**Title**
Different solutions to the regulation of cell type in yeast provide insight into evolutionary rewiring of transcriptional circuits



**Author**
Del Frate, Francesca

**Publication Date**
2022

Peer reviewed|Thesis/dissertation

Different solutions to the regulation of cell type in yeast provide insight into evolutionary rewiring of transcriptional circuits

by
Francesca Del Frate

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Barbara Panning
B18F20197C95417...

Barbara Panning
Chair

DocuSigned by:
Kaveh Ashrafi

Kaveh Ashrafi

DocuSigned by:
Alexander Johnson
E2BBE6A63C2745B...

Alexander Johnson

Committee Members

## Acknowledgments

It really takes a village to raise a PhD student, and I am no exception. It's difficult to describe all of the ways in which many people have supported me through this process, but I will do my best to mention the major contributions here. To my thesis advisor, Sandy Johnson, who made me realize that it was possible to take abstract topics like regulatory evolution and actually do the molecular biology to test hypotheses. He also made sure I learned the importance of having the proper controls and set an example for how to communicate dense, complex topics clearly and simply. Running a lab is a complex job, but Sandy always made sure we had an intellectually lively community of scientists. Starting with Kyle Fowler, who was my rotation mentor, I was introduced to a group of thoughtful and curious people who were always happy to help. Candace Britton and Liron Noiman added plenty more support, whether I needed help with preparing for my qualifying exam, a technique, or just wanted to talk to someone. Chein-der Lee, Carrie Graham, Naomi Ziv, Sheena Singh-Babak and Matt Lohse were always ready to explain a technique or help me interpret a result. Lucas Brenes, our technician at the time, was always ready give helpful advice and talk about a fun idea about an experiment. We all had so many amazing conversations, cracking jokes about current events, discussing various hobbies, and complaining about chaotically given seminar talks. A big thank you to the people who kept things running, Ananda Maya making sure we had all the glassware, media and miscellaneous lab supplies I could possibly need and Ethel-Enoex Gonodoo, who could efficiently get to the bottom of any administrative issue. I was really excited to have a new fellow graduate student when Haley Gause joined the year after me, and then another when Jenny Zhang joined the

year after. Niyati Rodricks joined as a technician, rounding out the crew. A few years later Megan Garber joined, and even though we've only overlapped for a year, she stepped in and started a side project working with me on the bioinformatic analyses that ended up in Chapter 1 of this thesis. She also took on the role of mentoring me through the process of wrapping up this project and writing the paper, helping me stay on track with writing and providing some brilliant edits on the paper. She also contributed intellectually, looking at the tripartite site with me and figuring out that the Mat$\alpha$2 site was actually reversed with respect to Mata1, a central point that helped us wrap up the story.

Of course I have to mention my amazing TETRAD class, the biggest class our program has had in years, which was full of amazing people that I learned so much from and that I could always count on for support. I have really been so lucky to meet such an amazing group of people through graduate school, a group of people that became my support network and family through the hard times of research through a pandemic as well as various other crises. In particular, my neighborhood friends and roommates, Ady Steinbach, Haley Gause, Elise Munoz, Luke Strauskulage, Eric Simental, Donovan Trinidad, Chris Carlson, Varun Bhadkamkar, and Dana Kennedy, who are always down for a quick drink at Pilsner Inn or a nice long afternoon at Dolores Park. And Henry Ng, Katie Augspurger, and Lili Kim, who lived a little further, but were always around to chat on campus. I also have to thank my college friends, without whom I wouldn't have made it to graduate school in the first place, particularly Elisabeth Meyer, Conrad Shock, and Ara Parikh, who were my Bay Area crew of college friends and have continued to be an amazing support network in my life. I also want to thank D'anne Duncan, who really

supported me with my transition into starting graduate school and choosing a lab, Peggy Ryan, who helped me learn how to manage my time and become more confident in my strengths, and Toni Hurley, our brilliant TETRAD program administrator, who is always there for us students. And to my committee, Kaveh Ashrafi and Barbara Panning, who were extremely supportive of my scientific and personal progress through the difficult times as well as the good times.

And now my family, my parents Renzo and Lourdes Del Frate, and my brother and sister, Enrico and Isabella Del Frate. To my parents, thank you for always supporting my curiosity and eclectic, nerdy interests, encouraging me to ask questions and share things I learned that I was excited about. And thank you for always being there for me when the going got tough, I always knew I could count on you when I needed support and I can't begin to explain how important that was. To my siblings, thank you for being my partners in crime, letting me tell you about things I was interested in and challenging my assumptions about the world and my own knowledge. It's been amazing to grow up with you, and see you two finding and taking your own paths.

Finally I do want to thank some of my earliest science mentors, Jack Szostak who took me on as an undergraduate at Harvard when I cold emailed him, interested in working on the origins of life. And Anders Björkbom and Noam Prywes, two of my direct mentors in the lab, who taught me all the techniques when I was just starting out and made me see that I too, could do experiments and discover things in a field I was interested in. Again, there are definitely many more people that have made

contributions both large and small, and I am grateful to all of you, I couldn't have done it without my village.

Bioinformatic searches in Figure1.2c, Figure 1.5, and supplementary figure 1.3 were done by Megan Garber with motifs I made, and she generated the figure panels for that work. All other experiments were designed and done by me and supervised by Alexander Johnson. Megan Garber made the major intellectual contribution of figuring out the orientation of the central Mat$\alpha$2 site within the tripartite site, which is in Chapter 1. Manuscript was written by me, in collaboration with Alexander Johnson, with Megan Garber contributing edits to Chapter 1.

# Different solutions to the regulation of cell type in yeast provide insight into evolutionary rewiring of transcriptional circuits

Francesca Del Frate

## Abstract

Though the outputs of regulatory circuits are conserved over long timescales, the exact mechanisms of regulation change comparatively frequently. One such example is the regulation of cell type in yeast, specifically the haploid specific genes. These are transcribed in both of the mating competent cell types, **a** and $\alpha$, and not in the diploid **a**/$\alpha$ cell type. The simplest and likely ancestral mode of regulation is direct repression of the haploid specific genes by the Mata1-Mat$\alpha$2 heterodimer in the **a**/$\alpha$ cell. However, this is not the only solution.

Here we discuss two examples where the output of the circuit has been maintained but the molecular mechanism is different in the regulation of haploid specific genes in yeast. After bioinformatic searches indicated the lack of a Mata1-Mat$\alpha$2 site in *GPA1*—one of the haploid specific genes in *Lachancea kluyveri*—further inspection revealed a tripartite Mata1-Mat$\alpha$2-Mcm1 site in *GPA1.* ChIPseq of Mat$\alpha$2 and reporter experiments testing the tripartite site confirmed that this gene is directly repressed by tripartite Mata1-Mat$\alpha$2-Mcm1, while confirming that the other haploid specific genes are repressed by Mata1-Mat$\alpha$2. Models made from existing structural data further supported that the three proteins could bind the tripartite site to co-repress *GPA1.* This depends on an ancestral gain of a domain on Mat$\alpha$2 that enables interaction with Mcm1.

In the other example, in the species *Wickerhamomyces. anomalous,* a lack of evidence for Mata1-Mat$\alpha$2 binding in the upstream regions of all haploid specific genes—except the transcription factor Rme1— indicated that Mata1-Mat$\alpha$2 regulation might be indirect for these genes. We knocked out Rme1, and by assaying the effect on mating and transcriptionally profiling the haploid specific genes with RNAseq, we found that two of the haploid specific genes are activated by Rme1. Further bioinformatic analysis suggests that this is direct regulation by Rme1. This is similar to indirect haploid specific gene regulation via Rme1 in another species, *K. lactis*, indicating that this likely happened more than once, and that *Rme1's* ancestral regulation by Mata1-Mat$\alpha$2 positioned it to acquire this new role in regulating haploid specific genes.

In both examples, transcriptional regulators already associated with the transcriptional circuit gained a new regulatory role with a few cis changes in target genes. Together, these examples illustrate how changes in regulatory circuits can build on each other to create new regulatory architectures, adding to our overall understanding of how transcriptional regulation shifts over time.

# Table of Contents

## List of Figures

## List of Tables

# Chapter 1

The ancestral gain of a protein-protein interaction preceded regulatory gain of three part repression in a single haploid specific gene in the yeast *Lachancea kluyveri*

## Introduction

Changes in transcription circuits over evolutionary timescales are a major source of phenotypic novelty. Two major sources of transcriptional plasticity have been well-documented: (1) changes in the cis-regulatory sequences of a gene, which can directly alter the pattern of expression of that gene (and indirectly affect the expression of other genes) and (2) the formation (and breaking) of cooperative interactions between different transcriptional regulators, which can directly affect the expression of many genes simultaneously[1–5]. Typically, the two types of changes are observed together in new circuit architectures. Both types of changes can occur without extensive pleiotropy; the former directly affects expression of only the gene in which it occurs, and the latter—because it is often due to the creation of a relatively weak protein-protein interaction in a part of the protein distinct from the DNA-binding domain—typically does not compromise the ancestral roles of the protein[1,6–8]. In contrast, changes in the intrinsic DNA-binding specificity of a conserved transcription regulator over evolutionary timescales seem to occur much less frequently. In the absence of gene duplication, such changes would likely compromise the existing roles of the protein and would be not be maintained.

While some evolutionary changes in transcription lead to dramatic new phenotypes, other studies indicate that the mechanisms of regulation can apparently drift between different molecular solutions while maintaining the same output[7,9]. Understanding these cases in detail provides an opportunity to understand the molecular principles behind transcription circuit plasticity. In this paper, we document

and explain a clear example of this type of plasticity in the regulation of the mating genes in the ascomycete (yeast) lineage.

We concentrate on a group of genes known as the haploid-specific genes, which are expressed in the two mating cell types (**a** and α) but repressed in the third cell type, the **a**/α cell (Figure 1.1a). The **a**/α cell is the product of the mating of an **a** cell and an α cell and itself is non-mating. The haploid-specific genes code for proteins required for both **a** and α cells to mate; for example, three code for the components of the trimeric G protein needed for pheromone signaling[10]. Their repression in the **a**/α cell therefore makes logical sense as their products are not needed, and could even be detrimental, in this cell type.

In many ascomycetes, the haploid-specific genes are repressed directly by a heterodimer of two homeodomain proteins, Mata1 and Matα2[11–13]. Again, this logic makes conceptual sense: Mata1 is made by a cells and Matα2 by α cells; only when the two proteins are synthesized together in **a**/α cells (the result of mating) does the heterodimer form and repress the haploid specific genes.

Although direct repression by the Mata1-Matα2 heterodimer is highly logical and greatly appealing in its simplicity, there are exceptions to this mechanism. In *Kluyveromyces lactis*, the repression of the haploid-specific genes is indirect: the Mata1-Matα2 heterodimer represses an activator of the haploid-specific genes but does not bind these genes directly[14]. And in *Wickerhamomyces anomalus,* the Mata1-Matα2 heterodimer requires a third protein, Mcm1, to repress at least one of the haploid-specific genes[9].

In this paper, we investigated regulation of the haploid-specific genes in *Lachancea kluyveri*, a species that branched from *S. cerevisiae* well after the *S. cerevisiae*-*W. anomalus* branchpoint (Figure 1.1b). We were drawn to this species because bioinformatic analyses indicated that one of the haploid-specific genes (*GPA1*, which codes for the alpha subunit of the trimeric G protein) appeared to lack a conventional Mata1-Mat$\alpha$2 heterodimer binding site, whereas other haploid genes in this species (and in many other species) clearly displayed this signature motif[14]. In this paper, we show that *GPA1* is not regulated in the conventional, deeply-conserved manner but is repressed in the **a**/$\alpha$ cell by three proteins working together, Mata1, Mat$\alpha$2, and Mcm1. In this three-part regulatory complex, we show that any pair of proteins is not sufficient to bring about repression due to non-optimal cis-regulatory sequences, resulting in the requirement for all three proteins. Mcm1 is produced in all three cell types, so the logic of regulation is preserved: repression occurs only in the **a**/$\alpha$ cell type, despite the idiosyncratic arrangement of proteins on DNA. We discuss possible evolutionary pathways that could have led to this unusual, non-canonical mechanism of regulation.

## Results

*Regulation of the haploid specific genes in Lachancea kluyveri*

To study the way in which the haploid -specific genes are regulated in *L. kluyveri*, we first performed an Rnaseq analysis to identify the haploid-specific genes by comparing gene expression across the three cell types: **a**, $\alpha$, and **a**/$\alpha$ (Figure 1.1a). Haploid-specific genes are defined here as genes that are expressed in **a** and $\alpha$ cells but not in **a**/$\alpha$ cells. We identified approximately 30 haploid specific genes, including those encoding the three subunits of the trimeric G-protein that mediates pheromone response (*GPA1, STE4, STE18*), the cyclin dependent kinase inhibitor (*FAR1*) that triggers cell cycle arrest as part of the mating response and *RME1*, a transcription regulator with a variety of functions (Figure 1.2a, Table 1.)[10]. These five genes are haploid specific genes in many other fungal species, indicating a deeply conserved expression pattern, and these are the genes we concentrate on for the remainder of the paper[9,10,14].

As discussed in the Introduction, the haploid-specific genes in most species are repressed by direct binding of the Mata1-Mat$\alpha$2 heterodimer, both subunits of which are synthesized only in the **a**/$\alpha$ cell[12,13]. To test whether this is the case in *L.kluyveri,* we performed a chromatin immunoprecipitation using tagged Mat$\alpha$2 in the **a**/$\alpha$ cell (Figure 1.2b and Supplementary Figure 1.1)**.** We identified seven high-confidence peaks including those spanning the upstream regions of *GPA1, STE4, STE18, FAR1 and RME1* (Figure 1.2B and Supplementary Figure 1.1). A bioinformatic search found conventional Mata1-Mat$\alpha$2 motifs upstream of only four of these genes(Figure

1.2c)(Supplementary Figure 1.2a). The exception was *GPA1* where, as discussed in the introduction, the motif appeared to be missing--even though *GPA1* exhibited clear haploid specific gene expression and an obvious Mat$\alpha$2 ChIP signal (Figure 1.2a, b, c). This apparent contradiction led us to manually examine the DNA sequence under the Mat$\alpha$2 ChIP peak. We identified a Mat$\alpha$2 DNA sequence motif and a Mata1 motif, but the orientation of the Mat$\alpha$2 motif was "backwards" relative to the Mata1 motif, and the spacing between the two motifs was three base pairs shorter than that of the conventional heterodimer site(Supplementary Figure 1.2b). These differences explain the failure of a position-weighted motif searching algorithm (based on the conserved heterodimer site) to highlight this site (Figure 1.2c). We also noticed a two-fold symmetric motif for Mcm1, a protein known to interact with Mat$\alpha$2 for a different role in the cell, repression of the a-specific genes in $\alpha$ cells(Figure 1.2d; Supplementary Figure 1.2c). Thus, it appeared as though three proteins (and three sequence motifs) were required to repress *GPA1* in *L. kluyveri*, while the other haploid-specific genes in this species contained all the hallmarks of regulation by the conventional Mata1-Mat$\alpha$2 heterodimer.

