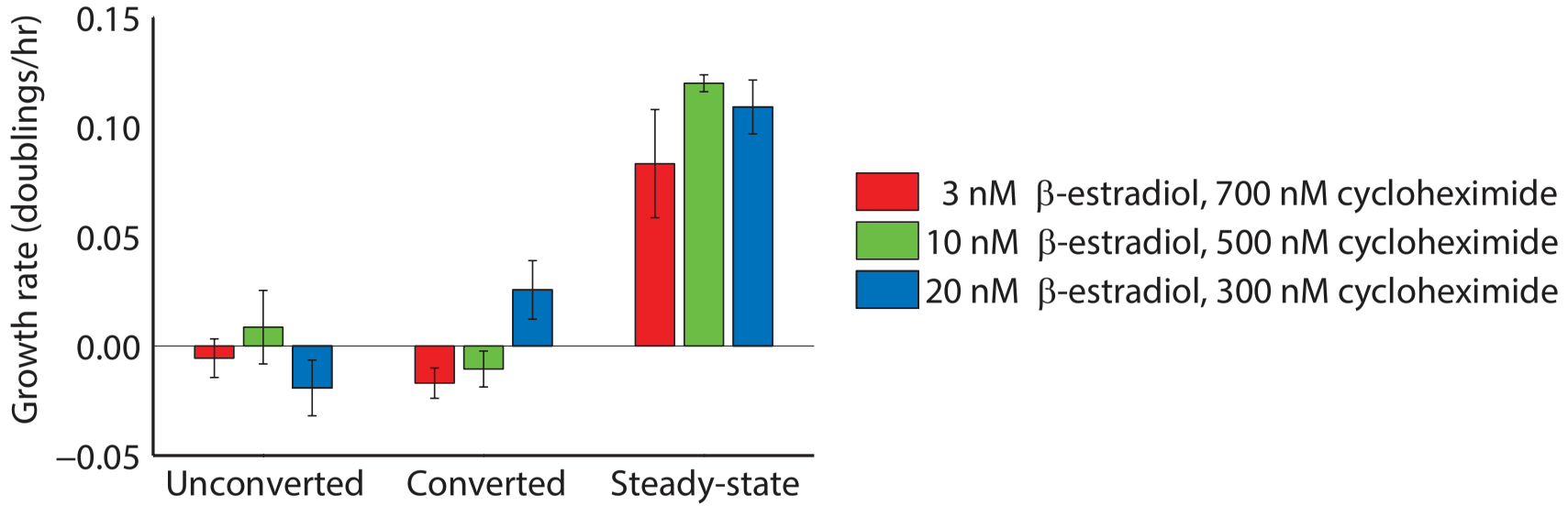

Figure 3.11: Both cell types are required for efficient culture growth in sucrose media containing cycloheximide.

Cultures consisting of only unconverted cells (yMEW193: *cyh2*^{r} Suc^{-} Cre^{-}), only converted cells (yMEW192c: *cyh2*^{s} Suc^{+}), or both cell types at a steady-state ratio (yMEW192) were grown in 0.5% sucrose minimal media at the indicated cycloheximide and β-estradiol concentrations. The fraction of converted cells at steady state were determined by flow cytometry to be 15% (red), 72% (green), and 92% (blue).

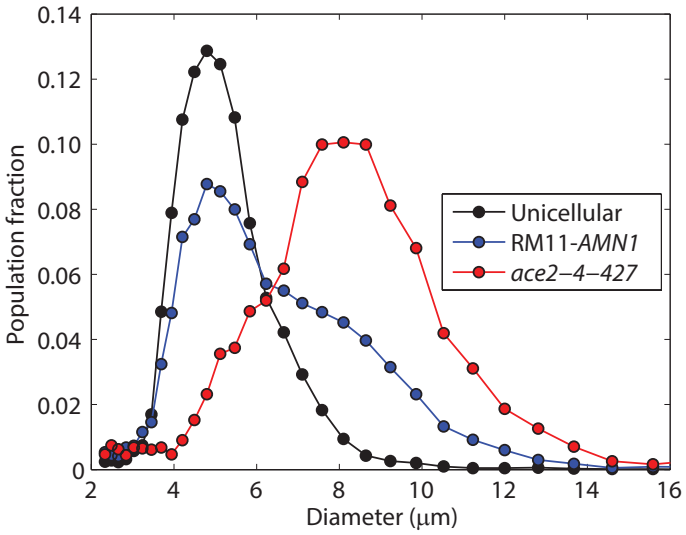

Figure 3.12: Distribution of cell clump diameters caused by daughter cell-specific mitotic exit network mutations.

The distribution of cell clump diameters for log phase cultures of unicellular (yMEW163), *AMN1* gain-of-function mutation (yMEW181), and *ace2* loss-of-function mutation (yMEW185) strains was determined by Coulter counting.

petitively binds to partners of Tem1 to halt mitotic exit network activity [7, 35]. Mutations in this pathway are known to cause failure of daughter cell separation and thus the formation of clumps of cells related by descent [5, 16, 34, 36][5]. These cell clumps do not grow indefinitely: they typically fragment pseudo-randomly due to a combination of buffeting in mixed media and eventual cell wall remodeling due to the activity of other chitinases.[6]

Mutations in mitotic exit network proteins were used to introduce "multicellularity" into our strains. Clump sizes attained were highly variable due to the random nature of fragmentation; clump volumes inferred from Coulter counter particle diameters (Figure 3.12 and Table 3.2) were smaller than those observed by microscopy (Figure 3.13) but reflected the

[5]This phenomenon is akin to clonal multicellularity. Aggregative multicellularity, where potentially unrelated cells join together with the help of cell surface proteins, is predicted to be less effective at preventing the spread of "cheating" mutants, as outlined in the introductory chapter. Aggregative multicellularity occurs naturally in budding yeast during the process of flocculation, and can be easily achieved by overexpressing *FLO* family proteins [32]

[6]It can be inferred that the fragmentation is not due to force exerted by continue growth alone, since much larger clump sizes can be achieved through experimental evolution[15, 25].

Table 3.2: Mean and median sizes of cell clumps in daughter cell-specific mitotic exit network mutants.

Strain	Mean clump diameter (μm)	Std. dev. in clump diameter (μm)	Median clump diameter (μm)
Unicellular	5.41	0.38	4.87
AMN1 from RM11	6.02	0.71	5.44
ace2-4-427	7.60	0.28	7.71

overall trends in clump size. A partial loss-of-function mutation in *ACE2* (*ace2-4-472* [15]), the most upstream component of the pathway tested, generated clumps of approximately 4-30 cells (Figure 3.13B). A gain-of-function mutation in *AMN1* and deletion of *CTS1* also caused clumping, albeit less severe (Figure 3.13C-D). Clumps produced in the *ace2* mutant were large enough to frequently contain both cell types when the conversion rate was sufficiently high (Figure 3.13E).

Working with these mitotic exit network mutations introduced several complications. In our strains, Cre recombinase is expressed from the *SCW11* promoter: conversion rate is therefore lower in mitotic exit network mutants with decreased Ace2 activity. Replacing the promoter of Cre was not attempted for two reasons: first, because the low expression level of this promoter reduces basal Cre activity; and secondly, because expression under the mitotic exit network allows conversion to occur on a "per cell division" basis rather than "per unit time," ensuring that time spent in stationary phase and other slow growth periods does not increase conversion. This issue was resolved by increasing the β-estradiol concentrations used to achieve a similar steady-state ratio between cell types to that observed in unicellular strains used as controls.

Another confounding issue was that Cdc28 is the kinase that obscures the Ace2 nuclear localization sequence [24, 26] through phosphorylation: overexpression of *CDC28* from the *ENO2* promoter might therefore be expected to exacerbate the partial loss of Ace2 activity in mitotic exit network mutants. Indeed, an *ace2-4-427 CDC28* excision strain appeared to

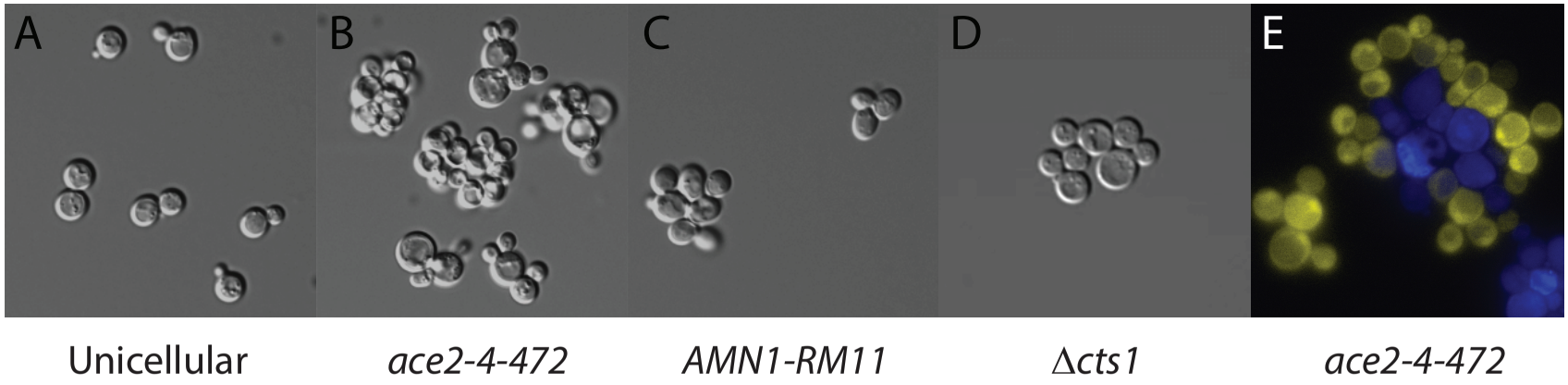

Figure 3.13: Clonal multicellularity is conferred by disruption of the daughter cell mitotic exit network or downstream effectors.

Yeast strains with functional *BUD4*, including our *BUD4*-corrected W303 strains (A), are primarily found as single cells in well-mixed liquid media. Yeast strains carrying (B) the partial loss-of-function mutation *ace2-4-472* [15], (C) *AMN1-RM11*, the gain-of-function mutation in *AMN1* found in strain RM11, or (D) a deletion of the chitinase *CTS1* experience frequent failure of daughter cell separation following cytokinesis, and therefore are typically found in multicellular clumps of cells related by descent. These clumps are large enough to often contain both cell types when our converting, multicellular strains [e.g. yMEW182, (E)] are cultured at high β-estradiol concentrations (here, flattened to demonstrate presence of both cell types).

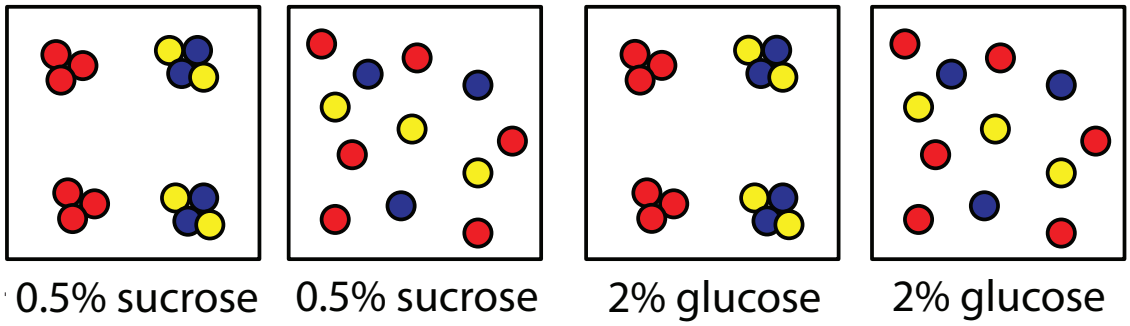

Figure 3.14: Growth conditions used for the competition assay in Figure 3.15

Unicellular (yMEW163) or multicellular (yMEW182) converting strains were passaged in the competition assay media until the steady-state ratio between cell types was reached (blue and yellow cells), then mixed with a Cre$^-$ "cheater" strain (red cells). The Cre$^-$ competitor strains were either unicellular (yMEW178) or multicellular (yMEW184), matched to the converting strain to minimize fitness differences. Competition assays were performed either in glucose (where there is no predicted fitness benefit for a converting strain) or sucrose media.

convert at a much lower rate than a corresponding *ace2-4-427 cyh2*^{r} excision strain (data not shown).

We predicted that Cre$^-$ mutants would be able to invade a unicellular, converting population growing in sucrose since the mutant cells would have equal access to monosaccharides but would never expend cell divisions on the production of Suc$^+$ daughters. By contrast, we hypothesized that Cre$^-$ mutants would be at a disadvantage when co-cultured with a multicellular, converting strain, since in this latter case the mutant cells would access monosaccharides only through the bulk medium while converting cells would benefit from proximity to Suc$^+$ cells in the same clump. To test these predictions, we initiated cultures from mixtures of converting and Cre$^-$ cells and measured the change in ratio between strains by flow cytometry as the cultures were passaged. To eliminate potential sources of fitness differences, the Cre$^-$ strain used in each assay was multicellular if the converting strain was multicellular, and unicellular otherwise (Figure 3.14). As a control, the competition assays were also performed in glucose media, where there is expected to be no advantage for the converting strain.

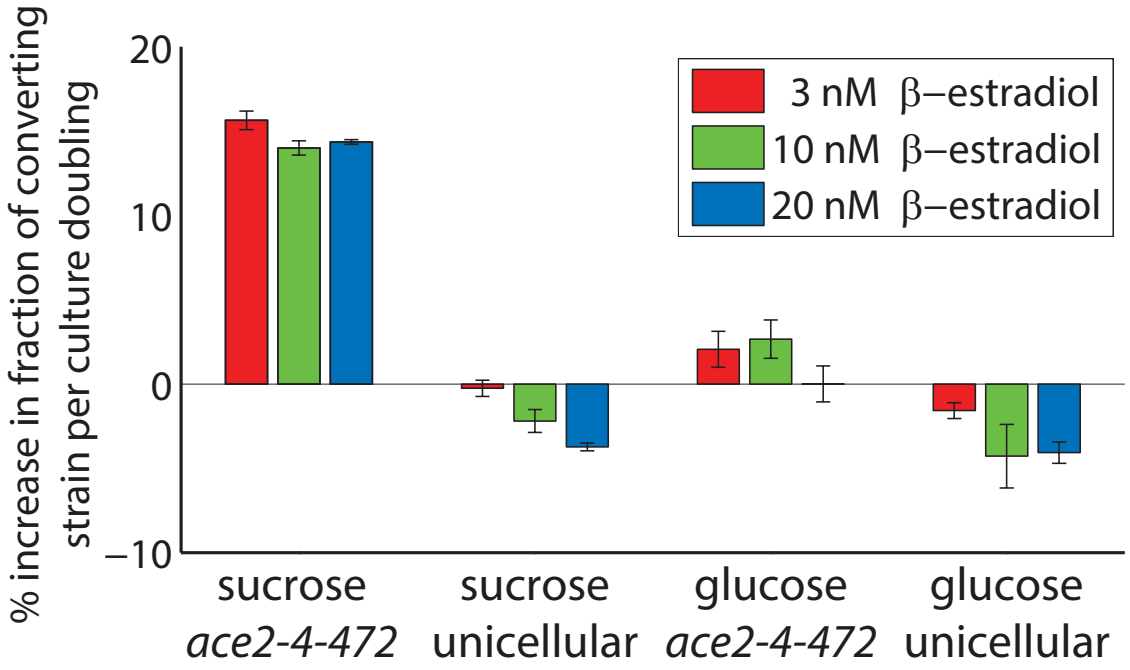

Figure 3.15: Multicellular *cyh2^r* excision converting strains are able to resist invasion by non-converting cheaters.

Cultures of a Cerulean$^+$ *cyh2^r* excision converting strain (unicellular: yMEW163; multicellular: yMEW182) and a Cerulean$^-$ Cre$^-$ "cheater" (unicellular: yMEW178; multicellular: yMEW184) were mixed at a 1:1 ratio, washed, and resuspended into 2% glucose or 0.5% sucrose minimal media containing 500 nM cycloheximide and the indicated concentration of β-estradiol. Samples were analyzed by flow cytometry during culture passaging to determine the change in ratio between strains with time. While the β-estradiol concentrations used for the unicellular and multicellular strains were identical, it should be noted that the conversion rates of the *ace2-4-427* P_{ENO2} strains were significantly lower at each β-estradiol concentration.

We found that the multicellular *cyh2^r* excision converting strain was able to compete successfully against the Cre$^-$ strain when grown in sucrose: the population fraction of Cre$^-$ cells decreased steadily and substantially with time (Figure 3.15). Consistent with expectation, no enrichment of the converting strain occurred in glucose media. The population fraction of the unicellular converting strain, however, remained steady when the conversion rate was very low (i.e., when the low conversion rate did not significantly decrease fitness) but lost ground as conversion rate increased, regardless of the media carbon source.

When the analogous experiment was conducted using the *CDC28* excision strains, no parameter regime could be found in which the multicellular converting strain had a significant fitness advantage over the non-converting strain (Figure 3.16 and data not shown). The population fraction of the multicellular converting strain did not change substantially with time while growing in sucrose, though it was rapidly outcompeted in glucose: this suggests a stabilizing effect to multicellulary for the converting phenotype similar to that seen in the *cyh2*^{r} excision strain (Figure 3.15).