*Testing the three-site hypothesis*

To test this model of tripartite regulation of *GPA1*, we mutated each of the three sites and measured the effects on repression. To avoid disturbing regulation of the endogenous *GPA1* gene (which could have consequences such as cell-cycle arrest), we created reporter constructs with the sequence upstream of *GPA1* driving the

expression of GFP, which we integrated into the genome(Figure 1.3a). *GPA1* is tightly repressed in the **a**/α cell, and to capture the full dynamic range of regulation, we used qPCR, rather than fluorescence, to directly measure transcript levels. Mutations to the three-part site included independently scrambling each of the three motifs and scrambling all three sites at once. In addition, we constructed a double point mutation in the Mcm1 motif, a change known to destroy binding of Mcm1 to DNA. We know from the expression data in the three cell types that both Mata1 and Matα2 proteins are required for repression of *GPA1*(Figure 1.2a, Table 1.1). We could not test Mcm1 in a similar way because it is essential; however the double point mutation in the Mcm1 binding motif is more specific to Mcm1 than is a scrambled site and thus links the protein to the site.

All of these manipulations disrupted repression, showing that all three sites are needed for proper regulation(Figure 1.3b). In contrast, expression in the α cell is relatively unaffected, so it is unlikely that the tripartite motif plays a major role in the activation of GPA1; rather, it seems to be dedicated solely to repressing the gene in **a**/α cells.

*How are the three proteins arranged on the GPA1 upstream region in L. kluyveri?*

Having demonstrated that all three motifs are needed for repression of GPA1 in **a**/α cells, we next considered how the three proteins might be arranged on this control region and whether this arrangement provided insights into this mode of regulation. As discussed above, the motif corresponding to Matα2 sits between the motifs

corresponding to Mata1 and Mcm1(Figure 1.2D, Figure 1.3D). Superposition of the preferred motif for each protein onto the three-part site therefore strongly indicated that Mat$\alpha$2 was located between Mcm1 and Mata1. When positioned on DNA using matches with their individual motifs, the spacing between Mcm1 and Mat$\alpha$2 is exactly the same as it is when the two proteins interact to repress the a-specific genes(Figure1.3D, Supplementary Figure 1.2c). It therefore seems very likely that the same arrangement of Mcm1 and Mat$\alpha$2 (which allows a favorable protein-protein interaction between the two proteins) occurs on both the a-specific genes (observed in many species) and, idiosyncratically, on the haploid-specific gene GPA1 in *L. kluyveri*. Regarding Mata1, inspection of the sequence showed a strong match to its motif. However, when all three proteins are placed on DNA to match their motifs, Mat$\alpha$2 is positioned "correctly" to interact with Mcm1, but is forced into a "backwards" orientation relative to Mata1, when compared with the conventional, heterodimer arrangement(Figure 1.3c,d,e,f; Supplementary Figure 1.2b). This change in orientation is accomplished by a change in the distance between the Mat$\alpha$2 motif and the Mata1 motif; it is shorter by three base pairs in the GPA1 site than in the conventional motif(Figure 1.3c,d; Supplementary Figure 1.2b).

To investigate this model further (in particular to determine if there are any steric clashes), we used the solved crystal structures of *S.cerevisiae* Mcm1-Mat$\alpha$2 bound to DNA and *S. cerevisiae* Mata1-Mat$\alpha$2 bound to DNA to position the tripartite complex on DNA(Supplementary Figure 1.3b)[15,16]. All three proteins are spatially well accommodated on their preferred motif, with the only remaining question being how Mata1 and Mat$\alpha$2 might interact on the GPA1 site given the differences in orientation

and spacing from the conventional heterodimer site. In *S. cerevisiae*, Mat$\alpha$2 interacts with Mata1 through a short alpha helix at the end of a flexible region; the helix forms only when the two proteins interact. Comparison of the *S. cerevisiae* Mat$\alpha$2 - Mata1 heterodimer structure to Mata1 and Mat$\alpha$2 as positioned on the *GPA1* site in the tripartite complex suggests that the short alpha helix of Mat$\alpha$2 can easily reach the same position of Mata1, indicating that, despite the spacing and orientation differences, the two proteins may interact in fundamentally the same way(Supplementary Figure 1.3a, b). However there must be a severe energetic cost to this altered, non-optimal configuration: when the *GPA1* Mata1-Mat$\alpha$2 site is tested alone (that is, when the Mcm1 motif is mutated) repression by Mata1 and Mat$\alpha$2 is deficient (see Figure 1.3b).

## Discussion

In this paper, we investigate a regulatory system that is deeply conserved in the fungal lineage, namely, repression of the haploid-specific genes by a heterodimer composed of one subunit of the homeodomain protein Mata1 and one subunit of the homeodomain protein Mat$\alpha$2. This is one of the simplest forms of regulation imaginable: One of the subunits (Mata1) is made in **a** cells and the other (Mat$\alpha$2) is made in $\alpha$ cells; only in **a**/$\alpha$ cells, which arise from mating (by cell fusion) between **a** and $\alpha$ cells, are both halves of the heterodimer made in the same cell and the haploid genes repressed. The haploid-specific genes include those that are needed for both **a** and $\alpha$ cells to mate; for example, they encode the components of the trimeric G protein needed for both cell types to respond to mating pheromones.

This simple regulatory scheme is found throughout the ascomycete lineage. This lineage represents approximately the same degree of divergence as that between humans and sponge; therefore, the conventional heterodimer regulatory scheme is widely used[17,18].

Despite its deep conservation and appealing simplicity, we show that this regulatory scheme has a notable variation observed in *L. Kluyveri.* In this species, most of the haploid-specific genes are regulated in the conventional manner, but one gene, *GPA1*, has a novel regulatory scheme that differs in several important ways from the conserved scheme. Specifically, we show that repression of *GPA1* in **a**/$\alpha$ cells of *L. kluyveri* requires binding of Mata1, Mat$\alpha$2, and a third protein Mcm1. When positioned on DNA using motif analysis and prior crystal structures, it becomes clear why no single pair of proteins suffice to bring about repression of *GPA1*, even though

two proteins are sufficient in other contexts(Figure 1.3b,c; Supplementary Figure 1.3a,b). As shown in Figure 1.4, Mcm1 and Mat$\alpha$2 are positioned on *GPA1* DNA exactly as they are when the two proteins carry out a different regulatory function, repression of the a-specific genes. This positioning results in a favorable contact between the two proteins, resulting in their cooperative binding to DNA. Despite this favorable orientation, Mcm1 and Mat$\alpha$2 cannot repress *GPA1* alone—Mata1 is also required (Figure 1.3). The reason for the failure of Mcm1 and Mat$\alpha$2 to work alone on *GPA1* is obvious from prior work: repression of the a-specific genes by these two proteins requires two binding sites for Mat$\alpha$2, one on each side of Mcm1. If one site is experimentally mutated, repression of a-specific genes is destroyed[7]. Thus, the configuration of Mcm1 and Mat$\alpha$2 on *GPA1* resembles a mutant a-specific gene regulatory site and, based on prior work, would not be expected to function, a prediction borne out by direct experiment (Figure 1.3b).

The Mata1-Mat$\alpha$2 pair is also insufficient to repress *GPA1*, and the likely reason for this is also clear. The orientation of the Mat$\alpha$2 subunit is “backwards” compared with the conventional Mat$\alpha$2-Mata1 heterodimer configuration found at haploid-specific genes (Figure 1.4; Figure 1.3c,d,e,f). In addition, the spacing between the Mat$\alpha$2 and Mata1 motifs is substantially altered from the conventional scheme(Supplementary Figure 1.2b, Figure 1.3c,d). Model building (based on the existing crystal structures) suggests that Mata1 and Mat$\alpha$2, as they are arranged on the *L. Kluyveri GPA1* regulatory region, could plausibly contact each other (through a short $\alpha$ helix on a flexible tether) as is observed in the structure of the conventional heterodimer; however, there must be a severe energetic cost to this altered arrangement because it cannot

support repression of *GPA1* in the absence of the Mcm1 binding sequence (Figure 3b)[16,19].

The arguments presented above explain, in energetic terms, why all three proteins are needed to repress the *L. Kluyveri GPA1* gene in **a**/α cells. But how might this novel arrangement have evolved? While we cannot provide a definitive answer, there are some important clues buried in the fungal lineage. At the point where *S. cerevisiae* and *W. anomalous* diverge (prior to the divergence of *S. cerevisiae* and *L. Kluyveri*) all of the protein-protein interactions needed for the three-part scheme on the *L. Kluyveri GPA1* gene were in place[9]. Specifically, the favorable contacts between Matα2 and Mcm1 and between Matα2 and Mata1 had evolved before these two branchpoints. Thus, the shift between the different modes of regulation could be brought about solely through changes in cis-regulatory sequences. Bioinformatic analysis shows that the conventional form of regulation by the Mata1-Matα2 heterodimer is found throughout the ascomycete lineage(Figure 1.5). For example, it applies to the haploid-specific genes in *S. cerevisiae*, in *Candida albicans* and (with the exception of *GPA1*) in *L. Kluyveri*. Given its widespread occurrence--particularly in species where the Matα2-Mcm1 interaction is absent—the conventional, heterodimer form of regulation is almost certainly the ancestral form. Accordingly, the three-part form of regulation is most likely a derived form of regulation. We had previously shown that a similar form of three-part regulation is also found in *W. anomalous*, but on a different haploid-specific gene, *RME1*(Figure 1.4). Based on motif analysis, it is not obvious how the three proteins are arranged on the *RME1* control region in *W. anomalous*. However, there are some sequence similarities (particular in the Mcm1 motifs) between the *W. anomalous RME1*

control region and that of *L. Kluyveri GPA1*, including a sequence that bears some similarity to the tripartite site, suggesting the possibility of divergence from a common ancestor.

This scenario is supported by bioinformatic analyses of neighboring species. A search with a tripartite site finds matches to that site within the upstream regions of *RME1* orthologs of species closely related to *W.anomalus* (Figure 1.5, Supplementary Figure 1.4). The most parsimonious interpretation is that the three-part form of regulation existed for both *GPA1* and *RME1* in an ancestor of *W. anomalous* and *L. Kluyveri* and, in *L. Kluyveri,* regulation of the *RME1* gene reverted to the conventional, heterodimer form. An alternative model holds that the three-part form of regulation arose independently in several different, closely related species. Although we cannot rule out this model, descent from a common ancestral three-part regulation seems more probable. We note that conversion between the heterodimer scheme and the three-part scheme (and back) requires only a few point mutations in the cis-regulatory sequences, and does not appear to require changes to any of the proteins.

Irrespective of the evolutionary pathway, this work highlights an important concept in gene expression: the same output (in this case, repression of the haploid genes in **a**/$\alpha$ cells) can be achieved by different mechanistic solutions and—over evolutionary time scales—the mechanism can drift from one solution to another while maintaining the same output. The key to this idea is that gene expression is typically controlled by assemblies of proteins binding cooperatively to control regions on DNA, and the energetics of assembly can be parceled out in different ways, resulting in different types of arrangements on DNA. For example, in the case described here, a

deficient binding site for the Mata1-Mat$\alpha$2 heterodimer is compensated by a favorable interaction with a third protein, Mcm1. This idea leads to a cautionary note on interpreting a particular gene expression strategy as somehow perfectly optimized. Instead, as evidenced by comparisons across species, a gene expression scheme is best regarded as a flexible set of possible mechanisms, linked by energetically feasible transitions.

## Methods

*Construct Cloning*

Constructs used in Figure 3 were made from TS185, a plasmid containing a hygromycin resistance cassette previously used to stably integrate a GFP transcriptional reporter at the *URA3* locus in *L.kluyveri*[20]*.* The plasmid included restriction sites for Age1 and BsiWI allowing for insertion of putative control sequences to test their effect on gene expression. Custom Geneblocks were designed and ordered from IDT for the 500 base pairs upstream of the *GPA1* transcriptional start site with different manipulations to the putative transcription regulator binding sites. Site manipulations are as pictured in figure panel 3a. These include a wild type *GPA1* upstream sequence, and *GPA1* upstream sequences where the putative sites for Mata1, Mat$\alpha$2, and Mcm1 are individually scrambled. Scrambled sites contained as many changes in the site as possible, while maintaining overall GC content. A sequence with all three of the putative sites scrambled was also constructed, and a gg→cc point mutation in conserved residues of the putative Mcm1 site was also included.

Constructs were made by restriction cloning. The TS185 vector and gene blocks were digested with Age1 and BsiWI-HF and ligated with the Fast-Link DNA Ligation Kit (Lucigen MBTOOL-010) and transformed into Stellar Competent Cells(Takara 636763).

*Strain Construction*

Construct plasmid DNA was linearized by NotI-HF and EcoRV-HF digest to prepare for transformation into yeast. *L. kluyveri* $\alpha$ cells were transformed by electroporation according to protocol published by Gojkovic with some modifications[21,22]. Instead of incubating cells in 1mL YPED for one hour at 25°C and plating onto selective media, 1mL YPED was added to cells after the pulse and this mixture was immediately plated to YPED plates and allowed to grow into a lawn overnight at 30°C. The next day, cells were replica plated to 400ug/mL Hygromycin plates, and allowed to grow for 24 hours. Colonies that arose in that time were patched to -Ura and 5-FOA plates for a second round of selection. Isolates that were Ura- and Hygromycin resistant were grown overnight in 2mL of YPED, and gDNA extracted with a modified Smash n Grab protocol[23]. Cells were spun down, resuspended in 200uL lysis buffer(2% v/v Triton X-100 1% v/v SDS 100 mM NaCl 10 mM Tris-Cl pH 8.0 1 mM EDTA pH 8.0) , 200uL phenol chloroform pH8(Fisher Scientific 68-051-00ML). 200uL 0.5mm glass beads(BioSpec Products 11079105) were added, and samples were lysed for 5 min in a benchtop vortexer using. After bead beating, sampes were spun down at 14,000rpm. 200uL of the supernatant was taken out and precipitated in 1mL of ethanol.

Strains were PCR validated to check both upstream and downstream flanks of the insertion at the *URA3* locus and to check for lack of the *URA3* open-reading frame. Three independent transformants were validated for each construct.

The three isolates of each validated $\alpha$ strain were mated to *L. kluyveri* **a** cells (LB76) by mixing roughly equal amounts of cells from fresh colonies of each cell type onto a fresh YPED plate then left for 3 hours at 30°C, and plated for single colonies.

Single colonies were patched onto SC-ura plates, 5-Foa plates, and Hyg plates. Isolates that were Ura+ and Hyg+ were validated as diploid a/α cells by PCR checks for both the MATa and MATα locus using extracted gDNA. One a/α strain per α isolate was validated and saved, so that each independent transformant would have a matched a/α strain.

The tagged Matα2 a/α strain, FDy18, used in the Chromatin Immunprecipitation experiment (see below) was generated from an existing strain used in a previous study (yLB96) which had a c-terminal 13x Myc tag on the endogenous Matα2 in an α cell(Baker et. al 2012). This strain was mated with a naïve strain (LBy76) of the a cell type as described above to generate the c-terminally Myc tagged Matα2 strain in the a/α. The untagged strain, FDy22, was generated from mating yLB76 and yLB77, the prototrophic a and α strains.

*RNA-Seq*

Cultures were inoculated from single colonies and grown overnight in YPED at 30° C, diluted back to an OD600 of 0.15 in the morning and harvested at an OD600 of 0.6-0.9 as is described in Nocedal et al.[24]. Three replicates of yLB76 (a cell) and yLB77(α cell) were from individual single colonies grown from the same streak. Three replicates from FDy22(a/α cell) were from 3 independently mated isolates. RNA was extracted using the RiboPure RNA purification kit (ThermoFisher AM1924). Total RNA quality was verified on the Agilent Tapestation. Total RNA was poly-a selected with the

NEBNext Poly(A) mRNA Magnetic Isolation Module(Neb E7490S). cDNA synthesis and library preparation was done with the NEBNext Ultra II Directional RNA Library Prep kit for Illumina (Neb E7760L). Quality and concentration of libraries were determined with the Agilent Tapestation. Libraries were pooled in equimolar amounts and sequenced using single end 65 base pair reads on an Illumina HiSeq 4000 in the UCSF Center for Advanced Technologies.

*RNA-Seq Analysis*

Quality of sequencing reads was determined using FastQC[25]. Filtering based on quality and trimming of reads was done using FastP[26]. The assembled genome for *Lachancea kluyveri NRRL Y*-12651 was downloaded from the y1000+ genome database database[17,27]; Trimmed reads were aligned to this reference genome using STAR[28]. A table with counts assigned to genes was generated from the alignments using Rsubread[29]. This count table was then used to determine differentially expressed genes using DESeq2[30]. DESeq2 was run with default parameters, resulting in a list of genes that were differentially expressed in yLB76(**a** cell) when compared to FDy22(**a**/$\alpha$ cell). The same was done for LB77($\alpha$ cell) compared to FDy22(**a**/$\alpha$ cell). Genes with an adjusted p-value < 0.1 from the differential expression outputs in the **a** cells vs. **a**/$\alpha$ cells and the $\alpha$ cells vs. **a**/$\alpha$ cells were plotted against each other in GraphPad Prism. Genes up-regulated with a $\log_2$ fold-change of 2 or more are taken as significantly enriched for expression in the haploid. Genes up-regulated greater than a $\log_2$ fold-change of 2 and an adjusted p-value < 0.1 in both differential expression comparisons (**a** vs. **a**/$\alpha$ and $\alpha$

vs. **a**/α) are considered significantly enriched for expression in the haploid. Expression in one comparison versus the other was plotted using Graphpad Prism.

*Chipseq*

Strains used were the tagged Matα2 **a**/α strain FDy18 described above and the untagged strain FDy22(see above for strain construction). Cells were grown and Chromatin Immunoprecipitation done as in previously published protocols, with some modifications as detailed below[20,24]. For cell lysis, cells were prepared as in the protocol except with 0.5mm Zirconia beads and were lysed by beadbeating in the Omni Bead Ruptor 12. Cells were bead beat with three 90s cycles, alternated with 90s of cooling samples on ice. Additionally, chromatin was sheared by sonication in a Diagenode Bioruptor Pico, 30s on, 30s off, for 25min. Antibody used was Invitrogen Anti-c-Myc Monoclonal (9E10.3) Antibody(ThermoFisher AHO0062) . Sepharose Protein G beads were replaced with 30uL of Dynabeads (ThermoFisher 10004D). Wash and incubation steps were otherwise the same as in published protocols.

Libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB E7645L), and library quality and concentration checked by Agilent Tapestation. Libraries were pooled in equimolar amounts for single end 65 base pair reads using an Illumina Hiseq4000 at the UCSF Center for Advanced Technology.

*Chipseq Analysis*

Reads were trimmed and aligned as described above for the RNAseq. BAM files were processed using DeepTools and Samtools and uploaded to the Integrated Genomics Viewer and the Integrated Genome Browser for visual inspection of data [31–34]Data was processed with the MACS2 program, with default settings except for those regarding duplicates[35]. Instead of removing all duplicate reads, the MACS2 function for keeping biologically relevant duplicates was used, an adjustment recommended for transcription factors with few targets in samples with high read depth.

*qPCR for reporter experiment(Figure 1.3)*

Cultures inoculated from single colonies from three independent genetic isolates of each of the 12 reporter strains(FDy27, FDy28, FDy30, FDy31, FDy32, FDy33, FDy34, FDy35, FDy36, FDy37, FDy38, FDy39, see above for strain construction). Cultures were grown overnight in YPED at 30$^{o}$C, then diluted back to an OD600 of 0.1 in the morning and harvested between an OD600 of 0.7-0.9. One isolate from each condition was grown on the same plate on the same day, so that replicates would account for variability between plates and days as well as between transformants. Cells were flash frozen in liquid nitrogen and RNA extracted using the MasterPure-Yeast-RNA Extraction Kit (Lucigen
MasterPure Yeast RNA Purification Kit MPY03100 ) and protocol was followed with one modification. After the isopropanol precipitation step, RNA was treated with TURBO DNA-free kit(ThermoFisher AM1907). RNA was reverse transcribed with the Superscript III Reverse transcriptase kit(ThermoFisher 18080044) with 250ng of random primers.

qPCR probes against GFP were designed using the NCBI Primer-BLAST tool. Previously verified probes for ACT1, a housekeeping gene between cell types in *L. kluyveri,* were used in this study[14]. cDNA was amplified with probes and 2x iTaq Universal SYBR Green Supermix(Bio-Rad 1725124) on Bio-rad CFX96 Real-Time PCR machine. Ct values were calculated with CFX Maestro software(bio-rad). *GFP* expression was normalized to *ACT1* and to overall expression of all samples. Expression from constructs with the various site mutants was compared to expression from the construct with the wild type sequence to calculate fold repression.

*Generation of Motifs*

Mata1-Mat$\alpha$2-6-Mcm1 motif: Sequences in the *L. kluyveri* genome in which ChIP signal was enriched were extracted using the Integrated Genomics Viewer, and inputted into Meme to generate de novo motifs[33,36]. This sequence was then used to generate a synthetic position specific weight matrix for the tripartite Mata1-Mat$\alpha$2-Mcm1 site in *L.kluyveri*. This consisted of flipping the orientation of the Mat$\alpha$2 motif relative to the Mata1 motif—so that the relative orientation and spacing of the two motifs matches that of the tripartite site upstream of *GPA1* in *L. kluyveri*— and adding an *S. cerevisiae* Mcm1 motif downloaded from the Jaspar database[37], setting the spacing to match the tripartite site in *L. kluyveri*.

Mata1- Mat$\alpha$2-5-Mcm1 motif: In regulation of a-specific genes, there are two different spacings between Mat$\alpha$2 and Mcm1, varying by one base pair. The *L. kluyveri* site has the wider spacing, so we edited the site to remove one basepair and also make a

version with the narrower spacing version, matching a putative tripartite site in *W.anomalus*.

Mata1- Mat$\alpha$2 motif: The 1000bp upstream regions of the 12 haploid specific genes in *S. cerevisae* were extracted using the SGD Sequence Resources Tool[38,39]. These were input into MEME with default settings to generate an Mata1- Mat$\alpha$2 motif[36].

*Bioinformatics search for binding sites upstream of haploid specific genes*

We identified orthologs across budding yeasts for the haploid specific genes, *GPA1, RME1, STE4, STE18, STE5,* by mining data made available by the Y1000 project[17] (see supplemental files 1-5). For each identified ortholog we used its coordinates and direction (positive or negative strand) to append an entry of 1000bp upstream of the gene of interest in its respective genome into a fasta file named after the gene of interest (see supplemental files 6-10). We applied FIMO[40] to search for the respective motifs specified in the text, figures, and visualized in Figure 1.5 with default options and a statistical threshold (p-values) of 1x10-2. The data were visualized by concatenating the highest scoring –log10(q-value), from each independent FIMO search. The resultant high scoring hits were visualized in a heatmap, where orthologs are sorted by their phylogenetic orientation as previously determined[17].

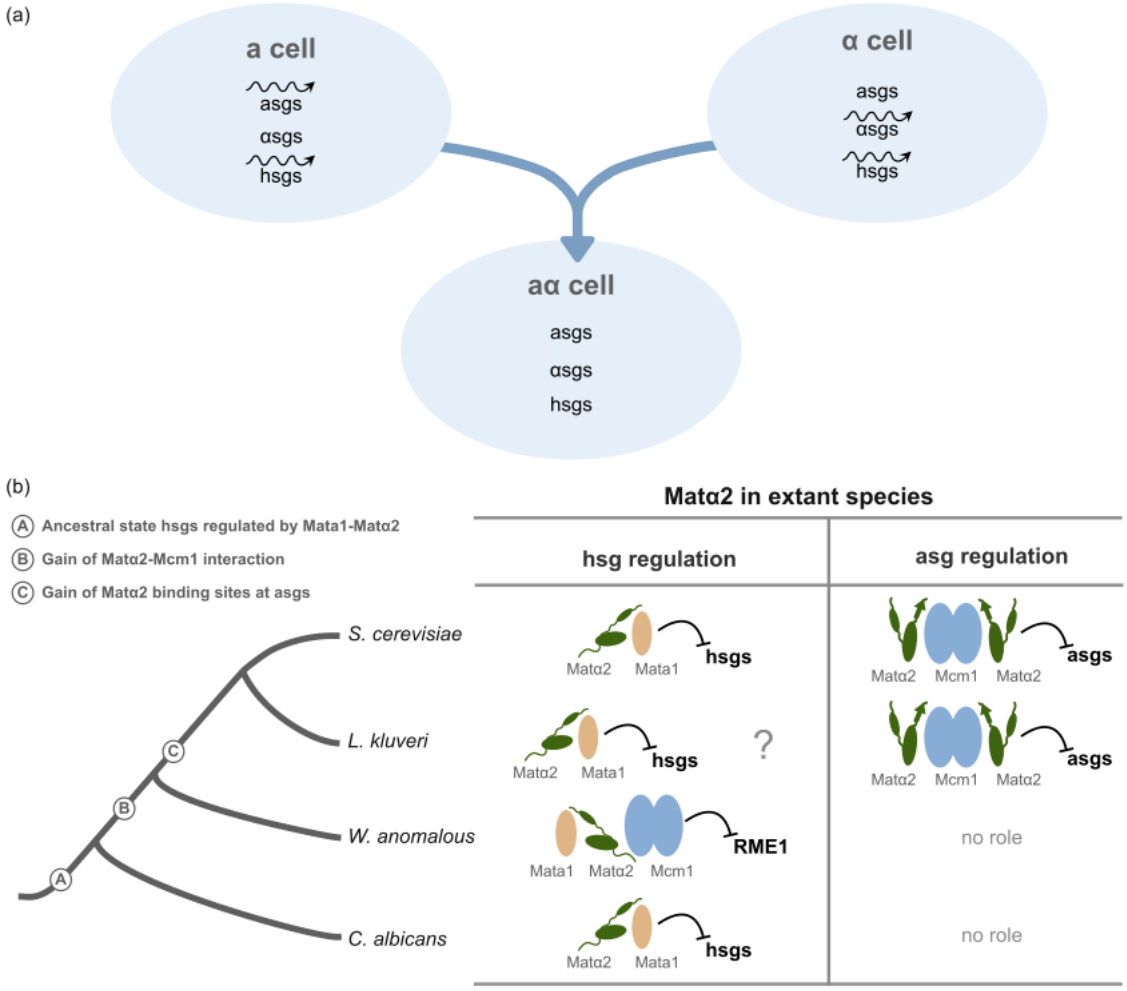
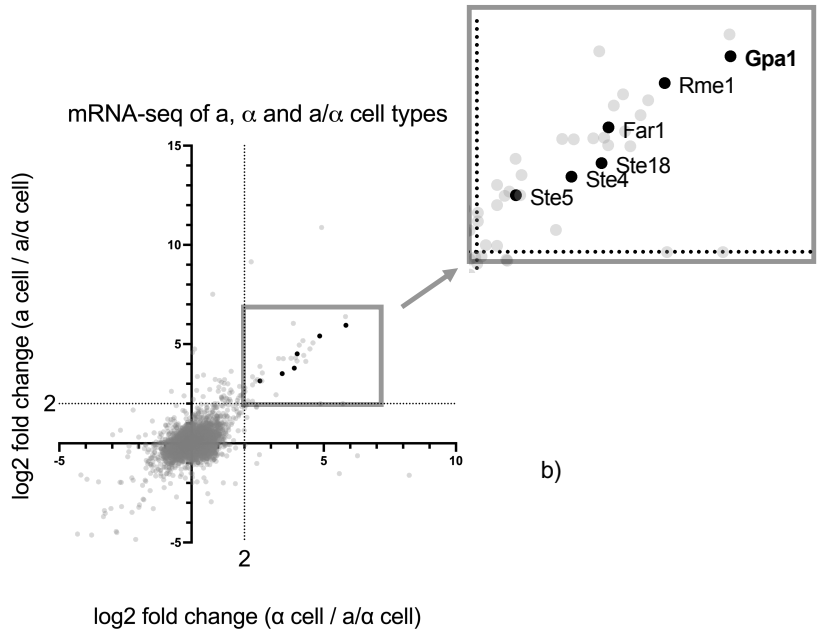