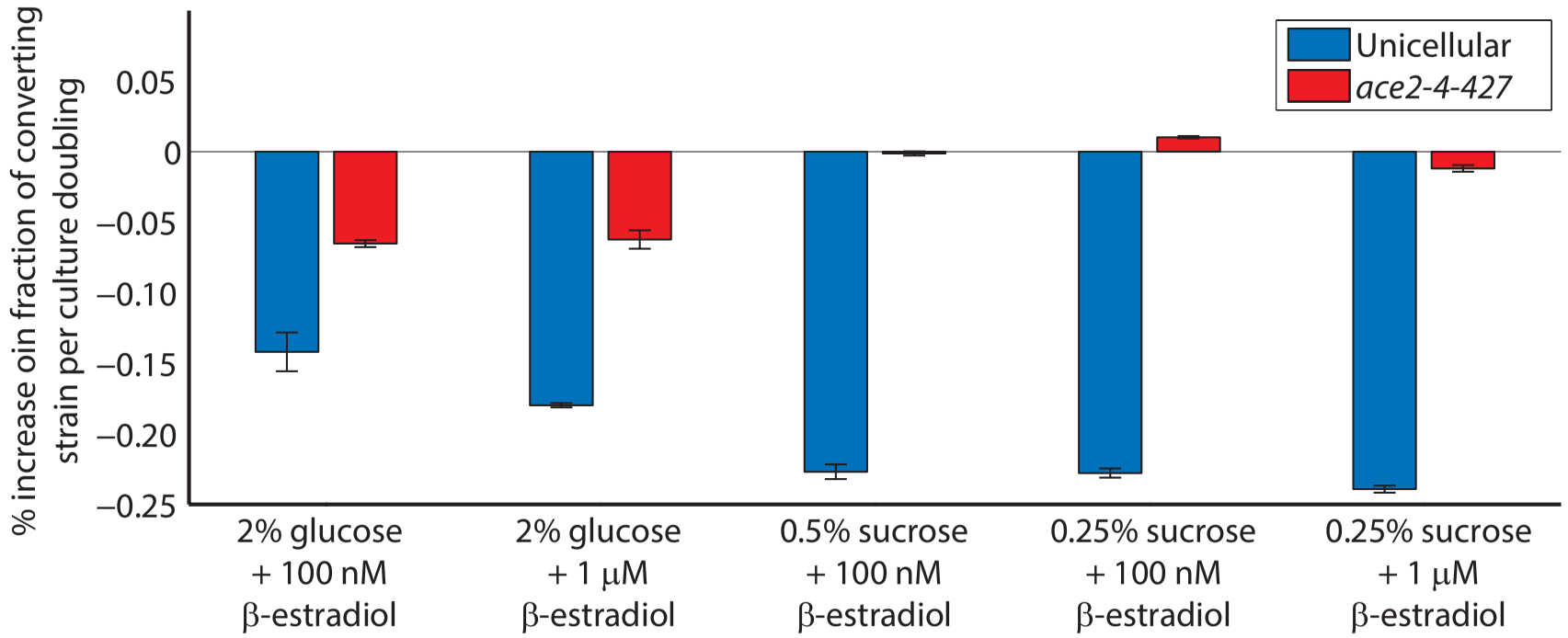

Figure 3.16: Multicellular *CDC28* excision converting strains partially resist invasion by non-converting cheaters.

Cultures of a Cerulean$^+$ *CDC28* excision converting strain (unicellular: yMEW170; multicellular: yMEW185) and a Cerulean$^-$ Cre$^-$ "cheater" (unicellular: yMEW178; multicellular: yMEW184) were mixed at a 1:1 ratio, washed, and resuspended into media containing the indicated carbon sources and concentrations of β-estradiol. Samples were analyzed by flow cytometry during culture passaging to determine the change in ratio between strains with time. While the β-estradiol concentrations used for the unicellular and multicellular strains were identical, it should be noted that the conversion rates of the *ace2-4-427* P_{ENO2} strains were significantly lower, as described above.

3.5 Discussion

We have developed a synthetic biological system to implement terminal differentiation in budding yeast. Differentiated cells in our system cannot divide indefinitely due to the excision of an essential cell cycle gene, but are otherwise healthy: they grow significantly in volume and continue secreting invertase. In long-term cultures, the fraction of differentiated cells reaches a steady-state which can be tuned (by altering Cre recombinase activity through changes in β-estradiol concentration) to achieve a ratio sufficient to support culture growth in sucrose media and to maximize the probability that both cell types are present in multicellular clumps.

Our choice of differentiated cell function was arbitrary: invertase secretion simply represents a general class of processes that require sharing differentiated cell products through the extracellular milieu. We have argued that this is the only kind of function which differentiated cells can perform in a unicellular species, since it does not require immediate physical contact between cells. The finding that our unicellular, terminally-differentiating strain is prone to invasion by non-differentiating mutants can therefore be applied generally to explain the dearth of extant unicellular, differentiating species. Furthermore, it is unlikely that such species would serve as intermediates for the evolution of differentiated multicellularity, since their short duration limits opportunities for the evolution of multicellularity[7].

The relegation of costly processes to somatic cells can allow germline cells to dedicate a larger fraction of their energy to reproduction, thus speeding up reproduction while maintaining viability through the activities of the soma. Division of labor – the separation of tasks between cell types – can also lead to improvements in the efficiency with which each task is performed through cellular specialization. These advantages have been credibly argued to be the driving forces behind the evolution of differentiation in some clades, including the cyanobacteria and volvocine algae (see Chapter 1). Our differentiation system

[7] By contrast, undifferentiated multicellular species have persisted in some cases for hundreds of millions of years [27].

partitions reproductive and secretory functions between two cell types, but it is important to note that this engineered division of labor is not actually advantageous, i.e., wildtype yeast strains which express invertase and Cdc28 (or $Cyh2^r$) in all cells grow faster than our differentiating strains. This is to be expected given the low fitness cost (< 1%)of secreting invertase relative to the fitness costs of differentiation, which are equal to the conversion rate.

While the non-optimality of differentiation in our system does not affect the conclusions we have drawn thus far, it does call to light an issue with evolution through a unicellular, differentiating intermediate: in order for such a species to arise, there presumably must be some benefit to differentiation itself, and this is difficult to achieve even by design. Are there any two processes performed in budding yeast that are so costly or mutually-antagonistic that their separation by cellular differentiation would be advantageous? Is the existence of such processes the rate-limiting step in the evolution of differentiation? Could division of labor in our strain be made beneficial by rationally choosing exogenous functions from other species? These questions will be addressed in depth in the concluding chapter.

It is unfortunate that in our system, mitotic exit network mutations used to introduce clonal multicellularity also affect the expression level of Cre recombinase: a clear extension of this work is to find an arrangement that avoids this issue. Attempts to design systems for expressing Cre from alternative promoters were not successful (data not shown). The cell cycle phase-limited and relatively weak expression of Cre recombinase from the *SCW11* promoter was likely critical for the low basal activity and wide range of inducible activity levels observed in our system: we are therefore indebted to Derek Lindstrom and Dan Gottschling for the significant effort they invested in the design, random PCR mutagenesis, and selection of the P_{SCW11}-*Cre-EBD78* they generously shared [18]. Furthermore, since expressing Cre from the *SCW11* promoter allows conversion rates to be remain constant per division rather than per unit time (a handy constraint when culture growth rates are allowed to vary, e.g., by reaching stationary phase or slowly converting sucrose to monosaccharides), it would be preferable to leave the regulation of Cre as-is and find alternative mitotic exit network muta-

tions downstream of Ace2. We attempted direct mutation of an effector chitinase, Cts1, and found that this produced very small clumps (Figure 3.13), but it is possible that by mutating multiple chitinases and glucanases, larger clump sizes could be achieved.

3.6 Bibliography

[1] K. Asakawa, S. Yoshida, F. Otake, and A. Toh-e. A novel functional domain of Cdc15 kinase is required for its interaction with Tem1 GTPase in *Saccharomyces cerevisiae*. *Genetics*, 157(4):1437–1450, 2001.

[2] A. J. Bardin, R. Visintin, and A. Amon. A mechanism for coupling exit from mitosis to partitioning of the nucleus. *Cell*, 102(1):21 – 31, 2000.

[3] A. Bishop, J. Ubersax, D. Petsch, D. Matheos, N. Gray, J. Blethrow, E. Shimizu, J. Tsien, P. Schultz, M. Rose, J. Wood, D. Morgan, and K. Shokat. A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature*, 407:395–401, 2000.

[4] C. A. Brown, A. W. Murray, and K. J. Verstrepen. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Current Biology*, 20(10):895 – 903, 2010.

[5] C. Cappellaro, V. Mrsa, and W. Tanner. New potential cell wall glucanases of *Saccharomyces cerevisiae* and their involvement in mating. *Journal of Bacteriology*, 180(19):5030–5037, 1998.

[6] M. Carlson and D. Botstein. Two differentially regulated mRNAs with different 5' ends encode secreted and intracellular forms of yeast invertase. *Cell*, 28(1):145–154, 1982.

[7] A. Colman-Lerner, T. E. Chin, and R. Brent. Yeast Cbk1 and Mob2 activate daughter-specific genetic programs to induce asymmetric cell fates. *Cell*, 107(6):739 – 750, 2001.

[8] P. R. Dohrmann, W. P. Voth, and D. J. Stillman. Role of negative regulation in promoter specificity of the homologous transcriptional activators Ace2p and Swi5p. *Molecular and Cellular Biology*, 16(4):1746–58, 1996.

[9] G. Giaever, A. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418(6896):387 – 391, 2002.

[10] A. I. Goranov, M. Cook, M. Ricicova, G. Ben-Ari, C. Gonzalez, C. Hansen, M. Tyers, and A. Amon. The rate of cell growth is governed by cell cycle stage. *Genes & Development*, 23(12):1408–1422, 2009.

[11] J. Gore, H. Youk, and A. van Ouenaarden. Snowdrift game dynamics and facultative cheating in yeast. *Nature*, 459(7244):253 – 256, 2009.

[12] D. Greig and M. Travisano. The Prisoner's Dilemma and polymorphism in yeast *SUC* genes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 271(Suppl 3):S25–S26, 2004.

[13] J. H. Koschwanez, K. R. Foster, and A. W. Murray. Sucrose utilization in budding yeast as a model for the origin of undifferentiated multicellularity. *PLoS Biol*, 9(8):e1001122, 2011.

[14] G. Johnston, J. Pringle, and L. Hartwell. Coordination of growth with cell division in the yeast *Saccharomyces cerevisiae*. *Experimental Cell Research*, 105(1):79 – 98, 1977.

[15] J. H. Koschwanez, K. R. Foster, and A. W. Murray. Improved use of a public good selects for the evolution of undifferentiated multicellularity. *eLife*, 2, 2013.

[16] M. J. Kuranda and P. W. Robbins. Chitinase is required for cell separation during growth of *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, 266(29):19758–67, 1991.

[17] M. Li and S. Elledge. SLIC: A method for sequence- and ligation-independent cloning. In J. Peccoud, editor, *Gene Synthesis*, volume 852 of *Methods in Molecular Biology*, pages 51–59. Humana Press, 2012.

[18] D. L. Lindstrom and D. E. Gottschling. The mother enrichment program: A genetic system for facile replicative life span analysis in *Saccharomyces cerevisiae*. *Genetics*, 183(2):413–422, 2009.

[19] M. S. Longtine, A. Mckenzie III, D. J. Demarini, N. G. Shah, A. Wach, A. Brachat, P. Philippsen, and J. R. Pringle. Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast*, 14(10):953–961, 1998.

[20] E. Mazanka, J. Alexander, B. J. Yeh, P. Charoenpong, D. M. Lowery, M. Yaffe, and E. L. Weiss. The NDR/LATS family kinase Cbk1 directly controls transcriptional asymmetry. *PLoS Biol*, 6(8):e203, 08 2008.

[21] M. D. Mendenhall, H. E. Richardson, and S. I. Reed. Dominant negative protein kinase mutations that confer a G1 arrest phenotype. *Proceedings of the National Academy of Sciences*, 85(12):4426–4430, 1988.

[22] M. D. Mendenhall and A. E. Hodge. Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews*, 62(4):1191–1243, 1998.

[23] A. W. Murray and T. Hunt. *The Cell Cycle: An Introduction*. Oxford University Press, 1993.

[24] C. O'Conalláin, M.-T. Doolin, C. Taggart, F. Thornton, and G. Butler. Regulated nuclear localisation of the yeast transcription factor Ace2p controls expression of chitinase (*CTS1*) in *Saccharomyces cerevisiae*. *Molecular and General Genetics MGG*, 262(2):275–282, 1999.

[25] W. C. Ratcliff, R. F. Denison, M. Borrello, and M. Travisano. Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences*, 2012.

[26] M. Sbia, E. J. Parnell, Y. Yu, A. E. Olsen, K. L. Kretschmann, W. P. Voth, and D. J. Stillman. Regulation of the yeast Ace2 transcription factor during the cell cycle. *Journal of Biological Chemistry*, 283(17):11135–11145, 2008.

[27] B. Schirrmeister, A. Antonelli, and H. Bagheri. The origin of multicellularity in cyanobacteria. *BMC Evolutionary Biology*, 11(1):45, 2011.

[28] B. E. Schirrmeister, J. M. de Vos, A. Antonelli, and H. C. Bagheri. Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proceedings of the National Academy of Sciences*, 110(5):1791–1796, 2013.

[29] M. Shirayama, Y. Matsui, and A. Toh-E. The yeast *TEM1* gene, which encodes a GTP-binding protein, is involved in termination of M phase. *Molecular and Cellular Biology*, 14(11):7476–7482, 1994.

[30] R. S. Sikorski and P. Hieter. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics*, 122(1):19–27, 1989.

[31] M. Simon, G. Faye, and B. Seraphin. *KIN28*, a yeast split gene coding for a putative protein kinase homologous to *cdc28*. *EMBO Journal*, 5:2697–2701, 1986.

[32] S. Smukalla, M. Caldara, N. Pochet, A. Beauvais, S. Guadagnini, C. Yan, M. D. Vinces, A. Jansen, M. C. Prevost, J.-P. Latgé, G. R. Fink, K. R. Foster, and K. J. Verstrepen. *FLO1* is a variable green beard gene that drives biofilm-like cooperation in budding yeast. *Cell*, 135:726–737, 2008.

[33] M. Valerio-Santiago and F. Monje-Casas. Tem1 localization to the spindle pole bodies is essential for mitotic exit and impairs spindle checkpoint function. *The Journal of Cell Biology*, 192(4):599–614, 2011.

[34] W. P. Voth, A. E. Olsen, M. Sbia, K. H. Freedman, and D. J. Stillman. Ace2, Cbk1, and Bud4 in budding and cell separation. *Eukaryotic Cell*, 4(6):1018–1028, 2005.

[35] Y. Wang, T. Shirogane, D. Liu, J. Harper, and S. J. Elledge. Exit from Exit: Resetting the cell cycle through Amn1 inhibition of G protein signaling. *Cell*, 112(5):697 – 709, 2003.

[36] G. Yvert, R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. M. Smith, R. Mackelprang, and L. Kruglyak. *Trans*-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genetics*, 35:57–64, 2003.

Chapter 4

Genome reidentification for the CAGI Personal Genome Project challenge

"C'est magnifique, mais ce n'est pas la guerre."

- Pierre François Joesph Bosquet

4.1 Abstract

The Critical Assessment of Genome Interpretation (CAGI) is an international, biannual competition designed to ascertain the state of the art in human genome interpretation. I participated in the CAGI 2013 challenge to identify the donors of genomes anonymously donated to the Personal Genome Project [2]. Participants in this challenge were provided with the whole genome sequences of seventy-seven individuals as well as the anonymized phenotypic profiles (including health history, ancestry, gender, and information on 239 traits) for all donors as well as 214 "decoy" volunteers with no corresponding genome sequence. I used a combination of ethnic admixture analysis, blood typing, Y chromosome and mitochondrial DNA haplotyping, causative allele identification, and supplemental information provided by volunteers to correctly identify the profiles corresponding to 32 of the 77 genomes provided, as determined through a blinded independent assessment. Many of the scripts and methods used in this analysis were shared with participants and colleagues through the Personal Genome Project Wiki[52].

4.2 Introduction

Genomics researchers rely on public databases of samples provided by donors who wish to remain anonymous. Breach of anonymity ("genome reidentification") is a serious risk which could potentially lead to discrimination against the donors or their families. Consent procedures that downplay this risk can mislead potential volunteers with irreversible consequences, but it is difficult for study organizers to accurately guess what could be done – now or in the future – to link anonymized genome sequences back to participants. For example, on the subject of genome reidentification, the consent form signed by sample donors in the 1000 Genomes Project [34] states that "we [the study organizers] believe this could happen only if somebody knew that you had given a sample to be studied for this project," yet this study group has repeatedly been the focus of proof-of-principle demonstrations of

de-anonymization [16, 32], including the recent discovery that dozens of participants can be identified by name [18]. The awareness raised by such demonstrations has informed the development of improved consent procedures [2, 31]. Since these demonstrations were performed by members of the scientific community acting under traditional codes of ethics, participant identities were never released publicly and the database hosting agencies were informed in advance of publication, allowing changes in data distribution that now afford participants protection against these reidentification methods [21].

The Critical Assessment of Genome Interpretation (CAGI) challenges researchers to develop and share state-of-the-art techniques in genome interpretation by applying them to real, unpublished datasets. The objective of the CAGI Personal Genome Project challenge is to match anonymous genome sequences with the phenotypic profiles provided by their donors, a task akin to genome reidentification. Gender, age, geographic location, family relationships, ethnicity, blood type, health history, and supplemental data provided by participants were used to identify 32 out of 77 genome-profile pairs. This targeted approach using freely-available software and minimal sophistication fared comparably to the best alternative method, as determined by an independent third-party assessor.

4.3 Methods

Supplemental data collection

An XML/HTML parser was used to collect the location, blood type, samples provided for sequencing, grandparents' countries of origin, stated ethnicity, stated gender, other genetic datasets, other health records, and the information for other participating blood relatives mentioned in the phenotypic profiles of interest. Some individuals did not appear to have submitted samples for analysis and were therefore excluded from further consideration. Full names were noted when provided in uploaded documents. Data obtained from participant profiles did not always match the information provided in the Challenge Profiles spread-

sheet: ethnicity was frequently changed, and two participants labeled female in the Challenge Profiles spreadsheet (hu00D419 and huC92BC9) self-identified as XY female or male on their online profiles. The Personal Genome Project forum [51] was also scanned for posts from participants claiming that their genomes had been withheld for the CAGI competition.

Ethnic admixture analysis

Called SNPs from GFF files were reformatted for use with DIYDodecad 2.1, a tool developed by Dionekes Pontikos for estimating ethnic admixture [13] which takes 23andme result files as input. The provided "globe13" calculator was used to estimate admixture from 13 populations (Siberian, Amerindian, West African, Paleo-African, Southwest Asian, East Asian, Mediterranean, Australasian, Arctic, West Asian, North European, South Asian, and East African). Admixture predictions were compared to participants' self-identified ethnicities, names, and grandparents' countries of origin where possible.

Analyzing supplemental genetic datasets

Participant-submitted 23andme and Family Tree DNA autosomal SNP profiling data obtained were collected as described above and compared to the SNP calls for anonymous genomes at 97 SNPs chosen for their high secondary allele frequency and compatibility across file types and processing dates (Table 4.1). Participants who provided 23andme or Family Tree DNA exome sequencing data were compared to the anonymous genomes at 25 SNPs.