**Figure 1.1: Regulation of Cell Type in Budding Yeast**

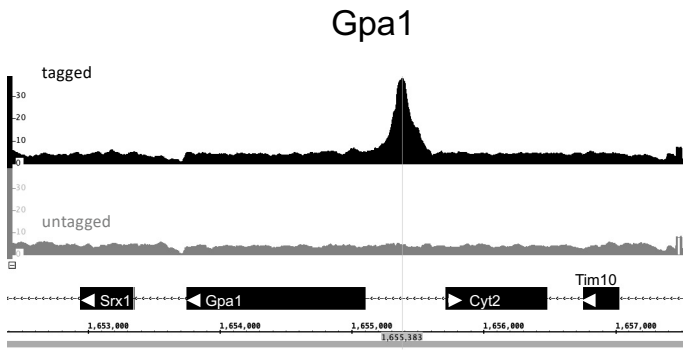
a) Three cell types in budding yeast; **a** and α cell types express the **a** specific and the α specific genes which are unique to them, and both **a** and α cell types express the haploid specific genes. When a and α cells mate, the resulting a/α cell does not express any of these genes..

b) Regulation of **a** specific genes and haploid specific genes by Matα2 and its binding partners Mata1 and Mcm1 in species related to *S. cerevisiae.* Haploid specific genes were directly regulated by the Mata1-Matα2 heterodimer in the ancestor of the species shown(indicated by circled A on the figure). On the branch leading to the extant species *W. anomalus* and *S.cerevisiae,* a protein-protein interaction gained between Matα2 and Mcm1 (see circled B) allowed for the addition of Mcm1 to haploid specific gene regulation in *W. anomalus,* and a gain of Matα2-Mcm1 repression in the a specific genes in the lineage leading to *S. cerevisiae(*see circled C)*. L. kluyveri* is a species on this lineage with Matα2-Mcm1 regulation of a-specific genes, leaving the question of how the haploid specific genes are regulated in this species.

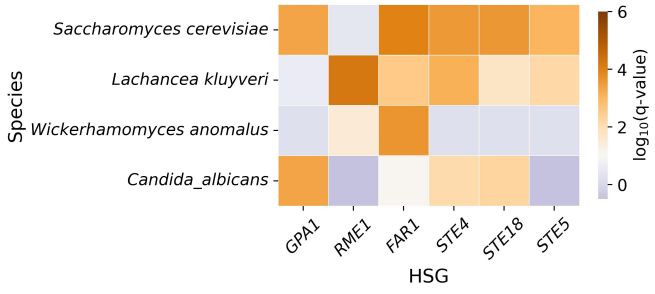
a)


mRNA-seq of a, α and a/α cell types
log2 fold change (a cell / a/α cell)
log2 fold change (α cell / a/α cell)
Gpa1
Rme1
Far1
Ste18
Ste4
Ste5
b)

b)


Gpa1
tagged
untagged
Srx1
Gpa1
Cyt2
Tim10

c)


Saccharomyces cerevisiae
Lachancea kluyveri
Wickerhamomyces anomalus
Candida_albicans
Species
GPA1
RME1
FAR1
STE4
STE18
STE5
HSG
log10(q-value)

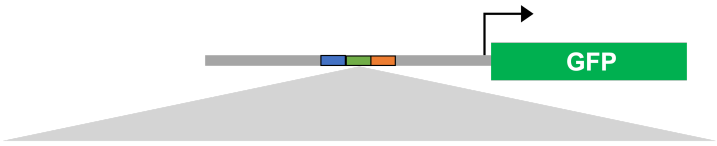
d)


GPA1
TACGTTTATTCCCAAATAGGAAAGTTACATTTTCGCATCAAAGCCTTCATA

**Figure 1.2: Haploid Specific Regulation of GPA1 in *L. kluyveri***

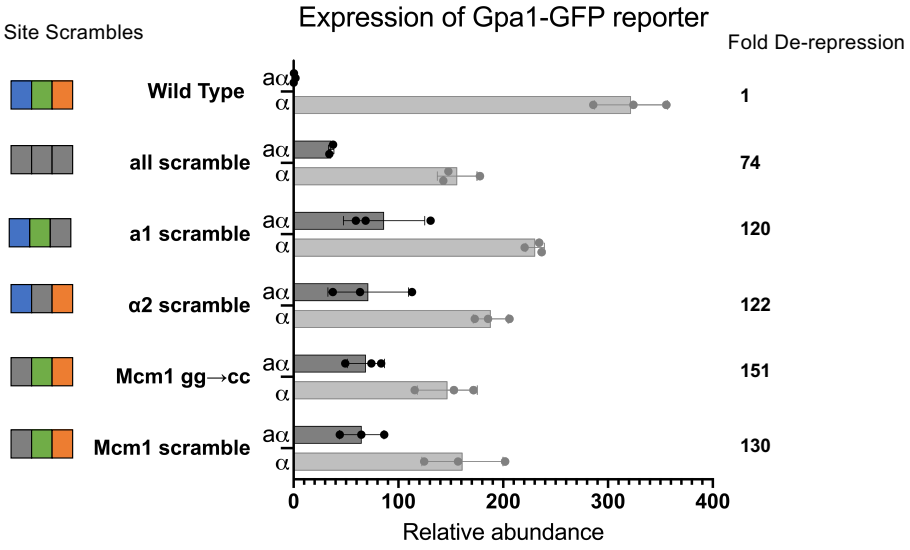
a) RNAseq of of a cell, α cell and a/α cell in *L. kluyveri.* Genes up log2fold or higher in both the a and α relative to the a/α are defined as haploid specific genes in *L. kluyveri.* Inset panel in the top right shows close up view of these upregulated genes. Conserved haploid specific genes *GPA1, RME1, STE4, STE18 AND STE5* are highlighted.
b) Chromatin immunoprecipitation of c-terminal myc tagged Matα2 shows significant enrichment at promoters of *GPA1*, the tagged strain is shown in black compared to the matched untagged strain in grey. Inspection of *GPA1* upstream regulatory region reveals Mcm1 site(blue) ~330bp upstream of orf next to putative α2 site(green) and a1 site(orange).
c) Bioinformatic search for Mata1-Matα2 motif in upstream regions of *GPA1, RME1,STE4, STE18, and FAR1* in *S. cerevisiae, W.anomalus* and *L. kluyveri. L. kluyveri GPA1* upstream region lacks a high scoring Mata1-Matα2 site, while the other five hsgs all have high scoring Mata1-Matα2 sites. Site score is log10 of qvalue.
d) Schematic of Mcm1, Matα2, and Mata1 sites found in *GPA1* promoter region. Mcm1 site indicated in blue, Matα2 site indicated in green, and Mata1 site indicated in orange. Residues that match highly conserved residues for these sites in *S. cerevisiae* are bolded.

a)

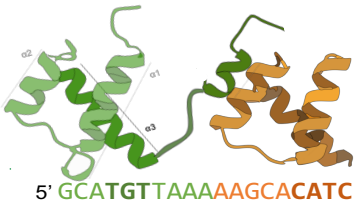



TACGTTTATTCCCAAATAGGAAAGTTACATTTTCGCATCAAAGCCTTCATA **Wild Type**
TACGTTTA--TACGACCTGAGTCCAAAATTTACTATCATGCAGCCTTCATA **All 3 sites scrambled**
TACGTTTATTCCCAAATAGGAAAGTTACATTACTATCATGCAGCCTTCATA **a1 site scrambled**
TACGTTTATTCCCAAATAGGATCCAAAATTTTTCGCATCAAAGCCTTCATA **α2 site scrambled**
TACGTTTA--TACGACCTGAGAAGTTACATTTTCGCATCAAAGCCTTCATA **Mcm1 site scrambled**
TACGTTTATTCCCAAATACCAAAGTTACATTTTCGCATCAAAGCCTTCATA **Mcm1 site gg→cc mutation**

b)




c)

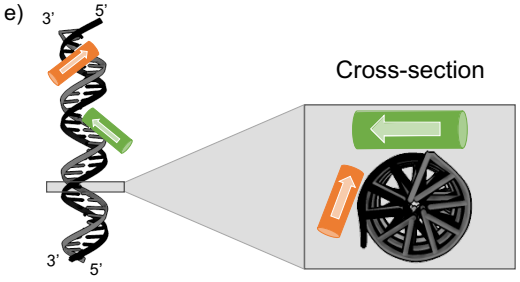


5' GCATGTTAAAAAGCACATC 3'
3' CGTACAATTTTTCGTGTAG 5'

d)

5' TTCCCAAATAGGAAAGTTACATTTTCGCATCAA 3'
3' AAGGGTTTATCCTTTCAATGTAAAAGCGTAGTT 5'

e)




f)

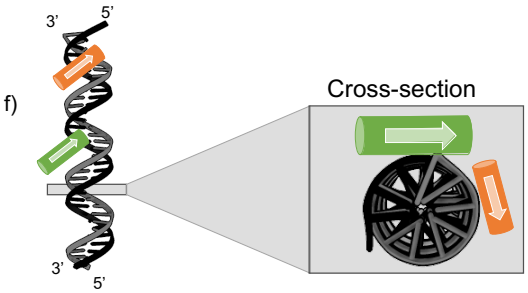


**Figure 1.3: Tripartite Regulation of GPA1**

a) *GFP* reporter constructs with wild type or differentially disrupted Gpa1 promoter sequences: all 3 putative sites scrambled, putative a1 site scrambled, putative α2 site scrambled, putative mcm1 site scrambled, and a gg→cc point mutation in putative Mcm1 site.

b) Expression of gfp transcript in *L. kluyveri* α vs the aα cell containing each construct. Expression is measured by qPCR with probes to *GFP* transcript and normalized to *ACT1*. There were six constructs: 1.wild type sequence, 2. all three of the putative sites scrambled, 3. a1 site scrambled, 4. α2 site scrambled, 5. Mcm1 site scrambled, 6.Mcm1 site with gg→cc point mutation in key residues. Expression of these constructs was then compared between the α and aα cell with three independent genetic isolates for constructs 1 and 3-6, and two independent genetic isolates for construct 2 of the three site scramble. Expression of the reporter transcript was measured by qPCR rather than measure of GFP fluorescence in order to detect the full dynamic range of Gpa1 from full wild type repression in the aα to wild type expression in the α. The experiment was done twice with cells grown on independent days with each of the genetic isolates. The bars correlate to mean expression between isolates, and the standard deviation is shown. Fold derepression is the relative derepression in construct a/α cells compared with wild type a/α cells, scaled to account for a different dynamic range of repression between constructs with different levels of expression in the α cell.

c) Cartoon of Mata1 and Matα2 proteins arranged along Mata1-Matα binding site found in *GPA1* gene in *S.cerevisiae,* arrangement based on structural and biochemical data(Goutte et al.; Li et al.). Highly conserved residues in the binding site are bolded. Matα1 and its binding site are green, and Mata1 and its binding site are orange. Mcm1 and its binding site are blue. The helices of the Matα2 DNA binding domain are labeled α1, α2, and α3, with α3 being the helix that makes key major groove contacts with residues in the DNA binding site..

d) Cartoon of model for interaction of Mcm1-Matα2-Mata1 on *GPA1* tripartite site in *L. kluyveri.* Arrangement is based on existing structural data about Mcm1-Matα2 binding

e) Diagram of Mata1 and Matα2 positioning in the major grooves of DNA of the binding site in the *GPA1* gene in *S.cerevisiae.* For both Mata1 and Matα2, it is the third helix of the DNA binding domain that makes key contacts with DNA in the major groove, for this reason, only the third helix of the DNA binding domains of Mata1 and Matα2 are shown, indicated by orange and green cylinders, respectively. Arrows point from N to C terminus of the helix to indicate relative orientation of the DNA binding domains on DNA. DNA is shown vertically with Matα2 and Mata1, the indicated cross section is taken looking up along the DNA from the end closest to Matα2. This cross section indicates the relative rotational positioning of the Mata1 and Matα2 proteins around the DNA helix. Proteins and DNA are not shown to scale.

f) Diagram of proposed Mata1 and Matα2 positioning in the major grooves of DNA on the tripartite site in *L. kluyveri*. As above, the orange cylinder designates the

third α helix of the DNA binding domain of Mata1 and the green cylinder designates the third α helix of the DNA binding domain of Matα2. Arrows point from N to C terminus of the helix to indicate relative orientation of the DNA binding domains on DNA. DNA is shown vertically, with Mata1 and Matα2 positioned on it. The cross section again is shown looking up at the DNA from the end closest to Matα2, showing the predicted relative rotational positioning of the two proteins around the DNA helix in *L. kluyveri.*

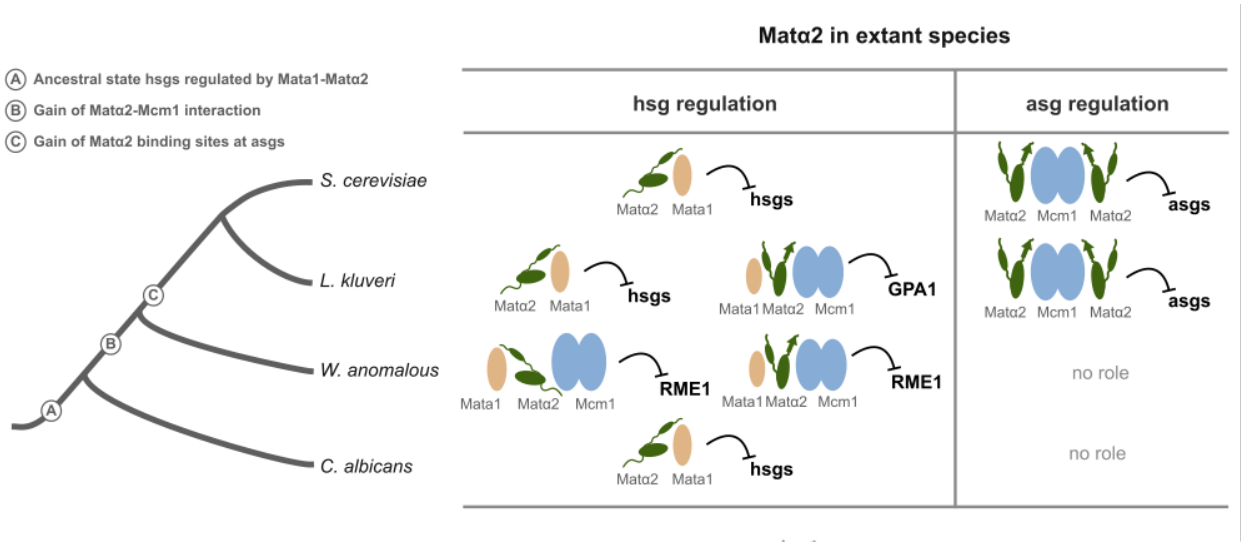



**Figure 1.4: Regulation of cell type by Matα2 and its binding partners**

Regulation of haploid specific genes in *GPA1* in *L. kluyveri* is by tripartite Mata1-Matα2-Mcm1. This requires the ancestral Matα2-Mcm1 interaction(B) . The rest of the conserved haploid specific genes are regulated by Mata1-Matα2 as in *S. cerevisiae.* Though it is for a different gene, this requirement for all three proteins for the repression of *GPA1* in *L. kluyveri* resembles the regulation of *RME1* in *W. anomalus.*

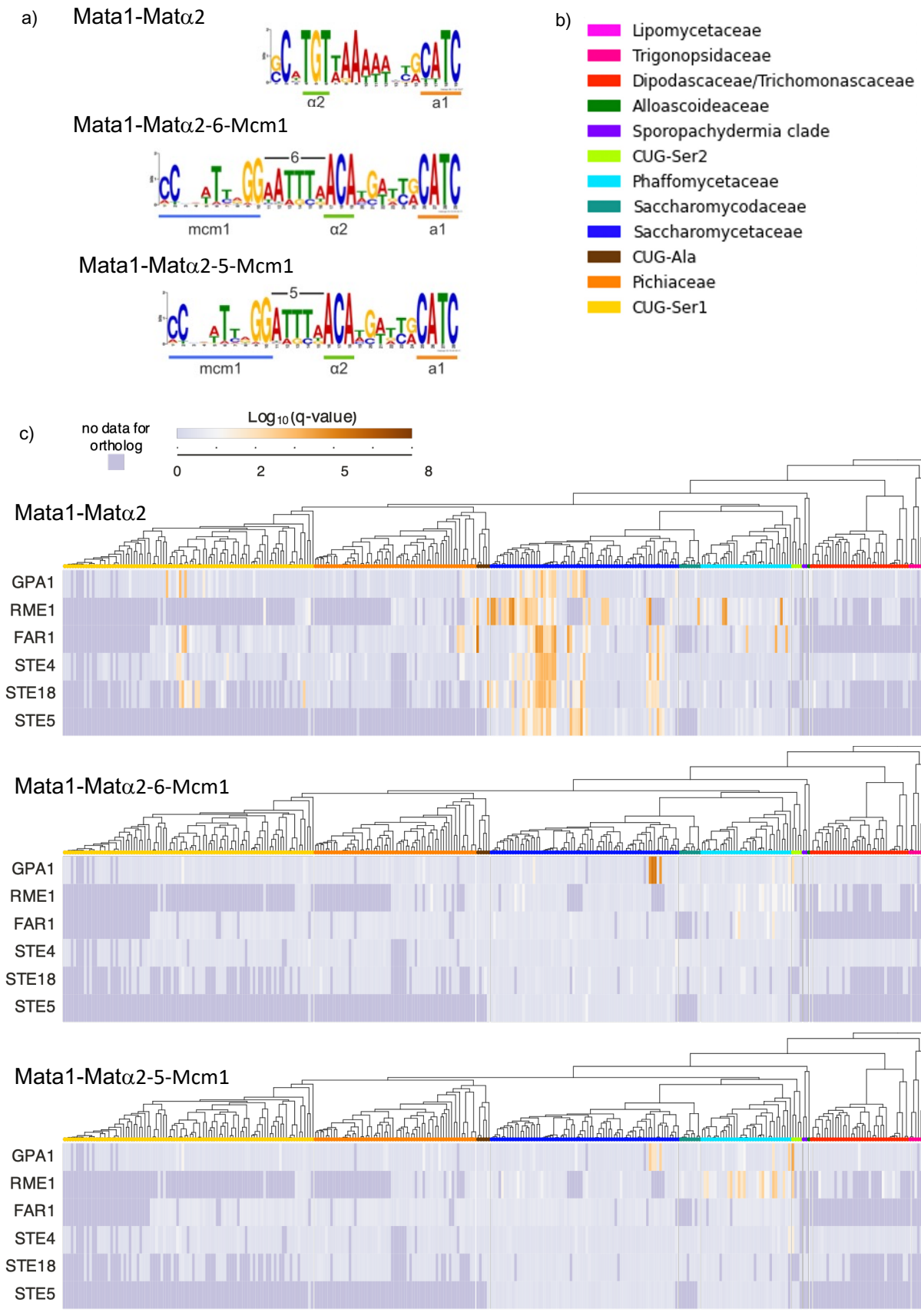
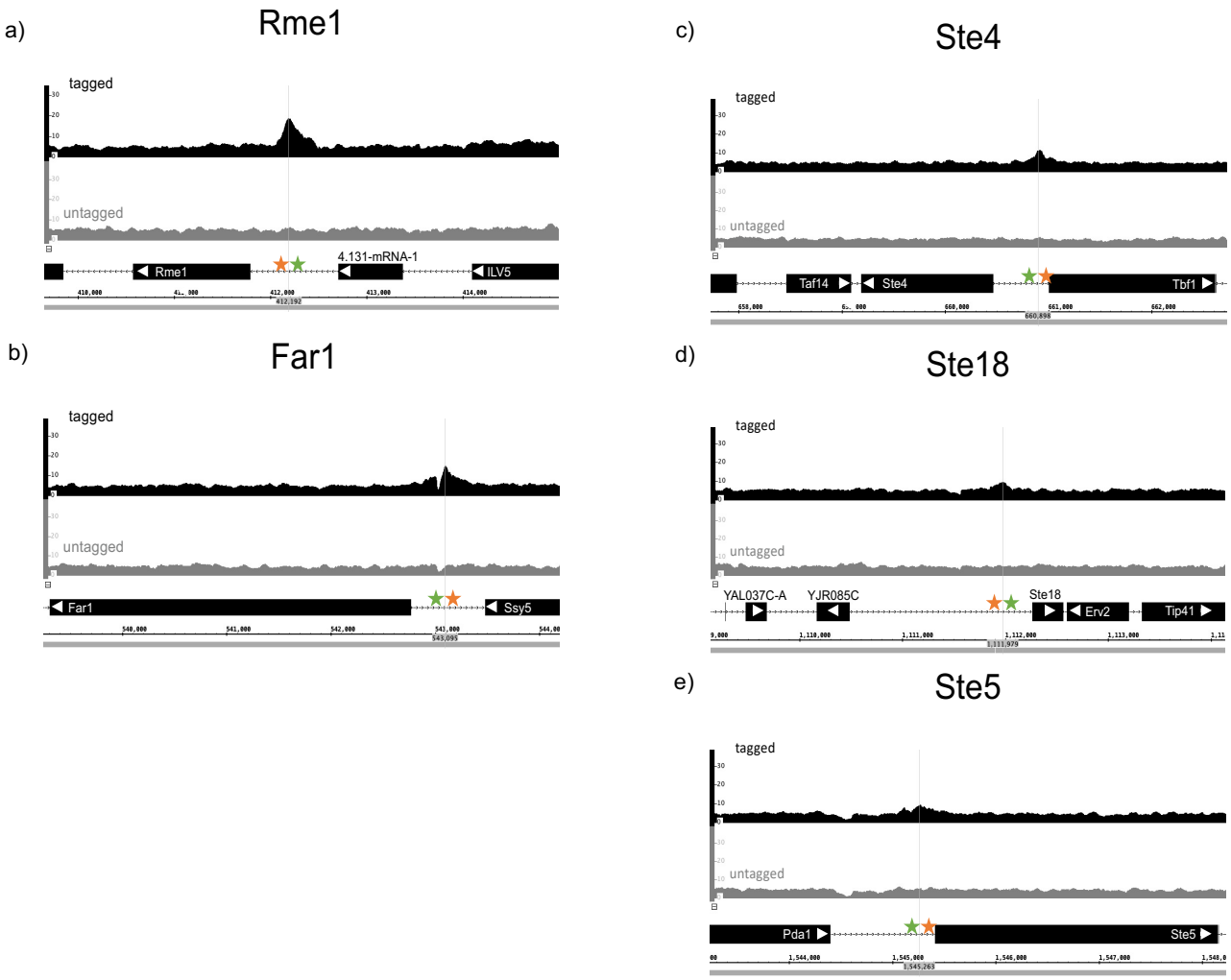

a)
Mata1-Matα2
α2
a1
Mata1-Matα2-6-Mcm1
mcm1
α2
a1
Mata1-Matα2-5-Mcm1
mcm1
α2
a1
b)
Lipomycetaceae
Trigonopsidaceae
Dipodascaceae/Trichomonascaceae
Alloascoideaceae
Sporopachydermia clade
CUG-Ser2
Phaffomycetaceae
Saccharomycodaceae
Saccharomycetaceae
CUG-Ala
Pichiaceae
CUG-Ser1
c)
no data for ortholog
Log10(q-value)
0
2
5
8
Mata1-Matα2
GPA1
RME1
FAR1
STE4
STE18
STE5
Mata1-Matα2-6-Mcm1
GPA1
RME1
FAR1
STE4
STE18
STE5
Mata1-Matα2-5-Mcm1
GPA1
RME1
FAR1
STE4
STE18
STE5

**Figure 1.5: Bioinformatic search in haploid specific genes across species**

a) Motifs used in the cross species bioinformatic search of haploid specific genes. From the top, these are the motif for Mata1-Mat$\alpha$2 in *S. cerevisiae,* a motif generated to reflect the tripartite site in *L.kluyveri*, and a modified tripartite motif based on a putative tripartite site seen in *W. anomalus RME1* promoter. The difference between these tripartite motifs is that the former has longer spacing between Mcm1 and Mat$\alpha$2 (6bp), while the latter has the shorter spacing between Mcm1 and Mat$\alpha$2. For this reason, the sites are referred to as Mata1-Mat$\alpha$2-6-Mcm1, and Mata1-Mat$\alpha$2-5-Mcm1. The Mcm1 portion of the motif is underlined in blue, the Mat$\alpha$2 portion of the motif is underlined in green, and the Mata1 portion of the motif is underlined in orange.

b) Legend for colors used to indicate different clades across the yeast tree

c) Motifs were used to search upstream regions of orthologs of the haploid specific genes across yeast species. The best possible match to the site is given a qvalue, a color is assigned based on that qvalue, with pale lavender indicating low significance, and dark orange-brown indicating high significance. From top to bottom of each panel are gene names, each row shows all of the scores for the orthologs of that gene across the species. A phylogenetic tree indicates the relatedness of the various species used for this study. Clades are indicated by color, from left to right: yellow for CUG-Ser1, orange for the Pichiaceae, dark brown for CUG-Ala, deep blue for the Saccharomycetaceae, teal for the Saccharomycodaceae, bright aqua blue for the Phaffomycetaceae, bright pale green for CUG-Ser2, deep violet for Sporopachydermia, dark green for the Alloascoideaceae, red for the Dipodascaceae/Trichomonascaceae, deep pink for the Trigonopsidaceae, and magenta for the Lipomycetaceae.