Y and mitochondrial haplogroup analysis

Family Tree DNA Y chromosome short tandem repeat (Y-STRs) counts were used to determine an individual's Y haplogroup using an automated haplogroup predictor developed by Whit Athey and Doug McDonald [1]. The Y haplogroups of the anonymous genomes were determined using Y chromosome SNPs described in Karafet et al., 2008[25].

Table 4.1: Variants used to match commercial SNP genotyping data to whole genome sequences

Position		Vars		Position		Vars		Position		Vars	
chr1	3507861	G	A	chr4	138134581	A	G	chr10	49739856	T	G
chr1	6756002	C	A	chr4	165661477	A	G	chr10	68224886	A	G
chr1	32157009	G	T	chr5	10536845	A	G	chr10	94893473	A	G
chr1	63628648	T	C	chr5	80942209	A	G	chr10	114154815	C	T
chr1	92334749	C	T	chr5	96161942	T	G	chr10	119335191	G	A
chr1	94537642	C	A	chr5	111948117	T	C	chr11	4703762	G	T
chr1	103088742	G	T	chr5	132758353	T	C	chr11	26752782	C	T
chr1	106128123	T	C	chr5	134543927	C	T	chr11	34998746	G	A
chr1	115619422	C	T	chr5	148587553	T	C	chr11	87610338	A	G
chr1	182115787	A	G	chr5	172719020	T	C	chr11	99241675	G	T
chr1	183604258	C	A	chr5	173899353	A	G	chr11	113230600	C	T
chr1	211892568	T	C	chr6	31273745	T	C	chr12	24016182	T	C
chr2	462799	A	G	chr6	31555130	A	G	chr12	55351215	C	T
chr2	20752612	T	G	chr6	33059996	T	G	chr12	115495279	G	A
chr2	29120733	C	T	chr6	73448086	A	G	chr13	52515354	A	A
chr2	58959112	G	A	chr6	126244682	T	C	chr13	67169478	C	T
chr2	61217542	T	C	chr6	137294656	C	T	chr14	22377212	C	T
chr2	75368797	C	T	chr6	154472327	T	C	chr15	80590198	T	G
chr2	128457941	A	G	chr6	169933335	G	A	chr15	97063357	C	T
chr2	152794981	G	A	chr7	29318397	T	C	chr16	2977262	A	G
chr2	190492014	T	C	chr7	100979310	C	T	chr16	48763325	G	T
chr2	223843894	T	C	chr7	147958844	C	T	chr17	56207731	A	G
chr2	230211434	T	C	chr7	150775306	G	A	chr18	13437993	A	G
chr3	30334755	A	G	chr8	2923969	A	G	chr18	25440258	T	C
chr3	56809628	A	G	chr8	17565061	T	C	chr18	51448585	T	C
chr3	69059023	A	C	chr8	17908116	A	G	chr19	39792460	G	A
chr3	139357093	G	A	chr8	29441961	T	C	chr20	20246306	A	G
chr3	172881577	C	T	chr8	101835049	G	A	chr20	25133966	G	T
chr3	177782520	A	C	chr8	120887041	G	A	chr20	48949175	G	A
chr4	8677043	C	T	chr9	15846112	C	T	chr20	62860980	A	G
chr4	70765416	G	A	chr9	86088871	C	T	chr21	16330894	T	C
chr4	72356146	C	T	chr9	108750147	T	G				
chr4	96076813	A	G	chr9	138151499	GG	CCGT				

Ninety-seven autosomal SNPs with the same reference and variant alleles in 23andme and CGI whole genome sequence calls were identified by trial and error. 23andme data were mapped to whole genome sequences through comparison of genotype at these loci.

James Lick's mthap utility [28] was used to determine the mitochondrial haplogroups of participants who provided Family Tree mitochondrial DNA sequencing results. The mitochondrial haplogroups of anonymous genomes were determined using a list of predictive SNPs maintained by Ian Logan [30]. A script was later produced to produce FASTA-formated mitochondrial DNA sequences for each anonymous genome and shared through the Personal Genome Project participant wiki [52]. (Variant calls in TSV format identify differences relative to the industry-standard revised Cambridge Reference Sequence, while the GFF files provided by the CAGI challenge include differences relative to the UCSC human reference genome, introducing confusion.) Mitochondrial sequences obtained this way were used to determine the mitochondrial haplogroups of the anonymous genome with additional specificity.

CYP2 genotyping

Due to the important role of cytochrome P450 in drug metabolism [17], several participants were able to provide CYP2 subunit sequencing results collected for medical reasons. The anonymous genomes were genotyped at relevant sites using CYP2 allele information summarized on SNPedia [29] and the Human Cytochrome P450 Allele Nomenclature Database [45].

Causative alleles of von Willebrand disease

Probable causative alleles in *VWF* were identified in the International Society on Thrombosis and Hameostasis Scientific and Standardization Committee's von Willebrand Factor variant database [20]. Anonymous genomes were checked for these causative alleles as well as for likely large deletions at the *VWF* locus (i.e. long stretches of homozygosity and more frequent no-call events due to lower coverage).

Table 4.2: Variants used to identify the most common *ABO* alleles

Allele	rs8176719	rs8176720	rs1053878	rs7853989	rs8176741	rs8176742
A1	C	T (ref)	G (ref)	G (ref)	A (ref)	C (ref)
A1v/A2	C	T (ref)	A	G (ref)	A (ref)	C (ref)
B	C	C	G (ref)	C	G	C (ref)
O1	- (ref)	T (ref)	G (ref)	G (ref)	A (ref)	C (ref)
O1v	- (ref)	C	G (ref)	G (ref)	A (ref)	T
O2	C	C	G (ref)	C	A (ref)	C (ref)

The *ABO* sequence in the UCSC human reference genome – against which all variants in the CAGI genomes are called – corresponds to the "O1" allele at six of the loci most predictive for blood type [41, 49].

Blood typing

Individuals were genotyped for previously-described predictive variants of ABO blood type [41, 49], summarized in Table 4.2. All individuals were found to be homozygous for three SNPs near the *RHD* locus (rs25629943, rs25628043, rs25628088), suggesting an error in calling, but Rhesus negative phenotype could still be accurately inferred using the recessive and highly-correlated allele rs590787. A script for determining blood type from variant calls at the *ABO* locus has been made available through the Personal Genome Project participant wiki [52].

Submission preparation

Entries to the CAGI PGP competition were submitted as a 77 x 291 matrix containing the assigned probabilities of assignment for every possible genome-profile combination. We assumed that the probability of a match was zero for profiles not associated with DNA samples or which had been eliminated using SNP genotyping data. We also assumed that the probability of a match was zero for genomes and profiles of opposite sex. Matches were given a probability of one if they were made on the basis of SNP genotyping data and/or haplogroup matching. All other determinations were considered subject to potential error and therefore

affected matches were assigned intermediate probabilites based on the estimated likelihood of participant or genotyping error.

A uniform prior was assigned based on the number of genomes which could potentially match each profile given the information above and the type of DNA sample submitted. (Participants who had provided blood samples were assigned a three-fold higher probability of participating relative to those who had submitted only saliva samples.) This probability was then weighted based on the number of profiles which each genome could potentially match. Standard deviations in probability estimates were determined by simulating one million profile-genome assignments consistent with the data available.

Two profiles – hu432EB5 and hu00D419 – had outsize influence on these probability assignments. hu00D419 describes herself as an XY female on the basis of 23andme test results, but may not have received a clinical diagnosis: her karyotype affects the number of possible genome-profile pairings for both males and females. hu432EB5 is confirmed participant with von Willebrand disease: limiting potential matches to genomes heterozygous for causative alleles could have a large effect on the apparent accuracy of the prediction. Four submissions were prepared to give best probability estimates for each possible hu00D419 karyotype and each liberal vs. conservative estimates of the likelihood that the mutant *VWF* alleles found in the anonymous genomes were causative.

4.4 Results

Each Personal Genome Project participant may update their online phenotypic profile to add information beyond that required by the study organizers. This information included other sequencing results collected for medical, genealogical, or recreational reasons; fitness monitoring logs; blood type; full names; sample collection logs; geographic locations; plaintext descriptions of ailments; links to the profiles of participating blood relatives; and information on ancestry. This data was systematically extracted from participating profiles using

a custom web trawler. Of 291 phenotypic profiles in the CAGI dataset, 108 were rapidly eliminated as potential genome donors due to failure to submit saliva or blood samples for sequencing[1].

The declining cost of recreational genotyping and volunteers' demonstrated interest in genomics are reflected in the high proportion (33%) of participants who have uploaded commercial sequencing data to their profiles. Twenty-seven individuals who uploaded 23andme SNP genotyping results, one individual who submitted Family Tree DNA SNP genotyping results, and one individual who submitted Family Tree DNA exome sequencing results were unambiguously matched to the genomes they provided (Table 4.3); 32 individuals were also eliminated as potential genome donors based on the sequencing results they provided (Table 4.4).

Y and mitochondrial haplogroups

With few exceptions [3], human mitochondrial DNA and non-pseudoautosomal regions of the Y chromosome do not undergo recombination [10, 15] and include regions exhibiting a wide range of substitution rates [19, 46, 55]. Variations in these sequences are therefore useful for tracking paternal and maternal lineages on timescales ranging from single generations to millennia [12, 44]. Similarity between Y chromosomes has traditionally been assessed based on short tandem repeat copy number variation due to the favorable mutation rate and low cost of genotyping these sites by PCR amplicon length, whereas mitochondrial DNA has historically been Sanger sequenced at known variable regions. For approximately one decade prior to the advent of "next generation" sequencing, Y and mitochondrial DNA haplotyping were the only direct-to-consumer DNA sequencing options available for genomics enthusiasts: many Personal Genome Project participants have thus taken these tests and provided their results through their online profiles. While Y and mitochondrial DNA hap-

[1]This issue was reported to CAGI during prediction assessment.

Table 4.3: Profiles matched to genomes using participant-provided autosomal genotyping data

Data provided	Profile identification number	Genome identification number
23andme autosomal SNP genotyping	hu1F73AB	52b
	hu3C06l1	60b
	hu90B053	815
	huCA14D2	737
	huE58004	df7
	huEDF7DA	812
	huDF04CC	ebc
	hu016B28	bd9
	hu19C09F	6a6
	hu4BE6F2	82a
	hu7S2f1D	084
	hu7B594C	648
	huA4F28l	3d4
	huB4D223	940
	huC434ED	581
	huD52556	f86
	hu4B0812	fb6
	hu448C4B	221
	hu627574	213
	hu72C17A	223
	huC3160A	574
	huED0F40	3c1
	hu032C04	ce7
	hu619F51	15f
	huC92BC9	6b7
	hu05FD49	805
23andme exome sequencing	hu2FEC01	368
Family Tree DNA autosomal SNP genotyping	hu599905	693

Files containing supplemental genotyping data were automatically collected from participant profiles using a custom web trawler and compared to anonymous genomes as described in the methods.

Table 4.4: Profiles eliminated from consideration using participant-provided autosomal genotyping data

Data provided	Profile identification number		
23andme autosomal SNP genotyping	huD58ABC	hu1097B2	hu1187FF
	hu1A4F2E	hu1BD549	hu4C3094
	hu868880	huBAA265	huBC03A7
	huD4F7DB	huDF9008	huEDEA65
	huF06AD0	hu57C9FD	hu2BC187
	hu30888B	hu394092	hu394755
	hu41F03B	hu48C4EB	hu5AE862
	hu82436A	hu840B0B	hu96713F
	huA720D3	huB4E01A	huB7EC37
	huB921C5	huE31062	huEAA57B
Family Tree DNA exome sequencing	huD0D79A		

Files containing supplemental genotyping data were automatically collected from participant profiles using a custom web trawler and compared to anonymous genomes as described in the methods.

lotypes are shared by many individuals in a population, they remain useful for eliminating possible matches between profiles and genomes.

Two otherwise-unassigned individuals provided their mitochondrial DNA sequences on their participant profiles; the mitochondrial DNA sequence for third was obtained from her sister's linked participant profile. One of these individuals (hu25E1EE) shared a mitochondrial haplotype with exactly one genome in the CAGI dataset: this link was further supported by the shared gender and ethnic admixture (see below) of the genome and participant profile. The remaining two individuals (huB4E01A and hu52B7E5) did not appear to share a mitochondrial haplotype with any of the CAGI genomes (see Table 4.5).

Y chromosome short tandem repeat copy number information provided by one participant (hu7123C1) was used to infer his Y haplogroup (J2b). Short nucleotide polymorphisms associated with this haplogroup were used to identify a possible match with one of

Table 4.5: Inferred mitochondrial haplogroups of CAGI genomes

Genome ID #	mtDNA haplogroup	Genome ID #	mtDNA haplogroup	Genome ID #	mtDNA haplogroup
b74	A2	fb6	H2a2a1	239	K1a3a2
c8a	A2ag	60b	H3	eb7	K1c2
598	H2a or H17	d7f	H4†	82a	K2b2
6d9	H2a or H46	368	H46b	8a3	N1b1a
15f	H1ac	df7	H4a1a1	4bd	N9a3
f86	H1bf or H1bh	e5b	H4a1a1	3d4	T1a1b
221	H10a1a1	c05	H5	414	T2a1a8
23a	H10b	407	H56b	6a6	T2b2b
fa7	H16b	ebc	H5a1	d76	T2be or T2b3c
af4	H1a1	40c	H6a1a3	6a5	T2b4b
5da	H1a3	c47	H6a1b	926	U4a
ce7	H1a3	732	HV0b, c, or d	39d	U4a3
737	H1b2a	dce	HV0b, c, or d	f9b	U4b1b
b62	H1bk	bd9	HV5	545	U4c1a
581	H1bo	52b	I1a1	3de	U5a1a1
213	H1c	815	J1b1a1a	693	U5a1a1d
181	H1f1	5dd	J1c2h	9bd	U5a1b1e
fb7	H1i1	838	J1c3i	6fb	U5a1c2a
b34	H1j1	c3b	J1c3j	805	U5a2c
bda	H1j1	246	J1c5b	223	U5a2d
396	H1n4	d56	J2b1a	88f	U8a1a1b
648	H1o	fc1	J2b1a6	574	V10
3c1	H1q1	1cf	K	812	V10
940	H27	903	K1a1b1a	073	V2b
9d1	H2a	8b2	K1a1b1b	084	X2b
6b7	H2a1e	912	K1a26		

Mitochondrial DNA variants relative to the revised Cambridge Reference Sequence (rCRS) were identified for each genome provided in the CAGI dataset and used to reconstruct mitochondrial DNA sequences for the corresponding genomes. Note absence of H6 (huB4E01A) and J1c (hu52B7E5) haplogroups. †This haplogroup assignment was used to link genome d7f to phenotypic profile hu5CD2C6 using uploaded data from her sibling (hu25E1EE).

the anonymous genomes, which was corroborated by additional health-related genetic profiling data. As with mitochondrial haplogroups, Y haplogroup frequency information was not available, so estimates of the probability of a false positive haplogroup pairing could not be made.

While microsatellite copy number variation has been successfully used to infer the surnames of individuals from their Y chromosome short tandem repeats [18], this technique unfortunately could not be employed because raw sequencing reads (and repeat count) were not reported for the competition. Furthermore, the single nucleotide polymorphism substitution rate is several orders of magnitude lower than the rate of change in microsatellite copy number, and thus such mutations are expected to correlate poorly with surnames.

Online statements of participation by genome donors

An online forum for Personal Genome Project volunteers has been organized by a participant. Several individuals whose genomes were withheld for the CAGI competition stated their participation on the forum and could be associated with a participant profile (Table 4.6). While three of these individuals had already been linked to an anonymous genome using participant-provided genotyping data, participation of the fourth (hu432EB5) was confirmed using this information, which limited the number of remaining profile-genome combinations substantially.

Moreover, hu432EB5 is the only participant to report von Willebrand disease type 2N, a partial-penetrance monogenic disorder caused by mutations in von Willebrand Factor (vWF) which specifically impede its binding to Factor VIII without decreasing overall vWF levels in the bloodstream [8, 33, 39]. Variants found in *VWF* in the anonymous CAGI genomes were compared to a catalog of missense mutations associated with von Willebrand disease type 2N curated by the research community [20]: unfortunately, no genome was homozygous or compound heterozygous for the most prevalent and predictive variants, nor for likely loss-of-function mutations (Table 4.7). A variant associated with type 1 von Wille-

Table 4.6: CAGI participants identified through the Personal Genome Project forum

Forum member name	Profile identification number	Means of identification
sharper	hu432EB5	Only von Willebrand disease type 2N patient
PGP-84	hu7123C1	Name and blog post
James Turner	hu016B28	Derry, NH location
huE58004	huE58004	Forum member name

The Personal Genome Project online forum was searched for posts containing the word "CAGI." Forum members who reported that their genome had been sequestered for CAGI were then associated with their participant profile by the method indicated.

brand disease (Y1584C) and considered dominant was found in two genomes, but type 1 and type 2N von Willebrand disease are easily distinguished clinically, so genomes with this variant were not considered likely matches for profile hu432EB5. Two genomes carrying the R854Q allele [27] were considered to be the most likely matches for hu432EB5.

Participant-uploaded clinical genotyping data

The human genome encodes approximately sixty cytochrome P450 (*CYP*) enzymes involved in the synthesis of cholesterol, fatty acids, and steroid hormones as well as the breakdown or bioactivation of exogenous compounds, including many medications [24]. Genetic variation in these enzymes can affect their efficacy and complicate the prescription of drugs they metabolize. For example, selecting an appropriate dosage of the anticoagulant warfarin, which targets certain *CYP* enzymes, is critical because under- and overdose can both cause lethal symptoms (stroke and myocardial infarction or internal bleeding, respectively), but due to genetic variation, the appropriate dose varies more than twenty-fold between patients. Three SNPs in the warfarin-metabolizing genes *CYP2C9* and *CYP4F2* explain 14% of this variation

Table 4.7: Variants associated with von Willebrand disease found in the anonymous CAGI genomes

vWF variant	von Willebrand disease type	Inheritance pattern	Reference	Genomes affected
R854Q	Type 2N	Recessive	[27]	213 3d4 9d1† eb7†
S1731T	Type 2N	Dominant	[42]	82a
Y1584C	Type 1	Dominant	[40]	073† 9bd†

Variants in vWF associated with von Willebrand's disease found in the anonymous CAGI genomes. Inheritance patterns reported in the literature remain speculative (see accompanying references).
† Genomes not assigned to profiles by other means.

[47], so *CYP* genotyping is routinely performed prior to warfarin prescription. Several Personal Genome Project participants were therefore able to provide genotyping data at these loci.

The *CYP2D6*, *CYP2C19*, and *CYP2C9* loci in the anonymous CAGI genomes were compared to *CYP* genotyping results uploaded by participants huC7B886 and hu5FA322 (Table 4.8). While six genomes were consistent with the genotype of *huC7B886*, none of these shared his gender and O- blood type, so this profile was ruled out as a potential match for any genome (Table 4.9). Participant hu5FA322's genotyping results, which included *CYP* loci as well as several genes with known involvement in porphyria (*SOD2*, *NAT1*, and *NAT2*), were consistent with one anonymous CAGI genome (Table 4.9), a match further supported by blood type, ethnicity, and sex. As with Y and mitochondrial haplogroups, insufficient data was available on allele frequencies to determine the probability of a false positive match.

Table 4.8: Participant-reported genotyping results in the CAGI PGP challenge

Locus	Allele	Identifying variant(s)	Reference
CYP2D6	CYP2D6*1	Reference allele	[45]
	CYP2D6*2	rs16947	[45]
		rs1135840	[45]
	CYP2D6*4	rs1065852	[45]
		rs28371703	[45]
		rs28371704	[45]
CYP2C9	CYP2C9*1	Reference allele	[45]
CYP2C19	CYP2C19*1	Reference allele	[45]
	CYP2C19*2	rs4244285	[45]
NAT1	NAT1*4	Reference allele	[22]
NAT2	NAT2*5B	rs1801280	[22]
		rs1203	[22]
SOD2	WT	Reference	[43]
	V16A	rs4880	[43]
HEXA	WT	Reference allele	[36]
	B1 variant	R178H/R178C	[48]
	739C→T	R247W	[50]
	745C→T	R249W	[7]
	805G→A	G269S	[38]
	IVS9	+1G→A	[6]
	E11	1273insTATC	[35]
	IVS12	+1G→C	[35]

Common names for alleles described in participant-provided genotyping data were matched to sequence variants that could be identified within the anonymous CAGI genomes.

Table 4.9: Clinical genotyping data provided by CAGI PGP participants

Profile	Locus	Genotype	Matching Genomes
huC7B886	CYP2D6	CYP2D6*1/*2	239, 23a, 912, c8a, d56, d76
	CYP2C9	CYP2C9*1/*1	
	CYP2C19	CYP2C19*2/*2	
hu5FA322	CYP2D6	CYP2D6*1/*4	b34†
	CYP2C9	CYP2C9*1/*1	
	CYP2C19	CYP2C19*1/*2	
	SOD2	V16A/+	
	NAT1	NAT1*4/*4	
	NAT2	NAT2*5B/*5B	

Clinical genotyping data provided by participants for individual loci. Allele nomenclature is described in Table 4.8.
†: This genome-profile pairing could not be ruled out by other means.

Participant hu1843FC reports that she is a carrier for Tay Sachs disease. None of the female anonymous genomes in the CAGI PGP dataset were carriers of the most common causative SNP/indel alleles at the *HEXA* locus (Table 4.8). hu1843FC was therefore ruled out as a potential genome donor.

Ethnic admixture analysis

Only 8 of the participant profiles claimed non-Caucasian ancestry. On the grounds that these individuals might be purposefully enriched among the sequenced genomes to increase the diversity of populations represented, we searched for genomes with non-Caucasian ethnic admixture. We used a Markov Chain Monte Carlo-based approach to identify the most likely boundaries between DNA segments arising from populations with distinct SNP frequencies. The primary advantage of this technique over estimation from representative SNPs is that the relative size of these segments reflects the timing of admixture, allowing us to discern individuals with grandparents of distinct heritage from those whose admixture occurred less recently. We identified three genomes with a significant proportion of African and Native

American admixture: we considered these genomes most likely to match profiles which reported African American or Latino ancestry. A post hoc analysis revealed that the relatively small segment size and composition in at least two of three participants was reflective of their Puerto Rican and Columbian ancestry (Figure 4.1). All profiles indicating Asian ancestry were also eliminated from consideration because no genome with significant Asian admixture could be identified.

Mendelian traits

Personal Genome Project participants were asked to report whether they possessed a number of traits with Mendelian inheritance. Unfortunately, preliminary investigations using known genome-profile pairings revealed that participant reporting and/or variant calling make many of these traits unreliable for genome-profile pairing.

For example, homozygosity for a haplotype consisting of the A allele at rs182549 and the T allele at rs4988235 is an excellent predictor of lactose intolerance in persons of European descent: 77% of affected individuals are reported to have this genotype [4, 14]. Twenty-five of seventy-seven genomes in the CAGI dataset were homozygous for this allele, but amongst those that could be linked back to profiles by definitive means, not a single individual reported lactose intolerance. Indeed, a retrospective analysis performed after the true pairings were announced showed that only one of the individuals homozygous for this haplotype had reported lactose intolerance, and one individual who did not possess this haplotype had also reported intolerance. This inconsistency may reflect either participant reporting error or the influence of alternative causative variants, but in either event, lactose intolerance was deemed too risky for use in supporting or refuting genome-profile pairings from this dataset.

Participant reporting error likely reflects ambiguity in diagnosis; we therefore focused our attention on the most salient and clinically-examined Mendelian traits. Color blindness, a condition for which children are routinely tested using pseudoisochromatic plates, was

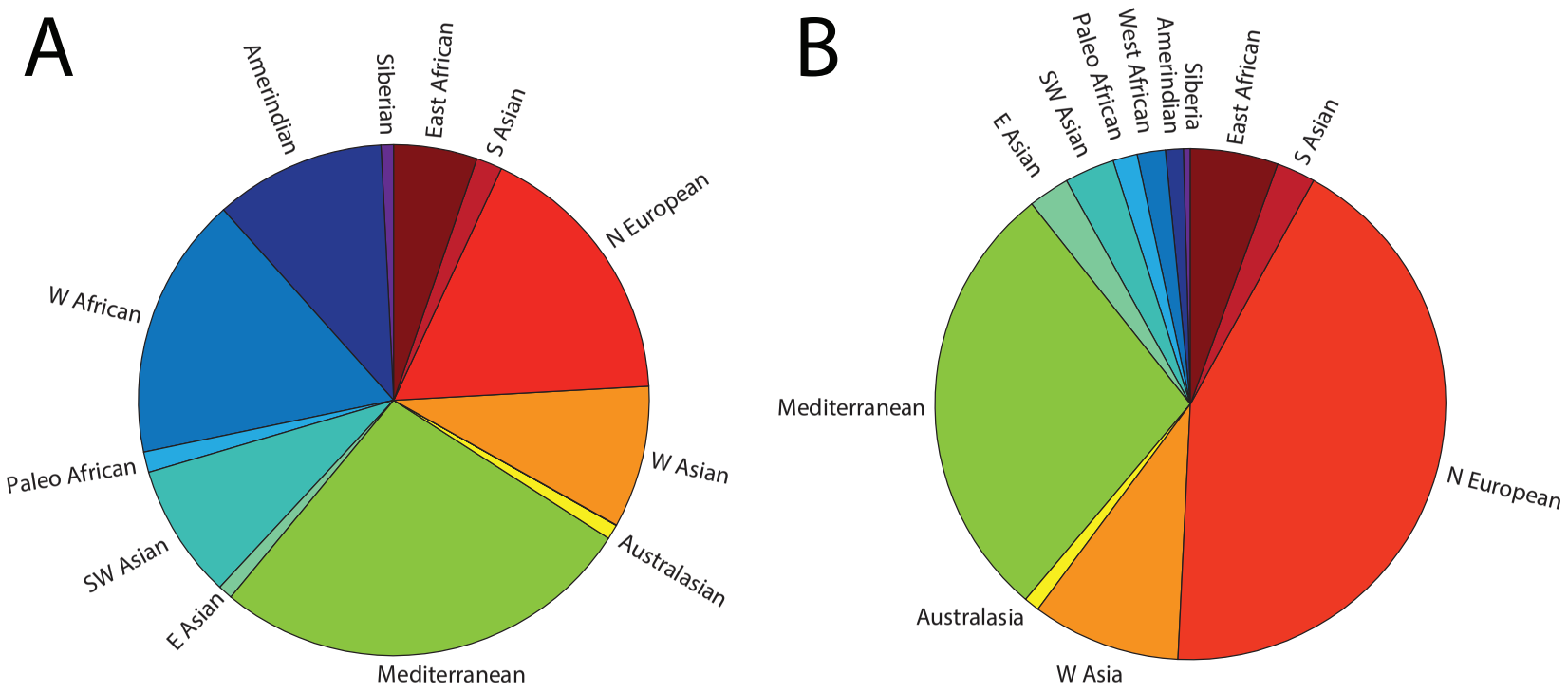

Figure 4.1: Determination of genome ethnicity by admixture analysis

(A) Results of admixture analysis for a participant of Puerto Rican heritage (huEA4EE5). Relatively high proportions of West African and Amerindian heritage are present.

(B) Representative result of admixture analysis for a participant of predominantly European heritage. Asian, African, and Amerindian segment calls are likely to be spurious.

therefore a prime candidate. Several forms of color blindness exist, of which dichromacy (caused by inactivation of one out of three cone opsins) is by far the most common, affecting 4.5% of Americans [11]. Dichromacy can result from unequal recombination between the adjacent red and green opsin loci (*OPN1LW* and *OPN1MW*, respectively) that deletes the intervening region: while the investments required to identify such gene rearrangements have prohibited large-scale analyses, smaller studies of ≈25 individuals suggest that this is by far the most common cause of dichromacy [37]. Because these opsin loci were created by a recent duplication event, sequencing reads from either locus have the potential to be mapped to both loci, which can obscure deletion of the intervening region when present. (Coverage information would reveal the deletion, but unfortunately raw reads were not provided for the anonymous genomes.) While eleven participant profiles reported colorblindness (including those of four actual genome donors), we were unable to detect plausible causative mutations and therefore could not use this trait.

By contrast, participant-reported blood types were employed with relative success to eliminate or support genome-profile pairings. The *ABO* locus encodes a glycosyltransferase which modifies the H antigen, an oligosaccharide later attached to glycoproteins. Non-functional alleles of the ABO protein ("O") are recessive to each of two codominant alleles ("A" and "B") that modify the H antigen in distinct ways. A fourth, hybrid "*cis*-AB" allele, the apparent result of recombination between A and B alleles, is also found at frequencies < 0.03% in some populations [54]. Variation at other sites can be epistatic to the *ABO* genotype. For example, the *FUT1* and *FUT2* loci encode homologous proteins responsible for H antigen production: their inactivation produces the Bombay phenotype, a set of serological properties mimicking "O" type blood [26]. Blood type prediction is further complicated by the fact that novel inactivating mutations in *ABO* and the Rhesus factor locus *RHD* can produce novel O and Rh-negative alleles. Despite these potential complications, we attempted to infer blood types for the CAGI genomes using previously-described correlations with six SNPs [41, 49][see Table 4.2]. Sixty-nine participants (including 36 true genome donors) re-

ported their blood type: a post-hoc analysis revealed only one case where our prediction differed from their stated phenotype. Using our blood type predictions, we were able to eliminate 1405 potential genome-profile pairings (45% of the possibilities remaining after other methods were employed.)

Hypoandrogenization and sex reversal

Participant hu00D419 stated on her participant survey that her 23andme results revealed her to be an XY female. Although it is unclear whether this diagnosis was confirmed by a medical professional, all genomes containing a Y chromosome were checked for mutations in genes with known association to Androgen Insensitivity Syndrome and sex reversal (Table 4.10). Genome 073 is heterozygous for the G146A mutation in NR5A1 associated with severe micropenis [53] and genome 40c is homozygous for the P392S mutation in AR associated with male infertility [23]. A DAX1 frameshift mutation associated with adrenal hypoplasa congenita [9] was also identified, but none of the participants who reported known infertility are a likely match for the affected genome, and no male participants reported other symptoms of this syndrome, which include delayed puberty, hypogonadism, and hypoglycemia. While haploinsufficiency in SOX9 is associated with sex reversal and skeletal abnormalities [5], there is no evidence that the SOX9 variant found in the CAGI dataset (PAP354Del) disrupts SOX9 function. None of the other mutations found are associated with sex reversal. hu00D419 could not be ruled out as a potential genome donor, and was considered alternately as a potential XX or XY match for separate competition entries.

4.5 Discussion

The true genome-profile pairings were revealed after the submission deadline, permitting a retrospective analysis of submission accuracy. Thirty-two matches had been correctly identified, including five that did not use SNP genotyping data. Only one true match had been

Table 4.10: Variants in androgenization-related genes found in male anonymous CAGI genomes

Gene	Function	Mutation(s)	Genomes	Zygosity
NR5A1	Steroidogenic nuclear receptor	G146A	073	heterozygous
AR	Androgen receptor	P392S	40c	homozygous
LHX9	Transcription factor, gonadal development	P355Shift	4bd 693 82a	heterozygous heterozygous heterozygous
		PP355AT	3de 926	heterozygous heterozygous
DAX1 (NR0B1)	Dosage-sensitive nuclear receptor	G190Shift	f9b	hemizygous
		G44S	bd9	hemizygous
WNT4	Ligand stimulating ovarian development	D239G	fb6	heterozygous
SOX9	Transcription factor promoting male development	PAP354Del	693	heterozygous

Genes involved in androgenization with variants in male anonymous CAGI genomes.

assigned a probability of zero: the participant in question was later revealed (by PGP data provider Madeleine Ball) to be transgendered. Predictions based on von Willebrand disease-associated variants, ethnic admixture, and blood type were accurate with the exception of one individual who reported blood type O- while our prediction algorithm predicted blood type AB+.

A common method of evaluating submission accuracy is the receiver operating characteristic (ROC) curve, which considers both sensitivity (the fraction of true positives) and precision (the fraction of false positives) for the prediction. The curve is constructed parametrically by plotting sensitivity vs. precision for matches assigned probabilities greater than or equal to θ, for $\theta \in (0,1)$. Random guessing is expected to produce equal sensitivity and precision, and would therefore produce a straight line ROC curve. The area underneath the curve (AUC) reflects the accuracy of the prediction: possible values fall between 0 and 1,

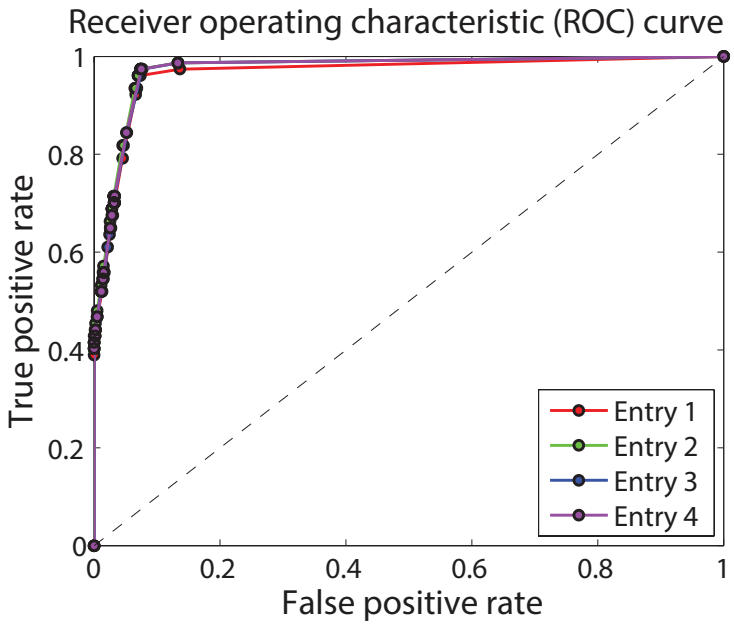

Figure 4.2: Receiver operating characteristic curve for submitted CAGI entries

Points on the receiver operating characteristic (ROC) curve represent the fraction of true pairings identified at a given threshold of confidence, plotted against the fraction of all possible false pairings admitted at the same confidence threshold. The expected ROC curve for a random guess is indicated as a dashed line. Area under the curve (AUC): 0.965. Mann-Whitney U: 58,914 of 1,719,410 ($p = 8 \times 10^{-122}$).

with values greater than 0.5 being better than chance. The ROC curve for our data has an AUC of 0.965 (Figure 4.2), exceeding that of all other entries. The Mann-Whitney U test (also known as the Wilcoxon rank-sum test) can be used to determine the likelihood of the null hypothesis that true and false matches do not differ significantly in probability. For our best entry, this null hypothesis was assigned a probability of $p = 8 \times 10^{-122}$, providing support for the accuracy of our predictions.

Submissions to the CAGI PGP challenge were externally assessed by Sean Mooney, director of the bioinformatics core and assistant professor at the Buck Institute: his comparison of the entries is provided in Table 4.11. Due to the ease with which SNP genotyping results could be used to identify genome-profile pairings, submission accuracy was considered for both the full dataset and the subset of participant profiles/genomes for which

no SNP genotyping results were reported. Our entries performed slightly better than the leading competitor (the Karchin lab) when "decoy" profiles (those without associated DNA samples) were included in the analysis, but this lead evaporated when the decoy profiles were removed from consideration, confirming that discovery of these decoys contributed significantly to the accuracy of our entries. The Karchin lab employed an orthogonal approach (construction of a Bayesian network based on SNPs identified through GWAS and the primary literature) to identify genome-profile pairings. Perhaps unsurprisingly, the overlap in unambiguously predicted genomes was low, with only two matches in common between our sets of predictions.

No explicit goal was posed in the CAGI Personal Genome Project challenge announcement. At least two teams viewed the competition as an attempt to identify algorithms that could accurately predict health and trait information from genome sequences, with broad implications for patient identification and preventative treatment or palliative care. For that purpose, the methods we employed would be of little value. However, we interpreted the CAGI PGP challenge as an exploration of the potential for genome reidentification, i.e., the re-association of publicly-shared genome sequences with their "anonymous" donors. Malicious use of genome sequences for discrimination against a research participant or their family is only possible once the participant's genome sequence has been associated with their full name: it is therefore critical for participants to understand the true risk of reidentification in order to give informed consent. We have contributed an assessment of the risk of genome reidentification that could be used to inform the consent process. Our techniques can be implemented on a private computer by an operator with little or no expertise[2]; they are therefore accessible to a large population of potential rabble-rousers. In this light, our entries provided complementary and important insights into the process of genome interpretation.

[2]For example, during the course of this project, we were able to find the full names of more than one dozen PGP participants just by unzipping their compressed data files.