**Supplementary Figure 1.1: Chromatin Immunoprecipitation of Mat$\alpha$2**
Chromatin immunoprecipitation of c-terminal myc tagged $\alpha$2 shows significant enrichment at promoters of genes regulated by a1$\alpha$2, the tagged strain is shown in black compared to the matched untagged strain in grey. Peaks centered either over a1$\alpha$2 motif or an a1$\alpha$2-Mcm1 motif. The motifs are indicated with stars, green for $\alpha$2, orange for a1. Y axis is normalized counts. Upstream regions of genes are as follows: a) RME1, b)FAR1 c)STE4, d)STE18, e)STE5

a) **Mata1-Matα2 sites** in haploid specific genes in *S. cerevisiae* and *L.kluyveri*

GCA**TGT**TAAAAAGCA**CATC** *S.Cerevisiae Gpa1*
CCA**TGT**TAAAAATCA**CATC**AA *L.Kluyveri Rme1*

b) **Tripartite site** in *GPA1 in L. kluyveri* aligned to **Matα2-Mcm1 site** in *a specific genes*

TT**CC**CAAATA**GG**AAAGTT**ACA**TTTTCG**CATC**AA *L.Kluyveri Gpa1*
GCA**TGT**TAAAAAGCA**CATC** *S.Cerevisiae Gpa1*

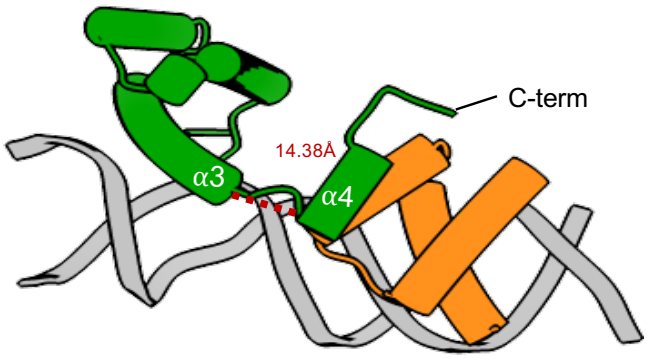
c) **Tripartite site** in *GPA1* in *L. kluyveri* aligned to **Mata1-Matα2 site**

TT**CC**CAAATA**GG**AAAGTT**ACA**TTTTCG**CATC**AA *L.Kluyveri Gpa1*
CA**TGT**ACTTA**CC**CAATTA**GG**AAATTT**ACA**TG *S.Cerevisiae Ste2*

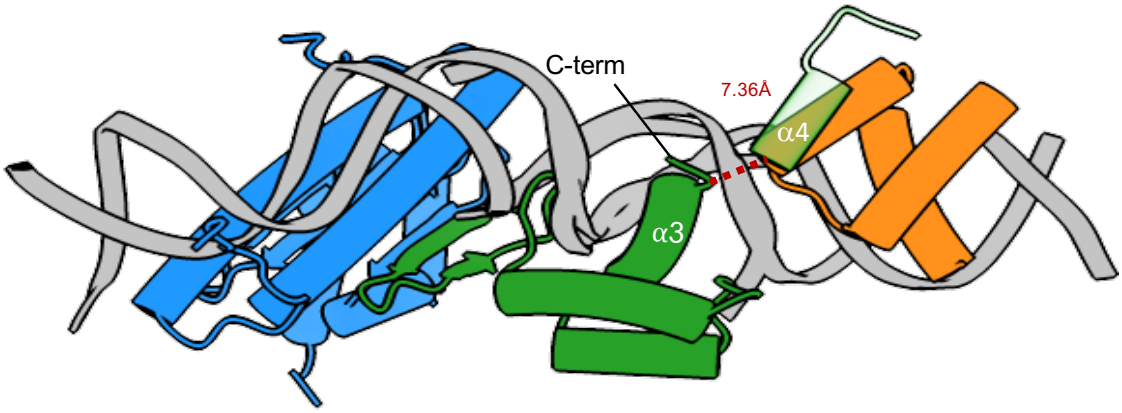
**Supplementary Figure 1.2: Alignments of *L. kluyveri* tripartite site**
Motifs are color coded as follows, orange for the Mata1 cis site, green for the Matα2 cis site, and blue for the Mcm1 cis site. Highly conserved residues within motifs are bolded.

a) Aside from *GPA1,* Mata1- Matα2 sites found in haploid specific genes in *L. kluyveri* align closely to Mata1- Matα2 sites in *S. cerevisiae* haploid specific genes. The representative Mata-Matα2 sites from *S. cerevisiae GPA1* and *L. kluyveri RME1* are shown aligned as an example of this.
b) The tripartite site found in *L. kluyveri GPA1* aligned to the Mata1-Matα2 site found in haploid specific genes in *S. cerevisiae.* In the tripartite site, the Mata1 and Matα2 sites are reversed relative to each other, and also have a 3 base pair difference in the spacing between them.
c) The tripartite site found in *L. kluyveri GPA1* aligned to the Matα2-Mcm1 site found in the a specific genes. The relative spacing between Matα2 and Mcm1 is the same between the two sites.

a)




b)

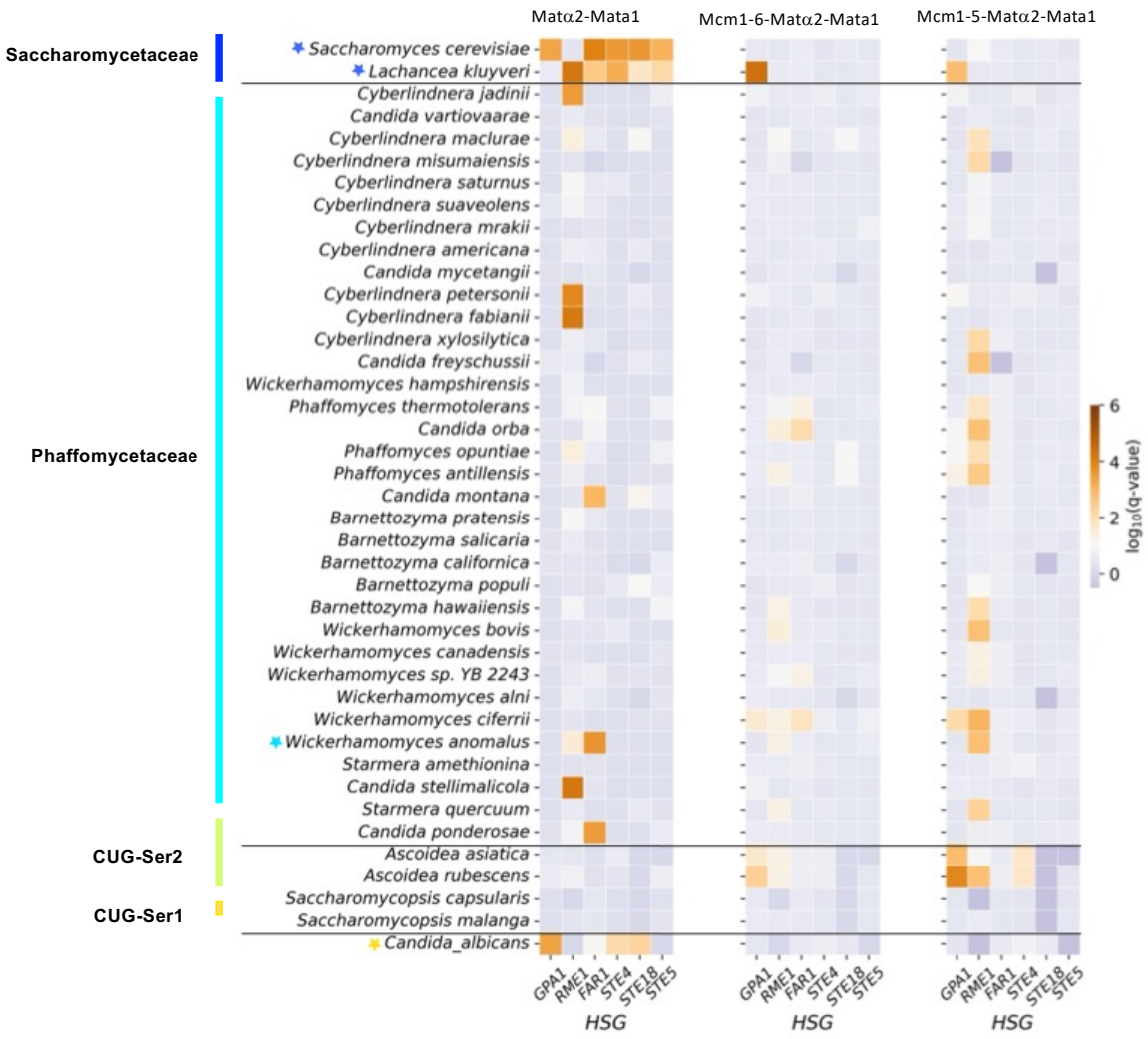



**Supplementary Figure 1.3: Modeling Mata1, Matα2, and Mcm1 binding in *L. kluyveri***

a) Structure of Mata1-Matα2 bound to DNA in *S. cerevisiae.* The third helix of the DNA binding domain of Matα2 is labeled α3 and the c-terminal helix that forms upon interaction with Mata1 is labeled α4. The c-terminal end of Matα2 is labeled. Mata1 is in orange, and Matα2 is in green. The distance between α3 and the structured part of the Mata1 interaction domain of Matα2 is measured as approximately 14.36 Å. Measurement line is in red, with the measured number also in red. This distance is spanned by a flexible linker region of four amino acids.

b) Superimposed crystal structure of Mcm1-Matα2 bound to DNA in *S.cerevisiae* and Mata1-Matα2 bound to DNA in *S. cerevisiae*. These superimposed structures illustrate the model for how Mata1, Matα2, and Mcm1 are predicted to

bind to DNA based on the orientation and spacing of their sites within the tripartite site. Mata1 is in orange, Mat$\alpha$2 is in green, and Mcm1 is in blue. The Mat$\alpha$2 included in this model is from the crystal structure of Mat$\alpha$2-Mcm1, and so does not include the c-terminal helix $\alpha$4. The third helix of the DNA binding domain is labeled $\alpha$3, the c-terminal end of this helix is labeled to orient the viewer. The $\alpha$4 helix from the Mata1- Mat$\alpha$2 structure(transparent green) is pictured interacting with Mata1 as it does in the Mata1-Mat$\alpha$2 structure. In order for Mat$\alpha$2 to interact with Mata1 in this model, the flexible linker region of 4 amino acids on Mat$\alpha$2 must span the space between the $\alpha$3 and $\alpha$4 helices of the protein, a distance measured here to be approximately 7.36 Å. The linker measures ~15-16Å when extended. The Mat$\alpha$2 linker can therefore plausibly form a small loop in order to accommodate the interaction with Mata1 in this orientation. Measurement indicated in red.




**Supplementary Figure 1.4: Results of tripartite search in the Phaffomycetae**
Subset from results of cross species search of the haploid specific genes *GPA1, STE4, STE18, FAR1, RME1,* and *STE5.* From left to right, the three panels show data from searches with the Mata1-Matα2 site, the tripartite site with longer spacing (6bp between Mcm1 and Matα2 sites), and the tripartite site with shorter spacing (5bp between Mcm1 and Matα2). This compares *S. cerevisae* and *L. kluyveri* to the species of the Phaffomycetae clade and the more diverged CUG-Ser2 clade, with *C.albicans* included as the outgroup species. This subset highlights the signal found in the Phaffomycetae clade for the tripartatite site with the shorter spacing(5 bases) in *RME1* across the clade. Also noticeable is the signal in *GPA1* in *A. asiatica* and both in *GPA1 and RME1* in *A. rubescens* in the diverged CUG-Ser2 clade.

**Table 1.1: RNAseq of of a cell, α cell and a/α cell in *L. kluyveri***

Genes differentially expressed with a fold change of 2 log2fold or more in both *L. kluyveri* haploid cell types when compared to the **a**/α cell type**.**

| **Gene name** | α cell/ **a**/α cell log2 fold change | **a** cell / **a**/α cell log2 fold change |
|---|---|---|
| *snap_masked-SAKL0C-processed-gene-5.65* | 5.82 | 6.38 |
| *GPA1* | 5.84 | 5.95 |
| *AGA2* | 4.92 | 10.88 |
| *Rme1* | 4.84 | 5.41 |
| *snap_masked-SAKL0C-processed-gene-2.12* | 4.59 | 5.06 |
| *augustus_masked-SAKL0B-processed-gene-5.93* | 4.48 | 4.76 |
| *augustus_masked-SAKL0E-processed-gene-9.41* | 4.21 | 5.18 |
| *snap_masked-SAKL0H-processed-gene-0.18* | 4.07 | 4.95 |
| *FUS3* | 4.24 | 4.43 |
| *augustus_masked-SAKL0B-processed-gene-4.131* | 4.32 | 4.13 |
| *FAR1* | 3.99 | 4.51 |
| *SST2* | 3.93 | 4.30 |
| *SUC2* | 3.99 | 4.15 |
| *ICS2* | 3.76 | 4.29 |
| *HSP12* | 3.85 | 6.04 |
| *STE18* | 3.89 | 3.79 |
| *FUI1* | 3.47 | 4.27 |
| *STE4* | 3.43 | 3.51 |
| *CTT1* | 3.29 | 4.27 |
| *ICL2* | 2.68 | 3.55 |
| *RTC3* | 2.59 | 3.88 |
| *snap_masked-SAKL0B-processed-gene-3.19* | 3.20 | 2.45 |
| *TMT1* | 2.66 | 3.14 |
| *STE5* | 2.59 | 3.14 |
| *UGA4* | 2.49 | 3.22 |
| *augustus_masked-SAKL0C-processed-gene-3.16* | 2.42 | 3.13 |
| *snap_masked-SAKL0E-processed-gene-0.50* | 2.31 | 2.94 |
| *snap_masked-SAKL0D-processed-gene-11.161* | 2.31 | 3.35 |
| *STE2* | 2.26 | 9.15 |
| *augustus_masked-SAKL0E-processed-gene-11.234* | 2.31 | 2.12 |
| *RCE1* | 2.14 | 2.13 |

**Table 1. 2: Peaks in Chromatin Immunoprecipitation of Matα2**

These were peaks identified with Macs2 in the replicate with stronger signal. The 7 peaks upstream of orfs with haploid specific gene expression are considered high confidence and are bolded. Note: *PRM1* had expression just under the 2 log2 fold cut off for haploid specific expression and so is counted here as haploid specific.

| orf_name | chromosome | peak_start coordinate | peak_end coordinate | qvalue | up in haploid/ **a**/α | a1α2 motif |
|---|---|---|---|---|---|---|
| MFALPHA | SAKL0A | 472742 | 472743 | 298.052 | No, α specific | yes |
| STE6 | SAKL0A | 894006 | 894007 | 89.1339 | No, a specific | no |
| **RME1** | SAKL0B | 412207 | 412208 | 240.606 | yes | yes |
| **PRM1** | SAKL0C | 194347 | 194348 | 209.789 | yes | yes |
| **FAR1** | SAKL0C | 543128 | 543129 | 106.208 | yes | yes |
| FAT1 | SAKL0D | 338385 | 338386 | 57.0073 | no | yes |
| SIM1 | SAKL0E | 724510 | 724511 | 49.3476 | no | yes |
| **STE18** | SAKL0F | 1111981 | 1111982 | 25.7163 | yes | yes |
| BRE2 | SAKL0G | 1013530 | 1013531 | 136.289 | no | yes |
| **GPA1** | SAKL0G | 1655385 | 1655386 | 803.298 | yes | no |
| **STE4** | SAKL0H | 660912 | 660913 | 51.5509 | yes | yes |
| **STE5** | SAKL0H | 1545260 | 1545261 | 29.7386 | yes | yes |
| IME4 | SAKL0H | 2181402 | 2181403 | 21.3828 | no | yes |