Table 4.11: Comparison of entries submitted for the CAGI PGP challenge

Submission	All Genomes		Without Genotyping		Without Decoys and Genotyping	
	Correct	Mean Rank	Correct	Mean Rank	Correct	Mean Rank
S1 (Anonymous1)	25	62.7	0	93.0	0	59.4
S2 (Silvio Tosatto)	0	114.1	0	104.1	0	66.6
S3 (Rachel Karchin)	32	26.9	5	40.8	6	25.4
S4 (Mary Wahl)	31	29.0	4	44.2	4	39.8
S5 (John Moult)	27	40.5	0	60.3	1	35.8
S6 (Anonymous1)	25	196.8	0	291.0	0	183.0
S7 (Silvio Tosatto)	0	111.2	0	104.4	0	67.0
S8 (Mary Wahl)	32	24.6	5	37.4	5	35.2
S9 (John Moult)	27	40.5	0	60.7	2	37.1
S10 (Silvio Tosatto)	0	123.7	0	124.4	0	79.3
S11 (Mary Wahl)	32	26.3	5	40.0	5	37.9
S12 (John Moult)	27	50.7	0	75.6	2	45.1
S13 (Silvio Tosatto)	1	104.4	1	107.5	1	68.5
S14 (Mary Wahl)	32	26.0	5	39.5	5	37.3
S15 (Silvio Tosatto)	0	110.0	0	108.7	0	69.2
S16 (Silvio Tosatto)	0	129.0	0	133.5	0	84.5

Submissions to the CAGI Personal Genome Project challenge were externally assessed by Sean Mooney. Best entries for our group (blue) and the Karchin lab (yellow) are highlighted.

4.6 Bibliography

[1] W. Athey and D. McDonald. Hapest5 haplogroup predictor. `http://dodecad.blogspot.com/2011/09/do-it-yourself-dodecad-v-21.html`.

[2] M. P. Ball, J. V. Thakuria, A. W. Zaranek, T. Clegg, A. M. Rosenbaum, X. Wu, M. Angrist, J. Bhak, J. Bobe, M. J. Callow, C. Cano, M. F. Chou, W. K. Chung, S. M. Douglas, P. W. Estep, A. Gore, P. Hulick, A. Labarga, J.-H. Lee, J. E. Lunshof, B. C. Kim, J.-I. Kim, Z. Li, M. F. Murray, G. B. Nilsen, B. A. Peters, A. M. Raman, H. Y. Rienhoff, K. Robasky, M. T. Wheeler, W. Vandewege, D. B. Vorhaus, J. L. Yang, L. Yang, J. Aach, E. A. Ashley, R. Drmanac, S.-J. Kim, J. B. Li, L. Peshkin, C. E. Seidman, J.-S. Seo, K. Zhang, H. L. Rehm, and G. M. Church. A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences*, 2012.

[3] H.-J. Bandelt, Q.-P. Kong, W. Parson, and A. Salas. More evidence for non-maternal inheritance of mitochondrial DNA? *Journal of Medical Genetics*, 42(12):957–960, 2005.

[4] T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6):1111 – 1120, 2004.

[5] W. Bi, W. Huang, D. J. Whitworth, J. M. Deng, Z. Zhang, R. R. Behringer, and B. de Crombrugghe. Haploinsufficiency of Sox9 results in defective cartilage primordia and premature skeletal mineralization. *Proceedings of the National Academy of Sciences*, 98(12):6698–6703, 2001.

[6] D. H. Brown, B. L. Triggs-Raine, M. J. McGinniss, and M. M. Kaback. A novel mutation at the invariant acceptor splice site of intron 9 in the *HEXA* gene [IVS9-1 G→T] detected by a PCR-based diagnostic test. *Human Mutation*, 5(2):173–174, 1995.

[7] Z. Cao, M. R. Natowicz, M. M. Kaback, J. S. Lim-Steele, E. M. Prence, D. Brown, T. Chabot, and B. L. Triggs-Raine. A second mutation associated with apparent β-hexosaminidase A pseudodeficiency: identification and frequency estimation. *American Journal of Human Genetics*, 53(6):1198–1205, 1993.

[8] A. Casonato, V. Daidone, G. Barbon, E. Pontara, I. Di Pasquale, L. Gallinaro, L. Marullo, and G. Bertorelle. A common ancestor more than 10,000 years old for patients with R854Q-related type 2N von Willebrand's disease in Italy. *Haematologica*, 98(1):147–152, 2013.

[9] J.-H. Choi, J.-Y. Park, G.-H. Kim, H. Y. Jin, B. H. Lee, J. H. Kim, C. H. Shin, S. W. Yang, and H.-W. Yoo. Functional effects of *DAX-1* mutations identified in patients with X-linked adrenal hypoplasia congenita. *Metabolism*, 60(11):1545 – 1550, 2011.

[10] H. J. Cooke, W. R. A. Brown, and G. A. Rappold. Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature*, 317(6039):687 – 692, 1985.

[11] S. Deeb. The molecular basis of variation in human color vision. *Clinical Genetics*, 67(5):369–377, 2005.

[12] M. Denaro, H. Blanc, M. J. Johnson, K. H. Chen, E. Wilmsen, L. L. Cavalli-Sforza, and D. C. Wallace. Ethnic variation in Hpa I endonuclease cleavage patterns of human mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 78(9):pp. 5768–5772, 1981.

[13] Dienekes. Do-It-Yourself Dodecad version 2.1. `http://dodecad.blogspot.com/2011/09/do-it-yourself-dodecad-v-21.html`.

[14] N. S. Enattah, T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, and I. Jarvela. Identification of a variant associated with adult-type hypolactasia. *Nature Genetics*, 30(2):233 – 237, 2002.

[15] R. E. Giles, H. Blanc, H. M. Cann, and D. C. Wallace. Maternal inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences*, 77(11):6715–6719, 1980.

[16] J. Gitschier. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap Project. *The American Journal of Human Genetics*, 84(2):251 – 258, 2009.

[17] F. P. Guengerich. Cytochrome P450 and chemical toxicology. *Chemical Research in Toxicology*, 21(1):70–83, 2008.

[18] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich. Identifying personal genomes by surname inference. *Science*, 339(6117):321–324, 2013.

[19] J. Haldane. The rate of spontaneous mutation of a human gene. *Journal of Genetics*, 31(3):317–326, 1935.

[20] D. Hampshire. International Society on Thrombosis and Haemostasis Scientific and Standardization Committee's von Willebrand Factor online database (vWFdb). `http://vwf.group.shef.ac.uk/`.

[21] E. C. Hayden. Privacy loophole found in genetic databases. *Nature News*, 2013.

[22] D. Hein, S. Boukouvala, D. Grant, R. Michin, and E. Sim. Changes in consensus arylamine N-acetyltransferase gene nomenclature. *Pharmacogenetics and Genomics*, 18(4):367–368, 2008.

[23] O. Hiort, P.-M. Holterhus, T. Horter, W. Schulze, B. Kremke, M. Bals-Pratsch, G. H. G. Sinnecker, and K. Kruse. Significance of mutations in the androgen receptor gene in males with idiopathic infertility. *The Journal of Clinical Endocrinology & Metabolism*, 85(8):2810–2815, 2000.

[24] M. Ingelman-Sundberg, S. C. Sim, A. Gomez, and C. Rodriguez-Antona. Influence of cytochrome P450 polymorphisms on drug therapies: Pharmacogenetic, pharmacoepigenetic and clinical aspects. *Pharmacology & Therapeutics*, 116(3):496 – 526, 2007.

[25] T. M. Karafet, F. L. Mendez, M. B. Meilerman, P. A. Underhill, S. L. Zegura, and M. F. Hammer. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research*, 2008.

[26] R. J. Kelly, L. K. Ernst, R. D. Larsen, J. G. Bryant, J. S. Robinson, and J. B. Lowe. Molecular basis for H blood group deficiency in Bombay (Oh) and para-Bombay individuals. *Proceedings of the National Academy of Sciences*, 91(13):5843–5847, 1994.

[27] P. A. Kroner, K. D. Friedman, S. A. Fahs, J. P. Scott, and R. R. Montgomery. Abnormal binding of factor VIII is linked with the substitution of glutamine for arginine 91 in von Willebrand factor in a variant form of von Willebrand disease. *Journal of Biological Chemistry*, 266(29):19146–9, 1991.

[28] J. Lick. MTHap. `http://dna.jameslick.com/mthap/`.

[29] J. Lick. SNPedia entry on *CYP2D6*. `http://snpedia.com/index.php/CYP2D6`.

[30] I. Logan. Haplogroup finder. `http://www.ianlogan.co.uk/haplogroup/instructs.htm`.

[31] J. E. Lunshof and M. P. Ball. Our genomes today: time to be clear. *Genome Medicine*, 5(6):1–2, 2013.

[32] J. E. Lunshof, R. Chadwick, D. B. Vorhaus, and G. M. Church. From genetic privacy to open consent. *Nature Reviews Genetics*, 9:406–411, 2008.

[33] C. Mazurier, J. Dieval, S. Jorieux, J. Delobel, and M. Goudemand. A new von Willebrand factor (vWF) defect in a patient with factor VIII (FVIII) deficiency but with normal levels and multimeric patterns of both plasma and platelet vWF. characterization of abnormal vWF/FVIII interaction. *Blood*, 75(1):20–26, 1990.

[34] G. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

[35] R. Myerowitz and F. C. Costigan. The major defect in Ashkenazi Jews with Tay-Sachs disease is an insertion in the gene for the alpha-chain of β-hexosaminidase. *Journal of Biological Chemistry*, 263(35):18587–9, 1988.

[36] R. Myerowitz. Tay-Sachs disease-causing mutations and neutral polymorphisms in the *HEXA* gene. *Human Mutation*, 9(3):195–208, 1997.

[37] J. Nathans, T. Piantanida, R. Eddy, T. Shows, and D. Hogness. Molecular genetics of inherited variation in human color vision. *Science*, 232(4747):203–210, 1986.

[38] R. Navon and R. L. Proia. The mutations in Ashkenazi Jews with adult GM2 gangliosidosis, the adult form of Tay-Sachs disease. *Science*, 243(4897):1471–1474, 1989.

[39] M. Nishino, J. Girma, C. Rothschild, E. Fressinaud, and D. Meyer. New variant of von Willebrand disease with defective binding to factor VIII. *Blood*, 74(5):1591–1599, 1989.

[40] L. A. O'Brien, P. D. James, M. Othman, E. Berber, C. Cameron, C. R. P. Notley, C. A. Hegadorn, J. J. Sutherland, C. Hough, G. E. Rivard, D. O'Shaunessey, D. Lillicrap, and the Association of Hemophilia Clinic Directors of Canada. Founder von Willebrand factor haplotype associated with type 1 von Willebrand disease. *Blood*, 102(2):549–557, 2003.

[41] S. K. Patnaik, W. Helmberg, and O. O. Blumenfeld. BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems. *Nucleic Acids Research*, 40(D1):D1023–D1029, 2012.

[42] A.-S. Ribba, I. Loisel, J.-M. Lavergne, I. Juhan-Vague, B. Obert, G. Cherel, D. Meyer, and J.-P. Girma. Ser968Thr mutation within the A3 domain of von Willebrand Factor (vWF) in two related patients leads to a defective binding of vWF to collagen. *Thrombosis and Haemostasis*, 86:848–854, 2001.

[43] J. S. Rosenblum, N. B. Gilula, and R. A. Lerner. On signal sequence polymorphisms and diseases of distribution. *Proceedings of the National Academy of Sciences*, 93(9):4471–4473, 1996.

[44] F. R. Santos, N. O. Bianchi, and S. D. Pena. Worldwide distribution of human Y-chromosome haplotypes. *Genome Research*, 6(7):601–611, 1996.

[45] S. C. Sim. The human cytochrome P450 (CYP) allele nomenclature database. `http://www.cypalleles.ki.se/index.htm`.

[46] P. Soares, L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Röhl, A. Salas, S. Oppenheimer, V. Macauley, and M. B. Richards. Correcting for purifying selection: an improved human mitochondrial molecular clock. *American Journal of Human Genetics*, 84(6):740 – 759, 2009.

[47] F. Takeuchi, R. McGinnis, S. Bourgeois, C. Barnes, N. Eriksson, N. Soranzo, P. Whittaker, V. Ranganath, V. Kumanduri, W. McLaren, L. Holm, J. Lindh, A. Rane, M. Wadelius, and P. Deloukas. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet*, 5(3):e1000433, 03 2009.

[48] A. Tanaka, K. Ohno, K. Sandhoff, I. Maire, E. W. Kolodny, A. Brown, and K. Suzuki. GM2-gangliosidosis B1 variant: analysis of β-hexosaminidase α gene abnormalities in seven patients. *American Journal of Human Genetics*, 46:329 – 339, 1990.

[49] J. L. Taylor-Cousar, M. A. Zariwala, L. H. Burch, R. G. Pace, M. L. Drumm, H. Calloway, H. Fan, B. W. Weston, F. A. Wright, M. R. Knowles, and for the Gene Modifier Study Group. Histo-blood group gene polymorphisms as potential genetic modifiers of infection and cystic fibrosis lung disease severity. *PLoS ONE*, 4(1):e4270, 01 2009.

[50] B. L. Triggs-Raine, E. H. Mules, M. M. Kaback, J. S. T. Lim-Steele, C. E. Dowling, B. R. Akerman, M. R. Natowicz, E. E. Grebner, R. Navon, J. P. Welch, C. R. Greenberg, G. H. Thomas, and R. A. Gravel. A pseudodeficiency allele common in non-Jewish Tay-Sachs carriers: implications for carrier screening. *American Journal of Human Genetics*, 51(4):793–801, 1992.

[51] J. Turner. Personal Genome Project participants forum. `http://forum.personal-genome.org`.

[52] J. Turner. Personal Genome Project wiki. `http://http://wiki.personal-genome.org/`.

[53] Y. Wada, M. Okada, T. Hasegawa, and T. Ogata. Association of severe micropenis with Gly146Ala polymorphism in the gene for steroidogenic factor-1. *Endocrine Journal*, 52(4):445–448, 2005.

[54] M. H. Yazer, M. L. Olsson, and M. M. Palcic. The *cis*-AB blood group phenotype: Fundamental lessons in glycobiology. *Transfusion Medicine Reviews*, 20(3):207 – 217, 2006.

[55] L. A. Zhivotovsky, P. A. Underhill, C. Cinnioğlu, M. Kayser, B. Morar, T. Kivisild, F. C. Rosaria Scozzari, G. Destro-Bisol, G. Spedini, G. K. Chambers, R. J. Herrera, K. K. Yong, D. Gresham, I. Tournev, M. W. Feldman, and L. Kalaydjieva. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Current Biology*, 19(17):1453–1457, 2009.

Chapter 5

Conclusions and Future Directions

"The cell is evolution's most brilliant invention and development is its triumphant elaboration."

- Lewis Wolpert

5.1 Summary of major results

In chapter one, we review the current hypotheses regarding the evolution of multicellularity and cellular differentiation in major clades throughout the tree of life. Many distinct selective pressures have favored the evolution of these traits, and the biological mechanisms which implement them are extraordinarily diverse. Despite this variation, common themes emerge: most multicellular organisms display differentiation, and in all cases where an informed conjecture is possible, it appears that multicellularity evolved first.

We hypothesized that unicellular, differentiating species would be short-lived due to the potential for invasion – and therefore, ultimately, reversion – by non-differentiating mutants. To test this theory, we engineered unicellular, differentiating budding yeast. We describe in chapter two the development of a synthetic cell type specification system that permits the independent control of conversion rate, relative growth rate, and cell type specific function. This strain behaves as expected according to mutation-selection balance theory under a variety of circumstances, including growth in well-mixed liquid media and on plates. In chapter three, our model system was extended to include strains in which differentiating cells permanently sacrifice their reproductive potential, thus drawing a closer analogy to the differentiation of somatic issue in naturally-occuring organisms. We used our strains to show that invasion of unicellular, differentiating strains by non-differentiating mutants does occur, and that this effect is slowed or reversed in multicellular strains. Our bioengineering-based approach limited relevant differences between compared strains to a few, known loci, a distinct advantage over previous comparative methods involving extant species which diverged tens or hundreds of millions of years ago. We argue that our results are generalizable and support the conclusion that evolution through a unicellular, differentiating intermediate is unlikely due to their short persistence time.

Chapter four recounts this author's attempts at genome reidentification under the guise of an international competition, the Critical Assessment for Genome Interpretation Personal Genome Project challenge. The concerned reader will be relieved to learn that

bioinformatics, social engineering, and internet trawling all remain predominantly ineffectual at determining the donors of anonymous genome sequences published online. A handful of research participants were, however, identified using techniques which could easily be implemented by non-experts using a personal computer and freeware programs, including Y chromosome/mitochondrial DNA haplotype matching, medical records, ethnic admixture, blood type, and relatives' uploaded genetic data. We echo the recommendations of prominent research ethicists that volunteers should be properly informed that deanonymization is a likely outcome now and in the future.

5.2 Specific Future Directions

5.2.1 Regulation of Differentiation by Local Cell Type Ratio

In our multicellular, differentiating yeast strains, the positioning of our differentiated cells is highly non-random. In a culture where approximately 50% of cells are differentiated, it is not rare to find smaller clumps that consist of only one cell type. This is the expected result of random fragmentation of larger clumps, and there is nothing that can fully prevent it; however, if the conversion rate were modulated by the fraction of nearby differentiated cells, the number of clumps which contain only unconverted cells could be decreased dramatically. This would likely be beneficial, since undifferentiated cells will grow and divide faster when there are differentiated cells nearby.

When our strains are grown in sucrose media, clumps containing differentiated (Suc^+) cells will experience higher monosaccharide concentrations. We would like the conversion rate in such clumps to be lower than in clumps that have no differentiated cells (and therefore low monosaccharide concentrations). Naïvely, one presumes that it is possible to achieve this by repressing Cre recombinase transcription when glucose is abundant, e.g. by placing Cre recombinase under regulation by the glucose-activated repressor Mig1 [19]. In a preliminary attempt, we added Mig1 consensus sites [19] to the *SCW11* promoter (the promoter currently used for Cre expression) at two positions which did not appear to overlap with nucleosome binding sites [15] (Figure 5.1A). This unfortunately did not lead to change in Cre activity with glucose level as evaluated from reporter construct activity (Figure 5.1B); in fact, the conversion rate appeared to be slightly higher in glucose media. Perhaps another positioning of the Mig1 binding sites, or placement of Cre under a promoter that is already glucose-repressed, would be sufficient to achieve the desired effect.

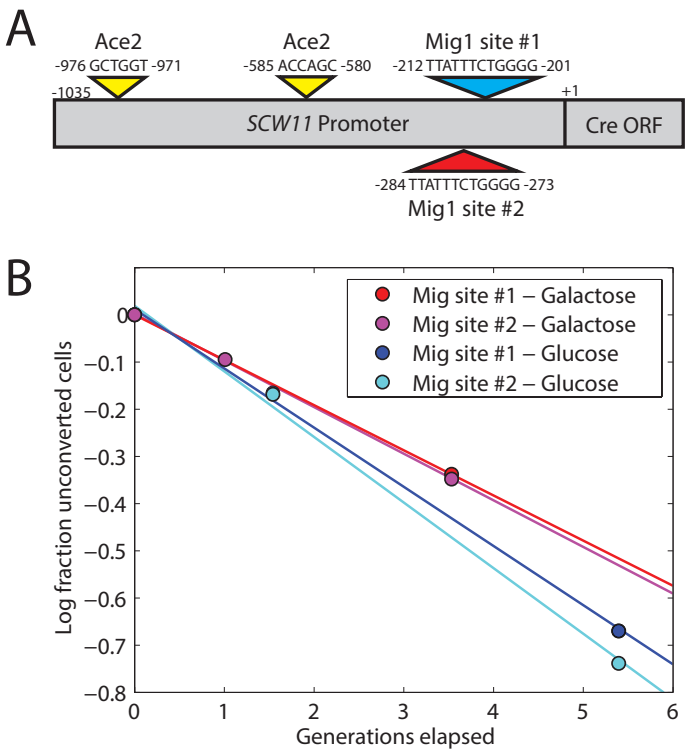

Figure 5.1: Mig1 binding sites do not confer glucose sensitivity to *SCW11* promoter

(A) Consensus Mig1 binding sites were introduced at nucleosome-free positions ≈200 bp from the transcriptional start, similar to Mig1 site positioning in *GAL1*. (B) Strains with Cre under modified *SCW11* promoters containing Mig site #1 (yMEW173) or #2 (yMEW174) were grown in glucose or galactose media containing 1 μM β-estradiol to determine conversion rate.

5.2.2 Two-Way Switch

In our current system, we make use of Cre recombinase to excise the genetic material between two loxP sites in the same orientation on the chromosome. When two loxP sites on the same chromosome are oriented in opposite directions, Cre recombinase will instead invert the intervening sequence. We can exploit this trait to make a strain that can convert freely between two gene expression states as shown in Figure 5.2. This type of conversion mimics stochastic phenotype switching, which has been proposed to be a preferable alternative to active sensing when changes in environment are rare [18]. Whereas permanent loss of one cell type is possible when conversions are irreversible, stable ratios are always reached for two-way switching, and the dynamics with which this steady state is approached may be quite different. Bryan Weinstein, a member of the Nelson lab, proposes to study the relationship between theory and experimental observations for a biological two-way switch similar to our strains.

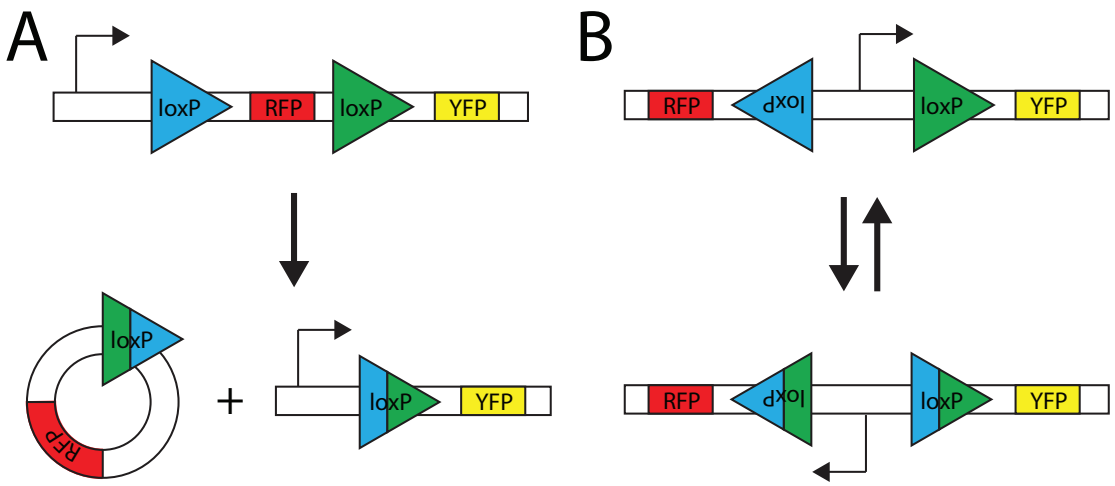

Figure 5.2: Changing relative loxP orientation permits reversible switching

(A) Recombination between loxP sites in the same orientation excises the intervening DNA. Since the reverse reaction would require loxP sites on two separate DNA fragments to come into contact again, the reaction is effectively irreversible.
(B) Recombination between loxP sites in opposite orientations inverts the intervening DNA; the reverse reaction is predicted to be equally likely. If the flipped sequence contains a promoter, then a different set of genes can be transcribed in each orientation.

Preliminary attempts to produce a two-way switch have been complicated by apparent interference between the promoter whose orientation flips and the nearby promoter for the construct's selective marker. When the two promoters face one another, expression from the flipping promoter decreases (Figure 5.3A); the situation resolves after conversion, when both promoters have the same orientation. Reversing the orientation of the selective marker cassette will likely alleviate the problem and is currently underway, but could not be completed in time for publication of this thesis. mCherry fluorescence was at least sufficient to observe by microscopy and to confirm that expression state is stable with time during colony growth in the absence of β-estradiol (Figure 5.3B).

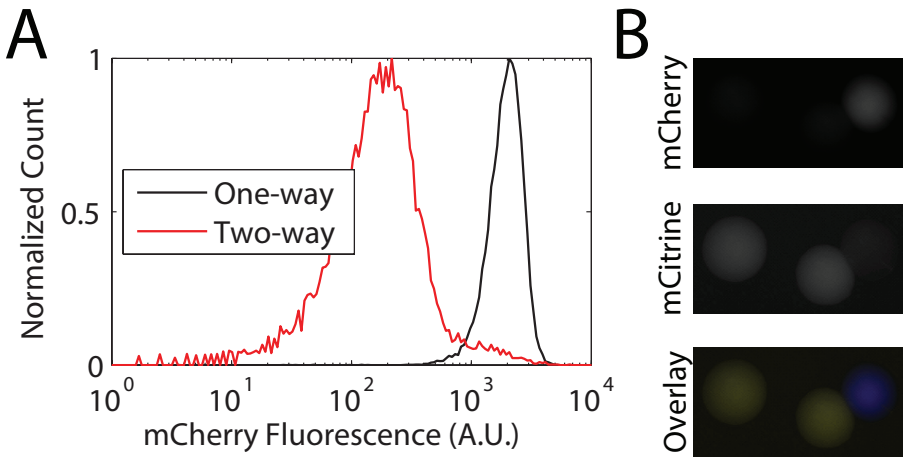

Figure 5.3: mCherry expression is lower and noisier in the two-way switch at present.

(A) mCherry fluorescence in unconverted cells containing the one-way switch construct (yMEW192) is higher and has a lower coefficient of variation than mCherry fluorescence from unconverted cells with the present version of the two-way switch (yMEW219).
(B) Colonies formed from a two-way switch culture containing both cell types. mCherry fluorescence, though low, is easily detected by microscopy. Fluorescence state appears stable as no mixed colonies are seen.

5.2.3 Beneficial Division of Labor

> "It is the great multiplication of the productions of all the different arts, in consequence of the division of labor, which occasions, in a well-governed society, that universal opulence which extends itself to the lowest ranks of the people."
> - Adam Smith, *The Wealth of Nations* (1776)

On the occasion of my qualifying exam some four years syne, I claimed that differentiation is beneficial when it separates functions that are mutually antagonistic – such as flagellar motility and mitotic division in the volvocine algae, or nitrogen fixation and photosynthesis in the cyanobacteria – so that each cell type can perform its task with greater efficiency. This is essentially the same argument proposed by Adam Smith to explain the economic advantages of the assembly line and other innovations of the industrial revolution. Division of labor, the term he coined for this strategy [25], is frequently employed in the evolution of multicellularity research community to describe the apparent cooperation between cell types through partitioning and specialization of tasks [10, 16, 23, 30].

In our differentiating strains, the two relevant types of "labor" are culture growth and nutrient production through invertase secretion. Although both of these processes require energy, there is no explicit antagonism between them. Moreover, the costs of invertase production are likely to be relatively low – less than 1% of a cell's energy expenditure [12], even under the relatively strong *ENO2* promoter – so we would predict that there would be little potential benefit for producing differentiated cells that perform only this function. As expected, our differentiating strains grow slower than their wildtype counterparts where all cells secrete invertase and divide indefinitely.

While this fact does not interfere with our interpretation of the results presented in chapter three, it does raise an important question: is there any form of division of labor which could be beneficial in yeast? Budding yeast are not motile and can't perform photosynthesis, so divisions of labor similar to those found in the volvocine algae, cyanobacteria, slime molds, and (hypothesized) urmetazoans are not possible. It is not obvious to this author what processes budding yeast possess that are in direct conflict with one another and can also be partitioned between two cell types. Barring inspiration from budding yeast's current lifestyle, it may still be possible to find a beneficial division of labor by growing yeast in a very different environment or by importing novel functions from other species.

In chapter one, we saw that the most basic type of differentiation in many forms of sessile algae and basal fungi is the production of a holdfast or rhizoid to secure the multicellular group in place while the remainder of cells have the potential to contribute reproductively. It would be an interesting exercise to show that a similar life cycle could be straightforwardly engineered. Budding yeast retain the ability to form rhizoid-like structures through invasive pseudohyphal growth [11]: we can imagine exploiting this feature to engineer a multicellular strain in which differentiated cells are primed for invasive growth by overexpression of negative regulators of Hog1 [24].

Suppose we grow this strain in a laboratory environment consisting of an agar plate covered in liquid media (Figure 5.4). Multicellular clumps settling on the surface of the

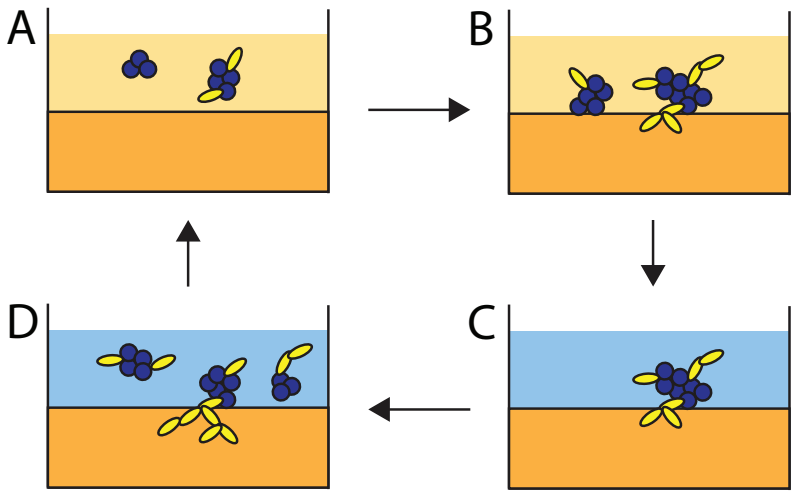

Figure 5.4: Culturing conditions which favor a multicellular strain with invasive pseudohyphal differentiated cells.

(A) Multicellular clumps containing undifferentiated (blue) and/or differentiated (yellow) cells are added to a new environment containing an agar plate (dark orange) under liquid media (light orange). (B) Clumps settle onto the agar under the force of gravity and begin to divide. If orientation permits, differentiated cells begin to invade the agar and root the multicellular clump in place. (C) The liquid media is changed to remove any clumps which are not rooted to the agar. (D) Continued growth eventually causes the clump to fragment, releasing new propagules into the liquid medium that can be passaged to fresh plates.

agar could become rooted in place through invasive growth by their differentiated cells: this phenotype could be selected by repeatedly changing the liquid media above the plate. Parts of the multicellular aggregate above the surface of the agar would occasionally break off (just as they do in our multicellular strains currently), releasing "propagules" into the liquid media that could be transferred to new agar plate environments. Differentiated cells will grow more slowly, and those which invade the agar are unlikely to eventually be released into the liquid medium: differentiated cells are therefore effectively somatic. Enrichment of this engineered strain on passaging could be used to demonstrate a benefit for differentiation of a rhizoid under these growth conditions.

Alternatively, beneficial division of labor could be implemented using costly functions that are non-native to *S. cerevisiae*. One especially costly process that can easily be ported to

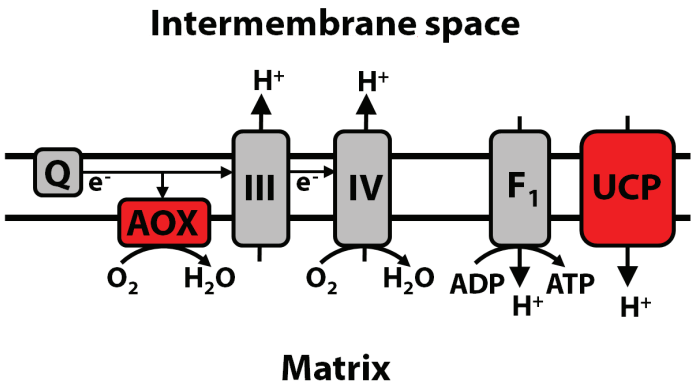

Figure 5.5: Alternative oxidase and uncoupling protein generating heat by interfering with ATP generation

Alternative oxidase (AOX) and uncoupling protein (UCP) inhibit ATP production by shunting electrons to oxygen reduction prematurely or diffusing an established proton gradient, respectively.

yeast is heat generation through decoupling of the mitochondrial electron transport chain (Figure 5.5). Alternative oxidase, an enzyme conserved across all domains of life, is induced by cold in some plants [14, 21]. Alternative oxidase shunts electrons away from the electron transport chain prematurely and thus inhibits the establishment of the mitochondrial proton gradient required for ATP production, which generates heat at significant cost to metabolic efficiency [27][1]. Uncoupling protein, the primary thermogenic enzyme in brown fat, diffuses the mitochondrial proton gradient and thus also interferes with ATP generation in order to release heat [26].

Highly thermogenic homologs of each of these proteins have successfully been cloned into *S. cerevisiae* and shown to inhibit their growth on non-fermentable carbon sources [9, 20]. It may in principle be possible for heat generation under cold conditions to serve as a somatic cell function in our differentiating strains, particularly when yeast are grown in colonies, where cells are very dense and heat loss to the air is poor. Alternative oxidase and uncoupling protein expression would be so much more costly to a differentiated cell's fitness than invertase expression that it would not be unreasonable to expect division of labor to be beneficial in this case.

[1]It also limits production of reactive oxygen species under this stressful condition: this was likely its ancestral function, and it is unclear which function accounts for the major benefit of alternative oxidase expression under cold conditions (reviewed in [27]).

5.2.4 Baupläne

> "The most general characteristic of organismic pattern is a surface-interior difference, and in some simple organisms such difference apparently constitutes the only persistent organismic pattern. [...] Surface-interior pattern is so obviously a reaction to environmental factors that it need not be explored further here."
>
> - Charles Manning Child, *Patterns and Problems of Development* (1941)

Charles Manning Child studied development in organisms such as hydroids and planarians which can reproduce through fragmentation followed by regeneration, a method shared by our *ace2* multicellular yeast strains. While his contemporaries studying sexually-reproducing species saw development as initiated by cytoplasmic determinants in the zygote and proceeding deterministically despite variation in growth conditions [28], Child championed the belief that environmental inhomogeneities were important for cell fate determination. (Child believed the regenerative responses he observed were due to influences of the environment, to which cells were suddenly exposed following injury [5].)

Child's hypothesis regarding the role of environmental inhomogeneity in development was long disfavored. The majority of data Child provided to support his views were visualizations of nutrient availability and respiration rate along axes of polarity [5], but as his contemporaries noted, this was more likely the consequence, rather than the cause, of developmental fate differences [3]. Furthermore, to borrow Blackstone's [3] phrasing, Child's investigations "took place in a mechanistic vacuum": contemporary techniques did not permit deconvolution of cause and effect. By the time molecular approaches permitted deeper understanding, the deterministic view of development had become engrained by force of example. The view that environment could influence development was also at odds with the popular notion, due to Waddington [28], that developmental programs had evolved to succeed reliably *in spite of* environmental variation and other "noise." Most damning, eventual analysis of mechanism did not support an environmental role for axiate patterning in many of the specific examples Child cited [4, 8].

Mechanistically-supported examples now abound, however, of environmental cues establishing axiate patterning under natural circumstances. A frequent theme is response to hypoxic conditions through the generation of reactive oxygen species (ROS) which can initiate signaling to alter gene expression [8]. Oral-aboral polarity in the sea urchin larva [1], polyp vs. stolon differentiation in hydroids [3], and rhizoid orientation in brown algae [6, 7] all rely on ROS concentration gradients established by inhomogeneity in mitochondrial distribution, differential exposure to oxygen, and/or metabolic rate dependence on nutrient or availability.

The reliance of axiate patterning on ROS signalling supports an increasingly popular notion that cellular differentiation can evolve through co-option of existing pathways for environmental sensation and response, a form of genetic assimilation [13, 29]. Consider, for example, the hypothesized planula-like urmetazoan, a multicellular organism multiple cells thick. Cells on the interior would experience lower oxygen and nutrient availability than those at the surface, which would trigger hypoxia and starvation-related signaling cascades. Differentiation could occur through evolution of a novel regulatory connections from those signaling pathways to other, previously-unrelated genes: for example, hypoxic conditions could trigger loss of flagella. Very few mutations would thus be required to create not only distinct cell types, but also a body plan or "bauplan": an ordered positioning of cell types along an axis of polarity (in this case, the radial axis).

I propose that this principle could be adapted to our multicellular differentiation system to produce a spherical bauplan. We will assume here that clumps are large enough[2] to experience amino acid starvation or other nutrient limitation conditions at their interior[3]. Then, interior cells could be made to differentiate at a higher rate by placing Cre recombinase under regulation from the general amino acid control transcriptional activator Gcn4 [2].

[2]We have used a simple *ace2* decrease-of-function mutation to implement multicellularity, but combinations of mutations are known which give much larger clump sizes [17], if necessary.

[3]As yeast are facultatively anaerobic, using hypoxic conditions as a differentiation cue may not be straightforward.