**Table 1.3: Strains used in this study**

| Strain | Species | Cell Type | Genotype | Source |
|---|---|---|---|---|
| yLB96 | Lachancea Kluyveri | α cell | MATα2-c-term 13x Myc tag, KanMX6 | Lauren Booth |
| yLB76 | Lachancea Kluyveri | **a** cell | Prototroph | Herskowitz Lab |
| yLB77 | Lachancea Kluyveri | α cell | Prototroph | Herskowitz Lab |
| FDFy18a | Lachancea Kluyveri | **a**/α cell | MATα2-c-term 13x Myc tag, KanMX6 | Francesca Del Frate |
| FDFy22a, b, c | Lachancea Kluyveri | **a**/α cell | prototroph | Francesca Del Frate |
| FDFy27a,b,c | Lachancea Kluyveri | α cell | Ura3::a1 site scramble-caGFP, Hyg | Francesca Del Frate |
| FDFy28a,b,c | Lachancea Kluyveri | α cell | Ura3::all site scramble-caGFP, Hyg | Francesca Del Frate |
| FDFy30a,b,c | Lachancea Kluyveri | α cell | Ura3::α2 site scramble-caGFP, Hyg | Francesca Del Frate |
| FDFy31a,b,c | Lachancea Kluyveri | α cell | Ura3::Mcm1 site scramble-caGFP, Hyg | Francesca Del Frate |
| FDFy32a,b,c | Lachancea Kluyveri | α cell | Ura3::Mcm1 gg-->cc site mutant-caGFP, Hyg | Francesca Del Frate |
| FDFy33a,b,c | Lachancea Kluyveri | α cell | Ura3::WT-caGFP, hyg | Francesca Del Frate |
| FDFy34a,b,c | Lachancea Kluyveri | **a**α cell | Ura3::a1 site scramble-caGFP, Hyg/ WT | Francesca Del Frate |
| FDFy35a,b,c | Lachancea Kluyveri | **a**α cell | Ura3::all site scramble-caGFP, Hyg/ WT | Francesca Del Frate |
| FDFy36a,b,c | Lachancea Kluyveri | **a**α cell | Ura3::α2 site scramble-caGFP, Hyg/ WT | Francesca Del Frate |
| FDFy37a,b,c | Lachancea Kluyveri | **a**α cell | Ura3::Mcm1 site scramble-caGFP, Hyg/WT | Francesca Del Frate |
| FDFy38a,b,c | Lachancea Kluyveri | **a**α cell | Ura3::Mcm1 gg-->cc site mutant-caGFP, Hyg/ WT | Francesca Del Frate |
| FDFy39a,b,c | Lachancea Kluyveri | **a**α cell | Ura3::WT-caGFP, hyg/ WT | Francesca Del Frate |

# Chapter 2:

Regulation of haploid specific genes in *Wickerhammomyces anomalus* requires the transcriptional regulator Rme1

## Introduction

As discussed in the previous chapter, mechanisms of gene regulation change frequently over time. These changes can result in different regulatory architectures, for example, simple direct regulation can shift to indirect regulation, with the addition of an intermediary regulator. This intercalation of a new regulator into an existing regulatiory circuit regulatory complexity, and has the potential to introduce phenotypic change[41]. One example of such an intercalation has been described in the regulation of cell type in yeast[14]. As described in the previous chapter, the ancestral regulation of haploid specific genes in the **a**/$\alpha$ cell is repression by direct binding of Mata1-Mat$\alpha$2. All of the haploid specific genes in this regulatory scheme are directly bound by Mata1-Mat$\alpha$2 in the **a**/$\alpha$ cell. *S. cerevisiae* and *C. albicans* have this ancestral form of regulation[12,13]. However, in the species *L. kluyveri,* Rme1, another transcriptional regulator, is intercalated into the regulation of the haploid specific genes[14]. In the haploid cell types, Rme1 is expressed and activates the haploid specific genes. In the **a**/$\alpha$ cell, *RME1* is directly repressed by Mata1-Mat$\alpha$2, and without Rme1 to activate them, the haploid specific genes are not transcribed. Another example is *W. anomalous,* where previous work shows that repression of the haploid specific genes requires binding of Mata1-Mat$\alpha$2 and a third protein, Mcm1, for repression[9]. Previous work conclusively supports direct repression of *RME1* by Mata1-Mat$\alpha$2 in the **a**/$\alpha$, but does not support direct Mata1- Mat$\alpha$2 repression of the other haploid specific genes. This evidence suggests a form of indirect regulation similar to that in *K. Lactis,* with Rme1 directly activating the haploid specific genes[14].

In this work, we show that Rme1 is required for the activation of two of the four highly conserved haploid specific genes, and is required for mating. It is intercalated into the regulation of the haploid specific genes in *W. anomalous.* This is an example of a change in the molecular mechanism of regulation of haploid specific genes. Specifically, this is an intercalation of a regulator into an existing regulatory circuit. Here we see another example of how a transcription factor ancestrally regulated by Mata1-Mat$\alpha$2 was able to take on a new role activating the haploid specific genes with the gain of a few cis regulatory sites.

## Results

In order to test the role of Rme1 in haploid specific gene regulation in *W. anomalus,* we knocked out *RME1* in the a cell type. We then tested the effect that knocking out *RME1* had on mating and on transcription of haploid specific genes in the a cell. If haploid specific genes require Rme1 for activation, then *RME*1 would also be required for mating, as many of the genes required for mating are haploid specific genes. We tested the mating competence of the RME1 deleted a cell compared to that of a wild type a cell(Figure 2.1a). No successful mating was observed in the RME1 knock out compared to a 57% mating efficiency with the wild type. While it is possible that some low level mating can occur in the RME1 knock out, there is clearly a severe mating defect in the absence of Rme1.

Transcriptional profiling of the Rme1 knock out a cell showed decreased expression of *GPA1* and *STE4* transcripts(Figure 2.1b). These were the only genes significantly down—log2fold or more—across all replicates. *MUP1* was upregulated in the knock out compared to the wild type. *RME1* canonically functions as an activator, but it is possible this effect is indirect, or that *RME1* is repressing transcription by another mechanism[42]. While it is known that the *MUP1* ortholog in *S. cerevisiae* is regulated by *STE12,* a transcriptional regulator associated with cell type regulation, it is not clear what role *MUP1* plays in *W. anomalus*[43].

To take an alternative approach, we turned to de novo motif finding to find cis regulatory sites in known haploid specific genes in *W. anomalus*(Figure 2.1c). We extracted 600 bp upstream of the transcription start site in the four core conserved haploid specific genes (*GPA1, STE4, STE18, FAR1*) of *W. anomalus.* With these

sequences, the Meme program generated a de novo motif resembling that of the *RME1* motif found in the 4 core hsgs in *K. lactis.* This further supports that Rme1 directly activates *GPA1* and *STE4* in the haploid, and indicates that perhaps Rme1 also activates *STE18* and *FAR1* by direct binding.

## Discussion

We showed that the transcription factor Rme1 is involved in the regulation of haploid specific genes in *W. anomalus.* It is required for full mating efficiency and also activates at least two highly conserved haploid specific genes, *GPA1* and *STE4*, which code for the alpha and beta subunits of the heterotrimeric G protein that responds to pheromone(Figure 2.1a,b). Ste4 dimerization with Ste18 is required for initiation of the mating cascade in response to pheromone, so Rme1-dependent Ste4 activation in the haploid is sufficient to explain the mating defect in the *RME1* knock out[10].

Bioinformatic analysis shows Rme1 cis regulatory motifs in the upstream control regions of *GPA1* and *STE4*, indicating that this regulation is most likely mediated by direct binding. Cis regulatory sites were also found in the control regions of in *Ste18* and *Far1*, indicating possible direct Rme1 binding, though transcriptional profiling did not provide strong evidence for Rme1 activation of these genes. However, this could be due to low basal expression of these genes in the haploid, which makes it more difficult to detect a decrease in expression in the absence of Rme1. Follow-up experiments with pretreatment of cells with $\alpha$ pheromone may increase signal and resolve this question. Alternatively, a more sensitive method like qPCR, or full mRNA seq (rather than 3' seq) could help with the issue of limited basal transcription. The bioinformatics findings also come with the caveat that the RME1 cis regulatory motif is relatively information poor, and so while appearance of the sites in these promoters supports a model of direct binding by RME1, further experiments would be needed to prove it. Another caveat is that without the complementary knock out in the $\alpha$ cell, we cannot entirely rule out that

the role of Rme1 may be different in the α cell, rather than general to both of the haploid cell types.

However, regardless of these remaining questions, it does appear that there is some similarity to the regulation of haploid specific genes in *K. lactis,* with direct repression of Rme1 in the a/α[14]. Rme1, in turn, acts as the direct activator of at least two highly conserved haploid specific genes in the haploid a cell. In the a/α, where Rme1 is repressed by the Mcm1-Mata1-Matα2 complex, these genes cannot be activated by Rme1, and are not transcribed.

Direct Mata1-Matα2 repression of haploid specific genes is a widespread regulatory scheme throughout yeast, so most likely the intercalation of Rme1 into regulation of haploid specific genes is derived. However, it is less clear when this intercalation occurred. Two species, *K. lactis* and *W. anomalus,* have direct evidence of Rme1 regulation of haploid specific genes, but many extant species sharing a common ancestor, such as *S. cerevisiae*, do not[9,14]. It seems more likely that the direct regulation of haploid specific genes by Mata1-Matα2 is widespread through this lineage, and that Rme1 regulation was gained more than once. If so, it is interesting to consider the role this transcription factor plays in the regulation of cell type in yeast. Rme1 has broadly conserved regulation as a haploid specific gene across this part of the yeast lineage. Its role in meiotic regulation in *S. cerevisiae* suggests that perhaps the ancestral regulation of this transcription factor sets a regulatory background where simple gain and loss of cis sites can allow it to acquire new roles in this transcriptional circuit[10]. This work provides more context to the question of how transcription factors intercalate into existing regulatory circuits, in this case, the conserved haploid specific gene regulation

of *RME1* meant the transcription factor was already part of the circuit, from that point, it would only require the gain of an Rme1 cis regulatory in order for Rme1 to intercalate into its regulation. Subsequently, the Mata1-Mat$\alpha$2-Mcm1 site could be lost, leaving the gene entirely dependent on Rme1 for haploid specific gene regulation. This is similar to what we hypothesize happened in *K. lactis* resulting in indirect repression of haploid specific genes via direct repression of the activator, Rme1, with the main difference that Rme1 in *K. lactis* is directly repressed by the Mata1-Mat$\alpha$2 heterodimer rather than the tripartite Mata1-Mat$\alpha$2-Mcm1. It is possible to find what looks like Mata1-Mat$\alpha$2 sites in some of the haploid specific genes, notably *GPA1* and *FAR1,* but the lack of evidence of direct binding of Mata1-Mat$\alpha$2 in these genes indicates that these sites may not be functional.

This work adds another example of Rme1 mediated Mata1-Mat$\alpha$2 regulation of the haploid specific genes and adds evidence to the idea that existing regulatory architecture can make some regulatory changes relatively easy. Rme1, as a haploid specific gene, was able to acquire the role of activating other haploid specific genes with a few cis regulatory changes, and thus added another level of complexity to haploid specific gene regulation in this species. It is known that as regulatory circuits change, transcription factors can gain and lose roles regulating groups of genes. This example, when combined with previous work in *K. lactis,* would suggest that this process may happen even more easily for transcription factors within a regulatory circuit, gaining and losing target genes within that circuit, while remaining a target for a higher level regulator in that circuit.

## Methods

*Strain Construction*

The Rme1 knock out vector(FDp7) was made by adding homology to the *RME1* locus to the CSB p121 knock out vector which was adapted for gene knock outs in *W.anomalus* from pCS.ΔLig4, a vector with an optimized Nourseothricin(NAT) marker developed for knock outs in the related species *W. Ciferri*[44].

To increase the likelihood of specific integration at the *RME1* locus rather than random integration by non-homologous end joining, long(1kb) homology arms were used. 1kb of homology was added upstream and downstream of the NAT resistance cassette. Upstream and downstream homology arms were PCR amplified from *W. anomalous* gDNA. Downstream homology was added by restriction digest and ligation with the Fast-Link DNA Ligation Kit (Lucigen MBTOOL-010) and transformed into Stellar Competent Cells(Takara 636763). Upstream homology arm was added with In-Fusion® Snap Assembly Master Mix(Takara 638947).

Linear fragment of transforming DNA was cut out of FDp7 by digest with Pme1. Cells were transformed with ~1ug of linearized DNA with a freeze thaw protocol designed for use in a closely related species *W.ciferri* that was found to also work well in *W.anomalus*. with one modification[9,44]. Cultures were allowed to recover for (some amount of time) at 30°C instead of plating directly to YPED and plated directly to selective media, YPED + 300ug/mL NAT. Colonies were patched to a new drug plate, and screened for lack of the *RME1* orf or against wild type control cells. These candidates were then screened again for both upstream and downstream flanks of the insertion site at the *RME1* locus.

*Mating Assay*

Strains used were FDy25(Rme1 ko, Nat resistant a cell) , FDy26( a cell with randomly integrated Nat marker) and CSB 342(WT α cell). Fresh cells were picked to water from colonies grown on a YPED plate at 30°C, a cells(FDy25 or FD26) were mixed with α cells(CSB 342) at a 200-fold excess of α cells over a cells. This mixture was then plated to YPED at a density of 100-150 cells per plate. Some of the mixture was also resuspended in YPED and after 3 hours, cells were imaged at 40x on a phase contrast microscope to visualize zygote formation.

Plated cells were allowed to grow on YPED plates at 30°C overnight. The next day, cells were replica plated to YPED plates with 300ug/ml Nourseothricin. This killed all α cells, leaving only a or a/α cells. To distinguish between a cells and a/α cells, cell type was checked with colony PCR, with primers against the Mata and Matα locus. Mating efficiency was calculated based on the fraction of Nat resistant colonies found to be a/α by PCR. For ease of comparison, this number was also converted to percent mating with Wild type a cell x Wild type α cell, and percent mating with the *RME1 ko* a cell x Wild type α cell.

*3' end sequencing*

Four replicates of FDy19 (a cell) , CSB 342(α cell), FDy24(a/α cell), and FDy25(RME1 ko a cell) were picked from individual single colonies grown from the same streak. Cultures were inoculated from single colonies and grown overnight in

YPED at 30°C, diluted back to an OD600 of 0.15 in the morning and harvested at an OD600 of 0.6-0.9 as is described in Nocedal et. al. [24]. RNA was extracted using the RiboPure RNA purification kit (ThermoFisher AM1924). RNA quality was verified on the Agilent Tapestation. cDNA was synthesized by priming off the poly-a tail and libraries were prepared for sequencing with the QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen A01173). Quality and concentration of libraries were determined with the Agilent Tapestation. Libraries were pooled in equimolar amounts and sequenced using single end 65 base pair reads on an Illumina HiSeq 4000 in the UCSF Center for Advanced Technologies.

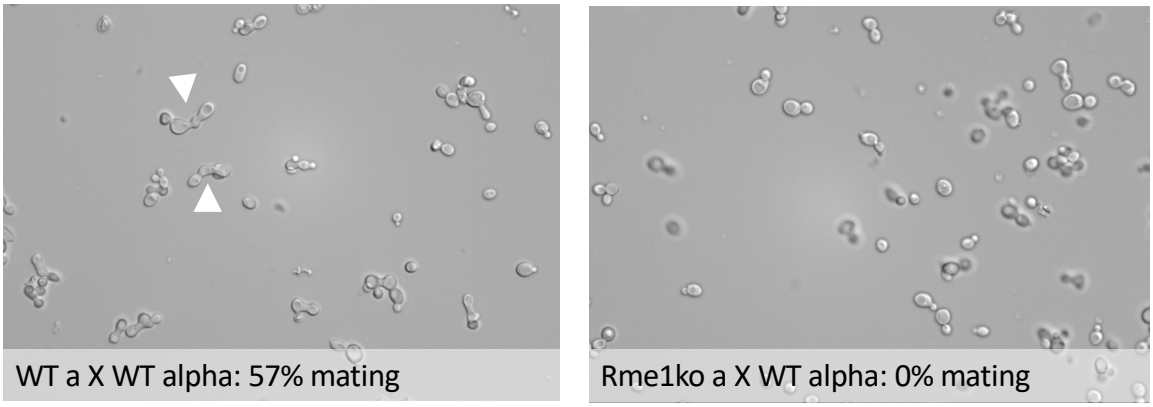
*3' sequencing analysis*

Quality of sequencing reads was determined using FastQC[25]. Filtering based on quality and trimming of reads was done using FastP [26]. Reads were aligned to the Wican1 genome assembly of *Wickerhamomyces Anomalus* NRRL Y-366-8[45] using STAR [28]. A table with counts assigned to genes was generated from the alignments using Rsubread[29]. This count table was then used to determine differentially expressed genes using DESeq2[30]. DESeq2 was run with default parameters, resulting in a list of genes that were differentially expressed in FDy19(**a** cell) when compared to FDy25(RME1 knock out **a** cell). Genes with a log2fold change of 1 or higher with an adjusted p value <0.1 were considered significantly differentially expressed.

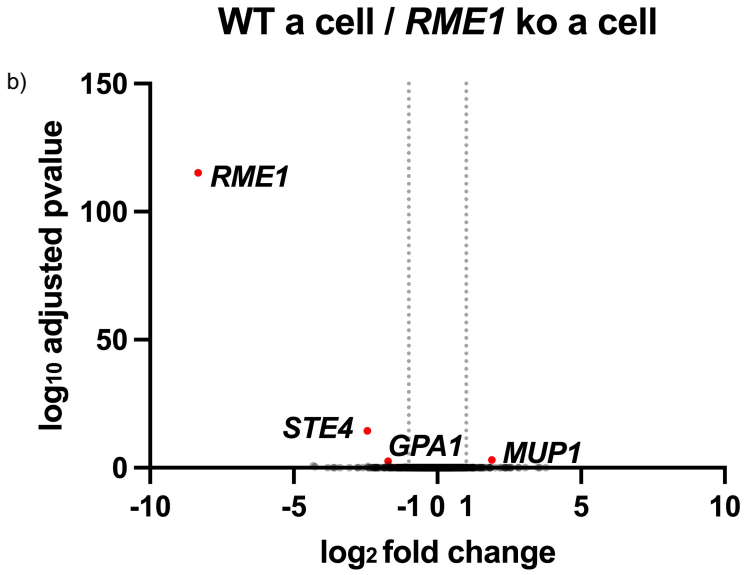
*Bioinformatics*

The sequences queried were extracted 600 bp upstream of the transcription start site in the four core conserved haploid specific genes(*GPA1, STE4, STE18, FAR1*) of *W.anomalus.* These sequences were input into MEME with all settings at default except that the program was allowed to search for any number of motif repetitions per query sequence, and look for up to 7 motifs rather than the standard 3[36]. Generated motifs were compared to a motif generated from the upstream regions of the 4 core haploid specific genes in *K. lactis.*

a)


WT a X WT alpha: 57% mating
Rme1ko a X WT alpha: 0% mating

b)


WT a cell / RME1 ko a cell
log10 adjusted pvalue
150
100
50
0
RME1
STE4
GPA1
MUP1
-10
-5
-1 0 1
5
10
log2 fold change

c)

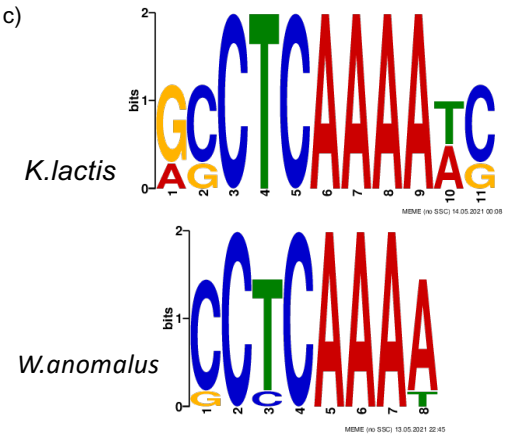

K.lactis
W.anomalus
bits

**Figure 2.1: Rme1 regulation of haploid specific genes in *W.anomalus***

a) The a cell with Rme1 deleted has a mating defect compared to wild type a cell when mated to a wild type α cell. WT a cells and *RME1* knock out a cells were mixed with a 200-fold excess of α cells. Cells were plated on selective media to kill the excess α cells, surviving cells would be either a or a/α cells. Mating was verified by PCR for the Mata and Matα locus and percentage mating was calculated based on the number of a cell (unsuccessful mating) compared to number of a/α cells (successful mating).
b) RNAseq of *RME1* knock out a cell compared to wild type a cell. All genes with a fold change of 2 fold or higher are indicated in red and labeled with the gene name. Vertical dotted lines on the x axis indicate the 2 fold cut off on the log2 scale. Only *GPA1* and *STE4* were significantly downregulated in the absence of *RME1* in the a cell. *Mup1* was significantly upregulated in the knock out compared with the wild type.
c) Bioinformatic analysis of upstream regions of four conserved haploid specific genes in *W.anomalus,* as well as their orthologs in *W.ciferri.* Extracted 600 bp upstream of the transcription start site in the four core conserved haploid specific genes(*GPA1, STE4, STE18, FAR1*) of *W.anomalus,* and also extracted the 600 bp upstream regions of their orthologs in the most closely related species with an available genome, *W. Ciferri.* Generated motifs were compared to a motif generated from the upstream regions of the 4 core haploid specific genes in *K. lactis.*

**Table 2.1: Effect of *RME1* knock out on *W. anomalus* mating**

Mating efficiency was calculated based on the fraction of Nat resistant colonies found to be a/$\alpha$ by PCR. For ease of comparison, this number was also converted to percent mating for both conditions.

| Strain | Percent mating | Mating efficiency |
|---|---|---|
| WT a cell x WT $\alpha$ cell | 108/188 (57%) | 0.574 |
| RME1 ko a cell x WT $\alpha$ cell | 0/279 (0%) | <0.00358 |

**Table 2.2: Strains used in this study**

| Strain | Species | Cell Type | Genotype | Source | Alternative names |
|---|---|---|---|---|---|
| FDFy19 | W. anomalus | a | prototroph | ATC via Candace Britton and Trevor Sorrells | CSBy5a,yTS19, ATCC 58044, NRRL Y-366-8, CBS 1984 |
| FDFy24 | W. anomalus | a/α | prototroph | Candace Britton | CSBy6, yTS145 |
| FDFy25 | W. anomalus | a | Rme1::Nat | Francesca Del Frate | None |

# References

1. Jarvela, A. M. C. & Hinman, V. F. Evolution of transcription factor function as a mechanism for changing metazoan developmental gene regulatory networks. *Evodevo* 6, 3 (2015).

2. Lynch, V. J. & Wagner, G. P. Resurrecting the Role of Transcription Factor Change in Developmental Evolution. *Evolution* 62, 2131–2154 (2008).

3. Sorrells, T. R. & Johnson, A. D. Making Sense of Transcription Networks. *Cell* 161, 714–723 (2015).

4. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8, 206–216 (2007).

5. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* 13, 59–69 (2012).

6. Stern, D. L. & Orgogozo, V. The Loci of Evolution: How Predictable is Genetic Evolution. *Evolution* 62, 2155–2177 (2008).

7. Baker, C. R., Booth, L. N., Sorrells, T. R. & Johnson, A. D. Protein Modularity, Cooperative Binding, and Hybrid Regulatory States Underlie Transcriptional Network Diversification. *Cell* 151, 80–95 (2012).

8. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* 134, 25–36 (2008).

9. Britton, C. S., Sorrells, T. R. & Johnson, A. D. Protein-coding changes preceded cis-regulatory gains in a newly evolved transcription circuit. *Science* 367, 96–100 (2020).

10. Herskowitz, I. A regulatory hierarchy for cell specialization in yeast. *Nature* 342, 749–757 (1989).

11. Tsong, A. E., Miller, M. G., Raisner, R. M. & Johnson, A. D. Evolution of a Combinatorial Transcriptional Circuit A Case Study in Yeasts. *Cell* 115, 389–399 (2003).

12. Srikantha, T. *et al.* TOS9 Regulates White-Opaque Switching in Candida albicans. *Eukaryot Cell* 5, 1674–1687 (2006).

13. Galgoczy, D. J. *et al.* Genomic dissection of the cell-type-specification circuit in Saccharomyces cerevisiae. *Proc National Acad Sci* 101, 18069–18074 (2004).

14. Booth, L. N., Tuch, B. B. & Johnson, A. D. Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature* 468, 959–963 (2010).

15. Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D. & Pabo, C. O. Crystal structure of a MATα2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* 67, 517–528 (1991).

16. Li, T., Stark, M. R., Johnson, A. D. & Wolberger, C. Crystal Structure of the MATa1/MATα2 Homeodomain Heterodimer Bound to DNA. *Science* 270, 262–269 (1995).

17. Shen, X.-X. *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* 175, 1533-1545.e20 (2018).

18. Taylor, J. W. & Berbee, M. L. Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia* 98, 838–849 (2006).

19. Tan, S. & Richmond, T. J. Crystal structure of the yeast MATα2/MCM1/DNA ternary complex. *Nature* 391, 660–666 (1998).

20. Sorrells, T. R., Booth, L. N., Tuch, B. B. & Johnson, A. D. Intersecting transcription networks constrain gene regulatory evolution. *Nature* 523, 361–365 (2015).

21. Faber, K. N., Haima, P., Harder, W., Veenhuis, M. & AB, G. Highly-efficient electrotransformation of the yeast Hansenula polymorpha. *Curr Genet* 25, 305–310 (1994).

22. Gojkovic, Z., Jahnke, K., Schnackerz, K. D. & Piškur, J. PYD2 encodes 5,6-dihydropyrimidine amidohydrolase, which participates in a novel fungal catabolic pathway1 1Edited by J. Karn. *J Mol Biol* 295, 1073–1087 (2000).

23. Hoffman, C. S. & Winston, F. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformaion of Escherichia coli. *Gene* 57, 267–272 (1987).

24. Nocedal, I., Mancera, E. & Johnson, A. D. Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. *Elife* 6, e23250 (2017).

25. Andrews. FastQC:  A Quality Control Tool for High Throughput Sequence Data. (2010).

26. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890 (2018).

27. Consortium, T. G. *et al.* Comparative genomics of protoploid Saccharomycetaceae. *Genome Res* 19, 1696–1709 (2009).

28. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).

29. Chisanga, D., Liao, Y. & Shi, W. Impact of gene annotation choice on the quantification of RNA-seq data. *Bmc Bioinformatics* 23, 107 (2022).

30. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014).

31. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44, W160–W165 (2016).

32. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 10, giab008 (2021).

33. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* 29, 24–26 (2011).

34. Freese, N. H., Norris, D. C. & Loraine, A. E. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics* 32, 2089–2095 (2016).

35. Gaspar, J. M. Improved peak-calling with MACS2. *Biorxiv* 496521 (2018) doi:10.1101/496521.

36. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intelligent Syst Mol Biology* 2, 28–36 (1994).

37. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 50, gkab1113- (2021).

38. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res* 40, D700–D705 (2012).

39. Engel, S. R. *et al.* The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3 Genes Genomes Genetics* 4, 389–398 (2013).

40. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018 (2011).

41. Gehring, W. J. & Ikeo, K. Pax 6: mastering eye morphogenesis and eye evolution. *Trends Genet* 15, 371–377 (1999).

42. van Werven, F. J. *et al.* Transcription of Two Long Noncoding RNAs Mediates Mating-Type Control of Gametogenesis in Budding Yeast. *Cell* 150, 1170–1181 (2012).

43. Zhou, W., Dorrity, M. W., Bubb, K. L., Queitsch, C. & Fields, S. Binding and Regulation of Transcription by Yeast Ste12 Variants To Drive Mating and Invasion Phenotypes. *Genetics* 214, 397–407 (2020).

44. Schorsch, C., Köhler, T. & Boles, E. Knockout of the DNA ligase IV homolog gene in the sphingoid base producing yeast Pichia ciferrii significantly increases gene targeting efficiency. *Curr Genet* 55, 381–389 (2009).

45. Riley, R. *et al.* Comparative genomics of biotechnologically important yeasts. *Proc National Acad Sci* 113, 9882–9887 (2016).

**Publishing Agreement**

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:

781CD4AB866F47F...

Author Signature

12/11/2022

Date