This placement is opposite to the location of somatic cells in other vaguely spherical organisms including *Volvox*, sponge larvae, and cnidarian planulae, but is not without precedent: Ratcliff et al. [22] have experimentally evolved strains where cells at the interior of clumps die at an accelerated rate, which induces fragmentation that effectively regulates clump size. Theoretically, the localization of differentiated cells could be reversed through introduction of a repressor of Cre recombinase under control of Gcn4.

Preliminary experimental investigations and modeling would be useful to determine what fitness benefit, if any, could be gained from introduction of a body plan in our multicellular strains. Even in the absence of a selective advantage, it remains an interesting question to what extent such a simple modification would introduce polarity given that clumps are constantly fragmenting anew.

5.3 Closing Remarks

We have engineered terminal cellular differentiation in budding yeast and used this system to explore evolutionary pressures underlying the coevolution of multicellularity and differentiation, thus using a synthetic approach to bring experimental validation to a field which for too long has relied on comparative analyses and historical speculation. Our multicellular, differentiating strains provide a starting point for experimental evolution or engineering of more complex traits including beneficial divisions of labor and body plans.

5.4 Bibliography

[1] C. Agca, W. H. Klein, and J. M. Venuti. Reduced O_2 and elevated ROS in sea urchin embryos leads to defects in ectoderm differentiation. *Developmental Dynamics*, 238(7):1777–1787, 2009.

[2] K. Arndt and G. R. Fink. GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5' TGACTC 3' sequences. *Proceedings of the National Academy of Sciences*, 83(22):8516–8520, 1986.

[3] N. W. Blackstone. Redox control and the evolution of multicellularity. *BioEssays*, 22(10):947–953, 2000.

[4] N. W. Blackstone. Charles Manning Child (1869fb1954): the past, present, and future of metabolic signaling. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 306B(1):1–7, 2006.

[5] C. M. Child. *Patterns and Problems of Development*. University of Chicago Press, 1941.

[6] S. M. Coelho, C. Brownlee, and J. H. Bothwell. Feedback control of reactive oxygen and Ca^{2+} signalling during brown algal embryogenesis. *Plant Signaling & Behavior 2008*, 3:570 – 572, 2008.

[7] S. M. Coelho, C. Brownlee, and J. H. Bothwell. A tip-high, ca^{2+}-interdependent, reactive oxygen species gradient is associated with polarized growth in *Fucus serratus* zygotes. *Planta*, 227(5):1037–1046, 2008.

[8] J. A. Coffman and J. M. Denegre. Mitochondria, redox signaling and axis specification in metazoan embryos. *Developmental Biology*, 308(2):266 – 280, 2007.

[9] P. Douette, P. Gerkens, R. Navet, P. Leprince, E. De Pauw, and F. Sluse. Uncoupling protein 1 affects the yeast mitoproteome and oxygen free radical production. *Free Radical Biology & Medicine*, 40:303–315, 2006.

[10] S. Gavrilets. Rapid transition towards the division of labor via evolution of developmental plasticity. *PLoS Comput Biol*, 6(6):e1000805, 06 2010.

[11] C. J. Gimeno, P. O. Ljungdahl, C. A. Styles, and G. R. Fink. Unipolar cell divisions in the yeast *S. cerevisiae* lead to filamentous growth: Regulation by starvation and RAS. *Cell*, 68(6):1077 – 1090, 1992.

[12] J. Gore, H. Youk, and A. van Ouenaarden. Snowdrift game dynamics and facultative cheating in yeast. *Nature*, 459(7244):253 – 256, 2009.

[13] B. K. Hall. *Evolutionary Developmental Biology.* Chapman & Hall, 1992.

[14] Y. Ito, D. Saisho, M. Nakazono, N. Tsutsumi, and A. Hirai. Transcript levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature. *Gene*, 203(2):121 – 129, 1997.

[15] C. Jiang and B. F. Pugh. A compiled and systematic reference map of nucleosome positions across the *Saccharomyces cerevisiae* genome. *Genome Biology*, 10(10):R109, 2009.

[16] D. L. Kirk. A twelve-step program for evolving multicellularity and a division of labor. *BioEssays*, 27(3):299–310, 2005.

[17] J. H. Koschwanez, K. R. Foster, and A. W. Murray. Improved use of a public good selects for the evolution of undifferentiated multicellularity. *eLife*, 2, 2013.

[18] E. Kussell and S. Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309(5743):2075–2078, 2005.

[19] L. L. Lutfiyya, V. R. Iyer, J. DeRisi, M. J. DeVit, P. O. Brown, and M. Johnston. Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae*. *Genetics*, 150(4):1377–1391, 1998.

[20] G. Mathy, R. Navet, P. Gerkens, P. Leprince, E. De Pauw, C. Sluse-Goffart, F. Sluse, and P. Douette. *Saccharomyces cerevisiae* mitoproteome plasticity in response to recombinant alternative ubiquinol oxidase. *Journal of Proteome Research*, 5:339–348, 2006.

[21] B. J. D. Meeuse. Thermogenic respiration in aroids. *Annual Review of Plant Physiology*, 26(1):117–126, 1975.

[22] W. C. Ratcliff, R. F. Denison, M. Borrello, and M. Travisano. Experimental evolution of multicellularity. *Proceedings of the National Academy of Sciences*, 2012.

[23] V. Rossetti, M. Filippini, M. Svercel, A. D. Barbour, and H. C. Bagheri. Emergent multicellular life cycles in filamentous bacteria owing to density-dependent population dynamics. *Journal of The Royal Society Interface*, 8(65):1772–1784, 2011.

[24] C. A. Shively, M. J. Eckwahl, C. J. Dobry, D. Mellacheruvu, A. Nesvizhskii, and A. Kumar. Genetic networks inducing invasive growth in *Saccharomyces cerevisiae* identified through systematic genome-wide overexpression. *Genetics*, 193(4):1297–1310, 2013.

[25] A. Smith. *The Wealth of Nations.* W. Strahan and T. Cadell, 1776.

[26] K. L. Townsend and Y.-H. Tseng. Brown fat fuel utilization and thermogenesis. *Trends in Endocrinology & Metabolism*, 25(4):168 – 177, 2014.

[27] G. C. Vanlerberghe and L. McIntosh. Alternative oxidase: From gene to function. *Annual Review of Plant Physiology and Plant Molecular Biology*, 48(1):703–734, 1997. PMID: 15012279.

[28] C. H. Waddington. Canalization of development and the inheritance of acquired characters. *Nature*, 150:563–565, 1942.

[29] C. H. Waddington. *The Strategy of the Genes: A Discussion of some Aspects of Theoretical Biology.* George Allen & Unwin, 1957.

[30] M. Willensdorfer. On the evolution of differentiated multicellularity. *Evolution*, 63(2):306 – 326, 2009.

Appendix A

Justthefacs: a freeware applet for flow cytometry analysis

"Just the facts, ma'am."

- Detective Sergeant Joe Friday

The prohibitive cost, large size, and long processing times of the flow cytometry analysis package FlowJo (TreeStar, Inc.) inspired the production of `justthefacs`, a Matlab applet designed for efficient batch analysis of flow cytometry data files. Routine procedures like side scatter vs. forward scatter exclusion (to remove cell aggregates from consideration, reducing variability in fluorescence measurements), thresholding (identifying cell populations with and without a fluorescent marker, e.g. for fitness assays) and fluorescence histogram generation can be quickly applied to many data files; results are easily exported to Excel or retained in Matlab for further analysis. Installaton requires placing the following three programs in the Matlab home directory:

- `justthefacs.m`, the GUI interface program
- `FCSFile.m`, the custom class definition
- `fca_readfcs.m`, an open source package produced by Laszlo Balkay, available through Matlab Central and included here to ensure reproducibility

```
1  function justthefacs
2  ArrayOfFCSFiles = {};
3     hColors = [ 0 0 0; 1 0 0; 1 0.5 0; 0 0.7 0; 0 1 1; 0 0 ...
          1; 1 0 1; 0.7 0.7 0.7;];
4     HistogramBins = logspace(0,5,200);
5     XTicks = [1 10 100 1000 10000 100000];
6     XLimits = [1 100000];
7
8     f = ...
          figure('Visible','off','MenuBar','none','Resize','off', ...
          ...
9     'Name','Just the FACS, ...
          man!','Position',[0,0,700,300],'Color',[1 1 1]);
10
11    hFileLabel = uicontrol('Style','text','String','Files of ...
          Interest', ...
12         'Position',[0 285 140 15]);
13    hFileList = uicontrol(f, ...
          'Style','listbox','String',{},'Value',1,'Callback', ...
14         {@ChangeFileSelection_Callback},'Position',[0 215 ...
              140 70]);
```

```
15      hAdd = ...
           uicontrol('Style','pushbutton','String','Add','Position', ...
           ...
16          [0,195,70,20],'Callback',{@AddButton_Callback});
17      hRemove = ...
           uicontrol('Style','pushbutton','String','Remove', ...
18          'Position',[70,195,70,20],'Callback',{@RemoveButton_Callback});
19      hAllFilesSettings = ...
           uicontrol('Style','pushbutton','String', ...
20          'Same ...
              bounds/settings','Position',[0,175,140,20],'Callback', ...
              ...
21          {@AllFilesSettingsButton_Callback});
22      hAllFilesHist = uicontrol('Style','pushbutton','String', ...
23          'Histogram ...
              All','Position',[0,155,140,20],'Callback', ...
24          {@AllFilesHistButton_Callback});
25      hAllFilesThreshold = ...
           uicontrol('Style','pushbutton','String', ...
26          'Threshold ...
              All','Position',[0,135,140,20],'Callback', ...
27          {@AllFilesThresholdButton_Callback});
28      hAllFilesMean = uicontrol('Style','pushbutton','String', ...
29          'Mean/Stdev ...
              All','Position',[0,115,140,20],'Callback', ...
30          {@AllFilesMeanButton_Callback});
31      hUpdateBounds = uicontrol('Style','pushbutton','String', ...
32          'Update Bounds','Position',[0,95,140,20],'Callback', ...
33          {@UpdateBoundsButton_Callback});
34      hFilterLabel = ...
           uicontrol('Style','text','String','Filter:','Position',[0 ...
           65 50 25]);
35      hFilterPopup = ...
           uicontrol('Style','popupmenu','String',{'No filters ...
           yet'},'Position',[50 70 90 ...
           20],'Callback',{@ChangeFilterSelection_Callback});
36      hThresholdLabel = ...
           uicontrol('Style','text','String','Threshold:','Position',[0 ...
           50 70 15]);
37      hThresholdEntry = ...
           uicontrol('Style','edit','String','0','Position',[70 ...
           50 70 15]);
38      hUpdateThreshold = ...
           uicontrol('Style','pushbutton','String','Update ...
           Threshold','Position',[0,30,140,20],'Callback', ...
39          {@UpdateThresholdButton_Callback});
```

```
40      hTotalLabel = uicontrol('Style','text','String','Cells ...
           Total:','Position',[0 15 70 15]);
41      hAboveLabel = uicontrol('Style','text','String','Cells ...
           Above:','Position',[0 0 70 15]);
42      hTotalEntry = ...
           uicontrol('Style','edit','String','0','Position',[70 ...
           15 70 15]);
43      hAboveEntry = ...
           uicontrol('Style','edit','String','0','Position',[70 ...
           0 70 15]);
44      hAxes = axes('Units','Pixels','Position',[190 70 190 190]);
45      xlabel(hAxes,'FSC-A');
46      ylabel(hAxes,'SSC-A');
47      hAxesFilter = axes('Units','Pixels','Position',[450 70 ...
           190 190]);
48      xlabel(hAxesFilter,'Filter of interest');
49      ylabel(hAxesFilter,'Count');
50      align([hFileList,hFileLabel],'Center','None');
51      set([f,hAboveLabel,hAllFilesThreshold,hAllFilesSettings, ...
52          hUpdateBounds,hTotalLabel,hTotalEntry,hAllFilesHist, ...
53          hAboveEntry,hFileLabel,hFilterLabel,hAxes,hAxesFilter,hAdd, ...
               ...
54          hRemove,hFileList,hFilterPopup,hThresholdLabel, ...
55          hThresholdEntry,hUpdateThreshold,hAllFilesMean], ...
56          'Units','normalized');
57      movegui(f,'center')
58      set(f,'Visible','on');
59      function UpdateThresholdCounts
60           [Total, Above] = ...
                ArrayOfFCSFiles{get(hFileList,'Value')}.threshold( ...
                ...
61              get(hFilterPopup,'Value'));
62           set(hTotalEntry,'String',num2str(Total));
63           set(hAboveEntry,'String',num2str(Above));
64      end
65      function PlotHistogram()
66          EventValues = ArrayOfFCSFiles{get(hFileList,'Value') ...
67             }.getfiltervalues(get(hFilterPopup,'Value'));
68          axes(hAxesFilter);
69          BinnedValues = histc(EventValues, HistogramBins);
70          BinnedValues = BinnedValues .* 1 / max(BinnedValues);
71          plot(HistogramBins,BinnedValues,'Color',[0 0 0]);
72          set(gca,'XScale','log');
73          xlim(XLimits);
74          a = ...
               ArrayOfFCSFiles{get(hFileList,'Value')}.FilterNames{ ...
               ...
```

```
75              get(hFilterPopup,'Value')};
76          title(strcat(a,' histogram'));
77          xlabel(a);
78          set(hAxesFilter,'XTick',XTicks)
79          ylabel('Normalized count');
80          line([ArrayOfFCSFiles{get(hFileList,'Value') ...
81              }.FilterThresholds(get(hFilterPopup,'Value')) ...
82              ArrayOfFCSFiles{get(hFileList,'Value') ...
83              }.FilterThresholds(get(hFilterPopup,'Value'))],[0 ...
                    1], ...
84              'Color',[0 0 1]);
85      end
86      function AllFilesSettingsButton_Callback(source,eventdata)
87          CurrentFilterName = ...
                ArrayOfFCSFiles{get(hFileList,'Value') ...
88              }.FilterNames{get(hFilterPopup,'Value')};
89          CurrentThreshold = ...
                ArrayOfFCSFiles{get(hFileList,'Value') ...
90              }.FilterThresholds(get(hFilterPopup,'Value'));
91          CurrentBoundsX =  ...
                ArrayOfFCSFiles{get(hFileList,'Value')}.BoundsX;
92          CurrentBoundsY =  ...
                ArrayOfFCSFiles{get(hFileList,'Value')}.BoundsY;
93          for i=1:length(ArrayOfFCSFiles)
94              ArrayOfFCSFiles{i}.BoundsX = CurrentBoundsX;
95              ArrayOfFCSFiles{i}.BoundsY = CurrentBoundsY;
96              FilterNumber = ...
                    find(ismember(ArrayOfFCSFiles{i}.FilterNames, ...
                    ...
97                      CurrentFilterName));
98              if isempty(FilterNumber)
99                  fprintf('The file %s does not have a %s ...
                        filter; ...
100                     could not set filter threshold.\n', ...
101                     ArrayOfFCSFiles{i}.FileName,CurrentFilterName);
102             else
103                 ArrayOfFCSFiles{i}.FilterThresholds(FilterNumber) ...
                        = CurrentThreshold;
104             end
105         end
106     end
107     function AllFilesHistButton_Callback(source,eventdata)
108         g = figure('Position',[200,200,450,300]);
109         LegendLabels = {};
110         CurrentFilterName = ...
                ArrayOfFCSFiles{get(hFileList,'Value')}.FilterNames{ ...
                ...
```

```
111             get(hFilterPopup,'Value')};
112         for i=1:length(ArrayOfFCSFiles)
113             FilterNumber = ...
                   find(ismember(ArrayOfFCSFiles{i}.FilterNames, ...
                   ...
114                 CurrentFilterName));
115             if isempty(FilterNumber)
116                 fprintf('The file %s does not have a %s ...
                       filter; ...
117                 could not plot on ...
                       histogram.\n',ArrayOfFCSFiles{i}.FileName, ...
                       ...
118                 CurrentFilterName);
119             else
120                 EventValues = ...
                       ArrayOfFCSFiles{i}.getfiltervalues(FilterNumber);
121                 BinnedValues = histc(EventValues, ...
                       HistogramBins);
122                 BinnedValues = BinnedValues .* 1 / ...
                       max(BinnedValues);
123                 plot(HistogramBins,BinnedValues,'Color', ...
124                     hColors(mod(i-1,8)+1,:)); hold on;
125                 LegendLabels{i} = ArrayOfFCSFiles{i}.FileName;
126             end
127         end
128         xlabel(strcat(CurrentFilterName,' (A.U.)'));
129         set(gca,'XTick',XTicks)
130         ylabel('Count (scaled)');
131         set(gca,'XScale','log');
132         xlim(XLimits);
133         legend(LegendLabels);
134         figure(f)
135     end
136     function AllFilesThresholdButton_Callback(source,eventdata)
137         TotalAndAbove = [];
138         Labels = {};
139         CurrentFilterName = ...
               ArrayOfFCSFiles{get(hFileList,'Value')}.FilterNames{ ...
               ...
140             get(hFilterPopup,'Value')};
141         for i=1:length(ArrayOfFCSFiles)
142             FilterNumber = ...
                   find(ismember(ArrayOfFCSFiles{i}.FilterNames, ...
143                 CurrentFilterName));
144             if isempty(FilterNumber)
145                 fprintf('The file %s does not have a %s ...
                       filter; ...
```

```
146                     could not ...
                           threshold.\n',ArrayOfFCSFiles{i}.FileName, ...
                           ...
147                     CurrentFilterName);
148             else
149                 [Total, Above] = ...
                       ArrayOfFCSFiles{i}.threshold(FilterNumber);
150                 TotalAndAbove(i,1:2) = [Total, Above];
151                 Labels{i} = ArrayOfFCSFiles{i}.FileName;
152             end
153         end
154         assignin('base', 'Labels', Labels);
155         assignin('base', 'TotalAndAbove', TotalAndAbove);
156     end
157     function AllFilesMeanButton_Callback(source,eventdata)
158         MeanAndStdev = [];
159         Labels = {};
160         CurrentFilterName = ...
               ArrayOfFCSFiles{get(hFileList,'Value')}.FilterNames{ ...
               ...
161              get(hFilterPopup,'Value')};
162         for i=1:length(ArrayOfFCSFiles)
163             FilterNumber = ...
                   find(ismember(ArrayOfFCSFiles{i}.FilterNames, ...
164                  CurrentFilterName));
165             if isempty(FilterNumber)
166                 fprintf('The file %s does not have a %s ...
                       filter; ...
167                       could not ...
                             threshold.\n',ArrayOfFCSFiles{i}.FileName, ...
                             ...
168                       CurrentFilterName);
169             else
170                 [Mean, Stdev] = ...
                       meanandstdev(ArrayOfFCSFiles{i},FilterNumber);
171                 MeanAndStdev(i,1:2) = [Mean, Stdev];
172                 Labels{i} = ArrayOfFCSFiles{i}.FileName;
173             end
174         end
175         assignin('base', 'Labels', Labels);
176         assignin('base', 'MeanAndStdDev', MeanAndStdev);
177     end
178     function AddButton_Callback(source,eventdata)
179         ListOfFiles = get(hFileList,'String');
180         FileNames = uigetfile('*.fcs','Add an FCS ...
               file','MultiSelect','on');
181         if isa(FileNames,'cell')
```

```
182             for i=1:length(FileNames)
183                 ListOfFiles{length(ListOfFiles) + 1} = ...
                        FileNames{i};
184                 ArrayOfFCSFiles{length(ArrayOfFCSFiles)+1} = ...
                        FCSFile(FileNames{i});
185             end
186             set(hFileList,'String',ListOfFiles);
187         elseif isa(FileNames,'numeric')
188             return;
189         else
190             ListOfFiles{length(ListOfFiles) + 1} = FileNames;
191             set(hFileList,'String',ListOfFiles);
192             ArrayOfFCSFiles{length(ArrayOfFCSFiles)+1} = ...
                    FCSFile(FileNames);
193         end
194     end
195     function RemoveButton_Callback(source,eventdata)
196         ListOfFiles = get(hFileList,'String');
197         SelectedFile = get(hFileList,'Value');
198         ListOfFiles(SelectedFile) = [];
199         ArrayOfFCSFiles(SelectedFile) = [];
200         set(hFileList,'Value',1);
201         set(hFileList,'String',ListOfFiles);
202         cla(hAxes);
203         cla(hAxesFilter);
204         set(hFilterPopup,'Value',1);
205         set(hFilterPopup,'String','No Filters Yet');
206         set(hTotalEntry,'String','0');
207         set(hAboveEntry,'String','0');
208         set(hThresholdEntry,'String','0');
209     end
210     function UpdateThresholdButton_Callback(source,eventdata)
211         ArrayOfFCSFiles{get(hFileList,'Value')}.FilterThresholds( ...
                ...
212             get(hFilterPopup,'Value')) = ...
                    str2num(get(hThresholdEntry,'String'));
213         PlotHistogram();
214         UpdateThresholdCounts;
215     end
216
217     function UpdateBoundsButton_Callback(source,eventdata)
218         hBounds = impoly(hAxes);
219         BoundPosition = getPosition(hBounds)';
220         ArrayOfFCSFiles{get(hFileList,'Value')}.setbounds( ...
221             BoundPosition(1,:), BoundPosition(2,:));
222         cla(hAxes);
223         ArrayOfFCSFiles{get(hFileList,'Value')}.showfscssc(hAxes);
```

```
224         PlotHistogram();
225         UpdateThresholdCounts;
226     end
227     function ChangeFileSelection_Callback(source,eventdata)
228         cla(hAxes);
229         ArrayOfFCSFiles{get(hFileList,'Value')}.showfscssc(hAxes);
230         set(hFilterPopup,'String', ...
               ArrayOfFCSFiles{get(hFileList,'Value')}.FilterNames);
231         set(hThresholdEntry,'String',num2str(ArrayOfFCSFiles{ ...
               ...
232             get(hFileList,'Value')}.FilterThresholds(get( ...
233             hFilterPopup,'Value'))));
234         cla(hAxesFilter);
235         PlotHistogram();
236         UpdateThresholdCounts;
237     end
238     function ChangeFilterSelection_Callback(source,eventdata)
239          set(hThresholdEntry,'String',num2str(ArrayOfFCSFiles{get( ...
               ...
240             hFileList,'Value')}.FilterThresholds(get( ...
241             hFilterPopup,'Value'))));
242          PlotHistogram();
243          UpdateThresholdCounts;
244     end
245 end
```

```
classdef FCSFile < handle
    properties
        FileName = '';
        BoundsX = [];
        BoundsY = [];
        FilterNames = {};
        FilterThresholds = [];
        FSCFilter = 0;
        SSCFilter = 0;
    end
    methods
        function FF = FCSFile(FileName)
            FF.FileName = FileName;
            [fcsdat, fcshdr] = fca_readfcs(FF.FileName);
            clear fcsdat;
            for i=1:fcshdr.NumOfPar
                FF.FilterNames{i} = fcshdr.par(1,i).name;
                FF.FilterThresholds(i) = 0;
            end
            for i=1:length(FF.FilterNames)
                if strcmp(FF.FilterNames{i},'FSC-A')
                    FF.FSCFilter = i;
                else
                    if strcmp(FF.FilterNames{i},'SSC-A')
                        FF.SSCFilter = i;
                    end
                end
            end
            FF.BoundsX = [0.5E5 0.5E5 1E5 1E5];
            FF.BoundsY = [1E4 5E4 5E4 1E4];
        end
      function FilterValues = getfiltervalues(FF,i)
          [fcsdat, fcshdr] = fca_readfcs(FF.FileName);
          InsideBounds = inpolygon(fcsdat(:,FF.FSCFilter), ...
             fcsdat(:,FF.SSCFilter),FF.BoundsX,FF.BoundsY);
          FilterValues = fcsdat(InsideBounds,i);
      end
      function setbounds(FF,NewBoundsX,NewBoundsY)
          FF.BoundsX = NewBoundsX;
          FF.BoundsY = NewBoundsY;
      end
      function [Total, Above] = threshold(FF,FilterNumber)
          [fcsdat, fcshdr] = fca_readfcs(FF.FileName);
          InsideBounds = FF.getfiltervalues(FilterNumber);
          Total = length(find(InsideBounds > 0));
          Above = length(find(InsideBounds > ...
             FF.FilterThresholds(FilterNumber)));
```

```
47        end
48        function [Mean, Stdev] = meanandstdev(FF,FilterNumber)
49            InsideBounds = FF.getfiltervalues(FilterNumber);
50            Mean = mean(InsideBounds);
51            Stdev = std(InsideBounds);
52        end
53        function showfscssc(FF,hAxes)
54            axes(hAxes);
55            [fcsdat, fcshdr] = fca_readfcs(FF.FileName);
56            fcsdat = fcsdat(1:min(10000,length(fcsdat)),:);
57            scatter(fcsdat(:,FF.FSCFilter), ...
                  fcsdat(:,FF.SSCFilter),1, 'k'); hold on;
58            hroi=fill(FF.BoundsX,FF.BoundsY,'r');
59            set(hroi,'FaceColor','none','EdgeColor',[1 0 0]);
60            xlim([0 250000]);
61            ylim([0 250000]);
62            xlabel('FSC-A');
63            ylabel('SSC-A');
64        end
65    end
66 end
```

```
1  function [fcsdat, fcshdr, fcsdatscaled] = fca_readfcs(filename)
2  % This file is required for operation of our code and therefore
3  % included here to facilitate reproducibility. - MEW, 2014
4  % Ver 2.5, Laszlo Balkay, balkay@pet.dote.hu
5  % 2006-2009 / University of Debrecen, Institute of Nuclear ...
      Medicine
6  if nargin == 0
7       [FileName, FilePath] = uigetfile('*.*','Select fcs2.0 ...
           file');
8       filename = [FilePath,FileName];
9       if FileName == 0;
10           fcsdat = []; fcshdr = [];
11           return;
12      end
13 else
14     filecheck = dir(filename);
15     if size(filecheck,1) == 0
16         hm = msgbox([filename,': The file does not ...
               exist!'], ...
17              'FcAnalysis info','warn');
18         fcsdat = []; fcshdr = [];
19         return;
20     end
21 end
```

```
[FilePath, FileNameMain, fext] = fileparts(filename);
FilePath = [FilePath filesep];
FileName = [FileNameMain, fext];
if  isempty(FileNameMain)
    currend_dir = cd;
    cd(FilePath);
    [FileName, FilePath] = uigetfile('*.*','Select FCS file');
     filename = [FilePath,FileName];
     if FileName == 0;
          fcsdat = []; fcshdr = [];
          return;
     end
     cd(currend_dir);
end
fid = fopen(filename,'r','b');
fcsheader_1stline   = fread(fid,64,'char');
fcsheader_type = char(fcsheader_1stline(1:6)');
if strcmp(fcsheader_type,'FCS1.0')
    hm = msgbox('FCS 1.0 file type is not ...
       supported!','FcAnalysis info','warn');
    fcsdat = []; fcshdr = [];
    fclose(fid);
    return;
elseif  strcmp(fcsheader_type,'FCS2.0') || ...
   strcmp(fcsheader_type,'FCS3.0')
    fcshdr.fcstype = fcsheader_type;
    FcsHeaderStartPos   = ...
       str2num(char(fcsheader_1stline(16:18)'));
    FcsHeaderStopPos    = ...
       str2num(char(fcsheader_1stline(23:26)'));
    FcsDataStartPos     = ...
       str2num(char(fcsheader_1stline(31:34)'));
    status = fseek(fid,FcsHeaderStartPos,'bof');
    fcsheader_main = ...
       fread(fid,FcsHeaderStopPos-FcsHeaderStartPos+1,'char');
    warning off MATLAB:nonIntegerTruncatedInConversionToChar;
    fcshdr.filename = FileName;
    fcshdr.filepath = FilePath;
    if fcsheader_main(1) == 12
        mnemonic_separator = 'FF';
    else
        mnemonic_separator = char(fcsheader_main(1));
    end
    if mnemonic_separator == '@';
        hm = msgbox([FileName,': The file can not be read ...
           (Unsupported FCS type: WinMDI histogram ...
           file)'],'FcAnalysis info','warn');
```

```
61          fcsdat = []; fcshdr = [];
62          fclose(fid);
63          return;
64      end
65      fcshdr.TotalEvents = ...
           str2num(get_mnemonic_value('$TOT',fcsheader_main, ...
           mnemonic_separator));
66      fcshdr.NumOfPar = ...
           str2num(get_mnemonic_value('$PAR',fcsheader_main, ...
           mnemonic_separator));
67      fcshdr.Creator = ...
           get_mnemonic_value('CREATOR',fcsheader_main, ...
           mnemonic_separator);
68      for i=1:fcshdr.NumOfPar
69          fcshdr.par(i).name = ...
               get_mnemonic_value(['$P',num2str(i),'N'], ...
70              fcsheader_main, mnemonic_separator);
71          fcshdr.par(i).range = ...
               str2num(get_mnemonic_value(['$P',num2str(i),'R'], ...
               ...
72              fcsheader_main, mnemonic_separator));
73          fcshdr.par(i).bit = ...
               str2num(get_mnemonic_value(['$P',num2str(i),'B'], ...
               ...
74              fcsheader_main, mnemonic_separator));
75          par_exponent_str= ...
               (get_mnemonic_value(['$P',num2str(i),'E'], ...
76              fcsheader_main, mnemonic_separator));
77          if isempty(par_exponent_str)
78              islogpar = ...
                   get_mnemonic_value(['P',num2str(i),'DISPLAY'], ...
                   ...
79                  fcsheader_main, mnemonic_separator);
80              if islogpar == 'LOG'
81                  par_exponent_str = '5,1';
82              else
83                  par_exponent_str = '0,0';
84              end
85          end
86          par_exponent= str2num(par_exponent_str);
87          fcshdr.par(i).decade = par_exponent(1);
88          if fcshdr.par(i).decade == 0
89              fcshdr.par(i).log = 0;
90              fcshdr.par(i).logzero = 0;
91          else
92              fcshdr.par(i).log = 1;
93              if (par_exponent(2) == 0)
```

```
94                fcshdr.par(i).logzero = 1;
95              else
96                fcshdr.par(i).logzero = par_exponent(2);
97              end
98          end
99      end
100     fcshdr.starttime = ...
           get_mnemonic_value('$BTIM',fcsheader_main, ...
           mnemonic_separator);
101     fcshdr.stoptime = ...
           get_mnemonic_value('$ETIM',fcsheader_main, ...
           mnemonic_separator);
102     fcshdr.cytometry = ...
           get_mnemonic_value('$CYT',fcsheader_main, ...
           mnemonic_separator);
103     fcshdr.date = ...
           get_mnemonic_value('$DATE',fcsheader_main, ...
           mnemonic_separator);
104     fcshdr.byteorder = ...
           get_mnemonic_value('$BYTEORD',fcsheader_main, ...
           mnemonic_separator);
105     fcshdr.datatype = ...
           get_mnemonic_value('$DATATYPE',fcsheader_main, ...
           mnemonic_separator);
106     fcshdr.system = ...
           get_mnemonic_value('$SYS',fcsheader_main, ...
           mnemonic_separator);
107     fcshdr.project = ...
           get_mnemonic_value('$PROJ',fcsheader_main, ...
           mnemonic_separator);
108     fcshdr.experiment = ...
           get_mnemonic_value('$EXP',fcsheader_main, ...
           mnemonic_separator);
109     fcshdr.cells = ...
           get_mnemonic_value('$Cells',fcsheader_main, ...
           mnemonic_separator);
110     fcshdr.creator = ...
           get_mnemonic_value('CREATOR',fcsheader_main, ...
           mnemonic_separator);
111 else
112     hm = msgbox([FileName,': The file can not be read ...
           (Unsupported FCS type)'],'FcAnalysis info','warn');
113     fcsdat = []; fcshdr = [];
114     fclose(fid);
115     return;
116 end
117 status = fseek(fid,FcsDataStartPos,'bof');
```

```
118 if strcmp(fcsheader_type,'FCS2.0')
119     if strcmp(mnemonic_separator,'\') || ...
           strcmp(mnemonic_separator,'FF')...
120             || strcmp(mnemonic_separator,'/')
121         if fcshdr.par(1).bit == 16
122             fcsdat = uint16(fread(fid,[fcshdr.NumOfPar ...
                   fcshdr.TotalEvents],'uint16')');
123             if strcmp(fcshdr.byteorder,'1,2')...
124                     || strcmp(fcshdr.byteorder, '1,2,3,4')
125                 fcsdat = ...
                       bitor(bitshift(fcsdat,-8),bitshift(fcsdat,8));
126             end
127         elseif fcshdr.par(1).bit == 32
128                 if fcshdr.datatype ~= 'F'
129                     fcsdat = (fread(fid,[fcshdr.NumOfPar ...
                           fcshdr.TotalEvents],'uint32')');
130                 else
131                     fcsdat = (fread(fid,[fcshdr.NumOfPar ...
                           fcshdr.TotalEvents],'float32')');
132                 end
133         else
134             bittype = ['ubit',num2str(fcshdr.par(1).bit)];
135             fcsdat = fread(fid,[fcshdr.NumOfPar ...
                   fcshdr.TotalEvents],bittype, 'ieee-le')';
136         end
137     elseif strcmp(mnemonic_separator,'!');
138         fcsdat_ = fread(fid,[fcshdr.NumOfPar ...
               fcshdr.TotalEvents],'uint16', 'ieee-le')';
139         fcsdat = zeros(fcshdr.TotalEvents,fcshdr.NumOfPar);
140         for i=1:fcshdr.NumOfPar
141             bintmp = dec2bin(fcsdat_(:,i));
142             fcsdat(:,i) = bin2dec(bintmp(:,7:16));
143         end
144     end
145     fclose(fid);
146 elseif strcmp(fcsheader_type,'FCS3.0')
147     if strcmp(mnemonic_separator,'|')
148         fcsdat_ = (fread(fid,[fcshdr.NumOfPar ...
               fcshdr.TotalEvents],'uint16','ieee-le')');
149         fcsdat = zeros(size(fcsdat_));
150         new_xrange = 1024;
151         for i=1:fcshdr.NumOfPar
152             fcsdat(:,i) = ...
                   fcsdat_(:,i)*new_xrange/fcshdr.par(i).range;
153             fcshdr.par(i).range = new_xrange;
154         end
155     else
```

```
156          fcsdat = fread(fid,[fcshdr.NumOfPar ...
                fcshdr.TotalEvents],'float32')';
157      end
158      fclose(fid);
159  end
160  fcsdatscaled = zeros(size(fcsdat));
161  for  i = 1 : fcshdr.NumOfPar
162      Xlogdecade = fcshdr.par(i).decade;
163      XChannelMax = fcshdr.par(i).range;
164      Xlogvalatzero = fcshdr.par(i).logzero;
165      if ¬fcshdr.par(i).log
166         fcsdatscaled(:,i)  = fcsdat(:,i);
167      else
168         fcsdatscaled(:,i) = ...
               Xlogvalatzero*10.^(double(fcsdat(:,i))/ ...
169             XChannelMax*Xlogdecade);
170      end
171  end
172  function mneval = ...
        get_mnemonic_value(mnemonic_name,fcsheader,mnemonic_separator)
173  if strcmp(mnemonic_separator,'\')  || ...
        strcmp(mnemonic_separator,'!') ...
174           || strcmp(mnemonic_separator,'|') || ...
                strcmp(mnemonic_separator,'@')...
175           || strcmp(mnemonic_separator, '/')
176      mnemonic_startpos = ...
         findstr(char(fcsheader'),mnemonic_name);
177      if isempty(mnemonic_startpos)
178          mneval = [];
179          return;
180      end
181      mnemonic_length = length(mnemonic_name);
182      mnemonic_stoppos = mnemonic_startpos + mnemonic_length;
183      next_slashes = ...
         findstr(char(fcsheader(mnemonic_stoppos+1:end)') ...
184          ,mnemonic_separator);
185      next_slash = next_slashes(1) + mnemonic_stoppos;
186      mneval = char(fcsheader(mnemonic_stoppos+1:next_slash-1)');
187  elseif strcmp(mnemonic_separator,'FF')
188      mnemonic_startpos = ...
         findstr(char(fcsheader'),mnemonic_name);
189      if isempty(mnemonic_startpos)
190          mneval = [];
191          return;
192      end
193      mnemonic_length = length(mnemonic_name);
194      mnemonic_stoppos = mnemonic_startpos + mnemonic_length ;
```

```
195      next_formfeeds = find( ...
            fcsheader(mnemonic_stoppos+1:end) == 12);
196      next_formfeed = next_formfeeds(1) + mnemonic_stoppos;
197      mneval = char(fcsheader(mnemonic_stoppos + 1 : ...
            next_formfeed-1)');
198  end
```